NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

RICARDO PEREIRA COELHO

Licenciado em Matemática

MÉTODOS DE APRENDIZAGEM AUTOMÁTICA PARA PREDIÇÃO DE EVENTOS DE COVID-19

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Universidade NOVA de Lisboa
Setembro, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

MÉTODOS DE APRENDIZAGEM AUTOMÁTICA PARA PREDIÇÃO DE EVENTOS DE COVID-19

RICARDO PEREIRA COELHO

Licenciado em Matemática

Orientadora: Isabel Cristina Maciel Natário

Professora Associada da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Júri

Presidente: Marta Cristina Vieira Faias Mateus

Professora Associada da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Arguente: Regina Maria Baltazar Bispo

Professora Auxiliar da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Orientadora: Isabel Cristina Maciel Natário

Professora Associada da Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Universidade NOVA de Lisboa
Setembro, 2022

Métodos de Aprendizagem Automática para Predição de Eventos de COVID-19

Este documento foi gerado com o processador (pdf/Xe/Lua)LaTeX e o modelo NOVAthesis (v6.9.14) [1].

Agradecimentos

Gostaria de agradecer à minha orientadora, Profª. Isabel Natário, pela proposta do tema e por todo o apoio e ajuda durante todo o período do desenvolvimento desta tese.

Também gostaria de expressar os meus sinceros agradecimentos à Bolsa de Investigação para Licenciado, no âmbito da Unidade de I&D "Centro de Matemática e Aplicaçlões" (CMA), na área cientifica da Matemática (Estatística), pelo o apoio financeiro para o desenvolvimento do presente estudo.

Para concluir, não posso deixar de agradecer à minha família, amigos e a todos os professores, em especial à minha professora de Matemática A do secundário, que intencionalmente ou sem querer, ajudaram-me a moldar a pessoa dedicada que sou hoje.

«You cannot teach a man anything; you can only help him discover it in himself.» (Galileo)

Resumo

O aparecimento da COVID-19, uma doença viral e infeciosa que apresenta uma alta taxa de propagação, resultou em grandes pressões sobre os sistemas de saúde, esgotando médicos, recursos e provocando a exaustão dos profissionais de saúde da linha da frente. Devido às suas consequências devastadores e ao seu desconhecimento, começaram-se a recolher e analisar dados dos pacientes que tinham testado positivo à COVID-19, tanto dos sintomas, como das suas COVID-19, variante do vírus, dados demográficos, dados geográficos do local de infeção, assim como as datas de infeção e de início de sintomas, para se tentar ajudar as unidades hospitalares a selecionarem os pacientes que deveriam ter um atendimento prioritário. Uma vez que estes conjuntos de dados rapidamente se tornaram gigantescos, utilizaram-se variados métodos, também de Aprendizagem Automática. Estes métodos, têm como principal objetivo a perceção de padrões e o relacionamento de dados, de uma forma automática a partir dos mesmos. Os métodos de Aprendizagem Automática, fazem uso de métodos estatísticos e probabilísticos de forma a criar algoritmos que consigam aprender diretamente de uma parte de dados, que são usados para treinar e avaliar os modelos conforme o seu desempenho preditivo. Estes métodos têm tido uma ampla utilização para problemas de classificação e em especial foram muito empregues no contexto da COVID-19, uma vez que desta doença nova não se sabia praticamente nada sobre a sua evolução ao longo do tempo, e aplicando-se diversos modelos de Aprendizagem Automática aos dados de saúde recolhidos, tentou-se fazer previsões para o número de casos diários, assim como também para a identificação dos pacientes, consoante os sintomas e as comorbilidades que estes apresentavam, que fossem de atendimento mais premente nos hospitais. Assim, o presente trabalho tem como objetivo a aplicação de diversos métodos de Aprendizagem Automática, entre os quais Regressão Logística, Modelos Aditivos Generalizados, Árvores de Classificação, Florestas Aleatórias e Redes Neuronais, em dados Portugueses de saúde de COVID-19, de forma a avaliar cada um destes modelos em termos preditivos e também tem como objetivo saber quais os sintomas e as doenças que estão mais relacionadas com a morte por COVID-19.

Palavras-chave: COVID-19, Aprendizagem Automática, Regressão Logística, Modelo Aditivo Generalizado, LASSO de Grupo, Árvores de Classificação, Florestas Aleatórias, Redes Neuronais

Abstract

The emergence of COVID-19, a viral and infectious disease that has a high rate of spread, has resulted in great pressures on health systems, depleting medical doctors, resources and causing the exhaustion of frontline health professionals. Due to its devastating consequences and lack of knowledge about it, data began to be collected and analyzed from patients who had tested positive for COVID-19, both on symptom data, comorbidities data, virus variant, demographic data, geographic data the site of infection, as well as the dates of infection and onset of symptoms, to try to help hospital units to select patients who should receive priority care. Once these datasets quickly became huge, they begam to be analysed with several methods, including various Machine Learning methods. These methods have as main objective the revelation of patterns and data relationships, automatically. Machine learning methods make use of statistics and probability in order to create the learning methods that are directly used to train and evaluate models according to their predictive performance. Machine Learning methods have a wide use for classification and especially have been much used in COVID-19 context, since this disease was new and practically nothing was known about its evolution in time. Several Machine Learning models were applied to the collected health data, in order to try to make predictions for the number of daily cases, as well as for the identification of most potentially problematic patients, depending on their symptoms and comorbilities, so that hospitals would pay more attention to those. This work have the objective of applying several Machine Learning methods, including Logistic Regression, Generalized Additive Models, Group LASSO, Classification Trees, Random Forests and Neural Networks, to the Portuguese COVID-19 data, in order to evaluate each one of these models in predictive terms and also have the objective to know which symptoms and comorbidities are most related with mortality from COVID-19.

Keywords: COVID-19, Machine Learning, Logistic Regression, Generalized Additive Model, Group LASSO, Classification Tree, Random Forest, Neural Networks

Índice

Índice de Figuras

Índice de Tabelas

Siglas

AIC	Critério de informação de *Akaike*
BIC	Critério de informação de *Bayesiano*
MSE	*Mean Squared Error*
OR	*Odds-Ratio*
RSS	*Residual Sum of Squares*
AUC	*Area under the ROC Curve*
CART	*Classification and Regression Trees*
COVID-19	*Coronavirus Disease* 2019
KNN	K-*nearest neighbors*
LASSO	*Least Absolute Shrinkage and Selection Operator*
OMS	Organização Mundial de Saúde
OOB	*Out-of-Bag*
ROC	*Receiver Operating Characteristic*
SARS-CoV-2	*Severe Acute Respiratory Syndrome Coronavirus* 2
SINAVE	Sistema Nacional de Vigilância Epidemiológica
SMOTE	*Synthetic Minority Oversampling Technique*
SVM	*Support Vector Machines*
XGBoost	*Extreme Gradient de Boosting*

1 Introdução

Em dezembro de 2019, surgiu em Wuhan, República da China, um surto de um síndrome respiratório agudo severo, provocado por uma nova forma de coronavírus, chamada de SARS-CoV-2. Em janeiro de 2020 a Organização Mundial de Saúde (OMS), declarou o surto do novo coronavírus uma emergência de saúde pública de interesse internacional. Em fevereiro de 2020 a OMS, selecionou um nome oficial para a doença infeciosa causada pelo novo coronavírus, denominando-a doença de coronavírus 2019, mais abreviadamente por COVID-19, sendo que em março de 2020 a OMS, declarou a pandemia de COVID-19 [2].

A COVID-19 é uma doença viral e infeciosa, que apresenta uma alta taxa de propagação, sendo que se espalhou rapidamente por todo o mundo, causando um grave problema de saúde pública global. Após um mês da OMS ter declarado a pandemia de COVID-19, já se contabilizavam mais de 1.3 milhões de casos em todo mundo e mais de 70000 mortes, a sua maioria entre a população envelhecida da Europa, onde o excesso de mortalidade por todas as causas subia acima do esperado num grande número de países [3]. A rápida disseminação da COVID-19 resultou em grandes pressões sobre os sistemas de saúde, esgotando médicos, recursos e provocando a exaustão dos profissionais de saúde da linha da frente.

Com consequências tão devastadoras e com tanta falta de conhecimento deste novo coronavírus, sentiu-se a necessidade de começar a recolher e analisar dados dos pacientes que tinham testado positivo à COVID-19, tanto dados dos sintomas, como das suas comorbilidades, a variante do vírus, dados demográficos, dados geográficos do local de infeção, assim como as datas de infeção e de início de sintomas. Estes dados foram recolhidos para se tentar perceber quais deveriam ser as pessoas que deveriam ter um atendimento prioritário, nas unidades hospitalares, já que a rápida disseminação da COVID-19 resultou em grandes pressões sobre todos os sistemas de saúde e no aumento da mortalidade e internamentos. Para se tratarem estes conjuntos de dados, que rapidamente se tornaram gigantescos utilizaram-se variados métodos, também de Aprendizagem Automática.

Os métodos de Aprendizagem Automática têm como principal objetivo a perceção de

padrões e o estabelecimento de relações entre dados, de uma forma automática, a partir destes. Os métodos de Aprendizagem Automática, fazem uso de métodos estatísticos e probabilísticos de forma a criar algoritmos que consigam aprender diretamente de uma parte de dados, que são usados para treinar e avaliar os modelos conforme o seu desempenho preditivo. Os métodos de Aprendizagem Automática têm tido uma ampla utilização para problemas de classificação e previsão, especialmente para aplicações de mercados financeiros e de saúde, sendo que estes métodos também foram muito empregues no contexto do COVID-19, uma vez que esta doença era nova e não se sabia praticamente nada sobre a sua evolução ao longo do tempo. Assim tem-se aplicado diversos modelos de Aprendizagem Automática aos dados de saúde recolhidos, de forma a tentar fazer previsões para o número de casos diários, assim como também para a identificação dos pacientes que causassem mais cuidados, consoante os sintomas e as comorbilidades que estes apresentassem de forma a que os hospitais lhes dessem uma maior atenção.

1.1 Motivação

Este trabalho aborda a questão da aplicação de métodos de Aprendizagem Automática aplicados a dados Portugueses de COVID-19. Ao abrigo do disposto no artigo 39.º do Decreto n.º 2-B/2020, de 2 de abril, a Direção-Geral da Saúde disponibilizou à Comunidade Científica e Tecnológica Portuguesa o acesso a dados de saúde pública do Sistema Nacional de Vigilância Epidemiológica (SINAVE) relativos a doentes infetados pelo novo coronavírus SARS-CoV-2. O presente estudo enquadra-se neste âmbito e alinha-se com o compromisso de análise destes dados em contexto de modelação estatística, para estimação e previsão da morbilidade e mortalidade por COVID-19 na população Portuguesa, assim como a determinação de fatores de risco inerentes.

1.2 Objetivo

O presente trabalho tem como objetivo a aplicação de diversos métodos de Aprendizagem Automática, entre os quais a Regressão Logística, modelos Aditivos Generalizados, Árvores de Classificação, Florestas Aleatórias e Redes Neuronais, a dados Portugueses de saúde de COVID-19, de forma a avaliar cada um destes modelos em termos preditivos e também tem como objetivo saber quais os sintomas e quais as comorbilidades que estão mais relacionadas com a morte por COVID-19. A base de dados disponível é constituída por 52 variáveis, sendo que estas se podem dividir em 6 grupos de dados: dados demográficos, dados geográficos, dados referentes a sintomas, dados referentes a comorbilidades, dados referentes a datas e o tipo de vírus. Estes dados são referentes a indivíduos que testaram positivo à COVID-19, em Portugal, entre março de 2020 e julho de 2021, correspondendo a um total de 909720 observações. O fenómeno a modelar é

representado por uma variável dicotómica, que indica o desfecho do caso, traduzindo se o indivíduo falece ou não.

1.3 Estrutura

Para além deste capítulo introdutório, a presente tese está organizada em outros 4 capítulos. O capítulo 2 é referente ao Estado da Arte, onde se abordam um amplo número de estudos na área deste trabalho, nos quais vários autores propõem métodos de Aprendizagem Automática para a previsão do risco de mortalidade por COVID-19. Apresentam-se os resultados conseguidos em cada um dos estudos revistos, assim como também as limitações de cada um. No capítulo 3 apresentamos os diferentes algoritmos de Aprendizagem Automática que são utilizados neste estudo, incluindo a Regressão Logística, Árvores de Decisão, modelos Aditivos Generalizados, Florestas Aleatórias, Redes Neuronais e LASSO de Grupo. Adicionalmente, neste capítulo ainda se abordam métodos para a redução do número de variáveis numa base dados, tais como o método *stepwise* ou métodos de contração, e também se abordam várias medidas de qualidade preditiva, assim como vários métodos para o balanceamento de uma base de dados. O capítulo 4 é onde se descreve a base de dados reais do número de indivíduos que testaram positivo à COVID-19, em Portugal, e onde implementamos os métodos descritos no capítulo anterior para o conjunto de dados de grande dimensão contendo dados demográficos, geográficos, sintomas, comorbilidades, datas e o tipo de vírus. Por fim, no capítulo 5, discutimos os resultados entre os diferentes modelos de Aprendizagem Automática utilizados, identificando qual tem o melhor desempenho preditivo e quais os sintomas e as comorbilidades que mais estão associadas à mortalidade por COVID-19.

2

Estado da Arte

Atendendo que o objetivo do presente trabalho é a aplicação de diversos métodos de Aprendizagem Automática, entre os quais a Regressão Logística, Modelos Aditivos Generalizados, Árvores de Classificação, Florestas Aleatórias e Redes Neuronais, a dados Portugueses de saúde em COVID-19, por forma a avaliar cada um destes modelos em termos preditivos, e também o objetivo de saber quais os sintomas e quais as comorbilidades que estão mais relacionadas com a morte por COVID-19, nesta secção, descrevem-se alguns estudos realizados neste contexto. Diferentes autores propuseram métodos de Aprendizagem Automática e outros métodos em problemas de previsão da mortalidade associada à COVID-19, assim como também de determinar quais os fatores associados a sintomas e comorbilidades mais relevantes, de modo a ajudar à tomada de decisões nas unidades de saúde hospitalares, selecionando quais os pacientes que devem de ter um atendimento prioritário, uma vez que sendo a COVID-19 uma doença com uma alta taxa de propagação, que no início desta pandemia muitas das unidades hospitalares ficaram completamente esgotadas. Neste capítulo também se apontam as limitações de cada um dos artigos revistos.

Mohammad Pourhomayoun e Mahdi Shakibi [2], apresentam um estudo com a construção de um modelo preditivo baseado em Inteligência Artificial e Aprendizagem de Automática, com o objetivo de determinar o risco de mortalidade e prever o risco de mortalidade de pacientes com COVID-19, onde o modelo desenvolvido pode ajudar os hospitais, assim como os sistemas de saúde a decidir quais os utentes que necessitam de receber um auxílio em primeiro lugar, ou seja, quem tem uma maior necessidade de ser hospitalizado primeiro, e assim fazer uma seleção dos pacientes quando os sistemas de saúde tiverem sobrecarregados pela superlotação tentando eliminar os atrasos na prestação dos cuidados necessários. O modelo construído prevê os riscos de mortalidade com base nas condições fisiológicas (comorbilidades), sintomas e informação demográfica de cada um dos pacientes. Na construção do modelo, os dois autores utilizam duas fases, o pré-processamento dos dados e a modelação. No pré-processamento dos dados tratam dos valores omissos, eliminam elementos dos dados redundantes e selecionam as variáveis mais informativas de entre todas. Após este passo os autores utilizam algoritmos

de Aprendizagem Automática para desenvolver um modelo preditivo para classificar os dados, calcular a probabilidade e o risco de mortalidade.

Para concretizar o objetivo, este estudo utilizou uma base de dados com 2670000 casos confirmados de COVID-19, de 146 países diferentes, incluindo 307382 observações contendo o sexo dos pacientes com uma média de idades de 44.75 anos. A base de dados original continha 32 variáveis de cada um dos pacientes, incluindo dados demográficos e fisiológicos. Na fase de limpeza dos dados, estes dois autores removeram as variáveis inúteis e redundantes. Dado que a base de dados tem valores omissos, os autores utilizaram técnicas de imputação de dados, tais como a substituição dos valores omissos pela média, mediana, moda ou a utilização da técnica do K-*nearest neighbors* (KNN), para lidar com os valores omissos. Os autores garantiram ainda que a base de dados era balanceada, isto é, garantiram que o número de óbitos era igual ao número de pacientes que sobreviveram. Este conjunto foi criado para treinar e testar o modelo, onde os dados do conjunto treino foram selecionados aleatoriamente e completamente separados dos dados treino.

Os autores consultaram uma equipa de médicos sobre a base de dados restringida a que chegaram, de modo a garantir que as variáveis que tinham selecionado eram as mais relevantes de entre as disponíveis da base de dados original. Assim, ao todo foram extraídas 112 variáveis da base de dados original, referentes aos sintomas, às notas médicas sobre o estado de saúde do paciente, dados demográficos e também a informações fisiológicas. Feito isto, os autores prosseguiram com a seleção das variáveis, com o objetivo de encontrar quais eram as variáveis mais informativas e eliminar as variáveis redundantes, de forma a reduzir a dimensionalidade e a complexidade do modelo. Para isto, utilizaram métodos de filtro univariados e multivariados, incluindo o coeficiente de correlação, entropia, teste do qui-quadrado, e o *score* de Fisher, assim como o também métodos de *wrapper*, de modo a elaborar uma ordenação de variáveis e selecionar o seu melhor subconjunto. Ficaram apenas com 57 variáveis das 112 variáveis anteriormente selecionadas, categorizadas em 3 grupos, tais como sintomas, comorbilidades e dados demográficos.

Seguidamente, os autores utilizaram diferentes métodos de Aprendizagem Automática para construirem o referido modelo preditivo. Os diferentes métodos utilizados foram a Regressão Logística, as Árvores de Decisão, as Florestas Aleatórias, as Redes Neuronais, o KNN e o *Support Vector Machines* (SVM).

Em todos os modelos elaborados, os autores utilizaram a validação cruzada para avaliar os modelos desenvolvidos. Na validação cruzada os autores também fazem a seleção das variáveis apenas nos dados de treino, para confirmar se as variáveis explicativas selecionadas através da validação cruzada correspondiam às variáveis explicativas selecionadas através do modelo de Aprendizagem Automática ajustados.

Para avaliar o desempenho dos modelos os autores utilizaram várias métricas, entre as quais a *accuracy*, para comparar o desempenho de todos os métodos de Aprendizagem Automática utilizados. Também para cada um dos modelos referidos anteriormente o autores geraram a curva ROC (*Receiver Operating Characteristic*) e de seguida calcularam

a área sob essa curva (AUC-*Area under the ROC Curve*), para avaliar o modelo desenvolvido, assim como também utilizaram a matriz de confusão para avaliarem os modelos desenvolvidos.

Os resultados obtidos indicam que o modelo da Redes Neuronais é o melhor modelo no que diz respeito à medida *accuracy*. Os autores ainda concluem que os modelos desenvolvidos são capazes de prever com elevada precisão o risco de mortalidade em pacientes com COVID-19, em que estes modelos levam em conta as comorbilidades, os sintomas e também as informações demográficas dos pacientes.

Este estudo, apresenta uma limitação, que reside no facto das variáveis explicativas terem sido selecionadas após o balanceamento da base de dados, sendo que desta forma se pode estar a descartar variáveis explicativas que sejam importantes para a modelação do risco de mortalidade. Por outro lado, os autores também levaram em conta em todos os modelos simultaneamente os dados referentes aos sintomas e às comorbilidades, o que pode significar problemas de multicolinearidade, já que podem existir sintomas que apresentem uma grande relação com as comorbilidades.

Noutro estudo desenvolvido por Khadijeh Moulaei et al. [4], os autores têm como objetivo a comparação de vários algoritmos de Aprendizagem Automática, para prever a mortalidade por COVID-19, de modo a escolher o algoritmo que apresente o melhor desempenho preditivo, para que este sirva como uma ferramenta preditiva para a ajuda de tomada de decisão dos pacientes que devem de receber assistência médica em primeiro lugar, uma vez que os pacientes hospitalizados com a infeção de COVID-19 estão sempre em risco de morte. Para a construção dos demais modelos foram utilizados dados dos pacientes hospitalizados no primeiro momento de admissão num certo hospital. Ao todo a base de dados é constituída por 1500 observações. Neste estudo a variável resposta é a mortalidade hospitalar por COVID-19. Os algoritmos utilizados pelos autores para modelar os dados foram as Florestas Aleatórias, a Regressão Logística, as Árvores de Decisão, o classificador *Naive Bayes*, o KNN e o algoritmo XGBoost (*Extreme Gradient de Boosting*). Para treinar e testar os diversos modelos, os autores utilizaram o método da validação cruzada com 10 grupos, onde as métricas utilizadas para a avaliação de cada um dos modelos foram a *accuracy*, *especificidade*, *precisão*, *sensibilidade* e a área sob a curva ROC.

Antes dos autores treinaram os respetivos modelos, balancearam a base de dados, uma vez que esta base de dados é desequilibrada/desbalanceada e os autores referem que uma das principais barreiras dos métodos de Aprendizagem Automática é o problema dos dados desequilibrados, fazendo com que os modelos de treino, apresentem resultados tendenciosos para a classe maioritária. Os autores balancearam a base de dados através da técnica sobreamostragem minoritária, mais conhecida como a técnica SMOTE (*Synthetic Minority Oversampling Technique*). Depois de balanceada a base de dados, os autores utilizaram a técnica da seleção de variáveis explicativas, para a seleção do melhor subconjunto de variáveis explicativas que entrariam nos modelos de treino. Neste caso, estes autores utilizaram a técnica da Avaliação do Atributo de rácio de Ganho de

Informação, (*Information Gain Ratio Attribute evaluation*). Após a utilização desta técnica das 54 variáveis explicativas da base de dados inicial, apenas se ficou com 38 variáveis, onde o autor refere que a dispneia e a contagem de plaquetas, são as variáveis com maior importância e menor importância ara prever a mortalidade por COVID-19, respetivamente. As variáveis selecionadas podem ser divididas em 6 categorias, variáveis demográficas, fatores de risco, manifestações clínicas, teste laboratoriais e planos terapêuticos. Posto isto, os autores treinaram e avaliaram os modelos em termos preditivos, através da validação cruzada, donde verificam que o melhor modelo em termos das métricas *accuracy*, *precisão*, *sensibilidade*, *especificidade* e área sob a curva ROC é a Floresta Aleatória. No entanto, os resultados dos modelos da Florestas Aleatórias, XGBoost e KNN, têm um bom desempenho de previsão, uma vez que a área sob a curva ROC destes modelos encontram-se acima de 0.96 e a eficiência destes é melhor que a do modelo de Regressão Logística quando treinado usando os mesmos parâmetros. Assim, o modelo das Florestas Aleatórias proposto pode ser adequadamente usado para prever o risco de mortalidade de pacientes hospitalizados com COVID-19 e maximizar o uso de recursos hospitalares restritos, onde este modelo poderia identificar automaticamente pacientes de alto risco já no momento da admissão ou durante o internamento. No entanto, este estudo apresenta algumas limitações entre as quais as observações com valores omissos, que foram eliminadas. Outra limitação é o facto da base de dados inicial ser desbalanceada, que foi contornada usando a técnica SMOTE para balancear a base de dados, pelo que quando os autores aplicam o método para a seleção de variáveis à base de dados balanceada, podem estar a deitar fora variáveis que sejam importantes para a modelação da morte por COVID-19, em relação às variáveis que obteriam se utilizassem a base de dados inicial sem as observações com valores omissos.

Cindy Feng, George Kephart e Elizabeth Juarez-Colunga [5], apresentam um trabalho que tem o mesmo objetivo que os anteriormente descritos, ou seja, a avaliação da precisão preditiva de diversos modelos de Aprendizagem Automática, entre os quais as Árvores de Classificação, Floresta Aleatória, Regressão Logística, Modelo Aditivo Generalizado, Análise Linear Discriminante e o XGBoost, construídos para prever o risco de mortalidade por COVID-19, onde os autores estão interessados em selecionar o melhor modelo preditivo para prever com precisão o risco de mortalidade entre os indivíduos infetados com COVID-19, de modo a priorizar o atendimento médico, mitigando a carga dos sistemas de saúde.

A base de dados com que estes autores trabalharam é constituída pelos casos confirmados de COVID-19, entre 21 de março de 2020 e 10 de dezembro de 2020, na cidade de Toronto no Canadá. Esta base de dados é constituída por 49216 observações de casos positivos de COVID-19, incluindo informação sobre a idade, género, se esteve sempre hospitalizado, se esteve sempre nos cuidados intensivos, a densidade populacional, temperatura média diária e renda média.

Os modelos preditivos utilizados foram a Regressão Logística, Modelo Aditivo Generalizado, a Análise Linear Discriminante, a Árvore de Decisão, a Floresta Aleatória e o XGBoost. No entanto, os autores no modelo de Regressão Logística consideram dois

tipos de modelos, o primeiro modelo é constituído por todas as variáveis explicativas, onde não utilizaram nenhum método para a redução do número de variáveis a considerar no modelo, e o segundo modelo é constituído por todas as variáveis explicativas e por todas as correspondentes interações de ordem dois, onde utilizaram o *Least Absolute Shrinkage and Selection Operator* (LASSO) para excluir as variáveis explicativas "desnecessárias", encolhendo os coeficientes destas exatamente para zero, produzindo assim um modelo mais parcimonioso. O parâmetro de regularização na Regressão de LASSO, foi escolhido através da minimização do erro de classificação incorreta, em termos da área sob a curva ROC, através da validação cruzada com 10 grupos. O Modelo Aditivo Generalizado é uma técnica de regressão não paramétrica que oferece uma maior flexibilidade na modelagem de efeitos não lineares das covariáveis, com recurso aos *splines* suavizadores. Neste caso, os autores estimaram o parâmetro de suavização através do método da validação cruzada. As Árvores de Decisão são um método popular alternativo à Regressão Logística, onde neste caso a profundidade máxima da Árvore de Classificação foi de que qualquer nó da árvore final tinha de ter pelo menos 100 observações, onde de seguida foi utilizada a poda da árvore obtida de forma a evitar o sobreajustamento aos dados e por outro lado para remover qualquer divisão que não melhorasse o ajustamento. As Florestas Aleatórias foram utilizadas pelos autores, uma vez que as árvores de Classificação tendem a sobreajustar o subconjunto de dados de treino, ou seja, as Árvores de Classificação tendem a ajustar-se muito bem ao conjunto de dados de treino, mas quando se utiliza o conjunto de dados de teste a árvore não é eficaz a fazer as previsões. Neste estudo, os autores utilizaram 1000 árvores para a construção das Florestas Aleatórias.

Neste estudo, os autores utilizaram a validação repetida de amostras divididas, para compararem a precisão preditiva de cada um dos métodos referidos anteriormente. Os dados foram divididos aleatoriamente em 80% para o conjunto de dados de treino e 20% para o conjunto de dados de teste, onde cada um dos modelos foi treinado com o conjunto de dados de treino e as previsões foram obtidas a partir dos dados de teste usando o modelo ajustado através da base de dados de treino. Este processo foi repetido 200 vezes, ou seja, cada um dos modelos preditivos foi ajustado usando o conjunto de dados de treino, sendo que o modelo resultante foi então usado para prever o risco de mortalidade, através da base de dados de teste. Estes resultados foram então resumidos nos 200 conjuntos de dados de teste. A validação repetida de amostras divididas avalia a robustez dos resultados e é menos provável de ser impactada por observações influentes em apenas algumas amostras de teste.

No entanto, para a validação da previsão, os autores também validaram os modelos com base nas previsões de k-passos à frente dos últimos k dias do período de observação $k = 7, 8, \ldots, 30$, ou seja, para cada uma das previsões de k-passos à frente, o conjunto de dados de treino incluía todos os dados anteriores aos k dias a serem previstos. Este passo, foi efetuado para cada um dos modelos, que foram ajustados ao conjunto de dados de treino sendo as previsões obtidas para os últimos k dias do conjunto de dados de teste.

A medida de desempenho dos modelos utilizados foi a métrica da área sob a Curva

ROC, onde valores elevados desta medida indicam uma melhor discriminação do modelo. No entanto, uma vez que esta medida não leva em conta a calibração, ou seja, a magnitude da discordância entre as respostas observadas e previstas, os autores utilizaram a pontuação de *Brier* de modo a quantificar o quão perto as previsões estão dos resultados reais, sendo que valores baixos de *Brier* indicam uma maior precisão do modelo.

Os autores chegam à conclusão que os modelos preditivos baseados em métodos de Aprendizagem Automática aplicado aos dados disponíveis, podem fornecer informações importantes para ajudarem o planeamento de recursos dos serviços de saúde, para lidarem melhor com a pandemia de COVID-19. Os autores concluem ainda que o uso de variáveis explicativas como a idade, variáveis explicativas hospitalares usadas para a COVID-19, sexo do utente, as suas características económicas e a densidade populacional nos modelos de Aprendizagem Automática, têm uma ótima capacidade de previsão do risco de mortalidade para a COVID-19. Por outro lado, os autores também concluem que dos modelos de Aprendizagem Automática baseados nas árvores, o modelo XGBoost é o melhor modelo, enquanto que os modelos baseados na regressão todos eles têm um bom poder preditivo.

No entanto, os autores também referem que o seu estudo tem limitações, sendo que uma das suas maiores limitações é a indisponibilidade de dados sobre as características clínicas dos pacientes, como por exemplo as comorbilidades. No entanto, apesar destas limitações os autores consideram que é possível prever o risco de mortalidade com um alto poder preditivo.

O estudo desenvolvido pelos autores Devin Incerti et al [6], tem como objetivo o desenvolvimento de um modelo de prognóstico para identificar e quantificar os fatores de risco para mortalidade, entre os pacientes internados em hospitais com COVID-19, nos Estados Unidos da América. Neste estudo, os autores utilizam uma base de dados constituída por 17086 pacientes hospitalizados com COVID-19 entre 20 de fevereiro de 2020 e 5 de junho de 2020, dados que derivam de todas as redes hospitalares do país. Esta base de dados contém também dados demográficos dos pacientes, dados das comorbilidades, dados de sinais vitais, dados de resultados laboratoriais, e uma variável de tempo, que mede o número de dias entre a data de confirmação do caso do paciente e a data do primeiro caso nos dados, em que esta variável serve para capturar a tendência ao longo do tempo. As variáveis que tinham um alto valor de número de observações omissas foram removidas. Por outro lado também se fez imputação para dados omissos, através da imputação multivariada pelas equações de *chained*.

O modelo desenvolvido pelos autores é o modelo de Regressão Logística, onde estes autores utilizando o LASSO de Grupo para primeiro selecionarem as variáveis explicativas a incluírem no modelo inicial, de modo a proteger o modelo construído do sobreajustamento. No entanto, segundo os autores, este método apenas removeu duas variáveis explicativas de todas as consideradas anteriormente, donde este modelo foi denominado por modelo completo. No entanto, para efeitos de comparação, os autores encaixaram mais quatro modelos parcimoniosos em que cada um destes era formado pelas seguintes

variáveis explicativas, um modelo que apenas tinha a variável explicativa idade, outro que apenas tinha as variáveis referentes às comorbilidades, outro que tinha todas as variáveis explicativas correspondentes aos dados demográficos com a variável data e por último outro modelo com todas as variáveis explicativas correspondentes aos dados demográficos, aos dados das comorbilidades e à variável data. As relações não lineares entre a variável resposta mortalidade e as variáveis explicativas contínuas foram modeladas utilizando *splines* cúbicos.

Para validar os modelos, os autores dividiram aleatoriamente a base de dados em dados de treino e em dados de teste, onde utilizaram 80% dos dados para treino e os restantes 20% para teste. A performance do modelo foi avaliada através da área sob a curva ROC e a pontuação de *Brier*. Por outro lado, os autores também avaliaram a calibração, para comparar as probabilidades previstas com as probabilidades reais.

Treinando e testando cada um dos modelos, os autores obtém que, em termos da métrica AUC, o modelo completo é o que apresenta um maior valor, seguindo-se o modelo somente com a variável idade, seguindo-se o modelo com as variáveis demográficas, comorbilidades e data, seguindo-se o modelo com todas as variáveis demográficas e data, enquanto que o modelo que teve o menor valor foi o modelo que apenas continha as variáveis referentes às comorbilidades. Já em termos da pontuação de *Brier*, o melhor modelo é o modelo completo, seguindo-se o modelo com as variáveis demográficas, comorbilidades e data, sendo que o pior modelo é mais uma vez o modelo que apenas continha as variáveis referentes às comorbilidades.

Do exposto, os autores concluem que o modelo mais parcimonioso, que é o modelo que apenas inclui a variável idade é altamente preditivo, ou sejam a idade é quase tão prognóstico quanto todas as outras informações sobre a demografia e as comorbilidades. No entanto, os autores referem que isto não significa que a idade por si só seja suficiente para a previsão, mas saber simplesmente a idade de um paciente é muito informativo. Os modelos com os sinais vitais e os resultados laboratoriais melhoram as previsões dos modelos que usam apenas a idade, aumentando significativamente a área sob a curva ROC e diminuindo a pontuação de *Brier*. Assim, os autores acabam por concluir que a variável explicativa idade é a variável explicativa que mais está associada à mortalidade, uma vez que o modelo que apenas a incluía era quase idêntico ao modelo contendo as informações demográficas e a data e a um modelo contendo as informações demográficas, data e as comorbilidades. Por outro lado, os sinais vitais e os resultados laboratoriais acrescentam informações prognósticas, para além da idade. No geral, os resultados sugerem que a idade, os sinais vitais e os resultados laboratoriais podem ser úteis para avaliar o prognóstico de um paciente hospitalizado.

Contudo, este estudo tem a limitação da existência de dados faltantes, pelo que os autores tentaram superar essa limitação utilizando imputação múltipla, pelo que pode ter existido alguma alteração da distribuição dos dados quando estes foram imputados.

Os autores Sumayh Aljameel et al. [7], desenvolveram um estudo no qual o objetivo é a obtenção de um modelo preditivo, para identificar os paciente que têm um maior risco

de morte por COVID-19.

Os autores utilizaram uma base de dados, que contém informações demográficas e dados clínicos de pacientes que testaram positivo à COVID-19 e que foram admitidos no Hospital Universitário King Fahad, Arábia Saudita, no período entre 30 de abril de 2020 e 24 de julho de 2020. Existem 287 registos de pacientes com COVID-19 no conjunto de dados, onde a variável resposta é a variável binária associada ao desfecho morte. A base de dados é constituída ainda por variáveis demográficas, sintomas, comorbilidades e por dados preliminares como, por exemplo, a temperatura corporal à entrada do hospital. Os autores acabaram por remover variáveis desta base de dados que tinham valores com uma baixa frequência e por outro lado apenas levaram para estudo todos os sintomas que tinham mais de 50% de ocorrência, sendo que os sintomas com ocorrências entre os 2% e os 49% foi criado uma nova variável em que foi atribuído um código único, com o nome de "outros sintomas". O mesmo foi feito para as comorbilidades, ou seja somente as comorbilidades com alta frequência é que foram usadas como variáveis explicativas, enquanto que as restantes foram incorporadas numa nova variável explicativa, que se chama "outras comorbilidades". Por outro lado, os autores ainda aplicaram a técnica do *K-means* para imputação dos dados omissos.

Neste estudo, os modelos preditivos considerados pelos autores foram a Regressão Logística, as Florestas Aleatórias e o XGBoost. Para treinar e testar cada um destes modelos, os autores utilizaram o método da validação cruzada, onde as métricas usadas para avaliar a performance preditiva de cada um destes modelos foram a *accuracy*, *precisão*, *especificidade*, *sensibilidade*, a medida-*F* e a área sob a curva ROC. No entanto, antes de treinarem e testarem cada um dos modelos, os autores constataram que a base de dados era desbalanceada, pelo que a balancearam através da técnica SMOTE. No entanto, os autores avaliaram os resultados tanto para os dados balanceados como para os dados desbalanceados.

O conjunto de variáveis explicativas utilizadas para treinar cada um dos modelos foram, o conjunto formado por todas as 25 variáveis explicativas, o conjunto formado pelas 20 primeiras variáveis explicativas mais correlacionadas com a variável de resposta, o conjunto formado pelas primeiras 15 variáveis explicativas mais correlacionadas com a variável de resposta e o conjunto formado pelas 10 primeiras variáveis explicativas mais correlacionadas com a variável de resposta, pelo que cada um dos modelos considerados foram treinados e testados 4 vezes, sendo que isto foi aplicado tanto para a base de dados desbalanceada como para a base de dados balanceada pela técnica SMOTE.

Os resultados obtidos indicam que a Floresta Aleatória superou todos os modelos quando se utilizou a técnica SMOTE para balancear a base de dados, em todas as métricas. Por outro lado, os autores também obtiveram que o modelo das Florestas Aleatórias, foi o melhor modelo quando utilizaram 20 variáveis explicativas, com a técnica SMOTE para balancearem a base de dados, sendo que a Floresta Aleatória alcançou o valor 1 na *especificidade* quando se utiliza as 15 primeiras variáveis explicativas. O modelo da Regressão Logística é o modelo que tem o desempenho inferior sobre os outros classificadores quando se utilizam as 20, 15 e as 10 primeiras variáveis explicativas

mais importantes, utilizando os dados balanceados pela técnica SMOTE. Os autores ainda referem que a variável explicativa idade é uma das variáveis explicativas mais importantes neste estudo, uma vez que está entre as 10 primeiras variáveis explicativas em todas as 25, tal como outros estudos que observaram que a idade é um dos principais fatores que ajudam a prever a gravidade dos casos. No entanto, os autores acabam por referir que os modelos do seu estudo precisam de ser validades usando vários conjunto de dados, onde também devem ser incluídas mais variáveis explicativas que foram consideradas importantes em outros estudos.

No entanto, também este estudo tem algumas limitações, sendo uma das principais o facto do autor propor o método do *K-means* para fazer imputação na base de dados, uma vez que este método utiliza a distância euclidiana como medida de distância e não nos podemos esquecer que no presente estudo existem variáveis categóricas, logo não é correto aplicar esta medida de distância a variáveis categóricas.

Outro estudo desenvolvido por Quazi Adibur et al [8], tem como objetivo o desenvolvimento de um modelo preditivo que seja capaz de estimar a probabilidade de morte duma mulher grávida diagnosticada com COVID-19, com base nos seguintes sintomas, dispneia, tosse, rinorreia, artralgia, e o diagnóstico de pneumonia. Neste estudo os métodos de Aprendizagem Automática utilizados para atingir o objetivo proposto foram o SVM, Árvores de Decisão, Florestas Aleatórias, XGBoost e Redes Neuronais Artificiais. A base de dados que os autores trabalham é selecionada de uma base de dados maior, formada por comorbilidades e sintomas de pessoas que testaram positivo à COVID-19, sendo que os autores selecionam apenas as observações correspondentes às mulheres grávidas.

Para treinar e testar cada um destes modelos, os autores dividiram a base de dados em duas bases de dados, a base de dados de treino e a base de dados de teste, onde a base de treino é constituída por 70% das observações dos dados iniciais e as restantes 30% para a base de dados de teste. No entanto, antes disso, os autores balancearam a base de dados inicial através da técnica SMOTE, onde esta técnica é uma técnica de sobreamostragem, onde o objetivo principal é gerar dados da classe minoritária, onde os autores deferiram que este algoritmo deveria ser aplicado até obterem 50% de observações na classe minoritária e 50% na classe maioritária. Os autores apenas levaram para o treino dos modelos os sintomas odinofagia, artralgia, rinorreia, pneumonia, tosse, dispneia e tipo de paciente, se é hospitalizado ou se é ambulatório, e o contacto COVID, sendo que a variável resposta é binária, indicando se a mulher grávida acabou por sobreviver ou não.

Para a avaliação de cada um dos modelos, os autores utilizam as métricas *accuracy*, *precisão*, medida-F e o *recall*. Neste estudo, os autores consideram a métrica mais importante a *precisão*, sendo que os modelos das Árvores de Decisão, Florestas Aleatórias, XGBoost e Redes Neuronais são os modelos ideais, uma vez que têm uma pontuação de 1 na *precisão*. No entanto, os autores referem que se considerarem a métrica medida-F, os melhores modelos são o XGBoost e as Redes Neuronais. Por outro lado, os autores ainda analisaram a relação entre a mortalidade de mulheres grávidas com COVID-19 com os sintomas odinofagia, calafrios, artralgia, rinorreia, pneumonia, tosse, dispneia e se o paciente era

ambulatório ou hospitalizado, sendo que para isso os autores utilizaram uma matriz de correlações, onde concluem que a rinorreia, artralgias, odinofagia e calafrios apresentam uma correlação muito forte com a mortalidade. Os autores ainda referem que os sintomas calafrios e rinorreia têm uma correlação positiva forte com o contato com COVID, sendo que a presença de rinorreia e calafrios é um identificador para a presença de COVID-19, que está profundamente correlacionado com a mortalidade. De referir que mais uma vez este artigo tem a desvantagem de se ter balanceado a base de dados através da técnica SMOTE, acrescentando-se observações à classe minoritária, geradas aleatoriamente através do algoritmo KNN, sendo que neste caso se perde a independência das observações.

Os autores Sheng Zhang et al. [9], desenvolveram também um trabalho com o objetivo de investigar fatores prognósticos para pacientes com COVID-19, usando métodos de Aprendizagem Automática, usando um modelo linear generalizado, a Regressão de LASSO e um modelo não linear que é as Redes Neuronais. Para atingir o objetivo proposto os autores utilizaram uma base de dados que contém os casos entre 29 de dezembro de 2019 a 2 de março de 2020 dos pacientes hospitalizados no Hospital de doenças infeciosas de Wuhan. A base de dados apenas continha os pacientes hospitalizados que tinham 18 ou mais anos, que testaram positivo à COVID-19, incluindo também dados demográficos, dados referentes a sintomas e às comorbilidades, o desfecho do caso, dados clínicos, incluindo a gravidade da doença, entre outros. No entanto, os autores excluíram todas observações que tivessem valores omissos, ficando com 1145 observações. Neste estudo a variável resposta indica a ocorrência de morte.

Para selecionar o subconjunto das variáveis explicativas que entrariam para o modelo da Regressão de LASSO e para o modelo das Redes Neuronais, os autores utilizaram o método de LASSO para a seleção do subconjunto das variáveis explicativas, que reduz automaticamente os coeficientes das variáveis explicativas não relevantes exatamente para zero, removendo assim variáveis irrelevantes. Para determinar o parâmetro de penalização, os autores utilizaram a validação cruzada. Feito isto, todas as variáveis explicativas com coeficientes diferentes de zero, foram utilizadas para a construção do modelo da Regressão de LASSO e das Redes Neuronais.

Para treinar e testar cada um dos modelos, os autores dividiram a base de dados em base de dados de treino e em base de dados de teste, onde utilizaram cerca de 70% das observações para treino e as restantes 30% para teste. Para avaliar a capacidade preditiva de cada um dos modelos, os autores utilizam a área sob a Curva ROC.

As conclusões que os autores chegam é que tanto a Regressão de LASSO como também as Redes Neuronais baseadas em LASSO são modelos bastantes poderosos para prever o prognóstico de pacientes com COVID-19, em termos do valor da área sob a curva ROC. Por fim, os autores também concluem que a idade é um dos fatores prognósticos para a mortalidade.

Neste estudo os autores referem que a seleção do subconjunto das variáveis para entrarem no modelo de treino é efetuada sobre os dados de treino, o que na minha modesta opinião, deveria ser feito sobre a base de dados completa, uma vez que esta tem

um maior número de observações, assim como conserva os padrões existentes nos dados. Depois de selecionadas o subconjunto das variáveis explicativas é que, treinaria cada um dos modelos com as variáveis explicativas selecionadas.

3

Métodos

Nas secções que se seguem, iremos denotar genericamente $\mathbf{x} = (x_1, \ldots, x_p)^T$, o vetor que contém as p variáveis explicativas, correspondendo x_j, $j = 1, \ldots, p$ à *j-ésima* variável explicativa. Cada uma destas p variáveis explicativas foi observada n vezes, $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, sendo x_{ij}, $i = 1, \ldots, n$, a *i-ésima* observação da variável explicativa x_j.

Do mesmo modo o vetor $\mathbf{y} = (y_1, \ldots, y_n)^T$ é o vetor que contém as observações da variável resposta Y, que no presente trabalho assumimos que Y$\sim$ Bernoulli. No caso univariado, assumimos que o vetor da covariável x é constituído pelas n observações, ou seja, $\mathbf{x} = (x_1, \ldots, x_n)^T$.

Neste capítulo, iremos descrever os métodos utilizados para a concretização do objetivo proposto, sendo que na secção 3.1 encontra-se descrito o teste de associação entre duas variáveis. Na secção 3.2, encontra-se descrito o modelo de Regressão Logística, tanto para o caso univariado como para o caso multivariado. Na secção 3.3, encontram-se descritos os métodos de contração, como a Regressão de *Ridge*, a Regressão de LASSO, e o LASSO de Grupo. Na secção 3.4, encontra-se uma descrição dos modelos Aditivos Generalizados, assim como também as funções de suavização *splines*. Na secção 3.5 descrevemos os modelos de Árvores de Regressão e as Árvores de Classificação, como também o método para poda, enquanto que na secção 3.6.1 descrevemos as Florestas Aleatórias. Na secção 3.7, descrevemos as Redes Neuronais. Na 3.8, descrevemos a avaliação da performance do modelo, na 3.9, introduzimos as métricas para a avaliação do desempenho preditivo dos modelos e por fim na secção 3.10, encontram-se descrito os problemas de classificação com dados desbalanceados e os diferentes métodos para o balanceamento dos dados.

3.1 Teste de Associação entre Duas Variáveis

3.1.1 Teste de Associação entre Duas Variáveis Categóricas

O teste do qui-quadrado pode ser utilizado para testar a hipótese nula dada em 3.1 da independência ou a homogeneidade entre duas variáveis [10].

$$\begin{cases} H_0 : \text{As Variáveis } X \text{ e } Y \text{ são independentes} \equiv \text{Não existe associação entre as variáveis } X \text{ e } Y \\ H_1 : \text{As Variáveis } X \text{ e } Y \text{ não são independentes} \equiv \text{Existe associação entre as variáveis } X \text{ e } Y \end{cases} \tag{3.1}$$

Este teste utiliza uma tabela de contingência para testar se duas dadas variáveis são associadas entre si, pelo que este teste também é chamado de teste de independência [10]. Considere-se que à variável aleatória X correspondem r níveis de valores e à variável aleatória Y correspondem c níveis. Isso permite-nos distribuir as suas N observações numa tabela de contingência $(r \times c)$. O número de observações associadas simultaneamente à linha i e à coluna j da tabela denotamos por O_{ij} [10]. Já o número total de observações da linha i da tabela denotamos por R_i, enquanto que o número total de observações da coluna j da tabela denotamos por C_j [10]. A probabilidade de uma observação pertencer à linha i e à coluna j da tabela é dada, sobre a hipótese de independência, é dada por $(R_i \times C_j)/N$, sendo que esta quantidade é denotada por E_{ij} [10].

O teste do qui-quadrado utiliza a estatística de teste de qui-quadrado (χ^2), dada por 3.2, onde sob a hipótese nula H_0 dada em 3.1, a estatística de teste χ^2 segue uma distribuição qui-quadrado com $(r-1)\times(c-1)$ graus de liberdades, como se encontra na representado em 3.2 [10].

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)\times(c-1)} \tag{3.2}$$

O *software* R tem a função *chisq.test()*, que calcula o valor observado da estatística de teste χ^2, sob a hipótese nula H_0 representada em 3.1. Para um valor do $p\text{-}value$ maior que o nível de significância α não se rejeita a hipótese nula ao nível de significância α, e pode-se concluir que a esse nível de significância é plausível assumir que as variáveis X e Y são independentes, ou seja, não existe uma associação entre a variável X e Y.

De referir que o valor do $p\text{-}value$, é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada numa amostra, sob a hipótese nula, H_0 [10].

3.1.2 Teste de Associação entre Uma Variável Categórica e Uma Variável Contínua

Uma forma de averiguar se uma variável contínua está associada com uma variável categórica, é através do modelo de Regressão Logística univariado, em que apenas se tem uma variável explicativa, onde a variável resposta é a variável categórica e a variável explicativa é a variável contínua. Assim, para ver se existe associação entre estas duas variáveis basta ajustar o modelo e fazer um teste de significância ao coeficiente da variável explicativa. Se não se rejeitar a hipótese nula, significa que as variáveis não têm relação, caso contrário pode-se afirmar que pode existir uma relação entre estas duas variáveis.

Na secção 3.2, descreve-se o modelo de Regressão Logística para o caso univariado, assim como o teste de hipóteses à significância dos parâmetros.

3.2 Regressão Logística

A Regressão Logística é um tipo especial de regressão, em que a variável resposta é dicotómica ou categórica (com mais do que duas classes) e as variáveis explicativas podem ser discretas ou contínuas. No caso da variável resposta ser dicotómica, atribui-se o valor 1 ao acontecimento sucesso e 0 ao acontecimento insucesso. A Regressão Logística modela a probabilidade de ocorrência do evento em análise [11].

3.2.1 Caso Univariado

3.2.1.1 O Modelo

Na Regressão Linear, temos que a variável resposta é contínua, enquanto que na Regressão Logística, tem-se que as variáveis resposta são categóricas ou dicotómicas e esta é uma das diferenças entre a Regressão Logística e a Regressão Linear [11]. Esta diferença é refletida na escolha do modelo paramétrico, assim como nos seus pressupostos.

Em qualquer regressão a quantidade chave é o valor médio da variável resposta dado o valor da variável explicativa. Esta quantidade é chamada de valor médio condicional e é expressa como $E(\mathrm{Y}|x)$, onde Y representa a variável resposta e x representa a variável explicativa [11].

Na Regressão Linear, admite-se que o valor médio condicional, pode ser expresso como uma equação linear em x, o chamado preditor linear, isto é:

$$E(\mathrm{Y}|x) = \beta_0 + \beta_1 x. \tag{3.3}$$

É de notar que esta quantidade pode assumir valores de $-\infty$ a $+\infty$ [11]. Já no caso em que a variável resposta é dicotómica, o valor médio condicional assume apenas valores entre 0 e 1, ou seja, tem-se que $0 \leq E(\mathrm{Y}|x) \leq 1 = \pi(x) \leq 1$, em que $\pi(x)$ é a probabilidade de sucesso associada à variável resposta dicotómica, $\pi(x) = P(\mathrm{Y} = 1|x)$ [11].

Assim, com base na função Binomial o modelo de Regressão Logística que usamos é [11]:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}. \tag{3.4}$$

Uma transformação que é central nos modelos de Regressão Logística é a transformação *logit*, usada com o objetivo de linearizar o modelo nos parâmetros, aplicando o logaritmo [11]. Esta transformação aplicada à probabilidade de sucesso permite relacioná-la com um preditor com valores reais, definindo-se como:

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x. \tag{3.5}$$

Esta transformação é importante, pois através desta transformação tem-se que o modelo possui diversas propriedades do modelo de Regressão Linear, isto é [11]:

1. A função *logit* calculada no valor médio condicional é linear nos parâmetros;
2. É contínua;
3. Os seus valores podem variar entre $-\infty$ a $+\infty$.

Outra das diferenças entre a Regressão Linear e a Regressão Logística, consiste na distribuição condicional da variável dependente [11]. No modelo de Regressão Linear assume-se que uma observação da variável resposta, pode ser expressa como $y = E(\mathrm{Y}|x)+\varepsilon$ [11]. A quantidade ε é chamado de erro e expressa os desvios de uma observação à média condicional [11]. Outra das suposições feitas habitualmente na Regressão Linear é que ε segue uma distribuição Normal com média 0 e que a variância σ^2 é constante, para todos os níveis da variável independente [11]. Isto quer dizer que a distribuição da variável dependente condicionada pela variável independente segue uma distribuição normal de média $E(\mathrm{Y}|x)$ e de variância constante, isto é,

$$\mathrm{Y}|x \sim N\Big(E(Y|x), \sigma^2\Big). \tag{3.6}$$

No modelo da Regressão Logística, temos que este caso não ocorre, pois uma vez que a variável dependente é dicotómica. Assim, nesta situação, pode-se expressar o valor da variável dependente dado a variável independente como $y = \pi(x) + \varepsilon$. Aqui a quantidade ε, apenas pode assumir dois valores possíveis que são:

$$\varepsilon = \begin{cases} -\pi(x), & \text{com probabilidade } \ 1 - \pi(x) \\ 1 - \pi(x), & \text{com probabilidade } \ \pi(x). \end{cases} \tag{3.7}$$

Neste caso, tem-se que ε tem uma distribuição com média 0 e variância igual a $\pi(x)\Big(1 - \pi(x)\Big)$. A distribuição condicional da variável dependente $(\mathrm{Y}|x)$ segue uma distribuição Bernoulli com probabilidade de sucesso $\pi(x)$.

O modelo de Regressão Logística descreve-se então como:

$$\begin{aligned} &\mathrm{Y}|x \sim \text{Bernoulli}\Big(\pi(x)\Big) \\ &\mathit{logit}\Big(\pi(x)\Big) = \beta_0 + \beta_1 x. \end{aligned} \tag{3.8}$$

3.2.1.2 Estimação do Modelo

Suponhamos que temos uma amostra com n observações independentes do par ordenado (y_i, x_i), $i = 1, 2, \ldots, n$, onde y_i corresponde à variável resposta para a *i-ésima* observação, onde a variável resposta é dicotómica e x_i à variável independente para a *i-ésima* observação. Para estimarmos o modelo de Regressão Logística, dado pela equação 3.8, necessitamos de estimar os parâmetros β_0 e β_1, uma vez que estes parâmetros são desconhecidos. No modelo de Regressão Linear o método usado para estimar os parâmetros desconhecidos é o método dos mínimos quadrados. Este método escolhe os valores de β_0 e β_1 que minimizam a soma de quadrados dos resíduos (*Residual Sum of Squares-RSS*), isto é, o β_0 e o β_1 são escolhidos de forma a minimizar a quantidade representada em 3.9 [11].

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \quad \text{onde} \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.9}$$

Com os pressupostos usuais da Regressão Linear, o método dos mínimos quadrados produz os estimadores com as propriedades estatísticas desejáveis. No entanto, quando o método dos mínimos quadrados é aplicado a uma variável resposta dicotómica, os estimadores obtidos não possuem as mesmas propriedades que os estimadores obtidos na Regressão Linear, ou seja, os estimadores não possuem as propriedades estatísticas desejáveis.

Quando os termos de erros são normalmente distribuídos, temos que o método geral de estimação que leva ao método dos mínimos quadrados no modelo de Regressão Linear coincide com o método de máxima verosimilhança. Para estimar os parâmetros desconhecidos no modelo de Regressão Logística, iremos basear-nos neste método. O método da máxima verosimilhança tem como objetivo determinar os valores dos parâmetros β_0 e β_1 que maximizam a plausibilidade de obter o conjunto de valores observados. Para se utilizar o método da máxima verosimilhança, primeiro teremos que construir a função de verosimilhança, onde esta expressa a probabilidade dos dados observados como função dos parâmetros desconhecidos, onde os estimadores de máxima verosimilhança são os valores que maximizam a função de verosimilhança. Assim, os parâmetros estimados serão aqueles que melhor se adequam aos dados observados.

Uma vez que:

1. A variável dependente Y, toma os valores 0 e 1;

2. $P(\mathrm{Y} = 1|x) = \pi(x)$ e $P(\mathrm{Y} = 0|x) = 1 - \pi(x)$,

e para todo o par (y_i, x_i), a contribuição deste na função de verosimilhança é dada pela seguinte expressão:

$$\pi(x_i)^{y_i}\Big(1 - \pi(x_i)\Big)^{(1-y_i)}, \quad \text{com} \quad y_i \in \{0, 1\}. \tag{3.10}$$

Assumindo a independência das observações, a função de verosimilhança é obtida à custa do produto dos termos dados na equação 3.10, como:

$$L(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{(1-y_i)}. \tag{3.11}$$

O método da máxima verosimilhança utiliza como estimador de $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ os valores que maximizam a expressão da equação 3.11. No entanto, do ponto de vista matemático é mais fácil trabalhar com o logaritmo da equação 3.11, pois com esta transformação passamos da função de verosimilhança do produto das probabilidades para a soma das probabilidades. Esta transformação tem o nome de log-verosimilhança e é definida como mostra a equação 3.12.

$$\ell(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})] = \sum_{i=1}^{n} \Big[y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)] \Big] \tag{3.12}$$

Para encontrar os valores de $\boldsymbol{\beta}$ que maximizam $\ell(\boldsymbol{\beta})$, diferencia-se $\ell(\boldsymbol{\beta})$ em ordem a β_0 e a β_1 (*scores*) e igualamos essas expressões a 0, ficando com as equações seguintes, que são conhecidas com as equações de verosimilhança:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{3.13}$$

$$\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0. \tag{3.14}$$

Deve-se ainda verificar que as segundas derivadas são negativas, assegurando os máximos. As equações 3.13 e 3.14 não são lineares nos parâmetros e portanto requerem métodos iterativos para a resolução destas equações. No *software R*, a função *glm()* utiliza o método iterativo dos mínimos quadrados ponderados para a estimação dos parâmetros do modelo.

3.2.1.3 Teste à Significância dos Coeficientes

Uma vez estimados os parâmetros do modelo, está-se interessado em analisar a significância destes, isto é, está-se interessado em testar quais as variáveis explicativas do modelo que estão efetivamente relacionadas com a variável resposta. No caso do modelo univariado, está-se interessado em testar:

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0. \tag{3.15}$$

Uma das formas de testar a significância de uma variável explicativa num dado modelo é ver se o modelo que a inclui explica mais a variável resposta do que o modelo que não a inclui [11]. Para isso, compara-se os valores observados da variável resposta com os previstos por cada um dos modelos, o que contém a variável explicativa e o que não a contém. No caso da Regressão Logística, a comparação do valor observado com o previsto é baseado na função de log-verosimilhança definida na equação 3.12, sendo a comparação entre o valor observado e previsto usando a função de verosimilhança baseada na expressão 3.16 [11].

$$D = -2 \ln \left[\frac{\text{(verosimilhança do modelo ajustado)}}{\text{(verosimilhança do modelo saturado)}} \right] \tag{3.16}$$

O modelo saturado, é o modelo onde o número de observações iguala o número de parâmetros, isto é, o modelo em que o valor médio de cada observação coincide com o valor observado, isto é, $E(\mathrm{Y}) = y_i$ [11]. A quantidade dentro dos parêntesis retos da equação 3.16 é chamada de razão de verosimilhança [11]. Usando menos duas vezes o logaritmo da razão de verosimilhança, é necessário obter uma quantidade, cuja a distribuição seja

conhecida e portanto possa ser utilizada para fins de testes de hipóteses. Este teste é chamado de teste de razão de verosimilhança (*Likelihood Ratio test*) [11]. O teste de razão de verosimilhança é baseado na equação 3.16, que atualizado com 3.12 resulta em:

$$D = -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right], \tag{3.17}$$

onde $\hat{\pi}_i = \hat{\pi}(x_i)$ [11]. A estatística D da equação 3.17 é chamada de *deviance* e para a Regressão Logística ela desempenha o mesmo papel que a soma de quadrados dos resíduos na Regressão Linear. No caso em que a variável resposta apenas tome os valores 0 e 1, tem-se que a função verosimilhança do modelo saturado é igual a 1, pois, por definição de modelo saturado tem-se que $\hat{\pi}_i = y_i$ e a função verosimilhança é [11]:

$$L(\text{modelo saturado}) = \prod_{i=1}^{n} y_i^{y_i}(1-y_i)^{(1-y_i)} = 1. \tag{3.18}$$

Assim, a *deviance* dada na equação 3.16 no caso da variável resposta seguir uma distribuição Bernoulli é [11]:

$$D = -2\ln(\text{verosimilhança do modelo ajustado}). \tag{3.19}$$

Para analisar a significância da variável independente compara-se o valor de D com e sem a variável independente. Assim fica-se com uma nova estatística de teste, que é a seguinte:

$$G = D(\text{modelo sem a variável}) - D(\text{modelo com a variável}). \tag{3.20}$$

A estatística G desempenha o mesmo papel na Regressão Logística que o numerador do teste parcial F na Regressão Linear [11]. Uma vez que, a função verosimilhança do modelo saturado é comum a ambos os valores de D, então G pode ser expresso como se tem representado na equação 3.21 [11].

$$G = -2\ln\left[\frac{\text{verosimilhança sem a variável}}{\text{verosimilhança com a variável}}\right] \tag{3.21}$$

Sob a hipótese nula H_0, representada em 3.15, a estatística G de 3.21 segue uma distribuição qui-quadrado com um grau de liberdade, ou seja, sob a hipótese nula, G segue a distribuição representada em 3.22 [11].

$$G = -2\ln\left[\frac{\text{verosimilhança sem a variável}}{\text{verosimilhança com a variável}}\right] \sim \chi_1^2 \tag{3.22}$$

De acordo com o teste de razão de verosimilhança, a hipótese nula H_0, representada em 3.15, é rejeitada a favor de H_1, a um nível de significância de α, se o valor observado da estatística G de 3.22 for superior ao quantil de probabilidade $(1-\alpha)$ de uma distribuição χ_1^2 [11].

Existe ainda outro teste que permite testar a hipótese nula H_0 representada em 3.15, que é o teste de *Wald* [11]. O teste de *Wald* é baseado em 3.23, onde $Var(\hat{\beta}_1) = \left[E\left(\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_1^2} \right) \right]^{-1}$ [11].

$$W^2 = \left(\frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}} \right)^2 \tag{3.23}$$

À estatística W representada em 3.23 dá-se o nome de estatística de *Wald* [11]. Sob a hipótese nula H_0 de 3.15, tem-se que a estatística W^2 de 3.23 segue uma distribuição qui-quadrado com um grau de liberdade, ou seja, sob a hipótese nula, W^2 segue a distribuição representada em 3.24 [11].

$$W^2 = \left(\frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}} \right)^2 \sim \chi_1^2 \tag{3.24}$$

De acordo com o teste de *Wald*, a hipótese nula H_0, representada em 3.15, é rejeitada a favor de H_1, a um nível de significância de α, se o valor observado da estatística W^2 de 3.24 for superior ao quantil de probabilidade $(1 - \alpha)$ de uma distribuição χ_1^2 [11].

3.2.2 Caso Multivariado

Na secção anterior abordámos o modelo de Regressão Logística univariado, ou seja, quando temos apenas uma única variável independente. Consideremos agora o caso onde temos p variáveis independentes expresso pelo seguinte vetor $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$.

3.2.2.1 O Modelo

No caso do modelo de Regressão Logística Múltipla, temos que:

$$P(\mathrm{Y} = 1|\mathbf{x}) = E(\mathrm{Y}|\mathbf{x}) = \pi(\mathbf{x}). \tag{3.25}$$

O *logit* da probabilidade de sucesso no modelo da Regressão Logística Múltipla é dada pela equação 3.26,

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} \right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p. \tag{3.26}$$

donde para o modelo de Regressão Logística Múltipla,

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}. \tag{3.27}$$

No caso do modelo de Regressão Logística Múltipla é possível existirem variáveis independentes discretas, contínuas e categóricas, como por exemplo, o sexo, a etnia, entre outras, não sendo adequado incluí-las no modelo como se fossem variáveis contínuas [11].

Utilizando uma escala numérica para representar os diversos níveis de uma variável categórica, tais valores são apenas identificadores e não possuem a sua significância numérica. Nessa situação, estas variáveis são substituídas por variáveis *dummy*, codificadas como 0 ou 1 para cada dado nível de uma variável categórica. Por exemplo, se tivermos uma variável categórica com 3 níveis, referentes a cor, por exemplo "Azul", "Verde" e "Vermelho", neste caso, são necessárias duas variáveis *dummy*. Uma codificação estratégica possível é que quando a covariável vale "Azul", as duas variáveis *dummy*, D_1 e D_2, seriam ambas iguais a zero, quando é "Verde", D_1 seria igual a 1 enquanto que a D_2 seria igual a 0 e quando a cor é "Vermelho", D_1 será igual a 0 e D_2 será igual a 1. Ao nível "Azul" chama-se nível de referência.

Posto isto, de um modo geral, se uma variável categórica tiver k níveis, então são necessárias $(k-1)$ variáveis *dummy*. O motivo para se utilizar uma a menos deve-se à questão de identificabilidade dos parâmetros [11]. Assim, se a $j-ésima$ variável independente x_j é uma variável categórica com k níveis, as $(k-1)$ variáveis *dummy* serão denotadas como D_{jl} e os coeficientes para estas variáveis são denotados por β_{jl}, com $l = 1, \ldots, k_j - 1$, pelo que o preditor para o modelo com p variáveis, sendo a *j-ésima* variável categórica, é dado pela equação 3.28.

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{l=1}^{k_j-1} \beta_{jl} D_{jl} + \beta_p x_p \tag{3.28}$$

3.2.2.2 Estimação do Modelo

Suponhamos que temos uma amostra com n observações independentes $(y_i, \mathbf{x}_i)$, com $i = 1, 2, \ldots, n$, onde y_i é a variável dependente dicotómica/categórica e $\mathbf{x}_i$ é o *i-ésimo* valor do vetor das variáveis independentes. Tal como no caso univariado para se ajustar o modelo é necessário obter-se estimativas para cada um dos β_j, $(j = 0, 1, 2, \ldots, p)$, isto é, é necessário estimar o vetor $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. O método de estimação usado no caso multivariado é o mesmo que no caso univariado, ou seja, o método utilizado é o de máxima verosimilhança. Assim, assumindo a independência entre as observações tem-se que a função de verosimilhança e de log-verosimilhança têm a mesma forma que a função 3.11 e 3.12, mas onde $\pi(\mathbf{x})$ é o que está definido em 3.27. Mais uma vez, para se obter os estimadores de máxima verosimilhança, ter-se-à que resolver um sistema de equações, com $(p+1)$ equações de verosimilhança, pois para se encontrar o vetor $\boldsymbol{\beta}$ que maximiza a função de verosimilhança, derivamos parcialmente a função de verosimilhança em ordem a cada um dos parâmetros, isto é, em ordem a cada β_j, com $j = 0, 1, \ldots, p$. As equações de verosimilhança são:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0 \tag{3.29}$$

derivando em ordem a β_0,

$$\sum_{i=1}^{n} x_{ij}[y_i - \pi(\mathbf{x}_i)] = 0 \tag{3.30}$$

derivando em ordem a β_j, com $j = 1, 2, ...p$.

Tal como na Regressão Logística univariada a solução das equações de verosimilhança requerem a utilização de um *software* adequado para se resolver numericamente o devido sistema de equações, dada a sua complexidade. A solução do sistema de equações denota-se por $\hat{\boldsymbol{\beta}}$. Os valores ajustados para o modelo de Regressão Logística são os $\hat{\pi}(\mathbf{x}_i)$, ou seja, é o valor da expressão 3.27 calculada usando os $\hat{\boldsymbol{\beta}}$ e $\mathbf{x}_i$.

Para além das estimativas de máxima verosimilhança dos parâmetros, pode-se obter a variância e a covariância dos parâmetros estimados, em que estas se obtêm a partir da matriz constituída pelas segundas derivadas parciais da função log-verosimilhança. A forma geral das derivadas parciais são:

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j^2} = \sum_{i=1}^{n} x_{ij}^2 \pi_i (1 - \pi_i) \tag{3.31}$$

e

$$\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi_i (1 - \pi_i) \tag{3.32}$$

para $j, l = 0, 1, 2, ..., p$ e onde $\pi_i = \pi(\mathbf{x}_i)$ [11].

À matriz $(p + 1) \times (p + 1)$ composta pelos valores simétricos das equações 3.31 e 3.32, que designamos por matriz de Informação de *Fisher*, é denotada por $I(\boldsymbol{\beta})$ [11]. A variância e a covariância dos coeficientes estimados são obtidos através da inversão da matriz $I(\boldsymbol{\beta})$, ou seja, $Var(\boldsymbol{\beta}) = I^{-1}(\boldsymbol{\beta})$ [11]. Na diagonal principal desta matriz, encontram-se a variância de $\hat{\beta}_j$ e fora da diagonal desta matriz encontra-se a covariância entre a variável $\hat{\beta}_j$ e $\hat{\beta}_l$ [11].

3.2.2.3 Teste à Significância dos Coeficientes

Tendo o modelo ajustado, é necessário avaliar a significância dos coeficientes de regressão ou a validade do modelo, de forma a identificar quais as variáveis que têm uma maior influência no modelo estimado. A validade do modelo pode ser acedida através da *deviance* e da avaliação de se os seus pressupostos são verificados, através de análise de resíduos.

Uma das formas de avaliar a significância dos coeficientes de regressão é através o teste da razão de verosimilhança (*Likelihood Ratio test*), em que se avalia a hipótese de todos os coeficientes β_j com exceção do termo independente β_0, serem simultaneamente nulos, ou seja, temos a seguinte hipótese nula, H_0 representada em 3.33.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0 \quad vs \quad H_1 : \exists_j : \beta_j \neq 0, \quad j = 1, ..., p. \tag{3.33}$$

Neste caso o teste de razão de verosimilhança é realizado da mesma maneira que no caso univariado, é baseado na estatística G dada na equação 3.21, onde a única diferença

é que os valores ajustados $\hat{\pi}$ sobre o modelo ajustado são baseados no modelo ajustado contendo os $(p+1)$ parâmetros de $\hat{\boldsymbol{\beta}}$.

Sob a hipótese nula, H_0 de 3.33, de que os coeficientes das p variáveis explicativas são iguais a zero, tem-se que a estatística G seque uma distribuição qui-quadrado com p graus de liberdade, como se encontra representado em 3.34.

$$G = -2\ln\left[\frac{\text{verosimilhança do modelo ajustado somente com o } \beta_0}{\text{verosimilhança do modelo ajustado com as } p \text{ variáveis}}\right] \sim \chi^2_p \tag{3.34}$$

Assim, de acordo com o teste de razão de verosimilhança a hipótese nula, H_0 representada em 3.33, é rejeitada a favor de H_1, a um nível de significância α, se o valor observado da estatística G de 3.34 for superior ao quantil de probabilidade $(1-\alpha)$ de uma distribuição χ^2_p.

Por outro lado, pode-se estar interessado em testar se apenas um subconjunto dos coeficientes das p variáveis explicativas são identicamente nulos simultaneamente, ou seja, está-se interessado em testar a hipótese nula H_0 de 3.35.

$$H_0 : \boldsymbol{\beta}_r = 0 \quad vs \quad H_1 : \boldsymbol{\beta}_r \neq 0 \tag{3.35}$$

O β_r de 3.35 representa um sub-vetor de $\boldsymbol{\beta}$, ou seja, apenas se está interessado em testar se um determinado subconjunto r dos p coeficientes são nulos simultaneamente.

Para se testar 3.35, pode-se recorrer à estatística de razão de verosimilhança, pois uma vez que esta é muitas vezes utilizada para comparar modelos que estão encaixados [11]. Se nos relembrarmos, o teste de razão de verosimilhança é baseado na comparação de um modelo contendo todas as variáveis explicativas com o modelo que não tinha nenhuma variáveis explicativas. Neste caso, compara-se o modelo sem as r variáveis explicativas que é um sub-modelo do modelo com todas as variáveis explicativas. Neste caso, sob a hipótese nula, H_0 representada em 3.35, a estatística de teste G segue uma distribuição qui-quadrado com $(p-r)$ graus de liberdade, como se encontra representado em 3.36.

$$G = -2\ln\left[\frac{\text{verosimilhança do modelo ajustado somente com as } (p-r) \text{ variáveis}}{\text{verosimilhança do modelo ajustado com as } p \text{ variáveis}}\right] \sim \chi^2_{(p-r)} \tag{3.36}$$

De acordo com o teste de razão de verosimilhança a hipótese nula, H_0 representada em 3.35, é rejeitada a favor de H_1 a um nível de significância α, se o valor observado da estatística G de 3.36 for superior ao quantil de probabilidade $(1-\alpha)$ de uma distribuição $\chi^2_{(p-r)}$.

Por outro lado, pode-se testar se cada um dos coeficientes individualmente é diferente de zero, para isso pode-se utilizar o teste de *Wald*, pois uma vez que em geral a estatística de *Wald* é mais utilizada para testar hipóteses nulas sobre componentes individuais.

O teste de *Wald* testa se cada um dos coeficientes é significativamente diferente de zero. Deste modo, o teste de *Wald* averigua se uma determinada variável independente

apresenta uma relação estatisticamente significativa com a variável dependente, isto é, estamos interessados em testar, a hipótese nula, H_0 dada em 3.37 [11],

$$H_0 : \beta_j = 0 \quad vs \quad H_1 : \beta_j \neq 0, \quad j = 1, ..., p \tag{3.37}$$

onde a estatística de teste sob a hipótese nula, H_0 dada em 3.37, está representada em 3.38.

$$W^2 = \left(\frac{\hat{\beta}_j}{\sqrt{Var(\hat{\beta}_j)}} \right)^2 \sim \chi_1^2 \tag{3.38}$$

À estatística W, como denotado anteriormente, damos o nome de estatística de *Wald*. Assim, hipótese nula é rejeitada a um nível de significância α se o quadrado do valor observado da estatística de *Wald* for superior ao quantil de probabilidade $(1 - \alpha)$ de uma distribuição χ_1^2.

3.2.2.4 Interpretação dos Coeficientes

Depois de se ajustado um modelo e após se ter testado a significância dos coeficientes estimados, é necessário interpretar os seus valores. No entanto, na Regressão Logística a interpretação dos coeficientes é diferente do que na Regressão Linear. Na Regressão Logística, utiliza-se os *Odds-Ratio* (*OR*), para interpretar os coeficientes do modelo ajustado.

O *OR* é um rácio de *chances* (*odds*) de um evento em relação a outro, em que a *chance* é definido por 3.39, onde p é uma probabilidade de evento.

$$odds = \frac{p}{1 - p} \tag{3.39}$$

Para a interpretação, vamos considerar que estamos sobre o caso univariado. Anteriormente vimos na subsecção 3.2.1, que $\pi(x_i) = P(Y_i = 1|x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$. A *chance* ou *odds* de sucesso é dada por:

$$odds(x_i) = \frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)} = \frac{\pi(x_i)}{1 - \pi(x_i)}. \tag{3.40}$$

Verificando-se a linearidade da função, tem-se 3.41

$$\ln\Big(odds(x_i)\Big) = \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_i. \tag{3.41}$$

Se as variáveis forem categóricas, a razão de *chances* ou *OR* compara chances com base nos níveis da covariável. Se a variável x_i for categórica com k níveis, sendo o nível de referência o primeiro ($x_i = 1$), por exemplo, temos:

$$\ln\Big(odds(x_i)\Big) = \beta_0 + \beta_2 D_2 + \cdots + \beta_k D_k. \tag{3.42}$$

Pode-se então calcular o *OR* de cada um desses níveis em relação ao nível de referência como:

$$\begin{aligned} OR_j &= \frac{odds(x_i = j)}{odds(x_i = 1)} = e^{\left(\ln\left(\frac{odds(x_i = j)}{odds(x_i = 1)}\right)\right)} \\ &= e^{\ln\left(odds(x_i=j)\right) - \ln\left(odds(x_i=1)\right)} = e^{\beta_0+\beta_j-\beta_0} = e^{\beta_j}. \end{aligned} \tag{3.43}$$

Este pode ser interpretado como quanto mais provável ou improvável (em termos de *chances*) é que o resultado esteja presente entre os indivíduos com $x_i = j$ em comparação com os indivíduos com $x_i = 1$ [11].

No caso de variáveis contínuas, o *OR* entre um aumento de uma unidade no valor da covariável pode ser interpretado como a alteração na *chance* de sucesso provocada por esse aumento, uma vez que:

$$g(x + 1) - g(x) = \beta_0 + \beta_1(x + 1) - \beta_0 - \beta_1(x) = \beta_1 \tag{3.44}$$

pelo que a razão de *chances* é dado por e^{β_1}.

3.2.2.5 Método para Aferir a Qualidade do Ajustamento de um Modelo

Na sub-subsecção 3.2.2.3, viram-se vários testes para testar se um dado parâmetro, ou um subconjunto de parâmetros do modelo eram ou não nulos. Existindo evidência estatística de tal, podem-se remover esses parâmetros, obtendo um submodelo do modelo inicial, ou seja, um modelo encaixado [11]. No entanto, existem vários critérios para averiguar se este submodelo é melhor ou não do que o modelo que continha os parâmetros eliminados.

Um dos critérios da seleção do melhor modelo é o Critério de informação de *Akaike*, denotado por *AIC* [11]. Este critério é muito utilizado para comparar modelos com diferentes números de parâmetros, em que a medida *AIC* é definida como 3.45 [11].

$$AIC = -2\ell + 2(p + 1) \tag{3.45}$$

Na equação 3.45, ℓ representa a log-verosimilhança do modelo ajustado e p é o número de coeficientes de regressão para covariáveis não constantes [11]. Note-se que já se viu que a *deviance* é dada por $D = -2\ell$, pelo que o *AIC*, pode ser representado por 3.46.

$$AIC = D + 2(p + 1) \tag{3.46}$$

Assim, observa-se que o critério *AIC* é baseado na log-verosimilhança, com a introdução de um fator de correção como modo de penalização da complexidade do modelo, em que esta penalização é representada pelo número de parâmetros $(p + 1)$ que são utilizados para ajustar o modelo [11]. Um valor baixo de *AIC* é considerado como representativo de um melhor ajustamento, ou seja, quanto menor for o *AIC* melhor o modelo se ajusta aos dados, e na seleção dos modelos deve-se ter como objetivo a minimização do *AIC*.

No entanto, repare-se que o *AIC* permite selecionar um modelo de entre um conjunto de modelos alternativos, fazendo uma análise comparativa, mas no entanto, não diz nada sobre a qualidade de ajustamento de um modelo em termos absolutos, ou seja, o critério permite apenas escolher qual o melhor modelo de entre os diversos candidatos mesmo que esse modelo não se ajuste bem aos dados.

Um outro critério para a seleção do melhor modelo é o Critério de informação de *Bayesiano*, denotado por *BIC* [11]. Este critério é uma das fermentas de estatística também muito utilizado para a seleção de modelos estatísticos. A medida *BIC* é dada por 3.47 [11].

$$BIC = -2\ell + (p+1)\ln(m) \tag{3.47}$$

Ou seja, através da observação da equação 3.47, também se oberva que o *BIC* é baseado na log-verosimilhança, no entanto a penalização deste é diferente da penalização no *AIC*, pois uma vez que o termo de penalização do *BIC* é dado pelo $(p+1)\ln(m)$, onde $(p+1)$ é o número de parâmetros do modelo e m é o número de observações. Logo o *BIC* pode também ser representado por 3.48, pois uma vez que a *deviance* $D = -2\ell$, como visto anteriormente.

$$BIC = D + (p+1)\ln(m) \tag{3.48}$$

3.2.2.6 Seleção do Melhor Modelo - Seleção das Covariáveis Via *AIC*

Quando se utiliza a análise estatística em Modelos Lineares Generalizados em problemas práticos, muitas vezes tem-se inicialmente um elevado número de covariáveis ou variáveis explicativas e tem-se interesse nessa análise saber qual o modelo mais parcimonioso, ou seja, o modelo que inclua o menor número de variáveis explicativas e ainda explique adequadamente a variabilidade da variável resposta.

Na Regressão Logística, como membro dos Modelos Lineares Generalizados, tem-se diferentes métodos para a seleção do melhor modelo, tais como o método de *Stepwise*, *Forward* e *Backward* [12]. Qualquer uma destas abordagens, envolvem a identificação de um subconjunto das p variáveis explicativas que estão mais relacionadas com a variável resposta. De seguida, descrevem-se cada um dos métodos referidos anteriormente.

No método de *Forward*, começa-se com o modelo nulo, ou seja, apenas se começa com o modelo que tem o intercepto, parâmetro (β_0), mas não tem nenhuma variável explicativa e calcula-se o *AIC*; denotamos esse modelo por M_0. De seguida, ajustam-se todos os modelos possíveis apenas com uma variável explicativa, isto é, se tivermos p variáveis explicativas, então iremos ajustar p modelos, e para cada um desses p modelos calcula-se o *AIC*. Dos p modelos ajustados, iremos selecionar o que tem o menor *AIC* e denotamo-lo por M_1. Caso o *AIC* do modelo M_1 seja maior que o *AIC* do modelo M_0 termina o processo e o melhor modelo é o M_0; caso contrário, prosseguimos e ajustam-se todos os modelos possíveis com duas variáveis explicativas, em que uma é a incluída no modelo M_1, ou seja, vamos aumentar o modelo M_1 em uma variável, pelo que iremos ajustar $(p-1)$ modelos.

Para cada um destes $(p-1)$ modelos calculamos o *AIC* e escolhemos o que tiver menor *AIC*, que denotamos por M_2. Se o *AIC* do modelo M_2 for maior que o *AIC* do modelo M_1 paramos e o melhor modelo é o modelo M_2; caso contrário prossegue-se do mesmo modo, até termos o *AIC* do modelo M_j maior do que o *AIC* do modelo M_{j-1}, onde $j = 1, \ldots, p$ e o melhor modelo é o modelo M_{j-1}. Caso não haja nenhum modelo M_j em que o seu *AIC* seja maior que o *AIC* do modelo M_{j-1}, com $j = 1, \ldots, p$, então é porque o melhor modelo é o modelo M_p, ou seja, é o modelo com todas as variáveis explicativas.

Na tabela 3.1, encontra-se um pequeno resumo do método de *Forward*.

Tabela 3.1: Método de Seleção de *Forward* [12].

Método de Seleção de *Forward*
1. Seja M_0 o modelo nulo, que não contém nenhuma variável explicativa;
2. Seja k=0;
3. Considera-se todos os $(p-k)$ modelos que aumenta em uma variável explicativa as variáveis explicativas do modelo M_k e para cada um destes $(p-k)$ modelos seleciona-se o que tiver menor *AIC*, que se denota por M_{k+1};
4. Se o *AIC* do modelo M_{k+1} for maior que o *AIC* do modelo M_k, stop e o melhor modelo é o modelo M_k. Caso contrário volta-se para o passo 3 e $k = k+1$. Se $k > p$ então tem-se que o melhor modelo é o modelo M_p, ou seja, é o modelo que contém as p variáveis explicativas;

No método *Backward*, começa-se com o modelo completo, isto é, ajusta-se o modelo usando todos as p variáveis explicativas, denotado por M_p. De seguida este método remove iterativamente a variável explicativa menos útil, uma de cada vez, ou seja, uma vez ajustado o modelo M_p, ajustam-se p modelos, em que nestes p modelos ajustados tem-se em cada um deles menos uma variável explicativa que o modelo M_p. Dos p modelos ajustados escolhe-se aquele que tiver o menor valor de *AIC* denotando-o por M_{p-1}. O próximo passo, consiste em ajustar agora $(p-1)$ modelos, em que estes $(p-1)$ modelos têm menos uma variável explicativa que o modelo M_{p-1}, ou seja, cada um destes $(p-1)$ modelos ajustados contém $(p-2)$ variáveis explicativas das p variáveis explicativas iniciais. Destes $(p-1)$ modelos ajustados, escolhe-se aquele que tiver o menor *AIC* e denotamo-lo por M_{p-2}. Este processo vai-se repetir até atingir-se o modelo nulo, ou seja, até atingirmos o modelo que apenas contém o intercepto e que é denotado por M_0. De uma forma genérica, depois de ajustado o modelo que contém todas as p variáveis explicativas M_p, para cada $k = p, \ldots, 1$, ajustam-se todos os k modelos que contém todas as variáveis explicativas menos uma das variáveis explicativas de M_k, para um total de $(k-1)$ variáveis explicativas e desses k modelos ajustados escolhe-se o que tiver o menor valor de *AIC* e denotamo-lo por M_{k-1}. Feito isto, fica-se com $(p+1)$ modelos, $M_0, \ldots, M_p$ e dentro deste $(p+1)$ modelos

escolhe-se apenas o que tiver o menor valor de *AIC*, o melhor modelo.

Na tabela 3.2, encontra-se um pequeno resumo do método de *Backward*.

Tabela 3.2: Método de Seleção de *Backward* [12].

Método de Seleção de *Backward*
1. Ajusta-se o modelo usando todos as p variáveis explicativas, que denotamos por M_p;
2. Depois, para $k = p, p-1, \ldots, 1$ ajustam-se todos os k modelos que contêm todas menos uma das variáveis explicativas do modelo M_k, ou seja, para cada um dos k modelos ajustados tem-se um total de $k-1$ variáveis explicativas.
3. Dos k modelos ajustados anteriormente, escolhe-se o que tem o menor valor do *AIC*, denotado por M_{k-1}.
4. Por último, escolhe-se o único melhor modelo entre $M_0, \ldots M_p$ usando o *AIC*.

O método de *Stepwise* é um método que começa com o ajuste do modelo apenas com o intercepto e que de seguida vai acrescentando ao modelo ajustado variáveis explicativas, tal como o método de *Fordward* faz, no entanto, após adicionar cada variável explicativa este método também pode remover algumas variáveis explicativas já antes incluídas, se estas não apresentarem uma melhoria na qualidade de ajuste do modelo, com um *AIC* menor [12]. Este procedimento repete-se até se encontrar o modelo final, em que este é o melhor modelo [12].

Em qualquer um destes métodos, obtém-se um modelo final, em que este é considerado o melhor modelo ajustado, ou seja, as variáveis explicativas que se encontram no modelo final, são as variáveis explicativas que melhor explicam a variável resposta, pelo que no modelo final temos o melhor subconjunto das p variáveis explicativas, que mais influência/importância têm para a análise do problema em questão.

3.2.2.7 Interações

Uma iteração é um produto entre duas ou mais variáveis explicativas. Uma iteração entre duas variáveis pode ser utilizada, quando duas variáveis explicativas estão associadas entre si, ou seja, dizemos que duas variáveis explicativas interagem quando o efeito de uma variável explicativa sobre a variável resposta depende do valor da outra variável explicativa e vise versa. No caso de se considerar a interação entre duas variáveis contínuas ou uma contínua e uma variável dicotómica, então a nova variável é somente o produto entre estas duas variáveis.

Já no caso de considerarmos a interação entre uma variável categórica com k níveis e uma variável contínua, então vão-se criar $(k-1)$ novas variáveis, em que estas $(k-1)$ novas variáveis, são as variáveis *dummy* para representar os k níveis e estas $(k-1)$ variáveis

dummy vão estar a multiplicar pela variável contínua. De forma análoga também se faz da mesma forma para uma variável categórica e uma variável dicotómica.

No entanto, quando se considera a iteração ente duas variáveis categóricas em que uma delas tem K níveis e a outra tem L níveis, então temos que são criadas $(K-1)\times(L-1)$ novas variáveis, em que estas são o produto entre as $(K-1)$ variáveis *dummy* criadas para representar a variável categórica com K níveis e as $(L-1)$ variáveis *dummy* criadas para representar a variável categórica com L níveis.

De notar que tudo o que foi relatado nesta secção, como os testes de significância aos parâmetros dos modelos, assim como os métodos de seleção de variáveis, se aplicam da mesma forma quando se considera a interação entre variáveis.

3.3 Métodos de Contração (*Shrinkage*)

Os métodos de seleção de variáveis descritos na sub-subsecção 3.2.2.6 da secção 3.2, são processos nos quais se seleciona um subconjunto de variáveis explicativas para serem integradas no modelo. No entanto, apesar desta técnica fornecer um modelo interpretável, o processo pode ser extremamente variável, uma vez que é um processo discreto, já que as variáveis são retidas ou descartadas dos modelos em cada passo, e pequenas alterações no conjunto de dados podem resultar em modelos bastantes diferentes [13]. Por este motivo, a precisão de previsão pode reduzir-se [13].

Uma das suposições clássicas dos modelos de Regressão Linear e da Regressão Logística é que não existem associações entre as covariáveis, isto é, todas as variáveis explicativas são independentes entre si. No entanto, na prática, muitas das vezes esta suposição é violada, ou seja, tem-se que existem uma ou mais variáveis explicativas que são correlacionadas entre si [14]. Quando tal acontece tem-se o problema de multicolinearidade. O problema de multicolinearidade leva a um aumento da variância das estimativas dos coeficientes, podendo determinar incorretamente a significância dos coeficientes incorretamente, dificultando a especificação do modelo correto [15].

Assim, uma forma de resolver os problemas descritos anteriormente são os métodos de contração, que são métodos contínuos e estáveis, em que o "contínuo" se refere ao facto de este método estimar todos os coeficientes das suas variáveis explicativas num só passo, não descarta variáveis na estimação.

Os métodos de contração envolvem o ajustamento do modelo envolvendo todas as suas p variáveis explicativas [12]. No entanto, alguns dos coeficientes estimados do modelo são reduzidos a zero em relação às estimativas tradicionais. Esta contração, também conhecida como regularização, tem o efeito de reduzir a variância. Dependendo do tipo de contração que é utilizado, alguns dos coeficientes podem ser estimados exatamente como zero, pelo que os métodos de contração também pode servir para realizar a seleção de variáveis, tais como os métodos descritos na sub-subsecção 3.2.2.6 [12].

Quando nos modelos de Regressão Linear ou nos modelos de Regressão Logística

existem variáveis explicativas correlacionadas entre si, tal pode implicar que os seus coeficientes sejam determinados incorretamente e as suas estimativas apresentem uma grande variância, ou seja este problema pode-nos estar a remeter para o seguinte, um coeficiente positivo muito grande de uma variável explicativa pode ser cancelado por um coeficiente negativo de uma outra variável explicativa, em que estas duas variáveis explicativas são correlacionadas, assim, ao impor uma restrição de tamanho aos coeficientes, este problema pode ser atenuado [16].

As duas técnicas mais conhecidas para reduzir os coeficientes de regressão em direção a zero são a Regressão de LASSO e a Regressão de *Ridge* [12].

Antes de se avançar, para a sua exposição, apresentam-se a norma ℓ_1 e ℓ_2 de um dado vetor $A = (a_1, \ldots, a_n)$, nas equações 3.49 e 3.50, respetivamente.

$$\ell_1 = ||A||_1 = \sum_{i=1}^{n} |a_i| \tag{3.49}$$

$$\ell_2 = ||A||_2 = \left(\sum_{i=1}^{n} a_i^2 \right)^{\frac{1}{2}} \tag{3.50}$$

3.3.1 Regressão de *Ridge*

Na Regressão Linear, as estimativas dos parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ são dadas pelo método dos mínimos quadrados, ou seja, os parâmetros são estimados usando os valores que minimizam a soma de quadrados dos resíduos (*Residual Sum of Squares*-*RSS*), isto é, os valores que minimizam a quantidade 3.51 [12].

$$RSS = ||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{3.51}$$

A Regressão de *Ridge* é muito semelhante à regressão dos mínimos quadrados, no entanto os coeficientes são dados por uma quantidade ligeiramente diferente. Na Regressão de *Ridge*, as estimativas dos coeficientes $\hat{\boldsymbol{\beta}}^{ridge}$, são os valores que minimizam a quantidade 3.52, onde $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ e $\tilde{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)^T$, ou seja, o $\tilde{\boldsymbol{\beta}}$ são os coeficientes do modelo excluindo o parâmetro β_0, o intercepto,

$$||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 + \lambda ||\tilde{\boldsymbol{\beta}}||_2^2. \tag{3.52}$$

Ou seja, a Regressão de *Ridge* estima os coeficientes do modelo recorrendo à penalização ℓ_2 para atingir o modelo ajustado ideal.

Uma vez que $||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$ é a norma ℓ_2 (quadrática) dos resíduos e $||\tilde{\boldsymbol{\beta}}||_2 = \sum_{i=1}^{p} \beta_j^2$ é a penalidade da norma ℓ_2 em $\tilde{\boldsymbol{\beta}}$, então a equação 3.52, pode ser representada por 3.53.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=RSS+\lambda\sum_{j=1}^{p}\beta_j^2 \tag{3.53}$$

O parâmetro $\lambda \geq 0$ é um parâmetro de penalização, sendo calculado à parte. A seleção de um bom valor de λ é crítico e uma das formas para o determinar é através da validação cruzada, sendo que este método é apresentado de uma forma mais detalhada na secção 3.8. Neste caso, divide-se aleatoriamente a amostra em k grupos de tamanho igual ou aproximadamente igual e, de seguida, escolhe-se o primeiro grupo para serem os dados de teste e os demais grupos para serem os dados de treino ajustando-se a Regressão de *Ridge* com $\lambda = \lambda_0$ e utiliza-se esse modelo para tentar prever os dados de teste, calculando-se então o correspondente erro quadrático médio (*Mean Squared Error*). Repete-se este procedimento outras $(k-1)$ vezes para os restantes grupos. Após terminar as k iterações, teremos calculados k erros de predição, pelo que se calcula a sua média. Este procedimento será realizado para distintos valores de λ_0, sendo que o melhor valor de λ é o valor para o qual a média dos k erros de predição for a menor dentro de todas as médias dos erros de predição para os diferentes valores de λ_0. De notar que o erro quadrático médio (*Mean Squared Error-MSE*) é dado por 3.54 [12].

$$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{3.54}$$

Ao contrário do método dos mínimos quadrados, que produzem apenas um conjunto de estimativas de $\hat{\boldsymbol{\beta}}$, a Regressão de *Ridge* produz um conjunto de diferentes estimativas dos coeficientes $\hat{\boldsymbol{\beta}}^{ridge}$, pois estes dependem do valor de λ, daí a dificuldade de determinar um bom valor de λ.

Do mesmo modo que o método dos mínimos quadrados, a Regressão de *Ridge* procura as estimativas dos parâmetros que se ajustem bem aos dados tornando o *RSS* pequeno. No entanto, o segundo termo da quantidade 3.53, $\lambda\sum_{j=1}^{p}\beta_j^2$, chamado de penalidade de contração (*shrinkage penalty*), e uma vez que este é pequeno quando $\beta_1,\ldots,\beta_p$ são próximos de zero, resultam num efeito de contração das estimativas de β_j para zero. Este parâmetro de ajuste λ serve para controlar o impacto relativo dos dois termos na estimação dos coeficientes de regressão [12].

Quando $\lambda = 0$, tem-se que o termo de penalidade não tem efeito, e a Regressão de *Ridge* irá produzir as estimativas dos mínimos quadrados. No entanto, quando $\lambda \to \infty$, tem-se que a penalidade de contração cresce, e as estimativas dos coeficientes da Regressão de *Ridge* aproximam-se de 0 [17].

Note-se que o termo de penalidade de contração é somente aplicado a $\beta_1,\ldots,\beta_p$ mas não é aplicado ao interceto β_0, uma vez que apenas se está interessado em reduzir a associação estimada de cada variável explicativa com a variável resposta e não se está interessado em reduzir β_0, pois este é simplesmente uma medida do valor médio da variável resposta, quando $x_{i1}=x_{i2}=\cdots=x_{ip}=0$, não estando vinculado em nenhum preditor [12]. Note-se também que se cada uma das variáveis explicativas se encontrarem

padronizadas, isto é, cada uma das variáveis explicativas tiverem média 0 e desvio padrão 1, antes de se aplicar a Regressão de *Ridge*, então a estimativa de β_0 terá a forma apresentada em 3.55 [12].

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.55}$$

A penalização do termo β_0 faria com que o procedimento dependesse da origem escolhida para Y, a variável resposta, ou seja, adicionar uma constante c a cada uma das observações y_i não resultaria simplesmente num deslocamento das previsões na mesma quantidade c [17].

No *software* R a biblioteca *glmnet* tem uma função que permite estimar os parâmetros da Regressão de *Ridge*, assim como também tem uma função para estimar o melhor parâmetro de penalização λ através da validação cruzada. No entanto, estas duas funções em vez de estimarem as estimativas dos parâmetros $\boldsymbol{\beta}$ da Regressão Linear pelo método dos mínimos quadrados estimam estes parâmetros pelo método da máxima verosimilhança, em que no caso da Regressão Linear usual coincidem. No entanto, no presente trabalho não se irá utilizar esta biblioteca, pelo que para mais informações pode-se consultar [18].

3.3.2 Regressão de LASSO

A Regressão de LASSO é um método de contração tal como a Regressão de *Ridge*, mas com diferenças importantes uma vez que, na Regressão de *Rigde* o modelo final inclui todas as p variáveis explicativas. Os métodos descritos na secção 3.2.2.6, geralmente selecionam modelos que envolvem apenas um subconjunto das p variáveis explicativas, enquanto que a Regressão de *Ridge* inclui todas as p variáveis no modelo final, o que é uma desvantagem óbvia deste método. Os coeficientes deste método serão exatamente zero apenas quando $\lambda = \infty$, uma vez que a penalidade $\lambda \sum \beta_j^2$ em 3.53 encolherá todos os coeficientes para zero, mas não definirá nenhum deles exatamente a zero. A Regressão de LASSO é uma alternativa recente à Regressão de *Ridge*, que tenta superar esta desvantagem, utilizando no termo de penalidade a norma ℓ_1 em vez da norma ℓ_2.

Assim, os coeficientes de LASSO minimizam a seguinte quantidade,

$$||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 + \lambda||\tilde{\boldsymbol{\beta}}||_1^2 = RSS + \lambda||\tilde{\boldsymbol{\beta}}||_1^2. \tag{3.56}$$

A equação da 3.56, pode ser ainda representada por 3.57, uma vez que, $||\tilde{\boldsymbol{\beta}}||_1^2 = \sum_{i=1}^{n} |\beta_j|$.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \tag{3.57}$$

Tal como a Regressão de *Ridge* a Regressão de LASSO reduz as estimativas dos coeficientes para zero, no entanto a Regressão de LASSO consegue estimar alguns dos coeficientes exatamente como zero, ou seja, tem-se que o termo de penalidade ℓ_1 tem o efeito de forçar algumas estimativas dos coeficientes a serem exatamente iguais a zero.

Tal como na Regressão de *Ridge*, na Regressão de LASSO o parâmetro λ também pode ser estimado através de validação cruzada, como foi descrito na Regressão de *Ridge* 3.3.1.

3.3.3 LASSO de Grupo (*Group* LASSO)

Em muitos problemas de Regressão Linear nem todas as variáveis explicativas são contínuas, pelo que, em muitos casos tanto se tem variáveis explicativas contínuas como também se tem variáveis explicativas categóricas, com níveis. No entanto, neste caso, tem-se que a solução dado pela Regressão de LASSO não é satisfatória, uma vez que, esta seleciona apenas individualmente as variáveis *dummy* e não o fator por completo [17]. Além disso, a solução de LASSO depende de como as variáveis *dummy* estão codificadas, dado que ao escolher diferentes contrastes para uma variável explicativa categórica, este em geral produzirá diferentes soluções, ou seja, ao escolher diferentes formas de codificar a variável explicativa categórica para *dummies*, a Regressão de LASSO produzirá em geral diferentes soluções [19]. Assim, de modo a superar este problema, o LASSO de Grupo, introduz uma extensão adequada à penalidade de LASSO.

Suponha-se que as p variáveis explicativas são divididos em G grupos, em que p_g é o número de variáveis no grupo g, com $g = 1, \ldots, G$, por exemplo, imagine-se que se tinha 3 variáveis explicativas, onde uma era contínua, a outra do tipo *dummy* e a última é do tipo categórica com 3 níveis. Neste exemplo, tem-se que $G = 3$ e que $p_1 = 1, p_2 = 1$ e $p_3 = 2$. O p_3 é igual a dois, pois quando se converte uma variável categórica com k níveis em *dummy*, apenas se fica com $(k-1)$ variáveis *dummy*, por causa da categoria de referência, tal como explicado em 3.2.2.1. Para a presente notação seja $\mathbf{x}_g$ a matriz que representa as variáveis explicativas correspondente ao g-*ésimo* grupo, com o correspondente vetor de coeficientes β_g.

O LASSO de Grupo minimiza o critério convexo [17],

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(||\mathbf{y} - \beta_0 - \sum_{g=1}^{G} \mathbf{x}_g \boldsymbol{\beta}_g||_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}_g||_2 \right). \tag{3.58}$$

Em 3.58 o termo $\sqrt{p_g}$ leva em conta os diferentes tamanhos dos grupos e a $||.||_2$ é a norma ℓ_2. Aqui o $\hat{\boldsymbol{\beta}}_g$, ou seja, o vetor dos $\boldsymbol{\beta}$ estimados do g-*ésimo* grupo, tem a propriedade de que se o g-*ésimo* grupo é selecionado, então $\hat{\boldsymbol{\beta}}_g \neq 0$, isto é, $\hat{\beta}_{gt} \neq 0$, para todo o $t = 1, \ldots, p_g$, e caso contrário o g-*ésimo* grupo não seja selecionado o $\hat{\boldsymbol{\beta}}_g = 0$, ou seja, $\beta_{gt} = 0$, para todo o $t = 1, \ldots, p_g$ [17].

A biblioteca *grpreg* do *software* R, tem uma função que permite estimar os parâmetros de 3.58, em que esta biblioteca utiliza o algoritmo do grupo descendente (*group descent algorithms*) [20], para ajustar os parâmetros do modelo do LASSO de Grupo. No entanto, para se estimar o parâmetro de penalização λ, pode-se estimar este parâmetro através da validação cruzada. Na subsecção 3.3.5, encontra-se a descrição do respetivo algoritmo e da biblioteca *grpreg*.

3.3.4 Regressão de LASSO: Caso da Regressão Logística

A penalidade ℓ_1 usada na Regressão de LASSO, secção 3.3.2 pode ser usada para a seleção de variáveis e contração em qualquer modelo de Regressão Linear. Na Regressão Logística, em vez de se minimizar o *RSS*, quer-se minimizar o simétrico da log-verosimilhança, pelo que os coeficientes do LASSO, no caso da Regressão Logística, quando a variável resposta segue uma distribuição de Bernoulli, minimizam a quantidade 3.59 [17].

$$\min_{(\beta_0,\boldsymbol{\beta}^T)\in R^{(p+1)}} \left(-\sum_{i=1}^{N} \left[y_i(\beta_0 + \boldsymbol{\beta}^T \mathbf{x}_i) - \ln(1 + e^{\beta_0+\boldsymbol{\beta}^T \mathbf{x}_i}) \right] + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{3.59}$$

Tal como na Regressão de LASSO em 3.3.2 não se penaliza o parâmetro β_0, representando $\boldsymbol{\beta}$ o vetor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

3.3.4.1 LASSO de Grupo: Caso da Regressão Logística

Tal como no caso da Regressão Linear, em muitos problemas de Regressão Logística, algumas das variáveis explicativas são categóricas, pelo que como referido anteriormente o LASSO não leva em conta o grupo das variáveis, pois este seleciona individualmente as variáveis *dummy* que compõem esta variável categórica e não seleciona o fator por completo. Assim, também se pode aplicar o LASSO de Grupo no caso da Regressão Logística.

No LASSO de Grupo no caso da Regressão Logística, em que a variável resposta segue uma distribuição de Bernoulli, os coeficientes do LASSO de Grupo minimizam a quantidade 3.60 [19].

$$-\sum_{i=1}^{n} \left[y_i \left(\beta_0 + \sum_{g=1}^{G} \mathbf{x}_g \boldsymbol{\beta}_g \right) - \ln \left[1 + e^{\beta_0 + \sum_{g=1}^{G} \mathbf{x}_g \boldsymbol{\beta}_g} \right] \right] + \lambda \sum_{g=1}^{G} \sqrt{p_g} ||\boldsymbol{\beta}_g||_2 \tag{3.60}$$

Tal como no LASSO de Grupo na Regressão Linear, visto na subsecção 3.3.3, $\mathbf{x}_g$ representa a matriz das variáveis que representam o *g-ésimo* grupo, onde $g = 1, \ldots, G$. O vetor $\boldsymbol{\beta}_g$ corresponde ao vetor de parâmetros correspondente ao *g-ésimo* grupo. Já o p_g, representa o número de variáveis que estão no *g-ésimo* grupo. Não esquecer que um grupo que seja constituída por um variável categórica com k níveis, o número de variáveis nesse grupo são as $(k-1)$ variáveis *dummy*.

Aqui, tal como no caso da Regressão de LASSO no caso da Regressão Logística, a única coisa que muda quando comparado com a Regressão Linear é que em vez de se ter o *RSS* tem-se o simétrico da log-verosimilhança. Neste caso, o parâmetro λ também pode ser estimado por validação cruzada, tal como já foi descrito anteriormente.

Aqui o $\hat{\boldsymbol{\beta}}_g$, ou seja, o vetor dos $\boldsymbol{\beta}$ estimados do *g-ésimo* grupo, tem a propriedade de que se o *g-ésimo* grupo é selecionado, então $\hat{\boldsymbol{\beta}}_g \neq 0$, isto é, $\hat{\beta}_{gt} \neq 0$, para todo o $t = 1, \ldots, p_g$, e caso contrário do *g-ésimo* grupo não ser selecionado o $\hat{\boldsymbol{\beta}}_g = 0$, ou seja, $\beta_{gt} = 0$, para todo o $t = 1, \ldots, p_g$ [17].

3.3.5 Biblioteca *grpreg*

A biblioteca *grpreg* do *software* R, tem uma função que permite estimar os parâmetros de 3.58 e de 3.60, onde esta biblioteca utiliza o algoritmo do grupo descendente (*group descent algorithms*) [20], para ajustar os parâmetros do modelo LASSO de Grupo. No entanto, para se estimar o parâmetro de penalização λ, usa-se a validação cruzada, onde se utiliza a função *cv.grpreg()*, desta biblioteca. Esta função tem como argumentos o número de grupos que se pretende considerar na validação cruzada que por defeito é 10, a matriz dos valores observados das variáveis explicativas $\mathbf{X}$, os valores observados da variável resposta $\mathbf{y}$, um vetor com o grupo das variáveis explicativas da matriz $\mathbf{X}$, a família da variável resposta, se é Normal ou Binomial, e também o tipo de penalização que se está interessado em aplicar ao modelo, em que o "*grLasso*", é a penalização por LASSO de Grupo. Nesta biblioteca, em vez de se minimizar as quantidades representadas em 3.58 e 3.60, otimiza-se a função objetivo representada em 3.61.

$$Q(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \frac{1}{n}L(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) + P_{\lambda}(\boldsymbol{\beta}) \tag{3.61}$$

No Apêndice A tem-se, descrito o algoritmo que esta biblioteca utiliza para estimar os parâmetros $\boldsymbol{\beta}$.

De referir que o λ selecionado pela validação cruzada é o que minimiza o erro de validação cruzada, neste caso baseado na *deviance*.

Na validação cruzada a função *cv.grpreg()* utiliza uma sequência de valores de λ que são determinados automaticamente numa escala de valores varia uniformemente na escala logarítmica sobre o intervalo dos valores de λ, onde o intervalo da escala dos valores de λ é a que está representada no Apêndice A.

Para cada valor de λ, o algoritmo divide os dados em 10 grupos, sendo que um dos grupos é considerado para teste e os restantes 9 grupos para treino. No primeiro passo, o primeiro grupo é considerado para teste enquanto que os restantes 9 grupos são considerados para treino sendo o modelo ajustado aos dados do grupo de treino. Este passo é repetido 10 vezes, pois têm-se 10 grupos. Para cada grupo de teste calcula-se o erro do modelo com base na *deviance*, sendo que no final faz-se a média destes 10 erros obtidos. Este processo é repetido 100 vezes, pois por defeito esta função considera 100 valores de λ diferentes. O valor de λ escolhido é aquele a que corresponde ao menor erro de validação cruzada.

Nesta função depois de ajustado o modelo, esta tem um objeto que é o λ_{min} que corresponde ao menor valor do erro de validação cruzada.

Para mais informações sobre esta biblioteca consultar [21]. Para mais informações sobre o algoritmo descrito na Apêndice A, consultar [20].

3.4 Modelos Aditivos Generalizados

3.4.1 Introdução

Os Modelos Aditivos Generalizados (GAMs) são Modelo Lineares Generalizados, em que as variáveis explicativas se podem relacionar de forma não linear, através de uma função suavizadora, com a média da variável resposta (ou sua função) [17].

Os Modelos Lineares Generalizados são relativamente simples de descrever e implementar e ainda têm a vantagem em termos de interpretação e inferência sobre outras abordagens [12]. No entanto os Modelos Lineares Generalizados podem ter limitações significativas no que diz respeito ao seu poder preditivo, uma vez que estes modelos são lineares nas variáveis explicativas e isto nem sempre acontece, pelo que estes modelos podem acabar por construir uma aproximação que às vezes acaba por ser pobre [12].

Os Modelos Aditivos Generalizados são uma extensão dos Modelos Lineares Generalizados, onde as variáveis explicativas não têm necessariamente de ter uma relação linear com a variável resposta [12].

Os Modelos Lineares Generalizados assumem a seguinte forma:

$$E(\mathrm{Y}|x_1, x_2, \ldots, x_p) = h(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p) \tag{3.62}$$

onde h é uma função de ligação.

Os Modelos Aditivos Generalizados têm a seguinte forma:

$$E(\mathrm{Y}|x_1, x_2, \ldots, x_p) = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) \tag{3.63}$$

onde as funções f_j com $k = 1, \ldots, p$, do modelo 3.63, são funções suavizadoras não especificadas nas variáveis que não assumirem a linearidade. Nas variáveis que assumirem a linearidade, tem-se que $f_j(x_j) = \beta_j x_j$ [17].

Tal como nos Modelos Lineares Generalizados, os Modelos Aditivos Generalizados, tanto podem ser aplicados a variáveis resposta quantitativas como a variáveis resposta qualitativas (com qualquer distribuição da família exponencial), como o caso de variáveis resposta dicotómicas [12].

Nas subsecções 3.4.2, 3.4.3, 3.4.5, 3.4.6 e 3.4.6, encontram-se descritos vários tipos de funções suavizadoras f, para o caso univariado, ou seja, para o caso em que apenas se tem uma única variável explicativa.

Já nas subsecções 3.4.8 e 3.4.9, encontram-se descrito o Modelo Aditivo Generalizado, para o caso da Regressão Linear e para o caso da Regressão Logística, respetivamente, enquanto que na subsecção 3.4.7, encontra-se descrito a seleção automática do parâmetro de suavização, para os *splines* de suavização.

3.4.2 Regressão Polinomial

A regressão polinomial estende o modelo de Regressão Linear através da adição de variáveis explicativas extras, em que estas são dadas como potências de cada uma das

variáveis explicativas originais. Esta é uma abordagem simples de fornecer um ajuste não linear aos dados. Por exemplo, no caso da regressão cúbica, tem-se que as variáveis explicativas que o modelo usa são as seguintes, x, x^2 e x^3.

Assim, a forma de estender o modelo de Regressão Linear simples para um modelo em que a variável explicativa não tem uma relação linear com a variável resposta é substituir o modelo dado pela equação 3.64,

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n \tag{3.64}$$

por uma função polinomial, dado pelo modelo da equação 3.65, em que o ε_i é o termo de erro (com os pressupostos habituais) [12].

$$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \varepsilon_i \tag{3.65}$$

O modelo da equação 3.65, é conhecido como regressão polinomial, em que para um grau suficientemente grande d, a regressão polinomial consegue produzir exatamente uma curva não linear [12].

Neste caso, tem-se que os parâmetros β_j, com $j = 0, \ldots, d$ do modelo da equação 3.65 podem ser facilmente estimados de igual modo como os parâmetros do modelo da equação 3.64, ou seja, através do método dos mínimos quadrados, uma vez que se está presente num caso da Regressão Linear simples, em que as variáveis explicativas são $x_i^1, x_i^2, \ldots, x_i^d$, onde $i = 1, \ldots, n$ [12].

Já a extensão da regressão polinomial no caso do modelo da Regressão Logística univariado, equação 3.4, é dada pelo modelo da equação 3.66 [12].

$$P(\mathrm{Y}_i = 1|x_i) = \pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i+\beta_2 x_i^2+\cdots+\beta_d x_i^d}}{1 + e^{\beta_0+\beta_1 x_i+\beta_2 x_i^2+\cdots+\beta_d x_i^d}} \tag{3.66}$$

3.4.3 Funções por Troços

As funções por troços convertem variáveis contínuas em variáveis categóricas ordenadas, ou seja, para este caso, dividem os valores da variável explicativa x em diferentes compartimentos e para cada um destes compartimentos ajustam uma constante diferente. Neste método cria-se k pontos de corte na variável explicativa x, denotados por $c_1, \ldots, c_k$, e de seguida criam-se $(k+1)$ novas variáveis fictícias, definidas pela expressão 3.67 [12].

$$\begin{cases} C_0(x) = \mathrm{I}(x < c_1), \\ C_i(x) = \mathrm{I}(c_i \leq x < c_{i+1}) \quad \text{com} \quad i = 1, \ldots, k-1 \\ C_k(x) = \mathrm{I}(x \geq c_k) \end{cases} \tag{3.67}$$

A função I(.) é uma função indicatriz, ou seja, esta função toma o valor 1 se a condição for verdadeira e toma o valor zero caso contrário [12]. Note-se que para qualquer que seja o valor de x, tem-se que $C_0(x) + C_1(x) + \cdots + C_k(x) = 1$, uma vez que o valor de x há-de estar

presente apenas num dos $(k+1)$ intervalos definidos anteriormente. Posto isto, tem-se que o modelo de Regressão Linear simples, que se encontra representado na equação 3.64, passa a ser dado pelo modelo da equação 3.68, quando se substitui a variável explicativa x pelas $(k+1)$ novas variáveis fictícias de x [12].

$$y_i = \beta_0 + \beta_1 C_1(x) + \cdots + \beta_k C_k(x) + \varepsilon_i \tag{3.68}$$

Observe-se que a variável fictícia $C_0(x)$ não se encontra representada no modelo da equação 3.68, uma vez que é redundante com o parâmetro do β_0, pois não se pode esquecer que estas novas variáveis fictícias funcionam como as variáveis categóricas, isto é, num modelo de regressão em que se tem uma variável explicativa categórica com l níveis, tem-se que se convertem $(l-1)$ níveis desta em $(l-1)$ variáveis *dummy*, sendo que a outra variável *dummy* está representada no termo do interceto do modelo, isto é, no parâmetro β_0 [12]. Neste caso, excluiu-se a variável $C_0(x)$, mas poderia optar-se por ficar com esta variável no modelo e excluir uma das restantes variáveis $C_i(x)$, com $i = 1, \ldots, k$, sendo que esta passaria a ser o termo do β_0 [12]. Também ainda se poderia excluir o termo do β_0 e incluir a variável fictícia $C_0(x)$ no lugar do termo do β_0 [12].

Mais uma vez pode-se utilizar o método dos mínimos quadrados para se ajustar o modelo da equação 3.68, isto é, pode-se utilizar o método dos mínimos quadrados para estimar os parâmetros $\beta_0, \beta_1, \ldots, \beta_k$ [12].

No caso do modelo da Regressão Logística, quando se ajusta este com as $(k+1)$ novas variáveis fictícias de x tem-se que o modelo da equação 3.4, passa a ser dado pelo modelo da equação 3.69 [12].

$$P(\mathrm{Y}_i = 1|x_i) = \pi(x_i) = \frac{e^{\beta_0+\beta_1 C_1(x)+\cdots+\beta_k C_k(x)}}{1 + e^{\beta_0+\beta_1 C_1(x)+\cdots+\beta_k C_k(x)}} \tag{3.69}$$

3.4.4 Representação de uma Função com Expansão de Base

Os modelos de regressão polinomiais e de regressão por troços vistos nas subsecções 3.4.2 e 3.4.3, respetivamente são um caso especial da abordagem de funções de base. A ideia de uma função de base é ter uma família de funções ou de transformações $b_1(), b_2() \ldots, b_k()$ que podem ser aplicadas a uma variável x e que a constituem [12]. Assim, em vez de se ajustar um modelo linear em x, ajusta-se o modelo representado em 3.70.

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_k(x_i) + \varepsilon_i \tag{3.70}$$

As funções que formam a base são fixas e conhecidas [12]. No caso da regressão polinomial, tem-se que as funções de base são do tipo $b_j(x_i) = x_i^j$, enquanto que na regressão por troços as funções de base são dadas por $b_j(x_i) = \mathrm{I}(c_j \leq x_i < c_{j+1})$ [12].

O modelo 3.70 pode ser visto como um modelo de regressão simples em que as variáveis explicativas são $b_1(x_i), b_2(x_i), \ldots, b_k(x_i)$. Assim, para estimar os parâmetros desconhecidos β_j, $j = 1, \ldots, k$, do modelo 3.70 pode-se utilizar o método dos mínimos quadrados [12].

Na subsecção 3.4.5, tem-se uma escolha muito especial para as funções de base, que dão origem aos *splines*.

3.4.5 Regressão por *Splines*

A regressão por *splines* é um método mais flexível que a regressão polinomial e também que as funções por troços, uma vez que a regressão por *splines*, divide a variável explicativa x em K regiões distintas e para cada uma destas regiões ajusta uma função polinomial ao conjunto de dados, pelo que a regressão por *splines* acaba por ser uma extensão da regressão polinomial e das funções por troços [12]. Cada um dos polinómios ajustados são restringidos de forma a que, para cada um dos nós os vários polinómios ajustados em cada uma das regiões se unam de uma forma suave [12].

3.4.5.1 Polinómios por Troços

Na regressão polinomial, ajusta-se um polinómio de um determinado grau em todo o intervalo da variação da variável explicativa x [12]. Por vezes, em vez de se ajustar um polinómio de grau elevado em todo o intervalo de variação da variável explicativa x pode ser útil a aplicação de polinómios de um grau mais baixo nas diferentes regiões do suporte de x, ou seja, tem-se que a regressão polinomial por troços envolve o ajustamento de polinómios de grau mais baixo nas diferentes regiões do suporte de x [12].

Por exemplo, um polinómio cúbico por troços, funciona ajustando o modelo representado em 3.65, onde $d = 3$, mas onde os coeficientes $\beta_0, \beta_1, \beta_2$ e β_3 diferem em diferentes partes do intervalo da variável explicativa x [12]. Os pontos em que os coeficientes mudam chamam-se *nós* (*knots*) [17]. Assim, um polinómio cúbico por troços sem *nós*, tem a forma da equação 3.65, enquanto que um polinómio cúbico por troços com um simples *nó* no ponto c tem a forma de 3.71 [12].

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i, & x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i, & x \geq c \end{cases} \tag{3.71}$$

Assim, neste caso tem que se ajustar duas diferentes funções polinomiais aos dados, uma para o subconjunto em que $x_i < c$, em que esta tem como coeficientes $\beta_{01}, \beta_{11}, \beta_{21}$ e β_{31} e a outra função polinomial para o subconjunto em que $x_i \geq c$, em que $\beta_{02}, \beta_{12}, \beta_{22}$ e β_{32} são os coeficientes a estimar. Cada uma das funções polinomiais, pode ser ajustada usando mais uma vez mais o método dos mínimos quadrados [12].

Em vez de se utilizar um único *nó*, pode-se levar em conta diferentes K *nós* em todo o intervalo de x, o que isto leva a um polinómio por troços mais flexível, mas isto levará a ajustar $(K + 1)$ funções polinomiais [12]. Um dos problemas dos polinómios por troços é que a função pode ser descontínua no *nó* [12]. Uma forma de remediar este problema, é ajustar o polinómio por troços adicionando a restrição de que a curva ajustada deve de ser contínua nos *nós*, ou seja, num determinado *nó*, não pode haver um salto na curva

ajustada [17]. Por exemplo, para se ultrapassar o problema da descontinuidade no *nó c* do exemplo do polinómio cúbico por partes representado em 3.71 é impor a condição que está explicita em 3.72 [17].

$$\beta_{01} + \beta_{11}c + \beta_{21}c^2 + \beta_{31}c^3 = \beta_{02} + \beta_{12}c + \beta_{22}c^2 + \beta_{32}c^3 \tag{3.72}$$

No entanto, com esta condição ainda se tem um problema, na junção da curvas no *nó c*, pois estamos perante o exemplo do polinómio por troços de 3.71, a curva apresenta uma forma de V, esta junção não parece natural, ou seja, a curva não é suavizada [12]. Assim, para se resolver este problema adiciona-se outras duas condições, em que se exige que a primeira e a segunda derivadas sejam contínuas no *nó c*, ou seja, exige-se que o polinómio por troços seja contínuo no *nó c*, mas que também que a curva seja suave nesse *nó* [12]. A curva ajustada neste caso, é chamada de *spline* cúbico [12]. Na sub-subsecção 3.4.5.2 define-se o conceito de *spline* de grau d.

3.4.5.2 *Splines*

Por definição, um *spline* de grau d é um polinómio por troços de grau d, contínua nos seus *nós* e em que as suas derivadas até á ordem $(d-1)$ são contínuas em cada um dos seus *nós*, ou seja, um *spline* de ordem d, com os *nós* ξ_j com $j = 1, \ldots, K$, é um polinómio por troços, contínuo em cada um dos seus *nós* ξ_j, e para além disso também se exige que as suas derivadas até à ordem d sejam contínuas nos seus *nós* ξ_j [12]. Por exemplo, como referido na sub-subsecção 3.4.5.1, um *spline* de grau $d = 3$, tem a primeira e a segunda derivadas contínuas e este *spline* é denominado por *spline* cúbico.

Assim, um *spline* de grau d pode ser representada como uma série de potências, dado pela equação 3.73, onde ξ_j, $j = 1, \ldots, K$ é uma sequência de *nós*, tal que $\xi_1 < \xi_2 < \ldots < \xi_K$ constitui uma partição do suporte de x.

$$S(x) = \sum_{j=0}^{d} \beta_j x^j + \sum_{j=1}^{K} \gamma_j (x - \xi_j)_+^d \tag{3.73}$$

Na equação 3.73, tem-se representado a expressão $(x - \xi_j)_+$, onde a notação desta expressão encontra-se representada na equação 3.74.

$$(x - \xi_j)_+ = \begin{cases} x - \xi_j, & x > \xi_j \\ 0, & \text{caso contrário} \end{cases} \tag{3.74}$$

Este *spline* de *nós* fixos definidos anteriormente também são conhecidos como *splines* de regressão [17].

Posto isto, tem-se que um *spline* cúbico com K *nós*, ξ_j com $j = 1, \ldots, K$ tem a representação dada por 3.75.

$$S(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{K} \gamma_j (x - \xi_j)_+^3 \tag{3.75}$$

Assim, pode-se utilizar o modelo dado em 3.70, para representar um *spline* de regressão, onde o *spline* tem grau d e com K *nós* pode ser modelado como se encontra em 3.76.

$$y_i = \beta_0 + \beta_1 x + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \sum_{j=1}^{K} \beta_{d+j}(x_i - \xi_j)_+^d + \varepsilon_i \tag{3.76}$$

Note-se que em 3.76, os $b_j()$ do modelo 3.70 são dados por 3.77.

$$\begin{cases} b_j(x_i) = x_i^j, & j = 1, \ldots, d \\ b_{d+k}(x_i) = (x_i - \xi_k)_+^d, & k = 1, \ldots, K \end{cases} \tag{3.77}$$

Por outro lado, também se observa que ao se ajustar um *spline* com grau d e com K *nós*, tem-se um modelo com $\Big((d+1)+K\Big)$ parâmetros, pelo que, o modelo tem agora $(d+K)$ variáveis explicativas. Mais uma vez, como se está presente num modelo simples de regressão, para se ajustar o modelo dado em 3.76, pode-se utilizar o método dos mínimos quadrados, em que neste caso se tem um termo de intercepto e $(d+K)$ preditores, dados em 3.77.

Os *splines* também se podem aplicar a uma variável resposta dicotómica. No entanto, neste caso o modelo a ajustar assume a forma de 3.78, onde os parâmetros $\beta_j, j = 1, \ldots, \Big((d+1)+K\Big)$, são estimados usando o mesmo procedimento que no caso univariado do modelo de Regressão Logística, como se viu na subsecção 3.2.1, onde nesta secção se encontra o método de estimação para o caso univariado do modelo da Regressão Logística.

$$\ln\left(\frac{P(\mathrm{Y}_i = 1|x_i)}{1 - P(\mathrm{Y}_i = 1|x_i)}\right) = \beta_0 + \beta_1 x + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \sum_{j=1}^{K} \beta_{d+j}(x_i - \xi_j)_+^d \tag{3.78}$$

Um *spline* natural é um *spline* de regressão, em que este tem restrições de limite adicionais, em que estas condições de limite adicionais são que a função deve de ser linear na região onde a variável explicativa x tem valores menores que o seu menor *nó*, assim, como também a função deve de ser contínua para os valores da variável explicativa x maiores que o seu maior valor de *nó* [12].

3.4.6 *Splines* de Suavização

Nesta secção apresenta-se uma outra nova abordagem de *splines* em que não é necessário especificar um conjunto de *nós* para a construção dos *splines*, evitando completamente o problema da seleção de *nós*.

Nesta nova abordagem, está-se interessado em ajustar uma curva suavizada ao conjunto de dados, ou seja encontrar uma dada função, denotemo-la por $g(x)$, que se ajuste relativamente bem ao conjunto de dados [12]. Assim, está-se interessado em obter uma função $g(x)$ de tal modo que a quantidade representada em 3.79, seja a mais pequena possível [12].

$$RSS = \sum_{i=1}^{n} \Big(y_i - g(x_i) \Big)^2 \tag{3.79}$$

A quantidade representada em 3.79, é simplesmente a soma de quadrados dos resíduos (*RSS*).

É possível tornar o *RSS* como zero, bastando simplesmente escolher a função g de modo que interpole todos os pontos y_i [12]. Porém neste caso, teríamos uma função flexível de mais, ou seja, a função estaria super-adaptada aos dados, e o que realmente queremos é uma função g que torne o *RSS* pequeno, mas que também seja suave [12].

Uma forma de garantir que a função g seja suave é encontrar uma função g que minimize a quantidade 3.80 [12].

$$RSS(g, \lambda) = \sum_{i=1}^{n} \Big(y_i - g(x_i) \Big)^2 + \lambda \int g''(t)^2 dt \tag{3.80}$$

A quantidade representada na equação 3.80 é a soma de quadrados dos resíduos adicionada de um termo de ajuste não negativo, ($\lambda \geq 0$) [17]. A função g que minimiza a quantidade 3.80, é conhecida como *spline* de suavização [12]. O termo $\sum_{i=1}^{n} (y_i - g(x_i))^2$ é a função de perda que encoraja g a ajustar-se bem aos dados, ou seja, este termo mede a proximidade de g aos dados, enquanto que o termo $\lambda \int g''(t)^2 dt$ é um termo de penalidade que penaliza a variabilidade de g, a curvatura da função, obrigando a função g a ser suave, e λ estabelece um compromisso entre os dois termos [17].

Na equação 3.80, tem-se que $g''(t)$ representa a segunda derivada da função g. A primeira derivada da função g mede a inclinação da função no ponto t, enquanto que a segunda derivada corresponde à variação da inclinação [12].

Quando $\lambda = 0$, então o termo de penalidade de 3.80 não tem efeito, pelo que a função g será muito instável e irá interpolar exatamente as observações treino [12]. Já quando $\lambda \to \infty$, g será perfeitamente suave, ou seja, será uma linha reta que passa o mais próximo possível às observações de treino [12]. De facto, neste caso, g será a reta linear dos mínimos quadrados, pois a função de perda em 3.80 equivale a minimizar a soma de quadrados dos resíduos[12]. Para um valor intermediário de λ, g aproximará as observações de treino, mas será um pouco suave [12].

Pode-se mostrar que a função $g(x)$ que minimiza a quantidade 3.80 tem algumas propriedades especiais, entre as quais, é um polinómio cúbico por partes com *nós* nos únicos valores de $x_1, \ldots, x_n$ e tem a primeira e a segunda derivadas contínuas em cada um destes *nós* [12]. Além disso esta função $g(x)$ ainda é linear nos *nós* extremos [12]. Ou seja, em outras palavras é um *spline* cúbico natural com os *nós* em $x_1, \ldots, x_n$. No entanto não é o mesmo *spline* cúbico natural que se obteria se aplicássemos a abordagem da função básica descrita em 3.4.5.2, com os *nós* em $x_1, \ldots, x_n$, pois neste caso é uma versão contraída do tal *spline*, onde o valor do parâmetro λ em 3.80 controla o nível de contração [12].

No caso da Regressão Logística, está-se interessado em maximizar a quantidade 3.81, em vez da quantidade de 3.80, ou seja, em vez de se ter a minimização do *RSS* com o

termo de penalidade, tem-se a maximização da log-verosimilhança com o parâmetro de penalidade [17],

$$\begin{aligned}&\sum_{i=1}^{n}\left(y_i \ln\left(\pi(x_i)\right) + (1-y_i)\ln\left(1-\pi(x_i)\right)\right) - \frac{1}{2}\lambda\int g''(t)^2 dt \\ &= \sum_{i=1}^{n}\left(y_i f(x_i) + (1-y_i)\ln(1+e^{f(x_i)})\right) - \frac{1}{2}\lambda\int g''(t)^2 dt\end{aligned} \tag{3.81}$$

pois uma vez que, o Modelo Aditivo Generalizado, no caso do modelo da Regressão Logística com uma simples variável quantitativa x é, dado por 3.82.

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = f(x) \tag{3.82}$$

3.4.7 Seleção Automática do Parâmetro de Suavização (*Smoothing*)

Os parâmetros de suavização para a regressão de *splines* abrangem o grau dos splines e o número e a alocação dos *nós* [12]. Para suavizar os *splines*, tem-se apenas o parâmetro de penalidade λ para selecionar, uma vez que os nós são em todos os pontos de treino $x_1, \ldots, x_n$. No entanto, é possível especificar o grau do *spline* em vez do parâmetro λ, e aí o parâmetro λ é calculado com base no grau do *spline* [17].

Seja $\hat{g}_\lambda$ a solução de 3.80, para um valor particular de λ, então podemos escrever $\hat{g}_\lambda$ como se encontra na equação 3.83 [12].

$$\hat{g}_\lambda = S_\lambda \mathbf{y} \tag{3.83}$$

O S_λ é a matriz de suavização, em que esta depende somente dos valores de x_i e de λ [12]. A equação 3.83 indica que o vetor de valores ajustados ao aplicar um *spline* de suavização aos dados pode ser escrito como uma matriz S_λ $(n \times n)$, (para a qual existe uma fórmula, mas não a iremos especificar, neste presente caso), vezes o vetor de resposta. Então o grau do *spline* de suavização é definido como 3.84 [12].

$$dg_\lambda = \sum_{i=1}^{n}\{S_\lambda\}_{ii} = tr(S_\lambda) \tag{3.84}$$

Na equação 3.84, o termo $tr(S_\lambda)$, significa o traço da matriz S_λ, ou seja, é a soma dos elementos da diagonal principal da matriz S_λ [12]. Pode-se mostrar que $dg_\lambda = tr(S_\lambda)$ é monótona em λ, pelo que para se suavizar *splines*, pode-se inverter a relação, e especificar λ através da fixação de dg [17]. Na prática isto pode ser alcançado através de métodos numéricos simples. O *software* R, tem a função *smooth.spline()*, que pode ser utilizada para especificar a quantidade de suavização, através do argumento df dessa função [17].

No entanto, se não se especificar o grau do *spline* de suavização, pode-se utilizar a técnica de validação cruzada com k grupos, para cada valor de λ utilizado, para se obter

o melhor valor de λ, em que o erro de validação cruzada com k grupos, para cada valor de λ é dado por 3.85 [12].

$$CV(\hat{g}_\lambda) = \frac{1}{K}\sum_{i=1}^{n}\left(y_i - \hat{g}_\lambda^{(-i)}(x_i)\right)^2 \tag{3.85}$$

A notação $\hat{g}_\lambda^{(-i)}(x_i)$ em 3.85, representa o valor ajustado para este *spline* de suavização avaliado em x_i, onde o ajuste usa todas as observações de treino, exceto a *i-ésima* observação (y_i, x_i) [12].

O valor de λ selecionado é aquele a que corresponde o menor valor do erro de validação cruzada.

3.4.8 Modelo Aditivo Generalizado - Caso da Regressão Linear

Um problema de Regressão Linear múltipla assume a forma de 3.86.

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{3.86}$$

Para permitir relações não lineares entre cada variável explicativa e a variável resposta, basta substituir cada componente linear de $\beta_j x_{ij}$ por uma função não linear (suave) f_j. Assim, obtemos o modelo na forma 3.87.

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i \tag{3.87}$$

O modelo representado em 3.87 é um exemplo de um Modelo Aditivo Generalizado. Este modelo é chamado de modelo aditivo, pois calcula-se separadamente f_j para cada x_j e, em seguida, soma-se todas as suas contribuições [12].

As funções f_j, com $j = 1, \cdots, p$, são funções não especificadas suavizadas, que podem ser *splines*, funções polinomiais, ou uma função linear quando a variável explicativa x_j assume a linearidade e neste caso $f_j(x_j) = \beta_j x_j$, ou combinações destas [12].

Dadas observações $(y_i, \mathbf{x}_i)$, um critério como a soma de quadrados dos resíduos penalizada, como em 3.80, pode ser especificada para este problema, sendo que neste caso a soma de quadrados dos resíduos penalizada encontra-se na equação 3.88,

$$PRSS(\alpha, f_1, f_2, \ldots, f_p) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} f_j(x_{ij})\right)^2 + \sum_{j=1}^{p}\lambda_j \int f_j''(t_j)^2 dt_j \tag{3.88}$$

onde $\lambda_j \geq 0$ são parâmetros de ajuste [17]. As funções $f_1, \ldots, f_p$ são estimadas pelo algoritmo *backfitting*, onde se pode consultar [22], para mais informações, sobre a estimação.

3.4.9 Modelo Aditivo Generalizado - Caso da Regressão Logística

Vimos anteriormente na secção 3.2, que o modelo de Regressão Logística era dado por:

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{3.89}$$

A parte esquerda da equação 3.89, é o logaritmo das *chances*, isto é, a equação 3.89 é o logaritmo da probabilidade condicional da variável dependente ser igual a 1 dados $\mathbf{x}$, $P(\mathrm{Y} = 1|\mathbf{x})$ versus a probabilidade condicional da variável dependente ser igual a 0 dados $\mathbf{x}$, $P(\mathrm{Y} = 0|\mathbf{x})$, como mencionada na secção 3.2. A equação 3.89 é uma função linear nas covariáveis, isto é, nas variáveis explicativas [12]. Assim, uma forma natural de estender o modelo 3.89, de forma a este permitir relações não lineares, é considerar o seguinte modelo [12]:

$$\ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p). \tag{3.90}$$

A diferença do modelo 3.89 para o modelo 3.90 é que no modelo 3.90, se uma dada covariável j, $j = 1, \ldots, p$ não apresentar uma relação linear, basta substituir a componente linear j do modelo 3.89, $\beta_i x_i$, por uma função não linear suavizante, $f_i(x_i)$, sendo que as covariáveis que apresentem uma relação linear o $f_j(x_j) = \beta_j x_j$.

Como referido anteriormente as funções f_j, com $j = 1, \ldots, p$, são funções não especificadas suavizadas, em que estas podem ser *splines*, funções polinomiais, ou uma função linear quando a variável explicativa x_j assume a linearidade e neste caso $f_j(x_j) = \beta_j x_j$, ou combinações destas.

As funções $f_1, \ldots, f_p$ são estimadas, através da maximização da log-verosimilhança penalizada, como dada em 3.91, pelo que algoritmo que é utilizado para a maximização deste é *backfitting* dentro de um o procedimento de *Newton-Raphson*, em que este é denotado por *Local Scoring Algorithm for the Additive Logistic Regression Model*, onde se pode consultar [22], para mais informações, sobre a estimação.

$$\sum_{i=1}^{n}\left(y_i\left(\beta_0 + \sum_{j=1}^{p} f_j(x_{ij})\right) + (1-y_i)\ln\left(1 + e^{\beta_0 + \sum_{j=1}^{p} f_j(x_{ij})}\right)\right) - \frac{1}{2}\lambda\int g''(t)^2 dt \tag{3.91}$$

Uma maneira de se comparar diferentes modelos Aditivos Generalizados, ajustados a um mesmo conjunto de dados, pode ser através da medida *AIC*, assim, como também se pode comparar um modelo de Regressão Logística, com um modelo Aditivo Generalizado através da medida do *AIC* [23].

Como visto anteriormente, temos que o Critério de informação de *Akaike*, é uma medida de ajuste para escolher o melhor modelo ajustado entre um dado número de modelos ajustados [24]. Na secção 3.2, o *AIC* era dado por 3.45, onde a agora a log-verosimilhança dessa equação corresponde à log-verosimilhança de um modelo Aditivo Generalizado. Tal como referido anteriormente, o modelo que apresentar o menor *AIC*, é aquele que melhor se ajusta aos dados [24].

3.4.10 Biblioteca *gam*

O *software* R tem a biblioteca *gam* que pode ser utilizado para ajustar um Modelo Aditivo Generalizado. A função desta biblioteca que permite ajustar um Modelo Aditivo Generalizado é a função $gam()$. Esta função utiliza o algoritmo *backfitting*, para ajustar os diferentes *splines* de suavização. Esta função tem como argumentos a "formula", onde este argumento é uma expressão da forma "response predictors". Outro dos argumentos é a família de distribuição da variável resposta, que pode ser a família gaussiana ou a Binomial, entre outras. No entanto, para se aplicar um *spline* de suavização na "formula" basta introduzir *s(nome da variável, df)*, onde *df* é o grau do *spline* de suavização.

Esta biblioteca, ainda tem a função *plot.gam()*, onde esta permite representar os efeitos dos *splines*, nas variáveis que são utilizadas.

Para mais informações, pode-se consultar "Package gam" [25].

3.5 Árvores de Decisão

3.5.1 Introdução

Como o nome indica uma Árvore de Decisão tem uma estrutura de árvore, na qual o modelo compreende uma série de decisões lógicas, semelhantes a um fluxograma, com os nós de decisões a indicarem a decisão a ser tomada num dado atributo/variável explicativa. Em termos matemáticos uma Árvore de Decisão é um grafo sem ciclos, com uma raiz fixa [26].

Os nós de decisões dividem-se em ramificações, que nos indicam as escolhas das decisões. A árvore termina com os nós folhas, que também se denominam de nós terminais, que denotam o resultado final, após uma combinação de várias decisões.

Na Figura 3.1, está representada a estrutura de uma Árvore de Decisão, na qual se pode observar que uma Árvore de Decisão é uma estrutura em forma de árvore que começa com o nó de raiz, que não tem nenhuma aresta de entrada, apenas tem arestas de saída. Todos os restantes nós têm exatamente uma aresta de entrada. Nos nós que têm uma aresta de entrada e de seguida têm duas ou mais arestas de saída, são chamados de nós de decisão, em que cada uma dessas arestas/ramificação de saída representa a decisão tomada. Os nós de decisão, também podem ser denotados por nós internos. Todos os restantes nós são chamados de nós folhas ou nós terminais, como referido anteriormente [27].

Uma Árvore de Decisão é um classificador expresso como uma partição recursiva do espaço das variáveis explicativas, uma vez que em cada nó interno se divide o espaço das variáveis explicativas em duas ou mais partes disjuntas.

Os métodos baseados em árvores particionam o espaço das variáveis explicativas num conjunto de retângulos, isto é, em várias regiões simples e de seguida ajustam um modelo simples, como uma constante, em cada uma destas. Estes métodos são conceptualmente simples, mas são muito poderosos e têm uma fácil interpretação. Uma vez que

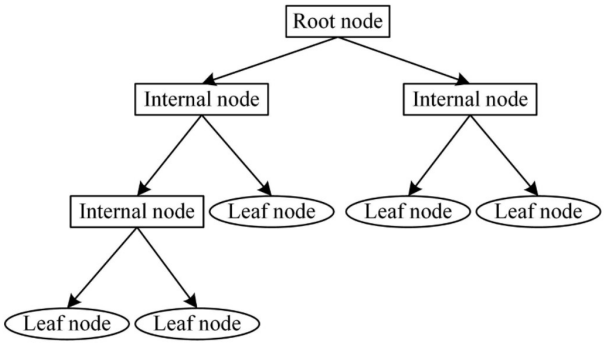


Figura 3.1: Exemplo de uma Árvore de Decisão [28].

o particionamento das variáveis explicativas é feito através de um conjunto de regras de divisão, então isto pode ser resumido a uma árvore, pelo que estes tipos de abordagem são conhecidos como métodos de Árvore de Decisão [12].

Existem diversos métodos de regressão e classificação baseados em árvores como o CART (*Classification and Regression Trees*) e o C4.5, sendo este último um grande concorrente do CART [17]. No entanto, no presente trabalho, apenas se considerou o algoritmo CART, pelo que não se descreve aqui o algoritmo C4.5. A razão pela qual apenas se considerou o algoritmo CART, foi devido ao facto de um estudo ("Performance Evaluation Among ID3, C4.5, and CART Decision Tree Algorithm" [29]) ter avaliado o desempenho de três algoritmos de Árvores de Decisão, os algoritmos ID3, C4.5 e o CART, sendo que os resultados foram sempre melhores no algoritmo CART, apesar destes 3 algoritmos apresentarem um grande potencial no desempenho de previsão. No entanto, o autor ainda conclui que o algoritmo C4.5, quando comparado a outros algoritmos, fornece os resultados mais precisos para um pequeno conjunto de dados. Por outro lado, uma das desvantagens do algoritmo C4.5 é que este é suscetível a ruídos [27].

As Árvores de Decisão podem ser aplicadas a problemas de regressão como também a problemas de classificação. Quando a variável resposta é quantitativa, estamos perante um problema de regressão, pelo que neste caso utilizam-se as Árvores de Regressão. Já as Árvores de Classificação são utilizadas quando a variável resposta é categórica.

3.5.2 Métodos CART

O algoritmo CART (*Classification and Regression Trees*) tanto pode ser utilizado para problemas de classificação como de regressão [26]. Este algoritmo é caracterizado pelo facto de construir árvores binárias, ou seja, cada nó de decisão tem exatamente duas ramificações possíveis. O algoritmo CART é baseado no método recursivo binário [27].

Este algoritmo utiliza como o critério de divisão o *índice de gini* quando se está sob

um problema de classificação. Já no caso de problemas de regressão este algoritmo utiliza como o critério de divisão a minimização do *RSS* [26].

3.5.2.1 Árvores de Regressão

Como referido anteriormente o algoritmo CART, quando aplicado a problemas de regressão, utiliza como critério de divisão a minimização do *RSS*. Uma vez que este algoritmo é baseado no método recursivo binário, então em cada uma das divisões tem-se duas novas regiões que particionam os dados, conforme a regra de decisão obtida.

O presente algoritmo começa no topo da árvore, ou seja no nó raiz com todos as observações, em que todas as observações pertencem a uma única região. De seguida divide sucessivamente os espaços das variáveis explicativas, em que cada divisão é indicada por dois novos ramos, em cada um dos nós de decisão. Este algoritmo é *greedy*, uma vez que para cada etapa do processo de construção da árvore a melhor divisão da árvore é feita em cada uma destas etapas, em vez de olhar para a frente e escolher uma divisão que levará a uma melhor árvore em alguma etapa futura.

Assim, para realizar a partição recursiva binária, primeiro seleciona-se a variável explicativa x_j e o ponto de corte s de modo que dividindo o espaço das variáveis explicativas, as regiões $\{\mathbf{x}|x_j < s\}$ e $\{\mathbf{x}|x_j \geq s\}$ levam à maior redução do *RSS*. Observe-se que a região $\{\mathbf{x}|x_j < s\}$ é o espaço da variável explicativa na qual a x_j assume um valor menor que s, enquanto que a região $\{\mathbf{x}|x_j \geq s\}$ é o espaço da variável explicativa na qual a x_j assume valores maiores ou iguais que s. Ou seja, considera-se todas as variáveis explicativas $\mathbf{x} = (x_1, \ldots, x_p)^T$ e todos os possíveis valores de corte s para cada uma das variáveis explicativas, e em seguida seleciona-se a variável explicativa e o ponto de corte s de modo que a árvore resultante tenha o valor do *RSS* mais baixo.

De um modo geral, para quaisquer j, com $j = 1, \ldots, p$ e s, define-se o par de semiplanos dado por 3.92 e procura-se o valor de j e de s que minimizam a equação 3.93.

$$R_1(j,s) = \{\mathbf{x}|x_j < s\} \quad \text{e} \quad R_2(j,s) = \{\mathbf{x}|x_j \geq s\} \tag{3.92}$$

$$RSS = \sum_{i:x_i \in R_1(j,s)} \left(y_i - \hat{y}_{R_1}\right)^2 + \sum_{i:x_i \in R_2(j,s)} \left(y_i - \hat{y}_{R_2}\right)^2 \quad \text{onde} \quad \mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) \tag{3.93}$$

Na equação 3.93, $\hat{y}_{R_1}$ é a média da variável resposta para as observações ditas de treino em $R_1(j,s)$ e $\hat{y}_{R_2}$ é a média da variável resposta para as observações de treino em $R_2(j,s)$.

De seguida, repete-se este processo, procurando a melhor variável explicativa e o melhor ponto de corte para dividir ainda mais os dados de modo a minimizar o *RSS* dentro de cada uma das regiões resultantes. No entanto desta vez, em vez de se dividir todo o espaço das variáveis explicativas, apenas se divide uma das duas regiões identificadas anteriormente, pelo que, no final deste passo fica-se com 3 regiões. Novamente procura-se dividir ainda mais cada uma dessas 3 regiões, de modo a minimizar o *RSS*. Este processo

contínua até que um critério de paragem seja alcançado. Existem vários critérios de paragem que podem ser utilizados, entre os quais [27]:

- Até que nenhuma das regiões contenha mais do que um número mínimo especificado de observações;
- Se todas as instâncias no conjunto de treino pertencerem a um único valor de y;
- O melhor critério de divisão, neste caso o *RSS*, não seja maior que um determinado limiar;
- Se a profundidade máxima da árvore for atingida, onde a definição de profundidade é apresentada de seguida.

A profundidade de um nó é definida, como o comprimento do caminho da raiz até ao nó, sendo que por defenição a profundidade da raiz é 0 e a profundidade de um nó descendente de um nó i que tenha profundidade k é $(k + 1)$. Já a profundidade máxima de uma árvore é o maior comprimento de todos os caminhos possíveis desde raiz até aos nós folhas. Por exemplo, na Figura 3.1, tem-se que a profundidade máxima da Árvore de Decisão é 3, pois este é o maior valor do comprimento de todos os caminhos possíveis desde raiz até ao nó folha, pois $\max(2, 2, 2, 3, 3) = 3$.

Depois de criadas cada uma das regiões $R1, \ldots, R_m$, prevê-se a variável resposta para uma determinada observação, dita de teste, usando a média das observações de treino na região à qual essa observação de teste pertence.

Assim, o algoritmo CART para o caso da regressão pode ser resumido como se encontra na Tabela 3.3.

Como referido anteriormente, o algoritmo CART, utiliza a partição recursiva binária, para a divisão do espaço das variáveis explicativas. No entanto, se na divisão estivermos perante uma variável explicativa binária, apenas uma divisão é possível, em que cada nível define um grupo. No entanto, para uma variável categórica com k níveis, existem $2^{k-1} - 1$ divisões possíveis [30].

3.5.2.2 Poda de uma Árvore de Regressão

Quando se constrói uma árvore de regressão com o processo descrito anteriormente, é muito provável que a árvore construída sobreajuste os dados (*overfitting*), apresentando um péssimo desempenho no ajuste dos dados de teste. Este péssimo desempenho pode ocorrer por que a árvore resultante pode ser muito complexa, ou seja, por ter uma grande profundidade [12]. Este problema pode ser resolvido pela poda da árvore construída.

Uma das formas de se podar uma árvore é através da poda de custo de complexidade (*Cost complexity pruning*). Neste método considera-se uma sequência de árvores indexadas por um parâmetro de ajuste não negativo α, em que para cada um dos valores de α faz corresponder uma sub-árvore $T \subset T_0$, (onde T_0 corresponde à árvore obtida antes da poda),

Tabela 3.3: Método CART caso Regressão.

Método CART caso Regressão

1. Comece com uma árvore vazia;
2. Seja R_0, o espaço das variáveis explicativas que é constituído pelos valores observados do suporte de cada variável explicativa $\mathbf{x} = (x_1, \ldots, x_p)^T$. Seja j=0;
3. Para $i = j + 1$ fazer o seguinte:
   a) Se $i = 1$, então de todas as variáveis explicativas $x_1, \ldots, x_p$ na partição R_0, selecione a variável explicativa x_j e o ponto de corte s, de modo que as regiões definidas em 3.92, levam à maior redução do *RSS*. Posto isto, obtêm-se duas novas regiões, definidas por 3.92. Denote-se essas duas regiões por R_{2i-1} e R_{2i}. Vá para o passo c);
   b) Se $i > 1$, então de todas as variáveis explicativas $x_1, \ldots, x_p$, que formam cada uma das i partições ainda não divididas, selecione a variável explicativa x_j e o ponto de corte s de uma dessas i partições, de modo a que as regiões definidas em 3.92, levam à maior redução do *RSS*. Posto isto, obtêm-se duas novas regiões, definidas por 3.92. Denote-se essas duas regiões por R_{2i-1} e R_{2i}. Vá para o passo c);
   c) Se algum dos critérios de paragem definidos anteriormente for atingido pára-se e vai-se para o passo 4. Caso contrário $j = j + 1$ e volta-se ao passo 3;
4. Para cada observação que cai na região R_j, com $j = 1, \ldots, m$, faz-se a previsão, que é simplesmente a média dos valores de resposta para as observações de treino na região R_j.

tal que a quantidade 3.94 é a mais pequena possível [12]. O parâmetro α é denominado por parâmetro de *complexidade*.

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T| \tag{3.94}$$

Na equação 3.94, o $|T|$ representa o número de nós terminais da árvore T, R_m é o retângulo, ou seja, o espaço das variáveis explicativas correspondentes ao *m-ésimo* nó terminal, e $\hat{y}_{R_m}$ é a variável reposta prevista associada ao R_m, isto é, a média das observações de treino em R_m. O parâmetro α controla um compromisso entre a complexidade das sub-árvores e o seu ajuste aos dados de treino. Quando $\alpha = 0$, então a sub-árvore T será simplesmente a árvore T_0, que é a árvore na sua maior extensão possível, pois a quantidade 3.94 apenas mede o erro de treino.

À medida que o valor de α aumenta existe um preço a pagar por se ter uma árvore com muitos nós terminais, pelo que a quantidade 3.94 tenderá a ser minimizada para uma sub-árvore menor [12].

Uma das formas de determinar o valor de α de forma a que a quantidade de 3.94 seja a menor possível, é através da validação cruzada, em que este algoritmo está descrito na Tabela 3.4. O erro de validação cruzada com k grupos, pode ser o erro quadrático médio (*MSE*), dado por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2. \tag{3.95}$$

Tabela 3.4: Poda da Árvore de Regressão [12].

Poda da Árvore
1. Use a divisão binária recursiva com o critério da minimização do *RSS*, para obter uma árvore complexa, com os dados de treino e utilize como critério de paragem, um dos critérios de paragem definidos antes na sub-subsecção 3.5.2.1;
2. Aplique a poda de custo de complexidade à árvore não podada, isto é, a árvore na sua maior extensão possível, de modo a obter uma sequência de sub-árvores melhores em função de α;
3. Use a validação cruzada em K grupos (*K-fold*) para escolher o parâmetro α, isto é, divide-se o conjunto de dados de treino em K grupos e para cada $k = 1, \ldots, K$:
a) Repete-se os passos 1 e 2 em todos os K grupos de dados de treino, expeto no *k-ésimo* grupo dos dados de treino;
b) Avalie o *MSE* nos dados do *k-ésimo* grupo, em função de α. Faça a média dos resultados para cada valor de α e escolha α para minimizar o erro médio;
4. Retorne à sub árvore do Passo 2 que corresponde ao valor escolhido de α.

3.5.2.3 Árvores de Classificação

Neste caso considere-se que se tem uma variável resposta categórica com K níveis, isto é, a variável resposta é constituída por $k = 1, \ldots, K$ classes.

Uma Árvore de Classificação é muito semelhante a uma Árvore de Regressão, apesar de uma Árvore de Decisão ser usada para prever uma resposta qualitativa [12]. Nas Árvores de Regressão viu-se que a variável resposta prevista para uma observação é dada pela média das observações de treino que pertencem ao mesmo nó terminal. No caso das Árvores de Classificação prevemos que cada observação pertence à classe mais comum de observações de treino na região à qual pertence, ou seja, os nós terminais da Árvore de Classificação são compostos pela classe mais representada da correspondente região.

O algoritmo CART para as Árvores de Classificação funciona exatamente como nas Árvores de Regressão, no entanto existe uma pequena alteração, pois como se viu anteriormente nas Árvores de Regressão o algoritmo CART utiliza como critério de divisão a

minimização do *RSS*, e esta medida não pode ser aplicada quando se tem como variável resposta uma variável categórica. Uma alternativa natural é a utilização da taxa de erro de classificação, uma vez que se pretende atribuir uma observação em uma determinada região à classe mais comum de observações de treino naquela região. A taxa de erro de classificação é simplesmente a fração dos dados de treino naquela região que não pertencem à classe mais comum, em que esta é dada pela fórmula de 3.96 [12].

$$E = 1 - \max_{k}(\hat{p}_{mk}) \tag{3.96}$$

Na fórmula 3.96, $\hat{p}_{mk}$ representa a proporção de observações de treino na *m-ésima* região que pertencem à *k-ésima* classe. No entanto, verifica-se que o erro de classificação não é suficientemente sensível para a construção da Árvores de Classificação, pelo que existem outras medidas que são preferíveis a esta. O algoritmo CART quando aplicado a problemas de classificação, utiliza como critério de divisão binário o *índice de gini*, para determinar qual a variável explicativa que deve de ser utilizada para fazer a divisão, ou seja, para determinar qual a variável explicativa que irá fazer a partição binária, estando o *índice de gini* definido em 3.97 [12]. A estratégia é escolher a variável explicativa cujo o *índice de gini* é o mínimo após a divisão.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \tag{3.97}$$

O *índice de gini* é uma medida de impureza que mede as divergências entre as distribuições de probabilidade da variável explicativa alvo [27]. Ou seja, o *índice de gini* mede o grau de heterogeneidade dos dados. Quando o valor do índice é igual a 0, o nó é puro e quando o valor se aproxima de 1, o nó é impuro.

Observe-se que a quantidade $\hat{p}_{mk}$ definida em 3.97, representa a proporção de observações de treino na *m-ésima* região que pertencem à *k-ésima* classe. Assim, o valor de $\hat{p}_{mk}$, pode ser dado pela fração do número total de observações que pertencem à *k-ésima* classe na *m-ésima* região a dividir pelo número total de observações da *m-ésima* região.

O valor de G pode ser interpretado como o valor de impureza da *m-ésima* partição em que também se pode denotar por $G = G(m)$. A verificação das divisões binárias é feita somando os valores de impureza de cada partição gerada pela divisão. Suponha-se que a divisão realizada na variável explicativa A particiona a *m-ésima* partição em m_1 e m_2, pois uma vez que o algoritmo CART utiliza a divisão binária. Então o valor do *índice de gini* da partição m, pode ser obtido usando a equação 3.98 [31].

$$G_A(m) = \frac{|m_1|}{|m|}\sum_{k=1}^{K}\hat{p}_{m_1k}(1-\hat{p}_{m_1k}) + \frac{|m_2|}{|m|}\sum_{k=1}^{K}\hat{p}_{m_2k}(1-\hat{p}_{m_2k}) = \frac{|m_1|}{|m|}G(m_1) + \frac{|m_2|}{|m|}G(m_2) \tag{3.98}$$

Na equação 3.98, o valor obtido de $G_A(m)$ é a impureza da variável explicativa A na *m-ésima* partição, o $G(m_1)$ é o valor da impureza da primeira partição da *m-ésima* partição, enquanto que $G(m_2)$ é o valor da impureza da segunda partição da *m-ésima* partição, onde

m é a *m-ésima* partição, m_1 é a primeira partição da *m-ésima* partição e m_2 representa a segunda partição da *m-ésima* partição.

Assim, a variável explicativa selecionada para efetuar a divisão binária num nó de decisão de entre um conjunto de variáveis explicativas é aquela que apresentar um menor valor do *índice de gini* (equação 3.98).

Para uma variável explicativa categórica, se esta for binária, então a divisão desta variável é simples, pois uma das partições terá a classe de sucesso (1), enquanto que a outra partição terá a classe de insucesso (0), sendo que a Árvore de Decisão é construída através da partição recursiva binária, como referido anteriormente. Se a variável explicativa for uma variável categórica com k níveis, então como visto no caso das Árvores de Regressão na sub-subsecção 3.5.2.1, existem $2^{k-1} - 1$ divisões possíveis, sendo que neste caso terá-se que testar cada uma das divisões possíveis, ou seja, ir-se-á obter $2^{k-1} - 1$ valores do *índice de gini*, nesta variável. Para uma variável explicativa contínua, tem de selecionar um ponto de corte s, de modo a que a divisão do espaço das variáveis explicativas, as regiões $\{\mathbf{x}|x_j < s\}$ e $\{\mathbf{x}|x_j \geq s\}$ levam á maior redução do *índice de gini* nesta partição para a variável explicativa x_j. Assim, se tivermos um conjunto de variáveis explicativas $\mathbf{x} = (x_1, \ldots, x_p)^T$, sejam elas contínuas ou categóricas, a variável selecionada para a divisão binária é aquela que apresentar o menor valor de 3.98.

A diminuição do nível de impureza obtido a partir de uma divisão binária da variável explicativa A pode ser obtida utilizando a equação 3.99.

$$\Delta G(A) = G(m) - G_A(m) \tag{3.99}$$

O valor obtido $\Delta G(A)$ da 3.99 é denominado pelo nível de impureza, o $G(m)$ representa a impureza do atributo A na *m-ésima* partição e $G_A(m)$ é a impureza da *m-ésima* partição.

Uma vez que a variável explicativa selecionada para a divisão binária é aquela que apresentar o menor valor de 3.98, isto é equivalente a ter que a variável selecionada para a divisão binária é aquela que apresentar uma maximização na redução de impureza, valor dado pela equação 3.99.

Na Tabela 3.5, encontra-se um esquema de como construir uma Árvore de Classificação de modo a prever cada uma das K classes da variável resposta Y.

3.5.2.4 Poda Árvore de Classificação

Do mesmo modo, a poda das árvores com o algoritmo CART nos problemas de classificação é idêntica à poda da árvores do algoritmo CART nos problemas de regressão (poda de custo de complexidade), no entanto, em vez de se considerar o modelo dado em 3.94, considera-se o modelo dado em 3.101. O modo de resolver este problema é idêntico à poda da árvore no caso da regressão que é dado na Tabela 3.4, mas em vez de se considerar o erro quadrático médio considerado na validação cruzada para cada um dos valores de α, no caso da classificação avalia-se a taxa de erro de classificação, em que esta é dada pela fórmula de 3.100 [26].

Tabela 3.5: Método CART caso Classificação.

Método CART caso Classificação

1. Comece com uma árvore vazia;
2. Seja R_0, o espaço das variáveis explicativas que é constituído pelos valores observados do suporte de cada variável explicativa $\mathbf{x} = (x_1, \ldots, x_p)^T$. Seja j=0;
3. Para $i = j + 1$ fazer o seguinte:
   a) De todas as variáveis explicativas $x_1, \ldots, x_p$ nas partições R_{i-1} selecione a variável explicativa x_j e o ponto de corte s, de modo que as regiões definidas em 3.92, levam á maior redução do *índice de gini* (equação 3.98). Posto isto, obtêm-se duas novas regiões, definidas por 3.92. Denote-se essas duas regiões por R_{2i-1} e R_{2i};
   b) Se $i > 1$, então de todas as variáveis explicativas $x_1, \ldots, x_p$, que formam cada uma das i partições ainda não divididas, selecione a variável explicativa x_j e o ponto de corte s de uma dessas i partições, de modo a que as regiões definidas em 3.92, do *índice de gini* (equação 3.98). Posto isto, obtêm-se duas novas regiões, definidas por 3.92. Denote-se essas duas regiões por R_{2i-1} e R_{2i}. Vá para o passo c);
   c) Se algum dos critérios de paragem definidos anteriormente for atingido pára-se e vai-se para o passo 4. Caso contrário $j = j + 1$ e volta-se ao passo 3;
4. Para cada observação que cai na região R_j, com $j = 1, \ldots, m$, faz-se a classificação, que é simplesmente a classe mais comum ocorrida nessa região.

$$R(T, \mathbf{X}_t) = \frac{|T_W|}{n_t} \tag{3.100}$$

Na equação 3.100, o $R(T, \mathbf{X}_t)$ é a taxa de erro de classificação da árvore T, em que T denota uma Árvore de Classificação, $|T_W|$ é a quantidade de observações de teste $\mathbf{X}_t$ classificadas incorretamente e n_t denota o número de observações de teste.

$$R_\alpha(T, \mathbf{X}) = R(T, \mathbf{X}_t) + \alpha|T| \tag{3.101}$$

Na equação 3.101 o valor de $R(T, \mathbf{X})$ é a taxa de erro de classificação da árvore T, $\mathbf{X}$ é a matriz dos valores das covariáveis e $|T|$ é o número de nós finais da árvore T e $\alpha \geq 0$ é o parâmetro de *complexidade* [26].

Na Tabela 3.6, encontra-se o algoritmo para determinar o valor de α que minimiza a quantidade dada na equação 3.101.

Tabela 3.6: Poda da Árvore de Classificação.

Poda da Árvore
1. Use a divisão binária recursiva, com o critério do *índice de gini* para obter uma árvore complexa, com os dados de treino e utilize como critério de paragem, um dos critérios de paragem definidos antes na sub-subsecção 3.5.2.1;
2. Aplique a poda de custo de complexidade à árvore não podada, isto é, à árvore na sua maior extensão possível, de modo a obter uma sequência de sub-árvores melhores em função de α;
3. Use a validação cruzada em K grupos (K-*fold*) para escolher o parâmetro α, isto é, divide-se o conjunto de dados de treino em K grupos e para cada $k = 1, \ldots, K$:
a) Repete-se os passos 1 e 2 em todos os K grupos de dados de treino, expeto no k-*ésimo* grupo dos dados de treino;
b) Avalie o $R_\alpha(\mathbf{x})$ (Taxa de Erro de Classificação) nos dados do k-*ésimo* grupo, em função de α. Faça a média dos resultados para cada valor de α e escolha α para minimizar o erro médio.
4. Retorne a sub árvore do Passo 2 que corresponde ao valor escolhido de α.

3.5.3 Vantagens e desvantagens dos algoritmo CART

O método CART tem as suas vantagens e as suas desvantagens, tal como os restantes métodos. Destacam-se as seguintes vantagens [27]: o algoritmo CART pode ser aplicado tanto a variáveis numéricas como a variáveis categóricas, o algoritmo CART identifica as variáveis mais significativas e elimina as variáveis não significativas e o algoritmo CART lida facilmente com *outliers*.

As desvantagens deste algoritmo são que o CART pode ter uma Árvore de Decisão instável, uma vez que uma modificação insignificantes da amostra de aprendizagem, como eliminar várias observações, pode causar mudanças na Árvore de Decisão, aumentando ou diminuindo a complexidade da árvore, incluindo mudanças nas variáveis de divisão e nos valores de corte, sendo que outra da desvantagem do algoritmo CART é que este divide apenas por uma variável [27].

3.5.4 Biblioteca *rpart*

A biblioteca *rpart* do *software* R tem a função *rpart()* com a qual se pode ajustar o modelo da Árvore de Decisão, quer para o caso da regressão quer para o caso da classificação. Esta função tem como argumentos os seguintes [32]:

- *minsplit* - número mínimo de observações que devem existir para que uma divisão seja tentada. Por defeito o valor é 20;

- *minbucket* - número mínimo de observações em qualquer nó terminal. Por defeito este valor é 7;

- *cp* - parâmetro complexidade, que é a melhoria mínima no modelo necessária em cada nó. É baseado no custo de complexidade (*Cost complexity pruning*), que no caso da classificação é dada pelo modelo 3.101 e para o caso da regressão é dao por 3.94. Por defeito este parâmetro toma o valor 0.01;

- *xval* - número de grupos da validação cruzada;

- *maxdep* - define a profundidade máxima de qualquer nó da árvore final, com o nó da raiz contando como profundidade 0. (Valores maiores que 30 o rpart dará resultados sem sentido num computador de 32 bits);

- *method* - método;

- *parms* - parâmetro opcional para a função de divisão, na qual a função *índice de gini* está disponível;

No entanto, existem mais argumentos que podem ser utilizados, mas neste trabalho apenas se usou alguns dos argumentos descritos anteriormente. Para mais informações sobre os restantes argumentos ver o documento "Package rpart" [32].

Adicionalmente o *software* R, tem uma biblioteca que permite ter uma representação visual de uma Árvore de Decisão, sendo que a biblioteca é o *rpart.plot*. A função que permite fazer isso é o *rpart.plot()* da biblioteca *rpart.plot*, donde para se ter uma representação de uma Árvore de Decisão, basta dar como entrada nesta função um objeto ajustado através da função *rpart()* da biblioteca *rpart*. A função *rpart.plot()* tem alguns argumentos que podem ser alterados, e que podem ser consultados no documento "Package rpart.plot" [33].

3.6 Floresta Aleatória

3.6.1 Introdução

As Florestas Aleatórias são um método de *ensemble* não paramétrico, no qual uma "floresta" de Árvores de Decisão é gerada pela reamostragem de dados de treino.

Um método *ensemble* é uma abordagem que combina muitos modelos simples de "blocos de construção" para obter um único modelo. Os modelos simples de blocos de construção, são muitas vezes também conhecidos como "aprendizes fracos", pois estes modelos simples podem levar a previsões medíocres. No caso da Floresta Aleatória, tem-se que os blocos de construção simples são Árvore de Regressão ou de Classificação.

As Florestas Aleatórias tanto podem ser utilizadas para classificação como para regressão.

Na subsecção 3.6.2, descrevem-se detalhadamente como funcionam as Florestas Aleatórias, enquanto que na subsecção 3.6.3 detalham-se as vantagens e as desvantagens das Florestas Aleatórias.

3.6.2 Método

Como descrito anteriormente, tem-se que as Florestas Aleatórias utilizam muitos modelos simples de 'blocos de construção', em que neste caso, cada um desses blocos de construção é uma Árvore de Regressão ou de Classificação, para obter um único modelo. No entanto, para a construção de cada uma dessas Árvores de Regressão ou de Classificação, é utilizada uma amostra aleatória dos dados de treino, ou seja, cada uma das Árvores de Regressão ou de Classificação é construída sob uma das amostras de treino *bootstrap*. No Apêndice B, pode-se encontrar devidamente explicado o que é uma amostra *bootstrap*.

Assim, tem-se que o método de amostra *bootstrap* é repetido muitas vezes, com cada uma dessas sub-amostras a gerarem uma única Árvore de Regressão ou de Classificação. Ao contrário do que se sucede na construção de uma Árvore de Decisão, cada vez que uma divisão numa árvore é considerada, uma amostra aleatória de m variáveis explicativas é escolhida como candidatos a divisão, de entre todo o conjunto completo de p variáveis explicativas, ou seja, em vez de se considerar todas as p variáveis explicativas como candidatas a divisão, como se faz nas Árvores de Decisão, apenas se levam em conta uma amostra aleatória de m variáveis explicativas. Na divisão apenas se pode usar uma das m variáveis explicativas escolhidas aleatoriamente, para a divisão. Geralmente, segundo vários investigadores os valores de m escolhidos são os seguintes [17]:

- Para classificação, por defeito o valor de m é dado pela parte inteira da $\sqrt{p}$ e o tamanho mínimo do nó é um;
- Para regressão, por defeito o valor de m é dado pela parte inteira de $\frac{p}{3}$ e o tamanho mínimo do nó é 5.

Ao construir uma Floresta Aleatória, para a divisão em cada uma das árvores, o algoritmo não pode considerar a maioria das variáveis explicativas disponíveis e isto pode parecer uma desvantagem, no entanto este processo pode ser visto como um passo muito inteligente, pois se supusermos que existe uma variável explicativa muito forte no conjunto dos dados, juntamente com outras variáveis explicativas moderadamente fortes, então no conjunto das árvores na Floresta Aleatória, a sua maioria ou todas as árvores usarão essa forte variável explicativa na divisão superior e consequentemente, todas as árvores serão bastantes semelhantes entre si [12]. Portanto, as previsões das árvores serão altamente correlacionadas [12].

Assim, as Florestas Aleatórias superam esse problema, forçando que em cada divisão nas árvores apenas se considere um subconjunto das variáveis explicativas para serem candidatas à divisão [12].

Cada uma das árvores na Floresta Aleatória é construída na sua maior extensão possível, isto é, não se leva em conta a poda. Cada uma das árvores individuais tomam uma decisão de classificação ou de previsão, isto é, cada uma das árvores pode ser uma Árvore de Classificação ou uma Árvore de Regressão, conforme se o problema inicial é de classificação ou de regressão. No caso da classificação, a classe final prevista de uma observação é feita através do voto maioritário para classificação, ou seja, para cada uma das árvores da Floresta Aleatória calcula-se a classe à qual essa observação pertence, sendo que a classe a que essa observação pertence na Floresta Aleatória é a classe que foi mais vezes retornada. Já no caso da regressão, a previsão é feita através da média ponderada dos resultados de previsão obtidos em cada uma das árvores da Floresta Aleatória. Na Tabela 3.7, encontra-se descrito um esquema do algoritmo de como é construída uma Floresta Aleatória, onde o valor de B expresso na Tabela 3.7, representa o número de árvores construídas na Floresta Aleatória e por conseguinte também é o número de amostras *bootstrap* construídas. Na prática o valor de B é um valor suficientemente grande de modo a que o erro estabilize [12].

Tabela 3.7: Algoritmo Floresta Aleatória [17].

Algoritmo Floresta Aleatória

1. Para $b = 1$ a B:
   a) Desenhe uma amostra *bootstrap* Z^* de tamanho n a partir dos dados de treino;
   b) Faça uma árvore da Floresta Aleatória T_b para os dados *bootstrap*, repetindo recursivamente as etapas a seguir para cada nó terminal da árvore, até que o tamanho mínimo do nó n_{min} seja alcançado;
      i. Selecione m variáveis aleatoriamente das p variáveis;
      ii. Escolha a melhor variável para efetuar a divisão entre as m;
      iii. Divida o nó em dois nós filhos;
2. Saída da construção de árvores $\{T_b\}_1^B$.
   Para fazer uma previsão num novo ponto x:
   Regressão: $\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.
   Classificação: Seja $\hat{C}_b(x)$ a predição da classe da *b-ésima* árvore da Floresta Aleatória.
   Então $\hat{C}_{rf}^B(x) = \textit{voto maioritário } \hat{C}_b(x)_1^B$.

As observações do conjunto de dados que não são selecionadas numa dada amostra *bootstrap*, são chamadas de observações *Out-of-Bag* (OOB). Estas observações são utilizadas para estimar a taxa de erro e também para estimar a importância das variáveis explicativas. Nas sub-subsecção 3.6.2.1 e 3.6.2.2, encontram-se descritos método da estimação dos erros, assim como também da estimação da importância das variáveis explicativas.

3.6.2.1 Estimação Erro OOB

Uma das características importantes das florestas aleatórias é a utilização das observações que estão fora da amostra *bootstrap* [17], as denominadas observações OOB.

Uma maneira muito direta de estimar o erro de teste do modelo da Floresta Aleatória, sem a necessidade de realizar validação cruzada com k grupos ou o método *Holdout*, este último encontra-se descrito na secção 3.8, é através do erro OOB, uma vez que vários estudos indicam que em média, durante o *bootstrap*, cerca de um terço das observações de treino são deixadas de fora em cada uma das amostras *bootstrap* utilizadas para ajustar uma árvore [34], pois basta lembrar que a chave do método da Floresta Aleatória é que as árvores são repetidamente ajustadas a amostras *bootstrap* das observações [12].

Assim, pode-se prever a resposta para a *i-ésima* observação usando cada uma das árvores utilizadas na construção da Floresta Aleatória, onde essa observação foi OOB. Isso irá produzir cerca de $\frac{B}{3}$ previsões para a *i-ésima* observação [12].

Logo, para se obter uma única previsão para a *i-ésima* observação pode-se calcular a média das respostas previstas, isto se, estivermos sobre um problema de regressão, ou pode-se obter por votação maioritária, este último se estivermos sobre um problema de classificação.

Uma previsão OOB pode ser obtida da maneira descrita anteriormente para cada uma das n observações, a partir das quais o erro OOB das n observações em conjunto, isto é o erro OOB geral, pode ser dado pelo erro quadrático médio (*MSE*), para o caso da regressão, enquanto que para o caso da classificação, utiliza-se o erro de classificação [12].

O erro OOB geral da Floresta Aleatória pode ser utilizado para determinar o número de árvores a serem consideradas na Floresta Aleatória, em que o número de árvores ótimo para a construção da Floresta Aleatória corresponde à estabilização do erro OOB. Isto pode ser obtido através de um gráfico, em que no eixo dos xx se tem o número de árvores e no eixo dos yy tem-se o correspondente erro OOB [17].

3.6.2.2 Medidas de Importância das Variáveis

Como referido anteriormente apesar das Florestas Aleatórias apresentarem melhores resultados em relação à previsão do que uma única simples árvore, as Florestas Aleatórias têm a desvantagem de serem difíceis de interpretar, uma vez que não nos podemos esquecer que as Florestas Aleatórias resultam da combinação de diversas Árvores de Decisão. Assim, quando se utiliza uma Floresta Aleatória com um grande número de Árvores de Decisão, não é possível representar de forma esquemática, tal como se obtém quando se utiliza as Árvores de Decisão, pois aí só se tem apenas uma árvore. Consequentemente, nas Florestas Aleatórias não se consegue ter uma perceção de quais são as variáveis mais importantes para a sua construção. As Florestas Aleatórias aumentam assim a precisão das previsões em detrimento da sua interpretabilidade.

No entanto, embora as Florestas Aleatórias sejam mais difíceis de interpretar, pode-se obter um resumo geral da importância de cada uma variável explicativa, usando o *RSS*,

para o caso da regressão ou o *índice de gini*, dado na equação 3.97 da secção 3.5, no caso da classificação [12].

No caso da regressão nas Florestas Aleatórias, pode-se registar a quantidade total do *RSS* que é diminuído devido às divisões de uma determinada variável explicativa, com a média de todas as Árvores de Regressão utilizadas para a construção da Floresta Aleatória. Da mesma forma, no caso da classificação na Floresta Aleatória, a importância de uma dada variável explicativa calcula-se através da média, sobre todas as Árvores de Decisão, das diminuições totais no *índice de gini* obtidas por divisões sobre essa variável explicativa em cada uma das Árvores de decisão utilizadas para a construção da Floresta Aleatória. De forma análoga, o procedimento é o mesmo para o caso da regressão nas Florestas Aleatórias, mas em vez do *índice de gini* utiliza-se o *RSS*. De referir que a diminuição total no *índice de gini* obtida pelas divisões que essa variável explicativa faz numa árvore, é a soma de cada diminuição do *índice de gini* sempre que essa variável explicativa é escolhida para dividir um nó interno ou nó de decisão na árvore.

3.6.3 Vantagens e Desvantagens

As Florestas Aleatórias, que mantêm muitas das vantagens das Árvores de Decisão, apresentam as seguintes vantagens [35]: o modelo funciona bem na maioria dos problemas; pode lidar com dados ruidosos ou ausentes e também lida bem com variáveis explicativas categóricas ou contínuas; seleciona apenas as variáveis explicativas mais importantes; pode ser utilizada em dados com um número extremamente grande de variáveis explicativas.

No entanto, as Florestas aleatórias apresentam as seguintes fraquezas [35]: ao contrário das Árvores de Decisão, o modelo não é facilmente interpretável; pode exigir algum trabalho para ajustar o respetivo modelo aos dados.

3.6.4 Biblioteca *randomForest*

A biblioteca *randomForest* do *software* R, tem a função *randomForest()* que pode ser utilizada para ajustar o modelo da Floresta Aleatória a um conjunto de dados. Esta função toma por defeito como número de variáveis candidatas, para cada uma das divisões, $\sqrt{p}$ no caso da classificação e $\frac{p}{3}$ no caso da regressão, sendo p o número de variáveis explicativas, em que o argumento referente ao número de variáveis candidatas da função *randomForest()* é o *mtry*. Por outro lado esta função também tem como argumento o *ntree*, que diz respeito ao número de árvores a construir, o valor deste argumento tem de ser dado pelo utilizador. No entanto, o número de árvores devem de ser suficientes de forma a estabilizar o erro, como referido anteriormente.

Outro argumento que esta função utiliza é o *replace*, que indica se queremos uma amostra com reposição *replace="TRUE"* ou não *replace="FALSE"*. Esta função por defeito considera a amostra com reposição, isto é, o *replace="TRUE"*. Outro argumento que esta função utiliza é o *sampsize*, que se refere ao número de observações da amostra, em que por defeito este é igual ao número de observações da base de dados de treino se nós

considerarmos a amostragem por reposição, ou seja se *replace="TRUE"*. Ou seja, por padrão a presente função faz amostras *bootstrap*, que foi o que utilizámos neste presente trabalho. No entanto se nós não quisermos considerar a amostra com reposição, esta função por defeito utiliza 63.2% das observações da base de dados de treino para a amostra. Outros dois argumentos importantes nesta função são o *nodesize* e o *maxnodes*, em que por defeito esta função considera o argumento *maxnodes* como "NULL" e no argumento *nodesize* o valor 5, isto se estivermos num problema de classificação, se for um problema de regressão o valor deste argumento é 1.

Esta função tem mais argumentos, que podem ser consultados em [32].

3.7 Redes Neuronais

3.7.1 Introdução

As Redes Neuronais inspiram-se no funcionamento do cérebro humano, procurando imitá-lo com neurónios artificiais conectados de uma forma semelhante à rede cerebral [36]. Uma Rede Neuronal modela a relação entre um conjunto de sinais de entrada e um sinal de saída usando um modelo derivado da nossa compreensão de como um cérebro humano responde a estímulos de entradas sensoriais. Tal como um cérebro, usa uma rede de células conectadas entre si, os neurónios, para criar um enorme processador paralelo. As Redes Neuronais usam uma rede de neurónios, também denotados por nós artificiais para resolver problemas de aprendizagem [37].

As Redes Neuronais podem ser aplicados a diversos problemas de aprendizagem, entre os quais problemas de classificação, problemas de previsão numérica, e até mesmo podem ser aplicadas para reconhecimento de padrões [38].

Um neurónio natural recebe sinais de entrada através de um processo bioquímico que permite que o impulso seja ponderado de acordo com a sua importância relativa ou frequência, acumulando sinais de entrada [37]. Quando a intensidade de um sinal excede um determinado limiar, o neurónio aciona o seu próprio sinal para o passar a um próximo neurónio [37].

Como mencionado anteriormente, uma Rede Neuronal consiste em unidades de processamento chamados neurónios. Um neurónio artificial tenta replicar a estrutura e o comportamento de um neurónio humano [36]. Como mencionado, um neurónio consiste em entradas e uma saída e o neurónio tem uma função que determina a ativação do neurónio [36]. Na Figura 3.2, tem-se uma ilustração da relação entre os sinais de entrada recebidos, variáveis explicativas $\mathbf{x} = (x_1, x_2, x_3)^T$ e o sinal de saída, variável resposta Y, de um modelo de um único neurónio artificial, podendo ser este modelo entendido em termos muitos semelhantes a um modelo biológico.

Tal como nos neurónios humanos, o sinal de cada entrada é ponderado por pesos $W = (w_1, w_2, w_3)$, de acordo com a sua importância. Os pesos w_i, $i = 1, 2, 3$, permitem que cada uma das entradas x_i, $i = 1, 2, 3$, contribua mais ou menos para a soma dos sinais

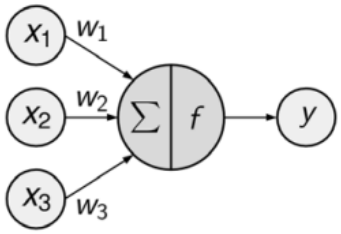


Figura 3.2: Exemplo de um modelo de um neurónio artificial [37].

de entrada. Os sinais de entrada são somados e este sinal é transmitido de acordo com uma função de ativação, denotada neste caso por f, ou seja, o total da soma dos sinais de entrada multiplicados pelos pesos é utilizado pela função de ativação f, e o resultado final dá um sinal de saída, $y(\mathbf{x})$ [37].

Assim, um típico neurónio artificial com p entradas, ou seja com p variáveis explicativas $\mathbf{x} = (x_1, \ldots, x_p)^T$ pode ser representada pela equação 3.102.

$$y(\mathbf{x}) = f\left(\sum_{i=1}^{p} w_i x_i\right) \tag{3.102}$$

As Redes Neuronais utilizam neurónios definidos por 3.102 como blocos de construção para construir modelos complexos de dados, ou seja, uma interconexão de neurónios individuais da forma de 3.102 formam uma Rede Neuronal. Uma Rede Neural Artificial pode ser definida em termos das seguintes características [35]:

- Uma **função de ativação**, que transforma uma combinação de sinais de entrada de um neurónio num único sinal de saída para ser transmitido futuramente na rede;
- A **topologia da rede** ou arquitetura, que descreve o número de neurónios no modelo, bem como o número de camadas e a forma pela qual elas estão conectadas;
- O **algoritmo de treino** que específica como os pesos da conexão são definidos para inibir ou excitar neurónios em proporção ao sinal de entrada.

Nas subsecções 3.7.2, 3.7.3 e 3.7.4 encontram-se descritas cada uma das características definidas acima de uma Rede Neuronal Artificial, respetivamente.

3.7.2 Função de Ativação

Uma função de ativação é um mecanismo na qual o neurónio artificial processa a informação e repassa essa dada informação por toda a rede. No caso biológico, a função de ativação pode ser imaginada como um processo que envolve somar todos os sinais de entrada e determinar se ele atinge o limiar de disparo. Neste caso, o neurónio transmite o sinal e, caso contrário, não faz nada. Já no caso da Rede Neuronal Artificial, isto é conhecido como um função de ativação de *threshold*, uma vez que resulta num sinal de saída apenas quando um *thershold* de entrada especificado é atingido.

Embora estas funções de ativação de *threshold* sejam interessantes devido ao seu paralelismo com a biologia, raramente são utilizadas em Redes Neuronais.

Uma função de ativação tem a função de transformar um sinal de entrada num sinal de saída. Se uma função de ativação não for utilizada numa Rede Neuronal, o sinal de saída seria praticamente uma função linear simples, um polinómio de grau 1. Embora uma equação linear seja simples e fácil de resolver, tem a limitação de não ter a capacidade de aprender e reconhecer padrões em dados complexos.

Existem várias funções de ativação que podem ser aplicadas entre as quais [39]:

- Função Binária por Passos;
- Função Linear;
- Função *Sigmoid*;
- Função *Tanh*;
- Função *ReLU*;
- Função *SoftMax*

Nas sub-subsecções 3.7.2.1, 3.7.2.2, 3.7.2.3, 3.7.2.4, 3.7.2.5 e 3.7.2.6, encontram-se descritas cada uma das funções de ativação mencionadas anteriormente, com base em [39]. Na sub-subsecção 3.7.2.7, tem-se descrito algumas sugestões para a escolha da função de ativação.

3.7.2.1 Função Binária por Passos

A função de binária por passos é definida matematicamente, como se encontra na fórmula 3.103.

$$f(x) = \begin{cases} 1, & \text{se } x \geq 0 \\ 0, & \text{se } x < 0 \end{cases} \tag{3.103}$$

A função de binária por passos é uma função de ativação bastante simples." deverá ser alterada para "A função de binária por passos é uma função de ativação bastante simples. No entanto, a derivada da função binária por passos é zero, (à exceção para o valor de $x = 0$, pois neste caso não existe derivada), pelo que esta função pode causar um impedimento na etapa de treino da Rede Neuronal [39].

3.7.2.2 Função Linear

A função linear é diretamente proporcional ao seu valor de entrada. Para se eliminar a desvantagem de gradiente nulo da função binária, pode-se usar a função linear com a representação em 3.104.

$$f(x) = ax \tag{3.104}$$

O valor da variável a pode ser uma constante qualquer diferente de zero escolhida pelo utilizador. Neste caso a derivada da função $f(x)$ é diferente de zero, pelo que os pesos e desvios serão atualizados durante a etapa do treino da Rede Neuronal, como se verá mais à frente no treino de uma Rede Neuronal.

3.7.2.3 Função *Sigmoid*

A função de ativação mais vezes utilizadas é a função *sigmoid*, pois é uma função não linear. A função *sigmoid* transforma os valores no intervalo $]0, 1[$. A função *sigmoid* que usaremos neste trabalho é definida como se encontra na equação 3.105.

$$f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \tag{3.105}$$

A função *sigmoid* é uma função contínua e diferenciável e a sua representação gráfica tem a forma de um "S". Além disso a função *sigmoid* não é simétrica em relação a zero, pelo que os valores devolvidos serão todos positivos, o que significa que os sinais de todos os valores dos sinais de saída dos neurónios serão todos positivos.

A função *sigmoid* essencialmente tenta empurrar os valores de saída para os extremos, o que é muito desejável quando se tenta classificar os valores para uma classe especifica [39]. Esta função de ativação é usada para classificação binária, uma vez que esta devolve valores no intervalo $]0, 1[$. Também se relembrámo-nos da Regressão Logística esta função é utilizada para converter uma função em linear em probabilidades entre 0 e 1.

3.7.2.4 Função *Tanh*

A função *tanh* é a função tangente hiperbólica. A função *tanh* é semelhante à função *sigmoid*, no entanto a função *tanh* é simétrica em torno da origem e os seus valores encontram-se no intervalo]-1,1[. A função *tanh* é definida tal como se encontra na equação 3.106.

$$f(x) = 2sigmoid(2x) - 1 = 2\frac{1}{1 + e^{-2x}} - 1 = \frac{e^{2x} - 1}{1 + e^{2x}} \tag{3.106}$$

Uma vez que a função *tanh* é simétrica em relação à origem, isso soluciona o problema da função *sigmoid*, pois neste caso os valores dos sinais de saída dos neurónios já podem ser positivos ou negativos.

A função *tanh* é contínua e diferenciável. Em comparação com a função *sigmoid*, o gradiente da função *tanh* é mais acentuado. A função *tanh* é preferível à função *sigmoid*, uma vez que o seu gradiente não se restringe a variar em uma determinada direção e também é centrado em 0.

3.7.2.5 Função *ReLU*

A função de ativação *ReLU*, onde *ReLU* significa unidade linear retificada, é uma função de ativação não linear que é muitas vezes utilizada em Redes Neuronais. A função *ReLU* é definida como se encontra na equação 3.107.

$$f(x) = \max(0, x) \tag{3.107}$$

A vantagem de se utilizar a função *ReLU* em relação a outras funções de ativação é que a função *ReLU* não ativa todos os neurónios ao mesmo tempo, isto significa que se a soma da combinação dos sinais de entrada multiplicados pelos seus pesos for inferior a 0, o sinal de saída do neurónio é zero pelo que o neurónio não é ativado. Isto significa que, ao mesmo tempo, apenas alguns dos neurónios serão ativados.

No entanto, a utilização desta função pode trazer problemas no treino da Rede Neuronal. O problema da utilização desta função é que em alguns casos, o valor do gradiente é nulo, pelo que os pesos e o *enviesamento* não serão atualizados durante uma das etapas *backpropagation* no treino da Rede Neuronal.

3.7.2.6 Função *SoftMax*

A função *Softmax* é uma combinação de várias funções *sigmoid*. Como a função *sigmoid* devolve valores no intervalo]0, 1[, estes valores podem ser tratados como a probabilidade de uma determinada classe. No entanto a função *sigmoid* apenas é capaz de lidar com problemas de classificação binários.

A função *Softmax* ao contrário da função *sigmoid* pode ser usada para problemas de classificação quando se tem presente mais de duas classes. A função *Softmax* devolve valores no intervalo]0, 1[. A função, para cada observação de todas as classes individuais, devolve a probabilidade de pertencer a uma determinada classe.

A função *Softmax* é representada pela equação 3.108, em que K é o número de classes (note-se que $K > 2$ uma vez que esta função é utilizada em problemas de classificação múltipla).

$$f(x_j) = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}}, \quad j = 1, \ldots, K \tag{3.108}$$

Quando se constrói um modelo de Rede Neuronal para classificação de múltiplas classes, a camada de saída da rede terá o mesmo número de neurónios que o número da classes alvo, como se virá mais à frente.

3.7.2.7 Escolha da Função de Ativação Correta

Nas sub-subsecções 3.7.2.1,3.7.2.2, 3.7.2.3, 3.7.2.4, 3.7.2.5 e 3.7.2.6 viram-se diferentes funções de ativação, em que uma das principais diferenças entre elas é o alcance do sinal de saída de cada uma delas, que tipicamente é um dos seguintes]0, 1[,]-1, 1[ou $\mathbb{R}$.

No entanto não existe uma regra para a seleção da função de ativação, pelo que a escolha desta depende do problema em estudo, por exemplo [39]:

- Para problemas de classificação, funções *sigmoid* ou uma combinação de funções *sigmoid* dão bons resultados;
- Funções *sigmoid* e funções *tanh* às vezes são evitadas devido ao problema de estas atingirem o valor próximo de zero no seu gradiente;
- A função *ReLU* é a função mais utilizada e tem um desempenho melhor do que outras funções de ativação na maioria dos casos;
- A função *ReLU* deve ser apenas usadas nas camadas ocultas;

Na construção de uma Rede Neuronal, podem-se experimentar diferentes de funções de ativação. No entanto, vários estudos demonstram que a função de ativação *sigmoid* e a função de ativação *tanh* não são adequadas para se aplicarem em camadas ocultas, uma vez que a inclinação da função torna-se muito pequena à medida que o valor de entrada da função é muito grande ou muito pequeno, o que, por sua vez, diminui a descida do gradiente. A função *ReLU* é a escolha muito preferida para Redes Neuronais com camadas ocultas, uma vez, que a derivada de função *ReLU* é 1 [39].

3.7.3 Topologia das Redes

A capacidade de aprendizagem de uma Rede Neuronal está enraizada na sua topologia, ou nos padrões e estruturas de interconexões dos neurónios. Existem diversas formas de arquitetura de redes, que podem ser diferenciadas por três características chaves:

- O número de camadas (layers);
- Como é que a informação na rede viaja, ou seja, se estas apenas viaja para a frente ou se também podem viajar para trás;
- O número de neurónios/nós dentro de cada camada da rede.

A topologia de uma determinada Rede Neuronal determina a sua capacidade de aprendizagem, sendo que geralmente redes maiores e mais complexas são capazes de identificar melhor padrões e limites de decisões complexas. No entanto, o poder de uma rede não é apenas uma função do seu tamanho, mas também da forma como as unidades são organizada [37].

3.7.3.1 Redes Neuronais de Camada Única

Como visto anteriormente, uma Rede Neuronal tem um vetor de entrada, com p variáveis $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T$, e constrói uma função usualmente não linear $f(\mathbf{x})$ para prever a resposta Y.

Em primeiro lugar aborda-se aqui o caso em que a variável resposta é quantitativa (regressão) e de seguida mostra-se as alterações que são efetuadas no caso da classificação, isto é, quando a variável resposta é categórica.

Para se entender, como é construída uma Rede Neuronal de camada única, apresentamos na Figura 3.3 um exemplo de uma Rede Neuronal com camada única [12].

Nesse exemplo tem-se representada uma Rede Neuronal *feed-forward* de camada única, para modelar uma variável resposta quantitativa, usando $p = 4$ variáveis explicativas.

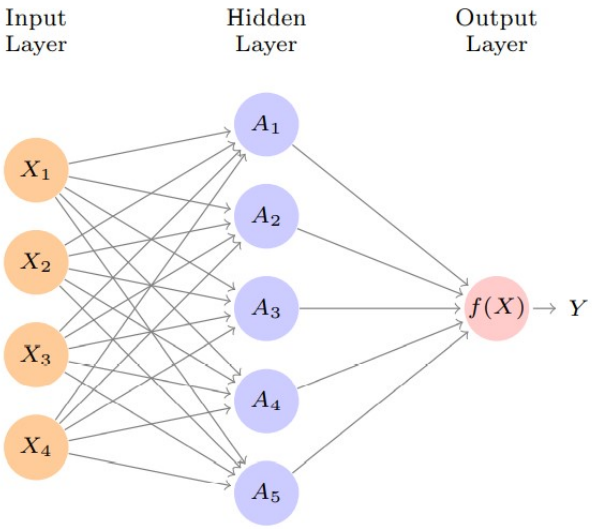


Figura 3.3: Rede Neuronal com camada única [12].

Como referido anteriormente, as variáveis explicativas compõem a camada de entrada, neste exemplo como se tem 4 variáveis explicativas então a camada de entrada é composta por 4 nós. As setas indicam que cada uma das variáveis explicativas que estão na camada de entrada estão a alimentar cada um dos k nós da camada oculta *(hidden)*, em que neste caso o $K = 5$. De salientar que o número de nós da camada oculta é escolhido pelo utilizador [12].

Cada um dos k, $k = 1, 2, 3, 4, 5$ nós da camada oculta, representa uma ativação A_k, $k = 1, 2, 3, 4, 5$, em que estas são calculadas como função das 4 variáveis explicativas, em que cada uma destas variáveis explicativas é referente a cada um dos 4 nós da camada de entrada.

Essas $K = 5$ ativações da camada oculta, alimentam de seguida a camada de saída, resultando num modelo de regressão linear nas $K = 5$ ativações.

O modelo de construção da Rede Neuronal de camada única é feito assim, no caso genérico, através de dois passos. No primeiro passo, tem-se que as K ativações A_k, com $k = 1, \ldots, K$ na camada oculta são calculadas como funções das p variáveis explicativas, em que estas são as variáveis de entrada no modelo. A ativação A_k é dada pela equação 3.109, onde $g()$ é uma função de ativação não linear, que esta pode ser uma das funções de ativações definidas anteriormente, na subsecção 3.7.2.

$$A_k = h_k(\mathbf{x}) = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_j\right) \tag{3.109}$$

De seguida, essas K ativações da camada oculta, alimentam a camada de saída, o que resulta num modelo de Regressão Linear nas K ativações, que pode ser representado pela equação 3.110, onde A_k nesta equação é o que está apresentado na equação 3.109.

$$f(\mathbf{x}) = \beta_0 + \sum_{k=1}^{K} \beta_k A_k \tag{3.110}$$

Posto isto, tem-se que um modelo de Rede Neuronal pode ser representado da forma da equação 3.111.

$$f(\mathbf{x}) = \beta_0 + \sum_{k=1}^{K} \beta_k h_k = \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_j\right) \tag{3.111}$$

A não linearidade da função de ativação g é essencial, pois uma vez que se ela fosse linear o modelo $f(\mathbf{x})$ em 3.111, colapsaria num modelo de Regressão Linear simples em $x_1, \ldots, x_p$. Além disso, ter uma função de ativação não linear permite que o modelo capture as não linearidades complexas e também os efeitos de interação [12].

O ajustamento de uma Rede Neuronal requer a estimação dos parâmetros desconhecidos, $\beta_0, \ldots, \beta_K$ e $w_{10}, \ldots w_{Kp}$. De referir que os parâmetros β_0 e w_{k0}, $k = 1, \ldots, K$ são chamados de *enviesamento*, em que estes têm o efeito de aumentar (*enviesamento* positivo) ou diminuir (*enviesamento* negativo) a entrada efetiva da função de ativação, ou seja, o *enviesamento* tem o poder de aumentar ou diminuir a soma de cada sinal de entrada multiplicado pelo respetivo peso num nó, para depois se aplicar esta quantidade resultante (sinal de entrada mais o *enviesamento*) à função de ativação. Já os parâmetros $\beta_1, \ldots, \beta_K$ e $w_{11}, w_{12} \ldots, w_{Kp}$ são os pesos, a força da entrada de cada um dos sinais de entrada num nó. Estes pesos podem ser positivos, negativos ou zero. Pesos negativos significam que o sinal é reduzido, pesos zero significam que não existe conexão entre os dois neurónios. O processo de ajustamento destes pesos é chamado de treino da Rede Neuronal.

No caso da variável resposta ser quantitativa, tipicamente é usada a perda de erro quadrático, de modo que os parâmetros estimados são os que minimizam a quantidade apresentada em 3.112, onde na subsecção 3.7.4 se encontram os detalhes de como realizar esta minimização.

$$\sum_{i=1}^{N} \Big(y_i - f(\mathbf{x}_i) \Big)^2 \tag{3.112}$$

No caso de um modelo de classificação o que mudaria, em relação a um problema de regressão, é somente a camada de saída, pois na camada de saída em vez de se ter apenas um nó, num problema de classificação com L classes, a camada de saída seria constituída por L nós, uma vez que quando se constrói uma rede ou um modelo com várias classes, a camada de saída da rede terá o mesmo número de nós/neurónios que o número de classes da variável resposta. No problema de classificação com L classes, tem-se que o *l-ésimo* nó da camada de saída modela a probabilidade da classe l, com $l = 1, \cdots, L$. Neste caso, tem-se L medições de y_l, $l = 1, \cdots, L$, em que cada uma destas y_l representa a *l-ésima* classe, isto é, tem-se L variáveis de saída, ou seja, as variáveis de saída são os y_l, $l = 1, \cdots, L$, em que cada uma destas está codificada como uma variável binária, isto é, apenas toma os valores 0 e 1 para a *l-ésima* classe.

No caso da classificação iremos ter L funções de saída, uma para cada uma das L classes. No entanto, conforme já foi mencionado, cada uma das l, $l = 1, \cdots, L$ unidades modela a probabilidade da classe, logo, tem-se que cada uma das l, $l = 1, \cdots, L$, funções de saída não pode ser modelada como no caso da regressão, ou seja, como se tem presente na equação 3.111, pois quer-se que as estimativas representem a probabilidade da classe, isto é, quer-se que $f_l(\mathbf{x}) = Pr(\mathrm{Y} = l|\mathbf{x})$, $l = 1, \cdots, L$, tal como no caso da Regressão Logística multinomial. Portanto, utiliza-se a função de ativação especial *Softmax*, que está definida em 3.113:

$$f_l(\mathbf{x}) = Pr(\mathrm{Y} = l|\mathbf{x}) = \frac{e^{Z_l}}{\sum_{r=1}^{L} e^{Z_r}} \tag{3.113}$$

onde Z_l é dado por 3.114.

$$Z_l = \beta_{l0} + \sum_{k=1}^{K} \beta_{lk} A_k \tag{3.114}$$

Para treinar uma Rede Neuronal, em que a resposta é qualitativa, procuram-se as estimativas dos coeficientes para os quais estes minimizam o simétrico da log-verosimilhança multinomial, isto é, que minimizam a quantidade 3.115.

$$-\sum_{i=1}^{N} \sum_{l=1}^{L} y_{il} \ln \Big(f_l(\mathbf{x}_i) \Big) \tag{3.115}$$

A quantidade de 3.115 também é conhecida como entropia cruzada, (*cross-entropty*) [12].

A função de ativação *Softmax* é uma combinação de várias funções *sigmoid*. Como visto na subsecção 3.7.2 uma função *sigmoid* apenas devolve valores no intervalo $]0, 1[$, e estes podem ser tratados como as probabilidades de uma determinada classe. A função *softmax* é utilizada para problemas de classificação de multiclasse, isto é, em problemas

de classificação em que o número de classes é maior que 2. Já as funções *sigmoid* são utilizadas para a classificação binária, isto é, quando a variável resposta apenas toma dois valores, 0 ou 1.

Assim, se apenas se estiver presente sob um problema de classificação binária, então apenas se terá uma função de saída, em que esta modela a probabilidade de sucesso, ou seja, a função de saída a ser utilizada é a função de ativação *sigmoid*, representada na equação 3.105. Posto isto, tem-se que a representação da Rede Neural quando a variável resposta é binária é análoga ao exemplo da Rede Neural apresentada na Figura 3.3, mas neste caso a função de saída f é dada por 3.116, onde A_K é definido em 3.109.

$$f(\mathbf{x}) = Pr(\mathrm{Y} = 1|\mathbf{x}) = \frac{e^{\beta_0+\sum_{k=1}^{K}\beta_k A_k}}{1+e^{\beta_0+\sum_{k=1}^{K}\beta_k A_k}} \tag{3.116}$$

Mais uma vez, para o ajustamento da Rede Neural no caso de classificação binária, envolve a estimação dos parâmetros presentes na equação 3.116, parâmetros $\beta_0, \ldots, \beta_K$ e $w_{10}, \ldots, w_{Kp}$. No entanto neste caso procuram-se as estimativas dos coeficientes que minimizam o simétrico da log-verosimilhança, ou seja, que minimizam a quantidade 3.117.

$$-\sum_{i=1}^{n}\left(y_i \ln\Big(f(\mathbf{x}_i)\Big) + (1-y_i)\ln\Big(1-f(\mathbf{x}_i)\Big)\right) \tag{3.117}$$

Na subsecção 3.7.4 encontra-se os detalhes de como realizar esta minimização.

3.7.3.2 Redes Neuronais de Multi-camadas

Nesta presente sub-subsecção, descreve-se a construção de uma Rede Neural de Multi-Camada, pois uma forma óbvia de criar redes mais complexas é através da adição de camadas ocultas. As Redes Neuronais modernas normalmente têm mais do que uma camada oculta e muitas vezes, têm muitos nós por camada. Na teoria, uma única camada oculta com um elevado número de nós tem a capacidade de aproximar a maioria das funções, no entanto, a tarefa de aprendizagem de descobrir uma boa solução é muito mais fácil numa rede com várias camadas ocultas, em que cada uma destas tem um tamanho modesto de nós.

Uma rede multi-camada adiciona uma ou mais camadas ocultas, em que estas processão os sinais dos nós de entrada antes destes sinais atingirem os nós de saída. A maioria das redes de multi-camada estão totalmente conectadas, isto é, cada um dos nós de uma camada oculta está conectado a um outro nó da próxima camada, no entanto, isto, não é obrigatório acontecer.

Na Figura 3.4, tem-se representada uma Rede Neuronal, que pertence a um exemplo de um problema de classificação de imagens, em que estas se referem à classe de números de 0 a 9 [12].

Como se pode observar através da Figura 3.4, tem-se que a camada de saída é constituída por 10 variáveis, em que cada uma dessas variáveis de saída corresponde à classe de

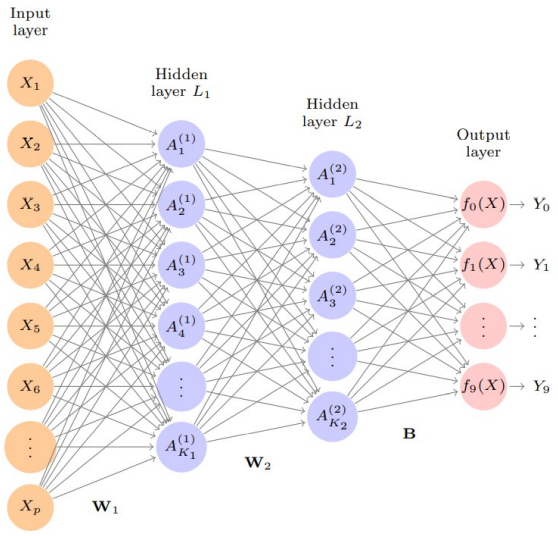


Figura 3.4: Rede Neuronal com camada Dupla [12].

cada um dos números de 0 a 9. Apenas se irá utilizar a Figura 3.4, para nos guiar como se constrói uma Rede Neural com duas camadas ocultas, para um problema de classificação binário, em que neste caso não vamos considerar as 10 classes presentes nesta figura, mas vamos considerar que a presente figura apenas é constituída por um nó na camada final tal como a Figura 3.3.

Tal como no caso da construção de uma Rede Neuronal de camada única oculta, a camada de entrada é constituída por todas as variáveis explicativas p, ou seja, os sinais de entrada são as p variáveis explicativas $\mathbf{x} = (x_1, \ldots, x_p)^T$. De seguida, cada um destes sinais, ou seja, cada uma destas variáveis explicativas são ponderados pelos pesos para alimentar cada um dos K_1 nós da primeira camada oculta, ou seja, o nó $k = 1, \ldots, K_1$ da primeira camada oculta é alimentado por cada uma das p variáveis explicativas ponderadas pelo seu respetivo peso $w_{kj}^{(1)}$, $j = 1, \ldots, p$. De seguida estes p sinais de entrada em cada um dos k nós, $k = 1, \ldots, K_1$ é somado e é adicionado em cada um dos k nós, $k = 1, \ldots, K_1$ da primeira camada oculta um peso $w_{k0}^{(1)}$ que é chamada de *enviesamento*, e que tem o efeito de aumentar ou diminuir a quantidade apresentada em 3.118, que é a soma de cada um dos sinais de entrada devidamente ponderados, para o nó k, da primeira camada oculta.

$$\sum_{j=1}^{p} w_{kj}^{(1)} x_j \tag{3.118}$$

De seguida é utilizada uma função de ativação, que permite passar um sinal para a próxima camada oculta. Assim, na primeira camada oculta, tem-se que a ativação $A_k^{(1)}$, para cada um dos $k = 1, \ldots, K_1$ nó, é dada pela equação 3.119.

$$A_k^{(1)} = h_k^{(1)} = g\left(w_{k0}^{(1)} + \sum_{j=1}^{p} w_{kj}^{(1)} x_j\right), \quad \text{para } k = 1, \ldots, K_1 \tag{3.119}$$

De seguida a segunda camada oculta é constituída por K_2 nós, em que cada um destes nós $r = 1, \ldots, K_2$ vai receber como sinais, cada uma das ativações da primeira camada oculta, ou seja, as ativações da primeira camada oculta vão servir como sinais de entrada na segunda camada oculta. No entanto, tal como a entrada dos sinais na primeira camada oculta eram ponderadas por pesos, conforme a sua importância, neste caso também se tem a mesma situação, ou seja, o nó $r, r = 1, \ldots, K_2$ da segunda camada oculta é alimentado por cada uma das K_1 ativações da primeira camada oculta ponderadas pelos seus respetivos pesos $w_{rk}^{(2)}, k = 1, \ldots, K_1$. Estes K_1 sinais de entrada em cada um dos r nós, $r = 1, \ldots, K_2$, é somado e é adicionado em cada um dos r nós da segunda camada oculta um peso $w_{r0}^{(2)}$ que é o *enviesamento*, onde este aumenta ou diminui a entrada efetiva na função de ativação. A função de ativação em cada um dos K_2 nós, permite passar o sinal, sendo que a ativação $A_r^{(2)}$ no *r-ésimo* nó na segunda camada oculta, $r = 1, \ldots, K_2$, é dada pela equação 3.120.

$$A_r^{(2)} = h_r^{(2)} = g\left(w_{r0}^{(2)} + \sum_{k=1}^{K_1} w_{rk}^{(1)} A_k^{(1)}\right), \text{ para } r = 1, \ldots, K_2 \tag{3.120}$$

Observe-se que cada uma das ativações na segunda camada oculta $A_r^{(2)} = h_r^{(2)}$ é uma função do vetor de entrada $\mathbf{x} = (x_1, \ldots, x_p)^T$, pois, uma vez que cada uma das $A_r^{(2)}$ ativações da segunda camada oculta é uma função da ativação $A_k^{(1)}$ da primeira camada oculta, em que cada uma destas por sua vez são uma função de $\mathbf{x}$. Logo, por este motivo, é que cada uma das ativações na segunda camada oculta $A_r^{(2)} = h_r^{(2)}$ é uma função do vetor de entrada $\mathbf{x}$.

Este processo, repete-se conforme o número de camadas ocultas selecionadas.

Na Rede Neuronal apresentada na Figura 3.4, a segunda camada oculta está a alimentar a camada de saída, ou seja, a última camada oculta de uma Rede Neuronal alimenta a camada de saída. Assim, através de uma cadeia de transformações, a rede é capaz de construir transformações bastantes complexas de $\mathbf{x}$ que, em última análise alimentam a camada de saída.

De referir que a notação descrita anteriormente como $A_r^{(2)}$ e $w_{rk}^{(2)}$ em 3.120, servem para nos indicar qual a camada que as ativações e os pesos pertencem, que neste caso pertencem à segunda camada oculta. Já a notação $\boldsymbol{W}_1$ na Figura 3.4 representa a matriz de pesos de dimensão $\Big((p+1) \times K_1\Big)$, que se alimenta da camada de entrada, para alimentar a primeira camada oculta, ou seja, em cada um dos nós da primeira camada oculta é alimentada através da combinação linear dos pesos com as variáveis explicativas. De referir que tal como no caso da Rede Neuronal de camada única, os peso $w_{r0}^{(1)}$, corresponde ao *enviesamento*, em que este tem o efeito de aumentar ou diminuir a entrada efetiva em cada uma das ativações $A_k^{(1)}, k = 1, \ldots, K_1$.

Cada elemento $A_k^{(1)}$ alimenta a segunda camada oculta, através da matriz de pesos $\boldsymbol{W}_2$, em que esta tem dimensão $\Big((K_1+1) \times K_2\Big)$. De notar que os peso $w_{r0}^{(2)}$, corresponde ao *enviesamento*, em que este tem o efeito de aumentar ou diminuir a entrada efetiva em cada uma das ativações $A_r^{(2)}, k = 1, \ldots, K_2$.

Chegando à camada de saída, tem-se um único nó. Tal como na Rede Neuronal de camada única a função de saída utilizada para modelar um problema de classificação binário é a função de ativação *sigmoid*.

A segunda camada oculta é que está a alimentar esta última camada, a camada de saída, assim, os sinais recebidos por este nó serão cada uma das ativações da segunda camada, ou seja, as ativações da primeira camada vão servir como sinais de entrada na camada de saída, em que cada um destes sinais é ponderado de acordo com a sua importância, ou seja, cada uma das ativações $A_r^{(2)}$, $r = 1, \ldots, K_2$ é ponderada pelo seu respetivo peso β_r, $r = 1, \ldots, K_2$. Posto isto cada um destes sinais será somado com os seus respetivos pesos no nó da camada de saída, sendo que é adicionado um termo β_0, em que este é o *enviesamento*, que tem o efeito de aumentar ou diminuir a entrada efetiva da função de ativação. Assim, a função de ativação *sigmoid* é utilizada, para dar o sinal final de saída deste nó, em que o sinal final será o representado na equação 3.121.

$$f(\mathbf{x}) = Pr(\mathrm{Y} = 1|\mathbf{x}) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{r=1}^{K_2} \beta_r A_r^{(2)}\right)}} \tag{3.121}$$

Tal como no caso da Rede Neuronal de Camada única, para se ajustar uma Rede Neuronal é necessário a estimação de cada um dos parâmetros, em que neste caso são os parâmetros da matriz $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ e o $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{K_2})^T$. No caso da classificação binária, mais uma vez é utilizada a minimização do simétrico da log-*verosimilhança* para estimar os respetivos parâmetros, ou seja, o objetivo é encontrar os parâmetros de $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ e de $\boldsymbol{\beta}$ que minimizam a quantidade representada em 3.122.

$$-\sum_{i=1}^{n} \left(y_i - \ln\left(f(\mathbf{x}_i)\right) + (1 - y_i) \ln\left(1 - f(\mathbf{x}_i)\right) \right) \tag{3.122}$$

A função f representada em 3.122 é a função f representada em 3.121.

Se estivéssemos presente sobre um problema de regressão, a camada de saída também era composta por um único nó, no entanto, em vez de se utilizar a função de ativação *sigmoid* para dar o sinal final, utilizar-se-ia a função de ativação linear, pois a função de ativação *sigmoid* é utilizada para converter os sinais de entrada desse nó num sinal de saída em que esse sinal representa uma probabilidade. Então o sinal final no caso da regressão em vez de ser dado por 3.121 é dado pela fórmula 3.123.

$$f(\mathrm{x}) = \beta_0 + \sum_{r=1}^{K_2} \beta_r A_r^{(2)} \tag{3.123}$$

No entanto, neste caso, o ajustamento da Rede Neuronal é feita à custa da minimização do erro quadrático, ou seja, o objetivo é encontra os parâmetros de $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ e de $\boldsymbol{\beta}$ que minimizem a quantidade 3.124.

$$\sum_{i=1}^{n} \left(y_i - f(\mathbf{x}_i) \right)^2 \tag{3.124}$$

3.7.3.3 Ligação entre os Nós

As Redes Neuronais, podem ser classificadas em dois grupos com base na conexão entre os nós, sendo estes grupos [40]:

1. Redes Neuronais *feed-forward*;
2. Redes Neuronais *recorrentes* (*Recurrent*).

Nas Redes Neuronais *feed-forward*, tem-se que a rede é alimentada ciclicamente, ou seja, os sinais de entrada propagam-se para a frente através da rede, camada por camada. Ou seja, nas Redes Neuronais *feed-forward*, tem-se que a informação flui apenas numa direção, da camada de entrada para a camada de saída, através de nós ocultos se estes existirem, sem círculos ou *loopbacks*. Exemplos de duas Redes Neuronais *feed-forward* estão nas Figura 3.3 e Figura 3.4, sendo que a primeira tem apenas uma camada oculta, enquanto que a segunda apresenta duas camadas ocultas.

Já numa Rede Neural *recorrente*, os neurónios formam um ciclo, ou seja, tem-se que os nós da camada oculta podem formar um ciclo. Assim, tem-se que o sinal de saída de uma camada, torna-se o sinal de entrada para a próxima camada, mas também se torna no sinal de entrada para si mesmo, formando um ciclo ou *loop* de *feedback*. Isso permite que a rede tenha memória sobre os estados anteriores e utilize isso para influenciar a sua saída atual, aumentando imensamente o poder das redes, pois permite que estas têm a capacidade de compreender uma sequência de eventos ao longo de um período de tempo. As Rede Neuronais *recorrentes* são utilizadas quando se tem uma sequência de eventos ao longo de um período de tempo, tornando-se muito útil para aplicações que envolvem o processamento de dados, que sejam uma sequência de eventos ao longo de um período de tempo, como por exemplo, o reconhecimento de fala, previsão do mercado de ações ou previsão do tempo [41].

3.7.4 Algoritmo de Treino - Ajustamento das Redes Neuronais

O ajustamento de Redes Neuronais é um processo bastante complexo, pelo que neste caso apenas se irá apresentar uma breve noção para o ajustamento de uma Rede Neuronal *feed-forward*, apenas constituída por uma única camada oculta. O ajustamento de Redes Neuronais mais complexas, generaliza este caso. Existem vários *software* estatísticos que implementam a estimação de uma Rede Neuronal. Por exemplo, *software* R, tem diversas bibliotecas que permitem estimar os parâmetros de uma Rede Neuronal, com várias camadas ocultas, entre os quais a biblioteca *neuralnet* e a biblioteca *keras*.

Como visto anteriormente, o modelo de Redes Neuronais tem parâmetros desconhecidos muitas vezes chamados de pesos, em que esses pesos são valores a serem estimados, ou seja, procuramos valores para eles de modo a que o modelo se ajuste bem aos dados de treino.

No modelo 3.111, tem-se que os parâmetros a serem estimados são $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$, assim como cada um dos $\boldsymbol{w}_k = (w_{k0}, w_{k1}, \ldots, w_{kp})^T$, $k = 1, \ldots, K$. Dadas as observações $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, pode-se ajustar o modelo resolvendo um problema não linear de mínimos quadrados, que se encontra representado na equação 3.125.

$$\min_{\{w_k\}_1^K, \boldsymbol{\beta}} \frac{1}{2} \sum_{i=1}^{n} \Big(y_i - f(\mathbf{x}_i) \Big)^2 \tag{3.125}$$

onde $f(\mathbf{x}_i)$, encontra-se representado na equação 3.126.

$$f(\mathbf{x}_i) = \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij} \right) \tag{3.126}$$

Apesar do objetivo de 3.125 parecer fácil de se concretizar, devido ao arranjo encaixado dos parâmetros e à simetria das unidades ocultas, não é assim tão fácil a sua minimização. Assim, tem-se que o problema em questão é um problema não convexo nos parâmetros, pelo que existem várias soluções. Para superar este problema, assim como o problema do sobreajustamento (*overfitting*), duas estratégias gerais podem ser empregues no ajustamento de Redes Neuronais. As duas estratégias gerais são as seguintes [12]:

- Aprendizagem Lenta: o modelo é ajustado de forma iterativa, mas um pouco lenta, usando o gradiente descendente. O processo de ajuste é então interrompido quando o sobreajustamento (*overfitting*) é detetado;
- Regularização (*Regularization*): penalidades são impostas nos parâmetros, usualmente utiliza-se LASSO ou *Ridge*, como visto na secção 3.3.

Pode-se reescrever o objetivo de 3.125 como em 3.127, onde $\boldsymbol{\theta}$ é um vetor que contém todos os parâmetros da equação 3.125.

$$R(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{n} \Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i) \Big)^2 \tag{3.127}$$

Na equação 3.127, tem-se explicita a dependência de f nos parâmetros.

No caso de classificação binária, o modelo encontra-se representado em 3.116, pelo que os parâmetros a serem estimados são $\boldsymbol{w}_k = (w_{k0}, \ldots, w_{kp})^T$ com $k = 1, \ldots, K$ e $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)^T$. Como referido anteriormente a estimação destes parâmetros era feita através da minimização do simétrico da log-verosimilhança, ou seja, pelo que o objetivo encontra-se representado na equação 3.128.

$$\min_{\{w_k\}_1^K, \boldsymbol{\beta}} \left\{ -\sum_{i=1}^{n} \Big(y_i \ln\Big(f(\mathbf{x}_i)\Big) + (1 - y_i) \ln\Big(1 - f(\mathbf{x}_i)\Big) \Big) \right\} \tag{3.128}$$

Na equação 3.128, a função $f(\mathbf{x}_i)$ é a função dada em 3.129.

$$f(\mathbf{x}_i) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\right)\right)}} \tag{3.129}$$

Tal como se fez anteriormente, pode-se reescrever o objetivo de 3.128, à custa de um vetor $\boldsymbol{\theta}$, que contém todos os parâmetros da equação 3.128. O objetivo de 3.128, escrito à custa de um vetor $\boldsymbol{\theta}$, é dado na equação 3.130 [12] .

$$R(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \left(y_i \ln\left(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) + (1 - y_i) \ln\left(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) \right) \tag{3.130}$$

O objetivo de 3.128 é um problema não convexo nos parâmetros, pelo que existem múltiplas soluções para o problema em questão [12].

Uma abordagem genérica para minimizar o $R(\boldsymbol{\theta})$ dado na equação 3.127 (caso da regressão) e $R(\boldsymbol{\theta})$ dado na equação 3.130 (caso da classificação binária) é através do gradiente descendente, chamado, de *backpropagation*, no caso das Redes Neuronais *feed-fordward* [17].

A ideia do gradiente descendente é muito simples e pode ser resumida como se encontra na Tabela 3.8.

Tabela 3.8: Gradiente Descendente [12].

Gradiente Descendente
1. Começe com uma estimativa para cada um dos parâmetros de $\boldsymbol{\theta}$, denotada por $\boldsymbol{\theta}^0$, e defina $t = 0$;
2. Iterar até que o objetivo da equação 3.125 não diminui mais:
a) Encontre um vetor δ que reflita uma pequena mudança em $\boldsymbol{\theta}$, ou seja, tal que $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \delta$ reduz o objetivo; i.e, tal que $R(\boldsymbol{\theta}^{t+1}) < R(\boldsymbol{\theta}^t)$;
b) Defina $t \leftarrow t + 1$.

3.7.4.1 *Backpropagation*

O algoritmo *backpropagation* é um algoritmo que utiliza como estratégia a retro-propagação de erros, pois este algoritmo itera através de muitos ciclos de dois processos. Cada iteração do algoritmo é conhecido como uma *época*. Como a rede não tem conhecimento à *priori*, normalmente os pesos são definidos aleatoriamente no início. Os ciclos incluem [37]:

- Uma fase *forward* na qual os neurónios são ativados em sequência a partir da camada de entrada para a camada de saída, aplicando os pesos de cada nó e a função de ativação. Ao atingir a camada final, um sinal de saída é produzido.

- Uma fase *backward* na qual o sinal de saída resultante da fase *forward* é comparado com o verdadeiro valor de destino nos dados de treino. A diferença entre o sinal de saída da rede e o verdadeiro valor resulta num erro que é propagado para trás na rede, para modificar os pesos de conexão entre os nós e reduzir o futuro erro.

O algoritmo *backpropagation*, utiliza a técnica do gradiente descendente.

O gradiente de $R(\boldsymbol{\theta})$ avaliado em algum valor atual de $\boldsymbol{\theta} = \boldsymbol{\theta}^m$, é o vetor de derivadas parciais nesse ponto, em que a sua representação é dada pela equação 3.131.

$$\nabla R(\boldsymbol{\theta}^m) = \frac{\partial R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\theta}=\boldsymbol{\theta}^m} \tag{3.131}$$

Isto dá-nos a direção no espaço de $\boldsymbol{\theta}$ em que $R(\boldsymbol{\theta})$ aumenta mais rapidamente. A ideia do gradiente descendente é mover um pouco $\boldsymbol{\theta}$ na direção oposta, ou seja, como se encontra representado na equação 3.132.

$$\boldsymbol{\theta}^{m+1} \leftarrow \boldsymbol{\theta}^m - \rho R(\boldsymbol{\theta}^m) \tag{3.132}$$

Para um valor suficientemente pequeno de ρ, este passo diminuirá o objetivo de $R(\boldsymbol{\theta})$, isto é $R(\boldsymbol{\theta}^{m+1}) \leq R(\boldsymbol{\theta}^m)$. Se o vetor do gradiente for identicamente igual a 0, podemos ter chegado a um mínimo do objetivo [12].

Resumindo, o algoritmo *backpropagation*, usa as derivadas da função de ativação de cada nó, para identificar a direção de cada um dos pesos de entrada através do gradiente, daí a importância das funções de ativação serem diferenciáveis. O gradiente sugere quão acentuadamente o erro será reduzido ou aumentado para uma mudança no peso. O algoritmo altera os pesos que resultam na maior redução do erro, por um valor conhecido como taxa de aprendizagem, ρ. Quanto maior a taxa de aprendizagem, mais rápido o algoritmo chegará ao objetivo, o que pode reduzir o tempo de treino, com o risco de passar o melhor valor que minimiza o objetivo [37].

Tanto no caso da regressão como no caso da classificação o $R(\boldsymbol{\theta})$ pode ser escrito como se encontra em 3.133, caso da regressão e 3.134, caso da classificação binária.

$$R(\boldsymbol{\theta}) = \sum_{i=1}^{n} R_i(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^{n} \Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)^2 \tag{3.133}$$

$$R(\boldsymbol{\theta}) = \sum_{i=1}^{n} R_i(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \Big(y_i \ln\Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) + (1 - y_i) \ln\Big(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)\Big) \tag{3.134}$$

De notar que a função $f_{\boldsymbol{\theta}}$ em 3.133 é dada por 3.126. Já o $f_{\boldsymbol{\theta}}$ de 3.134 tem a representação de 3.129.

Uma vez que, os $R(\boldsymbol{\theta})$ tanto no caso da regressão (equação 3.133) como no caso da classificação binária (equação 3.134) são uma soma, então os seus gradientes também são uma soma sobre as n observações. Assim, para simplificar iremos calcular as suas derivadas em apenas num desses termos. Considere-se o $R_i(\boldsymbol{\theta})$ representado em 3.135, para o caso da regressão e o $R_i(\boldsymbol{\theta})$ representado em 3.136, para o caso da classificação binária.

$$R_i(\boldsymbol{\theta}) = \frac{1}{2}\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)^2 \tag{3.135}$$

$$R_i(\boldsymbol{\theta}) = -\Big(y_i \ln\Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) + (1-y_i)\ln\Big(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)\Big) \tag{3.136}$$

No apêndice C, detalha-se o cálculo das derivados dos $R_i(\boldsymbol{\theta})$ tanto para o caso da regressão como para o caso da classificação binária.

No caso da regressão as derivadas parciais de $R(\boldsymbol{\theta})$ com respeito a β_k, $k = 1, \ldots, K$ são dadas pela equação 3.137, enquanto que as derivadas parciais de $R(\boldsymbol{\theta})$ com respeito a w_{kj}, $k = 1, \ldots, K$ e $j = 1, \ldots, p$ são dadas pela equação 3.138.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k} = -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) g(z_{ik}) \tag{3.137}$$

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} = -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) \beta_k g'(z_{ik}) x_{ij} \tag{3.138}$$

Já as derivadas parciais de $R(\boldsymbol{\theta})$ com respeito a β_0 é apresentada na equação 3.139, enquanto que a derivadas parciais de $R(\boldsymbol{\theta})$ com respeito a w_{k0}, $k = 1, \ldots, K$, são dadas pela equação 3.140.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_0} = -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) \tag{3.139}$$

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{k0}} = -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) \beta_k \tag{3.140}$$

No caso da classificação binária as derivadas parciais de $R_i(\boldsymbol{\theta})$, apresentam-se nas equações 3.141 a 3.144.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k} = \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\Big) g\Big(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\Big) \tag{3.141}$$

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} = \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\Big) g'(z_{ik}) x_{ij} \tag{3.142}$$

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_0} = \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\Big) \tag{3.143}$$

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{k0}} = \Big(f_\theta(\mathbf{x}_i) - y_i\Big) g'(z_{ik}) \tag{3.144}$$

Dadas as derivadas, uma atualização do gradiente descendente na $(r+1)$-*ésima* iteração tem a forma de 3.145,

$$\begin{cases} \beta_k^{(r+1)} = \beta_k^{(r)} - \rho \sum_{i=1}^n \dfrac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k^{(r)}} \\ w_{kj}^{(r+1)} = w_{kj}^{(r)} - \rho \sum_{i=1}^n \dfrac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}^{(r)}} \end{cases} \tag{3.145}$$

onde ρ é a taxa de aprendizagem.

Assim, usando 3.145, as atualizações podem ser implementadas com um algoritmo em duas fases, o algoritmo *backpropagation* que descrevemos inicialmente. Na primeira fase, fase *forward*, os pesos atuais são fixos e através de cada uma das funções f no caso da regressão função dada em 3.126 e no caso da classificação binária função dada em 3.129, prevê-se o seu valor, ou seja, é produzido um sinal final na rede. Na segunda fase, fase *backward*, o valor previsto é comparado com o verdadeiro valor, através de $\Big(y_i - \hat{f}_\theta(\mathbf{x}_i)\Big)$ e $\Big(\hat{f}_\theta(\mathbf{x}_i) - y_i\Big)$ os erros de regressão e de classificação, respetivamente para a regressão e classificação binária. Posto isto, os erros são propagados para trás na rede, através da fórmula 3.145 modificando os pesos da conexão entre os nós.

3.7.4.2 *Gradiente Descendente Estocástico*

O gradiente descendente geralmente leva muitas iterações até atingir um mínimo local, e na prática existem uma série de abordagens que aceleram o processo.

Geralmente quando o valor de n é grande, em vez de se somar todas as n observações de cada uma das derivadas $R_i(\theta)$, apenas se calcula uma pequena fração delas cada vez que calculamos um passo do gradiente descendente. Este processo é conhecido como o *gradiente descendente estocástico*, e é muito utilizado para ensinar Redes Neuronais profundas [12].

3.7.4.3 Ajuste da Rede

O número de camadas ocultas e o número de nós por cada camada oculta são dados pelo utilizador. No entanto, alguns investigadores utilizam validação cruzada para estimar o número ótimo de camadas ocultas e o número de neurónios por camada oculta [17].

3.7.5 Biblioteca *Keras*

Neste presente trabalho foi utilizada a biblioteca *keras* do *software* R, para se ajustar uma Rede Neuronal aos dados. Esta biblioteca, tem como funções de ativação todas aquelas que se viu nesta secção. Como método para a estimação dos parâmetros esta biblioteca tem o *gradiente descendente estocástico* como opção e este foi utilizado neste trabalho. Por defeito

utiliza uma taxa de aprendizagem de $\rho = 0.01$. Os pesos iniciais com que esta biblioteca começa são 0 para os *enviesamentos* e os restantes pesos são inicializados por defeito através do iniciador uniforme *Glorot*. Este iniciador uniforme seleciona valores de uma distribuição uniforme em $[-limit, limit]$, onde $limit = \sqrt{6/(t_1 + t_2)}$, onde t_1 representa o número de unidades de entrada num nó e t_2 representa o número de unidades de saída do nó.

Por defeito esta biblioteca, utiliza 10 épocas e 32 observações para atualizar o gradiente. Para mais informação, veja-se o documento "Package keras" [42].

3.8 Estimação da Performance Futura dos Modelos

3.8.1 O Método *Holdout*

O método *holdout* particiona uma base de dados em duas base de dados, a base de dados de treino e a base de dados de teste. Na Figura 3.5, tem-se uma representação esquemática de como o presente método funciona.

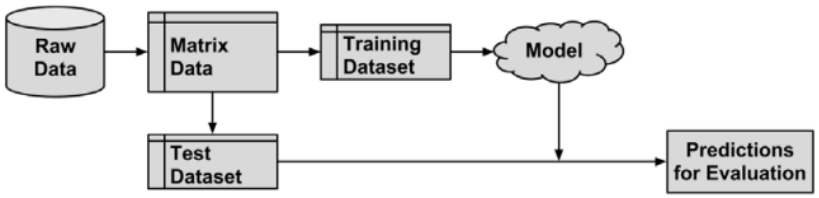


Figura 3.5: Método *Holdout* [37].

Conforme está representado na Figura 3.5, a base de dados de treino é utilizada para estimar o modelo, que de seguida é aplicado ao conjunto de dados de teste para gerar previsões para avaliação. Normalmente, cerca de um terço dos dados vão ser para teste, enquanto os restantes dois terços dos dados serão utilizados para treinar o modelo, mas esta proporção pode variar, dependendo da quantidade de dados disponíveis. A divisão da base de dados original na base de dados de treino e de teste é feita aleatoriamente, garantindo que estes dois conjuntos não têm diferenças sistemáticas [37].

Um dos problemas deste método é que cada uma das partições pode ter diferente representatividade dos valores da variável resposta. Em alguns casos, como por exemplo, se uma dada classe de uma variável resposta categórica apresentar uma proporção muito baixa na base de dados, quando se divide esta base de dados em dados de treino e teste, isto pode-nos levar à omissão dessa classe no conjunto de dados de treino, e neste caso, está-se perante um problema significativo, uma vez que o modelo não pode aprender essa classe. Uma forma de reduzir as chances deste problema ocorrer é utilizar uma técnica chamada de amostragem aleatória estratificada. Embora em média, uma amostra aleatória contenha aproximadamente a mesma proporção de cada uma das classes da variável resposta que no conjunto de dados completos, a amostra aleatória estratificada garante

que as partições aleatórias geradas tenham aproximadamente a mesma proporção de cada classe como no conjunto completo [37].

No caso da amostra aleatória estratificada, em vez de se selecionar as observações aleatoriamente para a base de dados de treino e as restantes para a base de dados de teste, vai-se a cada uma das classes da variável resposta e seleciona-se a percentagem de observações de cada uma das classes que irão pertencer à base de dados de treino. No entanto, é necessário que a percentagem para a divisão, em cada uma das classes seja a mesma, para garantir que a base de dados de treino tenha a mesma proporção de cada uma das classes que na base de dados original, e o mesmo para a base de dados de teste.

Embora a amostra aleatória estratificada garanta aproximadamente a mesma proporção de cada uma das classes, não garante outros tipos de representatividade, pois algumas amostras podem ter poucos casos difíceis de prever ou ter muitos valores discrepantes. Isto é especialmente problemático para conjuntos de dados pequenos [37].

3.8.2 Validação Cruzada

A repetição do método *holdout* é a base da técnica conhecida como validação cruzada com k grupos, (*k-fold cross-validation ou k-fold CV*), mas em vez de colecionar amostras aleatórias repetidas que poderiam utilizar as mesmas observações mais de uma vez, a *k-fold CV* divide os dados em k partições aleatórias completamente separadas em que são chamadas de grupos. A proporção de observações é igual para cada uma dos k grupos. Esta técnica tornou-se na técnica mais utilizada para estimar o desempenho de um modelo [37].

Validação cruzada com k grupos, divide o conjunto das observações em k grupos, em que estas têm aproximadamente o mesmo número de observações. De seguida, o primeiro grupo é tratado como o conjunto de dados de teste, enquanto que os restantes $(k-1)$ grupos são utilizados para treinar o modelo. Depois de treinado o modelo com os $(k-1)$ grupos, avalia-se o desempenho do modelo com o primeira grupo, em que este é referente aos dados de teste. Concluindo este passo, passa-se para o segundo grupo, sendo que este desta vez será tratado como o conjunto de dados de teste, enquanto os restantes $(k-1)$ grupos serão tratados como dados de treino, para treinar o modelo. Depois de treinado o modelo, avalia-se agora o desempenho do modelo com os dados de teste, em que estes são os dados do segundo grupo. Este processo é repetido k vezes. No final, existem k valores do desempenho do modelo, uma em cada um dos k grupos. Com esses k valores do desempenho do modelo, calcula-se a sua média, ficando assim com o desempenho global do modelo utilizando a validação cruzada com k grupos. Observe-se que as medidas para o desempenho do modelo pode ser qualquer uma das métricas para a avaliação futura do modelo que se encontram na secção 3.9. Assim, a estimativa do desempenho do modelo, pelo método da validação cruzada utilizando k grupos, é dada pela fórmula 3.146.

$$CV_k = \frac{1}{k} \sum_{i=1}^{k} M_i \tag{3.146}$$

O M_i utilizado na fórmula 3.146, diz respeito ao valor da métrica utilizada para avaliar o desempenho do modelo referente ao *i-ésimo* grupo, $i = 1, \ldots, k$, ou seja, o *i-ésimo* grupo é os dados de teste.

Na Tabela 3.9, tem-se um resumo do processo da avaliação do desempenho global de um modelo, usando validação cruzada com k grupos e utilizando uma métrica M, que pode ser uma das métricas apresentadas na secção 3.9.

Tabela 3.9: Método validação cruzada com k grupos para a avaliação do desempenho de um modelo.

Método validação cruzada com k grupos para a avaliação do desempenho global de um modelo
1. Escolha um valor de k e uma métrica M para avaliar o desempenho do modelo;
2. Para $i = 1, \ldots, k$ fazer:
a) treino o respetivo modelo com a base de dados de treino, em que esta é constituída por todas as observações de todos os grupos à exceção do grupo i, em que este grupo será os dados de teste;
b) Avalie o desempenho do modelo obtido em *a)* com os dados de teste, em que estes são os dados do grupo i. Seja M_i o valor do desempenho do modelo, com a métrica M;
3. Depois de ter percorrido todos os passos de 2. k vezes, calcule a seguinte quantidade $CV_k = \frac{1}{k} \sum_{i=1}^{k} M_i$, em que esta é a estimativa de validação cruzada com k grupos.

O valor de k pode ser um valor natural qualquer, mas normalmente é comum usar validação cruzada com $k = 10$ grupos, uma vez que, evidências empíricas sugerem que existe um beneficio adicional muito reduzido, quando se usa um número maior [37]. Uma outra evidência para apenas se utilizar apenas $k = 10$ grupos é a vantagem computacional, pois a validação cruzada com k grupos requer que o modelo se ajuste k vezes e isto pode ser computacionalmente dispendioso, uma vez que, muitos dos métodos de aprendizagem de estatística têm um procedimento de ajuste computacionalmente intensivo e portanto ao executar a validação cruzada com um valor de k muito grande, isto pode apresentar problemas computacionais [12].

3.9 Métodos de Avaliação do Desempenho dos Modelos de Classificação

Existem diversos métodos que permitem avaliar se um dado modelo ajustado aos dados de treino possui boas qualidades preditivas quando ajustado a dados que nunca tinha visto antes, ou seja, queremos avaliar se o modelo ajustado nos dados de treino consegue prever corretamente novos dados ou, dito de outro modo, queremos avaliar a sua capacidade preditiva, isto é, se consegue prever corretamente o maior número de vezes possíveis.

3.9.1 Matriz de Confusão

A matriz de confusão é uma tabela que categoriza as previsões de acordo com a sua correspondência ao valor real nos dados [37]. Uma das dimensões da tabela indica as possíveis categorias de valores previstos, enquanto que a outra dimensão indica o mesmo valor para os valores reais [37]. No entanto, como no presente trabalho, se está sobre um problema de classificação binário, ou seja, com apenas duas classes, a classe de sucesso e a classe de insucesso, apenas iremos tratar das matrizes (2×2). A Tabela 3.10, representa uma matriz de confusão para o caso binário.

Tabela 3.10: Matriz de confusão para o caso binário.

		Valores Previstos	
		0	1
Valores	0	TN	FP
Verdadeiros	1	FN	TP

Quando o valor previsto é igual ao valor real, esta é uma classificação correta. As previsões corretas caem na diagonal da matriz de confusão (indicada por TN e TP) [37]. As células da matriz fora da diagonal (indicadas por FN e FP) indicam os casos em que o valor previsto difere do valor real [37]. Essas são previsões incorretas. Existem medidas para avaliar os desempenhos de um modelo de classificação que são baseados nas contagens de previsões que caem dentro e fora da diagonal desta matriz de confusão. No entanto a classe de interesse é conhecida como classe **positiva**, enquanto que a classe de menor interesse é conhecida como classe **negativa** [37]. Por exemplo, num problema de classificação binário, normalmente a classe de sucesso é classe de maior interesse, pelo que neste caso pode-se designar a classe de sucesso como classe **positiva**, enquanto que a classe de insucesso é a classe **negativa**.

A relação entre as previsões da classe de sucesso e a classe de insucesso, podem ser representadas como a matriz de confusão representada na Tabela 3.10, que tabula se as previsões se enquadram em uma das quatro categorias seguintes [37]:

- **Verdadeiros Positivos (TP)**: A classe de sucesso/positiva é classificada corretamente;
- **Verdadeiros Negativos (TN)**: A classe de insucesso/negativa é classificada corretamente;
- **Falsos Positivos (FP)**: A classe de insucesso/negativa é classificada incorretamente como a classe de sucesso/positiva;
- **Falsos Negativos (FN)**: A classe de sucesso/positiva é classificada incorretamente como a classe de insucesso/negativa.

De referir, que os termos TP, TN, FP e FN referem-se ao número de vezes que as previsões do modelo se enquadravam em cada uma destas categorias.

Nas subsecções 3.9.2,3.9.3, 3.9.4, 3.9.5 e 3.9.6, encontram-se varias métricas para avaliar o desempenho do modelo com base na matriz de confusão (2×2), para classificação binária.

3.9.2 *Sensibilidade* e *Especificidade*

A *sensibilidade* de um modelo, que também é conhecido como a taxa dos verdadeiros positivos, mede a proporção de positivos que se encontram corretamente classificados [37]. Na equação 3.147, tem-se representada a fórmula para calcular a *sensibilidade* de um modelo, sendo que esta é calculada como o número dos verdadeiros positivos a dividir pelo número total de positivos nos dados, ou seja, é o número dos verdadeiros positivos a dividir pela soma dos verdadeiros positivos com os falsos negativos.

$$sensibilidade = \frac{TP}{TP + FN} \tag{3.147}$$

Já a *especificidade* de um modelo, que também é conhecida como a taxa dos verdadeiros negativos, mede a proporção de negativos que se encontram corretamente classificados [37]. Na equação 3.148, tem-se representada a fórmula para o cálculo da *especificidade* de um modelo, sendo que esta é calculada como o número dos verdadeiros negativos a dividir pelo número total de negativos nos dados, ou seja pelo, é o número dos verdadeiros negativos a dividir pela soma dos verdadeiros negativos com os falsos positivos.

$$especificidade = \frac{TN}{TN + FP} \tag{3.148}$$

Os valores da sensibilidade e da especificidade variam entre 0 e 1, sendo que os valores mais desejável de ambos são valores próximos de 1 [37].

3.9.3 *Precisão* e *Recall*

A *precisão*, também conhecida como valor preditivo positivo, é definida como a proporção de positivos que são verdadeiramente positivos, ou seja, a proporção de elementos

que o modelo diz ser positiva e que realmente são positivas [37]. Assim, a *precisão* é calculada como a fração de elementos verdadeiramente positivos dividida pelo número total de unidades que foram classificadas como positivas pelo modelo, ou seja, os elementos verdadeiramente positivos são os elementos que foram classificados pelo modelo como positivos e na realidade também o são, enquanto que os falsos positivos são os elementos que são classificados pelo modelo como positivos e na realidade são negativos. Na equação 3.149, tem-se representada a fórmula de cálculo da *precisão*, podendo-se referir que a *precisão* diz-nos quanto podemos confiar no modelo quando ele prevê um elemento como positivo [37].

$$precisão = \frac{TP}{TP + FP} \tag{3.149}$$

O *recall* mede a precisão preditiva do modelo para a classe positiva [37]. Intuitivamente o *recall* mede a capacidade do modelo encontrar todos os elementos positivos no conjunto de dados, ou seja de todos os elementos positivos dos dados, qual a fração de elementos que estão corretamente classificados como positivos, pelo que se pode afirmar que o *recall* é uma medida de quão completos são os resultados [37]. Na equação 3.150, está a representada a fórmula de cálculo do *recall*, donde se observa que o *recall* é a fração dos elementos verdadeiramente positivos dividida pelo número de total de elementos positivos.

$$recall = \frac{TP}{TP + FN} \tag{3.150}$$

Pode-se observar que a fórmula de cálculo do *recall* em 3.150, é a mesma da *sensibilidade*, no entanto a interpretação do *recall* difere face à interpretação do valor da sensibilidade. Um modelo com um valor elevado de *recall*, diz-nos que a maioria dos elementos que pertencem à classe positiva estão a ser identificados corretamente pelo modelo, pelo que um modelo que apresente um valor elevado de *recall*, significa que tem uma ampla amplitude, pois está a capturar grande parte dos elementos que pertencem à classe positiva [37].

3.9.4 *Accuracy*

A *accuracy* é uma métrica que mede a proporção do número de observações que foram corretamente previstas, isto é, pode-se entender a *accuracy* como a probabilidade da previsão do modelo estar correta [37]. De uma forma simples, pode-se considerar a escolha aleatória de um elemento e prever a classe a que este elemento pertence, sendo que a *accuracy* é a probabilidade de que previsão esteja correta [43].

Assim, a *accuracy* é uma medida de desempenho global que mede o quão o modelo está a prever corretamente o conjunto de dados, ou seja, é a proporção de classificações corretas, tanto de casos positivos como negativos [37]. Deste modo, a fórmula do cálculo da *accuracy* é dada pelo número total de observações previstas corretamente, em que esta quantidade é dada pela soma do número de verdadeiros positivos com o número de

verdadeiros negativos, a dividir pelo número total de previsões. Na equação 3.151, tem-se a representação da fórmula do cálculo da *accuracy*, com base na matriz de confusão.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.151}$$

A *accuracy* tem a vantagem de ser uma métrica intuitiva e muito fácil de entender. No entanto, como se verá na secção 3.10, esta medida pode induzir a uma conclusão errónea quanto ao desempenho de um modelo, quando empregue em bases de dados com classes desequilibradas [44].

A *accuracy* assume valores entre 0 e 1, sendo que valores próximo de 1 indicam que o modelo de um modo global está a classificar as observações nas classes corretas, enquanto que um valor próximo de 0 indica que o modelo não está a classificar corretamente as observações [37]. Como referido anteriormente, esta métrica é uma métrica global, pois com esta métrica não se consegue ter a perceção de como o modelo se comporta individualmente em termos de classificação em cada uma das classes [37]. A quantidade que falta para chegar a 1 é chamada de taxa de erro de classificação, ou também de taxa de classificação incorreta, (*misclassification rate*), e esta é dada pela fórmula 3.152 [37].

$$taxa\ de\ erro\ de\ classificação = 1 - accuracy = \frac{FP + FN}{TP + TN + FP + FN} \tag{3.152}$$

3.9.5 Estatística *Kappa*

A estatística *kappa*, ajusta a *accuracy* considerado a possibilidade de uma previsão estar correta apenas por acaso [37]. Os valores de *kappa* variam entre 0 e 1, onde o valor 1 indica uma concordância perfeita entre as previsões do modelo e os verdadeiros valores, o que é uma ocorrência rara. Valores menores indicam uma concordância imperfeita [37].

Na equação 3.153, encontra-se representada a fórmula para calcular a estatística *kappa* [37]. Nessa fórmula Pr refere-se à proporção de concordância real (a) e esperada (e) entre o classificador e os verdadeiros valores.

$$kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{3.153}$$

Neste caso, o $Pr(a)$ é simplesmente a soma da proporção do número de observações que no total foram classificadas corretamente, ou seja, é a soma entre o quociente do número de verdadeiros positivos e o número total de observações classificadas e o quociente entre o número de verdadeiros negativos e o número total de observações classificadas, ou seja, $Pr(a)$ é dado pela fórmula 3.154 [37].

$$Pr(a) = \frac{TP}{TN + TP + FP + FN} + \frac{TN}{TN + TP + FP + FN} = \frac{TP + TN}{TN + TP + FP + FN} \tag{3.154}$$

Num problema de classificação binário $Pr(e) = Pr(e_1) + Pr(e_2)$, onde $Pr(e_1)$ é probabilidade de que as previsões concordem aleatoriamente com os valores verdadeiros da classe

positiva e $Pr(e_2)$ é a probabilidade de que as previsões concordem aleatoriamente com os valores verdadeiros da classe negativa [37]. A suposição é que os dois classificadores (previsão do modelo e valor verdadeiro da classe) são independentes [37]. Nesse caso, as probabilidades $Pr(e_1)$ e $Pr(e_2)$ são calculadas multiplicando-se a parcela das classes real e a parcela da classe prevista. Assim $Pr(e_1)$ é dado pela fórmula 3.155, enquanto que o $Pr(e_2)$ é dado pela fórmula 3.156 [37].

$$Pr(e_1) = \frac{FP + TP}{TN + TP + FP + FN} \times \frac{FN + TP}{TN + TP + FP + FN} \tag{3.155}$$

$$Pr(e_2) = \frac{TN + FP}{TN + TP + FP + FN} \times \frac{TN + FN}{TN + TP + FP + FN} \tag{3.156}$$

Uma interpretação comum dos valores de *kappa* é a seguinte [37]:

- Má concordância - valores abaixo de 0.2;
- Concordância medíocre - valores entre 0.2 a 0.4;
- Concordância moderada - valores entre 0.4 a 0.6;
- Boa Concordância - valores entre 0.6 a 0.8;
- Concordância muito boa - valores entre 0.8 a 1.0.

3.9.6 Medida-*F* (*F-measure*)

A métrica medida-*F* é uma medida de desempenho do modelo que combina a *precisão* e o *recall* num único valor [37]. A métrica medida-*F*, também é conhecida como *F-score* ou ainda como *F1* [45].

A medida-*F* combina a *precisão* e o *recall* usando a média harmónica. A média harmónica é usada em vez da média aritmética, uma vez que a *precisão* e o *recall* são expressos como proporções entre 0 e 1 [37]. A fórmula de cálculo para a medida-*F*, encontra-se representada em 3.157 [45].

$$\text{medida-}F = \frac{2 \times precisão \times recall}{recall + precisão} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.157}$$

O valor da medida-*F* varia no intervalo entre]0, 1[, em que valores elevados da medida-*F*, ou seja valores próximos de 1, indicam um alto desempenho de classificação, enquanto valores baixos da medida-*F*, isto é, valores próximos de 0 demonstram um desempenho mau de classificação [45].

3.9.7 Biblioteca *Caret*

A biblioteca *caret* do *software* R, tem uma função que nos dá a possibilidade de cálculo de todas as métricas anteriormente descritas. A função da biblioteca *caret* que nos dá as métricas é a função *confusionMatrix()*, em que esta função tem como argumentos de

entrada as classes observadas das observações dos dados de teste, assim como também um vetor contendo as classes previstas pelo modelo ajustado com os dados de treino, quando se aplica neste as observações aos dados de teste. Ou seja, os argumento *data* da função *confusionMatrix()*, refere-se às classes previstas do modelo ajustado aos dados de treino quando aplicado às observações dos dados de teste e o argumento *reference* é as classes efetivas das observações dos dados de teste. O argumento *mode*="everything", refere-se ao facto de que todas as medidas anteriormente descritas são dadas no mesmo output. Para mais informação consultar [46].

3.9.8 Curva ROC

A curva ROC (*Receiver Operating Characteristic*) é um gráfico bidimensional no qual a taxa dos verdadeiros positivos representa-se no eixo dos yy e a taxa dos falsos positivos representa-se no eixo dos xx [37]. Estes valores são equivalentes à *sensibilidade* e (1 − *especificidade*), respetivamente. O gráfico também é conhecido como o gráfico de *sensibilidade*/*especificidade* [37].

A curva ROC é usada para fazer um equilíbrio entre os benefícios, ou seja, os verdadeiros positivos, e os custos, ou seja, os falsos positivos, ou seja dito de um outro modo, a curva ROC é usada para examinar a compensação entre a deteção de verdadeiros positivos, evitando os falsos positivos [37].

Na Figura 3.6, encontra-se uma representação de uma curva ROC.

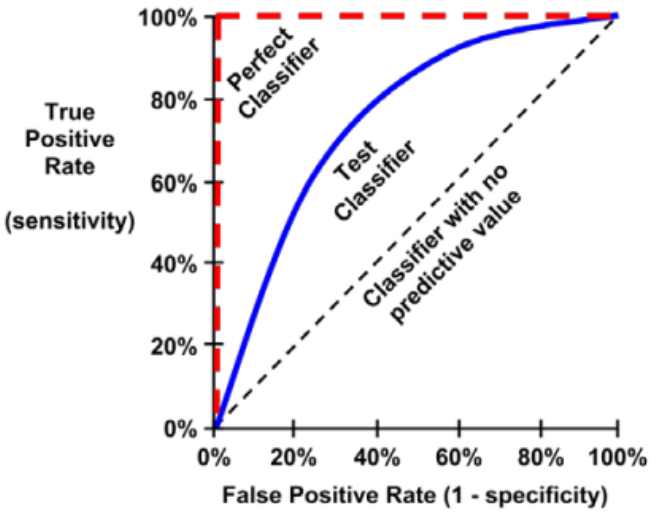


Figura 3.6: Curva ROC [37].

Os pontos que compõem a curva ROC indicam a taxa de verdadeiros positivos em vários limites de falsos positivos. Para se construir uma curva ROC, as previsões de um classificador são classificadas pela probabilidade estimada do modelo da classe positiva, em que estas probabilidades são ordenadas do valor mais elevado para o mais baixo [37]. Começando na origem, o impacto de cada previsão na taxa de verdadeiros positivos e

na taxa de falsos positivos resultará numa curva traçada verticalmente (para previsão correta) ou horizontalmente (para um previsão incorreta) [37].

Na Tabela 3.11, encontra-se o algoritmo para a construção da curva ROC [45].

Tabela 3.11: Algoritmo construção curva ROC [45].

Algoritmo construção Curva ROC

1. Seja ($S_{test} = \{y_1, \ldots, y_N\}$) um conjunto de observações para teste, onde N é o número total de observações para teste; Seja $f(i)$ um classificador que classifica a i-ésima observação para a classe positiva ou negativa, P e N' representam o número total de observações que têm classificação positiva e negativa, respetivamente, no conjunto de teste;

2. Considere-se uma matriz A $(N \times 3)$ onde a *i-ésima* linha da primeira coluna refere-se à classe da *i-ésima* observação dos dados de teste. Na segunda coluna a *i-ésima* linha refere-se à probabilidade dada pelo classificador f à *i-ésima* observação nos dados de teste. A terceira coluna a *i-ésima* linha indica a classe prevista pelo classificador f na *i-ésima* observação dos dados de teste, ou seja, num problema binário se $f(i) > 0.5$ a observação é classificada como positiva e é classificada como negativa caso contrário.

3. Seja S_{ord} a matriz A ordenada pelas linhas, consoante as probabilidades estimadas pelo modelo, da maior para a menor. Seja $S1_{ord}$, $S2_{ord}$ e $S3_{ord}$, cada uma das 3 colunas da matriz S_{ord}, ou seja, $S1_{ord}$ é a primeira coluna, $S2_{ord}$ a segunda coluna e $S3_{ord}$ a terceira coluna;

4. Acrescente uma linha inicial na matriz S_{ord} onde na primeira linha da segunda coluna $S2_{ord}$ se coloca o valor ∞ e nas restantes duas colunas a primeira linha é vazia;

5. $FP \leftarrow 0$, $TP \leftarrow 0$, $f_{prev} \leftarrow -\infty$ e $ROC = []$;

6. Seja $|S_{ord}|$ o número de linhas da matriz S_{ord}. Para $i = 1$ a $|S_{ord}|$ fazer:

   a) Se $S3_{ord}(i) \neq f_{prev}$ então $ROC(i) = \left(\frac{FP}{N'}, \frac{TP}{P}\right)$ e $f_{prev} = S3_{ord}(i+1)$;

   b) Se $S1_{ord}(i)$ é uma observação positiva então $TP = TP + 1$ e $ROC(i) = \left(\frac{FP}{N'}, \frac{TP}{P}\right)$;

   c) Se $S1_{ord}(i)$ é uma observação negativa então $FP = FP + 1$ e $ROC(i) = \left(\frac{FP}{N'}, \frac{TP}{P}\right)$;

7. Por fim unem-se todos os pontos resultantes anteriormente através de uma curva.

Através da observação da curva ROC pode-se ter uma noção do tipo de classificador. Para ilustrar esse conceito, três classificadores hipotéticos são contrastados no gráfico da

Figura 3.6. Primeiro, a linha diagonal do canto inferior esquerdo ao canto superior direito do diagrama representa um classificador sem valor preditivo. Esse tipo de classificador deteta verdadeiros positivos e falsos positivos exatamente na mesma taxa, o que implica que o classificador não pode discriminar entre os dois. Esta é a linha de base pela qual outros classificadores podem ser julgados. As curvas ROC próximas a esta linha indicam modelos que não são muito úteis. Da mesma forma, o classificador perfeito tem uma curva que passa pelo ponto com 100% de taxa de verdadeiros positivos e 0% de taxa de falsos positivos. Ele é capaz de identificar corretamente todos os verdadeiros positivos antes de classificar incorretamente qualquer resultado negativo. A maioria dos classificadores do mundo real é semelhante ao classificador de teste, eles caem em algum lugar na zona entre perfeito e inútil.

3.9.8.1 Área sob a Curva ROC - AUC

Uma das desvantagens das curvas ROC é que na comparação de diferentes classificadores através da curva ROC pode-se ter uma grande dificuldade em fazê-lo e isto deve-se ao facto que que não existe nenhum escalar que represente o desempenho esperado. Assim, a área sob curva ROC (AUC) é uma métrica que é utilizada para calcular a área sob a curva ROC. Os valores da área sob a curva ROC estão limitados entre 0 e 1 mas, na realidade, não existe nenhum classificador com a área sob a curva ROC menor que 0.5 [45]. Do exposto, resulta que se uma curva ROC, estiver próxima do classificador perfeito, então a área sob esta curva ROC é próxima de 1, enquanto que se uma curva ROC, estiver próxima do classificador sem valor preditivo, a área sob a curva ROC é próxima dos 0.5. Uma interpretação dos valores da área sob a curva ROC de um classificador são as seguintes [37]:

- AUC com valores entre 0.9 e 1.0, resultam num excecional classificador;
- AUC com valores entre 0.8 e 0.9, resultam num bom/excelente classificador;
- AUC com valores entre 0.7 e 0.8, resultam num razoável classificador;
- AUC com valores entre 0.6 e 0.7, resultam num classificador pobre;
- AUC com valores entre 0.5 e 0.6, resultam num classificador não discriminante.

Na Figura 3.7, está representado um exemplo do valor da área sob a curva ROC dada por dois classificadores, o classificador A e o classificador B. Através da sua análise, pode-se constatar que a área sob a curva ROC do classificador B é maior do que a área sob a curva ROC do classificador A, pelo que se pode concluir que o classificador B alcança um melhor desempenho quando comparado com o classificador A.

A área sombreada a cinza é comum a ambos os classificadores, enquanto que a área sombreada a rosa representa a área em que o classificador B supera o classificador A, e a área a azul representa a área em que o classificador A supera o classificador B.

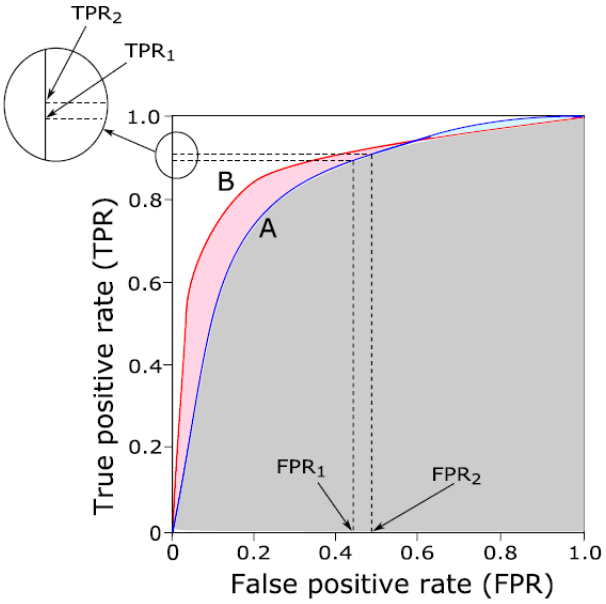


Figura 3.7: Exemplo da métrica área sob a curva ROC (AUC) [45].

Assim, é possível que um classificador com a área sob a curva ROC mais baixo, supere um classificador com a área sob a curva ROC mais elevado em determinadas regiões especificas, pois vê-se neste exemplo que o classificador B supera o classificador A, exceto em FPR> 0.6.

Para além disso é possível ocorrer o caso em que dois classificadores com duas curvas ROC diferentes apresentem o mesmo valor da área sob as suas curvas ROC. Assim, por esta razão, o valor da área sob a curva ROC pode ser enganosa, pelo que na prática a área sob a curva ROC deve ser sempre utilizada em combinação com a representação da curva ROC [37].

3.9.8.2 Biblioteca *pROC*

A biblioteca *pROC* do *software* R, tem uma função que constrói a curva ROC de um dado classificador. A função *roc()* desta biblioteca utiliza como elementos de entrada um vetor das verdadeiras classes, em que este argumento é o "response", enquanto que o argumento que recebe um vetor com as probabilidades preditas é o "data".

Para se representar uma curva ROC pode-se utilizar a função *plot.roc()*, desta biblioteca, em que esta função recebe como argumente um objeto da função *roc()*. Por outro lado, esta função também tem como argumento o "print.auc", onde este argumento refere-se ao facto do gráfico que o *software* R faz dá a área sob a curva ROC.

No entanto, para mais detalhes pode-se consultar "Package pROC"[47].

3.10 Problemas da Classificação com Classes Desequilibradas

Uma das principais barreiras dos algoritmos de Aprendizagem Automática (*Machine Learning*) é o problema do desequilíbrio/desbalanceamento de dados entre classes [4]. Este problema ocorre quando uma variável resposta categórica tem mais observações numa dada classe do que nas restantes classes, isto é, existe uma grande diferença entre o número de observações em cada uma das classes. Num problema de classificação em que a variável resposta é binária, tem-se um problema de desequilibro na base de dados se existir uma grande diferença entre a proporção do número de dados em cada uma das duas classes. A classe que tiver o maior número de observações iremos a designar por classe **maioritária** e a classe com o menor número de observações é a classe **minoritária**.

Algumas das métricas vistas anteriormente na secção 3.9, tais como a *accuracy* e a taxa de erro costumam dar-nos conclusões erróneas sobre a avaliação do modelo no geral, pois quando se tem um problema de classificação, normalmente o classificador tende a classificar mais vezes para a classe maioritária do que para a classe minoritária. Como visto anteriormente, as métricas *accuracy* e a taxa de erro levam os verdadeiros positivos e os verdadeiros negativos em conta, mas uma vez que a classe maioritária tem um maior número de dados então a *accuracy* pode ser enviesada para a classe maioritária, independentemente da classe minoritária com o número de observações mais baixo, o que leva a um desempenho péssimo destas métricas [44].

No entanto, numa base de dados desequilibradas em vez de se considerar a métrica *accuracy* para avaliar o desempenho do modelo no global, pode-se utilizar a *accuracy balanceada*, dada por 3.158, que tenta combater as desvantagens do problema da métrica *accuracy* em base de dados desequilibrados.

$$accuracy\ balanceada = \frac{especificidade + sensibilidade}{2} \tag{3.158}$$

Uma outra métrica também muitas vezes utilizada em problemas em que a base de dados é desequilibrada é o média geométrica (G-mean), onde esta métrica agrega a *sensibilidade* e a *especificidade* de acordo com 3.159, pois o principal objetivo de um classificador é melhorar a *sensibilidade* sem sacrificar a *especificidade*. No entanto, os objetivos de *sensibilidade* e *especificidade* são muitas vezes conflituosos, o que pode não funcionar bem, especialmente quando o conjunto de dados está desequilibrado/desbalanceado 3.159 [45].

$$GM = \sqrt{especificidade \times sensibilidade} \tag{3.159}$$

No entanto, para além destas métricas também existem métodos de amostragem que permitem balancear/equilibrar uma base de dados. Na subsecção 3.10.1, iremos ver alguns métodos de amostragem que permitem balancear bases de dados, pois por exemplo alguns dos modelos de aprendizagem automática, como as Árvores de Classificação costumam ter o pressuposto de que as classes da variável resposta é equilibrada, pelo que uma das soluções quando isto não ocorre é balancear a base de dados [48].

3.10.1 Métodos de amostragem

Os métodos de amostragem são um pré-processamento de dados que lidam com o problema do desequilibro entre as classes da variável resposta, construindo um conjunto de dados de equilibrados na variável resposta [49]. Os métodos de amostragem incluem a subamostragem (*undersampling*) e a sobreamostragem (*oversampling*) [49].

3.10.1.1 Sobreamostragem

O método de sobreamostragem é um método que aumenta o número de dados na classe minoritária, até o número de observações da classe minoritária atingir o mesmo número ou a mesma proporção de observações da classe maioritária, mantendo intactas as observações da classe maioritária.

Duas abordagens principais de sobreamostragem são a sobreamostragem aleatória e a técnica SMOTE (*Synthetic Minority Oversampling Technique*) [5]. A sobreamostragem aleatória balanceia os dados através da replicação aleatória de observações na classe minoritária, isto é, de todas as observações que pertencem à classe minoritária este método escolhe aleatoriamente um determinado número dessas observações e replica-as no conjunto de dados. O número de observações escolhidas aleatoriamente, corresponde à diferença entre o número de observações que pertencem à classe maioritária e o número de observações que pertencem à classe minoritária.

A técnica SMOTE é uma técnica de amostragem, na qual a classe minoritária é sobre amostrada, criando assim exemplos sintéticos em vez de sobre amostra com alternância. A classe minoritária é sobre amostrada combinando cada amostra de classe minoritária e introduzindo exemplos sintéticos ao longo dos segmentos de reta que unem todos os k vizinhos adjacentes da classe minoritária. Dependendo da quantidade de sobreamostragem necessária, os vizinhos dos k vizinhos mais próximos são escolhidos aleatoriamente [50].

O método mais simples para aumentar o tamanho da classe minoritária corresponde à sobreamostragem aleatória, ou seja, um método não heurístico que equilibra a distribuição de classes por meio da replicação aleatória de exemplos positivos. No entanto, como esse método replica exemplos existentes na classe minoritária, o sobreajustamento (*overfitting*) é mais provável de ocorrer.

3.10.1.2 Subamostragem

O método da subamostragem é um método muito eficiente quando se tem o problema de classes desequilibradas [49]. A subamostragem equilibra os dados através da remoção observações na classe maioritária [44]. Ou seja, este método equilibra as classes de uma variável resposta, em que esta tem um desequilibro entre as classes, removendo algumas observações da classe maioritária. Um dos métodos de subamostragem mais utilizado é a subamostragem aleatória [49].

A subamostragem aleatória é um método de subamostragem que exclui aleatoriamente observações que pertencem à classe maioritária de forma a balancear os dados. Dada a sua simplicidade, este método costuma ter um desempenho muito bom, quando aplicado a um conjunto de dados desequilibrado [51]. Assim, a subamostragem aleatória remove aleatoriamente observações da classe maioritária, até que a proporção de observações na classe maioritária seja idêntica à proporção de observações na classe minoritária, ou seja, a subamostragem aleatória é um método inverso ao de sobreamostragem aleatória.

No entanto, esta técnica como as outras técnicas referidas anteriormente, também tem as suas vantagens e limitações.

Uma das principais desvantagens deste método é que a subamostragem pode levar à perda de informação útil ao remover padrões significativos nos dados, uma vez que grande parte dos dados da classe maioritária é descartada [44].

4

Resultados

Ao abrigo do disposto no artigo 39.º do Decreto n.º 2-B/2020, de 2 de abril, a Direção-Geral da Saúde disponibilizou à Comunidade Científica e Tecnológica Portuguesa o acesso a dados de saúde pública do SINAVE relativos a doentes infetados pelo novo coronavírus SARS-CoV-2.

O presente estudo enquadra-se neste âmbito e alinha-se com o compromisso de análise destes dados em contexto de modelação estatística, para estimação e previsão da morbilidade e mortalidade por COVID-19 na população Portuguesa, assim como a determinação de fatores de risco inerentes.

Os dados disponibilizados referem-se a todos os casos de COVID-19 confirmados laboratorialmente e notificados, por data, em Portugal durante o primeiro ano e pouco de pandemia (março de 2020 a julho de 2021), bem como alguma informação clínica associada.

Trata-se de uma base de dados de grandes dimensões, com alguns problemas inerentes que serão aqui destacados, à qual iremos aplicar os métodos descritos anteriormente de forma a conseguir responder aos objetivos propostos, e cujos resultados se descrevem neste capítulo. Contudo, previamente, faz-se uma análise exploratória do conjunto de dados.

Na secção 4.1, elabora-se uma análise preliminar à base de dados disponibilizada pelo SINAVE, com casos de COVID-19 confirmados laboratorialmente e notificados, por data, em Portugal, durante o período de março de 2020 a julho de 2021. Nesta secção, elabora-se uma análise estatística principalmente aos sintomas e às comorbilidades. Na secção 4.2, faz-se uma limpeza à base de dados, devido ao elevado número de valores omissos, como se virá na análise da base de dados. Na secção 4.3, treina-se os diferentes modelos, assim, como se apresentam os resultados da estimação dos modelos de Regressão Logística, Regressão Logística com interações e o modelo Aditivo Generalizado. Por fim na secção 4.4, apresentam-se os resultados preditivos de cada um destes modelos.

4.1 Análise Preliminar dos Dados

A presente base de dados é constituída por 52 variáveis, que se podem dividir em 5 grupos:

1. Dados pessoais, incluindo um identificador único de cada pessoa e os seus dados demográficos e geográficos;
2. Datas: data de confirmação (data em que o sistema tem conhecimento do caso); data do início dos sintomas (data em que o utente teve início dos sintomas); data de óbito (data em que o utente morre);
3. Existência de comorbilidades e sua apresentação;
4. Existência de sintomas e sua apresentação;
5. Tipo Vírus e tipo de teste.

Nas Tabelas, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 e 4.7, apresentam-se uma pequena descrição das variáveis referente ao utente, aos dados geográficos, às datas, às comorbilidades, aos sintomas sobre o utente e também sobre a variante do vírus, assim como o tipo de teste empregue, respetivamente.

Tabela 4.1: Descrição variáveis referentes aos utentes (CG="Categórica"e CT="Contínua").

Nome da Variável	Descrição	Tipo	Valores
id_pessoa	Identificação do indivíduo		
sexo_utente	Sexo do indivíduo	CG	"M"=Masculino "F"=Feminino
idade_utente_a_data _validacao	Idade do indivíduo	CT	anos
morte	Indica o caso de desfecho do caso	CG	"0"= Sobrevive "1"=Faleceu
pais_nacionalidade _utente	Indica o país da nacionalidade do indivíduo	CG	Nome do país
regra_de_confirmacao	Indica a notificação do caso se foi por notificação médica (NM), notificação laboratorial (NL) ou por caso.	CG	NL NM Caso
durante_o_periodo _de_incubacao_esteve _fora_de_portugal	Indica se o indivíduo esteve ou não fora de Portugal	CG	"0"=Não "1"=Sim
qual_o_pais	Indica o nome do país que o individuo esteve, quando esteve fora de Portugal	CG	Nome do país

Tabela 4.2: Descrição variáveis referentes aos dados demográficos (CG="Categórica").

Nome da Variável	Descrição	Tipo	Valores
codigo_concelho_morada_utente	Código do concelho da morada do indivíduo	CG	308 códigos dos concelhos
descricao_concelho_morada_utente	Nome do concelho da morada do indivíduo	CG	308 nomes dos concelhos
codigo_concelho_ocorrencia	Código do concelho da ocorrência do caso	CG	308 códigos dos concelhos
descricao_concelho_ocorrencia	Nome do concelho da ocorrência do caso	CG	308 nomes dos concelhos

Tabela 4.3: Descrição variáveis referentes às datas.

Nome da Variável	Descrição	Tipo	Valores
data_confirmado	Data da confirmação do caso	Data	dia-mês-ano
data_inicio_sintomas	Data do início dos sintomas	Data	dia-mês-ano
data_obito	Data de óbito do indivíduo	Data	dia-mês-ano

Tabela 4.4: Descrição variáveis referentes às comorbilidades (CG="Categórica").

Nome da Variável	Descrição	Tipo	Valores
doenca_neurologica_ou _neuromuscular_cronica	Indica se o indivíduo apresenta ou não a comorbilidade doença neurológica ou neuromuscular crónica	CG	"0"=Não "1"=Sim
neoplasia	Indica se o indivíduo apresenta ou não a comorbilidade neoplasia	CG	"0"=Não "1"=Sim
vih_outras_ imunodeficiencias	Indica se o indivíduo apresenta ou não a comorbilidade VIH ou outras imunodeficiência	CG	"0"=Não "1"=Sim
diabetes	Indica se o indivíduo apresenta ou não a comorbilidade diabetes	CG	"0"=Não "1"=Sim
doenca_pulmonar_cronica	Indica se o indivíduo apresenta ou não a comorbilidade doença pulmonar crónica	CG	"0"=Não "1"=Sim
asma	Indica se o indivíduo apresenta ou não a comorbilidade asma	CG	"0"=Não "1"=Sim
doenca_hematologicas_ cronicas	Indica se o indivíduo apresenta ou não a comorbilidade doença hematológica crónica	CG	"0"=Não "1"=Sim
patologia_hepatica	Indica se o indivíduo apresenta ou não a comorbilidade patologia hepática	CG	"0"=Não "1"=Sim
deficiencia_neurologica_ cronicas	Indica se o indivíduo apresenta ou não a comorbilidade deficiência neurológica crónica	CG	"0"=Não "1"=Sim
doenca_renal_cronica	Indica se o indivíduo apresenta ou não a comorbilidade doença renal crónica	CG	"0"=Não "1"=Sim
insuficiencia_renal_aguda	Indica se o indivíduo apresenta ou não a comorbilidade insuficiência renal aguda	CG	"0"=Não "1"=Sim
insuficiencia_cardiaca	Indica se o indivíduo apresenta ou não a comorbilidade insuficiência cardíaca	CG	"0"=Não "1"=Sim
coagulopatia_de_consumo	Indica se o indivíduo apresenta ou não a comorbilidade coagulopatia de consumo	CG	"0"=Não "1"=Sim
o_doente_apresenta_ comorbilidades	Indica se o indivíduo apresenta ou não comorbilidades	CG	"0"=Não "1"=Sim

Tabela 4.5: Descrição variáveis referentes aos sintomas (CG="Categórica").

Nome da Variável	Descrição	Tipo	Valores
historia_de_febre_ou_ calafrios	Indica se o indivíduo apresenta ou não o sintoma febre ou calafrios	CG	"0"=Não "1"=Sim
dispneia	Indica se o indivíduo apresenta ou não o sintoma dispneia	CG	"0"=Não "1"=Sim
tosse_seca_ou_produtiva	Indica se o indivíduo apresenta ou não o sintoma tosse	CG	"0"=Não "1"=Sim
coriza	Indica se o indivíduo apresenta ou não o sintoma coriza	CG	"0"=Não "1"=Sim
odinofagia	Indica se o indivíduo apresenta ou não o sintoma odinofagia	CG	"0"=Não "1"=Sim
cefaleia	Indica se o indivíduo apresenta ou não o sintoma cefaleia	CG	"0"=Não "1"=Sim
dor_abdominal	Indica se o indivíduo apresenta ou não o sintoma dor abdominal	CG	"0"=Não "1"=Sim
artralgia	Indica se o indivíduo apresenta ou não o sintoma artralgia	CG	"0"=Não "1"=Sim
dor_no_peito	Indica se o indivíduo apresenta ou não o sintoma dor no peito	CG	"0"=Não "1"=Sim
mialgias	Indica se o indivíduo apresenta ou não o sintoma mialgias	CG	"0"=Não "1"=Sim
nauseas_vomitos	Indica se o indivíduo apresenta ou não o sintoma náuseas ou vómitos	CG	"0"=Não "1"=Sim
diarreia	Indica se o indivíduo apresenta ou não o sintoma diarreia	CG	"0"=Não "1"=Sim
irritabilidade_confusao	Indica se o indivíduo apresenta ou não o sintoma irritabilidade confusão	CG	"0"=Não "1"=Sim
taquicardia	Indica se o indivíduo apresenta ou não o sintoma taquicardia	CG	"0"=Não "1"=Sim

Tabela 4.6: Continuação descrição variáveis referentes aos sintomas (CG="Categórica").

Nome da Variável	Descrição	Tipo	Valores
fraqueza_geral_ou_ astneia	Indica se o indivíduo apresenta ou não o sintoma astneia ou fraqueza geral	CG	"0"=Não "1"=Sim
coma	Indica se o indivíduo apresenta ou não o sintoma coma	CG	"0"=Não "1"=Sim
auscultacao_pulmonar_ anomala	Indica se o indivíduo apresenta ou não o sintoma auscultação pulmonar anómala	CG	"0"=Não "1"=Sim
convulsoes	Indica se o indivíduo apresenta ou não o sintoma convulsões	CG	"0"=Não "1"=Sim
radiografia_pulmonar_ com_alteracoes	Indica se o indivíduo apresenta ou não o sintoma radiografia pulmonar com alterações	CG	"0"=Não "1"=Sim
pneumonia	Indica se o indivíduo apresenta ou não o sintoma pneumonia	CG	"0"=Não "1"=Sim
apresentacao_da_doenca	Indica se o indivíduo é sintomático ou assintomático	CG	"Sintomático" "Assintomático"

Tabela 4.7: Descrição variáveis referentes ao tipo de vírus e teste (CG="Categórica").

Nome da Variável	Descrição	Tipo	Valores
virus_variant	Indica a variante do vírus	CG	Nome da variante do vírus
virus_variant_other	Indica outras variantes do vírus não mencionadas na variável *virus_variant*	CG	Nome da variante do vírus
analise_1	Indica o tipo de teste efetuado se foi ANTIGEN ou PCR	CG	ANTIGEN PCR

Esta base de dados inclui todos os casos de COVID-19 notificados laboratorialmente, desde o início da pandemia em Portugal, início de março de 2020 a meados de julho de 2021. No entanto, esta não contém informação sobre o desfecho do caso, como por exemplo, cura, hospitalização simples ou hospitalização em cuidados intensivos. Contudo, esta base de dados tem uma variável associada à data de óbito, então pode-se criar uma variável que se refere ao desfecho derradeiro, isto é, se o utente acaba por falecer ou não, ou seja, criou-se uma variável binária que toma o valor 1 caso o utente tenha falecido, tendo a data de morte associada, e 0 caso o utente tenha sobrevivido.

Nas Tabelas 4.8, 4.9, 4.10, 4.11, 4.12 e 4.13, apresentam-se as variáveis referente ao utente, os dados geográficos, às datas, aos sintomas, às comorbilidades sobre o utente e também sobre a variante do vírus, assim como o tipo de teste empregue, respetivamente. Apresenta-se em cada tabela a percentagem de valores omissos (NA's) para cada uma das variáveis.

Tabela 4.8: Variáveis referentes aos utentes.

Variável	**% de NA's**	**Variável**	**% de NA's**
id_pessoa	0.00	*sexo_utente*	0.00
idade_utente_a_data_validacao	7.81	*morte*	0.00
pais_nacionalidade_utente	1.33	*regra_de_confirmacao*	0.000
durante_o_periodo_de_incubacao _esteve_fora_de_portugal	74.57	*qual_o_pais*	99.38

Tabela 4.9: Variáveis referentes aos dados geográficos.

Variável	**% de NA's**
codigo_concelho_morada_utente	2.64
descricao_concelho_morada_utente	2.64
codigo_concelho_ocorrencia	40.69
descricao_concelho_ocorrencia	40.69

Tabela 4.10: Variáveis referentes às datas.

Variável	**% de NA's**
data_confirmado	0.01
data_inicio_sintomas	65.76
data_obito	0.00

Através da análise de cada uma das tabelas anteriormente descritas, constatamos que a base de dados tem muitos problemas no que diz respeito ao preenchimento de dados. Assim, foi necessário fazer alguma limpeza na presente base de dados. Contudo, antes da limpeza, faz-se uma primeira análise preliminar dos dados.

Tabela 4.11: Variáveis referentes aos sintomas.

Variável	% de NA's	Variável	% de NA's
historia_de_febre_ou_calafrios	74.85	*dispneia*	81.28
tosse_seca_ou_produtiva	72.54	*coriza*	79.81
odinofagia	80.53	*cefaleia*	77.43
dor_abdominal	83.95	*artralgia*	40.90
dor_no_peito	83.62	*mialgias*	76.98
nauseas_vomitos	40.90	*diarreia*	40.90
irritabilidade_confusao	40.90	*taquicardia*	40.90
fraqueza_geral_ou_astneia	81.74	*coma*	40.90
auscultacao_pulmonar_anomala	40.90	*convulsoes*	40.90
radiografia_pulmonar_com_alteracoes	40.90	*pneumonia*	93.07
apresentacao_da_doenca	57.90		

Tabela 4.12: Variáveis referentes às comorbilidades.

Variável	% de NA's
doenca_neurologica_ou_neuromuscular_cronica	40.90
neoplasia	40.90
vih_outras_imunodeficiencias	40.90
diabetes	40.90
doenca_pulmonar_cronica	40.90
asma	40.90
doenca_hematologicas_cronicas	40.90
patologia_hepatica	40.90
deficiencia_neurologica_cronicas	40.90
doenca_renal_cronica	40.90
insuficiencia_renal_aguda	40.90
insuficiencia_cardiaca	40.90
coagulopatia_de_consumo	40.90
o_doente_apresenta_comorbilidades	75.54

Tabela 4.13: Variáveis referentes ao tipo de vírus e teste.

Variável	% de NA's
virus_variant	99.18
virus_variant_other	99.70
analise_1	7.84

O gráfico da Figura 4.1 mostra o número de novos casos confirmados (chamados apenas de casos), desde o dia 3 de março de 2020 até ao dia 11 de julho de 2021, sendo que 16 de janeiro de 2021 foi o dia em que se atingiu o maior número de casos diários, 16856 novos casos confirmados. Através da análise da Figura 4.1, pode-se observar que se registaram as seguintes 4 vagas/ondas:

- 1ª vaga ocorreu no final de março de 2020 e início de abril de 2020;
- 2ª vaga teve início em outubro de 2020 e durou até ao princípio de dezembro;
- 3ª vaga ocorreu após o Natal de 2020 até ao meados de março de 2021;
- 4ª vaga teve início nos finais de junho de 2021.

Também se pode observar que o referido gráfico tem diferentes oscilações e isto pode ser devido às diferentes medidas adotadas pelo Governo de Portugal, desde o início da pandemia até à data do último caso confirmado, incluído neste estudo.

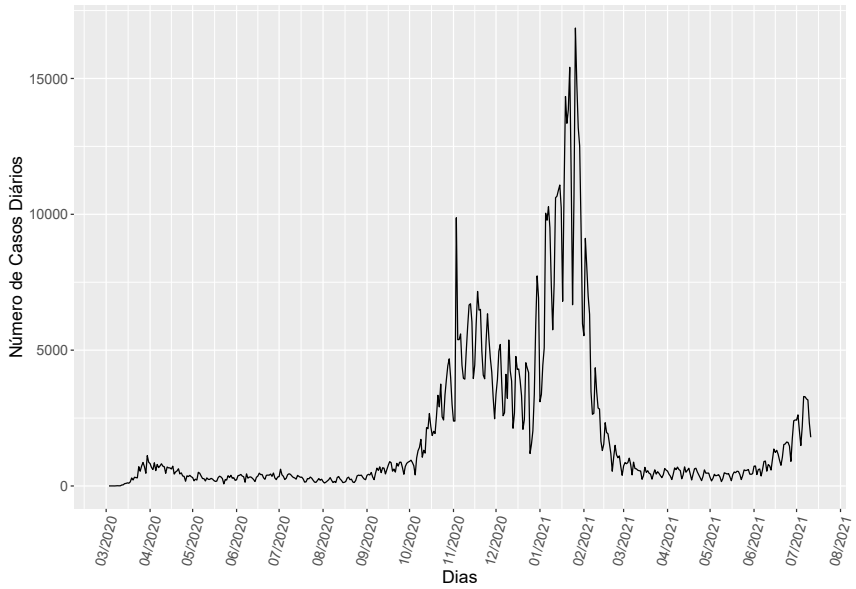


Figura 4.1: Número de casos diários, base de dados original.

No gráfico de barras da Figura 4.2, encontra-se representada a percentagem de casos confirmados por mês da presente base de dados.

Através da análise da Figura 4.2, verifica-se que é durante o período de outono e inverno que há o maior número de casos, sendo que o mês de janeiro do ano de 2021 representa cerca de 33% dos casos totais, seguindo-se o mês de novembro e dezembro de 2020, com cerca de 16% e de 13%, respetivamente. De referir também que a diminuição do número de casos de 33% no mês de janeiro de 2021 para 8% em fevereiro de 2021, bem como outras grandes diferenças devem-se às medidas adotadas pelo Governo de Portugal, de forma a mitigar a propagação do vírus tais como o dever de recolher obrigatório assim como as diversas medidas de restrições que tomou ao longo do tempo.

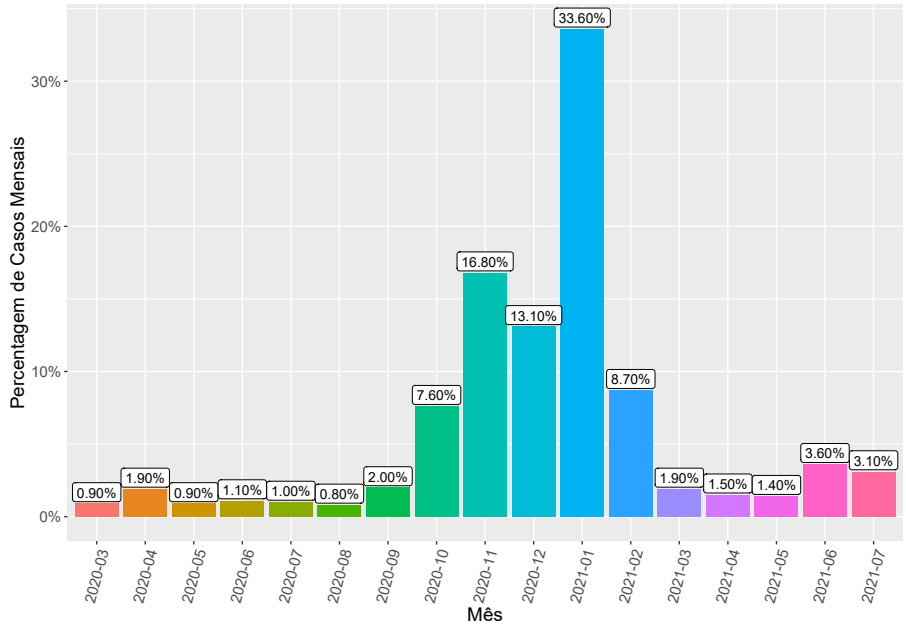


Figura 4.2: Percentagem de casos mensais, base de dados original.

Nas Figuras 4.3 e 4.4, encontra-se o número de casos confirmados, em cada um dos 20 distritos de Portugal por cada 100 mil habitantes, no período de tempo considerado. Através da análise das Figuras 4.3 e 4.4, conclui-se que são os distritos de Lisboa, Setúbal, Braga, Porto e Faro que apresentam maior número de casos por cada 100 mil habitantes, sendo que o distrito de Lisboa com 7697 casos por 100 mil habitantes é o distrito com maior número de casos por 100 mil habitantes. Já os distritos que apresentam o menor número de casos por cada 100 mil habitantes são Leiria, Açores e os dois distritos do interior Guarda e Portalegre. O distrito dos Açores é o que apresenta menor número de casos por 100 mil habitantes, 192 casos por 100 mil habitantes, como se observa na Figura 4.4a. O distrito da Madeira apresenta cerca de 3559 casos por cada 100 mil habitantes, como se observa na Figura 4.4b.

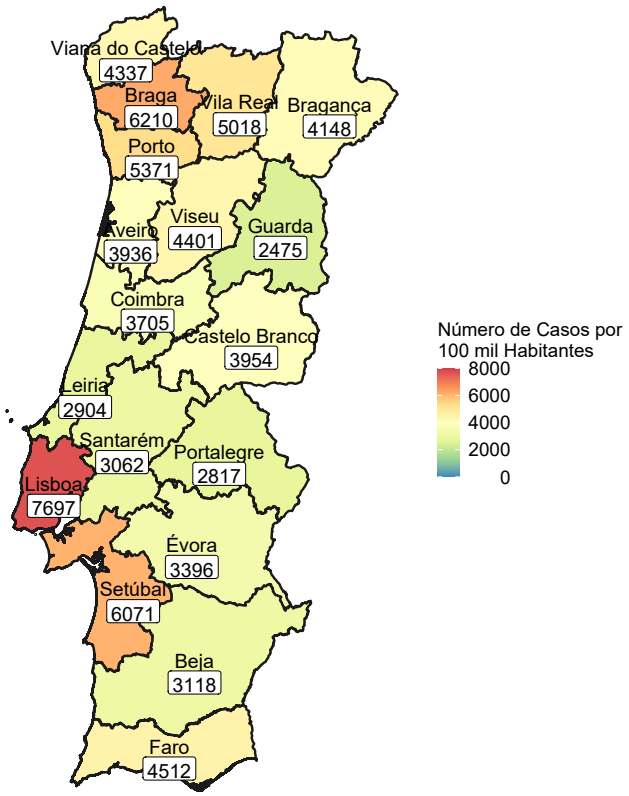
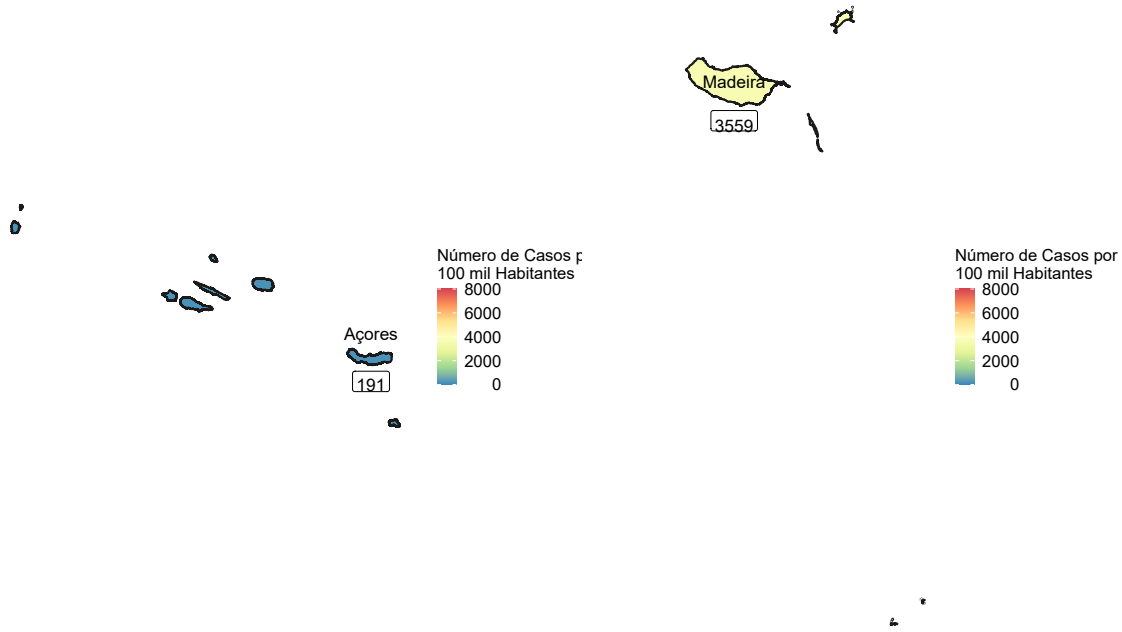


Figura 4.3: Número de casos por 100 mil habitantes por distrito em Portugal Continental.



(a) Região Autónoma dos Açores. (b) Região Autónoma da Madeira.

Figura 4.4: Número de casos por 100 mil habitantes por distrito em Portugal, Regiões Autónomas.

Na Figura 4.5, representa-se a distribuição por sexo dos indivíduos que testaram positivo à COVID-19, verificando-se que a presente base de dados tem uma maior representação do sexo do feminino com uma diferença de aproximadamente 10%.

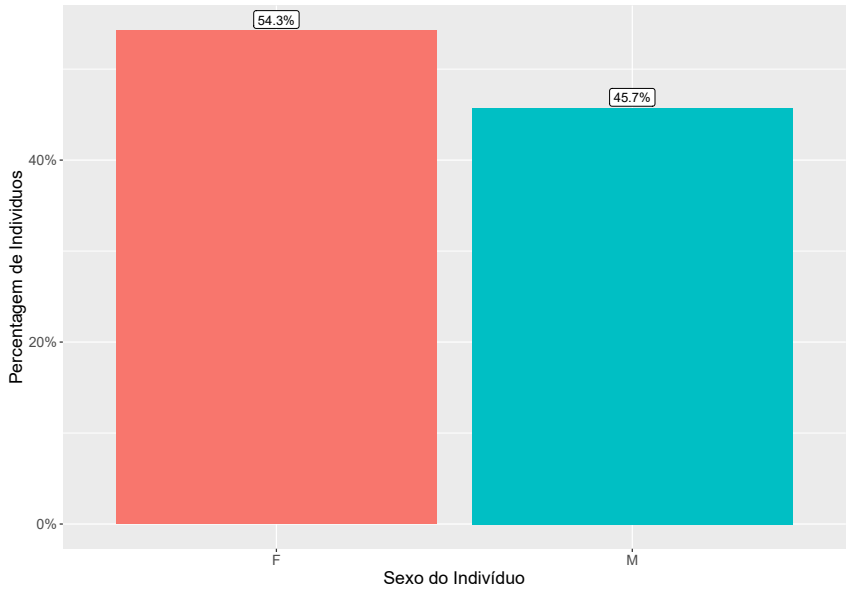


Figura 4.5: Distribuição Percentual de indivíduos do sexo masculino (M) e do sexo feminino (F).

No gráfico da Figura 4.6, representa-se a percentagem de indivíduos que testaram

positivo à COVID-19, por faixa etária. Pode-se concluir que a faixa etária que tem uma maior representação de indivíduos é a classe dos [40 – 49] anos, sendo que esta representa cerca de 17% dos casos de COVID-19. As classes etárias dos [20 – 29] anos, dos [30 – 39] anos, e dos [50 – 59] anos, têm praticamente a mesma representação, cerca de 15% dos indivíduos. Já as classes etárias dos [70 – 79] anos, dos ≥ 80 anos e dos [0 – 9] anos, são as que têm um menor número de casos, sendo que cada uma destas classes representa, respetivamente cerca de 6%, 6% e 7% dos casos totais de COVID-19. As classes etárias dos [10 – 19] anos e dos [60 – 69] anos, também têm uma distribuição muito semelhante, sendo que cada uma destas representa cerca de 10% dos casos totais de COVID-19.

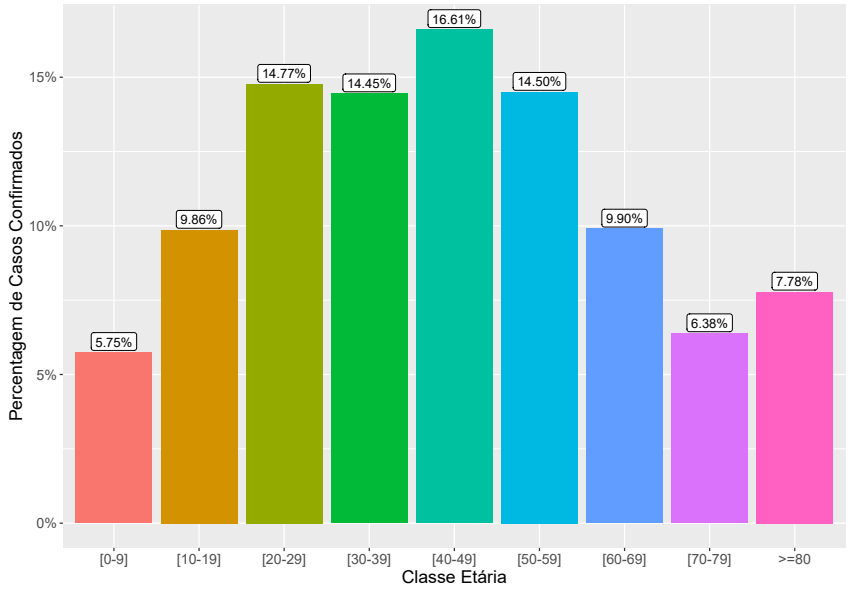


Figura 4.6: Percentagem de casos por classe etária.

O gráfico da Figura 4.7, retrata o número de mortes diárias associada à COVID-19. Pode-se observar que no início há um aumento significativo de mortes, sendo que estas depois tendem a diminuir e a estabilizar, sendo que a partir de meados de outubro há novamente um aumento significativo de mortes por COVID-19, atingindo-se o máximo de mortes no dia 30 de janeiro de 2021. A partir deste dia o número de mortes por COVID-19 começa a diminuir, estabilizando.

O gráfico da Figura 4.8, mostra a percentagem de mortes associadas à COVID-19 por cada mês, sendo que se pode concluir que são nos meses de novembro a fevereiro que houve o maior número de mortes, sendo que também são nestes meses que se tem o maior número de casos mensais, como mostra o gráfico da Figura 4.2.

A Figura 4.9, mostra a distribuição das mortes entre os indivíduos que testaram positivo à COVID-19, sendo que a maioria não morreu, apenas cerca de 2% dos indivíduos acabaram por falecer, sendo que os restantes 98% sobreviveram.

Na Figura 4.10, representa-se por sexo a percentagem de indivíduos que faleceram e que sobreviveram, não existindo grande diferença entre géneros, em ambos os casos

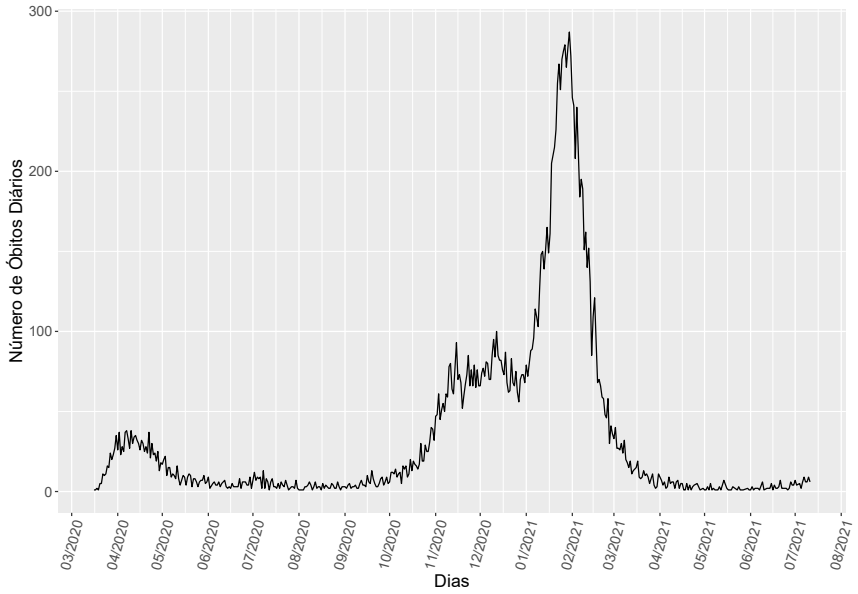


Figura 4.7: Número de óbitos diários, base de dados original.

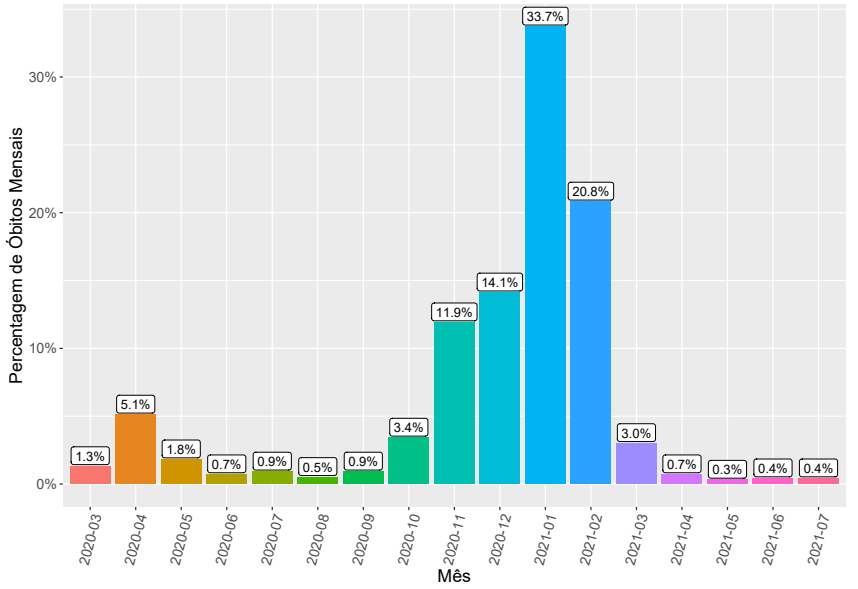


Figura 4.8: Percentagem de mortes mensais, base de dados original.

tem-se que cerca de 2% dos indivíduos acabaram por falecer.

Na Figura 4.11, tem-se uma representação da distribuição etária dos indivíduos que faleceram e os que sobreviveram. Daqui, pode-se deduzir que de todos os indivíduos que testaram positivo à COVID-19, 16.6% sobreviveram e pertencem à classe etária dos $[40-49]$ anos, 14.5% sobreviveram e encontram-se respetivamente nas classes etárias dos $[20-29]$ anos, $[30-39]$ anos e $[50-59]$ anos, 6% sobreviveram e encontram-se na classe etária dos $[70-79]$ anos e cerca de 7% sobreviveram e pertencem à classe etária dos 80 ou mais anos. Somente cerca de 1% dos indivíduos é que têm 80 ou mais anos e acabaram por falecer. De todos os indivíduos, somente cerca de 6% destes pertencem à classe etária

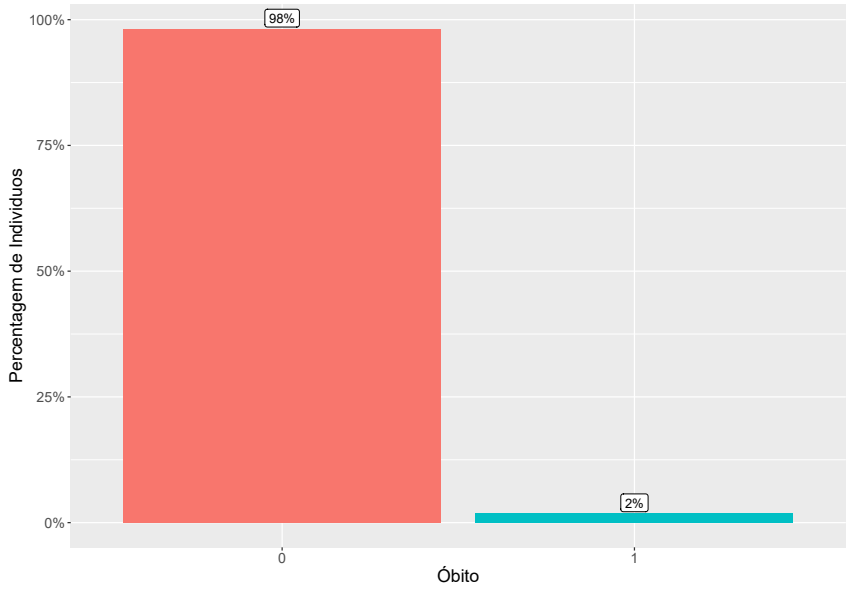


Figura 4.9: Percentagem de óbitos.

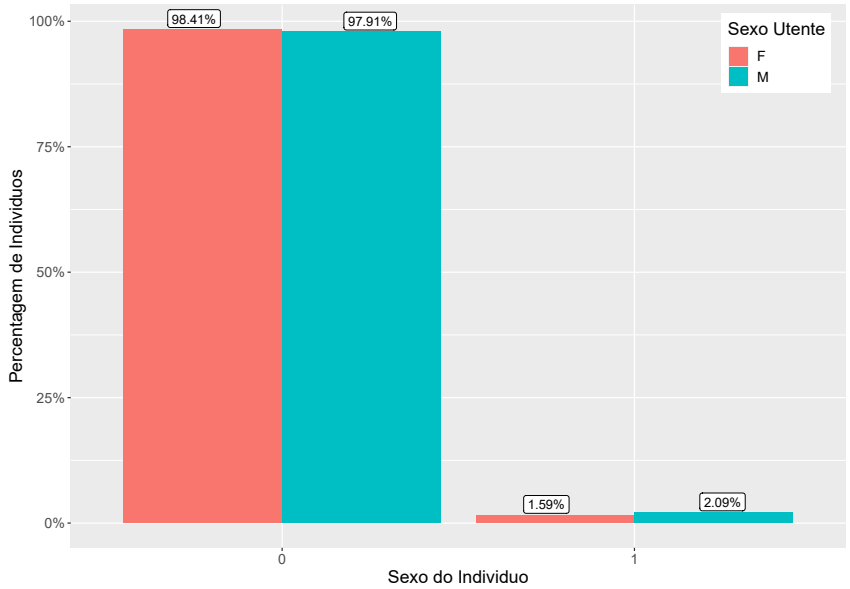


Figura 4.10: Distribuição de óbitos por sexo.

dos [0-9] anos e sobreviveram.

Na Figura 4.11, ainda se constata que 0.05% dos indivíduos acabaram por falecer e pertenciam à classe etária dos [50 – 59] anos, 0.16% pertenciam à classe etária dos [50 – 59] anos, 0.30% tinham entre 70 a 79 anos e 1.22% tinham 80 anos ou mais. Assim, ainda se pode concluir que, á medida que a classe etária é mais velha, tem-se um maior número de óbitos face à classe etária anterior.

O gráfico da Figura 4.12, mostra a proporção de indivíduos que acabaram por falecer em cada uma das classes etárias. As conclusões que se retiram da análise da Figura 4.12, são as mesmas que já foram mencionadas anteriormente, isto é, a classe etária dos >= 80

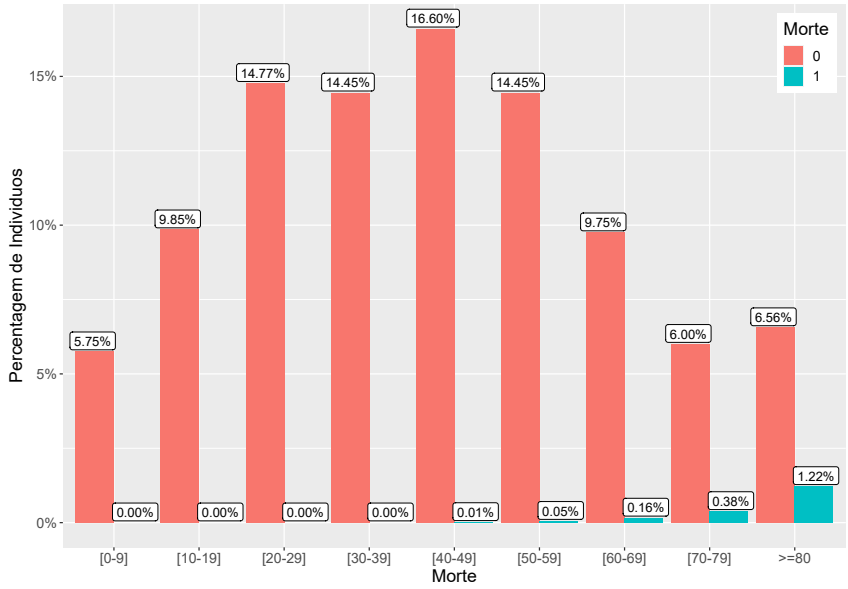


Figura 4.11: Distribuição de óbitos por classe etária face ao total de indivíduos, "1"=óbito.

anos é a classe que apresenta o maior número de mortes, seguindo-se a classe etária dos [70 − 79] anos, e assim sucessivamente.

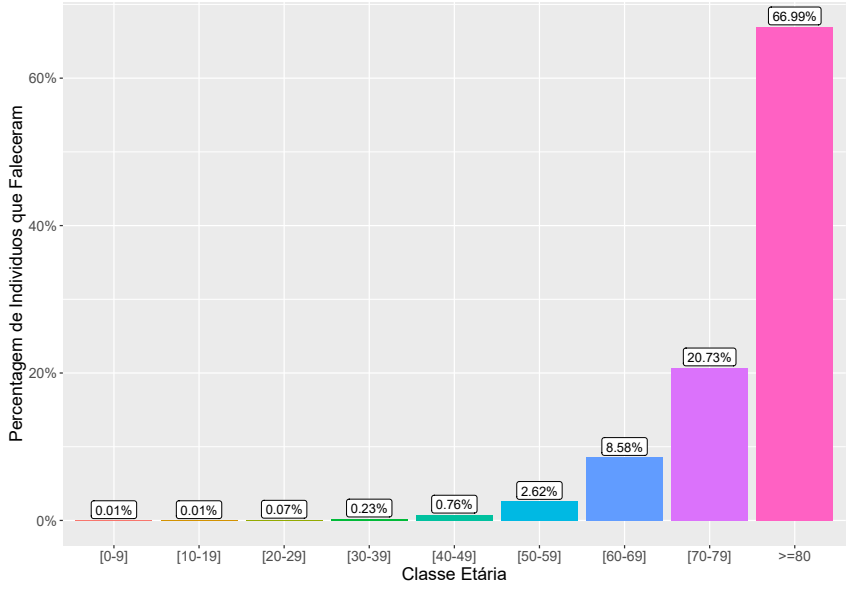


Figura 4.12: Percentagem de óbitos por classe etária.

Em suma, das conclusões retiradas da análise da Figura 4.12 e das conclusões referidas da análise da Figura 4.6, conclui-se que apesar das classes etárias mais avançadas apresentarem o menor número de casos face às classes etárias mais jovens, estas apresentam uma maior mortalidade, quando comparada com as classes etárias mais jovens.

A Figura 4.13, mostra ainda a percentagem de indivíduos que sobreviveram e faleceram, por cada uma das classes etárias e o que se pode concluir através da sua análise é que

nas diversas classes etárias tem-se sempre um maior número de sobreviventes, sendo que na classe etária dos [70 − 79] anos e na de >= 80anos é onde se observa maior percentagem de óbitos. Já nas restantes classes etárias, têm-se que praticamente todos os casos sobreviveram.

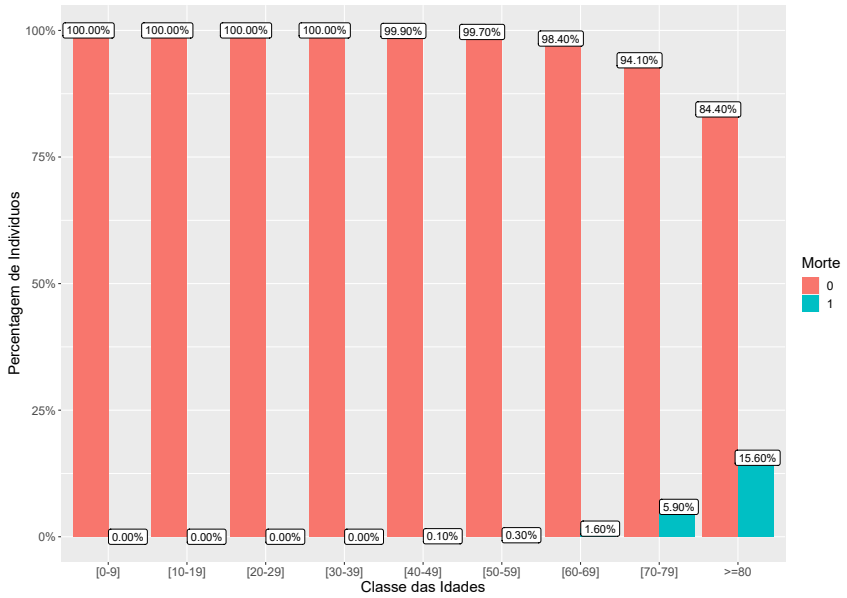


Figura 4.13: Percentagem de óbitos e não óbitos por classe etária, "1"=óbito.

Nas Figuras 4.14 e 4.15, pode-se observar o número de mortes por cada 100 mil habitantes, em cada um dos 20 distritos de Portugal. Constata-se que os 20 distritos que têm maior número de mortes por cada 100 mil habitantes, são os distritos de Bragança, Lisboa, Évora, Coimbra e Setúbal, com respetivamente 217, 155, 136, 135 e 133 mortes por cada 100 mil habitantes. Os distritos que apresentam o menor número de mortes são a Madeira e os Açores, com 27 e 8 mortes por cada 100 mil habitantes, respetivamente.

Em suma, fazendo uma análise das Figuras 4.3 e 4.4 e das Figuras 4.14 e 4.15, pode-se observar que apesar do distrito de Faro apresentar um número elevado de casos por cada 100 mil habitantes, este distrito é um dos distritos que apresenta menor número de mortes por cada 100 mil habitantes. Para o distrito da Madeira, também se pode observar que apesar deste apresentar um maior número de casos por cada 100 mil habitantes em relação a vários distritos de Portugal Continental, este é o distrito que apresenta o menor número de mortes por 100 mil habitantes. O distrito de Bragança é o 8º distrito que apresenta maior número de casos por cada 100 mil habitantes, no entanto é o que tem o maior número de mortes por cada 100 mil habitantes. Já no distrito da Guarda passa-se o contrário, pois este é o 2º distrito que apresenta menor número de casos por cada 100 mil habitantes, no entanto é o 3º distrito que apresenta o maior número de mortes por cada 100 mil habitantes. Comparando os distritos de Braga e de Santarém, observamos que o distrito de Braga é o 3º distrito que tem maior número de casos por cada 100 mil habitantes e tem cerca do dobro dos casos do distrito de Santarém, no entanto estes dois distritos apresentam o

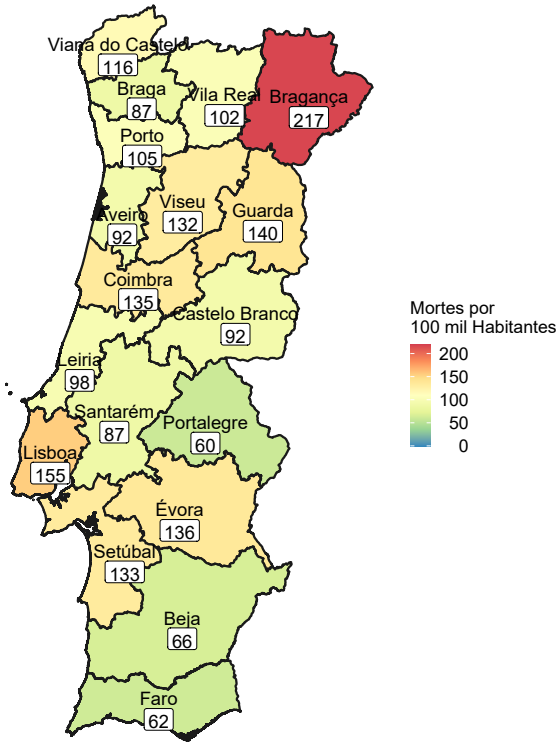
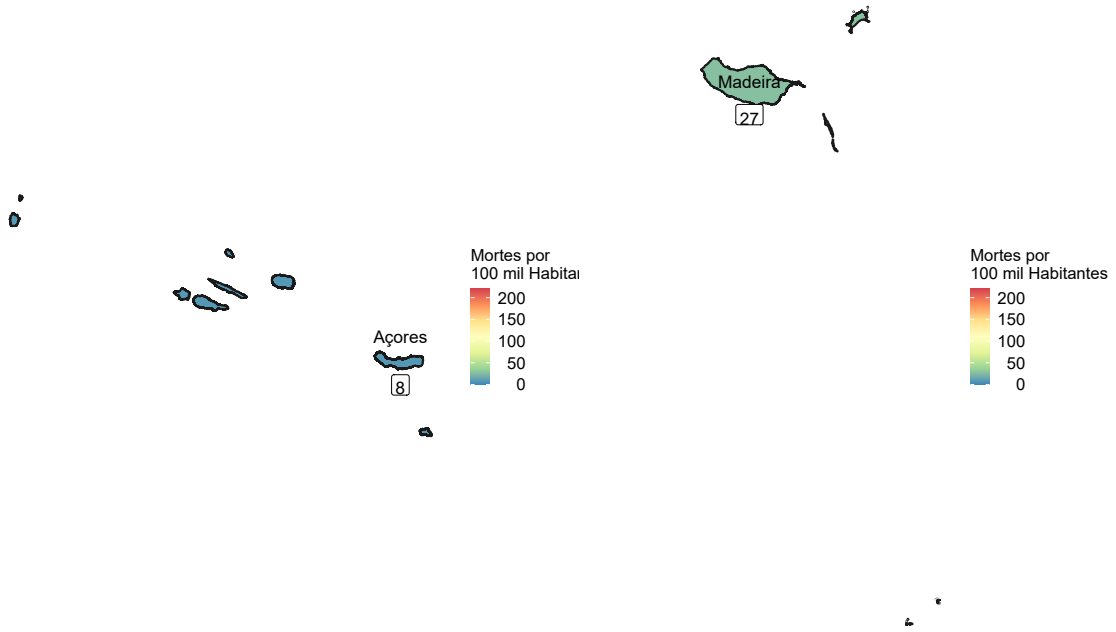


Figura 4.14: Número de óbitos por cada 100 mil habitantes em Portugal Continental.

mesmo número de mortes por cada 100 mil habitantes.



(a) Região Autónoma dos Açores. (b) Região Autónoma da Madeira.

Figura 4.15: Óbitos por cada 100 mil habitantes em Portugal, Regiões Autónomas.

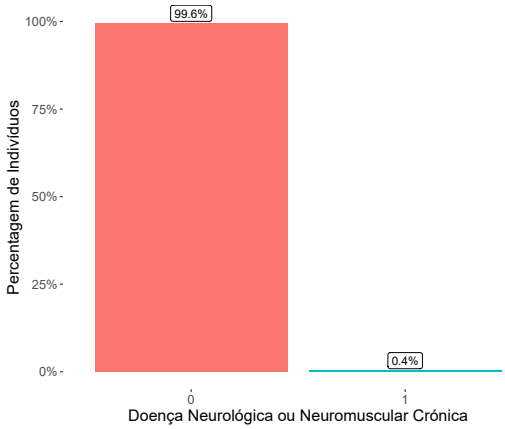
4.1.1 Comorbilidades

Nesta subsecção, apresentamos uma análise estatística às variáveis das comorbilidades, que estão na presente base de dados. Dado que as variáveis referentes às comorbilidades são variáveis dicotómicas que representam se um indivíduo apresenta ou não a comorbilidade, então na análise estatística iremos estudar a distribuição percentual dos indivíduos que têm a comorbilidade e também iremos relatar a distribuição dos óbitos pela comorbilidade.

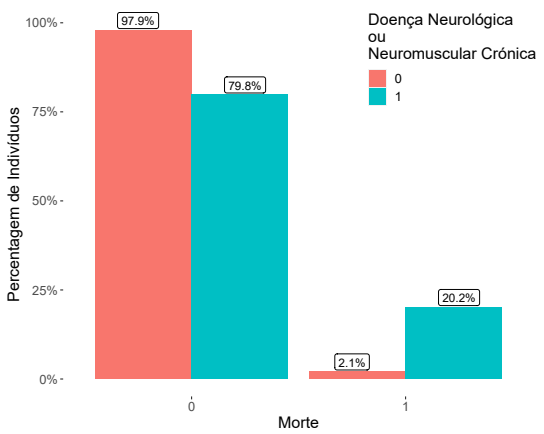
4.1.1.1 Doença Neurológica ou Neuromuscular Crónica

O gráfico da Figura 4.16a, apresenta a distribuição percentual dos indivíduos da base de dados com a doença neurológica ou neuromuscular crónica, sendo que se pode concluir que apenas 0.4% dos indivíduos da base de dados é que têm a comorbilidade doença neurológica ou neuromuscular crónica.

No gráfico da Figura 4.16b, apresenta-se a distribuição dos óbitos pela comorbilidade doença neurológica ou neuromuscular crónica. Dos indivíduos que apresentam a doença, cerca de 80% sobrevivem sendo que os restantes 20% acabam por falecer. Já no que diz respeito aos indivíduos que não apresentam a dada doença, apenas cerca de 2% é que acabam por falecer, pelo que este valor é um valor bastante inferior quando comparado com o valor homólogo para os pacientes com a comorbilidade.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade doença neurológica ou neuromuscular crónica.



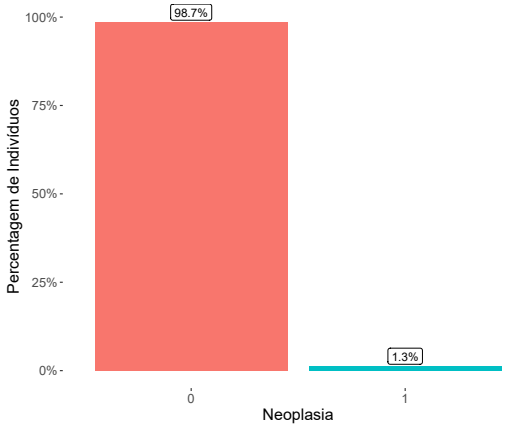
(b) Distribuição dos óbitos pela comorbilidade doença neurológica ou neuromuscular crónica.

Figura 4.16: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

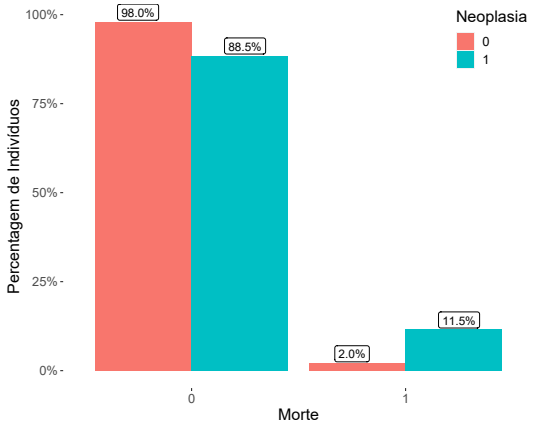
4.1.1.2 Neoplasia

O gráfico da Figura 4.17a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade neoplasia, onde se pode concluir que cerca de 1.3% dos indíviduos é que têm a comorbilidade neoplasia.

O gráfico da Figura 4.17b, apresenta a distribuição percentual dos óbitos pela comorbilidade neoplasia. Da sua análise, pode-se concluir que a percentagem de indivíduos que acabam por falecer é muito mais elevada nos indivíduos que têm neoplasia do que nos indivíduos que não apresentam a referida doença.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade neoplasia.



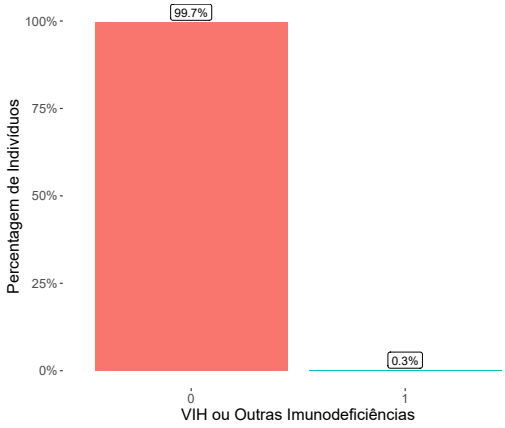
(b) Distribuição dos óbitos pela comorbilidade neoplasia.

Figura 4.17: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

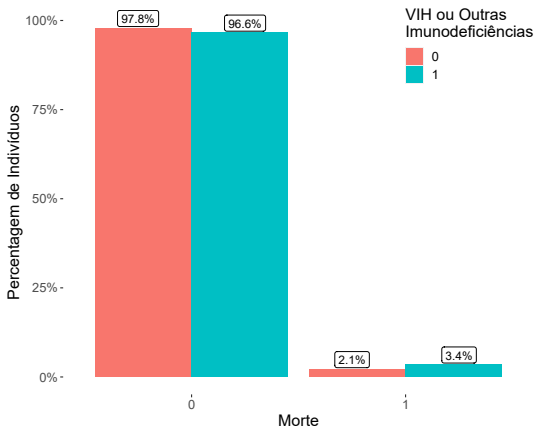
4.1.1.3 VIH ou Outra Imunodeficiência

O gráfico da Figura 4.18a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade VIH ou outra imunodeficiência, onde se pode concluir que apenas 0.3% destes é que têm a comorbilidade VIH ou outra imunodeficiência.

No gráfico da Figura 4.18b, apresenta-se a distribuição dos óbitos pela comorbilidade VIH ou outras imunodeficiência. Da análise da Figura 4.18b, deduz-se que a percentagem de indivíduos que acabam por falecer e que apresentam a doença é muito próxima da percentagem de indivíduos que faleceram e não apresentam a doença, no entanto esta última apresenta um valor menor, que é de cerca 2%, enquanto que a percentagem de indivíduos que acabam por falecer e têm a VIH ou outra imunodeficiência é de cerca de 3%.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade VIH ou outra imunodeficiência.



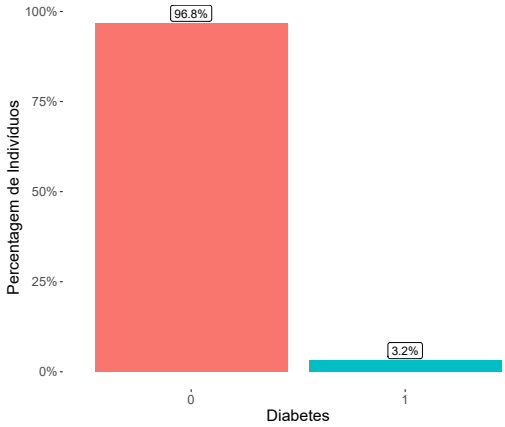
(b) Distribuição dos óbitos pela comorbilidade VIH ou outra imunodeficiência.

Figura 4.18: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

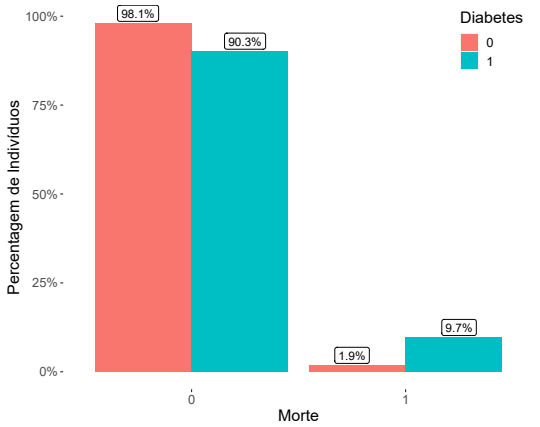
4.1.1.4 Diabetes

O gráfico da Figura 4.19a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade diabetes, onde se pode concluir que cerca de 3.2% destes é que têm a comorbilidade diabetes.

No gráfico da Figura 4.19b, apresenta-se a distribuição dos óbitos pela comorbilidade diabetes, sendo que da sua análise concluí-se que percentagem de mortes é muito mais elevada nos indivíduos que apresentam a comorbilidade diabetes do que os que não apresentam.



(a) Distribuição percentual de indivíduos com (1) e sem (0) diabetes.



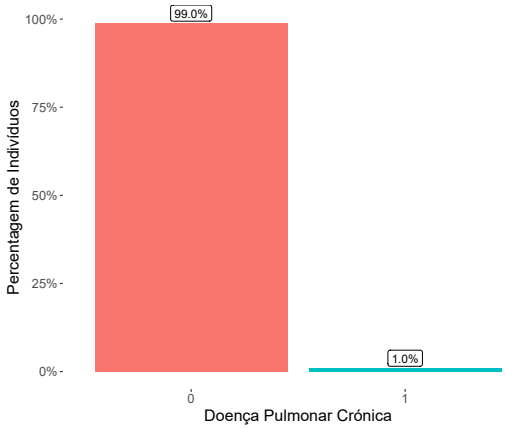
(b) Distribuição dos óbitos pela comorbilidade diabetes.

Figura 4.19: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

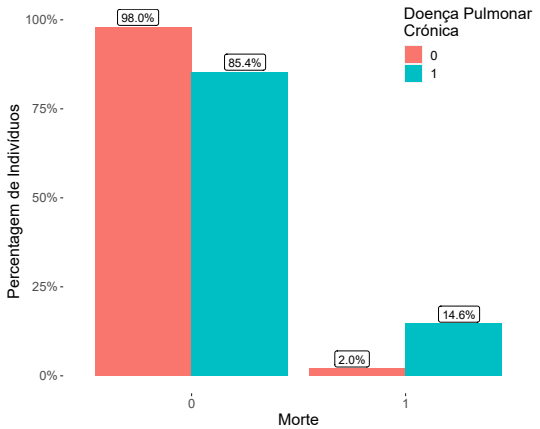
4.1.1.5 Doença Pulmonar Crónica

O gráfico da Figura 4.20a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade doença pulmonar crónica, onde se conclui apenas 1.0% dos indivíduos é que têm a comorbilidade doença pulmonar crónica.

No gráfico da Figura 4.20b, tem-se apresentada a distribuição dos óbitos pela comorbilidade doença pulmonar crónica. Da sua análise, concluí-se que percentagem de mortes é muito mais elevada nos indivíduos que apresentam a comorbilidade doença pulmonar crónica do que os que não apresentam.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade doença pulmonar crónica.



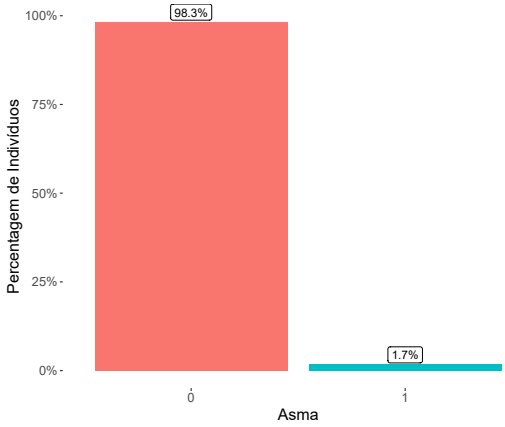
(b) Distribuição dos óbitos pela comorbilidade doença pulmonar crónica.

Figura 4.20: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

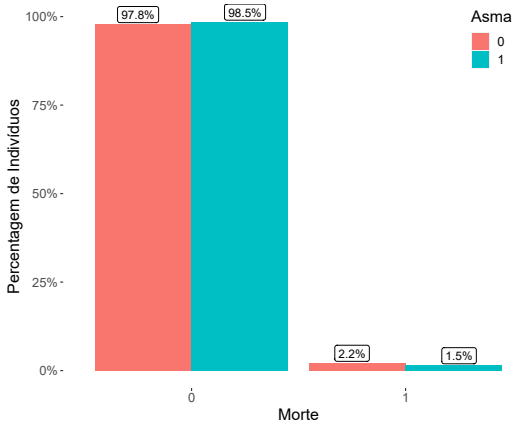
4.1.1.6 Asma

O gráfico da Figura 4.21a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade asma, onde se conclui que cerca 2% dos indivíduos é que têm a comorbilidade asma.

A Figura 4.21b, mostra a distribuição dos óbitos pela comorbilidade asma. Através da análise da Figura 4.21b, conclui-se que a percentagem de indivíduos que acabam por falecer é mais baixa nos indivíduos com a comorbilidade asma do que nos indivíduos que não têm a comorbilidade asma.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade asma.



(b) Distribuição dos óbitos pela comorbilidade asma.

Figura 4.21: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

4.1.1.7 Doença Hematológicas Crónicas

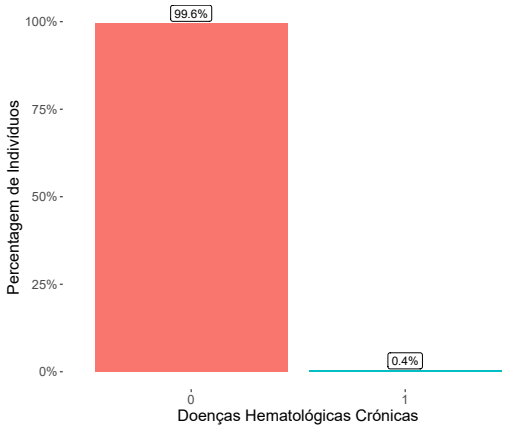
O gráfico da Figura 4.22a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade doença hematológica crónica, onde se conclui que apenas uma parte residual destes indivíduos é que apresentam esta comorbilidade.

O gráfico da Figura 4.22b, apresenta a distribuição dos óbitos pela comorbilidade doença hematológica crónica, donde se conclui que a percentagem de indivíduos que acabam por falecer é mais elevada nos indivíduos com a comorbilidade doença hematológica crónica do que nos indivíduos que não têm a comorbilidade doença hematológica crónica.

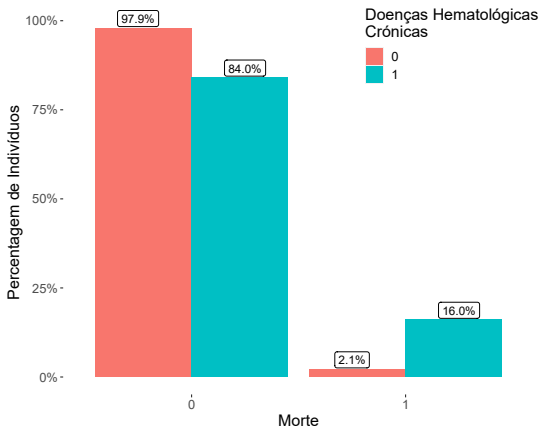
4.1.1.8 Patologia Hepática

O gráfico da Figura 4.23a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade patologia hepática, onde se conclui que apenas 0.3% dos indivíduos é que apresentam esta comorbilidade.

No gráfico da Figura 4.23b, apresenta-se a distribuição dos óbitos pela comorbilidade patologia hepática, onde se conclui que a percentagem de mortes dos indivíduos que apresentam a comorbilidade patologia hepática é muito mais elevada do que os indivíduos que não a apresentam.

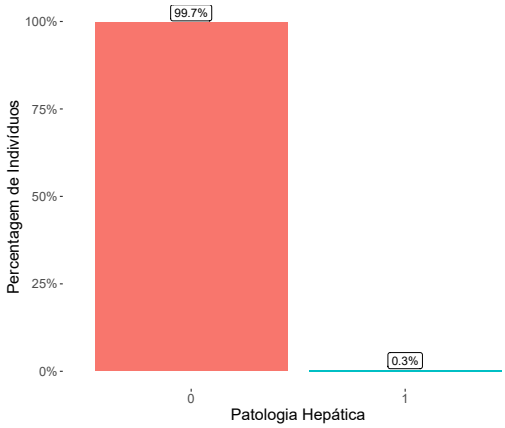


(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade doença hematológica crónica.

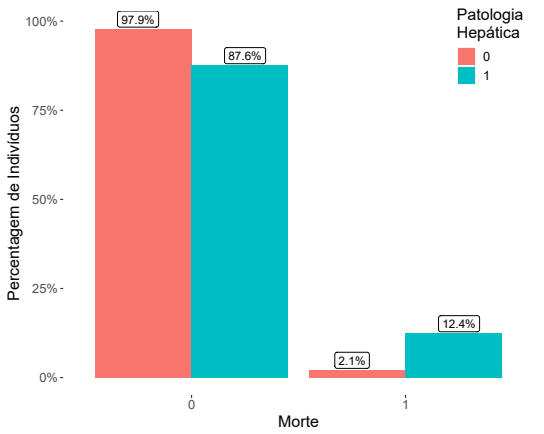


(b) Distribuição dos óbitos pela comorbilidade doença hematológica crónica.

Figura 4.22: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade patologia hepática.



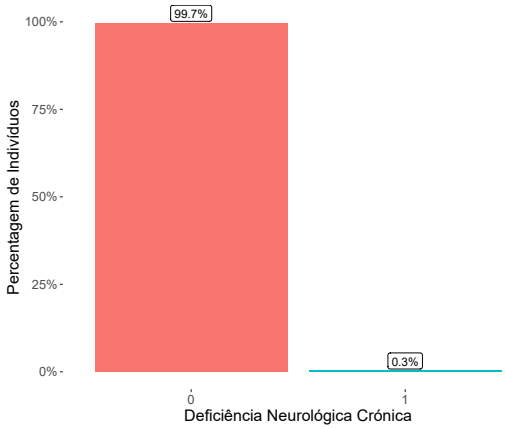
(b) Distribuição dos óbitos pela comorbilidade patologia hepática.

Figura 4.23: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

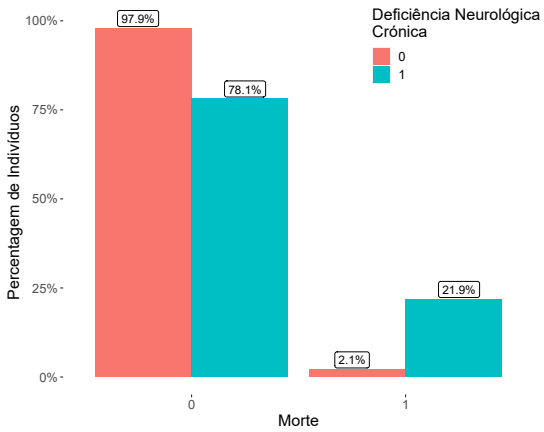
4.1.1.9 Deficiência Neurológica Crónica

O gráfico da Figura 4.24a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade deficiência neurológica crónica, onde se conclui que apenas 0.3% dos indivíduos é que apresentam esta comorbilidade.

A Figura 4.24b, mostra a distribuição dos óbitos pela comorbilidade deficiência neurológica crónica. Através da análise da Figura 4.24b, pode-se concluir que acabam por falecer mais indivíduos que apresentam patologia hepática dos que acabam por falecer e não têm a patologia hepática, uma vez que, a percentagem de indivíduos que têm a patologia hepática é bastante superior à percentagem de indíviduos que não apresentam patologia hepática e que acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade deficiência neurológica crónica.



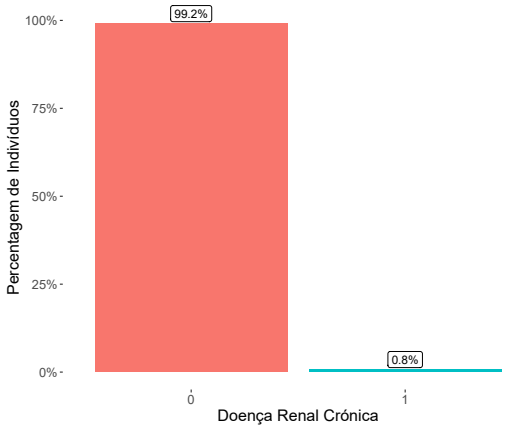
(b) Distribuição dos óbitos pela comorbilidade deficiência neurológica crónica.

Figura 4.24: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

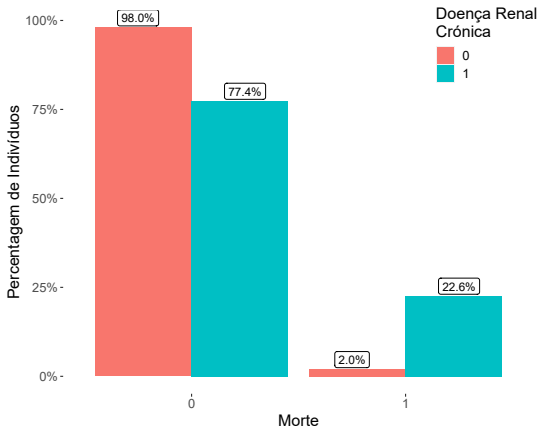
4.1.1.10 Doença Renal Crónica

O gráfico da Figura 4.25a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade doença renal crónica, onde se conclui que apenas 0.8% dos indivíduos é que apresentam esta comorbilidade.

Na Figura 4.25b, tem-se apresentada a distribuição dos óbitos pela comorbilidade doença renal crónica, donde se conclui que a percentagem de indivíduos que apresentam a doença renal crónica e que acabam por falecer é mais elevada face aos indivíduos que não apresentam a referida doença e que acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade doença renal crónica.



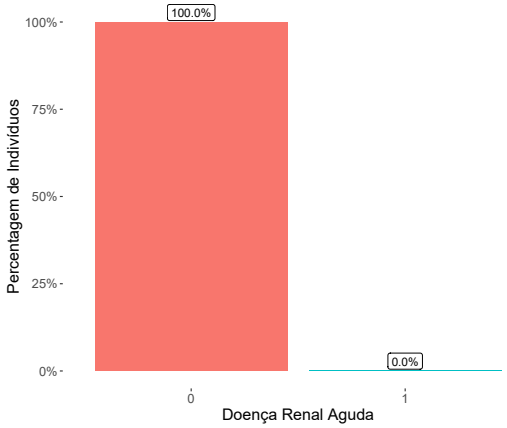
(b) Distribuição dos óbitos pela comorbilidade doença renal crónica.

Figura 4.25: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

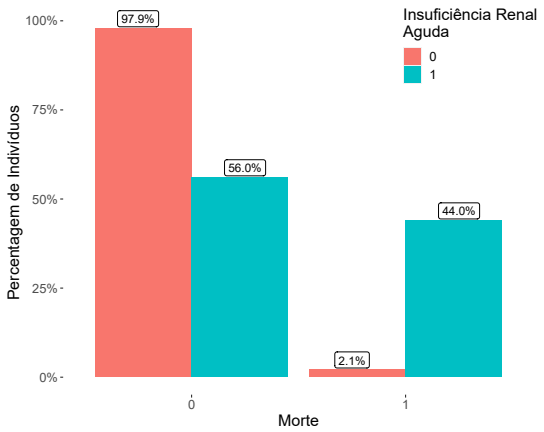
4.1.1.11 Insuficiência Renal Aguda

O gráfico da Figura 4.26a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade insuficiência renal aguda, onde se conclui que praticamente nenhum dos indivíduos apresenta a referida comorbilidade.

O gráfico da Figura 4.26b, mostra a distribuição dos óbitos pela comorbilidade insuficiência renal aguda. Da análise da Figura 4.26b, conclui-se que a percentagem de indivíduos que acabam por falecer e têm a comorbilidade insuficiência renal aguda é mais elevada face aos indivíduos que acabam por falecer e não apresentam a referida comorbilidade.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade insuficiência renal aguda.



(b) Distribuição dos óbitos pela comorbilidade insuficiência renal aguda.

Figura 4.26: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

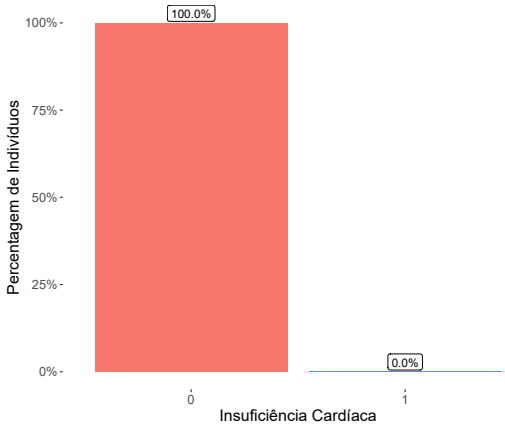
4.1.1.12 Insuficiência Cardíaca

O gráfico da Figura 4.27a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade insuficiência cardíaca, onde se conclui que praticamente nenhum dos indivíduos apresenta a referida comorbilidade.

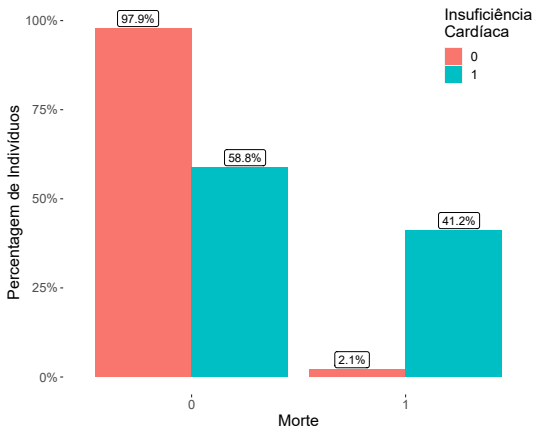
No gráfico da Figura 4.27b, tem-se apresentada a distribuição dos óbitos pela comorbilidade insuficiência cardíaca, que através da sua análise, conclui-se que a percentagem de indivíduos que acabam por falecer e têm a comorbilidade insuficiência cardíaca é bastante mais elevada face aos indivíduos que acabam por falecer e não apresentam a referida comorbilidade.

4.1.1.13 Coagolopatia de Consumo

O gráfico da Figura 4.28a, apresenta a distribuição percentual dos indivíduos da base de dados com a comorbilidade coagolopatia de consumo, onde se conclui que praticamente nenhum dos indivíduos apresenta a referida comorbilidade.



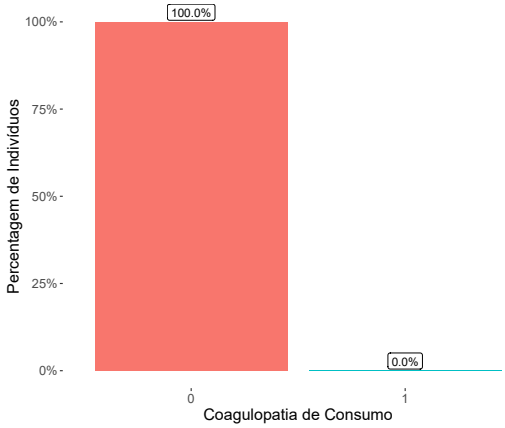
(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade insuficiência cardíaca.



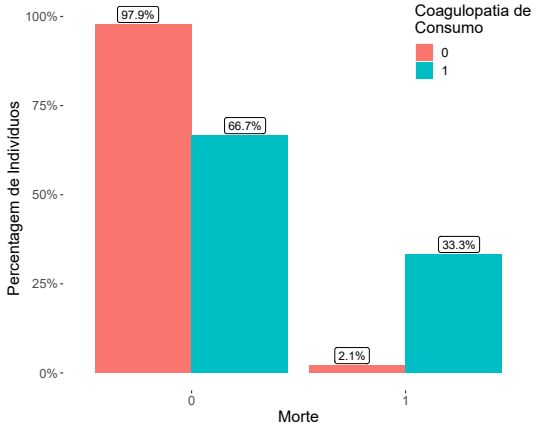
(b) Distribuição dos óbitos pela comorbilidade insuficiência cardíaca.

Figura 4.27: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

O gráfico da Figura 4.28b, apresenta a distribuição dos óbitos pela comorbilidade coagolopatia de consumo, donde se observa que existe uma maior percentagem de indivíduos que acabam por falecer quando estes apresentam a comorbilidade coagolopatia de consumo, dos que acabam por falecer e não apresentam coagolopatia de consumo.



(a) Distribuição percentual de indivíduos com (1) e sem (0) a comorbilidade coagolopatia de consumo.



(b) Distribuição dos óbitos pela comorbilidade coagolopatia de consumo.

Figura 4.28: Distribuição pela comorbilidade (a) e distribuição dos óbitos pela comorbilidade (b).

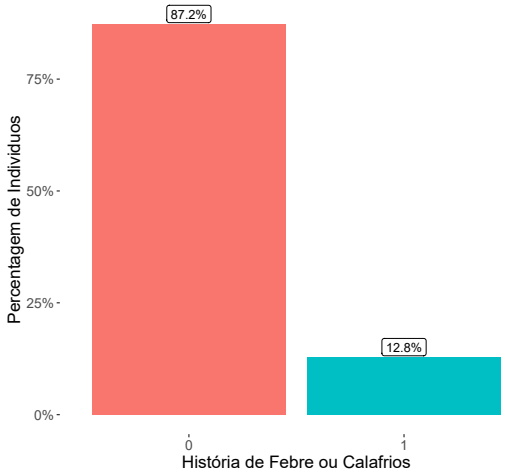
4.1.2 Sintomas

Nesta subsecção, apresentamos uma análise estatística às variáveis dos sintomas, que estão na presente base de dados. Dado que as variáveis referentes aos sintomas são variáveis dicotómicas que representam se um indivíduo apresenta ou não o sintoma, então na análise estatística iremos estudar a distribuição percentual dos indivíduos que têm o sintoma e também iremos relatar a distribuição dos óbitos pelo sintoma.

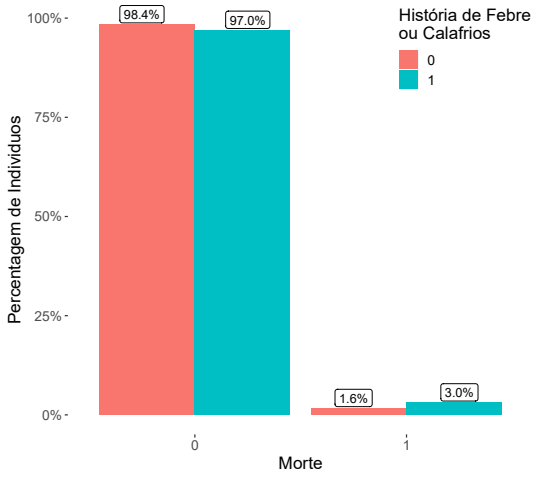
4.1.2.1 História de Febre ou Calafrios

O gráfico da Figura 4.29a apresenta a distribuição percentual dos indivíduos da base de dados com o sintoma história de febre ou calafrios, sendo que cerca de 12.8% dos indivíduos têm este sintoma.

No gráfico da Figura 4.29b, tem-se representado a distribuição dos óbitos pelo sintoma história de febre ou calafrios. Da sua análise, pode-se concluir que a percentagem de indivíduos que acabaram por falecer dado que estes apresentavam o sintoma história de febre ou calafrios é de 3%, sendo este valor mais elevado, quando comparado com os indivíduos que faleceram e não apresentavam o sintoma história de febre ou calafrios.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma história de febre ou calafrios.



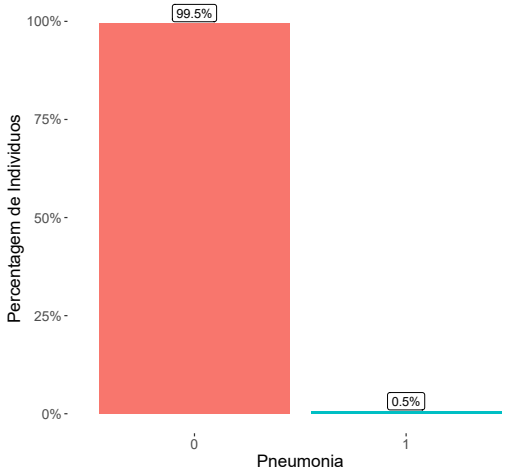
(b) Distribuição dos óbitos pelo sintoma história de febre ou calafrios.

Figura 4.29: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

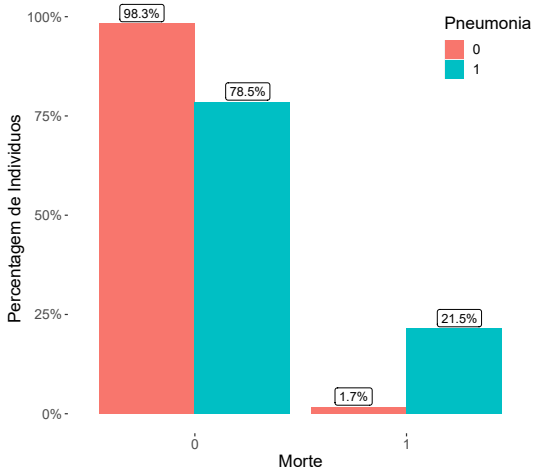
4.1.2.2 Pneumonia

O gráfico da Figura 4.30a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma pneumonia, sendo que apenas 0.5% dos indivíduos é que têm este sintoma.

No gráfico da Figura 4.30b, tem-se representado a distribuição dos óbitos pelo sintoma pneumonia. Da sua análise, pode-se deduzir que existe uma diferença muito grande entre a percentagem de indivíduos que apresentam a pneumonia e que acabam por falecer face à percentagem de indivíduos que não apresentam o dado sintoma e que também acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma pneumonia.



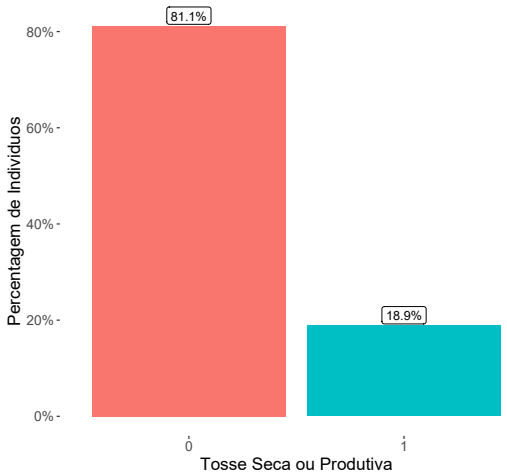
(b) Distribuição dos óbitos pelo sintoma pneumonia.

Figura 4.30: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

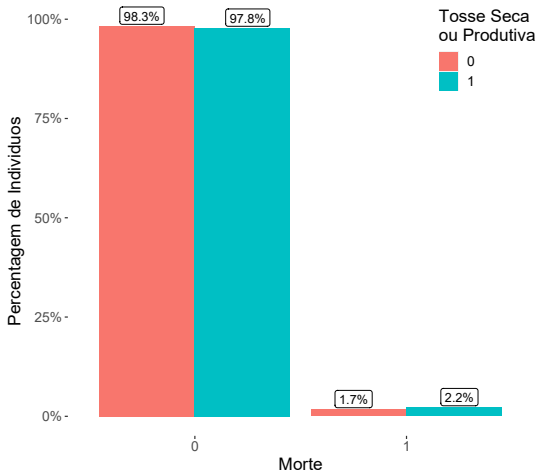
4.1.2.3 Tosse Seca ou Produtiva

O gráfico da Figura 4.31a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma tosse seca ou produtiva, sendo que cerca de 19% dos indivíduos têm este sintoma.

No gráfico da Figura 4.31b, tem-se representado a distribuição dos óbitos pelo sintoma tosse seca ou produtiva. Da sua análise, pode-se deduzir que a percentagem de indíviduos que acabam por falecer e apresentam o sintoma tosse seca ou produtiva é idêntico ao seu valor homólogo para os indivíduos que não apresentam o referido sintoma tosse seca ou produtiva, sendo em ambos os casos cerca de 2%.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma tosse seca ou produtiva.



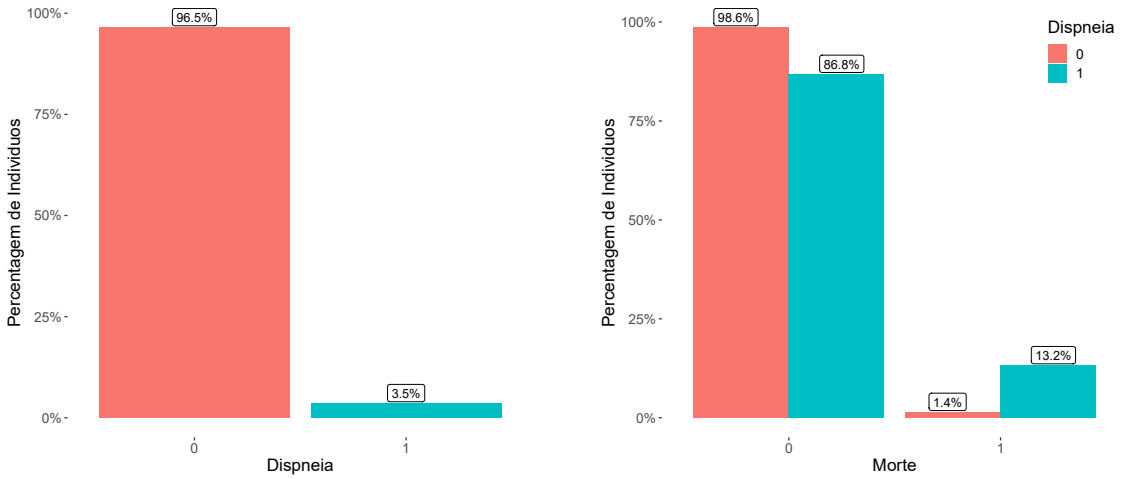
(b) Distribuição dos óbitos pelo sintoma tosse seca ou produtiva.

Figura 4.31: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.1.2.4 Dispneia

O gráfico da Figura 4.32a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma dispneia, sendo que cerca de 3.5% dos indivíduos têm este sintoma.

No gráfico da Figura 4.32b, tem-se representado a distribuição dos óbitos pelo sintoma dispneia. Da sua análise, pode-se deduzir que a percentagem de indíviduos que acabam por falecer e apresentam o sintoma dispneia é mais elevada face aos indivíduos que não apresentam o referido sintoma e que também acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma dispneia.

(b) Distribuição dos óbitos pelo sintoma dispneia.

Figura 4.32: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.1.2.5 Coriza

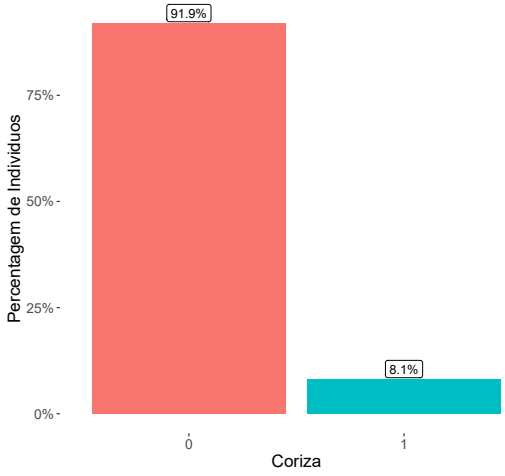
O gráfico da Figura 4.33a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma coriza, sendo que cerca de 8.1% dos indivíduos têm este sintoma.

No gráfico da Figura 4.33b, tem-se representado a distribuição dos óbitos pelo sintoma coriza. Da sua análise, pode-se deduzir que a percentagem de indíviduos que acabam por falecer e apresentam o sintoma coriza é mais baixa que a percentagem de indivíduos que não apresentam o referido sintoma e que acabam por falecer.

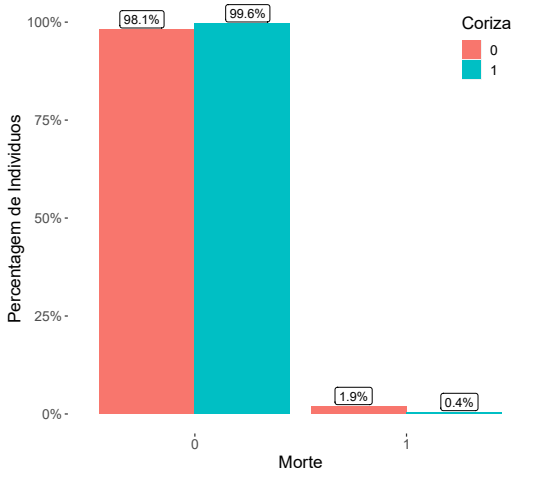
4.1.2.6 Odinofagia

O gráfico da Figura 4.34a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma odinofagia, sendo que cerca de 6.7% dos indivíduos têm este sintoma.

No gráfico da Figura 4.34b, tem-se representado a distribuição dos óbitos pelo sintoma odinofagia. Da sua análise, pode-se deduzir que a percentagem de indíviduos que acabam por falecer e apresentam o sintoma odinofagia é mais baixa que a percentagem de indivíduos que não apresentam o referido sintoma e que acabam por falecer.

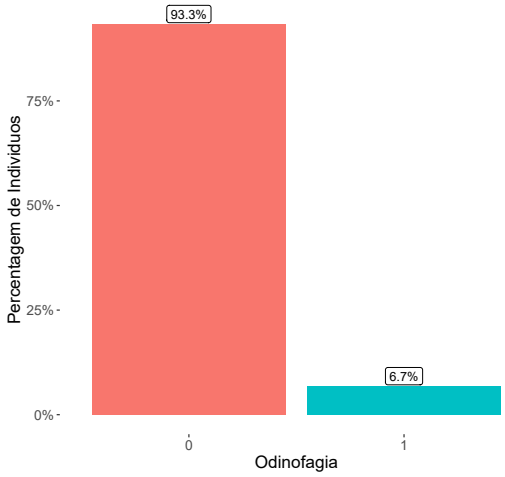


(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma coriza.

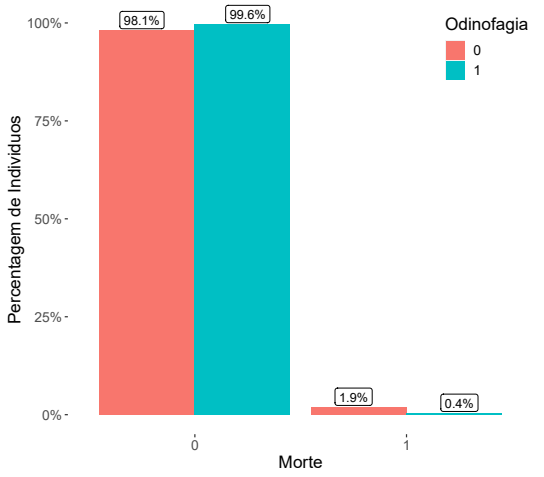


(b) Distribuição dos óbitos pelo sintoma coriza.

Figura 4.33: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma odinofagia.



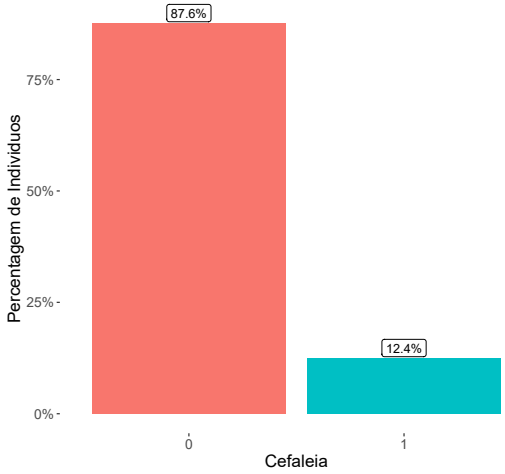
(b) Distribuição dos óbitos pelo sintoma odinofagia.

Figura 4.34: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

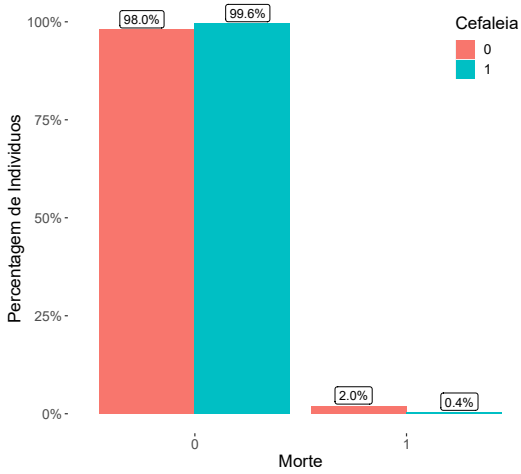
4.1.2.7 Cefaleia

O gráfico da Figura 4.35a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma cefaleia, sendo que cerca de 12% dos indivíduos têm este sintoma.

No gráfico da Figura 4.35b, tem-se representado a distribuição dos óbitos pelo sintoma cefaleia. Da sua análise, pode-se deduzir que a percentagem de indíviduos que acabam por falecer e apresentam o sintoma cefaleia é mais baixa que a percentagem de indivíduos que não apresentam o referido sintoma e que acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma cefaleia.



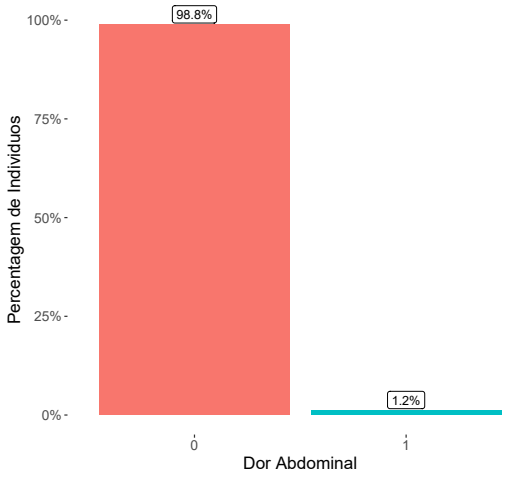
(b) Distribuição dos óbitos pelo sintoma cefaleia.

Figura 4.35: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

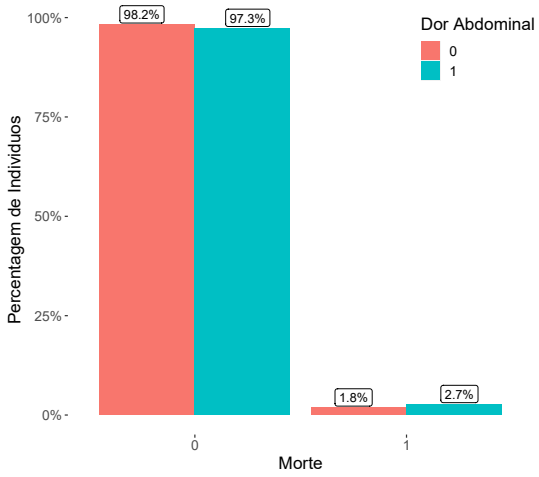
4.1.2.8 Dor Abdominal

O gráfico da Figura 4.36a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma dor abdominal, sendo que apenas 1.2% dos indivíduos é que têm este sintoma.

No gráfico da Figura 4.36b, tem-se representado a distribuição dos óbitos pelo sintoma dor abdominal. Da sua análise, observa-se que a percentagem de indivíduos que acabam por falecer é mais elevado nos indivíduos que apresentam o sintoma da dor abdominal do que os que não apresentam, sendo que dos indivíduos que não apresentam o sintoma cerca de 2% acabam por falecer, enquanto que nos indivíduos que apresentam o sintoma em estudo, cerca de 3% é que acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma dor abdominal.



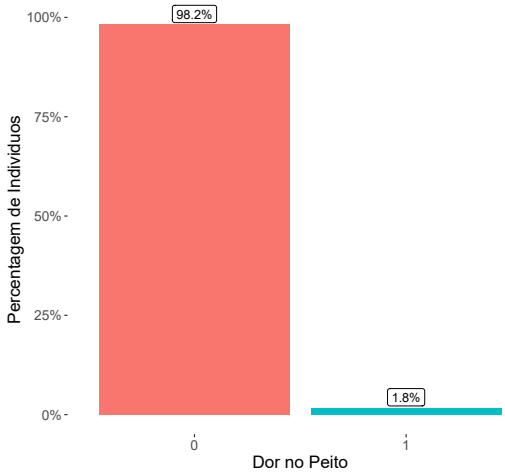
(b) Distribuição dos óbitos pelo sintoma dor abdominal.

Figura 4.36: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

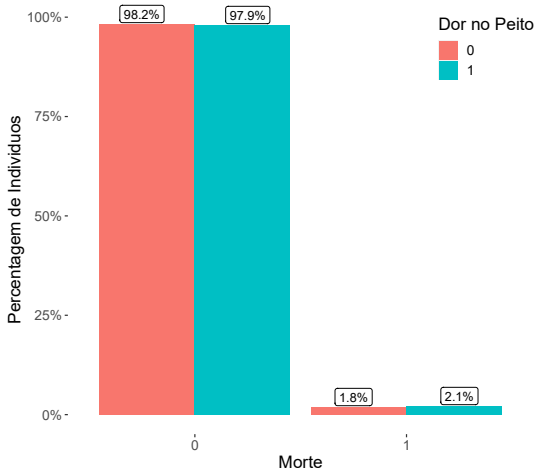
4.1.2.9 Dor no Peito

O gráfico da Figura 4.37a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma dor no peito, sendo que apenas de 1.8% dos indivíduos é que têm este sintoma.

No gráfico da Figura 4.37b, tem-se representado a distribuição dos óbitos pelo sintoma dor no peito. Da sua análise, observa-se que a percentagem de indivíduos que apresenta o sintoma dor no peito e que acabaram por falecer é muito idêntica à percentagem de indivíduos que acabaram por falecer e não apresenta o referido sintomas, uma vez que em ambos os casos, cerca de 2% dos indivíduos é que acabam por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma dor no peito.



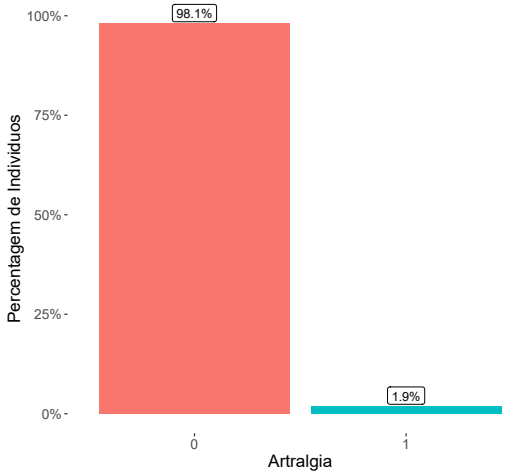
(b) Distribuição dos óbitos pelo sintoma dor no peito.

Figura 4.37: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

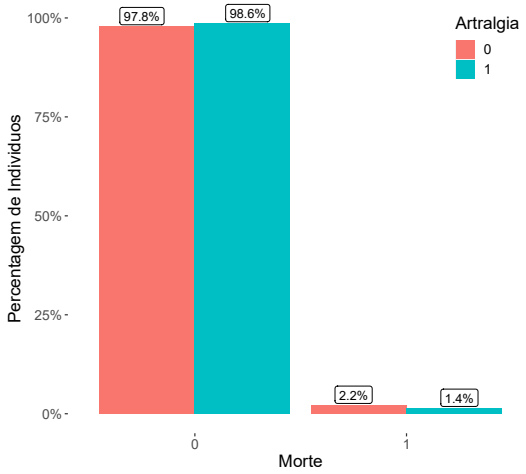
4.1.2.10 Artralgia

O gráfico da Figura 4.38a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma artralgia, sendo que cerca de 2.0% dos indivíduos apresentam o sintoma.

No gráfico da Figura 4.38b, tem-se representado a distribuição dos óbitos pelo sintoma artralgia. Da sua análise, observa-se que a percentagem de indivíduos que apresentam o sintoma artralgia e que acabaram por falecer é mais baixa do que a percentagem de indivíduos que acabaram por falecer e não apresentam o referido sintomas.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma artralgia.



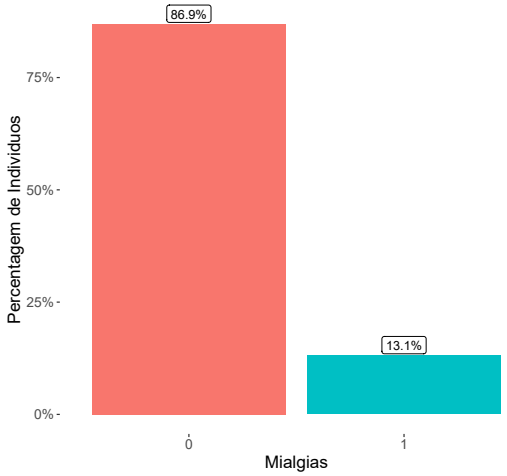
(b) Distribuição dos óbitos pelo sintoma artralgia.

Figura 4.38: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

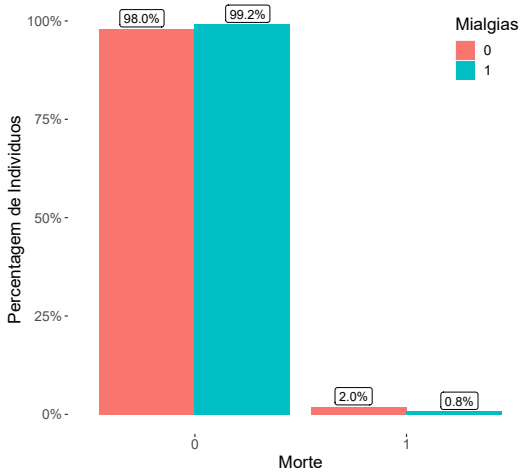
4.1.2.11 Mialgias

O gráfico da Figura 4.39a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma mialgias, sendo que cerca de 13.0% dos indivíduos apresentam o sintoma.

No gráfico da Figura 4.39b, tem-se representado a distribuição dos óbitos pelo sintoma mialgias. Da sua análise, pode-se concluir que o número de indivíduos que acabam por falecer é mais elevado nos indivíduos que não apresentam o sintoma mialgias do que nos indivíduos que apresentam o referido sintoma.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma mialgias.



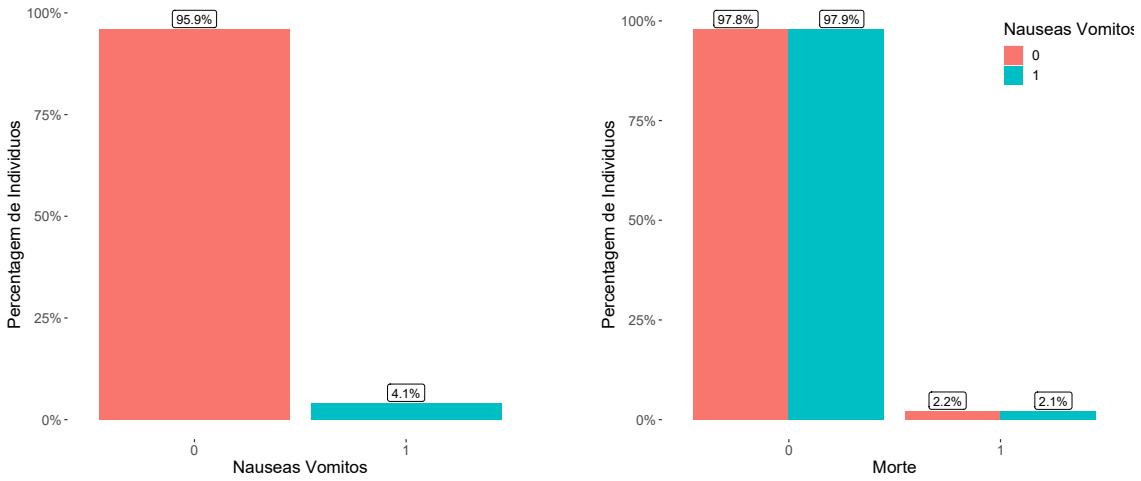
(b) Distribuição dos óbitos pelo sintoma mialgias.

Figura 4.39: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.1.2.12 Náuseas ou Vómitos

O gráfico da Figura 4.40a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma náuseas ou vómitos, sendo que cerca de 4.1% dos indivíduos apresentam o sintoma.

No gráfico da Figura 4.40b, tem-se representado a distribuição dos óbitos pelo sintoma náuseas ou vómitos. Da sua análise, pode-se concluir que a percentagem de indivíduos que faleceram e apresentam o sintoma em estudo é idêntica á percentagem de indivíduos que acabaram por falecer e não apresentam o sintoma náuseas ou vómitos, sendo que em ambos os casos cerca de 2% dos indivíduos é que acabaram por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma náuseas ou vómitos.

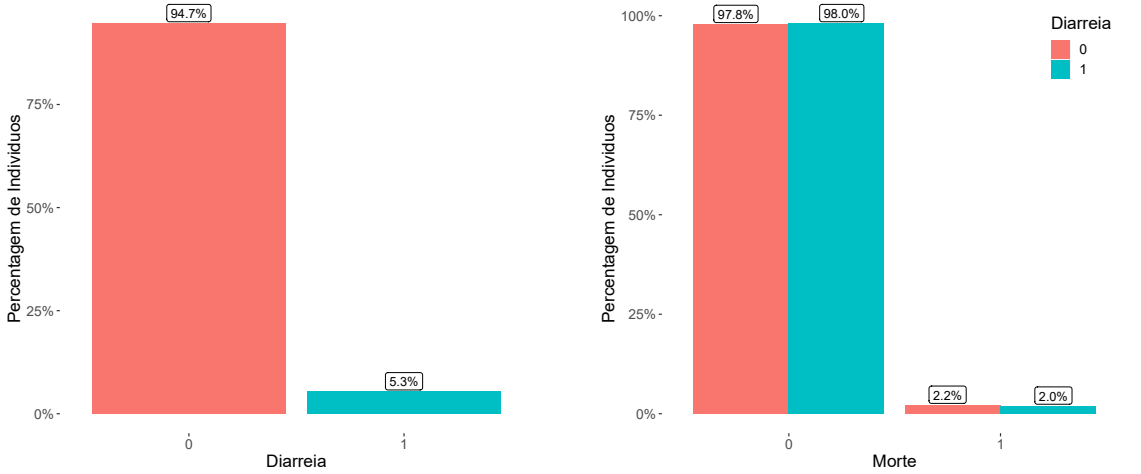
(b) Distribuição dos óbitos pelo sintoma náuseas ou vómitos.

Figura 4.40: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.1.2.13 Diarreia

O gráfico da Figura 4.41a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma diarreia, sendo que cerca de 5.3% dos indivíduos apresentam o sintoma.

No gráfico da Figura 4.41b, tem-se representado a distribuição dos óbitos pelo sintoma diarreia. Da sua análise, pode-se concluir que a percentagem de indivíduos que faleceram e apresentam o sintoma em estudo é idêntica á percentagem de indivíduos que acabaram por falecer e não apresentam o sintoma diarreia, sendo que em ambos os casos cerca de 2% dos indivíduos é que acabaram por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma diarreia.

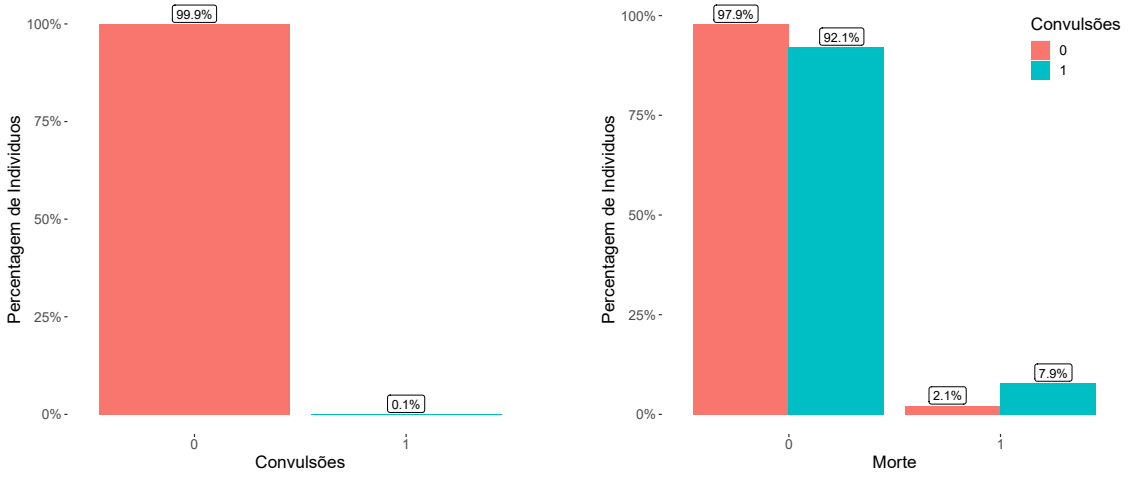
(b) Distribuição dos óbitos pelo sintoma diarreia.

Figura 4.41: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.1.2.14 Convulsões

O gráfico da Figura 4.42a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma convulsões, sendo que apenas 0.1% dos indivíduos é que apresentam o sintoma.

No gráfico da Figura 4.42b, tem-se representado a distribuição dos óbitos pelo sintoma convulsões. Da sua análise, pode-se concluir que a percentagem de indivíduos que faleceram e apresentam o sintoma convulsões é maior do que a percentagem de indivíduos que acabaram por falecer e não apresentam o sintoma em causa.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma convulsões.

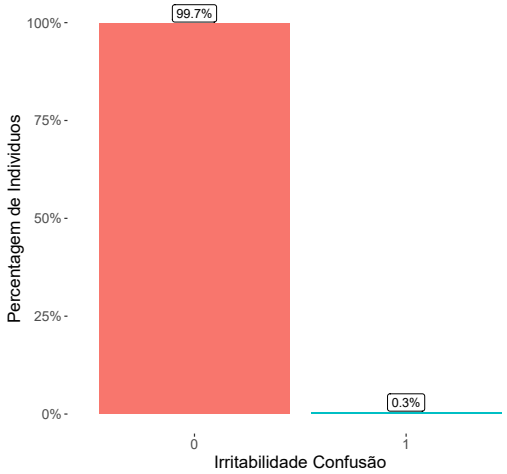
(b) Distribuição dos óbitos pelo sintoma convulsões.

Figura 4.42: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

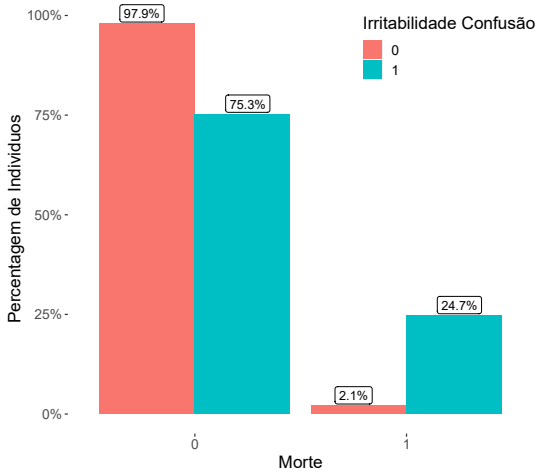
4.1.2.15 Irritabilidade Confusão

O gráfico da Figura 4.43a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma irritabilidade confusão, sendo que apenas 0.3% dos indivíduos é que apresentam o sintoma.

No gráfico da Figura 4.43b, tem-se representado a distribuição dos óbitos pelo sintoma irritabilidade confusão. Da sua análise, pode-se concluir que a percentagem de indivíduos que faleceram e apresentam o sintoma irritabilidade confusão é bastante superior à percentagem de indivíduos que acabaram por falecer e não apresentam o sintoma em causa.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma irritabilidade confusão.



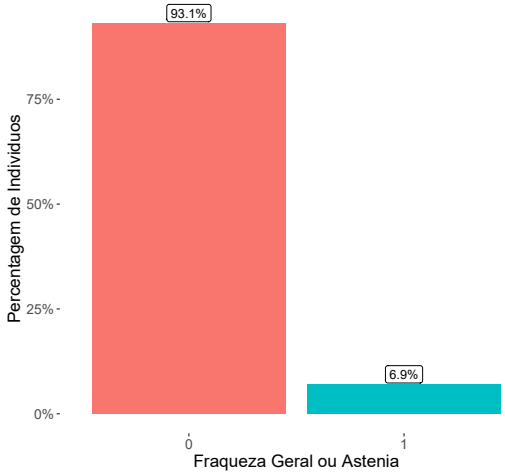
(b) Distribuição dos óbitos pelo sintoma irritabilidade confusão.

Figura 4.43: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

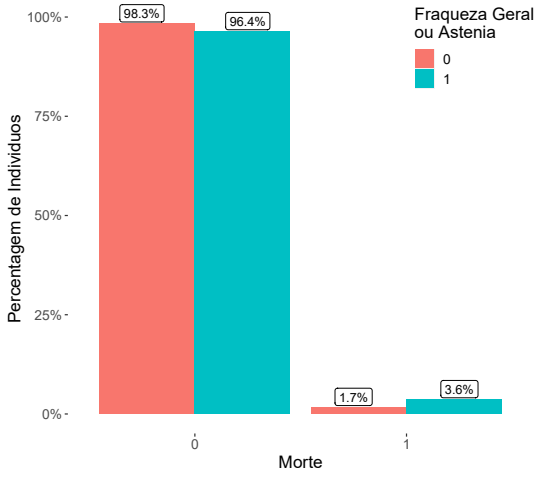
4.1.2.16 Fraqueza Geral ou Astenia

O gráfico da Figura 4.44a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma fraqueza geral ou astneia, sendo que cerca de 7.0% dos indivíduos é que apresentam o sintoma.

No gráfico da Figura 4.44b, tem-se representado a distribuição dos óbitos pelo sintoma fraqueza geral ou astneia. Da sua análise, observa-se que a percentagem de indivíduos que não apresentam o sintoma em causa e que acabaram por falecer é cerca de 2%, sendo este valor menor quando comparado com o valor homólogo para os indivíduos que apresentem o sintoma em causa, pois neste último caso apenas cerca de 4% dos indivíduos é que acabaram por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma fraqueza geral ou astneia.



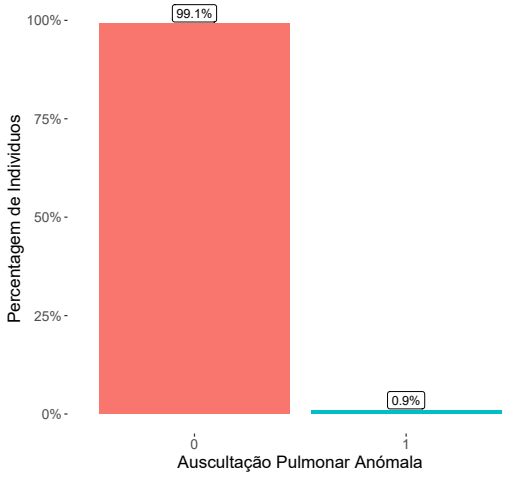
(b) Distribuição dos óbitos pelo sintoma fraqueza geral ou astneia.

Figura 4.44: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

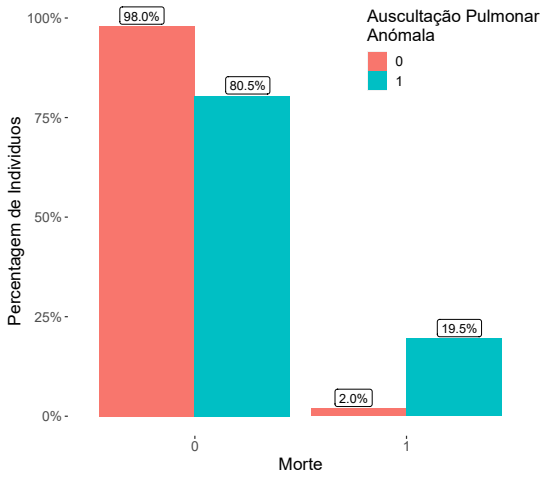
4.1.2.17 Auscultação Pulmonar Anómala

O gráfico da Figura 4.45a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma auscultação pulmonar anómala, sendo que apenas 0.9% dos indivíduos é que apresentam o sintoma.

No gráfico da Figura 4.45b, tem-se representado a distribuição dos óbitos pelo sintoma auscultação pulmonar anómala. Da sua análise, observa-se que a percentagem de indivíduos que apresentam o sintoma auscultação pulmonar anómala e que acabaram por falecer é bastante superior quando comparado com o seu valor homólogo para os indivíduos que não apresentam o sintoma em causa.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma auscultação pulmonar anómala.



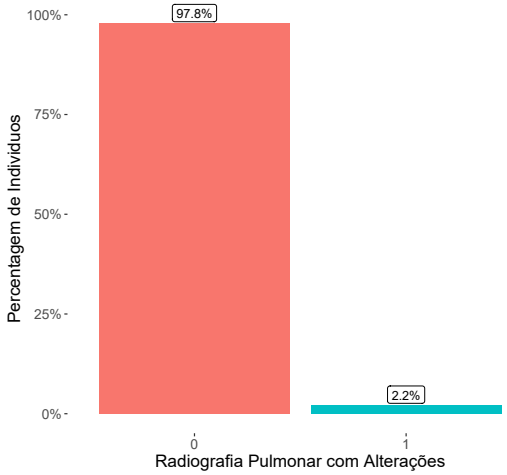
(b) Distribuição dos óbitos pelo sintoma auscultação pulmonar anómala.

Figura 4.45: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

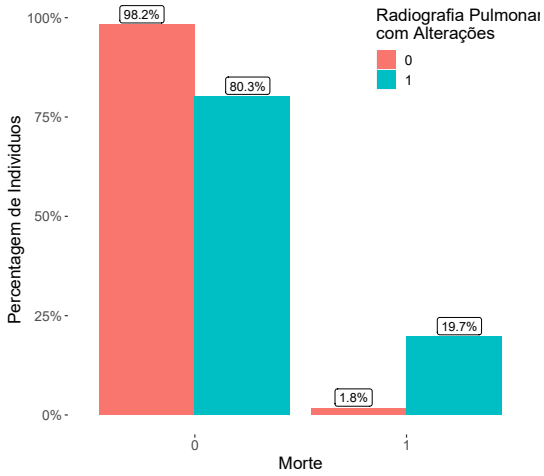
4.1.2.18 Radiografia Pulmonar com Alterações

O gráfico da Figura 4.46a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma radiografia pulmonar com alterações, sendo que cerca de 2.2% dos indivíduos é que apresentam o sintoma.

No gráfico da Figura 4.46b, tem-se representado a distribuição dos óbitos pelo sintoma radiografia pulmonar com alterações. Da sua análise, concluí-se que existe uma diferença significativa entre a proporção de indivíduos que apresentam o sintoma radiografia pulmonar com alterações e que acabaram por falecer e a proporção de indivíduos que não apresentam o sintoma em causa e que acabaram por falecer, pois, uma vez que, dos indivíduos que apresentam o sintoma em causa, cerca de 20% acabaram por falecer, enquanto que dos indivíduos que não apresentam o sintoma em estudo, somente cerca de 2% é que acabaram por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma radiografia pulmonar com alterações.



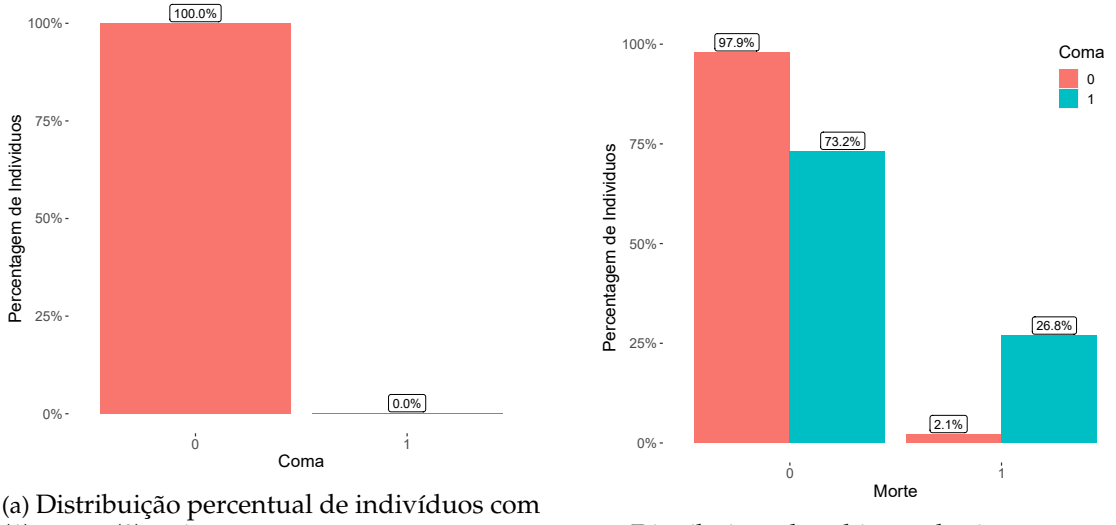
(b) Distribuição dos óbitos pelo sintoma radiografia pulmonar com alterações.

Figura 4.46: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.1.2.19 Coma

O gráfico da Figura 4.47a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma coma, sendo que praticamente nenhum dos indivíduos apresenta o sintoma em causa.

No gráfico da Figura 4.47b, tem-se representado a distribuição dos óbitos pelo sintoma coma. Da sua análise, concluí-se que existe uma diferença significativa entre a proporção de indivíduos que apresentam o sintoma coma e que acabaram por falecer e a proporção de indivíduos que não apresentam o sintoma em causa e que não acabaram por falecer, pois, uma vez que, dos indivíduos que apresentam o sintoma em causa, cerca de 27% acabaram por falecer, enquanto que dos indivíduos que não apresentam o sintoma em estudo, somente cerca de 2% é que acabaram por falecer.



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma coma.

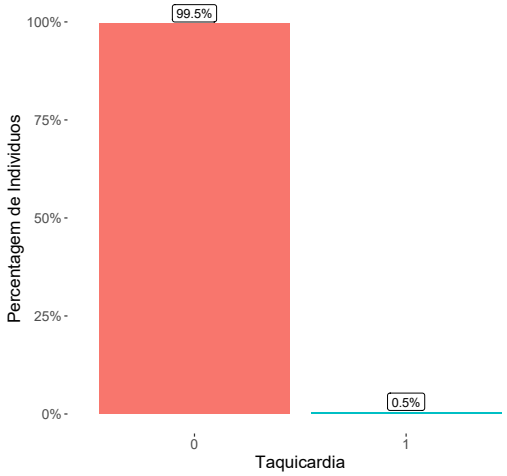
(b) Distribuição dos óbitos pelo sintoma coma.

Figura 4.47: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

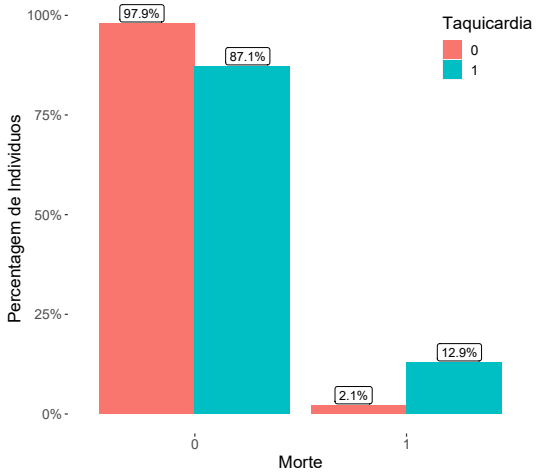
4.1.2.20 Taquicardia

O gráfico da Figura 4.48a apresenta distribuição percentual dos indivíduos da base de dados com o sintoma taquicardia, sendo que penas 0.5% dos indivíduos é que apresentam o sintoma em causa.

No gráfico da Figura 4.48b, tem-se representado a distribuição dos óbitos pelo sintoma coma. Da sua análise, observa-se que existe uma diferença considerável entre os indivíduos que acabaram por falecer e que não apresentam o sintoma taquicardia, com os indivíduos que acabaram por falecer e que apresentam o sintoma taquicardia, sendo que dos indivíduos que não apresentam o sintoma em causa, cerca de 2% é que acabaram por falecer, enquanto que dos indivíduos que apresentam o sintoma taquicardia, cerca de 13% é que acabaram por falecer. .



(a) Distribuição percentual de indivíduos com (1) e sem (0) o sintoma taquicardia.



(b) Distribuição dos óbitos pelo sintoma taquicardia.

Figura 4.48: Distribuição por sintoma (a) e distribuição dos óbitos por sintoma (b).

4.2 Limpeza da Base de Dados

Na secção 4.1, constatou-se que a presente base de dados tinha diversos problemas, no que diz respeito ao preenchimento dos dados, como se pode observar nas tabelas 4.8, 4.9, 4.10, 4.11, 4.12, e 4.13.

Uma vez que as variáveis referentes ao vírus que se encontram na Tabela 4.13 têm uma percentagem muito elevada de valores omissos cerca de 99%, optou-se por remover estas variáveis da base de dados. Apesar da variável *analise_1*, ter apenas cerca de 7% dos seus valores omissos, decidiu-se remover esta variável, uma vez que esta variável é referente ao tipo de teste, como por exemplo, o teste PCR ou Antigénio e, neste trabalho apenas se está interessado em estudar quais os sintomas e quais as comorbilidades que se relacionam com a variável *morte* por COVID-19.

Do mesmo modo e pelas mesmas razões que se eliminou as variáveis referentes aos vírus, também se optou por fazer o mesmo para a variável *qual_o_pais* (país de origem do paciente, quando esteve fora de Portugal), que tem cerca de 1% dos dados preenchidos, conforme se pode observar na Tabela 4.8.

Também se optou por remover a variável referente à data do início dos sintomas, que se encontra na Tabela 4.10, pois esta variável tem cerca de 65% dos valores omissos. Adicionalmente se compararmos esta variável com a variável referente à data de confirmação, o que se observa é uma translação no tempo entre as duas datas, ou seja, tem-se que a data de confirmação reflete a data do início dos sintomas, mas com um determinado desfasamento no tempo, como se pode observar na Figura 4.49. No entanto a variável *data_confirmado* foi convertida numa nova variável de tempo *data_confirmado1*, que representa a diferença entre o número de dias da data do primeiro caso e a data de confirmação em questão.

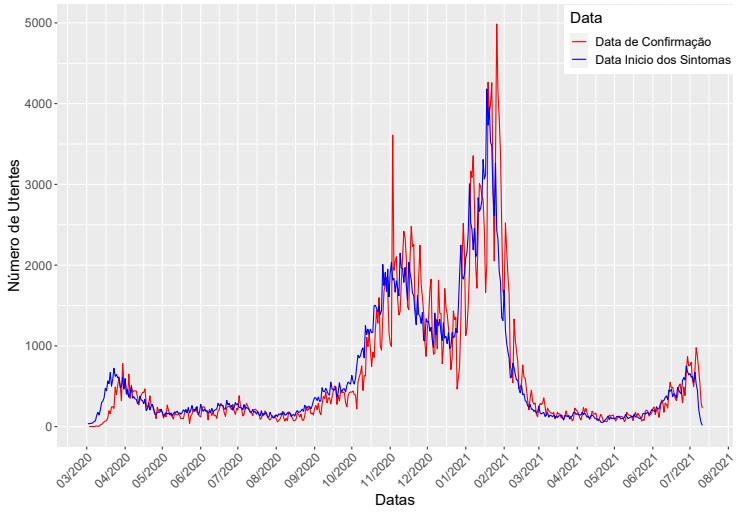


Figura 4.49: Data início dos sintomas *vs* data de confirmação.

Dado que no presente estudo não se está interessado em modelar o número de infeções/casos ao longo do tempo, isto é, não se está interessado em modelar o número de infeções de um determinado dia do ano e, como referido anteriormente a data de

confirmação reflete a data do início dos sintomas, estando apenas esta desfasado durante um período de tempo da data de confirmação, não se levou em conta neste estudo a variável data início dos sintomas.

A variável *durante_o_periodo_de_incubacao_esteve_fora_de_portugal* que se encontra na Tabela 4.8, foi removida, apesar desta variável estar em associação com a variável morte, uma vez que o valor do $p\text{-}value$ do teste χ^2, é de 3.53×10^{-19}, pelo que a um nível de significância de 5% existem evidências estatísticas de rejeitar a hipótese nula da independência.

Por outro lado, também se ajustou um modelo de Regressão Logística, em que se tentou explicar a variável resposta *morte* com a variável explicativa *durante_o_periodo_de_incubacao_esteve_fora_de_portugal*. Depois de ajustado o modelo, testou-se a hipótese nula representada em 4.1, recorrendo ao teste de *Wald*, obteve-se um valor do $p\text{-}value$ de 1.04×10^{-15}, pelo que a um nível de significância de 5% existem evidências estatística que nos permitem rejeitar a hipótese nula, pelo que a um nível de significância de 5%, é plausível assumir que o coeficiente da variável explicativa *durante_o_periodo_de_incubacao_esteve_fora_de_portugal* é diferente de 0.

$$\begin{cases} H_0 : \beta_{durante\_o\_periodo\_de\_incubacao\_esteve\_fora\_de\_portugal} = 0 \\ H_1 : \beta_{durante\_o\_periodo\_de\_incubacao\_esteve\_fora\_de\_portugal} \neq 0 \end{cases} \tag{4.1}$$

Ou seja, tem-se que a variável explicativa *durante_o_periodo_de_incubacao_esteve_fora_de_portugal* explica alguma parte das mortes, sem no entanto sabermos o país onde o indivíduo se encontrava. Apesar das conclusões tiradas anteriormente, não nos podemos abstrair de só se ter um número de observações comuns na variável morte e na variável *durante_o_periodo_de_incubacoo_esteve_fora_de_portugal* de apenas de cerca de 25%, como se observa na Tabela 4.8. Assim, quando se utiliza o teste do χ^2 e também quando se constrói o modelo e se utiliza o teste *Wald* para se testar se o coeficiente de regressão é nulo, apenas se está a trabalhar sobre um máximo de cerca de 25% das observações, pelo que se pode estar a criar um padrão que não existe na realidade. Logo, dada a quantidade de valores omissos da variável *durante_o_periodo_de_incubacao_esteve_fora_de_portugal*, decidiu-se remover esta variável no presente estudo.

Também se removeu a variável *pais_nacionalidade_utente*, apesar de existirem evidências estatísticas de que esta variável está em associação com a variável *morte*, uma vez que o valor do $p\text{-}value$ do teste do χ^2 é de 0.0005. A decisão de se remover esta variável deve-se ao facto de não se estar interessado em modelar qual a nacionalidade do utente que está mais associado à morte por COVID-19. Por outro lado, das observações desta variável que não tinham valores omissos, temos que cerca de 92% destas observações são de indivíduos que têm a nacionalidade Portuguesa, sendo que dos indivíduos que acabaram por falecer cerca de 98% destes tinham nacionalidade Portuguesa, daí venha também o facto de se rejeitar a hipótese nula da independência do teste do χ^2, pois uma vez que grande parte das observações pertencerem a uma única classe da nacionalidade.

As variáveis que dizem respeito à morada do utente, quer à descrição (variável *descricao_concelho_morada_utente* - nome do concelho) quer ao código da morada do utente (variável *codigo_concelho_morada_utente* - código do concelho), também foram removidas da presente base de dados. O mesmo também foi efetuado para as variáveis que dizem respeito ao concelho de ocorrência, quer à descrição do concelho (variável *descricao_concelho_ocorrencia* - nome do concelho de ocorrência) quer ao código do concelho (variável *codigo_concelho_ocorrencia* - código do concelho de ocorrência). No entanto, estas 4 variáveis da base de dados, foram convertidas em 4 novas variáveis, sendo duas delas sobre o distrito de ocorrência e as restantes duas sobre o distrito de morada do utente, ou seja, as variáveis *descricao_concelho_ocorrencia* e *codigo_concelho_ocorrencia* foram convertidas em *descricao_distrito_ocorrencia* e *codigo_descricao_ocorrencia*, respetivamente, enquanto que as variáveis *descricao_concelho_morada_utente* e *codigo_concelho_morada_utente* foram convertidas em *descricao_distrito_morada_utente* e *codigo_descricao_morada_utente*, respetivamente. O motivo, pelo qual se converteu os concelhos em distritos, deve-se ao facto da diminuição dos níveis de categorias nestas variáveis, uma vez que nas variáveis referentes aos concelhos temos 309 categorias enquanto que nos dados referentes aos distritos somente temos 20 categorias. No entanto, após a conversão dos concelhos em distritos, optou-se somente por levar em conta as variáveis referentes aos distritos de ocorrência, uma vez que grande parte dos utentes têm a descrição da morada de distrito igual à descrição do distrito de ocorrência, sendo este valor cerca de 98%. Este facto pode-se dever a deslocações ou ao seu local de trabalho. Ou seja, temos que praticamente todos os indivíduos que moram num dado distrito tiveram o caso de ocorrência no seu distrito de morada, pelo que utilizar as variáveis referentes à morada do utente em distrito é praticamente o mesmo que considerar o distrito de ocorrência. Posto isto, apenas se optou por considerar as variáveis *descricao_distrito_ocorrencia* e *codigo_descricao_ocorrencia*, uma vez que só estamos interessados no local de ocorrência, apesar destas apresentarem um maior número de valores omissos quando comparado com os concelhos. No entanto, também se observa que grande parte das variáveis que se encontram na base de dados têm cerca de 41% dos valores omissos, pelo que quando removermos essas observações fica-se com uma base de dados de tamanho inferior a 60% do tamanho da base de dados original. Também se verificou que todas as observações omissas correspondentes às variáveis dos distrito de ocorrência correspondiam exatamente às mesmas observações nas variáveis explicativas que têm também os 40.9% de valores omissos.

A variável correspondente à apresentação ou não de comorbilidades de um dado indivíduo que se encontra na Tabela 4.12 foi removida, uma vez que esta variável tem cerca de 76% dos valores omissos. Por outro lado, esta variável pode ter problemas de colinearidade com as demais variáveis referentes às comorbilidades apresentadas na Tabela 4.12, uma vez que esta variável está em associação com as restantes variáveis referentes às comorbilidades. Tal pode ser observado na Tabela 4.14, referente ao teste do χ^2 entre a variável *o_doente_apresenta_comorbilidades* e cada uma das restantes variáveis referentes às comorbilidades. Como referido anteriormente, tem-se que a variável

o_doente_apresenta_comorbilidades, está em associação com cada uma das variáveis restantes das comorbilidades, uma vez que, para cada um dos teste do χ^2 entre a variável *o_doente_apresenta_comorbilidades* e cada uma das restantes variáveis das comorbilidades rejeita-se a hipótese nula, da independência entre as duas variáveis em análise.

Tabela 4.14: Teste χ^2 entre a variável *o_doente_apresenta_comorbilidades* e cada uma das variáveis das comorbilidades.

Variável	Estatística	valor do $p\text{-}value$	Decisão
doenca_neurologica_ou _neuromuscular_cronica	4841.98	0.00	Rejeitar H_0
neoplasia	16337.80	0.00	Rejeitar H_0
vih_outras_imunodeficiencia	3058.11	0.00	Rejeitar H_0
diabetes	40644.95	0.00	Rejeitar H_0
doenca_pulmonar_cronica	12599.37	0.00	Rejeitar H_0
asma	20976.53	0.00	Rejeitar H_0
doenças_hematologicas_cronicas	4257.80	0.00	Rejeitar H_0
patologia_hepatica	3395.21	0.00	Rejeitar H_0
deficiencia_neurologica_cronica	3841.69	0.00	Rejeitar H_0
doenca_renal_cronica	9800.46	0.00	Rejeitar H_0
insuficiência_renal_aguda	447.84	0.00	Rejeitar H_0
insuficiencia_cardiaca	349.51	0.00	Rejeitar H_0
coagulopatia_de_consumo	25.73	0.00	Rejeitar H_0

Do mesmo modo e pela mesma razão, também se optou por eliminar a variável *apresentacao_da_doenca*, que representa se um dado indivíduo é assintomático ou se é sintomático. Uma das razões que nos levou à exclusão desta variável foi o facto de esta variável ter cerca de 60% dos valores omissos. Para além disso, também se efetuou um teste do χ^2 entre a a variável *apresentacao_da_doenca* e cada uma das restantes variáveis referentes aos sintomas que se encontram na Tabela 4.11, encontrando-se os resultados obtidos do teste do χ^2 e as suas conclusões encontram-se na Tabela 4.15.

Da análise da Tabela 4.15, pode-se concluir que cada uma das variáveis referentes aos sintomas está em associação com a variável *apresentacao_da_doenca*, uma vez que, em cada um dos testes do χ^2 entre a *apresentacao_da_doenca* e cada uma das variáveis dos sintomas restantes, a um nível de significância de 5%, existem evidências estatísticas que nos permitem rejeitar a hipótese nula da independência, basta observar que os valores do $p\text{-}value$ em cada um dos teste está abaixo do nível de significância de 5%. Também se ajusta um modelo de Regressão Logística em que a variável resposta é a variável *apresentacao_da_doenca* e as variáveis explicativas são as restantes variáveis referentes aos sintomas que se encontram na Tabela 4.11, para se testar se todos os coeficientes das variáveis explicativas eram diferentes de 0, no seu conjunto. Através do teste de razão de verosimilhança, obteve-se um valor do $p\text{-}value$ de 2.2×10^{-10}, pelo que a um nível de significância de 5%, existem evidências estatísticas que nos permitem rejeitar a hipótese nula, de que todos os coeficientes das variáveis explicativas são exatamente igual a 0. Logo, a um nível de significância de 5% é plausível assumir que existem alguns coeficientes

Tabela 4.15: Teste χ^2 entre a variável *apresentacao_da_doenca* e cada uma das variáveis dos sintomas.

Variável	Estatística	valor do *p-value*	Decisão
historia_de_febre_ou_calafrios	46178.16	0.00	Rejeitar H_0
pneumonia	1401.29	0.00	Rejeitar H_0
tosse_seca_ou_produtiva	80204.75	0.00	Rejeitar H_0
dispneia	9983.32	0.00	Rejeitar H_0
coriza	24776.13	0.00	Rejeitar H_0
odinofagia	20259.28	0.00	Rejeitar H_0
cefaleia	43562.81	0.00	Rejeitar H_0
dor_abdominal	2758.95	0.00	Rejeitar H_0
dor_no_peito	4675.25	0.00	Rejeitar H_0
artralgia	2935.19	0.00	Rejeitar H_0
mialgias	47164.57	0.00	Rejeitar H_0
nauseas_vomitos	6272.41	0.00	Rejeitar H_0
diarreia	8638.83	0.00	Rejeitar H_0
convulsoes	22.74	0.00	Rejeitar H_0
irritabilidade_confusao	272.78	0.00	Rejeitar H_0
fraqueza_geral_ou_astenia	21364.04	0.00	Rejeitar H_0
auscultacao_pulmonar_anomala	1244.08	0.00	Rejeitar H_0
radiografia_pulmonar_com_alteracoes	3134.06	0.00	Rejeitar H_0
coma	18.39	0.00	Rejeitar H_0
taquicardia	555.36	0.00	Rejeitar H_0

diferentes de zero, pelo que a variável *apresentacao_da_doenca* pode ser explicada através das restantes variáveis referentes aos sintomas que se encontram na Tabela 4.11. Assim, pelas conclusões que se obtiveram através do teste do χ^2 e também do teste da razão de verosimilhança do modelo ajustado, optou-se por excluir a variável *apresentacao_da_doenca*, uma vez que poderíamos ter um problema de multicolinearidade.

Também se optou por eliminar a variável *regra_de_confirmação*, uma vez que esta diz respeito ao tipo de confirmação, isto é, se esta foi feita através de notificação médica, por notificação laboratorial ou se por caso. No entanto efetuou-se o teste χ^2 entre a variável *regra_de_confirmação* e a variável *morte*, pelo que a um nível de significância de 5% o resultado do teste remete para a rejeição da hipótese nula, uma vez que o valor do *p-value* do teste χ^2 é de 0.000, ou seja, existem evidências estatísticas que nos permitem rejeitar a hipótese nula, pelo que a um nível de significância de 5%, é plausível assumir que a variável *regra_de_confirmação* está em associação com a variável *morte*. Apesar do teste do χ^2 nos inferir que a variável *regra_de_confirmação* e a variável *morte* estão em associação, optou-se por eliminar a variável *regra_de_confirmação*, uma vez que não se está interessado no presente trabalho em estudar se o tipo de confirmação explica a morte de um indivíduo com COVID-19, mas sim em ver quais os sintomas e quais as comorbilidades estão mais relacionadas com a variável *morte*.

Posto isto, apenas se ficaram com as variáveis na base de dados descritas nas Tabelas 4.16, 4.17, 4.18, 4.19 e 4.20, apesar de algumas destas ainda apresentarem um elevado

número de valores omissos, principalmente as variáveis que se referem aos sintomas.

Tabela 4.16: Variáveis referentes aos sintomas que ficaram na base de dados.

Variável	% de NA's	Variável	% de NA's
historia_de_febre_ou_calafrios	74.85	*dispneia*	81.28
tosse_seca_ou_produtiva	72.54	*coriza*	79.81
odinofagia	80.53	*cefaleia*	77.43
dor_abdominal	83.95	*artralgia*	40.90
dor_no_peito	83.62	*mialgias*	76.98
nauseas_vomitos	40.90	*diarreia*	40.90
irritabilidade_confusao	40.90	*taquicardia*	40.90
fraqueza_geral_ou_astneia	81.74	*coma*	40.90
auscultacao_pulmonar_anomala	40.90	*convulsoes*	40.90
radiografia_pulmonar_com_alteracoes	40.90	*pneumonia*	93.07

Tabela 4.17: Variáveis referentes às comorbilidades que ficaram na base de dados.

Variável	% de NA's
doenca_neurologica_ou_neuromuscular_cronica	40.90
neoplasia	40.90
vih_outras_imunodeficiencias	40.90
diabetes	40.90
doenca_pulmonar_cronica	40.90
asma	40.90
doenca_hematologicas_cronicas	40.90
patologia_hepatica	40.90
deficiencia_neurologica_cronicas	40.90
doenca_renal_cronica	40.90
insuficiencia_renal_aguda	40.90
insuficiencia_cardiaca	40.90
coagulopatia_de_consumo	40.90

Tabela 4.18: Variáveis referentes às datas que ficaram na base de dados.

Variável	% de NA's
data_confirmado1	0.01

Tabela 4.19: Variáveis referentes aos dados geográficos que ficaram na base de dados.

Variável	% de NA's
codigo_distrito_ocorrencia - Código do Distrito	2.64
descricao_distrio_ocorrencia - Nome do Distrito	2.64

Tabela 4.20: Variáveis referentes aos utente que ficaram na base de dados.

Variável	% de NA's	Variável	% de NA's
morte	0.00	*sexo_utente*	0.00
idade_utente_a_data_validacao	7.81		

Uma vez que algumas das variáveis referentes aos sintomas têm cerca de 41% dos valores omissos e dado que as observações que têm o preenchimento nos sintomas são as mesmas observações que têm também o preenchimento nas comorbilidades, optou-se por fazer uma imputação nos dados referentes aos sintomas. A imputação somente foi feita nas variáveis dos sintomas que têm mais de 50% dos valores omissos, onde esta imputação consiste em substituir os valores omissos (que não estão preenchidos) pelo valor 0, uma vez que se supôs que para um dado indivíduo se um dado sintoma não se encontra preenchido é porque o indivíduo não apresenta esse sintoma.

Feita essa imputação eliminaram-se todos os valores omissos, isto é, excluíram-se todas as observações em que estas tinham pelo menos um valor omisso em alguma variável. Posto isto, das 909720 observações iniciais apenas se ficou com 466458 observações, sendo que esta nova base de dados iremos denominar por base de dados restrita.

Na base de dados restrita que se acabou de obter, verificou-se que para cada uma das comorbilidades a distribuição percentual de indivíduos com e sem a respetiva comorbilidade, manteve-se idêntica à distribuição percentual de indivíduos com e sem essa comorbilidade na base de dados original. Para a distribuição dos óbitos pela comorbilidade também se obteve a mesma conclusão, ou seja, a distribuição dos óbitos de uma dada comorbilidade na base de dados restrita é idêntica à distribuição dos óbitos dessa mesma comorbilidade na base de dados original. Já nos sintomas, verificou-se que praticamente em todos os sintomas a distribuição percentual de indivíduos com e sem um dado sintoma se manteve igual à distribuição percentual de indivíduos com e sem esse mesmo sintoma na base de dados original, sendo que apenas se obteve valores mais discrepantes nos sintomas mialgias, história de febre ou calafrios, tosse seca ou produtiva e cefaleia. Consequentemente, quando se analisa a distribuição dos óbitos por cada um dos sintomas na base de dados restrita, concluiu-se que a distribuição dos óbitos por cada um dos sintomas se manteve a mesma face à distribuição dos óbitos por cada um dos sintomas na base de dados original, que é o que nos interessa, dado que o nosso objetivo é modelar as mortes através tanto dos sintomas como das comorbilidades.

4.3 Base de Dados Restrita

A base de dados restrita obtida anteriormente, foi dividida em duas bases dados, a base de dados restrita das comorbilidades e a base de dados restrita dos sintomas. A base de dados restrita dos sintomas é constituída pelas seguintes variáveis, a variável resposta *morte* e as seguintes variáveis explicativas, a variável *idade_utente_a_data_validacao*

, a variável *sexo_utente*, a variável *data_confirmado1*, a variável *descricao_distrito_ocorrencia* e todas as variáveis referentes aos sintomas, que se encontram na Tabela 4.16. Já a base de dados restrita das comorbilidades é constituída pelas seguintes variáveis, a variável resposta *morte* e as seguintes variáveis explicativas, a variável *idade_utente_a_data_validacao* , a variável *sexo_utente*, a variável *data_confirmado1*, a variável *descricao_distrito_ocorrencia* e todas as variáveis referentes às comorbilidades, que se encontram na Tabela 4.17.

O motivo pelo qual se divide a base de dados restrita em duas diferentes bases de dados, uma base de dados restrita dos sintomas e a outra a base de dados restrita das comorbilidades é pelo facto de neste trabalho se estar interessado em analisar o poder preditivo dos modelos de Regressão Logística, de Regressão Logística com interações, do LASSO de Grupo, do modelo Aditivo Generalizado, das Árvores de Classificação, da Floresta Aleatória e das Redes Neuronais *feed-forward*, consoante os sintomas e as comorbilidades que os indivíduos possam apresentar, sendo que se decidiu não misturar sintomas e comorbilidades, uma vez que, talvez possam existir sintomas que estejam relacionadas com algumas das comorbilidades, pelo que poderíamos ter problemas de multicolinearidade nos diversos modelos. Assim, uma forma de mitigar este problema é separar os dados em dados referentes aos sintomas e dados referentes às comorbilidades.

Os modelos de Regressão Logística, de Regressão Logística com interações e o modelo Aditivo Generalizado são modelos que tanto podem ser utilizados para estimação, como também para classificação, enquanto que os modelos de LASSO de Grupo, das Árvores de Classificação, da Floresta Aleatória e das Redes Neuronais são utilizados mais frequentemente para classificação [12].

Assim, nos modelos de Regressão Logística, de Regressão Logística com interações e do modelo Aditivo Generalizado, também iremos analisar a estimação conseguida por cada um destes modelos.

4.3.1 Dados Comorbilidades

Antes de aplicar os diferentes modelos aos dados das doenças, elaborou-se o teste do χ^2, a Regressão Logística múltipla com o *stepwise* e a Regressão do LASSO de Grupo, de forma a reduzir o número das variáveis explicativas que se irão levar para análise nos diversos modelos. O teste χ^2 apenas foi aplicado entre a variável resposta *morte* e cada uma das variáveis explicativas à exceção da variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *data_confirmado1*, uma vez que estas duas variáveis são contínuas. A Tabela 4.21, mostra quais as variáveis que cada um dos 3 métodos elimina.

Através da análise da Tabela 4.21, observa-se que apenas o método do *stepwise* aplicado na Regressão Logística elimina uma variável explicativa, a *coagulopatia_de_consumo*. O critério utilizado para eliminar uma dada variável explicativa na análise subsequente, é que pelo menos dois dos métodos descritos anteriormente eliminassem a mesma variável. Dado que neste caso, nenhum dos métodos elimina a mesma variável, tem-se que nenhuma das variáveis foi removida. Por isto, tem-se que a base de dados restrita correspondente

Tabela 4.21: Variáveis eliminadas e não eliminadas referentes às comorbilidades ("X" corresponde que a variável é eliminada pelo respetivo método e "−" que a variável não é eliminada, pelo respetivo método).

Variável	**Regressão Logística com o *stepwise***	**Lasso de Grupo**	**Teste χ^2**
idade_utente_a_data_validacao	–	–	–
sexo_utente	–	–	–
data_confirmado1	–	–	–
descricao_distrito_ocorrencia	–	–	–
doenca_neurologica_ou_neuromuscular_cronica	–	–	–
neoplasia	–	–	–
vih_outras_imunodeficiencias	–	–	–
diabetes	–	–	–
doenca_pulmonar_cronica	–	–	–
asma	–	–	–
doenca_hematologicas_cronicas	–	–	–
patologia_hepatica	–	–	–
deficiencia_neurologica_cronica	–	–	–
doenca_renal_cronica	–	–	–
insuficiencia_renal_aguda	–	–	–
insuficiencia_cardiaca	–	–	–
coagulopatia_de_consumo	X	–	–

às comorbilidades é constituída por todas as variáveis explicativas que se encontram na Tabela 4.21.

Na secção 3.10, referiu-se quais eram os problemas de classificação numa base de dados desequilibrada/desbalanceada, sendo que um dos problemas que se referiu é que muitas das métricas descritas na secção 3.9 costumam dar-nos conclusões erróneas sobre a avaliação do modelo construído.

A base de dados restrita das comorbilidades não é equilibrada, ou seja, é desbalanceada, uma vez que existe uma grande diferença entre as duas categorias da variável resposta, pois a categoria que representa o não falecimento de um indivíduo têm cerca de 98% das observações, enquanto que a categoria que representa o falecimento de um indivíduo têm cerca de 2% das observações da presente base de dados. Assim, foi necessário equilibrar/balancear a presente base de dados, de modo a treinar e a testar os diferentes modelos para avaliar cada um destes em termos preditivos de uma forma correta. Para se balancear a base de dados, utilizou-se o método da subamostragem aleatória, que elimina observações da classe maioritária, em que neste caso a classe maioritária é a classe que representa o não falecimento de um indivíduo. O motivo pelo qual se considera um método de subamostragem em vez de um método de sobreamostragem, é que os métodos de sobreamostragem efetuam o balanceamento de uma base de dados através da sobreamostragem de observações que pertencem à classe minoritária no caso do método

da sobreamostragem aleatória, enquanto que na técnica SMOTE as novas observações da classe minoritária não são apenas cópias das observações que pertencem à classe minoritária como se faz no método da sobreamostragem aleatória. Na técnica SMOTE o algoritmo gera novas observações sintéticas para a classe minoritária como se referiu na sub-subsecção 3.10.1.1. No entanto, em ambas as técnicas de sobreamostragem referidos, passaríamos a ter uma base de dados onde as observações deixariam de ser independentes, o que para muitos dos modelos considerados um dos seus pressupostos é a independência das observações. Posto isto, construiu-se uma nova base de dados, em que a iremos denotar por base de dados balanceada das comorbilidades, onde esta tem 20470 observações, da base de dados restrita das comorbilidades.

Para se treinar e testar cada um dos modelos, de forma a averiguar a capacidade preditiva de cada um destes e compara-los em termos preditivos entre si, utilizou-se o método *holdout* para dividir a base de dados balanceada das comorbilidades em base de dados de treino e em base de dados de teste. Neste caso, a base de dados de treino é constituída por 80% das observações da base de dados balanceada das comorbilidades enquanto que a base de dados referente ao teste é constituída pelas restantes 20%. De referir que a base de dados balanceada das comorbilidades têm as mesmas variáveis que a base de dados restrita das comorbilidades.

Nas sub-subsecções 4.3.1.1, 4.3.1.2, 4.3.1.3, 4.3.1.4, 4.3.1.5, 4.3.1.6 e 4.3.1.7, encontram-se os resultados dos modelos de Regressão Logística, Regressão Logística com interações, o LASSO de Grupo, modelo Aditivo Generalizado, Árvores de Classificação, Floresta Aleatória e Redes Neuronais *feed-forward*, respetivamente, onde se encontram os modelos finais e os parâmetros utilizados para se treinar cada um dos modelos com a base de dados de treino das comorbilidades.

Como referido anteriormente os modelos de Regressão Logística, de Regressão Logística com interações e o modelo Aditivo Generalizado, são muitas vezes utilizados para estimar os coeficientes das variáveis explicativas, de forma a averiguar a sua significância, para aferir se as variáveis explicativas estão ou não relacionadas com a variável resposta, pelo que nas sub-subsecções referentes a estes modelos também se apresentam os resultados de estimação destes modelos, assim como os diferentes testes à significância dos seus coeficientes.

4.3.1.1 Regressão Logística

Antes de se treinar o modelo de Regressão Logística com os dados de treino, ajustou-se um modelo de Regressão Logística com o *stepwise* aos dados restritos das comorbilidades (não balanceada), para se selecionar qual o melhor subconjunto das variáveis explicativas que se levaria em conta para o modelo de treino com a Regressão Logística. O motivo pelo qual o subconjunto das variáveis explicativas consideradas no modelo de treino serem as variáveis explicativas selecionadas a partir do método do *stepwise* e dos referidos testes de significâncias, a partir do ajustamento do modelo de Regressão Logística ajustado

aos dados restritos das comorbilidades não balanceados, é o simples facto da base de dados balanceada das comorbilidades ter sido obtida pelo método da subamostragem aleatória, que como referido anteriormente elimina de forma aleatória observações da classe maioritária. Uma das desvantagens deste método é o facto de poder levar à perda de informações úteis ao remover padrões significativos nos dados, uma vez que grande parte das observações referentes à classe maioritária foi descartada. Assim, se aplicássemos o método do *stepwise* à base de dados balanceada das comorbilidades, este método poderia estar a eliminar variáveis que seriam importantes para a modelação. Uma forma de mitigar a desvantagem do método da subamostragem aleatória para o balanceamento dos dados, é então selecionar as variáveis explicativas a partir dos dados desbalanceados e depois treinar o modelo com os dados de treino das comorbilidades, a partir das variáveis explicativas selecionadas desta forma.

Assim, começando com o ajustamento do modelo de Regressão Logística e aplicando o método do *stepwise* aos dados restritos das comorbilidades, através do *software* R, este método apenas remove a variável explicativa *coagulopatia_de_consumo*, conforme já se tinha visto anteriormente na Tabela 4.21.

Na Tabela I.1 do Anexo I, encontram-se os resultados obtidos com da Regressão Logística com o método do *stepwise*, aplicada na base de dados restrita das comorbilidades, onde se observa que apesar dos resultados do *stepwise*, as variáveis explicativas *asma*, *vih_outras_imunodeficiencias*, *data_confirmado1* e alguns dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, não são estatisticamente significantes a um nível de significância de 5%, pelo que se realizou o teste de razão de verosimilhança sobre a nulidade do conjunto dos coeficientes destas variáveis explicativas referidas anteriormente. Realizando o teste de razão de verosimilhança ao conjunto dos coeficientes das variáveis explicativas *asma*, *vih_outras_imunodeficiencias*, *data_confirmado1* e sobre os diferentes níveis da variável explicativas *descricao_distrito_ocorrencia*, obteve-se um valor observado da estatística de teste de 243.07 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes destas variáveis explicativas são conjuntamente identicamente igual a 0, pelo que não é plausível assumir que o conjunto dos coeficientes das variáveis explicativas *asma*, *vih_outras_imunodeficiencias*, *data_confirmado1* e dos diferentes níveis da variável explicativas *descricao_distrito_ocorrencia* sejam conjuntamente iguais a 0. Posto isto, realizou-se um teste de razão de verosimilhança sobre os diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, onde se obteve um valor observado da estatística de teste de 235.58 e um valor do $p\text{-}value$ de 2.2×10^{-16}, donde se conclui que existem evidências estatísticas que nos permitam rejeitar a hipótese nula a um nível de significância de 5%, pelo que não é plausível assumir que o conjunto dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* sejam conjuntamente iguais a 0. De seguida, realizou-se o teste de razão de verosimilhança sobre o conjunto dos coeficientes das variáveis explicativas *asma*, *vih_outras_imunodeficiencias* e *data_confirmado1*, onde se obteve um valor observado da estatística de teste de 8.69 e um valor do $p\text{-}value$

de 0.0336, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes das variáveis explicativas *asma*, *vih_outras_imunodeficiencias* e *data_confirmado1* sejam conjuntamente iguais a 0. Seguidamente, realizou-se o teste de razão de verosimilhança sobre o conjunto dos coeficientes das variáveis explicativas *asma* e *vih_outras_imunodeficiencias*, onde se obteve um valor observado da estatística de teste de 5.52 e um valor do $p\text{-}value$ de 0.0633, pelo que a um nível de significância de 5%, não existem evidências estatísticas que nos permitam rejeitar a hipótese nula de que o conjunto dos coeficientes das variáveis explicativas *asma* e *vih_outras_imunodeficiencias* sejam conjuntamente iguais a 0, pelo que é plausível assumir que o conjunto dos coeficientes das variáveis explicativas *asma* e *vih_outras_imunodeficiencias* são conjuntamente iguais a 0. Posto isto, removeu-se estas duas variáveis explicativas. Na Tabela I.2, do Anexo I, encontram-se os resultados obtidos através da Regressão Logística deste novo modelo. Da análise dos valores dos $p\text{-}values$ do teste de *Wald*, valores da coluna Pr(>|z|), na Tabela I.2, observa-se que todas as variáveis explicativas, à exceção da variável explicativa *data_confirmado1* e de alguns dos níveis da variável explicativa *descricao_distrito_ocorrencia*, são estatisticamente significantes a um nível de significância de 5%, pelo que se aplicou um teste de razão de verosimilhança, sobre os coeficientes da variável explicativa *data_confirmado1* e os diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*. Do teste de razão de verosimilhança, aplicado ao conjunto dos coeficientes da variável explicativa *data_confirmado1* e aos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, obteve-se um valor observado da estatística de teste de 237.55 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5%, existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes da variável explicativa *data_confirmado1* e dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são conjuntamente iguais a 0. Posto isto, aplicou-se o teste de razão de verosimilhança, sobre os coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, onde se obteve um valor observado da estatística de teste de 235.52 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5%, existem evidências estatísticas que nos permitem rejeitar a hipótese nula, de que o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são conjuntamente iguais a 0, ou seja, não é plausível assumir que o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são conjuntamente iguais a 0. De seguida, realizou-se um teste de razão de verosimilhança sobre a variável explicativa *data_confirmado1*, onde se obteve um valor observado da estatística de teste de 3.18 e um valor do $p\text{-}value$ de 0.07464, onde se conclui que a um nível de significância de 5%, não existem evidências estatísticas que nos permitam rejeitar a hipótese nula de que o coeficiente da variável explicativa *data_confirmado1* seja identicamente igual a 0, pelo que se removeu esta variável do presente modelo.

Assim, ajustou-se um novo modelo de Regressão Logística, onde não se levou em conta todas as variáveis anteriormente removidas. Na Tabela I.3, do Anexo I, encontram-se os

resultados obtidos com este novo modelo com a Regressão Logística.

Da análise dos valores do $p\text{-}values$ do teste de *Wald*, que são os valores da coluna Pr(>|z|) que se encontra na Tabela I.3, observa-se que todas as variáveis explicativas à exceção de alguns dos níveis da variável explicativa *descricao_distrito_ocorrencia*, são estatisticamente significantes a um nível de significância de 5%, pelo que se realizou um teste de razão de verosimilhança, sobre os coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, donde se obteve um valor observado da estatística de teste de 234.37 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5%, existem evidências estatísticas que nos permitem rejeitar a hipótese nula, de que o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são conjuntamente iguais a 0. Posto isto, não se removeu esta variável explicativa do presente modelo.

Do exposto, dado que não existem evidências estatísticas que nos permitem inferir que o conjunto dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são conjuntamente iguais a 0 e dado que todas as restantes variáveis explicativas do presente modelo são estatisticamente significantes a um nível de significância de 5%, então temos que o modelo final é o modelo é constituído por todas as variáveis explicativas da base de dados restrita das comorbilidades à exceção das variáveis explicativas *coagulopatia_de_consumo*, *asma*, *vih_outras_imunodeficiencias* e *data_confirmado1*. Este será o modelo de treino ajustado à base de dados de treino das comorbilidades, que é constituído por todas as variáveis explicativas da base de dados restrita das comorbilidades à exceção das variáveis explicativas *coagulopatia_de_consumo*, textitasma, *vih_outras_imunodeficiencias* e *data_confirmado1*.

Na Tabela I.4, apresentam-se os *OR* dos coeficientes das variáveis explicativas, que estão presentes no modelo final. A partir dos valores dos *OR*, pode-se tirar as seguintes conclusões:

- A *chance* de morte por COVID-19 estimada é maior 1.12 vezes para cada ano de aumento de idade;
- As *chances* de falecer devido à COVID-19 entre os indivíduos de sexo masculino é 2.12 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos do sexo feminino, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *doenca_renal_cronica* é 2.13 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doenca_renal_cronica*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *neoplasia* é 1.82 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *neoplasia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *diabetes* é 1.38 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *diabetes*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *deficiencia_neurologica_cronica* é 1.75 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *deficiencia_neurologica_cronica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *patologia_hepatica* é 2.51 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *patologia_hepatica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *doenca_pulmonar_cronica* é 1.49 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doenca_pulmonar_cronica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *insuficiencia_renal_aguda* é 3.31 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *insuficiencia_renal_aguda*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *doenca_neurologica_ou_neuromuscular_cronica* é 1.48 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doenca_neurologica_ou_neuromuscular_cronica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *doencas_hematologicas_cronicas* é 1.43 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doencas_hematologicas_cronicas*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com a comorbilidade *insuficiencia_cardiaca* é 1.57 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *insuficiencia_cardiaca*, neste presente estudo;

4.3.1.2 Regressão Logística com interações

No modelo de Regressão Logística construído na sub-subsecção 4.3.1.1, apenas se consideraram variáveis explicativas, sem interações. Neste caso, para a construção deste modelo, consideram-se as seguintes interações:

- A interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia*;
- Todas as interações entre a variável explicativa *idade_utente_a_data_validacao* e todas as variáveis explicativas referentes às comorbilidades;
- Todas as interações entre todas as variáveis explicativas referentes às doenças.

As duas primeiras interações consideradas têm como objetivo averiguar se existe uma relação entre variável explicativa *idade_utente_a_data_validacao* com os distritos, onde a população é mais idosa, assim como também se pretende analisar a influência da idade com cada uma das comorbilidades.

Para a seleção das variáveis explicativas para o modelo de treino, considerou-se mais uma vez as variáveis selecionadas a partir do método do *stepwise* e dos testes à significância dos coeficientes das demais variáveis explicativas, a partir do ajustamento aos dados da base de dados restrita das comorbilidades não balanceada. Assim, para se manter a melhor relação possível entre os dados utilizou-se a base de dados restrita das comorbilidades, sendo que depois de selecionadas as variáveis explicativas neste processo, ajusta-se o modelo final obtido aos dados de treino das comorbilidades, para se treinar o respetivo modelo. No entanto, neste caso, ao contrário do modelo de Regressão Logística construído na sub-subsecção 4.3.1.1, iremos considerar um nível de significância de 10% para o teste de *Wald*, devido ao elevado número de variáveis explicativas do modelo inicial, uma vez que se utilizássemos um nível de significância de 5%, o processo de ajustamento e de testagem dos coeficientes com o teste de razão de verosimilhança tornaria este processo muito exaustivo. No entanto, no teste de razão de verosimilhança considerou-se o nível habitual de significância de 5%.

Começando então com a construção do modelo e aplicando o *stepwise* na Regressão Logística aplicado na base de dados restrita e considerando as interações descritas anteriormente, o resultado final deste método encontra-se na Tabela I.5, do Anexo I.

Da Tabela I.5, pode-se observar que praticamente todas as variáveis são estatisticamente significantes, somente a interação *idade_utente_a_data_validacao:asma* é que não é estatisticamente significante a um nível de significância de 10%, para além disso, também existem algumas interações entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* que não são estatisticamente significantes a um nível de significância de 10%, como também alguns dos níveis da variável explicativa *descricao_distrito_ocorrencia* não são estatisticamente significantes a um nível de significância de 10%. Posto isto, aplicou-se um teste de razão de verosimilhança, aos coeficientes da variável explicativa *descricao_distrito_ocorrencia*, aos coeficientes das interações entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* e ao coeficiente da interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *asma*, obtendo-se um valor observado da estatística teste de 288.52 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% existem

evidências estatísticas que nos permitam rejeitar a hipótese nula de que o conjunto dos coeficientes sejam nulos, isto é, não existem evidências estatísticas que nos permitam referir que o conjunto dos coeficientes das variáveis explicativas *descricao_distrito_ocorrencia*, das interações entre *idade_utente_a_data_validacao* e a *descricao_distrito_ocorrencia* e a interação entre *idade_utente_a_data_validacao* e a *asma* sejam exatamente igual a 0. Assim, decidiu-se aplicar 3 testes de razão de verosimilhança, um ao conjunto dos coeficientes das variáveis explicativas *descricao_distrito_ocorrencia*, outro ao conjunto dos coeficientes das interações entre as variáveis explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* e outro ao coeficiente da interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *asma*.

Começando com o teste de razão de verosimilhança, aplicado aos coeficientes das interações entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia*, obtém-se um valor observado de observado da estatística de teste de 49.41 e um valor do $p\text{-}value$ de 0.0002, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula, isto é, a um nível de significância de 5% não existem evidências estatísticas de que o conjunto de coeficientes das variáveis explicativas correspondentes às interações entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* sejam 0. Logo, é plausível assumir que o conjunto dos coeficientes das interações entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* é diferente de 0.

Aplicando, o teste de razão de verosimilhança aos coeficientes da variável explicativa *descricao_distrito_ocorrencia*, obteve-se um valor observado da estatística de teste de 62.52 e um valor do $p\text{-}value$ de 1.527×10^{-6}, pelo que existem evidências estatística que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes da variável explicativa *descricao_distrito_ocorrencia* sejam exatamente iguais a 0. Logo, não existem evidências estatísticas que nos permitam afirmar que o conjunto dos coeficientes da variável explicativa *descricao_distrito_ocorrencia* sejam nulos. Por fim, aplicando o teste de razão de verosimilhança ao coeficiente da interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *asma*, obtém-se um valor observado da estatística de teste de 2.64 e um valor do $p\text{-}value$ de 0.1043, pelo que a um nível de significância habitual de 5%, não existem evidências estatísticas que nos permitam rejeitar a hipótese nula de que o coeficiente desta interação seja nula, ou seja, é plausível assumir que o coeficiente desta interação é exatamente igual a 0. Como a um nível de significância de 5% é plausível assumir que o coeficiente da interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *asma* é nulo, decidiu-se remover esta interação do modelo.

Assim, removendo esta interação do modelo, e ajustando o modelo sem essa interação, obtém-se o resultado que se encontra na Tabela I.6 do Anexo I, pelo que este é o modelo final quando considerado sobre a base de dados restrita das comorbilidades, pois tem-se que todas as variáveis explicativas são estatisticamente significantes a um nível de significância de 10%, exceto as variáveis explicativas correspondentes aos níveis da variável

explicativa *descricao_distrito_ocorrencia* e os coeficientes das interações entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia*. No entanto, quando se aplica o teste de razão de verosimilhança aos coeficientes dos níveis da variável explicativa *descricao_distrito_ocorrencia*, obtém-se um valor observado da estatística de teste de 62.6 e um valor do $p\text{-}value$ de 1.491×10^{-6}, pelo que a um nível de significância de 5% existem evidências que nos permitam rejeitar a hipótese nula de que o conjunto dos coeficientes da variável explicativa *descricao_distrito_ocorrencia* sejam exatamente iguais a 0. Aplicando o teste de razão de verosimilhança, da interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia*, obtém-se um valor observado da estatística de teste de 49.48 e um valor do $p\text{-}value$ de 0.0002, pelo que a um nível de significância de 5% não é plausível assumir que o conjunto dos coeficientes da interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* sejam iguais a 0. Do exposto resulta que não existem evidências estatísticas que nos permitam retirar a variável explicativa *descricao_distrito_ocorrencia* e a variável explicativa referente à interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia*.

Dado que não existem evidências estatísticas que nos permitam retirar a variável explicativa *descricao_distrito_ocorrencia* e a variável explicativa referente à interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia* e que os restantes coeficientes das restantes variáveis explicativas são estatisticamente significantes a um nível de significância de 10% para o teste de *Wald*, então temos que o modelo final é o modelo que tem as variáveis explicativas que se encontram na Tabela I.6 do Anexo I, onde esta diz respeito aos resultados obtidos deste novo modelo.

Posto isto, o modelo de treino leva em conta todas as variáveis explicativas que se encontram na Tabela I.6 do Anexo I, onde se ajusta este modelo aos dados de treino das comorbilidades, para se treinar o respetivo modelo.

4.3.1.3 LASSO de Grupo - Caso da Regressão Logística

O LASSO de Grupo foi apenas aplicada aos dados de treino, uma vez que este método tem o objetivo de encolher os coeficientes das variáveis explicativas em direção a zero, simplificando consequentemente o modelo e, quando aplicado em bases de dados de tamanho diferente, é claro que as estimativas dos coeficientes são diferentes, podendo em duas bases de dados de tamanho diferentes que contenham as mesmas variáveis explicativas, numa delas estimar um coeficiente duma variável explicativa como 0 e na outra base de dados estimar o coeficiente dessa mesma variável explicativa como diferente de 0.

Neste estudo utilizam-se variáveis explicativas mistas, isto é, tanto se tem variáveis categóricas como variáveis contínuas. Neste caso, as 17 variáveis explicativas que são

utilizadas para construir o modelo estão divididas em $G = 17$ grupos, em que cada um dos grupos é referente a cada uma das variáveis explicativas. No entanto, o número de variáveis no grupo da variável explicativa *descricao_distrito_ocorrencia*, é de $p_l = 19$ variáveis *dummy*, pois a variável *descricao_distrito_ocorrencia* é uma variável categórica com 20 categorias, em que cada uma destas categorias é um distrito. Todos os restantes restantes grupos são constituídos por uma única variável.

Para se ajustar o modelo aos dados de treino, utilizou-se a biblioteca *grpreg* do *software* R. No entanto, conforme visto na sub-subsecção 3.3.4.1 da secção 3.3, é necessário estimar o parâmetro de ajuste λ, para depois se estimar cada um dos coeficientes β. Para se estimar estes parâmetros, recorreu-se ao método da validação cruzada com 10 grupos, sendo utilizado o critério da *deviance* para a seleção do melhor parâmetro λ. Assim, utilizou-se a função *cv.grpreg()* da biblioteca *grpreg* para se obter o melhor valor de λ, em que este é o valor a que corresponder ao menor valor da *deviance* no final da validação cruzada com 10 grupos, isto é, o valor de λ para o qual a média da *deviance* em cada um dos 10 grupos da validação cruzada seja a menor de todos os parâmetros λ testados. Na função *cv.grpreg()* da presente biblioteca, o valor de λ que retorna o menor erro da validação cruzada com k grupos é dado pelo $lambda.min$, em que este é um valor que a função *cv.grpreg()* contém depois de ajustado o modelo, com os dados de treino.

O valor de λ que retorna o menor valor de erro da validação cruzada com 10 grupos, foi de $\lambda = 0.0011$. Com este valor de λ, apenas as variável *asma* e a variável *coagulopatia_de_consumo* é que têm os parâmetros dos seus coeficientes estimados como 0, tal como se observa na Tabela I.7, que se encontra no Anexo I, onde esta diz respeito à estimação dos coeficientes do LASSO de Grupo, no modelo de treino. Todos os restantes coeficientes das demais variáveis explicativas são diferentes de 0. Posto isto, já se tem um modelo treinado com os dados de treino, apenas nos falta averiguar se o modelo ajustada é bom ou não.

4.3.1.4 Modelo Aditivo Generalizado

O modelo Aditivo Generalizado foi utilizado, para se tentar captar a não linearidade do efeito das variáveis explicativas *idada_utente_a_data_validacao* e *data_confirmado1* ao longo do tempo. Neste presente caso, o modelo Aditivo Generalizado, tanto foi ajustado na base de dados restrita das comorbilidades como também na base de dados balanceada, uma vez que como se balanceou a base de dados restrita das comorbilidades, houve uma perda dos dados e possivelmente dos seus padrões, e assim para se averiguar se existe alguma concordância entre o grau escolhido dos *splines* de suavização utilizados nas variáveis explicativas *idada_utente_a_data_validacao* e *data_confirmado1*, assim como também nas variáveis explicativas selecionadas em cada um dos modelos, para depois se levar em conta para o modelo de treino, já que o modelo de treino vai ser ajustado numa base de dados bastante menor.

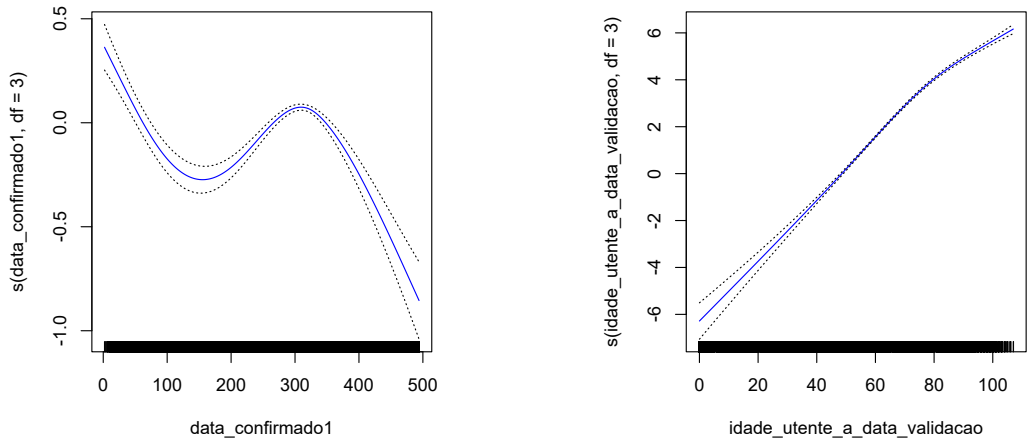
Assim, começou-se por ajustar diferentes modelos Aditivo Generalizado aos dados restritos das comorbilidades, sendo que se modelou as variáveis explicativas *data_confirmado1* e *idade_utente_data_validacao*, desde o grau de *spline* de suavização de ordem 2 até à ordem 3, como também se considerou os modelos onde a variável explicativa *idade_utente_data_validacao* é modelada linearmente e a variável explicativa *data_confirmado1* é modelada com um *spline* de suavização de grau 2 e 3, como também se considerou o seu oposto, ou seja, também se considerou as combinações onde a variável explicativa *data_confirmado1* é modelada linearmente e a variável explicativa *idade_utente_data_validacao* é modelada com um *spline* de suavização de grau 2 e 3. Para se selecionar o melhor modelo ajustado de entre de todos os modelos ajustados, utilizou-se a medida *AIC*, sendo que o melhor modelo escolhido é aquele que apresentar o menor valor de *AIC*. Na Tabela 4.22, encontram-se os *AIC* dos diversos modelos, onde se consideraram uma combinação dos graus dos *splines* de suavização nas variáveis explicativas *data_confirmado1* e *idade_utente_a_data_validacao*. Utilizou-se a função *gam()* da biblioteca *gam* do *software* R para se ajustar estes diversos modelos.

Tabela 4.22: Resultados do *AIC* dos diversos modelos Aditivos Generalizados ajustados na base de dados restrita das comorbilidades.

Grau do *spline*		
idade_utente_a_data_validacao	*data_confirmado1*	*AIC*
-	2	62506.26
2	-	62399.27
2	2	62253.53
-	3	62312.12
3	-	62345.9
3	2	62200.88
2	3	62063.17
3	3	62011.29

Através da análise da Tabela 4.22, conclui-se que o melhor modelo em termos do *AIC* é o modelo que tem grau 3 para o *spline* de ambas as variáveis explicativas, ou seja, o melhor modelo em termos do *AIC* é o modelo que tem um *spline* de ordem 3 para a variável *idade_utente_a_data_validacao* e também tem um *spline* de ordem 3 para a variável explicativa *data_confirmado1*. Não obstante, quando se efetuam os gráficos dos efeitos não lineares destas duas variáveis, que se encontram na Figura 4.50, sendo que a Figura 4.50a representa o efeito da variável explicativa *data_confirmado1* e a Figura 4.50b representa o efeito da variável explicativa *idada_utente_a_data_validacao*, observa-se que o efeito do *spline* de suavização de grau 3 na variável explicativa *idada_utente_a_data_validacao* é linear, pelo que se optou por modelar esta variável explicativa de forma linear. Já o efeito do *spline* de suavização de grau 3 produz um efeito não linear na variável explicativa *data_confirmado1*, pelo que se optou por modelar esta variável explicativa de forma não linear.

O *output* da função *gam()* da biblioteca *gam* do *software* R, com os resultados de ajuste



(a) Efeito *data_confirmado1*

(b) Efeito *idada_utente_a_data_validacao*

Figura 4.50: Efeitos das variáveis explicativas *data_confirmado1* e *idada_utente_a_data_validacao*

é o que se encontra na Tabela I.8, que se encontra no Anexo I, onde se observa que o *splines* de suavização de grau 3 aplicado à variável explicativa *data_confirmado1*, é estatisticamente significante a um nível de significância de 5%, pelo que esta variável explicativa ajusta-se bem ao *spline* de suavização de grau 3.

Através da análise da Tabela I.8, que se encontra no Anexo I, observamos que os coeficientes das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias*, *asma* e alguns dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* não são estatisticamente significantes para um nível habitual de significância de 5%. Assim, utilizou-se o teste da razão de verosimilhança para se testar se existem evidências estatísticas que nos permitam inferir que os coeficientes destas variáveis explicativas sejam simultaneamente identicamente nulos. Assim, realizando o teste de razão de verosimilhança sobre os coeficientes das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias*, *asma* e dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, obteve-se um valor observado da estatística de teste de 250.47, e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias*, *asma* e dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* sejam conjuntamente iguais a 0. Posto isto, realizou-se um teste de razão de verosimilhança, sobre os coeficientes dos diversos níveis da variável explicativa *descricao_distrito_ocorrencia*, para se averiguar se o conjunto dos coeficientes dos diferentes níveis desta variável explicativa são conjuntamente iguais a 0. Da realização do teste, obteve-se um valor observado da estatística de teste de 244.29 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são todos

iguais a 0, pelo que não existem evidências estatísticas que nos permitam retirar a variável explicativa *descricao_distrito_ocorrencia* do presente modelo. Uma vez que, os coeficientes das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias* e *asma* não são estatisticamente significantes a um nível de significância de 5%, quando se aplica o teste de *Wald*, (coluna do Pr(>|z|), da Tabela I.8, do Anexo I), então aplicou-se o teste de razão de verosimilhança sobre os coeficientes destas 3 variáveis explicativas, donde se obteve um valor observado da estatística de teste de 6.2 e um valor do $p\text{-}value$ de 0.1023, pelo que a um nível de significância $\alpha = 0.05$ é plausível assumir que os 3 coeficientes destas variáveis explicativas são conjuntamente iguais a 0, ou seja, é plausível assumir que os coeficientes das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias* e *asma* são conjuntamente iguais a 0. Posto isto, removeu-se estas 3 variáveis explicativas do modelo e ajustou-se o modelo de novo. Pelo que o *output* obtido através da função *gam()* da biblioteca *gam* do *software* R encontra-se na Tabela I.9. Através da análise dos valores da coluna Pr(>|z|) da Tabela I.9, do Anexo I, que é referente aos valores do $p\text{-}value$ do teste de *Wald* nas variáveis explicativas sem serem ajustadas pelos *splines* de suavização, observa-se que todas as variáveis explicativas à exceção de alguns dos níveis da variável explicativa *descricao_distrito_ocorrencia*, são estatisticamente significantes a um nível de significância de 5%. Assim, realizou-se um teste de razão de verosimilhança sobre os coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, onde se obteve um valor observado da estatística de teste de 244.26 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são conjuntamente iguais a 0. Assim, não existem evidências estatísticas que nos permitem retirar a variável explicativa *descricao_distrito_ocorrencia* do presente modelo.

Uma vez que, não existem evidências estatísticas que nos permitem retirar a variável explicativa *descricao_distrito_ocorrencia* do presente modelo e que as restantes variáveis explicativas são estatisticamente significantes a um nível de significância de 5%, então tem-se que o modelo final ajustado à base de dados restrita das comorbilidades, é constituída por todas as variáveis explicativas à exceção das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias* e *asma*, onde a variável explicativa *data_confirmado1* é modelada através de um *spline* de suavização de grau 3.

Ajustando agora o modelo Aditivo Generalizado à base de dados balanceada das comorbilidades, com as diversas sequências de graus dos *splines* de suavização para as variáveis explicativas *data_confirmado1* e *idade_utente_a_data_validacao*, de modo a selecionar os graus que melhor se ajustam a estes dados para estas duas variáveis explicativas, de forma averiguar se são os mesmos que no modelo ajustado à base dados restrita das comorbilidades, os resultados obtidos dos diversos modelos Aditivos Generalizados encontram-se na Tabela 4.23.

Tal como nos resultados do modelo Aditivo Generalizado, quando aplicado na base de dados restrita das comorbilidades, neste caso também se obtém que o melhor modelo em

Tabela 4.23: Resultados do *AIC* dos diversos modelos Aditivos Generalizados ajustados na base de dados balanceada das comorbilidades.

Grau do *spline*		
idade_utente_a_data_validacao	*data_confirmado1*	*AIC*
-	2	12214.46
2	-	12214.81
2	2	12169.45
-	3	12158.43
3	-	12193.44
3	2	12148.35
2	3	12114.69
3	3	12093.86

termos do *AIC* é o modelo onde a variável explicativa *idade_utente_a_data_validacao* é modelada com um *spline* de suavização de grau 3 e a variável explicativa *data_confirmado1* também é modelada com um *spline* de suavização de grau 3. No entanto, tal como no modelo ajustado com a base de dados restrita das comorbilidades, quando se efetua o gráfico do efeito do *spline* de suavização de grau 3 na variável explicativa *idade_utente_a_data_validacao*, também se constata que este apresenta um efeito linear, pelo que se optou por modelar esta variável de forma linear. Já a variável explicativa *data_confirmado1* foi modelada com um *spline* de suavização de grau 3. Na Tabela I.10, do Anexo I, encontram-se os resultados do *output* obtido através da função *gam()* da biblioteca *gam* do *software* R, referente a este modelo.

Através da análise dos valores do $p\text{-}value$ do teste de *Wald*, aplicados a cada um das variáveis explicativas não modeladas com os *splines* de suavização, valores da coluna Pr(>|z|), da Tabela I.10 que se encontra no Anexo I, observa-se que as variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias* e *asma* não são estatisticamente significantes a um nível de significância de 5%, tal como se tinha no modelo ajustado à base de dados restrita das comorbilidades. Assim, aplicou-se um teste de razão de verosimilhança, sobre o conjunto dos coeficientes destas 3 variáveis explicativas, onde se obteve um valor observado da estatística de teste de 0.97 e um valor do $p\text{-}value$ de 0.8088, pelo que a um nível de significância de 5% não existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que o conjunto dos coeficientes destas 3 variáveis explicativas são conjuntamente iguais a 0. Posto isto, existem evidências estatísticas que nos permitem remover estas 3 variáveis explicativas do modelo ajustado à base de dados balanceada, pelo que ficamos com um modelo que contém as mesmas variáveis explicativas que o modelo final Aditivo Generalizado ajustado à base de dados restrita das comorbilidades. No entanto, através da análise da Tabela I.10, do Anexo I, observa-se que para além destas 3 variáveis explicativas não serem estatisticamente significantes a um nível de significância de 5%, também a variável explicativa *insuficiencia_cardiaca* não é significante. No entanto, não se realizou o teste de razão de verosimilhança sobre as demais variáveis explicativas

que não fossem estatisticamente significantes, uma vez que a base de dados balanceada é obtida através do método de subamostragem aleatória, que remove observações da classe maioritária de forma aleatória, sendo que uma das desvantagens deste método é a perda de padrões que existem entre as variáveis na base de dados com um maior número de observações.

Do exposto, resulta que o modelo Aditivo Generalizado que é treinado na base de dados de treino das comorbilidades é formado por todas as variáveis explicativas à exceção das variáveis explicativas *coagulopatia_de_consumo*, *vih_outras_imunodeficiencias* e *asma*, onde a variável explicativa *idade_utente_a_data_validacao* é modelada linearmente e a variável explicativa *data_confirmado1* é modelada através de um *spline* de suavização de grau 3.

Na Figura 4.51, encontra-se representado o efeito não linear, sobre a probabilidade estimada de mortalidade, da variável explicativa *data_confirmado1*, referente ao modelo final obtido na base de dados restrita das comorbilidades.

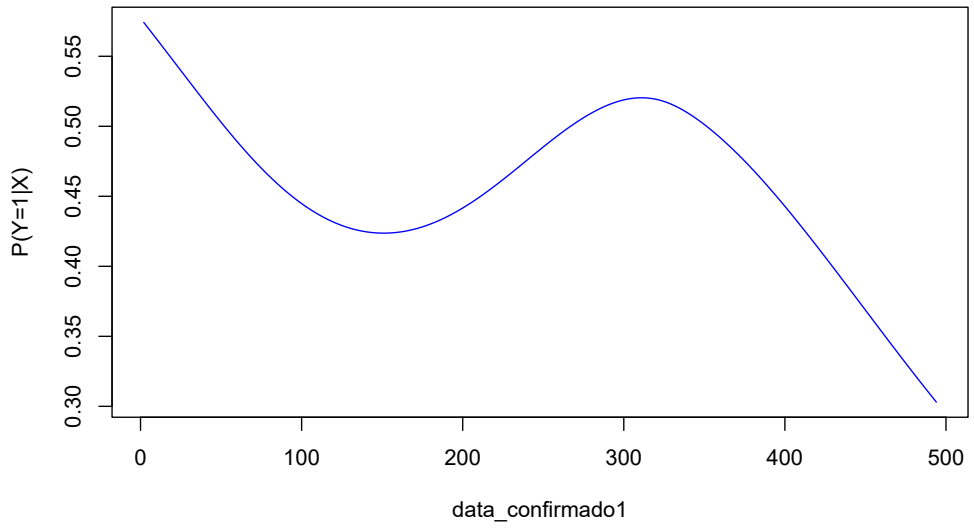


Figura 4.51: Efeito da variável explicativa *data_confirmado1* na probabilidade estimada de mortalidade.

Através da análise da Figura 4.51, pode-se concluir que a variável explicativa *data_confirmado1* possui um efeito não linear, uma vez que através da análise desta figura verifica-se que desde o primeiro caso de confirmação até meados de setembro (cerca de 150 dias após o primeiro caso confirmado), existe um efeito de decréscimo da chance de morte com o avançar da data de caso, sendo que a partir de meados de setembro até ao início/metade do mês de janeiro, (300/320 dias após o primeiro caso confirmado), existe um efeito de aumento da chance de morte com o avançar da data de caso, sendo que a partir de meados do mês de janeiro até ao último caso confirmado na presente base de

dados que é 500 dias após o primeiro caso, existe novamente um efeito de decréscimo da chance de morte com o avançar da data de caso. Estes padrões estão muito certamente associados com as medidas de contenção que o Governo de Portugal instituiu, como por exemplo no período entre janeiro de 2021 a meados de março de 2021, o confinamento, assim como também no período do início da pandemia, entre meados de março de 2020 a meados de maio de 2020, que impôs o confinamento total sendo que, em ambos estes períodos, tem-se um decréscimo do número de casos. Ou seja, da análise do efeito não linear da variável explicativa *data_confirmado1*, pode-se concluir que a *chance* de um indivíduo falecer durante o período entre os meados de setembro e janeiro é mais elevada do que os restantes períodos, uma vez que este período, é caracterizado pelo número crescente de casos.

Através da análise dos *OR*, que se encontram na Tabela I.11, do Anexo I, pode-se concluir o seguinte sobre as variáveis explicativas com os efeitos paramétricos:

- A *chance* de morte por COVID-19 estimada é maior 1.12 vezes para cada ano de aumento de idade;
- As *chances* de falecer devido à COVID-19 entre os indivíduos do sexo masculino é 2.12 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que são do sexo feminino, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *neoplasia* é 1.82 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *neoplasia*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *diabetes* é 1.37 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *diabetes*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *doenca_neurologica_ou_neuromuscular_cronica* é 1.50 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doenca_neurologica_ou_neuromuscular_cronica*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *doenca_pulmonar_cronica* é 1.49 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doenca_pulmonar_cronica*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *patologia_hepatica* é 2.51 vezes maior do que as *chances* de falecer

devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *patologia_hepatica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *doencas_hematologicas_cronicas* é 1.42 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doencas_hematologicas_cronicas*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *doenca_renal_cronica* é 2.11 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *doenca_renal_cronica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *deficiencia_neurologica_cronica* é 1.77 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *deficiencia_neurologica_cronica*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *insuficiencia_renal_aguda* é 3.26 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *insuficiencia_renal_aguda*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem a comorbilidade *insuficiencia_cardiaca* é 1.57 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem a comorbilidade *insuficiencia_cardiaca*, neste presente estudo;

4.3.1.5 Árvore de Classificação

Antes de se treinar o modelo da Árvore de Classificação com os dados de treino referente às comorbilidades, primeiro aplicou-se este modelo à base de dados balanceada das comorbilidades com todas as variáveis referentes às comorbilidades consideradas para os ajustamentos dos modelos. A Árvore de Classificação apenas pode ser aplicada aos dados balanceados, uma vez que um dos pressupostos deste modelo é que as classes da variável resposta têm de ser equilibradas entre si, ou seja, cada um das classes tem de ter um número de observações idênticas.

A aplicação da Árvore de Classificação aos dados balanceados com todas as variáveis da base de dados balanceada, encontra problemas com as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1* nas divisões necessárias, com conclusões que não são corretas. Quando se utiliza estas duas variáveis explicativas, uma vez que como estamos a utilizar o algoritmo CART que utiliza a divisão recursiva binária, ou seja, a divisão em cada uma das variáveis leva à criação de dois novos ramos, em algumas das divisões na variável explicativa *descricao_distrito_ocorrencia*, a árvore está a juntar distritos

que tiveram muito poucas mortes com distritos que tiveram muitas mortes e a classificar este ramo como mortes, em vez de apenas considerar os distritos que tiveram muitas mortes. Esta questão, pode dever-se ao facto da base de dados das comorbilidades ter sido balanceada, pelo que quando as observações foram removidas, ficamos com uma base de dados em que alguns dos distritos que tinham muitas mortes passam a ter um número idêntico de mortes a alguns dos outros distritos com poucas mortes, Por exemplo, um dos ramos de divisão da Árvore de Classificação quando utilizada a variável explicativa *descricao_distrito_ocorrencia* continha os seguintes distritos, Beja, Braga, Bragança, Coimbra, Évora, Faro, Guarda, Leiria, Lisboa, Portalegre, Porto Santarém, Setúbal e Viseu e este ramo terminava com uma folha em que esta classificava como morte os indivíduos que tinham sido infetados em cada um destes distritos, contudo os distritos de ocorrência Beja, Faro, Leiria, Santarém, Portalegre e Braga são dos distritos que têm o menor número de mortes, enquanto que os restantes distritos descritos no ramo de divisão são os que têm um maior número de mortes, tal como se pode confirmar na Figura 4.14, ou seja, neste presente caso, tem-se que a Árvore de Classificação não divide corretamente os distritos. Já a variável *data_confirmado1* foi removida pelo facto desta variável também se ter perdido um grande número de observações, pelo que os dias que apresentavam o maior número de casos, com o balanceamento da base de dados, ficaram com um número de observações idênticas aos dias que apresentavam um menor número de casos, ou seja, a proporcionalidade de casos diários entre a base de dados restrita e balanceada não se manteve. Por estes motivos não se levou em conta as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, da base de dados balanceada das comorbilidades na construção da Árvore de Classificação.

Recorrendo à função *rpart()* da biblioteca *rpart* do *software* R, construiu-se a Árvore de Classificação de tamanho máximo, onde o critério de paragem utilizado foi que o número mínimo de observações em qualquer uma das folhas (nós terminais) tinha de ter pelo menos 7 observações. O critério de divisão utilizado na construção da Árvore de Decisão foi o *índice de gini*. Depois de se construir Árvore de Classificação de tamanho máximo utilizou-se a poda de custo de complexidade (Cost complexity pruning), onde o parâmetro de ajuste α, dado na equação 3.94 da secção 3.5, é de $\alpha = 0.0002$.

No Anexo I, encontra-se representada na Figura I.1, a Árvore de Classificação obtida quando aplicada na base de dados balanceada das comorbilidades sem as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*. As variáveis explicativas incluídas na construção da Árvore de Classificação, após se efetuar a poda são:

- *idade_utente_a_data_validacao*;
- *sexo_utente*;
- *doenca_renal_cronica*;
- *neoplasia*;
- *doenca_pulmonar_cronica*;

- *patologia_hepatica*;
- *doenca_hematologica_cronica*.

No Anexo I na sub-subsecção I.1.5.1, encontram-se as diversas classificações que resultam do modelo da Árvore de Classificação, quando aplicada à base de dados balanceada das comorbilidades, onde se observa que as variáveis explicativas *idade_utente_a_data_validacao* e *sexo_utente* são muitas vezes utilizadas na divisão dos ramos da árvores final obtida.

O modelo de treino, foi aplicado com recurso mais uma vez à função *rpart()* da biblioteca *rpart* do *software* R, onde no modelo de treino se levou em conta todas as variáveis explicativas à exceção das variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, sendo que o critério de paragem para a construção da Árvore de Classificação de tamanho máximo é mais uma vez que o número mínimo de observações em qualquer uma das folhas, tinha de ter pelo menos 7 observações. Mais uma vez, se utilizou o *índice de gini* como o critério utilizado na construção da Árvore de Classificação. Depois de se construir a Árvore de Classificação de tamanho máximo utilizou-se, mais uma vez, a poda de custo de complexidade (*Cost complexity pruning*), para se obter uma Árvore de Classificação final, sendo que o valor do parâmetro de ajuste α é de $\alpha = 0.00024$. As variáveis explicativas incluídas na construção da Árvore de Classificação, após a realização da poda da Árvore de tamanho máximo ajustada à base de dados de treino das comorbilidades, são as seguintes:

- *idada_utente_a_data_validacao*;
- *sexo_utente*;
- *diabetes*;
- *neoplasia*;
- *doenca_renal_cronica*;
- *doenca_pulmonar_cronica*;
- *asma*.

4.3.1.6 Floresta Aleatória

Tal como no método da Árvore de Classificação, neste método também se excluíram as variáveis explicativas *data_confirmado1* e *descricao_distrito_ocorrencia*, pelos mesmos motivos apresentados no método da Árvore de Classificação, uma vez que as Florestas Aleatórias utilizam as Árvores de Classificação como os modelos simples de blocos de construção. Antes de se treinar o modelo da Floresta Aleatória com os dados de treino das comorbilidades, primeiro ajustou-se o modelo da Floresta Aleatória aos dados balanceados das comorbilidades, para discutir qual o número de árvores que deveriam ser

utilizadas no modelo de treino. Assim, recorrendo à função *randomForest()* da biblioteca *randomForest* do *sofware* R, ajustou-se o modelo da Floresta Aleatória com um número de sub-Árvores de 250, para se analisar se o erro OOB geral estabilizava, pois se o erro OOB geral se estabilizar com este número de sub-Árvores, então poderemos usá-lo para treinar o modelo. Na Figura 4.52, encontra-se o erro OOB da Floresta Aleatória ajustada aos dados balanceados das comorbilidades, considerando as 250 Árvores de Classificação.

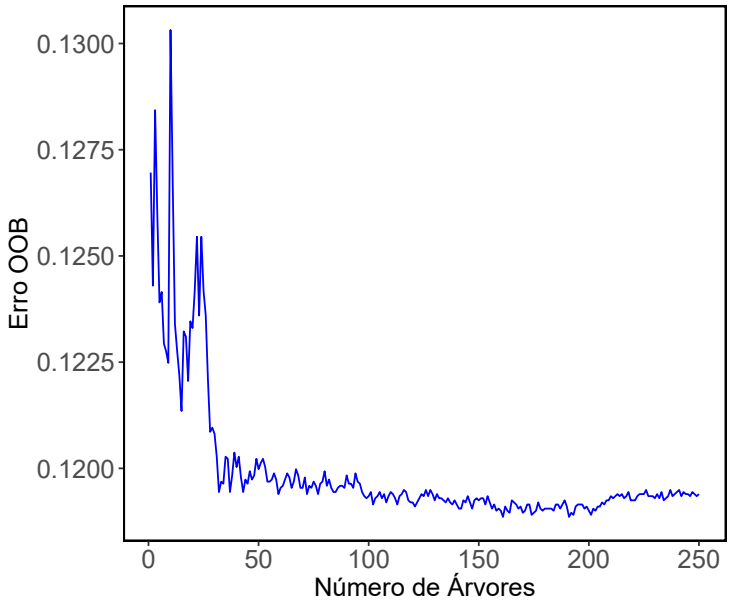


Figura 4.52: Erro OOB no modelo Floresta Aleatória, base de dados balanceada das comorbilidades.

Através da análise da Figura 4.52, observa-se que o erro OOB geral, tende-se a estabilizar, pois a partir das 150 – 200 Árvores, o erro OOB geral não apresenta grandes variações. Assim, o modelo de Floresta Aleatória foi treinado com os dados de treino das comorbilidades, considerando também as 250 Árvores de Classificação.

No gráfico da Figura 4.53, encontra-se a importância de cada uma das variáveis explicativas através da diminuição média do *índice de gini*, onde se observa que as variáveis explicativas que este modelo, quando ajustado na base de dados balanceada das comorbilidades, considera como sendo as mais importantes são, por ordem de importância, as seguintes:

- *idade_utente_a_data_validacao*;
- *diabetes*;
- *doenca_renal_cronica*;
- *doenca_pulmonar_cronica*;
- *neoplasia*;

- *sexo_utente*;
- *doenca_neurologica_ou_neuromuscular_cronica*.

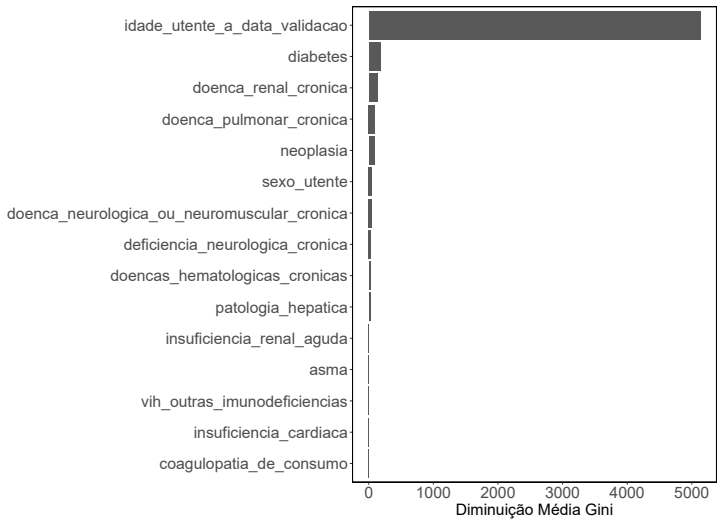


Figura 4.53: Importância variáveis explicativas no modelo Floresta Aleatória, base de dados balanceada das comorbilidades.

Por outro lado, através da análise do gráfico da Figura 4.53, observa-se que as variáveis explicativas *coagulopatia_de_consumo*, *insuficiencia_cardiaca* e *vih_outras_imunodeficiências* não são muito relevantes.

Do exposto, resulta que o modelo de treino, aplicado à base de dados de treino das comorbilidades, leva em conta todas as variáveis explicativas à exceção das variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmacao1*, sendo constituído por 250 Árvores de Classificação como o bloco de construção do modelo.

4.3.1.7 Redes Neuronais

Neste presente trabalho, no caso das comorbilidades, apenas se aplicam as Redes Neuronais *feed-forward*.

Para se treinar o modelo das Redes Neuronais, com a base de dados de treino das comorbilidades, primeiro tem que se selecionar o número de camadas ocultas e também o número de neurónios em cada uma das camadas ocultas. Para isto, elaboraram-se diversos modelos e em cada um desses modelos construídos utilizou-se a validação cruzada com 10 grupos para calcular a estimativa de desempenho de cada um dos modelos, sendo que a métrica que se utilizou na validação cruzada foi a métrica *accuracy*. Ou seja, o número de camadas ocultas e o número de neurónios de cada uma das camadas ocultas irá ser selecionado do modelo que apresentar a maior estimativa da validação cruzada utilizando a métrica *accuracy*.

Os diversos modelos considerados foram todas as combinações possíveis entre o número de camadas ocultas a variar entre 1 e 2 e o número de nós em cada uma das

camadas ocultas variar entre 1 e 50, onde as funções de ativação escolhidas tanto para a primeira camada oculta como para a segunda camada oculta foram as funções de ativação *ReLU* e para a camada de saída utilizou-se a função de ativação *sigmoid*. Ou seja, em primeiro lugar consideraram-se todas as Redes Neuronais com uma única camada, onde o número de neurónios de cada uma destas Redes Neuronais variava entre 1 e 50, com as funções de ativação referidas. Depois consideraram-se todos os modelos com duas camadas ocultas, onde o número de neurónios em cada uma destas camadas ocultas variavam entre 1 e 50, em que as funções de ativação utilizadas em cada uma das camadas ocultas é a função de ativação *ReLU* e a função de ativação utilizada na camada de saída é a função de ativação *sigmoid*.

Depois, em cada um dos modelos referidos anteriormente utilizou-se a validação cruzada com 10 grupos para se calcular a estimativa de desempenho de cada um dos modelos, sendo que a métrica utilizada foi a *accuracy*. Uma vez que a métrica utilizada é a *accuracy*, então na validação cruzada iremos utilizar a base de dados balanceada das comorbilidades, uma vez que na secção 3.10 se referiu que esta métrica não é viável quando a base de dados é desequilibrada.

De referir que neste presente trabalho, apenas se consideraram as funções de ativação *ReLU* para as camadas ocultas, em vez de se testarem mais modelos com outras diferentes funções de ativação. Este facto deve-se sobretudo ao tempo computacional deste processo ser longo, uma vez que só a utilização das funções de ativação *ReLU* para as camadas ocultas resultou já num tempo computacional de cerca de 12 horas.

Para se ajustar cada um dos modelos anteriormente descritos recorreu-se à biblioteca *Keras* do *sofware* R. Para se ajustar cada uma das Redes Neuronais, utilizou-se o *enviesamento*, onde as estimativas iniciais dos *enviesamento* são 0 e as estimativas iniciais dos restantes parâmetros são através do indicador uniforme *Glorot*, já que se utilizou o método de estimação do *gradiente descendente estocástico*.

Posto isto, verificou-se que o melhor modelo de Redes Neuronais selecionado é o que apresenta duas camadas ocultas, sendo que o número de neurónios na primeira camada oculta é de 20 neurónios e o número de neurónios na segunda camada oculta é de 15 neurónios, uma vez que a estimativa de validação cruzada com 10 grupos é de 0.8712, utilizando a métrica *accuracy*.

Desta forma, o modelo de Redes Neuronais foi treinado com a base de dados de treino das comorbilidades, onde se considerou uma Rede Neuronal composta por uma camada de entrada, duas camadas ocultas e uma camada de saída. O número de neurónios na camada de entrada é de 32, uma vez que a variável *descricao_distrito_ocorrencia* é uma variável categórica com 20 níveis, ou seja, passando esta variável para variáveis *dummy* ficamos com 19 níveis, que se soma às restantes 13 variáveis explicativas, pois a base de dados de treino das comorbilidades é constituída por 14 variáveis explicativas. Já o número de neurónios nas duas camadas ocultas é de 20 e 15, respetivamente, sendo que em ambas as camadas ocultas se utilizou a função de ativação *ReLU*. Por último a camada de saída é apenas constituída por um neurónio, sendo utilizada nesta camada a função

de ativação *sigmoid*. Na estimação da Rede Neuronal com os dados de treino utilizou-se o *enviesamento*, onde as estimativas iniciais dos *enviesamento* são 0 e as estimativas iniciais dos restantes parâmetros são feitos através do indicador uniforme *Glorot*,uma vez que se utilizou o método de estimação do *gradiente descendente estocástico*.

4.3.2 Dados Sintomas

Do mesmo modo que se elaborou uma análise preliminar para a redução do número das covariáveis a considerar antes da aplicação os diversos modelos de Aprendizagem Automática aos dados das comorbilidades, nos dados referentes aos sintomas também se fez o mesmo, isto é, antes de se aplicar os diferentes modelos de Aprendizagem Automática aos dados dos sintomas, aplicou-se o teste do χ^2, a Regressão Logística múltipla com o *stepwise* e a Regressão do LASSO de Grupo, de forma a reduzir o número de variáveis explicativas iniciais que se utilizarão nos diferentes modelos. O teste do teste do χ^2 apenas foi efectuado entre as variáveis explicativas categóricas e a variável resposta *morte*. A Tabela 4.24, mostra quais as variáveis que cada um dos 3 métodos elimina.

Da análise da Tabela 4.24, observa-se que o método do *stepwise* aplicado na Regressão Logística elimina 4 variáveis explicativas que são a *data_confirmado1*, a *dor_abdominal*, a *diarreia* e as *convulsoes*. Já o teste do χ^2, elimina também 4 variáveis explicativas que são a *tosse_seca_ou_produtiva*, a *dor_no_peito*, a *nauseas_vomitos* e a *diarreia*. Já o método do LASSO de Grupo, não eliminou nenhuma das variáveis. Para a eliminação das variáveis no estudo subsequente, utilizou-se o mesmo critério que a eliminação das variáveis referentes às comorbilidades, ou seja, apenas se eliminaram as variáveis em que pelo menos dois dos métodos sugerissem a sua eliminação. Dado que o LASSO de Grupo não eliminou nenhuma das variáveis e que apenas a variável explicativa *diarreia* é que é eliminada simultaneamente pelo método do *stepwise* e do teste do χ^2, então somente esta variável é que foi removida. Assim, a base de dados restrita dos sintomas é constituída por todas as variáveis explicativas indicadas na Tabela 4.24 à exceção da variável explicativa *diarreia*, contendo também a variável resposta *morte*.

Na secção 3.10, referiu-se quais eram os problemas de classificação numa base de dados desequilibrada/desbalanceada, sendo que um dos problemas que se referiu é que muitas das métricas descritas na secção 3.9 costumam dar conclusões erróneas sobre a avaliação do modelo construído.

Uma vez que, esta base de dados não é equilibrada, ou seja, é desbalanceada, pois existe uma grande diferença entre as duas categorias da variável resposta, foi necessário equilibrar/balancear a presente base de dados, de modo a treinar e a testar os diferentes modelos para os avaliar em termos preditivos de uma forma correta. Para se balancear a base de dados, utilizou-se o método da subamostragem aleatória. O motivo pelo qual se considera um método de subamostragem em vez de um método de sobreamostragem, é que os métodos de sobreamostragem efetuam o balanceamento de uma base de dados através da sobreamostragem de observações que pertencem à classe minoritária no caso

Tabela 4.24: Variáveis eliminadas e não eliminadas referentes aos sintomas ("X"corresponde que a variável é eliminada pelo respetivo método e "−"que a variável não é eliminada, pelo respetivo método).

Variável	**Regressão Logística com o *stepwise***	**Lasso de Grupo**	**Teste χ^2**
idade_utente_a_data_validacao	–	–	–
sexo_utente	–	–	–
data_confirmado1	X	–	–
descricao_distrito_ocorrencia	–	–	–
historia_de_febre_ou_calafrios	–	–	–
pneumonia	–	–	–
tosse_seca_ou_produtiva	–	–	X
dispneia	–	–	–
coriza	–	–	–
odinofagia	–	–	–
cefaleia	–	–	–
dor_abdominal	X	–	–
dor_no_peito	–	–	X
artralgia	–	–	–
mialgias	–	–	–
nauseas_vomitos	–	–	X
diarreia	X	–	X
convulcoes	X	–	–
irritabilidade_confusao	–	–	–
fraqueza_geral_ou_astneia	–	–	–
auscultacao_pulmonar_anomala	–	–	–
radiografia_pulmonar_com _alteracoes	–	–	–
coma	–	–	–
taquicardia	–	–	–

do método da sobreamostragem aleatória, enquanto que na técnica SMOTE as novas observações da classe minoritária não são apenas cópias das observações que pertencem à classe minoritária como se faz no método da sobreamostragem aleatória. Na técnica SMOTE o algoritmo gera novas observações sintéticas para a classe minoritária como se referiu na sub-subsecção 3.10.1.1. No entanto, em ambas as técnicas de sobreamostragem referidos, passaríamos a ter uma base de dados onde as observações deixariam de ser independentes, o que para muitos dos modelos considerados um dos seus pressupostos é a independência das observações. Posto isto, construiu-se uma nova base de dados, em que iremos denotar por base de dados balanceada dos sintomas, que contém 20470 observações da base de dados restrita dos sintomas.

Para se treinar e testar cada um dos modelos, para depois se averiguar a capacidade preditiva destes e compará-los em termos preditivos, utilizou-se o método *holdout* para dividir a base de dados balanceada dos sintomas em base de dados de treino e em

base de dados de teste. Neste caso, a base de dados de treino é constituída por 80% das observações da base de dados balanceada dos sintomas enquanto que a base de dados referente ao teste é constituída pelas restantes 20% das observações. De referir que a base de dados balanceada dos sintomas tem as mesmas variáveis que a base de dados restrita dos sintomas.

Nas sub-subsecções 4.3.2.1, 4.3.2.2, 4.3.2.3, 4.3.2.4, 4.3.2.5, 4.3.2.6 e 4.3.2.7, encontram-se os modelos de Regressão Logística, a Regressão Logística com interações, o LASSO de Grupo, o modelo Aditivo Generalizado, as Árvores de Classificação, a Floresta Aleatória e as Redes Neuronais *feed-forward*, respetivamente, onde se descrevem os modelos finais para se aplicar à base de dados de treino dos sintomas.

Como referido anteriormente os modelos de Regressão Logística, de Regressão Logística com interações e o modelo Aditivo Generalizado, são muitas vezes utilizados para estimar os coeficientes das variáveis explicativas, de forma a averiguar a sua significância, para aferir se as variáveis explicativas estão ou não relacionadas com a variável resposta, pelo que nas sub-subsecções referentes a estes modelos também se apresentam os resultados de estimação destes modelos, assim como os diferentes testes à significância dos seus coeficientes.

4.3.2.1 Regressão Logística

Antes de se treinar o modelo de Regressão Logística nos dados de treino referente aos sintomas, primeiro ajustou-se um modelo de Regressão Logística com o método de *stepwise* aos dados restritos dos sintomas, com o objetivo de se selecionar o melhor subconjunto de entre todas as variáveis explicativas referidas anteriormente, que se levaria em conta para o modelo de treino com a Regressão Logística. A Regressão Logística com o método do *stepwise* é aplicada na base de dados restrita dos sintomas, uma vez que se aplicássemos este método na base de dados balanceada dos sintomas, poderia-se estar a eliminar variáveis explicativas que sejam importantes. No entanto, é de esperar que algumas das variáveis explicativas selecionadas para o modelo de treino não sejam estatisticamente significantes, quando se ajusta o modelo com essas variáveis explicativas selecionadas para a base de dados balanceada dos sintomas. Assim, o modelo de treino é aplicado aos dados de treino dos sintomas, onde as variáveis explicativas utilizadas são as variáveis explicativas que o método do *stepwise* não elimina e que sejam estatisticamente significantes, a um nível de significância de 5%.

Assim, aplicando o método do *stepwise*, através do *software* R, os resultados que se obtêm são os que se encontram na Tabela I.12, que se encontra no Anexo I.

Das variáveis explicativas utilizadas no início do modelo, o método do *stepwise* remove as seguintes:

- *convulsoes*;
- *dor_abdominal*;

- *data_confirmado1*.

Por outro lado, através da análise dos valores dos $p\text{-}values$ referentes ao teste de *Wald*, coluna $\Big(\Pr(>|z|)\Big)$, constata-se que grande parte dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* não são estatisticamente significantes a um nível de significância de 5%, ou seja, a um nível de significância de 5%, cada um dos coeficientes destes níveis podem ser considerados nulos de forma individualmente. Assim, aplicando o teste de razão de verosimilhança para averiguar se existem evidências estatísticas se o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* podem ser considerados conjuntamente nulos, obteve-se um valor observado da estatística de teste de 4217.15, sendo que o valor do $p\text{-}value$ que se obteve é de 2.2×10^{-16}, ou seja, a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula de que todos os coeficientes sejam conjuntamente nulos. Por outro lado, quando se compara o *AIC* do modelo com a variável explicativa *descricao_distrito_ocorrencia*, com o *AIC* do modelo sem a variável explicativa *descricao_distrito_ocorrencia*, verifica-se que existe um aumento do valor do *AIC*, pois o modelo com a referida variável explicativa tem um *AIC* de 59562.97, enquanto que o modelo sem a variável explicativa *descricao_distrito_ocorrencia* tem um valor de *AIC* de 59562.97, pelo que em termos do *AIC*, verifica-se que o modelo com a variável explicativa *descricao_distrito_ocorrencia* ajusta-se melhor aos dados do que o modelo sem esta variável explicativa. Posto isto, a variável explicativa *descricao_distrito_ocorrencia* ficou no presente modelo.

Uma vez que, as restantes variáveis explicativas são estatisticamente significantes a um nível de significância de 5%, pois o valor do $p\text{-}value$ do teste do *Wald*, coluna $\Big(\Pr(>|z|)\Big)$, quando aplicado a cada uma das restantes variáveis explicativas é menor que 0.05, então temos que o presente modelo é o modelo final da Regressão Logística, uma vez que não existem evidências estatísticas que nos permitam retirar a variável explicativa *descricao_distrito_ocorrencia* do presente modelo e dado que as restantes variáveis explicativas são estatisticamente significantes a um nível de significância de 5%. Assim, todas as variáveis explicativas descritas na Tabela I.12, do Anexo I são mantidas no presente modelo, sendo este o modelo final da Regressão Logística.

Na Tabela I.13, do Anexo I, encontram-se representados os *OR* de cada uma das variáveis explicativas do modelo.

Através da análise dos valores dos *OR* de cada uma das variáveis explicativas podem-se tirar as seguintes conclusões:

- A chance de morrer por COVID-19 é maior 1.11 vezes por cada ano de aumento na idade dos indivíduos.

- As *chances* de falecer devido à COVID-19 entre os indivíduos do sexo masculino é 2.02 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos do sexo feminino, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *dispneia* é 3.25 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *dispneia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem *radiografia_pulmonar_com_alteracoes* é 1.88 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem *radiografia_pulmonar_com_alteracoes*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *mialgias* é 0.59 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *mialgias*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *historia_de_febre_ou_calafrios* é 1.51 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *historia_de_febre_ou_calafrios*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *coriza* é 0.55 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *coriza*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *cefaleia* é 0.61 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *cefaleia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *tosse_seca_ou_produtiva* é 0.80 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *tosse_seca_ou_produtiva*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *taquicardia* é 1.65 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *taquicardia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *coma* é 3.77 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *coma*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *odinofagia* é 0.67 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *odinofagia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *irritabilidade_confusao* é 1.37 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *irritabilidade_confusao*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *fraqueza_geral_ou_astenia* é 1.13 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *fraqueza_geral_ou_astenia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *dor_no_peito* é 0.81 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *dor_no_peito*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *pneumonia* é 1.20 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *pneumonia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *nauseas_vomitos* é 0.87 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *nauseas_vomitos*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *auscultacao_pulmonar_anomala* é 0.88 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *auscultacao_pulmonar_anomala*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos com o sintoma *artralgia* é 1.24 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *artralgia*, neste presente estudo;

Posto isto, treinou-se o modelo de Regressão Logística, através da base de dados de treino dos sintomas, em que as variáveis explicativas selecionadas para treinar este modelo foram todas as variáveis que o método do *stepwise* selecionou na base de dados restrita dos sintomas, ou seja, foram todas as variáveis explicativas que se encontram na Tabela I.12, no Anexo I.

4.3.2.2 Regressão Logística com interações

Na Regressão Logística com interações, as interações que se consideraram foram as seguintes:

- Interação entre a variável explicativa *idade_utente_a_data_validacao* e a variável explicativa *descricao_distrito_ocorrencia*;

- Interação entre a variável explicativa *idade_utente_a_data_validacao* e cada uma das variáveis explicativas referentes aos sintomas;

- Interação entre todas as variáveis explicativas referentes aos sintomas;

As duas primeiras interações consideradas têm como objetivo averiguar se existe uma relação entre variável explicativa *idade_utente_a_data_validacao* com os distritos, onde a população é mais idosa, assim como também se pretende analisar a influência da idade com cada um dos sintomas.

Neste modelo, utilizou-se o método do *stepwise* para se selecionar o melhor subconjunto de todas as variáveis explicativas que melhor se ajustam ao presente modelo. Neste caso, tal como no método de Regressão Logística anteriormente descrito, as variáveis explicativas que entram no modelo de treino são as variáveis explicativas que resultam do método *stepwise* e do teste de significância a cada um dos coeficientes das variáveis selecionadas pelo método do *stepwise* na base de dados restrita dos sintomas. Neste caso, ajusta-se este modelo com *stepwise* na base de dados restrita dos sintomas, para se tentar manter as variáveis explicativas mais importantes, pois se aplicássemos o método do *stepwise* ao modelo de Regressão Logística com interações na base de dados balanceada dos sintomas, muito provavelmente as variáveis explicativas selecionadas para o modelo de treino poderiam ser outras variáveis explicativas que o método do *stepwise* não selecionaria quando aplicada á base de dados restrita dos sintomas.

No entanto, neste caso, ao contrário do modelo de Regressão Logística construído na sub-subsecção 4.3.2.1, iremos considerar um nível de significância de 10% para o teste de *Wald*, devido ao elevado número de variáveis explicativas do modelo inicial, uma vez que se utilizássemos um nível de significância de 5%, o processo de ajustamento e de testagem dos coeficientes com o teste de razão de verosimilhança, tornaria este processo muito exaustivo. No entanto, no teste de razão de verosimilhança considerou-se o nível habitual de significância de 5%.

Assim, aplicando o método do *stepwise* no modelo de Regressão Logística com interações, através do *software* R, obtém-se o modelo final que se encontra nas Tabelas I.14 e I.15 do Anexo I.

Através da análise das Tabelas I.14 e I.15, pode-se observar que as variáveis explicativas *dispneia:odinofagia*, *cefaleia:pneumonia*, *idade_utente_a_data_validacao:dor_no_peito*, *dor_no_peito:descricao_distrito_ocorrencia* e *idade_utente_a_data_validacao:descricao_distrito_ocorrencia*, não são estatisticamente significantes a um nível de significância de 10%, pelo que se aplicou o teste de razão de verosimilhança aos coeficientes destas variáveis explicativas, obtendo-se um valor observado da estatística de teste de 59.06 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% não existem evidências

estatísticas que nos permitam afirmar que o conjunto dos coeficientes destas variáveis explicativas sejam identicamente iguais a 0.

Aplicando o teste de razão de verosimilhança às variáveis explicativas *dispneia:odinofagia*, *cefaleia:pneumonia*, *idade_utente_a_data_validacao:dor_no_peito* e *dor_no_peito*, obtém-se um valor observado da estatística de teste 6.8396 e um valor do $p\text{-}value$ de 0.1446, pelo que a um nível de significância de 5% é plausível assumir que o conjunto dos coeficientes destas variáveis explicativas são conjuntamente iguais a 0, pelo que estas variáveis explicativas foram removidas do modelo.

Ajustando de novo o modelo, sem as variáveis explicativas *dispneia:odinofagia*, *cefaleia:pneumonia*, *idade_utente_a_data_validacao:dor_no_peito* e *dor_no_peito*, verifica-se que a variável explicativa *pneumonia:fraqueza_geral_ou_astenia* não é estatisticamente significante a um nível de significância de 10%, pelo que se aplicou o teste de razão de verosimilhança ao coeficiente desta variável explicativa, tendo-se obtido um valor observado da estatística de teste de 2.6843 e um valor do $p\text{-}value$ de 0.103, onde a um nível de significância de 5% é plausível assumir que o coeficiente desta variável é identicamente igual a 0, pelo que esta variável foi removida do modelo.

Ajustando de novo o modelo, mas sem a variável explicativa *pneumonia:fraqueza_geral_ou_astenia*, verifica-se que todas as variáveis explicativas são estatisticamente significantes a um nível de significância de 10%, à exceção dos diferentes níveis das variáveis explicativas *descricao_distrito_ocorrencia* e *idade_utente_a_data_validacao:descricao_distrito_ocorrencia*. Posto isto, elaborou-se o teste de razão de verosimilhança aos coeficientes destas duas variáveis explicativas, onde se obteve um valor observado da estatística de teste de 251.93 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5%, não é plausível assumir que o conjunto dos coeficientes dos diferentes níveis destas duas variáveis explicativas sejam conjuntamente iguais a 0. No entanto, quando se aplica o teste de razão de verosimilhança aos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, obtém-se um valor observado da estatística teste de 2.223 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5%, não é plausível assumir que o conjunto dos coeficientes dos diferentes níveis desta variável explicativa sejam conjuntamente iguais a 0, e o mesmo se conclui para o teste de razão de verosimilhança quando aplicado à variável explicativa *idade_utente_a_data_validacao:descricao_distrito_ocorrencia*, onde, neste caso, se obtém um valor observado da estatística da teste de 38.948 e um valor do $p\text{-}value$ de 0.004487. Posto isto, não existem evidências estatísticas que nos permitam inferir que os coeficientes dos diferentes níveis das variáveis explicativas *descricao_distrito_ocorrencia* e *idade_utente_a_data_validacao:descricao_distrito_ocorrencia*, sejam conjuntamente iguais a zero, pelo que, se acabou por manter estas duas variáveis no modelo. Uma vez que as restantes variáveis são estatisticamente significantes, a um nível de significância de 10%, tem-se que o modelo final, considerando as interações é o modelo que se encontra na Tabela I.16 e I.17, do Anexo I, sendo que as variáveis explicativas que se levaram em conta para o modelo de treino são as que se encontram na Tabela I.16 e I.17, do Anexo I. Posto isto, treinou-se o modelo com os dados de treino dos sintomas.

4.3.2.3 LASSO de Grupo - Caso de Regressão Logística

O LASSO de Grupo foi apenas aplicado aos dados de treino, uma vez que este método tem o objetivo de encolher os coeficientes das variáveis explicativas em direção a zero e, quando aplicado em bases de dados diferentes, é claro que as estimativas dos coeficientes são diferentes, podendo em duas bases de dados de tamanho diferentes que contenham as mesmas variáveis explicativas, numa delas estimar um coeficiente duma variável explicativa como 0 e na outra base de dados estimar o coeficiente dessa mesma variável explicativa como diferente de 0. Por este motivo, o método do LASSO de Grupo foi aplicado logo diretamente nos dados de treino.

Neste caso, o LASSO de Grupo foi aplicado, uma vez que na base de dados de treino referente aos sintomas existem variáveis explicativas mistas, ou seja, tanto existem variáveis categóricas como variáveis contínuas. No presente caso, as 23 variáveis explicativas que são utilizadas para construir este modelo estão divididas em $G = 23$ grupos, em que cada um dos grupos é referente a cada uma das variáveis explicativas. No entanto, o número de variáveis no grupo da variável explicativa *descricao_destrito_ocorrencia*, é de $p_l = 19$ variáveis *dummy*, pois esta variável é uma variável categórica com 20 categorias, em que cada uma destas categorias é um distrito. Todos os restantes grupos são constituídos por uma única variável.

Posto isto, utilizou-se a função *cv.grpreg()* da biblioteca *grpreg* do *software* R, para se estimar em primeiro lugar o parâmetro de ajuste λ, onde se utilizou a validação cruzada com 10 grupos, sendo utilizado o critério da *deviance* para a seleção do melhor parâmetro λ. O melhor valor do parâmetro de ajuste λ, que retorna o menor erro da validação cruzada é de $\lambda = 0.0012$, sendo que na Tabela I.18 do Anexo I, encontram-se os coeficientes estimados das diferentes variáveis explicativas utilizadas no treino deste modelo, para o melhor valor de λ obtido.

Este método quando aplicado na base de dados de treino dos sintomas, estima os coeficientes das variáveis explicativas *data_confirmado1*, *convulsoes* e *nauseas_vomitos* e *auscultacao_pulmonar_anomala* como exatamente 0, sendo que os restantes coeficientes das restantes variáveis explicativas são estimados como diferentes de 0, como se observa na Tabela I.18 do Anexo I.

4.3.2.4 Modelo Aditivo Generalizado

O modelo Aditivo Generalizado foi utilizado para se tentar captar a não linearidade do efeito das variáveis explicativas *idada_utente_a_data_validacao* e *data_confirmado1* ao longo do tempo. Tal como no caso do modelo Aditivo Generalizado aplicado no caso das comorbilidades, neste caso também iremos aplicar o modelo Aditivo Generalizado tanto à base de dados restrita dos sintomas como na base de dados balanceada dos sintomas, para averiguar se existe concordância entre o grau escolhido dos *splines* de suavização utilizados nas variáveis explicativas *idada_utente_a_data_validacao* e *data_confirmado1*, assim como também nas variáveis explicativas selecionadas em cada um dos modelos, para

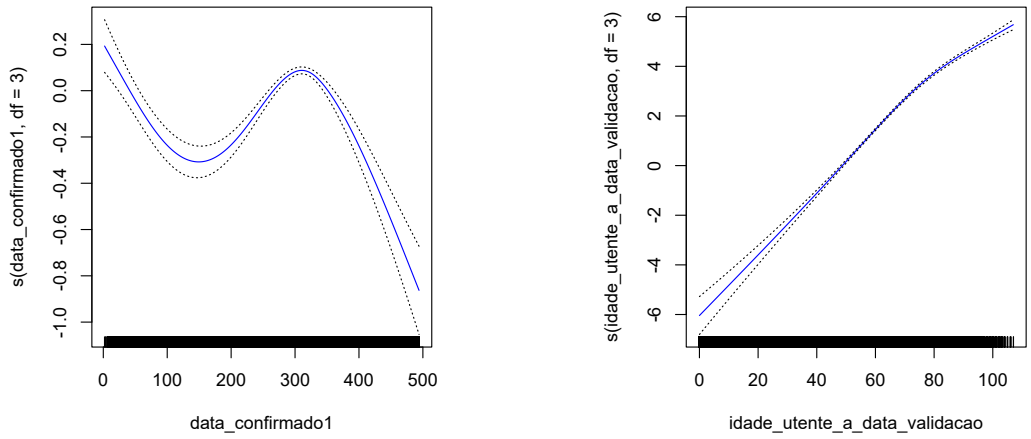
depois se levar em conta no modelo de treino, já que o modelo de treino vai ser feito numa base bastante menor. Começando por ajustar diversos modelos Aditivo Generalizado à base de dados restrita dos sintomas para a escolha do grau dos *splines* de suavização, consideraram-se todas as combinações possíveis desde grau 2 até 3 nas variáveis explicativas *idade_utente_data_validacao* e *data_confirmado1*, assim como também se considerou as combinações onde a variável explicativa *idade_utente_data_validacao* é modelada linearmente e a variável explicativa *data_confirmado1* é modelada com um *spline* de suavização de grau 2 e 3, como também se considerou o seu oposto, ou seja, também se considerou as combinações onde a variável explicativa *data_confirmado1* é modelada linearmente e a variável explicativa *idade_utente_data_validacao* é modelada com um *spline* de suavização de grau 2 e 3. A medida *AIC* foi a medida utilizada para se compararem os diferentes modelos com as diferentes combinações mencionadas anteriormente, sendo o melhor modelo escolhido é aquele que apresentar o menor valor de *AIC*. Na Tabela 4.25, encontram-se os *AIC* de cada um dos modelos Aditivos Generalizados ajustados na base de dados restrita dos sintomas, onde se utilizou mais uma vez a função *gam()* da biblioteca *gam* do *sofware* R, para se ajustar cada um dos modelos Aditivos Generalizados à base de dados restrita dos sintomas, assim como também à base de dados balanceada dos sintomas.

Tabela 4.25: Resultados do *AIC* dos diversos modelos Aditivos Generalizados ajustados na base de dados restrita dos sintomas.

Grau do *spline*		
idade_utente_a_data_validacao	*data_confirmado1*	*AIC*
-	2	59416.95
2	-	59329.93
2	2	59187.12
-	3	59233.26
3	-	59285.1
3	2	59143.34
2	3	59007.99
3	3	58965.01

Através da análise dos diferentes valores de *AIC* que se encontram na Tabela 4.25, observa-se que o melhor ajuste quando aplicado na base de dados restrita dos sintomas, corresponde a *splines* de grau 3 tanto para a variável explicativa *idada_utente_a_data_validacao* como também para a variável explicativa *data_confirmado1*. Não obstante, quando se efetuam os gráficos dos efeitos não lineares destas duas variáveis, que se encontram na Figura 4.54, sendo que a Figura 4.54a representa o efeito da variável explicativa *data_confirmado1* e a Figura 4.54b representa o efeito da variável explicativa *idada_utente_a_data_validacao*, observa-se que o efeito do *spline* de suavização de grau 3 na variável explicativa *idada_utente_a_data_validacao* é linear, pelo que se optou por modelar esta variável explicativa de forma linear. Já o efeito do *spline* de suavização de grau 3 produz um efeito não linear na variável explicativa *data_confirmado1*, pelo que se optou por modelar esta variável explicativa de

forma não linear.



(a) Efeito *data_confirmado1* (b) Efeito *idada_utente_a_data_validacao*

Figura 4.54: Efeitos das variáveis explicativas *data_confirmado1* e *idada_utente_a_data_validacao*

Na Tabela I.19 do Anexo I, encontra-se o modelo Aditivo Generalizado ajustado aos dados restritos dos sintomas, onde a variável explicativa *data_confirmado1* foi modelada com um *spline* de suavização de grau 3 e a variável explicativa *idada_utente_a_data_validacao* foi modelada de forma linear.

Na Tabela 4.26, encontram-se os valores do *AIC*, para os diferentes *splines* de suavização utilizados para as variáveis explicativas *idada_utente_a_data_validacao* e *data_confirmado1*, aplicados na base de dados balanceada dos sintomas. Tal como no modelo Aditivo Generalizado ajustado à base de dados restrita dos sintomas, o melhor modelo é o modelo Aditivo Generalizado em que os *splines* de suavização das variáveis explicativas *idada_utente_a_data_validacao* e *data_confirmado1* têm respetivamente grau 3.

No entanto, tal como no modelo ajustado com a base de dados restrita dos sintomas, quando se efetua o gráfico do efeito do *spline* de suavização de grau 3 na variável explicativa *idade_utente_a_data_validacao*, também se constata que este apresenta um efeito linear, pelo que se optou por modelar esta variável de forma linear. Já a variável explicativa *data_confirmado1* foi modelada com um *spline* de suavização de grau 3. Na Tabela I.20 do Anexo I, encontra-se o modelo Aditivo Generalizado ajustado aos dados balanceados dos sintomas, onde a variável explicativa *data_confirmado1* foi modelada com um *spline* de suavização de grau 3 e a variável explicativa *idada_utente_a_data_validacao* foi modelada de forma linear.

Através da análise da Tabela I.19 do Anexo I, referente ao modelo ajustado aos dados restritos dos sintomas, observa-se que as variáveis explicativas *auscultacao_pulmonar_anomala* e praticamente todos os níveis da variável explicativa *descricao_distrito_ocorrencia* não são estatisticamente significantes a um nível de significância de 5%, pelo que se elaborou um teste de razão de verosimilhança ao conjunto dos coeficientes destas duas variáveis

Tabela 4.26: Resultados do *AIC* dos diversos modelos Aditivos Generalizados ajustados na base de dados balanceada dos sintomas.

Grau do *spline*		
idade_utente_a_data_validacao	*data_confirmado1*	*AIC*
-	2	11502.2
2	-	11501.61
2	2	11457.56
-	3	11449
3	-	11483.57
3	2	11439.85
2	3	11405.89
3	3	11388.5

explicativas, tendo-se obtido um valor observado da estatística de teste 239.36 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% não existem evidências estatísticas que nos permitam concluir que o conjunto das estimativas dos coeficientes da variável explicativa *auscultacao_pulmonar_anomala* e dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* sejam exatamente nulos. Posto isto, testou-se se o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* são simultaneamente iguais a zero, pelo que se utilizou mais uma vez o teste de razão de verosimilhança, onde se obteve um valor observado da estatística de teste de 235.35 e um valor do $p\text{-}value$ de 2.2×10^{-16}, pelo que a um nível de significância de 5% existem evidências estatísticas que nos permitem rejeitar a hipótese nula, ou seja, não existem evidências estatísticas que nos permitam concluir que o conjunto dos coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* sejam conjuntamente iguais a 0. Do exposto, não se retirou esta variável do modelo. Desta forma, realizou-se o teste de razão de verosimilhança ao coeficiente da variável explicativa *auscultacao_pulmonar_anomala*, onde se obteve um valor observado da estatística de teste de 2.97 e um valor do $p\text{-}value$ de 0.0851, pelo que a um nível de significância de 5% não existem evidências estatísticas que nos permitam rejeitar a hipótese nula, ou seja, é plausível assumir que o coeficiente da variável explicativa *auscultacao_pulmonar_anomala* é exatamente igual a 0. Posto isto, removeu-se esta variável explicativa do modelo e ajustou-se de novo o modelo.

No novo ajuste do modelo, pode-se observar através da Tabela I.21 do Anexo I, que todas as variáveis que foram modeladas linearmente, somente alguns níveis da variável explicativa *descricao_distrito_ocorrencia*, são estatisticamente significantes a um nível de significância de 5%, pelo que se realizou um teste de razão de verosimilhança sobre os diferentes níveis da variável explicativa *descricao_distrito_ocorrencia*, tendo-se obtido um valor observado da estatística de teste 236.4 e um valor do $p\text{-}value$ de 2.2×10^{-16}, onde se conclui que existem evidencias estatísticas que nos permitem rejeitar a hipótese nula a um nível de significância de 5%, ou seja, que não é plausível assumir que o conjunto dos

coeficientes dos diferentes níveis da variável explicativa *descricao_distrito_ocorrencia* sejam exatamente iguais a 0.

Assim, a variável explicativa *descricao_distrito_ocorrencia* não foi removida do modelo, pelo que o modelo final obtido, quando se ajusta o modelo Aditivo Generalizado à base de dados restrita dos sintomas é constituída por todas as variáveis explicativas à exceção da variável explicativa *auscultacao_pulmonar_anomala*. Na Tabela I.21, do Anexo I, encontram-se as estimativas dos coeficientes das variáveis explicativas do modelo Aditivo Generalizado final obtido, quando ajustado à base de dados restrita dos sintomas.

Por outro lado, quando se ajusta o modelo Aditivo Generalizado à base de dados balanceada dos sintomas, com a variável explicativa *data_confirmado1* modelada com um *spline* de suavização de grau 3 e a variável explicativa *idada_utente_a_data_validacao* modelada de forma linear, também se observa que a variável explicativa *auscultacao_pulmonar_anomala* não é estatisticamente significante a um nível de significância de 5%, pelo que se aplicou o teste de razão de verosimilhança para se averiguar se existe a concordância neste modelo para também eliminar a variável explicativa *auscultacao_pulmonar_anomala*, face ao modelo final ajustado com a base de dados restrita dos sintomas. Aplicando um teste de razão de verosimilhança sobre o coeficiente da variável explicativa *auscultacao_pulmonar_anomala*, obtém-se um valor observado da estatística de teste de 0.31 e um valor do $p\text{-}value$ de 0.5789, pelo que a um nível de significância de 5%, não existem evidências estatísticas que nos permitam rejeitar a hipótese nula, pelo que é plausível assumir que o coeficiente desta variável explicativa é exatamente igual a zero. Posto isto, conclui-se que também se pode remover a variável explicativa *auscultacao_pulmonar_anomala*, do modelo ajustado à base de dados balanceada dos sintomas, pelo que este modelo agora obtido tem as mesmas variáveis que o modelo final ajustado à base de dados restrita dos sintomas.

Assim, uma vez que não se conseguiu remover mais nenhuma variável do modelo ajustado à base de dados restrita dos sintomas, o modelo de treino é constituído por todas as variáveis explicativas do modelo final obtido do ajustamento à base de dados restrita dos sintomas.

Na Figura 4.55, encontra-se representado o efeito não linear, sobre a probabilidade estimada de mortalidade, da variável explicativa *data_confirmado1*, referente ao modelo final obtido na base de dados restrita dos sintomas.

Através da análise da Figura 4.55, pode-se concluir que a variável explicativa *data_confirmado1*, possui um efeito não linear, uma vez que através da análise da Figura 4.55 verifica-se que desde o primeiro caso de confirmação até meados de setembro, (cerca de 150 dias após o primeiro caso confirmado), existe um efeito de decréscimo da chance de morte com o avançar da data de caso, sendo que a partir de meados de setembro até ao início/metade do mês de janeiro, (300/320 dias após o primeiro caso confirmado), existe um efeito de aumento da chance de morte com o avançar da data de caso, sendo que a partir de meados do mês de janeiro até ao último caso confirmado na presente base de dados que é 500 dias após o primeiro caso, existe novamente um efeito de decréscimo da chance de morte com o avançar da data de caso. Estes padrões estão muito certamente

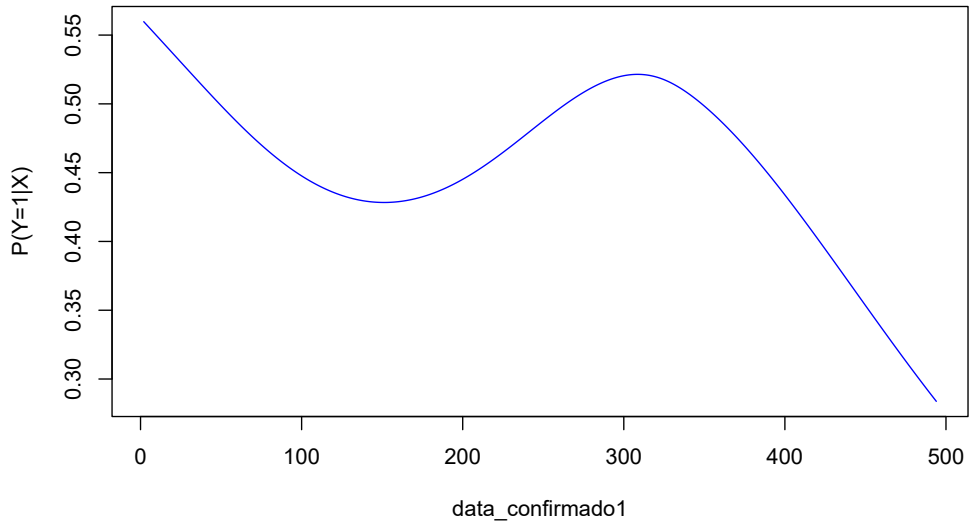


Figura 4.55: Efeito da variável explicativa *data_confirmado1* na probabilidade estimada de mortalidade.

associados com as medidas de contenção que o Governo de Portugal instituiu, como por exemplo no período entre janeiro de 2021 a meados de março de 2021, o confinamento, assim como também no período do início da pandemia, entre meados de março de 2020 a meados de maio de 2020 que impôs o confinamento total, em ambos estes períodos, tem-se um decréscimo do número de casos. Ou seja, da análise do efeito não linear da variável explicativa *data_confirmado1*, pode-se concluir que a *chance* de um indivíduo falecer durante o período entre os meados de setembro e janeiro é mais elevada do que os restantes períodos, uma vez que este período, é caracterizado pelo número crescente de casos.

Através da análise dos *OR*, que se encontra na Tabela I.22, do Anexo I, pode-se concluir o seguinte sobre as variáveis explicativas com os efeitos paramétricos:

- A chance de morrer por COVID-19 é maior 1.11 vezes por cada ano de aumento na idade dos indivíduos.

- As *chances* de falecer devido à COVID-19 entre os indivíduos do sexo masculino é 2.01 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que são do sexo feminino, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *historia_de_febre_ou_calafrios* é 1.52 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *historia_de_febre_ou_calafrios*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *pneumonia* é 1.22 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *pneumonia*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *dispneia* é 3.21 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *dispneia*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *artralgia* é 1.25 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *artralgia*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *irritabilidade_confusao* é 1.39 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *irritabilidade_confusao*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *fraqueza_geral_ou_astenia* é 1.13 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *fraqueza_geral_ou_astenia*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *radiografia_pulmonar_com_alteracoes* é 1.84 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *radiografia_pulmonar_com_alteracoes*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *coma* é 3.79 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *coma*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *taquicardia* é 1.64 vezes maior do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *taquicardia*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *tosse_seca_ou_produtiva* é 0.80 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *tosse_seca_ou_produtiva*, neste presente estudo;
- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *coriza* é 0.55 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *coriza*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *odinofagia* é 0.68 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *odinofagia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *cefaleia* é 0.61 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *cefaleia*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *dor_no_peito* é 0.80 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *dor_no_peito*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *mialgias* é 0.59 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *mialgias*, neste presente estudo;

- As *chances* de falecer devido à COVID-19 entre os indivíduos que apresentem o sintoma *nauseas_vomitos* é 0.87 vezes menor do que as *chances* de falecer devido à COVID-19 entre os indivíduos que não apresentem o sintoma *nauseas_vomitos*, neste presente estudo;

4.3.2.5 Árvore de Classificação

Antes de se treinar o modelo da Árvore de Classificação nos dados de treino referente aos sintomas, primeiro aplicou-se este modelo à base de dados balanceada dos sintomas com todas as variáveis referentes aos sintomas considerados para os ajustamentos do modelo. A Árvore de Classificação apenas pode ser aplicada a dados balanceados, uma vez que um dos pressupostos deste modelo é que as classes da variável resposta têm de ser equilibradas entre si.

A aplicação da Árvore de Classificação aos dados balanceados com todas as variáveis da base de dados balanceada dos sintomas, encontra problemas com as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, nas divisões necessárias, com conclusões que não são corretas. Quando se utiliza estas duas variáveis explicativas, uma vez que como estamos a utilizar o algoritmo CART que utiliza a divisão recursiva binária, ou seja, a divisão em cada uma das variáveis leva à criação de dois novos ramos, em algumas das divisões na variável explicativa *descricao_distrito_ocorrencia*, a árvore está a juntar distritos que tiveram muito poucas mortes com distritos que tiveram muitas mortes e a classificar este ramo como mortes, em vez de apenas considerar os distritos que tiveram muitas mortes. Esta questão, pode dever-se ao facto da base de dados dos sintomas ter sido balanceada, pelo que quando as observações foram removidas, ficamos com uma base de dados em que alguns dos distritos que tinham muitas mortes passam a ter um número

idêntico de mortes a alguns dos outros distritos com poucas mortes. Por exemplo, um dos ramos de divisão da Árvore de Classificação quando utilizada a variável explicativa *descricao_distrito_ocorrencia* continha os seguintes distritos, Beja, Braga, Bragança, Coimbra, Évora, Faro, Guarda, Leiria, Lisboa, Portalegre, Porto Santarém, Setúbal e Viseu e este ramo terminava com uma folha em que esta classificava como morte os indivíduos que tinham sido infetados em cada um destes distritos, contudo os distritos de ocorrência Beja, Faro, Leiria, Santarém, Portalegre e Braga são dos distritos que têm o menor número de casos, enquanto que os restantes distritos descritos no ramo de divisão são os que têm um maior número de mortes, tal como se pode confirmar na Figura 4.14, ou seja, neste presente caso, tem-se que a Árvore de Classificação não divide corretamente os distritos. Já a variável *data_confirmado1* foi removida pelo facto desta variável também se ter perdido um grande número de observações, pelo que os dias que apresentavam o maior número de casos, com o balanceamento da base de dados restrita dos sintomas, ficaram com um número de observações idênticas aos dias que apresentavam um menor número de casos, ou seja, a proporcionalidade de casos diários entre a base de dados restrita e balanceada não se manteve. Por estes motivos não se levou em conta as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, da base de dados balanceada dos sintomas na construção da Árvore de Classificação.

Recorrendo à função *rpart()* da biblioteca *rpart* do *software* R, construiu-se a Árvore de Classificação de tamanho máximo, com o critério de paragem de que o número mínimo de observações em qualquer uma das folhas (nós terminais) tem de ter pelo menos 7 observações. O critério de divisão utilizado na construção da Árvore de Decisão foi o *índice de gini*. Depois de se construir Árvore de Classificação de tamanho máximo utilizou-se a poda de custo de complexidade (Cost complexity pruning), onde o parâmetro de ajuste α, dado na equação 3.94 da secção 3.5, é de $\alpha = 0.0002$.

No Anexo I, encontra-se representada na Figura I.2, a Árvore de Classificação quando aplicada à base de dados de balanceada sem as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, sendo que as variáveis explicativas incluídas na construção da Árvore de Classificação, após se efetuar a poda são:

- *idada_utente_a_data_validacao*;
- *sexo_utente*;
- *radiografia_pulmonar_com_alteracoes*;
- *dispneia*;
- *tosse_seca_ou_produtiva*;
- *historia_de_febre_ou_calafrios*;
- *mialgias*;

- *odinofagia*;
- *cefaleia*;
- *coriza*.

No Anexo I na sub-subsecção I.2.5.1, encontram-se as diversas classificações que resultam do modelo da Árvore de Classificação, quando aplicada à base de dados balanceada dos sintomas.

O modelo de treino foi aplicado com recurso mais uma vez à função *rpart()* da biblioteca *rpart* do *software* R, em que no modelo de treino levou-se em conta todas as variáveis explicativas à exceção das variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, sendo que o critério de paragem para a construção da Árvore de Classificação de tamanho máximo foi mais uma vez, que o número mínimo de observações em qualquer uma das folhas, (nós terminais), tinha de ter pelo menos 7 observações. Mais uma vez utilizou-se o *índice de gini*, como o critério para a construção da Árvore de Classificação. Depois de se construir Árvore de Classificação de tamanho máximo utilizou-se, novamente a poda de custo de complexidade (Cost complexity pruning), para se obter uma Árvore de Classificação final, sendo que o valor do parâmetro de ajuste α é de $\alpha = 0.0002$. As variáveis explicativas incluídas na construção da Árvore de Classificação final na base de dados de treino, após a poda são as seguintes:

- *idada_utente_a_data_validacao*;
- *sexo_utente*;
- *radiografia_pulmonar_com_alteracoes*;
- *dispneia*;
- *cefaleia*;
- *dor_no_peito*;
- *tosse_seca_ou_produtiva*;
- *historia_de_febre_ou_calafrios*;
- *mialgias*;
- *fraqueza_geral_ou_astneia*;
- *coriza*.

4.3.2.6 Floresta Aleatória

No modelo da Floresta Aleatória, tal como no modelo da Árvore de Classificação também se optou por não levar em conta as variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1*, pelos mesmos motivos que no modelo da Árvore de Classificação, já que as Florestas Aleatória utilizam as Árvores de Classificação como os modelos simples de blocos de construção. Antes de se treinar o modelo da Floresta Aleatória com os dados de treino dos sintomas, primeiro ajustou-se o modelo da Floresta Aleatória aos dados balanceados dos sintomas, para discutir qual o número de árvores que deveriam de ser utilizados no modelo de treino. Assim, recorrendo à função *randomForest()* da biblioteca *randomForest* do *sofware* R, ajustou-se o modelo da Floresta Aleatória com um número de sub-Árvores de 250, para se analisar se o erro OOB geral estabilizava, pois se o erro OOB geral se estabilizar com este número de sub-Árvores, então poderemos treinar o modelo com este número de sub-Árvores. Na Figura 4.56, encontra-se o erro OOB da Floresta Aleatória ajustada aos dados balanceados dos sintomas, considerando as 250 Árvores de Classificação.

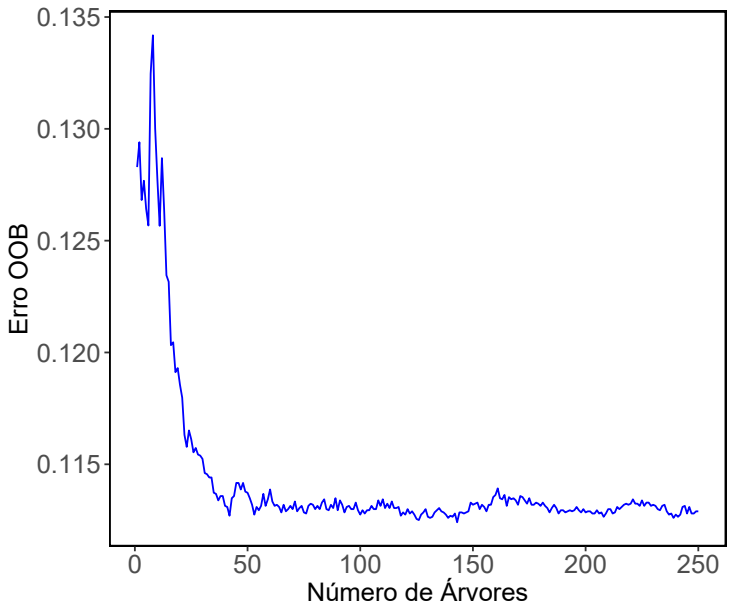


Figura 4.56: Erro OOB Floresta Aleatória, base de dados balanceada dos sintomas.

Através da análise da Figura 4.56, observa-se que o erro OOB geral, tende-se a estabilizar, pois a partir das 100 Árvores, o erro OOB geral não apresenta grandes variações. Uma vez que o erro OOB geral da Floresta Aleatória ajustada aos dados balanceados dos sintomas se considera as 250 Árvores de Classificação, então o modelo da Floresta Aleatória foi treinado com os dados de treino dos sintomas, considerando as 250 Árvores de Classificação.

No gráfico da Figura 4.57, encontra-se a importância de cada uma das variáveis explicativas através da diminuição média do *índice de gini*, onde se observa que as variáveis explicativas que este modelo, quando ajustado na base de dados balanceada dos sintomas, considera como sendo as mais importantes são as seguintes:

- *idade_utente_a_data_validacao*;
- *dispneia*;
- *radiografia_pulmonar_com_alteracoes*;
- *cefaleia*;
- *coriza*;
- *mialgias*;
- *odinofagia*;
- *pneumonia*;
- *sexo_utente*.

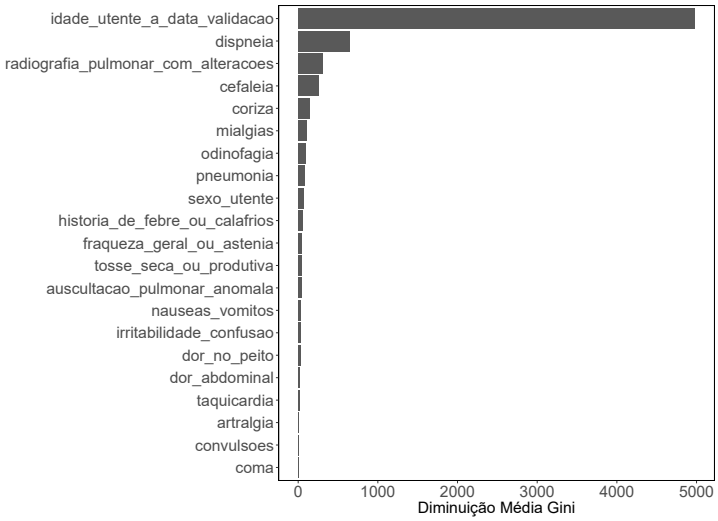


Figura 4.57: Importância variáveis explicativas Floresta Aleatória, base de dados balanceada dos sintomas.

Por outro lado, através da análise do gráfico da Figura 4.57, observa-se que as variáveis explicativas *convulsoes*, *coma* e *artralgia* não são muito relevantes.

Posto isto, tem-se que o modelo de treino, aplicado à base de dados de treino dos sintomas, leva em conta todas as variáveis explicativas à exceção das variáveis explicativas *descricao_distrito_ocorrencia* e *data_confirmado1* sendo constituído por 250 Árvores de Classificação como o bloco de construção deste modelo.

4.3.2.7 Redes Neuronais

Neste presente trabalho, no caso dos sintomas, apenas se consideraram as Redes Neuronais *feed-forward*.

Para se treinar o modelo das Redes Neuronais, com a base de dados de treino dos sintomas, primeiro tem que se selecionar o número de camadas ocultas e também o número de neurónios em cada uma das camadas ocultas. Para isto, elaboraram-se diversos modelos, e para isso cada um dos diversos modelos construídos utilizou-se a validação cruzada com 10 grupos para se calcular a estimativa de desempenho de cada um dos modelos, sendo que a métrica que se utilizou na validação cruzada foi a métrica *accuracy*.

Os diversos modelos considerados foram todas as combinações possíveis entre o número de camadas ocultas a variar entre 1 e 2 e o número de nós em cada uma das camadas ocultas a variar entre 1 e 50, onde as funções de ativação escolhidas tanto para a primeira camada oculta como para a segunda camada oculta foram as funções de ativação *ReLU* e para a camada de saída utilizou-se a função de ativação *sigmoid*. Ou seja, em primeiro lugar consideraram-se todas as Redes Neuronais com uma única camada oculta, onde o número de neurónios de cada uma destas Redes Neuronais variava entre 1 e 50, e a função de ativação utilizada na primeira camada oculta é a função de ativação *ReLU* e a função de ativação utilizada na camada de saída é a função de ativação *sigmoid*. Depois de considerar estes modelos, consideraram-se todos os modelos com duas camadas ocultas, onde o número de neurónios em cada uma destas camadas ocultas variavam entre 1 e 50, com as mesmas funções de ativação anteriormente escolhidas.

Assim, resumindo em cada um dos modelos referidos anteriormente utilizou-se a validação cruzada com 10 grupos para se calcular a estimativa de desempenho de cada um dos modelos, sendo que a métrica utilizada foi a *accuracy*. Uma vez que a métrica utilizada é a *accuracy*, então na validação cruzada iremos utilizar a base de dados balanceada dos sintomas, uma vez que na secção 3.10 se referiu que esta métrica não é viável quando a base de dados é desequilibrada.

De referir que neste presente trabalho, apenas se consideraram as funções de ativação *ReLU* para as camadas ocultas, em vez de se testarem mais modelos com outras diferentes funções de ativação. Este facto deve-se sobretudo ao tempo computacional deste processo ser um pouco longo, uma vez que somente utilizando as funções de ativação *ReLU* para as camadas ocultas temos um tempo computacional de cerca de 12 horas.

Para se ajustar cada um dos modelos anteriormente descritos recorreu-se à biblioteca *Keras* do *sofware* R. Para se ajustar cada uma das Redes Neuronais, utilizou-se o *enviesamento*, onde as estimativas iniciais dos *enviesamento* foi 0 e as estimativas iniciais dos restantes parâmetros foram dadas através do indicador uniforme *Glorot*, já que se utilizou o método de estimação do *gradiente descendente estocástico*.

Posto isto, tem-se que o melhor modelo de Redes Neuronais selecionado, é o que apresenta duas camadas ocultas, sendo que o número de neurónios na primeira camada oculta é de 30 neurónios e o número de neurónios na segunda camada oculta é de 20 neurónios, sendo a estimativa de validação cruzada com 10 grupos foi de 0.8558, utilizando a métrica *accuracy*.

Desta forma, o modelo de Redes Neuronais foi treinado com a base de dados de treino dos sintomas, onde se considerou uma Rede Neuronal composta por uma camada

de entrada, duas camadas ocultas e uma camada de saída. O número de neurónios na camada de entrada é de 41, uma vez que a variável *descricao_distrito_ocorrencia* é uma variável categórica com 20 níveis, ou seja, passando esta variável para variáveis *dummy* ficamos com 19 níveis a que soma as restantes 22 variáveis explicativas, pois a base de dados de treino dos sintomas é constituída por 23 variáveis explicativas. Já o número de neurónios nas duas camadas ocultas é de 30 e 20, respetivamente, sendo que em ambas as camadas ocultas se utilizou a função de ativação *ReLU*. Por último a camada de saída é apenas constituída por um neurónio, sendo utilizada nesta camada a função de ativação *sigmoid*. Na estimação da Rede Neuronal, com os dados de treino utilizou-se o *enviesamento*, onde as estimativas iniciais dos *enviesamento* foram 0 e as estimativas iniciais dos restantes parâmetros foram dadas através do indicador uniforme *Glorot*.

4.4 Resultados Preditivos dos Modelos

Após os treinos de cada um dos modelos descritos anteriormente, nas subsecções 4.4.1 e 4.4.2, encontram-se os resultados em termos preditivos de cada um dos diferentes modelos, onde se utilizaram as métricas *accuracy*, *sensibilidade*, *especificidade*, *precisão*, *recall*, *kappa*, medida-*F* (*F1*) e a área sob a curva ROC, para as comparações referentes tanto aos sintomas como às comorbilidades.

4.4.1 Resultados Preditivos dos Modelos Referentes às Comorbilidades

Na Tabela 4.27, encontram-se os resultados das diferentes métricas utilizadas para a avaliação da performance preditiva de cada um dos diferentes modelos treinados, com a base de dados de treinos das comorbilidades.

Tabela 4.27: Métricas para avaliação dos modelos referentes às comorbilidades.

Métrica	*accuracy*	*sensibilidade*	*especificidade*	*precisão*	*recall*	*kappa*	*F1*
Regressão Logística	0.8918	0.8611	0.9210	0.9118	0.8611	0.7831	0.8857
Regressão Logística com Interações	0.8930	0.8581	0.9262	0.9169	0.8581	0.7855	0.8865
Modelo Aditivo Generalizado	0.8908	0.8586	0.9214	0.9121	0.8586	0.7812	0.8845
Lasso de Grupo	0.8920	0.8601	0.9224	0.9132	0.8601	0.7836	0.8858
Árvore de Decisão	0.8815	0.8495	0.9119	0.9015	0.8495	0.7626	0.8748
Floresta Aleatória	0.8891	0.8425	0.9333	0.9231	0.8425	0.7776	0.8810
Redes Neuronais	0.8696	0.8370	0.9005	0.8887	0.8370	0.7386	0.8621

Através da análise dos valores das diferentes métricas, que se encontram na Tabela 4.27, verifica-se que todos os modelos têm um valor da métrica estatística *kappa* entre os 0.73 e os 0.79, donde se conclui que existe uma boa concordância entre as previsões dos diversos modelos e os verdadeiros valores. Também através da análise da métrica *F1*, verifica-se que

o valor desta métrica em todos os modelos é elevado, dado que estão próximos do valor 0.9, pelo que se pode afirmar que todos os modelos têm um alto desempenho de classificação. Do mesmo modo para a métrica *accuracy*, que é uma medida de avaliação do desempenho global do modelo, verifica-se que em todos os modelos considerados os valores da *accuracy* estão próximos de 0.9, pelo que se pode considerar que todos os modelos estão a classificar as observações na classe correta, donde estes modelos são modelos que apresentam um bom desempenho globalmente. No entanto, pode-se observar que o modelo das Redes Neuronais é o modelo que apresenta o menor valor nas métricas *accuracy*, estatística *kappa* e $F1$, pelo que se pode considerar que este modelo é o que apresenta o menor desempenho global de classificação, através da análise destas métricas.

Quando se faz uma análise individual em termos preditivos da classe de sucesso, que é a nossa classe de interesse, verifica-se que os modelos Árvores de Decisão, Florestas Aleatórias e Redes Neuronais são os modelos que apresentam o menor valor da métrica *sensibilidade*, pelo que estes são os modelos que têm um menor desempenho na classificação da classe de sucesso em relação aos outros modelos, sendo que o melhor modelo preditivo para prever a classe de sucesso é o modelo de Regressão Logística, uma vez que foi neste modelo que se obteve um valor maior valor na métrica *sensibilidade*, sendo este valor de 0.8611.

Já no que diz respeito à métrica *recall*, observamos que os modelos Árvores de Decisão, Florestas Aleatórias e Redes Neuronais, são os modelos que não conseguem identificar tantas observações que pertencem à classe de sucesso quanto as existentes, em comparação com os demais modelos, sendo que o modelo de Regressão Logística é o modelo que consegue identificar mais observações que pertencem à classe positiva, de entre as existentes nos dados de teste, uma vez que foi no modelo de Regressão Logística que se obteve o valor mais elevado da métrica *recall*.

Por outro lado, o modelo de Regressão Logística apresenta um valor da métrica *precisão* de 0.9118, ou seja, isto quer dizer que de todos os elementos que o modelo classificou como pertencente à classe positiva cerca de 91.2% desses valores eram realmente positivos. No entanto, o melhor valor da métrica *precisão* é obtida no modelo das Florestas Aleatórias que é de 0.9231 que é um valor mais elevado que o valor obtido no modelo da Regressão Logística, ou seja, de todas as observações que o modelo da Floresta Aleatória classificou como pertencentes à classe positiva 92.31% delas pertenciam realmente à classe positiva. No obstante, o modelo das Florestas Aleatórias apresenta um valor menor nas métricas *sensibilidade* e *recall*, quando comparado com o modelo de Regressão Logística. Ou seja, apesar do modelo das Florestas Aleatórias apresentar um maior valor na métrica *precisão*, este modelo não consegue captar tantas observações que realmente pertencem á classe positiva quando comparado com o modelo de Regressão Logística, ou seja, o modelo das Florestas Aleatórias é menos amplo que o modelo de Regressão Logística, pois nas florestas Aleatórias, obtivemos que de todas as observações que realmente pertencem à classe positiva somente cerca de 84.3% destas é que estão a ser classificadas como pertencendo à classe positiva, enquanto que no modelo de Regressão Logística temos que

de todas as observações que realmente pertencem á classe positiva cerca de 86.1% destas é que estão a ser classificadas como pertencendo à classe positiva. Para o modelo do LASSO de Grupo, passa-se o mesmo, pois este modelo também apresenta um valor mais elevado na métrica *precisão* que o valor obtido no modelo da Regressão Logística, mas nas métricas *sensibilidade* e *recall* o LASSO de Grupo apresenta valores inferiores aos valores obtidos nestas métricas no modelo de Regressão Logística.

Na Figura 4.58, encontra-se representada as curva ROC de cada um dos modelos, assim como a área sob a curva ROC (AUC), de cada um destes modelos.

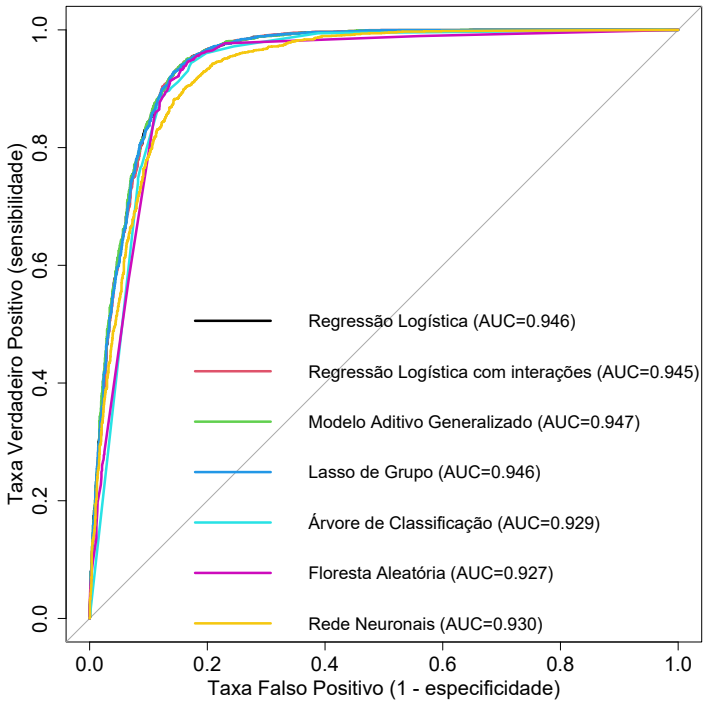


Figura 4.58: Curva ROC e AUC dos diversos modelos considerados para os dados das comorbilidades.

Através da análise da 4.58, observamos que em termos da métrica AUC, cada um dos modelos considerados têm ótima capacidade preditiva, sendo que os modelos de Regressão Logística, Regressão Logística com interações, o modelo Aditivo Generalizado e o LASSO de Grupo, são os modelos que apresentam os valores mais altos de AUC, sendo que cada um destes modelos apresenta um valor de AUC de 0.94.

Do exposto, resulta que os modelos de Regressão Logística, Regressão Logística com interações e os modelos Aditivos generalizados, podem ser considerados os melhores modelos em termos preditivos, levando em conta a avaliação global do desempenho do modelo, sendo que em termos de predição da classe de sucesso, o melhor modelo de entre os demais é o modelo de Regressão Logística, pois é modelo que apresenta um maior valor das métricas *recall* e *sensibilidade*, e por outro lado na métrica *precisão*, este modelo apresenta um valor de 0.9118, que é um valor bastante elevado.

4.4.2 Resultados Preditivos dos Modelos Referentes aos Sintomas

Na Tabela 4.28, encontram-se os resultados das diferentes métricas utilizadas para a avaliação da performance preditiva de cada um dos diferentes modelos treinados, com a base de dados de treino dos sintomas.

Tabela 4.28: Métricas para avaliação dos modelos referentes aos sintomas.

Métrica	*accuracy*	*sensibilidade*	*especificidade*	*precisão*	*recall*	*kappa*	*F1*
Regressão Logística	0.8964	0.8656	0.9257	0.9171	0.8656	0.7924	0.8906
Regressão Logística com interações	0.8964	0.8616	0.9295	0.9207	0.8616	0.7924	0.8902
Modelo Aditivo Generalizado	0.8964	0.8626	0.9286	0.9198	0.8626	0.7924	0.8903
Lasso de Grupo	0.8962	0.8636	0.9271	0.9184	0.8636	0.7919	0.8902
Árvore de Decisão	0.8920	0.8546	0.9276	0.9181	0.8546	0.7836	0.8852
Floresta Aleatória	0.8947	0.8521	0.9352	0.9259	0.8521	0.7889	0.8874
Redes Neuronais	0.8676	0.8420	0.8919	0.8809	0.8420	0.7347	0.8610

Através da análise dos valores das diferentes métricas, que se encontram na Tabela 4.28, verifica-se que todos os modelos têm um valor da métrica estatística *kappa* entre os 0.73 e os 0.80, donde se conclui que existe uma boa concordância entre as previsões dos diversos modelos e os verdadeiros valores. Também através da análise da métrica *F1*, verifica-se que o valor desta métrica em todos os modelos è elevado, dado que estes estão próximos do valor 0.9, pelo que se pode afirmar que todos os modelos têm um alto desempenho de classificação. Do mesmo modo para a métrica *accuracy*, que é uma medida de avaliação do desempenho global do modelo, verifica-se que em todos os modelos considerados os valores da *accuracy* estão próximos de 0.9, pelo que se pode considerar que todos os modelos estão a classificar as observações na classe correta, pelo que estes modelos são modelos que apresentam um bom desempenho globalmente. No entanto, pode-se observar que o modelo das Redes Neuronais é o modelo que apresenta o menor valor nas métricas *accuracy*, estatística *kappa* e $F1$, pelo que se pode considerar que este modelo é o que apresenta o menor desempenho global de classificação, através da análise destas métricas.

No entanto, quando se faz uma análise individual em termos preditivos da classe de sucesso, verifica-se que os modelos Árvores de Decisão, Florestas Aleatórias e Redes Neuronais são os modelos que apresentam o menor valor da métrica *sensibilidade*, pelo que estes modelos são os modelos que têm um menor desempenho na classificação da classe de sucesso, sendo que o melhor modelo preditivo para prever a classe de sucesso é o modelo da Regressão Logística.

Já no que diz respeito à métrica *recall*, observamos que os modelos Árvores de Decisão, Florestas Aleatórias e Redes Neuronais, são os modelos que não conseguem identificar tantas observações que pertencem à classe de sucesso quanto as existentes, em comparação

com os demais modelos, sendo que o modelo de Regressão Logística é o modelo que consegue identificar mais observações nessas condições.

Na Figura 4.58, encontra-se representada as curva ROC de cada um dos modelos, assim como também se encontra representada a área sob a curva ROC (AUC), de cada um destes modelos.

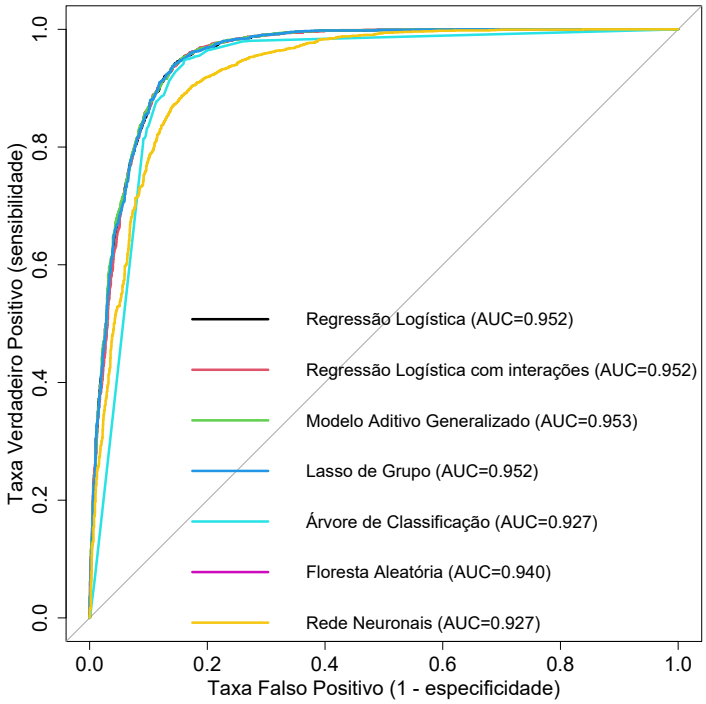


Figura 4.59: Curva ROC e AUC dos diversos modelos considerados para os dados dos sintomas.

Através da análise da Figura 4.59, observamos que em termos da métrica AUC, cada um dos modelos considerados têm ótima capacidade preditiva, sendo que os modelos de Regressão Logística, Regressão Logística com interações, o modelo Aditivo Generalizado e o LASSO de Grupo, são os modelos que apresentam os valores mais altos de AUC, sendo que os modelos Regressão Logística, a Regressão Logística considerando as interações e o LASSO de Grupo apresentam um valor de AUC de 0.952, enquanto que o modelo o modelo Aditivo Generalizado apresenta um valor de AUC de 0.953.

Do exposto, resulta que o modelos de Regressão Logística, Regressão Logística com interações e os modelos Aditivos Generalizados, podem ser considerados os melhores modelos em termos preditivos, levando em conta a avaliação global do desempenho do modelo, sendo que em termos de predição da classe de sucesso, o melhor modelo de entre os demais é o modelo de Regressão Logística.

5

Conclusões

Como referido anteriormente, o principal objetivo desta tese é aplicar diferentes modelos de Aprendizagem de Automática, para a estimação e previsão da mortalidade por COVID-19 na população Portuguesa, recorrendo à base de dados disponibilizada pelo SINAVE. Os dados disponibilizados referem-se a todos os casos de COVID-19 confirmados laboratorialmente e notificados, por data, em Portugal durante o primeiro ano e pouco de pandemia (início de março de 2020 a julho de 2021), bem como alguma informação clínica associada. Nestes dados disponibilizados, as variáveis podem ser divididas em 6 grupos, entre os quais, dados demográficos, dados referentes a datas, tal como a data de início dos sintomas, data da confirmação de infeção ou a data de óbito, dados sobre o desfecho (falecimento ou não), dados referentes a sintomas, dados referentes a comorbilidades, dados ao tipo do vírus e dados demográficos, tais como a descrição do distrito onde o paciente reside e a descrição do concelho em que o utente foi infetado.

O presente trabalho, tem algumas limitações e alguns pressupostos. As limitações deste trabalho, predem-se com o facto de na base de dados existirem várias variáveis com elevado número de observações omissas, o que foi ultrapassado através da remoção de algumas das variáveis explicativas que tinham um maior número desse tipo de observações. Apesar de muitos dos estudos destacados no capítulo 2 optarem por efetuar imputação da base de dados em situações análogas, neste trabalho não se optou pela imputação, uma vez que existem algumas variáveis com praticamente todos os valores omissos. Nestes dados, existem variáveis referentes aos sintomas que têm um elevado número de observações omissas, pelo que para essas variáveis, neste estudo, fez-se a suposição que se essas variáveis não estavam preenchidas é porque o indivíduo não tinha esses sintomas. Feita estas imputação, excluíram-se todas as observações que tinham pelo menos um valor omisso em alguma variável. Feita a parte da limpeza da base de dados e a suposição referida anteriormente ficou-se com uma base de dados com 466458 (51%) observações, compreendendo variáveis demográficas, datas, sintomas, comorbilidades e geográficas.

Em muitos estudos revistos comparáveis ao que aqui se apresenta, os autores modelaram a mortalidade/morbilidade por COVID-19 incluindo simultaneamente sintomas e comorbilidades nos seus modelos. Ao contrário desses estudos, decidimos separar a

explicação da mortalidade por sintomas e comorbilidades, uma vez que facilmente se está sob a presença de multicolinearidades, uma vez que alguns dos sintomas podem estar associados a alguma das comorbilidades, afetando os ajustes de alguns modelos como sendo a Regressão Logística. Posto isto dividiu-se a base de dados em base de dados de sintomas e base de dados das comorbilidades, em que cada uma destas incluía também as variáveis demográficas, geográficas e datas. No entanto, ambas as bases de dados consideradas têm a limitação de serem desbalanceadas, o que nos modelos de Aprendizagem Automática pode resultar em conclusões enganadoras nos resultados preditivos dos modelos. Para se ultrapassar esta limitação, muitos dos artigos mencionados no capítulo 2 utilizam a técnica SMOTE. Neste caso, não se utilizou a técnica SMOTE, uma vez que este método gera novas observações sintéticas para a classe minoritária, com base nas observações existentes na classe minoritária, pelo que passamos a ter uma base de dados em que as observações deixam de ser independentes e para muitos dos modelos considerados um dos seus pressupostos é a independência das observações. A técnica que nós utilizamos foi a técnica da subamostragem aleatória, que remove observações aleatoriamente da classe maioritária, sendo que neste caso as observações que ficam na base de dados são na mesma independentes. No entanto, este método tem a desvantagem de eliminar padrões nos dados, pelo que se optou por selecionar as variáveis explicativas relevantes através da base de dados maior, não balanceada.

Um dos outros pressupostos que se está a assumir é que praticamente nenhum dos indivíduos do presente estudo é vacinado, uma vez que os dados com que estamos a trabalhar referem-se entre o início de março de 2020 a meados de julho de 2021 e Portugal no fim desse período apenas tinha cerca de 70% da população vacinada com a primeira dose, ou seja, ainda não estavam totalmente vacinados (Direção Geral de Saúde).

Apesar das limitações, todos os modelos construídos referentes ao estudo das comorbilidades, são bons em termos preditivos, sendo que os modelos de Regressão Logística, Regressão Logística com interações, o modelo Aditivo Generalizado e o LASSO de Grupo são os modelos que apresentam melhores desempenhos preditivos em termos da área sob a curva ROC. Adicionalmente, pela avaliação da métrica *accuracy*, uma métrica de avaliação do desempenho global do modelo, todos os modelos estão a classificar genericamente as observações na classe correta, pelo que estes modelos são modelos que apresentam globalmente um bom desempenho, sendo que o modelo da Regressão Logística com interações é o que apresenta o melhor desempenho. Aliás, este modelo é o que apresenta os melhores valores nas métricas *accuracy*, estatística *kappa* e na métrica $F1$. Por outro lado, se estivermos interessados apenas na classificação da classe de interesse, morte por COVID-19 neste caso, o melhor modelo preditivo é o modelo da Regressão Logística, pelo que este modelo é o modelo que classifica mais corretamente as observações nesta classe. Assim, o modelo da Regressão Logística foi o melhor modelo para prever a classe de interesse, morte por COVID-19 neste caso, entre os demais considerados, para a base de dados das comorbilidades.

Já nos modelos que dizem respeito aos sintomas, constata-se que através da análise de

todas as métricas todos os modelos construídos são bons em termos preditivos, uma vez que todos os modelos apresentam um valor da área sob a curva ROC maior que 0.9, sendo que os modelos de Regressão Logística, Regressão Logística com interações, o modelo Aditivo Generalizado e o LASSO de Grupo são os modelos que apresentam os melhores desempenhos. Pela avaliação da métrica *accuracy*, todos os modelos estão a classificar as observações na classe correta, pelo que estes são modelos que apresentam um bom desempenho global, sendo que o modelo modelo Aditivo Generalizado, o modelo de Regressão Logística e o Regressão Logística com interações são os que apresentam o melhor valor, pelo que podem ser considerados os melhores modelos em termos preditivos, levando em conta a avaliação global do desempenho do modelo. Não obstante, se estivermos interessados apenas na classificação da classe de interesse, morte por COVID-19, o melhor modelo preditivo é o modelo da Regressão Logística, uma vez que este é o modelo que classifica mais corretamente as observações nesta classe.

As variáveis explicativas idade e sexo são duas variáveis explicativas que estão muito relacionadas com a mortalidade por COVID-19, sendo que a idade é um fator bastante determinante, como se constatou através da análise dos *OR* no modelo de Regressão Logística e no modelo Aditivo Generalizado. Por outro lado, as variáveis idade e sexo, são variáveis que se revelam significativas em todos os modelos finais da Regressão Logística e da Regressão Logística com interações. Nos estudos referidos no capítulo 2, muitos dos autores também chegaram à conclusão de que a idade de um paciente era um fator importante que está associado à mortalidade por COVID-19.

Com base nos modelos finais da Regressão Logística, Regressão Logística com interações e o modelo Aditivo Generalizado ajustados à base de dados das comorbilidades, constata-se que as comorbilidades referentes aos diabetes, neoplasia, doença neurológica ou neuromuscular crónica, doença pulmonar crónica, patologia hepática, doença hematológica crónica, doença renal crónica, a deficiência neurológica crónica, a insuficiência cardíaca e a insuficiência renal crónica são as comorbilidades que estão mais associadas à mortalidade de indivíduos que testem positivo à COVID-19, uma vez que estas variáveis estão presentes nos modelos finais, e se revelam significativas. Por outro lado, a interação entre a idade do paciente e a comorbilidades diabetes, também se mostra relevante, assim como a interação entre a comorbilidade diabetes e a comorbilidade neoplasia, uma vez que estas interações se encontram no modelo final da Regressão Logística com interações e são significativas.

Do mesmo modo, os sintomas dispneia, radiografia pulmonar com alterações, febre ou calafrios, taquicardia, se o utente está em coma, irritabilidade confusão, fraqueza geral ou astneia, pneumonia, as artralgias, náuseas ou vómitos, odinofagia, tosse seca ou produtiva, cefaleia, mialgias e a coriza destacam-se como sendo os sintomas que estão mais associadas à mortalidade de indivíduos que testem positivo à COVID-19, uma vez que estas variáveis estão presentes nos modelos de regressão finais e se revelam significativas (Regressão Logística, Regressão Logística com interações e modelo Aditivo Generalizado). Nos estudos referidos no capítulo 2, os autores também referem que sintomas como a

artralgia, a pneumonia, odinofagia, tosse e a dispneia são fatores relevantes, tal como obtivemos neste estudo.

De referir ainda que neste estudo, os resultados preditivos das diferentes métricas consideradas para o modelo das Redes Neuronais ficam bastante distantes dos resultados obtidos pelos autores dos artigos mencionados no capítulo 2, sendo que neste estudo podemos considerar esses modelos como o pior modelo em termos preditivos, tanto para os sintomas como para as comorbilidades, em contraste com os estudos vistos no capítulo 2, em que este é genericamente o melhor modelo em termos preditivos do que os modelos de regressão.

Não se determinaram os *OR* no modelo de Regressão Linear com interações, que se poderiam comparar com os *OR* do modelo de de Regressão Logística e com o modelo Aditivo Generalizado, de modo a chegar a uma conclusão sobre quais as comorbilidades e os sintomas que poderiam ter um papel de proteção à mortalidade por COVID-19, ficando como trabalho futuro. Adicionalmente, também se poderá vir a averiguar se estes modelos ainda continuam a ter um bom comportamento preditivo com dados referentes a pacientes vacinados, assim como ver se os fatores das comorbilidades e dos sintomas ainda se mantém os mesmos.

Bibliografia

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/master/template.pdf (ver p. ii).

[2] M. Pourhomayoun e M. Shakibi. «Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making». Em: *Smart Health* 20.December 2020 (2021), p. 100178. ISSN: 23526483. DOI: 10.1016/j.smhl.2020.100178. URL: https://doi.org/10.1016/j.smhl.2020.100178 (ver pp. 1, 4).

[3] R. R. Assessment. «Coronavirus disease 2019 (COVID-19) in the EU/EEA and the UK–ninth update». Em: *European Centre for Disease Prevention and Control: Stockholm* (2020) (ver p. 1).

[4] K. Moulaei et al. «Comparing machine learning algorithms for predicting COVID-19 mortality». Em: *BMC Medical Informatics and Decision Making* 22.1 (2022), pp. 1–12. ISSN: 14726947. DOI: 10.1186/s12911-021-01742-0. URL: https://doi.org/10.1186/s12911-021-01742-0 (ver pp. 6, 94).

[5] C. Feng, G. Kephart e E. Juarez-Colunga. «Predicting COVID-19 mortality risk in Toronto, Canada: a comparison of tree-based and regression-based machine learning methods». Em: *BMC Medical Research Methodology* 21.1 (2021), pp. 1–15. ISSN: 14712288. DOI: 10.1186/s12874-021-01441-4 (ver pp. 7, 95).

[6] D. Incerti et al. «Prognostic model to identify and quantify risk factors for mortality among hospitalised patients with COVID-19 in the USA». Em: *BMJ Open* 11.4 (2021), pp. 1–11. ISSN: 20446055. DOI: 10.1136/bmjopen-2020-047121 (ver p. 9).

[7] S. S. Aljameel et al. «Machine Learning-Based Model to Predict the Disease Severity and Outcome in COVID-19 Patients». Em: *Scientific Programming* 2021 (2021) (ver p. 10).

[8] Q. A. R. Adib et al. «Prediction Model for Mortality Analysis of Pregnant Women Affected with COVID-19». Em: *24th International Conference on Computer and Information Technology, ICCIT 2021* (2021). DOI: 10.1109/ICCIT54785.2021.9689824 (ver p. 12).

[9] S. Zhang et al. «Identification and validation of prognostic factors in patients with COVID-19: A retrospective study based on artificial intelligence algorithms». Em: *Journal of Intensive Medicine* 1.2 (2021), pp. 103–109. ISSN: 2667100X. DOI: 10.1016/j.jointm.2021.04.001. URL: https://doi.org/10.1016/j.jointm.2021.04.001 (ver p. 13).

[10] W. Conover. *Practical nonparametric statistics*. 3. ed. Wiley series in probability and statistics. Wiley, 1999. VIII, 584. ISBN: 0471160687 (ver pp. 15, 16).

[11] A. J. Scott, D. W. Hosmer e S. Lemeshow. *Applied Logistic Regression.* Vol. 47. 4. 1991, p. 1632. ISBN: 9780470582473. DOI: 10.2307/2532419 (ver pp. 17, 18, 20–28).

[12] T. Hastie et al. *An introduction to statistical learning (2nd ed.)* Vol. 102. 2021, p. 618. ISBN: 9780387781884. arXiv: 1011.1669v3 (ver pp. 28–34, 38–47, 49, 51–54, 59–62, 69–73, 77–79, 81, 84, 145, 207).

[13] R. Shrinkage. «Regression Shrinkage and Selection via the Lasso Author (s): Robert Tibshirani Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 58 , No . 1 (1996), Published by : Wiley for the Royal Statistical Society Stable URL». Em: 58.1 (2016), pp. 267–288 (ver p. 31).

[14] F. Khan et al. «A Comparison of Autometrics and Penalization Techniques under Various Error Distributions: Evidence from Monte Carlo Simulation». Em: *Complexity* 2021 (2021). ISSN: 10990526. DOI: 10.1155/2021/9223763 (ver p. 31).

[15] G. M. Lee, K. Yeon e H. Kim. «A Comparative Study on Robust Regression Methods». Em: *Iarjset* 3.8 (2016), pp. 166–170. DOI: 10.17148/iarjset.2016.3830 (ver p. 31).

[16] M.-b. Geostatistics, P. J. Diggle e. «Regularization Paths for Generalized Linear Models via Coordinate Descent». Em: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. ISSN: 0014-4886. arXiv: 0908.3817 (ver p. 32).

[17] T. Hastie, R. Tibshirani e J. Friedman. *The Elements of Statistical Learning*. 2ª ed. Vol. 26. Springer Series in Statistics 4. New York, NY: Springer New York, 2009, pp. 505–516. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7. URL: http://link.springer.com/10.1007/978-0-387-84858-7 (ver pp. 33–36, 38, 41, 42, 44–46, 49, 59–61, 78, 81).

[18] R. T. Jerome Friedman Trevor Hastie. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*. R package version 4.1.4. 2022. URL: https://CRAN.R-project.org/package=glmnet (ver p. 34).

[19] L. Meier, S. Van De Geer e P. Bühlmann. «The group lasso for logistic regression». Em: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 70.1 (2008), pp. 53–71. ISSN: 13697412. DOI: 10.1111/j.1467-9868.2007.00627.x (ver pp. 35, 36).

[20] P. Breheny e J. Huang. «Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors». Em: *Statistics and Computing* 25.2 (2015), pp. 173–187. ISSN: 15731375. DOI: 10.1007/s11222-013-9424-2. arXiv: 1209.2160 (ver pp. 35, 37, 204).

[21] R. K. Patrick Breheny Yaohui Zeng. *grpreg: Regularization Paths for Regression Models with Grouped Covariates*. R package version 3.4.0. 2021. URL: https://CRAN.R-project.org/package=grpreg (ver pp. 37, 204, 205).

[22] T. Hastie e R. Tibshirani. «Generalized Additive Models». Em: *Statistical Science* 1.3 (1986), pp. 297–310. DOI: 10.1214/ss/1177013604. URL: https://doi.org/10.1214/ss/1177013604 (ver pp. 46, 47).

[23] S. K. Sapra. «Generalized Additive Models in Business and Economics». Em: *International Journal of Advanced Statistics and Probability* 1.3 (2013), pp. 64–81. DOI: 10.14419/ijasp.v1i3.1022 (ver p. 47).

[24] S. Clifford et al. «Using the Generalised Additive Model to model the particle number count of ultrafine particles». Em: *Atmospheric Environment* 45.32 (2011), pp. 5934–5945. ISSN: 13522310. DOI: 10.1016/j.atmosenv.2011.05.004. URL: http://dx.doi.org/10.1016/j.atmosenv.2011.05.004 (ver p. 47).

[25] T. Hastie. *gam: Generalized Additive Models*. R package version 1.20.1. 2022. URL: https://CRAN.R-project.org/package=gam (ver p. 48).

[26] P. D. Mining e P. Hagenlocher. «Decision Tree Learning». Em: () (ver pp. 48–50, 55, 56).

[27] S. Singh. «Comparative Study Id3 , Cart And C4 . 5 Decision Tree Algorithm : A Survey». Em: *International Journal of Advanced Information Science and Technology (IJAIST)* 3.7 (2014), pp. 47–52. DOI: 10.15693/ijaist/2014.v3i7.47-52 (ver pp. 48, 49, 51, 54, 57).

[28] X. Meng et al. «Construction of decision tree based on C4.5 algorithm for online voltage stability assessment». Em: *International Journal of Electrical Power and Energy Systems* 118.July 2019 (2020), p. 105793. ISSN: 01420615. DOI: 10.1016/j.ijepes.2019.105793. URL: https://doi.org/10.1016/j.ijepes.2019.105793 (ver p. 49).

[29] F. M. Javed Mehedi Shamrat et al. «Performance Evaluation Among ID3, C4.5, and CART Decision Tree Algorithm». Em: *Lecture Notes in Networks and Systems* 317.March 2021 (2022), pp. 127–142. ISSN: 23673389. DOI: 10.1007/978-981-16-5640-8_11 (ver p. 49).

[30] N. Speybroeck. «Classification and regression trees». Em: *International Journal of Public Health* 57.1 (2012), pp. 243–246. ISSN: 1420911X. DOI: 10.1007/s00038-011-0315-z (ver p. 51).

[31] F. J. Reynara, S. Carolina e I. N. Simbolon. «The Comparison of C4. 5 and CART (Classification and Regression Tree) Algorithm in Classification of Occupation for Fresh Graduate». Em: *Proceedings of the 4th International Conference on Vocational Education and Technology, IConVET 2021, 27 November 2021, Singaraja, Bali, Indonesia*. 2022 (ver p. 54).

[32] A. Liaw e M. Wiener. «Classification and Regression by randomForest». Em: *R News* 2.3 (2002), pp. 18–22. URL: https://CRAN.R-project.org/doc/Rnews/ (ver pp. 57, 58, 63).

[33] S. Milborrow. *rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'*. R package version 3.1.1. 2022. URL: https://CRAN.R-project.org/package=rpart.plot (ver p. 58).

[34] B. Efron e R. J. Tibshirani. *An Introduction to the Bootstrap*. 1993. ISBN: 9780412042317. DOI: 10.1007/978-1-4899-4541-9 (ver pp. 61, 206).

[35] T. Fischetti et al. *R: Data Analysis and Visualization*. Packt Publishing, 2016. ISBN: 9781786460486. URL: https://books.google.pt/books?id=KnTZDQAAQBAJ (ver pp. 62, 64).

[36] H. Kukreja et al. «An introduction to artificial neural network». Em: *Int J Adv Res Innov Ideas Educ* 1 (2016), pp. 27–30 (ver p. 63).

[37] B. Lantz. *Machine Learning with R*. Packt Publishing, 2013, p. 396. ISBN: 9781782162148. URL: https://edu.kpfu.ru/pluginfile.php/278552/mod_resource/content/1/MachineLearningR__Brett_Lantz.pdf (ver pp. 63, 64, 68, 78, 79, 82–93).

[38] P. Gaur. «Neural networks in data mining». Em: *International Journal of Electronics and Computer Science Engineering (IJECSE, ISSN: 2277-1956)* 1.03 (2012), pp. 1449–1453 (ver p. 63).

[39] S. Sharma, S. Sharma e A. Athaiya. «Activation Functions in Neural Networks». Em: *International Journal of Engineering Applied Sciences and Technology* 04.12 (2020), pp. 310–316. DOI: 10.33564/ijeast.2020.v04i12.054 (ver pp. 65, 66, 68).

[40] F. Haddadi et al. «Intrusion detection and attack classification using feed-forward neural network». Em: *2nd International Conference on Computer and Network Technology, ICCNT 2010* (2010), pp. 262–266. DOI: 10.1109/ICCNT.2010.28 (ver p. 76).

[41] A. Shrestha e A. Mahmood. «Review of deep learning algorithms and architectures». Em: *IEEE Access* 7 (2019), pp. 53040–53065. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2912200 (ver p. 76).

[42] J. Allaire e F. Chollet. *keras: R Interface to 'Keras'*. R package version 2.9.0. 2022. URL: https://CRAN.R-project.org/package=keras (ver p. 82).

[43] M. Grandini, E. Bagli e G. Visani. «Metrics for multi-class classification: an overview». Em: *arXiv preprint arXiv:2008.05756* (2020) (ver p. 87).

[44] S. M. Abd Elrahman e A. Abraham. «A Review of Class Imbalance Problem». Em: *Journal of Network and Innovative Computing* 1 (2013), pp. 332–340. URL: www.mirlabs.net/jnic/index.html (ver pp. 88, 94–96).

[45] A. Tharwat. «Classification assessment methods». Em: *Applied Computing and Informatics* 17.1 (2018), pp. 168–192. ISSN: 22108327. DOI: 10.1016/j.aci.2018.08.003 (ver pp. 89, 91–94).

[46] M. Kuhn. *caret: Classification and Regression Training*. R package version 6.0-92. 2022. URL: https://CRAN.R-project.org/package=caret (ver p. 90).

[47] X. Robin et al. «pROC: an open-source package for R and S+ to analyze and compare ROC curves». Em: *BMC Bioinformatics* 12 (2011), p. 77 (ver p. 93).

[48] A. Ghatak. *Machine Learning with R*. 2017, pp. 1–210. ISBN: 9789811068089. DOI: 10.1007/978-981-10-6808-9 (ver p. 94).

[49] V. Ganganwar. «An overview of classification algorithms for imbalanced datasets». Em: *International Journal of Emerging Technology and Advanced Engineering* 2.4 (2012), pp. 42–47. ISSN: 2250-2459. URL: http://www.ijetae.com/files/Volume2Issue4/IJETAE_0412_07.pdf (ver p. 95).

[50] P. Vuttipittayamongkol, E. Elyan e A. Petrovski. «Knowledge-Based Systems». Em: *Knowledge-Based Systems* 212 (2021), p. 106631. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2020.106631. URL: https://doi.org/10.1016/j.knosys.2020.106631 (ver p. 95).

[51] K. M. Hasib et al. «A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem». Em: *Journal of Computer Science* 16.11 (2020), pp. 1546–1557. ISSN: 15526607. DOI: 10.3844/JCSSP.2020.1546.1557 (ver p. 96).

[52] P. Sadowski. «Notes on Backpropagation». Em: *Department of Computer Science University of California Irvine* 1.2 (2017), pp. 1–4 (ver p. 207).

[53] M. Bod. «A guide to recurrent neural networks and backpropagation». Em: *Rnn Dan Bpnn* 2.2 (2001), pp. 1–10 (ver p. 207).

A

Biblioteca *grpeg* Estimação da Função Objetivo 3.61

Como referido na secção 3.3 esta biblioteca, em vez de se minimizar as quantidades representadas em 3.58 e 3.60, esta biblioteca otimiza a função objetivo representada em A.1.

$$Q(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \frac{1}{n}L(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) + P_\lambda(\boldsymbol{\beta}) \tag{A.1}$$

O $P_\lambda(\boldsymbol{\beta})$, na equação A.1 é dado pela equação A.2

$$P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{l=1}^{L} \sqrt{p_l}||\mathbf{x}_l||_2 \tag{A.2}$$

onde $\boldsymbol{\beta}_l$ denota o vetor dos coeficientes de regressão correspondente ao *l-ésimo* grupo, e $\mathbf{x}_l$ é a matriz das variáveis que representam o grupo l.

Na formula de A.1, tem-se que a função de perda L é a *deviance*, ou seja, a quantidade L é -2 vezes a log-verosimilhança, da distribuição especificada da variável de resposta, no caso da variável resposta seguir uma distribuição Bernoulli.

No caso da variável de resposta ter uma distribuição binomial, tem-se que a função de custo L é dada por A.3.

$$L(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = -2\left(\sum_{i=1}^{n} y_i \ln\left(\hat{\pi}_i\right) + (1-y_i)\ln\left(1-\hat{\pi}_i\right)\right) \tag{A.3}$$

onde $\hat{\pi}_i = \dfrac{e^{\eta_i}}{1+e^{\eta_i}}$ onde η_i é a combinação linear dos preditores, isto é, $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ip}$. Como visto anteriormente a $P(\mathrm{Y}_i = 1|\eta_i) = \hat{\pi}_i$

Já no caso da variável de resposta ter uma distribuição normal, tem-se que a função de custo L é dada por A.4.

$$L(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \frac{1}{2}||\mathbf{y}-\mathbf{x}\boldsymbol{\beta}||_2^2 = \frac{1}{2}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)\right) \tag{A.4}$$

No entanto, aqui apenas se encontra descrito para o caso em que a variável de resposta tem uma distribuição binomial, pois para mais informação, consultar [21].

Na Tabela A.1, encontra-se descrito o algoritmo para a estimação dos parâmetros $\boldsymbol{\beta}$.

Tabela A.1: Algoritmo Grupo Descendente (*group descent algorithms*), para Regressão Logística com a Penalidade Grupo Lasso [20].

Algoritmo Grupo Descendente para Regressão Logística com a Penalidade Grupo Lasso

1. Repetir:
   a) $\eta \leftarrow \mathbf{x}\boldsymbol{\beta}$;
   b) $\pi \leftarrow \{e^{\eta_i}/(1+e^{\eta_i})\}_{i=1}^{n}$.
   c) $\tilde{r} \leftarrow (\mathbf{y} - \pi)/v$;
   d) Para $j = 1, \ldots, J$:
      i. $z_j = \mathbf{x}_j^T \tilde{r} + \beta_j$
      ii. $\beta_j' \leftarrow S(vz_j, \lambda_j)/v$
      iii. $\tilde{r}' \leftarrow (\tilde{r} - \mathrm{x}_j^T)(\beta_j' - \beta_j)$
2. Até se atingir a convergência.

Na Tabela A.1, $v \leftarrow \max_i\{\nabla^2 L_i(\eta_i^*)\}$. O $S(v\mathbf{z}_j, \lambda_j)$ é dado por A.5, onde para mais informações consultar .[20]

$$L(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) = \frac{1}{2}||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2 = \frac{1}{2}\left(\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right) \right. \tag{A.5}$$

É claro que o algoritmo apresentado na Tabela A.1, necessita de valores iniciais de λ e de $\boldsymbol{\beta}$ para começar a iterar. Neste caso, uma vez que se utiliza validação cruzada, para se encontrar um valor de λ ótimo, então temos que $\lambda \in [\lambda_{min}, \lambda_{max}]$. Nesta biblioteca o $\lambda_{min} = 0.0001$ se o número de variáveis explicativas for menor que o número de observações e $\lambda_{min} = 0.05$, no caso contrário por padrão. No caso da Regressão Logística o valor de $\lambda_{max} = \max_j(v||z_j||)$, onde os z_j apenas são calculados em relação a β_0, ou seja, somente ajustando o modelo com o β_0. Assim, o algoritmo começa com o λ_{max} onde $\hat{\boldsymbol{\beta}} = 0$. Com este valor de $\lambda = \lambda_{max}$ o algoritmo começa e tentar encontrar os melhores valores de $\boldsymbol{\beta}$ de modo a que minimizem A.1, onde o número máximo de iterações é 1000 por padrão, ou até que cada coeficiente estimado seja menor que 0.0001, por padrão. Posto isto já se tem um valor de $\boldsymbol{\beta}$. Por padrão a biblioteca, utiliza 100 valores diferentes de λ entre o λ_{max} e λ_{min}. Para o próximo valor de λ que o algoritmo começa, este começa com os valores de β iguais aos valores de $\boldsymbol{\beta}$ estimados com o valor de λ anterior, e isto repete-se sucessivamente, ou seja, os $\boldsymbol{\beta}$ estimados com o valor de $\lambda = \lambda_{min}$ serão os valores inicias de $\boldsymbol{\beta}$ para o segundo valor de λ [20].A sequência de valores de λ são determinados

automaticamente numa escala de valores lambda que varia uniformemente na escala logarítmica sobre o intervalo dos valores λ [21].

B

Amostra *Bootstrap*

Para a definição de uma amostra *bootstrap*, recorremos a [34].

Seja $\hat{F}$ uma distribuição empírica, onde cada um dos seus valores observados x_i, $i = 1, \ldots, n$ têm probabilidade $\frac{1}{n}$.

Uma amostra *bootstrap* é definida por uma amostra aleatória de tamanho n extraída de $\hat{F}$, chamada de $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$,

$$\hat{F} \rightarrow (x_1^*, x_2^*, \ldots, x_n^*) \tag{B.1}$$

A notação de x^* não é o conjunto de dados reais de x, mas sim uma versão aleatória ou uma reamostra de x. Ou seja, a amostra *bootstrap* $x_1^*, x_2^*, \ldots, x_n^*$ é uma amostra aleatória de tamanho n retirada da população com n objetos $(x_1, x_2, \ldots, x_n)$.

Assim, podemos ter $x_1^* = x_7, x_2^* = x_3, x_3^* = x_3, x_4^* = x_2, \ldots, x_n^* = x_7$. O conjunto de dados *bootstrap* consiste nos membros do conjunto de dados original $(x_1, x_2, \ldots, x_n)$, sendo que alguns destes valores aparecem zero vezes, outros aparecem uma vez, outros duas vezes, e por ai por diante.

C

Redes Neuronais

Suponhamos, que $\mathbf{x} = (x_1, \ldots, x_p)^T$, é o vetor que contém as p variáveis explicativas, correspondendo x_j, $j = 1, \ldots, p$ à *j-ésima* variável explicativa. Cada uma destas p variáveis explicativas foi observada n vezes, $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$, sendo x_{ij}, $i = 1, \ldots, n$, a *i-ésima* observação da variável explicativa x_j.

Do mesmo modo o vetor $\mathbf{y} = (y_1, \ldots, y_n)^T$ é o vetor que contém as observações da variável resposta Y, onde a consideremos contínua no caso da regressão e binária no caso discreto. Suponha-se ainda que se está presente sobre uma Rede Neuronal *feed-forward*, com uma única camada oculta com K nós, $k = 1, \ldots, K$. Já vimos que a estrutura representativa de uma rede neuronal no caso da regressão era a mesma que no caso da classificação binária, o que apenas, muda é a função de ativação na camada de saída, em que no caso de regressão é a função linear e no caso da classificação binária é a função *sigmoid*.

As deduções que aqui se encontram foram baseadas em [12], para o caso da regressão e em [52] e [53] para o caso da classificação binária.

Vimos na subsecção 3.7.4, que os parâmetros a serem estimados eram os $\boldsymbol{w}_k = (w_{k0}, \ldots, w_{kp})^T$ e $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$, $k = 1, \ldots, K$.

C.0.1 Caso da Regressão

No caso da regressão vimos na subsecção 3.7.4 que a função $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ é dada por:

$$f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\right) \tag{C.1}$$

e que $R_i(\theta)$ era dado por C.2.

$$\frac{1}{2}\left(y_i - \beta_0 - \sum_{k=1}^{K} \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kh} x_{ij}\right)\right)^2 \tag{C.2}$$

Para simplificar, considere-se que z_{ik} é dado pela equação C.3.

$$z_{ik} = w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij} \tag{C.3}$$

Neste caso, tem-se que aplicando as regras de derivação da cadeia (*chain rule*), a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$ é dada por :

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k} = \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k} \tag{C.4}$$

Ora, a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$ é dada pela expressão de C.5, pois uma vez que neste caso o $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ funciona como uma variável fixa, ou seja, pode-se considerar $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ como sendo uma variável x tal como na análise matemática.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} = -\frac{1}{2} 2\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)' \Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) = -1\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) \tag{C.5}$$

Já a derivada parcial, $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k}$ é dada pela expressão C.7, pois uma vez que, fixando um valor de k, $k = 1, \cdots, K$ tem-se que $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ é dado pela expressão C.6.

$$f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \beta_0 + \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\right) \tag{C.6}$$

Logo, $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k}$ é dada pela expressão C.7, onde o z_{ik} é dado pela expressão C.3

$$\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k} = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\right) = g(z_{ik}) \tag{C.7}$$

Assim, uma vez que a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$ pode ser rescrita pela expressão C.4, então tem-se que pelas expressões C.5 e C.7 a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$, para $k = 1, \ldots, K$, é dada pela expressão C.8.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k} = \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k} = -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) g(z_{ik}) \tag{C.8}$$

Para $k = 0$, é fácil de verificar que a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_0}$ é dada por C.9, uma vez que, $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k} = 1$.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_0} = \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_0} = -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big) \tag{C.9}$$

Da mesma forma, pelas regra de derivação da cadeia (*chain rule*), tem-se que a derivada parcial $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}}$, pode ser dada pela expressão C.10.

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} = \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} \frac{\partial g(z_{ik})}{\partial z_{ik}} \frac{\partial z_{ik}}{\partial w_{kj}} \tag{C.10}$$

Ora a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$ é dada pela expressão C.5.

Já a $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})}$ é dada pela expressão C.12, uma vez que a expressão de $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ em C.7, pode ser escrita à custa do z_{ik} como se encontra na expressão C.11, pois tem-se que z_{ik} é dada pela expressão C.3.

$$f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \beta_0 + \beta_k g\left(w_{k0} + \sum_{j=1}^{p} w_{kj} x_{ij}\right) = \beta_0 + \beta_k g(z_{ik}) \tag{C.11}$$

$$\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} = \beta_k \tag{C.12}$$

A $\frac{\partial g(z_{ik})}{\partial z_{ik}}$ é dada pela expressão C.13, pois temos que a função g é uma função de ativação, em que esta é o utilizador a escolher. Assim, sendo, tem-se a sua derivada parcial na sua generalização.

$$\frac{\partial g(z_{ik})}{\partial z_{ik}} = g'(z_{ik}) \tag{C.13}$$

A $\frac{\partial z_{ik}}{\partial w_{kj}}$ é dada pela expressão C.14, uma vez que, o z_{ik} definido na expressão C.3 depende dos w_{kj}

$$\frac{\partial z_{ik}}{\partial w_{kj}} = x_{ij} \tag{C.14}$$

Assim, a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}}$ é dada por C.15.

$$\begin{aligned} \frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} &= \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} \frac{\partial g(z_{ik})}{\partial z_{ik}} \frac{\partial z_{ik}}{\partial w_{kj}} \\ &= -\Big(y_i - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)\beta_k g'(z_{ik}) x_{ij} \end{aligned} \tag{C.15}$$

Já na $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{0j}}$ é fácil de mostrar que esta é dada por C.16, uma vez que a única coisa que muda é a $\frac{\partial z_{ik}}{\partial w_{0j}}$ em que a $\frac{\partial z_{ik}}{\partial w_{kj}} = 1$.

$$\begin{aligned} \frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} &= \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} \frac{\partial g(z_{ik})}{\partial z_{ik}} \frac{\partial z_{ik}}{\partial w_{kj}} \\ &= -\Big(y_i - f_{\theta}(x_i)\Big)\beta_k g'(z_{ik}) \end{aligned} \tag{C.16}$$

C.0.2 Caso de Classificação Binária

Na subsecção 3.7.4, vimos que $R_i(\boldsymbol{\theta})$ é dado por C.17

$$R_i(\boldsymbol{\theta}) = -\left(y_i \ln\left(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right) + (1-y_i)\ln\left(1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)\right) \tag{C.17}$$

Neste presente caso, as derivadas parciais $\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$ e de $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}}$ calculam-se da mesma forma que no caso da regressão, no entanto, na classificação binária, tem-se que a função $R_i(\boldsymbol{\theta})$ é dado por C.17 e também se tem uma outra alteração na função final f, pois no caso da classificação binária, o resultado que queremos que retorne são probabilidades.

As alterações no caso da classificação binária, fase a um problema de regressão é que a função $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ é a função de *sigmoid*, enquanto que no caso da regressão é a função identidade, ou seja, no caso da classificação binária, tem-se que $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ é dado pela expressão C.18, onde z_{ik} é dado pela expressão C.19, como se viu na subsecção 3.7.4.

$$f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \frac{1}{1+e^{-\left(\beta_0+\sum_{k=1}^{K}\beta_k g\left(w_{k0}\sum_{j=1}^{p} w_{kj}x_{ij}\right)\right)}} = \frac{1}{1+e^{-s_i}} \tag{C.18}$$

$$z_{ik} = \beta_0 + \sum_{k=1}^{K}\beta_k g\left(w_{k0}\sum_{j=1}^{p} w_{kj}x_{ij}\right) \tag{C.19}$$

Assim, para se calcular a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$ é necessário calcular-se as $\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$ e $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i}{\partial \beta_k}$, como se encontra na expressão C.8. No caso da $\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$, atendendo à expressão C.17, tem-se que o y_i funciona como uma constante, logo só é necessário derivar o $\ln(f_{\boldsymbol{\theta}}(\mathbf{x}_i))$, em que $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ funciona como uma variável. Assim, deste modo tem-se que $\frac{\partial \ln(f_{\boldsymbol{\theta}}(\mathbf{x}_i))}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} = \frac{1}{f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$, pois uma vez que, neste caso como o $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ funciona como uma variável, então derivar o $\ln(f_{\boldsymbol{\theta}}(\mathbf{x}_i))$ em ordem a $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ é igual a derivar o $\ln(x)$ em ordem a x. Logo, tem-se que a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$ é dada pela expressão C.20

$$\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} = -\left(\frac{y_i}{f_{\boldsymbol{\theta}}(\mathbf{x}_i)} - \frac{1-y_i}{1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)}\right) = \left(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\right) \tag{C.20}$$

Para se calcular a $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k}$, uma vez que temos que $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$ está em função de z_{ik} e este por sua vez é em função de β_k, então utilizando a regra de derivação da função composta, tem-se que a é dada pela expressão C.21.

$$\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k} = \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial z_{ik}}\frac{\partial z_{ik}}{\partial \beta_k} \tag{C.21}$$

Ora a $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial z_{ik}}$ é dada pela expressão C.22, em que neste caso o z_{ik} funciona como uma simples variável x.

$$
\begin{aligned}
\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial z_{ik}} &= \left(\left(1+e^{-z_{ik}}\right)^{-1}\right)' \\
&= (-1)\left(1+e^{-z_{ik}}\right)'\left(1+e^{-z_{ik}}\right)^{-2} \\
&= (-1)(-1)e^{-z_{ik}}\left(1+e^{-z_{ik}}\right)^{-2} \\
&= e^{-z_{ik}}\left(1+e^{-z_{ik}}\right)^{-2} \\
&= \left(1+e^{-z_{ik}}\right)^{-1}\left(e^{-z_{ik}}\left(1+e^{-z_{ik}}\right)^{-1}\right) \\
&= \left(1+e^{-z_{ik}}\right)^{-1}\left(1-\left(1+e^{-z_{ik}}\right)^{-1}\right) \\
&= f_{\boldsymbol{\theta}}(\mathbf{x}_i)(1-f_{\boldsymbol{\theta}}(\mathbf{x}_i))
\end{aligned}
\tag{C.22}
$$

A $\dfrac{\partial z_{ik}}{\partial \beta_k}$, é dada pela expressão C.23, pois basta fixar um valor para k, onde $k = 1, \cdots, K$.

$$
\frac{\partial z_{ik}}{\partial \beta_k} = g\left(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij}\right) = g\Big(z(x_{ik})\Big) \tag{C.23}
$$

Posto isto, tem-se que $\dfrac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k}$ é dada pela expressão C.24, pois visto que já foram calculadas as restantes derivadas parciais que eram necessárias para o cálculo desta, em que as derivadas parciais que eram necessárias para o cálculo desta encontram-se na expressão C.21

$$
\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k} = \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial z_{ik}}\frac{\partial z_{ik}}{\partial \beta_k} = f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big(1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)g\left(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij}\right) \tag{C.24}
$$

Uma vez que $\dfrac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$ é dada pela expressão C.8 e as $\dfrac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$ e $\dfrac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial \beta_k}$ já se encontram calculadas nas expressões C.20 e C.24, respetivamente, então tem-se que através de algum cálculo algébrico a $\dfrac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k}$ pode ser dada pela expressão C.25.

$$
\begin{aligned}
\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k} &= \left(\frac{-y_i}{f_{\boldsymbol{\theta}}(\mathbf{x}_i)} + \frac{(1-y_i)}{1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)}\right)\Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big(1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)\Big)g\left(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij}\right) \\
&= \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\Big)g\left(w_{k0} + \sum_{j=1}^{p} w_{kj}x_{ij}\right)
\end{aligned}
\tag{C.25}
$$

No entanto a $\dfrac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_0}$ é dada pela equação C.26, uma vez que a $\dfrac{\partial z_{ik}}{\partial \beta_0} = 1$.

$$
\begin{aligned}
\frac{\partial R_i(\boldsymbol{\theta})}{\partial \beta_k} &= \left(\frac{-y_i}{f_{\boldsymbol{\theta}}(\mathbf{x}_i)} + \frac{(1-y_i)}{1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)}\right)\Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big(1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)\Big)\Big) \\
&= \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\Big)
\end{aligned}
\tag{C.26}
$$

Já a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}}$ também é calculada da mesma forma que no caso da regressão, isto é, tal como se encontra na expressão C.10. A $\frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}$ já se encontra calculada, e pode-se encontra esta na expressão C.20. No caso da $\frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})}$, esta está explicita na expressão C.28, onde se encontra aí a sua dedução, em que aí se fixa k, $k = 1, \cdots, K$. De notar que a função $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$, pode ser escrita em função do $g(z_{ik})$, em que basta tomar o z_{ik}, tal como este representado na expressão C.3. Assim, tem-se que a expressão de $f_{\boldsymbol{\theta}}(\mathbf{x}_i)$, que se encontra na expressão C.18, pode ser escrita como a que se encontra na expressão C.27.

$$\begin{aligned} f_{\boldsymbol{\theta}}(\mathbf{x}_i) &= \frac{1}{1 + e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g\left(w_{k0} \sum_{j=1}^{p} w_{kj} x_{ij}\right)\right)}} \\ &= \frac{1}{1 + e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g(z_{ik})\right)}} \end{aligned} \tag{C.27}$$

$$\begin{aligned} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} &= \left(\frac{1}{1 + e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g(z_{ik})\right)}}\right)' \\ &= \left(\left(1 + e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g(z_{ik})\right)}\right)^{-1}\right)' \\ &= (-1)\left(-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g(z_{ik})\right)\right)' \left(1 + e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g(z_{ik})\right)}\right)^{-2} \left(e^{-\left(\beta_0 + \sum_{k=1}^{K} \beta_k g(z_{ik})\right)}\right) \\ &= \beta_k \left(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\left(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)\right) \end{aligned} \tag{C.28}$$

As $\frac{\partial g(z_{ik})}{\partial z_{ik}}$ e $\frac{\partial z_{ik}}{\partial w_{kj}}$, têm a mesma forma que no caso da regressão, isto é, têm a mesma expressão que C.13 e C.14, respetivamente.

Dado que que já se tem todas as parcelas necessárias para calcular a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}}$, então através de algum cálculo algébrico, tem-se que $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}}$ pode ser dada pela expressão C.29.

$$\begin{aligned} \frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} &= \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} \frac{\partial g(z_{ik})}{\partial z_{ik}} \frac{\partial z_{ik}}{\partial w_{kj}} \\ &= \left(\frac{-y_i}{f_{\boldsymbol{\theta}}(\mathbf{x}_i)} + \frac{(1 - y_i)}{1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)}\right)\left(f_{\boldsymbol{\theta}}(\mathbf{x}_i)\left(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i)\right)\right)\beta_k g'(z_{ik}) x_{ij} \\ &= \left(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i\right) g'(z_{ik}) x_{ij} \end{aligned} \tag{C.29}$$

No entanto a $\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{0j}}$ é dada pela equação C.30, uma vez que a $\frac{\partial z_{ik}}{\partial w_{0j}} = 1$.

$$
\begin{aligned}
\frac{\partial R_i(\boldsymbol{\theta})}{\partial w_{kj}} &= \frac{\partial R_i(\boldsymbol{\theta})}{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \frac{\partial f_{\boldsymbol{\theta}}(\mathbf{x}_i)}{\partial g(z_{ik})} \frac{\partial g(z_{ik})}{\partial z_{ik}} \frac{\partial z_{ik}}{\partial w_{kj}} \\
&= \left(\frac{-y_i}{f_{\boldsymbol{\theta}}(\mathbf{x}_i)} + \frac{(1-y_i)}{1-f_{\boldsymbol{\theta}}(\mathbf{x}_i)} \right) \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) \Big(1 - f_{\boldsymbol{\theta}}(\mathbf{x}_i) \Big) \Big) \beta_k g'(z_{ik}) \\
&= \Big(f_{\boldsymbol{\theta}}(\mathbf{x}_i) - y_i \Big) g'(z_{ik})
\end{aligned}
\tag{C.30}
$$

I

Anexo 1

I.1 Dados Comorbilidades

I.1.1 Regressão Logística - Dados Desequilibrados

Tabela I.1: Resultado da Regressão Logística com o método *stepwise* na base de dados restrita das comorbilidades.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.5023	0.2935	-39.20	0.0000
idade_utente_a_data_validacao	0.1110	0.0009	121.40	0.0000
sexo_utenteM	0.7515	0.0226	33.18	0.0000
doenca_renal_cronica1	0.7519	0.0483	15.57	0.0000
neoplasia1	0.5971	0.0470	12.70	0.0000
descricao_distrito_ocorrenciaAveiro	-0.1565	0.2856	-0.55	0.5837
descricao_distrito_ocorrenciaBeja	0.1047	0.3077	0.34	0.7336
descricao_distrito_ocorrenciaBraga	-0.3523	0.2849	-1.24	0.2162
descricao_distrito_ocorrenciaBragança	-0.2999	0.2905	-1.03	0.3021
descricao_distrito_ocorrenciaCastelo Branco	-0.5194	0.2964	-1.75	0.0797
descricao_distrito_ocorrenciaCoimbra	0.1424	0.2865	0.50	0.6190
descricao_distrito_ocorrenciaÉvora	0.1357	0.2969	0.46	0.6475
descricao_distrito_ocorrenciaFaro	-0.3032	0.2903	-1.04	0.2962
descricao_distrito_ocorrenciaGuarda	0.0881	0.2955	0.30	0.7655
descricao_distrito_ocorrenciaLeiria	0.1682	0.2882	0.58	0.5595
descricao_distrito_ocorrenciaLisboa	-0.0383	0.2828	-0.14	0.8922
descricao_distrito_ocorrenciaMadeira	-1.2614	0.4011	-3.14	0.0017
descricao_distrito_ocorrenciaPortalegre	-0.2400	0.3200	-0.75	0.4532
descricao_distrito_ocorrenciaPorto	-0.1150	0.2833	-0.41	0.6848
descricao_distrito_ocorrenciaSantarém	0.0112	0.2886	0.04	0.9690
descricao_distrito_ocorrenciaSetúbal	0.0787	0.2842	0.28	0.7819
descricao_distrito_ocorrenciaViana do Castelo	-0.3583	0.2904	-1.23	0.2172
descricao_distrito_ocorrenciaVila Real	-0.4225	0.2956	-1.43	0.1529
descricao_distrito_ocorrenciaViseu	-0.1868	0.2875	-0.65	0.5160
diabetes	0.3209	0.0346	9.27	0.0000
deficiencia_neurologica_cronica	0.5485	0.0815	6.73	0.0000
patologia_hepatica	0.9134	0.0971	9.40	0.0000
doenca_pulmonar_cronica	0.4003	0.0484	8.26	0.0000
insuficiencia_renal_aguda	1.1629	0.1778	6.54	0.0000
doenca_neurologica_ou_neuromuscular_cronica	0.3741	0.0756	4.95	0.0000
doencas_hematologicas_cronicas	0.3522	0.0812	4.34	0.0000
insuficiencia_cardiaca	0.4382	0.2038	2.15	0.0316
data_confirmado1	-0.0002	0.0001	-1.77	0.0765
vih_outras_imunodeficiencias	0.3188	0.1841	1.73	0.0833
asma	-0.1604	0.1000	-1.60	0.1089
AIC		62653.4		

Tabela I.2: Resultado da Regressão Logística ajustado na base de dados restrita das comorbilidades, sem as variáveis explicativas *vih_outras_imunodeficiencias*, *asma* e *coagulapatia_de_consumo* .

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.5005	0.2935	-39.18	0.0000
idade_utente_a_data_validacao	0.1110	0.0009	121.55	0.0000
sexo_utenteM	0.7532	0.0226	33.28	0.0000
doenca_renal_cronica	0.7545	0.0483	15.62	0.0000
neoplasia	0.5990	0.0470	12.75	0.0000
descricao_distrito_ocorrenciaAveiro	-0.1585	0.2857	-0.55	0.5790
descricao_distrito_ocorrenciaBeja	0.1035	0.3078	0.34	0.7367
descricao_distrito_ocorrenciaBraga	-0.3539	0.2850	-1.24	0.2143
descricao_distrito_ocorrenciaBragança	-0.3007	0.2907	-1.03	0.3008
descricao_distrito_ocorrenciaCastelo Branco	-0.5202	0.2965	-1.75	0.0793
descricao_distrito_ocorrenciaCoimbra	0.1404	0.2866	0.49	0.6242
descricao_distrito_ocorrenciaÉvora	0.1346	0.2970	0.45	0.6504
descricao_distrito_ocorrenciaFaro	-0.3060	0.2904	-1.05	0.2920
descricao_distrito_ocorrenciaGuarda	0.0854	0.2956	0.29	0.7728
descricao_distrito_ocorrenciaLeiria	0.1657	0.2883	0.57	0.5656
descricao_distrito_ocorrenciaLisboa	-0.0401	0.2829	-0.14	0.8873
descricao_distrito_ocorrenciaMadeira	-1.2674	0.4012	-3.16	0.0016
descricao_distrito_ocorrenciaPortalegre	-0.2413	0.3201	-0.75	0.4510
descricao_distrito_ocorrenciaPorto	-0.1167	0.2834	-0.41	0.6804
descricao_distrito_ocorrenciaSantarém	0.0095	0.2887	0.03	0.9738
descricao_distrito_ocorrenciaSetúbal	0.0768	0.2844	0.27	0.7870
descricao_distrito_ocorrenciaViana do Castelo	-0.3595	0.2905	-1.24	0.2158
descricao_distrito_ocorrenciaVila Real	-0.4244	0.2957	-1.44	0.1512
descricao_distrito_ocorrenciaViseu	-0.1896	0.2876	-0.66	0.5097
diabetes	0.3187	0.0346	9.21	0.0000
deficiencia_neurologica_cronica	0.5488	0.0815	6.74	0.0000
patologia_hepatica	0.9206	0.0970	9.49	0.0000
doenca_pulmonar_cronica	0.3955	0.0483	8.19	0.0000
insuficiencia_renal_aguda	1.1774	0.1777	6.63	0.0000
doenca_neurologica_ou_neuromuscular_cronica	0.3756	0.0756	4.97	0.0000
doencas_hematologicas_cronicas	0.3547	0.0811	4.37	0.0000
insuficiencia_cardiaca	0.4343	0.2038	2.13	0.0331
data_confirmado1	-0.0002	0.0001	-1.79	0.0740
AIC		62654.9		

Tabela I.3: Resultado da Regressão Logística ajustado na base de dados restrita das comorbilidades - Modelo Final.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.5541	0.2916	-39.63	0.0000
idade_utente_a_data_validacao	0.1111	0.0009	121.90	0.0000
sexo_utenteM	0.7520	0.0226	33.25	0.0000
doenca_renal_cronica	0.7560	0.0483	15.65	0.0000
neoplasia	0.6010	0.0470	12.80	0.0000
descricao_distrito_ocorrenciaAveiro	-0.1671	0.2852	-0.59	0.5579
descricao_distrito_ocorrenciaBeja	0.0842	0.3071	0.27	0.7840
descricao_distrito_ocorrenciaBraga	-0.3662	0.2845	-1.29	0.1979
descricao_distrito_ocorrenciaBragança	-0.3190	0.2900	-1.10	0.2714
descricao_distrito_ocorrenciaCastelo Branco	-0.5414	0.2958	-1.83	0.0672
descricao_distrito_ocorrenciaCoimbra	0.1239	0.2860	0.43	0.6648
descricao_distrito_ocorrenciaÉvora	0.1176	0.2964	0.40	0.6916
descricao_distrito_ocorrenciaFaro	-0.3315	0.2896	-1.14	0.2525
descricao_distrito_ocorrenciaGuarda	0.0640	0.2950	0.22	0.8282
descricao_distrito_ocorrenciaLeiria	0.1464	0.2877	0.51	0.6107
descricao_distrito_ocorrenciaLisboa	-0.0573	0.2822	-0.20	0.8390
descricao_distrito_ocorrenciaMadeira	-1.3036	0.4004	-3.26	0.0011
descricao_distrito_ocorrenciaPortalegre	-0.2637	0.3194	-0.83	0.4090
descricao_distrito_ocorrenciaPorto	-0.1259	0.2829	-0.45	0.6562
descricao_distrito_ocorrenciaSantarém	-0.0081	0.2881	-0.03	0.9775
descricao_distrito_ocorrenciaSetúbal	0.0556	0.2837	0.20	0.8447
descricao_distrito_ocorrenciaViana do Castelo	-0.3766	0.2899	-1.30	0.1938
descricao_distrito_ocorrenciaVila Real	-0.4430	0.2951	-1.50	0.1333
descricao_distrito_ocorrenciaViseu	-0.2087	0.2870	-0.73	0.4672
diabetes	0.3200	0.0346	9.25	0.0000
deficiencia_neurologica_cronica	0.5619	0.0812	6.92	0.0000
patologia_hepatica	0.9221	0.0970	9.50	0.0000
doenca_pulmonar_cronica	0.3959	0.0483	8.20	0.0000
insuficiencia_renal_aguda	1.1985	0.1774	6.75	0.0000
doenca_neurologica_ou_neuromuscular_cronica	0.3921	0.0751	5.22	0.0000
doencas_hematologicas_cronicas	0.3574	0.0811	4.41	0.0000
insuficiencia_cardiaca	0.4480	0.2039	2.20	0.0280
AIC		62656.07		

Tabela I.4: *OR* variáveis explicativas do modelo final da Regressão Logística referentes às comorbilidades.

		Intervalo Confiança	
Variável	***Odds-Ratio***	**2.5%**	**97.5%**
idade_utente_a_data_validacao	1.12	1.12	1.12
sexo_utenteM	2.12	2.03	2.22
doenca_renal_cronica1	2.13	1.94	2.34
neoplasia1	1.82	1.66	2.00
descricao_distrito_ocorrenciaAveiro	0.85	0.48	1.48
descricao_distrito_ocorrenciaBeja	1.09	0.60	1.99
descricao_distrito_ocorrenciaBraga	0.69	0.40	1.21
descricao_distrito_ocorrenciaBragança	0.73	0.41	1.28
descricao_distrito_ocorrenciaCastelo Branco	0.58	0.33	1.04
descricao_distrito_ocorrenciaCoimbra	1.13	0.65	1.98
descricao_distrito_ocorrenciaÉvora	1.12	0.63	2.01
descricao_distrito_ocorrenciaFaro	0.72	0.41	1.27
descricao_distrito_ocorrenciaGuarda	1.07	0.60	1.90
descricao_distrito_ocorrenciaLeiria	1.16	0.66	2.03
descricao_distrito_ocorrenciaLisboa	0.94	0.54	1.64
descricao_distrito_ocorrenciaMadeira	0.27	0.12	0.60
descricao_distrito_ocorrenciaPortalegre	0.77	0.41	1.44
descricao_distrito_ocorrenciaPorto	0.88	0.51	1.53
descricao_distrito_ocorrenciaSantarém	0.99	0.56	1.74
descricao_distrito_ocorrenciaSetúbal	1.06	0.61	1.84
descricao_distrito_ocorrenciaViana do Castelo	0.69	0.39	1.21
descricao_distrito_ocorrenciaVila Real	0.64	0.36	1.14
descricao_distrito_ocorrenciaViseu	0.81	0.46	1.42
diabetes1	1.38	1.29	1.47
deficiencia_neurologica_cronica1	1.75	1.50	2.06
patologia_hepatica1	2.51	2.08	3.04
doenca_pulmonar_cronica1	1.49	1.35	1.63
insuficiencia_renal_aguda1	3.31	2.34	4.69
doenca_neurologica_ou_neuromuscular_cronica1	1.48	1.28	1.71
doencas_hematologicas_cronicas1	1.43	1.22	1.68
insuficiencia_cardiaca1	1.57	1.05	2.33

I.1.2 Modelo Regressão Logística com interações

Tabela I.5: Resultado da Regressão Logística com interações final com o *stepwise*, ajustado na base de dados restrita das comorbilidades

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.7059	1.5825	-6.13	0.0000
idade_utente_a_data_validacao	0.0875	0.0199	4.41	0.0000
sexo_utenteM	0.4887	0.1497	3.26	0.0011
doenca_renal_cronica1	4.2981	0.3501	12.28	0.0000
neoplasia1	2.8698	0.3309	8.67	0.0000
descricao_distrito_ocorrenciaAveiro	-1.9999	1.6075	-1.24	0.2135
descricao_distrito_ocorrenciaBeja	-0.8475	1.7215	-0.49	0.6225
descricao_distrito_ocorrenciaBraga	-3.1200	1.6037	-1.95	0.0517
descricao_distrito_ocorrenciaBragança	-0.9575	1.6780	-0.57	0.5683
descricao_distrito_ocorrenciaCastelo Branco	-2.9189	1.7222	-1.69	0.0901
descricao_distrito_ocorrenciaCoimbra	-1.5653	1.6164	-0.97	0.3328
descricao_distrito_ocorrenciaÉvora	-1.6402	1.7092	-0.96	0.3373
descricao_distrito_ocorrenciaFaro	-1.9290	1.6250	-1.19	0.2352
descricao_distrito_ocorrenciaGuarda	-1.7886	1.7208	-1.04	0.2986
descricao_distrito_ocorrenciaLeiria	-1.7257	1.6269	-1.06	0.2888
descricao_distrito_ocorrenciaLisboa	-2.1995	1.5841	-1.39	0.1650
descricao_distrito_ocorrenciaMadeira	-5.6066	2.6133	-2.15	0.0319
descricao_distrito_ocorrenciaPortalegre	-2.3237	1.9106	-1.22	0.2239
descricao_distrito_ocorrenciaPorto	-3.0429	1.5913	-1.91	0.0558
descricao_distrito_ocorrenciaSantarém	-1.0011	1.6234	-0.62	0.5375
descricao_distrito_ocorrenciaSetúbal	-1.9954	1.5939	-1.25	0.2106
descricao_distrito_ocorrenciaViana do Castelo	-2.4427	1.6656	-1.47	0.1425
descricao_distrito_ocorrenciaVila Real	-2.8748	1.7076	-1.68	0.0923
descricao_distrito_ocorrenciaViseu	-2.1937	1.6310	-1.34	0.1786
diabetes1	0.9622	0.2682	3.59	0.0003
deficiencia_neurologica_cronica1	5.0196	0.5012	10.01	0.0000
patologia_hepatica1	2.7067	0.5769	4.69	0.0000
doenca_pulmonar_cronica1	2.3225	0.3675	6.32	0.0000
insuficiencia_renal_aguda1	1.3476	0.2207	6.11	0.0000
doencas_hematologicas_cronicas1	2.4093	0.5464	4.41	0.0000
idade_utente_a_data_validacao:doenca_renal_cronica1	-0.0403	0.0042	-9.50	0.0000
idade_utente_a_data_validacao:deficiencia_neurologica_cronica1	-0.0532	0.0061	-8.78	0.0000
idade_utente_a_data_validacao:neoplasia1	-0.0273	0.0042	-6.46	0.0000
idade_utente_a_data_validacao:doenca_neurologica_ou_neuromuscular_cronica1	0.0047	0.0009	5.30	0.0000
neoplasia1:diabetes1	-0.4521	0.1088	-4.15	0.0000
idade_utente_a_data_validacao:doenca_pulmonar_cronica1	-0.0230	0.0046	-5.04	0.0000
doenca_renal_cronica1:doenca_pulmonar_cronica1	-0.4043	0.1195	-3.38	0.0007
patologia_hepatica1:doenca_neurologica_ou_neuromuscular_cronica1	-0.8600	0.4630	-1.86	0.0633
idade_utente_a_data_validacao:doencas_hematologicas_cronicas1	-0.0237	0.0068	-3.50	0.0005
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaAveiro	0.0233	0.0202	1.16	0.2478
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBeja	0.0118	0.0216	0.54	0.5861
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBraga	0.0348	0.0201	1.73	0.0834
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBragança	0.0095	0.0209	0.46	0.6486
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCastelo Branco	0.0299	0.0214	1.40	0.1626
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCoimbra	0.0217	0.0203	1.07	0.2837
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaÉvora	0.0225	0.0213	1.06	0.2913
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaFaro	0.0202	0.0204	0.99	0.3229
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaGuarda	0.0236	0.0214	1.10	0.2709
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLeiria	0.0239	0.0204	1.17	0.2424
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLisboa	0.0271	0.0199	1.36	0.1727
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaMadeira	0.0549	0.0325	1.69	0.0908
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPortalegre	0.0261	0.0236	1.11	0.2689
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPorto	0.0369	0.0200	1.85	0.0645
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSantarém	0.0128	0.0204	0.63	0.5289
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSetúbal	0.0260	0.0200	1.30	0.1940
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViana do Castelo	0.0263	0.0208	1.26	0.2072
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaVila Real	0.0307	0.0213	1.44	0.1504
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViseu	0.0254	0.0204	1.24	0.2143
doenca_renal_cronica1:neoplasia1	-0.4447	0.1364	-3.26	0.0011
doenca_renal_cronica1:patologia_hepatica1	-0.6809	0.2276	-2.99	0.0028
diabetes1:doencas_hematologicas_cronicas1	-0.4076	0.1736	-2.35	0.0189
doenca_renal_cronica1:insuficiencia_renal_aguda1	-0.9093	0.3405	-2.67	0.0076
idade_utente_a_data_validacao:patologia_hepatica1	-0.0212	0.0079	-2.70	0.0069
doenca_pulmonar_cronica1:insuficiencia_cardiaca1	1.1042	0.4078	2.71	0.0068
doenca_pulmonar_cronica1:vih_outras_imunodeficiencias1	-1.2540	0.5747	-2.18	0.0291
doenca_renal_cronica1:diabetes1	-0.2025	0.0930	-2.18	0.0294
diabetes1:patologia_hepatica1	-0.4965	0.2027	-2.45	0.0143
idade_utente_a_data_validacao:diabetes1	-0.0069	0.0034	-2.07	0.0384
insuficiencia_renal_aguda1:doenca_neurologica_ou_neuromuscular_cronica1	0.8609	0.4504	1.91	0.0559
deficiencia_neurologica_cronica1:patologia_hepatica1	-0.9113	0.5193	-1.75	0.0793
idade_utente_a_data_validacao:sexo_utenteM	0.0032	0.0018	1.72	0.0862
idade_utente_a_data_validacao:asma1	-0.0020	0.0013	-1.60	0.1102
AIC		62273.98		

Tabela I.6: Resultado da Regressão Logística com interações sem a interação entre as variáveis *idade_utente_a_data_validacao* e *asma*, ajustado na base de dados restrita das comorbilidades

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-9.7077	1.5821	-6.14	0.0000
idade_utente_a_data_validacao	0.0875	0.0199	4.41	0.0000
sexo_utenteM	0.4930	0.1497	3.29	0.0010
doenca_renal_cronica1	4.2988	0.3501	12.28	0.0000
neoplasia1	2.8694	0.3309	8.67	0.0000
descricao_distrito_ocorrenciaAveiro	-2.0046	1.6071	-1.25	0.2123
descricao_distrito_ocorrenciaBeja	-0.8507	1.7211	-0.49	0.6211
descricao_distrito_ocorrenciaBraga	-3.1232	1.6033	-1.95	0.0514
descricao_distrito_ocorrenciaBragança	-0.9585	1.6776	-0.57	0.5678
descricao_distrito_ocorrenciaCastelo Branco	-2.9246	1.7219	-1.70	0.0894
descricao_distrito_ocorrenciaCoimbra	-1.5703	1.6160	-0.97	0.3312
descricao_distrito_ocorrenciaÉvora	-1.6454	1.7089	-0.96	0.3356
descricao_distrito_ocorrenciaFaro	-1.9304	1.6246	-1.19	0.2347
descricao_distrito_ocorrenciaGuarda	-1.7933	1.7205	-1.04	0.2973
descricao_distrito_ocorrenciaLeiria	-1.7282	1.6265	-1.06	0.2880
descricao_distrito_ocorrenciaLisboa	-2.2021	1.5837	-1.39	0.1644
descricao_distrito_ocorrenciaMadeira	-5.6088	2.6138	-2.15	0.0319
descricao_distrito_ocorrenciaPortalegre	-2.3266	1.9104	-1.22	0.2233
descricao_distrito_ocorrenciaPorto	-3.0476	1.5908	-1.92	0.0554
descricao_distrito_ocorrenciaSantarém	-1.0047	1.6230	-0.62	0.5359
descricao_distrito_ocorrenciaSetúbal	-1.9979	1.5935	-1.25	0.2099
descricao_distrito_ocorrenciaViana do Castelo	-2.4489	1.6652	-1.47	0.1414
descricao_distrito_ocorrenciaVila Real	-2.8801	1.7074	-1.69	0.0916
descricao_distrito_ocorrenciaViseu	-2.1958	1.6306	-1.35	0.1781
diabetes1	0.9569	0.2682	3.57	0.0004
deficiencia_neurologica_cronica1	5.0223	0.5012	10.02	0.0000
patologia_hepatica1	2.7019	0.5769	4.68	0.0000
doenca_pulmonar_cronica1	2.3193	0.3673	6.31	0.0000
insuficiencia_renal_aguda1	1.3508	0.2206	6.12	0.0000
doencas_hematologicas_cronicas1	2.4118	0.5464	4.41	0.0000
idade_utente_a_data_validacao:doenca_renal_cronica1	-0.0403	0.0042	-9.50	0.0000
idade_utente_a_data_validacao:deficiencia_neurologica_cronica1	-0.0532	0.0061	-8.79	0.0000
idade_utente_a_data_validacao:neoplasia1	-0.0274	0.0042	-6.46	0.0000
idade_utente_a_data_validacao:doenca_neurologica_ou_neuromuscular_cronica1	0.0047	0.0009	5.31	0.0000
neoplasia1:diabetes1	-0.4505	0.1088	-4.14	0.0000
idade_utente_a_data_validacao:doenca_pulmonar_cronica1	-0.0230	0.0046	-5.05	0.0000
doenca_renal_cronica1:doenca_pulmonar_cronica1	-0.4018	0.1195	-3.36	0.0008
patologia_hepatica1:doenca_neurologica_ou_neuromuscular_cronica1	-0.8623	0.4629	-1.86	0.0625
idade_utente_a_data_validacao:doencas_hematologicas_cronicas1	-0.0238	0.0068	-3.51	0.0004
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaAveiro	0.0234	0.0202	1.16	0.2468
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBeja	0.0118	0.0216	0.55	0.5850
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBraga	0.0349	0.0201	1.73	0.0831
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBragança	0.0095	0.0209	0.46	0.6483
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCastelo Branco	0.0299	0.0214	1.40	0.1616
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCoimbra	0.0218	0.0203	1.07	0.2825
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaÉvora	0.0226	0.0213	1.06	0.2900
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaFaro	0.0201	0.0204	0.99	0.3230
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaGuarda	0.0236	0.0214	1.10	0.2701
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLeiria	0.0239	0.0204	1.17	0.2421
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLisboa	0.0271	0.0199	1.36	0.1725
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaMadeira	0.0549	0.0325	1.69	0.0910
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPortalegre	0.0262	0.0236	1.11	0.2683
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPorto	0.0370	0.0200	1.85	0.0642
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSantarém	0.0128	0.0204	0.63	0.5280
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSetúbal	0.0260	0.0200	1.30	0.1938
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViana do Castelo	0.0263	0.0208	1.26	0.2060
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaVila Real	0.0307	0.0213	1.44	0.1497
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViseu	0.0254	0.0204	1.24	0.2142
doenca_renal_cronica1:neoplasia1	-0.4415	0.1363	-3.24	0.0012
doenca_renal_cronica1:patologia_hepatica1	-0.6795	0.2275	-2.99	0.0028
diabetes1:doencas_hematologicas_cronicas1	-0.4065	0.1736	-2.34	0.0192
doenca_renal_cronica1:insuficiencia_renal_aguda1	-0.9108	0.3404	-2.68	0.0075
idade_utente_a_data_validacao:patologia_hepatica1	-0.0212	0.0079	-2.69	0.0071
doenca_pulmonar_cronica1:insuficiencia_cardiaca1	1.0978	0.4079	2.69	0.0071
doenca_pulmonar_cronica1:vih_outras_imunodeficiencias1	-1.2473	0.5747	-2.17	0.0300
doenca_renal_cronica1:diabetes1	-0.2008	0.0930	-2.16	0.0308
diabetes1:patologia_hepatica1	-0.4970	0.2027	-2.45	0.0142
idade_utente_a_data_validacao:diabetes1	-0.0069	0.0034	-2.06	0.0395
insuficiencia_renal_aguda1:doenca_neurologica_ou_neuromuscular_cronica1	0.8618	0.4504	1.91	0.0557
deficiencia_neurologica_cronica1:patologia_hepatica1	-0.9096	0.5193	-1.75	0.0798
idade_utente_a_data_validacao:sexo_utenteM	0.0031	0.0018	1.70	0.0898
AIC		62274.62		

I.1.3 LASSO de Grupo (*Group* LASSO)

Tabela I.7: Estimativas LASSO de Grupo.

Variáveis Explicativas	Estimativas
(Intercept)	-9.0456
idade_utente_a_data_validacao	0.1186
sexo_utenteM	0.7246
descricao_distrito_ocorrenciaAveiro	0.5726
descricao_distrito_ocorrenciaBeja	0.5489
descricao_distrito_ocorrenciaBraga	0.4461
descricao_distrito_ocorrenciaBragança	0.6878
descricao_distrito_ocorrenciaCastelo Branco	0.3134
descricao_distrito_ocorrenciaCoimbra	0.7009
descricao_distrito_ocorrenciaÉvora	0.8255
descricao_distrito_ocorrenciaFaro	0.4251
descricao_distrito_ocorrenciaGuarda	1.0191
descricao_distrito_ocorrenciaLeiria	0.7758
descricao_distrito_ocorrenciaLisboa	0.6465
descricao_distrito_ocorrenciaMadeira	0.1347
descricao_distrito_ocorrenciaPortalegre	0.2256
descricao_distrito_ocorrenciaPorto	0.5872
descricao_distrito_ocorrenciaSantarém	0.8445
descricao_distrito_ocorrenciaSetúbal	0.7401
descricao_distrito_ocorrenciaViana do Castelo	0.3281
descricao_distrito_ocorrenciaVila Real	0.2848
descricao_distrito_ocorrenciaViseu	0.6774
data_confirmado1	-0.0003
neoplasia	0.9148
diabetes	0.4712
vih_outras_imunodeficiencias	0.06858
doenca_neurologica_ou_neuromuscular_cronica	0.8659
asma	0.0000
doenca_pulmonar_cronica	0.5348
patologia_hepatica	1.2764
doencas_hematologicas_cronicas	0.4931
doenca_renal_cronica	0.9273
deficiencia_neurologica_cronica	1.4453
insuficiencia_renal_aguda	1.7183
insuficiencia_cardiaca	0.2869
coagulopatia_de_consumo	0.0000

I.1.4 Modelo Aditivo Generalizado (GAM)

Tabela I.8: Modelo inicial GAM ajustado à base de dados restrita dos comorbilidades.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.41	0.29	-38.84	0.00
idade_utente_a_data_validacao	0.11	0.00	120.59	0.00
sexo_utenteM	0.75	0.02	33.13	0.00
descricao_distrito_ocorrenciaAveiro	-0.23	0.29	-0.80	0.42
descricao_distrito_ocorrenciaBeja	0.07	0.31	0.21	0.83
descricao_distrito_ocorrenciaBraga	-0.41	0.29	-1.45	0.15
descricao_distrito_ocorrenciaBragança	-0.36	0.29	-1.25	0.21
descricao_distrito_ocorrenciaCastelo_Branco	-0.59	0.30	-1.99	0.05
descricao_distrito_ocorrenciaCoimbra	0.06	0.29	0.22	0.82
descricao_distrito_ocorrenciaÉvora	0.08	0.30	0.28	0.78
descricao_distrito_ocorrenciaFaro	-0.34	0.29	-1.16	0.25
descricao_distrito_ocorrenciaGuarda	0.03	0.30	0.10	0.92
descricao_distrito_ocorrenciaLeiria	0.10	0.29	0.33	0.74
descricao_distrito_ocorrenciaLisboa	-0.07	0.28	-0.25	0.80
descricao_distrito_ocorrenciaMadeira	-1.21	0.40	-3.00	0.00
descricao_distrito_ocorrenciaPortalegre	-0.32	0.32	-0.99	0.32
descricao_distrito_ocorrenciaPorto	-0.16	0.28	-0.58	0.56
descricao_distrito_ocorrenciaSantarém	-0.04	0.29	-0.12	0.90
descricao_distrito_ocorrenciaSetúbal	0.03	0.28	0.10	0.92
descricao_distrito_ocorrenciaViana_do_Castelo	-0.43	0.29	-1.47	0.14
descricao_distrito_ocorrenciaVila_Real	-0.50	0.30	-1.70	0.09
descricao_distrito_ocorrenciaViseu	-0.26	0.29	-0.91	0.36
s(*data_confirmado1*, df = 3)	-0.00	0.00	-1.33	0.18
neoplasia	0.60	0.05	12.74	0.00
diabetes	0.32	0.03	9.21	0.00
vih_outras_imunodeficiencias	0.33	0.19	1.76	0.08
doenca_neurologica_ou_neuromuscular_cronica	0.40	0.08	5.31	0.00
asma	-0.16	0.10	-1.57	0.12
doenca_pulmonar_cronica	0.40	0.05	8.27	0.00
patologia_hepatica	0.91	0.10	9.38	0.00
doencas_hematologicas_cronicas	0.35	0.08	4.30	0.00
doenca_renal_cronica	0.74	0.05	15.37	0.00
deficiencia_neurologica_cronica	0.57	0.08	7.00	0.00
insuficiencia_renal_aguda	1.17	0.18	6.49	0.00
insuficiencia_cardiaca	0.46	0.20	2.23	0.03
coagulopatia_de_consumo	0.02	0.61	0.04	0.97
AIC		62312.12		

Tabela I.9: Modelo final GAM ajustado à base de dados restrita das comorbilidades.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.20	0.29	-38.09	0.00
idade_utente_a_data_validacao, df = 3	0.11	0.00	108.39	0.00
sexo_utenteM	0.75	0.02	33.23	0.00
descricao_distrito_ocorrenciaAveiro	-0.23	0.29	-0.81	0.42
descricao_distrito_ocorrenciaBeja	0.06	0.31	0.21	0.83
descricao_distrito_ocorrenciaBraga	-0.41	0.29	-1.45	0.15
descricao_distrito_ocorrenciaBragança	-0.36	0.29	-1.25	0.21
descricao_distrito_ocorrenciaCastelo_Branco	-0.59	0.30	-1.99	0.05
descricao_distrito_ocorrenciaCoimbra	0.06	0.29	0.22	0.83
descricao_distrito_ocorrenciaÉvora	0.08	0.30	0.28	0.78
descricao_distrito_ocorrenciaFaro	-0.34	0.29	-1.17	0.24
descricao_distrito_ocorrenciaGuarda	0.03	0.30	0.09	0.93
descricao_distrito_ocorrenciaLeiria	0.09	0.29	0.32	0.75
descricao_distrito_ocorrenciaLisboa	-0.07	0.28	-0.25	0.80
descricao_distrito_ocorrenciaMadeira	-1.21	0.40	-3.01	0.00
descricao_distrito_ocorrenciaPortalegre	-0.32	0.32	-0.99	0.32
descricao_distrito_ocorrenciaPorto	-0.17	0.28	-0.58	0.56
descricao_distrito_ocorrenciaSantarém	-0.04	0.29	-0.13	0.90
descricao_distrito_ocorrenciaSetúbal	0.03	0.28	0.10	0.92
descricao_distrito_ocorrenciaViana_do_Castelo	-0.43	0.29	-1.48	0.14
descricao_distrito_ocorrenciaVila_Real	-0.51	0.30	-1.71	0.09
descricao_distrito_ocorrenciaViseu	-0.26	0.29	-0.92	0.36
s(*data_confirmado1*, df = 3)	-0.00	0.00	-1.35	0.18
neoplasia	0.60	0.05	12.79	0.00
diabetes	0.32	0.03	9.15	0.00
doenca_neurologica_ou_neuromuscular_cronica	0.41	0.08	5.34	0.00
doenca_pulmonar_cronica	0.40	0.05	8.21	0.00
patologia_hepatica	0.92	0.10	9.47	0.00
doencas_hematologicas_cronicas	0.35	0.08	4.34	0.00
doenca_renal_cronica	0.75	0.05	15.43	0.00
deficiencia_neurologica_cronica	0.57	0.08	7.01	0.00
insuficiencia_renal_aguda	1.18	0.18	6.66	0.00
insuficiencia_cardiaca	0.45	0.20	2.22	0.03
AIC		62311.62		

Tabela I.10: Modelo GAM ajustado à base de dados balanceada das comorbilidades.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.86	0.67	-13.17	0.00
idade_utente_a_data_validacao	0.12	0.00	68.74	0.00
sexo_utenteM	0.78	0.05	16.28	0.00
descricao_distrito_ocorrenciaAveiro	0.14	0.66	0.21	0.83
descricao_distrito_ocorrenciaBeja	0.39	0.70	0.55	0.58
descricao_distrito_ocorrenciaBraga	-0.07	0.66	-0.11	0.91
descricao_distrito_ocorrenciaBragança	0.31	0.68	0.46	0.65
descricao_distrito_ocorrenciaCastelo_Branco	-0.31	0.68	-0.45	0.65
descricao_distrito_ocorrenciaCoimbra	0.36	0.67	0.54	0.59
descricao_distrito_ocorrenciaÉvora	0.52	0.69	0.76	0.45
descricao_distrito_ocorrenciaFaro	-0.00	0.67	-0.01	0.99
descricao_distrito_ocorrenciaGuarda	0.44	0.69	0.63	0.53
descricao_distrito_ocorrenciaLeiria	0.38	0.67	0.57	0.57
descricao_distrito_ocorrenciaLisboa	0.28	0.66	0.42	0.68
descricao_distrito_ocorrenciaMadeira	-0.44	0.80	-0.55	0.58
descricao_distrito_ocorrenciaPortalegre	-0.23	0.72	-0.32	0.75
descricao_distrito_ocorrenciaPorto	0.17	0.66	0.25	0.80
descricao_distrito_ocorrenciaSantarém	0.44	0.67	0.66	0.51
descricao_distrito_ocorrenciaSetúbal	0.38	0.66	0.57	0.57
descricao_distrito_ocorrenciaViana_do_Castelo	-0.11	0.67	-0.17	0.87
descricao_distrito_ocorrenciaVila_Real	-0.29	0.68	-0.42	0.67
descricao_distrito_ocorrenciaViseu	0.32	0.67	0.48	0.63
s(*data_confirmado1*, df = 3)	-0.00	0.00	-1.55	0.12
neoplasia	0.99	0.12	8.14	0.00
diabetes	0.47	0.08	5.58	0.00
vih_outras_imunodeficiencias	0.24	0.34	0.70	0.48
doenca_neurologica_ou_neuromuscular_cronica	0.85	0.28	3.08	0.00
asma	-0.04	0.19	-0.22	0.83
doenca_pulmonar_cronica	0.61	0.13	4.56	0.00
patologia_hepatica	1.70	0.27	6.24	0.00
doencas_hematologicas_cronicas	0.51	0.24	2.12	0.03
doenca_renal_cronica	1.01	0.15	6.85	0.00
deficiencia_neurologica_cronica	1.61	0.31	5.13	0.00
insuficiencia_renal_aguda	2.77	1.10	2.53	0.01
insuficiencia_cardiaca	0.58	0.82	0.70	0.48
coagulopatia_de_consumo	6.16	39.35	0.16	0.88
AIC		12158.43		

Tabela I.11: *OR* variáveis explicativas referentes às comorbilidades com efeitos paramétricos do modelo modelo GAM final ajustado à base de dados restrita das comorbilidades.

		Intervalo Confiança	
Variável	***Odds-Ratio***	**2.5%**	**97.5%**
idade_utente_a_data_validacao		1.12	1.11
1.12			
sexo_utenteM	2.12	2.03	2.22
descricao_distrito_ocorrenciaAveiro	0.79	0.45	1.39
descricao_distrito_ocorrenciaBeja	1.07	0.58	1.95
descricao_distrito_ocorrenciaBraga	0.66	0.38	1.16
descricao_distrito_ocorrenciaBragança	0.69	0.39	1.23
descricao_distrito_ocorrenciaCastelo_Branco	0.55	0.31	0.99
descricao_distrito_ocorrenciaCoimbra	1.06	0.61	1.87
descricao_distrito_ocorrenciaÉvora	1.09	0.61	1.95
descricao_distrito_ocorrenciaFaro	0.71	0.40	1.26
descricao_distrito_ocorrenciaGuarda	1.03	0.58	1.84
descricao_distrito_ocorrenciaLeiria	1.10	0.62	1.93
descricao_distrito_ocorrenciaLisboa	0.93	0.53	1.62
descricao_distrito_ocorrenciaMadeira	0.30	0.14	0.65
descricao_distrito_ocorrenciaPortalegre	0.73	0.39	1.37
descricao_distrito_ocorrenciaPorto	0.85	0.49	1.48
descricao_distrito_ocorrenciaSantarém	0.96	0.55	1.70
descricao_distrito_ocorrenciaSetúbal	1.03	0.59	1.80
descricao_distrito_ocorrenciaViana_do_Castelo	0.65	0.37	1.15
descricao_distrito_ocorrenciaVila_Real	0.60	0.34	1.08
descricao_distrito_ocorrenciaViseu	0.77	0.44	1.35
neoplasia	1.82	1.66	2.00
diabetes	1.37	1.28	1.47
doenca_neurologica_ou_neuromuscular_cronica	1.50	1.29	1.74
doenca_pulmonar_cronica	1.49	1.35	1.63
patologia_hepatica	2.51	2.08	3.04
doencas_hematologicas_cronicas	1.42	1.21	1.67
doenca_renal_cronica	2.11	1.92	2.32
deficiencia_neurologica_cronica	1.77	1.51	2.08
insuficiencia_renal_aguda	3.26	2.30	4.62
insuficiencia_cardiaca	1.57	1.05	2.34

I.1.5 Árvore de Classificação-Dados Balanceados

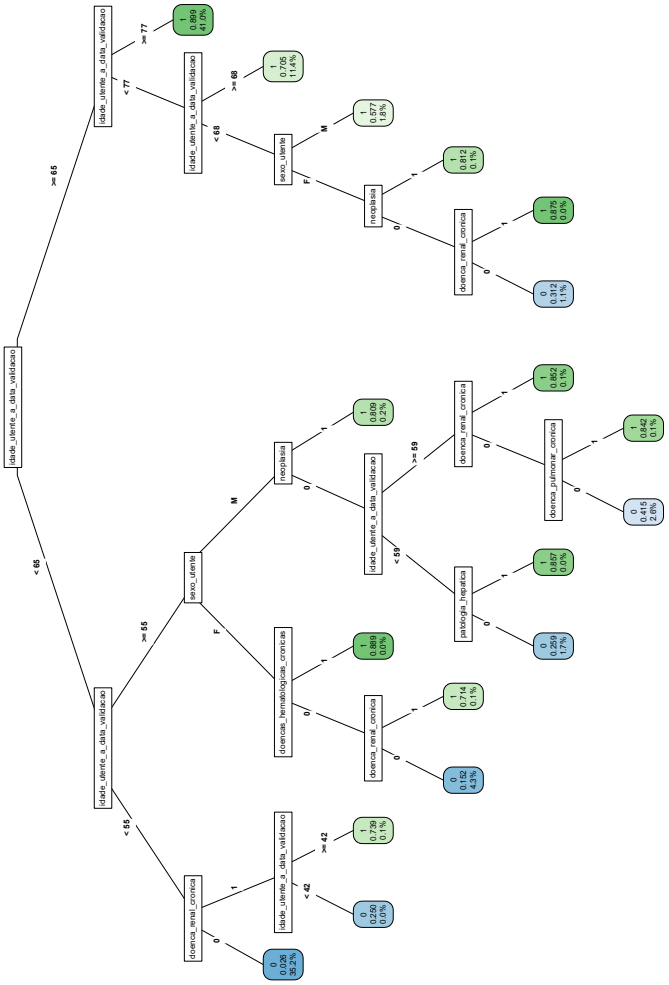


Figura I.1: Árvore de Decisão final.

I.1.5.1 Descrição dos resultados da Árvore de Classificação

Através da observação da Árvore de Classificação representada na Figura I.1, pode-se observar que a variável explicativa *idade_utente_a_data_validacao* é das variáveis explicativas mais vezes utilizadas para o processo da divisão binária de um nó, pelo que, se pode afirmar que esta é uma variável bastante importante para explicar o falecimento de um indivíduo que tenha testado positivo à COVID-19, como se virá de seguida. Através da análise da Figura I.1, podem-se observar que:

- Se um indivíduo tiver mais de 77 anos, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos e se tiver 68 ou mais anos, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se tiver menos 68 e se o seu sexo for o masculino, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se tiver menos 68, se o seu sexo for o feminino e apresentar a comorbilidade neoplasia, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se tiver menos 68, se o seu sexo for o feminino, não apresentar a comorbilidade neoplasia e apresentar a comorbilidade doença renal crónica, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se tiver menos 68, se o seu sexo for o feminino, não apresentar a comorbilidade neoplasia e não apresentar a comorbilidade doença renal crónica, então a Árvore de Classificação indica-nos prognóstico sebrevivência;
- Se um indivíduo tiver menos 65 anos, se tiver menos de 55 anos e não apresentar a comorbilidade doença renal crónica, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver menos 65 anos, se tiver menos de 55 anos, se apresentar a comorbilidade doença renal crónica e se tiver 42 ou mais anos, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver menos 65 anos, se tiver menos de 55 anos, se apresentar a comorbilidade doença renal crónica e se tiver menos 42 anos, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o feminino e apresentar a comorbilidade doença hematológica crónica, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o feminino, não apresentar a comorbilidade doença hematológica crónica e apresentar a doença renal crónica, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o feminino, não apresentar a comorbilidade doença hematológica crónica e não apresentar a doença renal crónica, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o masculino e apresentar a comorbilidade neoplasia, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o masculino, não apresentar a comorbilidade neoplasia, se tiver menos de 59 anos e apresentar a comorbilidade patologia hepática, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o masculino, não apresentar a comorbilidade neoplasia, se tiver menos de 59 anos e não apresentar a comorbilidade patologia hepática, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o masculino, não apresentar a comorbilidade neoplasia, se tiver 59 ou mais anos e se apresentar a comorbilidade doença renal crónica, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o masculino, não apresentar a comorbilidade neoplasia, se tiver 59 ou mais anos, se não apresentar a comorbilidade doença renal crónica e se apresentar a comorbilidade doença pulmonar crónica, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos 65 anos, se tiver 55 ou mais anos, se o seu sexo for o masculino, não apresentar a comorbilidade neoplasia, se tiver 59 ou mais anos, se não apresentar a comorbilidade doença renal crónica e se não apresentar a comorbilidade doença pulmonar crónica, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

I.2 Dados Sintomas

I.2.1 Regressão Logística - Dados Desequilibrados

Tabela I.12: Resultado da Regressão Logística com o método *stepwise* na base de dados restrita dos sintomas.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.9834	0.2951	-37.22	0.0000
idade_utente_a_data_validacao	0.1023	0.0009	108.39	0.0000
dispneia	1.1796	0.0302	39.08	0.0000
sexo_utente	0.7009	0.0232	30.21	0.0000
radiografia_pulmonar_com_alteracoes	0.6330	0.0401	15.80	0.0000
mialgias	-0.5272	0.0432	-12.21	0.0000
descricao_distrito_ocorrenciaAveiro	-0.1591	0.2879	-0.55	0.5805
descricao_distrito_ocorrenciaBeja	0.1248	0.3105	0.40	0.6877
descricao_distrito_ocorrenciaBraga	-0.3614	0.2871	-1.26	0.2081
descricao_distrito_ocorrenciaBragança	-0.1226	0.2926	-0.42	0.6752
descricao_distrito_ocorrenciaCastelo Branco	-0.4227	0.2988	-1.41	0.1571
descricao_distrito_ocorrenciaCoimbra	0.0070	0.2888	0.02	0.9807
descricao_distrito_ocorrenciaÉvora	0.2611	0.2991	0.87	0.3827
descricao_distrito_ocorrenciaFaro	-0.1464	0.2923	-0.50	0.6166
descricao_distrito_ocorrenciaGuarda	-0.1555	0.2986	-0.52	0.6024
descricao_distrito_ocorrenciaLeiria	0.2165	0.2905	0.75	0.4561
descricao_distrito_ocorrenciaLisboa	0.0139	0.2848	0.05	0.9610
descricao_distrito_ocorrenciaMadeira	-1.1970	0.4048	-2.96	0.0031
descricao_distrito_ocorrenciaPortalegre	-0.1460	0.3224	-0.45	0.6506
descricao_distrito_ocorrenciaPorto	-0.0741	0.2855	-0.26	0.7953
descricao_distrito_ocorrenciaSantarém	0.0465	0.2909	0.16	0.8730
descricao_distrito_ocorrenciaSetúbal	0.0911	0.2863	0.32	0.7502
descricao_distrito_ocorrenciaViana do Castelo	-0.3164	0.2927	-1.08	0.2796
descricao_distrito_ocorrenciaVila Real	-0.4561	0.2983	-1.53	0.1263
descricao_distrito_ocorrenciaViseu	-0.2502	0.2898	-0.86	0.3879
historia_de_febre_ou_calafrios	0.4098	0.0287	14.27	0.0000
coriza	-0.6020	0.0686	-8.77	0.0000
cefaleia	-0.5008	0.0554	-9.05	0.0000
tosse_seca_ou_produtiva	-0.2180	0.0284	-7.67	0.0000
taquicardia	0.4988	0.0914	5.46	0.0000
coma	1.3276	0.2341	5.67	0.0000
odinofagia	-0.3985	0.0765	-5.21	0.0000
irritabilidade_confusao	0.3138	0.0838	3.74	0.0002
fraqueza_geral_ou_astenia	0.1232	0.0335	3.67	0.0002
dor_no_peito	-0.2143	0.0687	-3.12	0.0018
pneumonia	0.1829	0.0584	3.13	0.0017
nauseas_vomitos	-0.1433	0.0585	-2.45	0.0143
auscultacao_pulmonar_anomala	-0.1327	0.0615	-2.16	0.0309
artralgia	0.2169	0.1022	2.12	0.0338
AIC		59562.97		

Tabela I.13: *OR* variáveis explicativas do modelo final da Regressão Logística referentes aos sintomas.

		Intervalo Confiança	
Variável	***Odds-Ratio***	**2.5%**	**97.5%**
idade_utente_a_data_validacao	1.11	1.11	1.11
dispneia1	3.25	3.07	3.45
sexo_utenteM	2.02	1.93	2.11
radiografia_pulmonar_com_alteracoes1	1.88	1.74	2.04
mialgias1	0.59	0.54	0.64
descricao_distrito_ocorrenciaAveiro	0.85	0.49	1.50
descricao_distrito_ocorrenciaBeja	1.13	0.62	2.08
descricao_distrito_ocorrenciaBraga	0.70	0.40	1.22
descricao_distrito_ocorrenciaBragança	0.88	0.50	1.57
descricao_distrito_ocorrenciaCastelo_Branco	0.66	0.36	1.18
descricao_distrito_ocorrenciaCoimbra	1.01	0.57	1.77
descricao_distrito_ocorrenciaÉvora	1.30	0.72	2.33
descricao_distrito_ocorrenciaFaro	0.86	0.49	1.53
descricao_distrito_ocorrenciaGuarda	0.86	0.48	1.54
descricao_distrito_ocorrenciaLeiria	1.24	0.70	2.19
descricao_distrito_ocorrenciaLisboa	1.01	0.58	1.77
descricao_distrito_ocorrenciaMadeira	0.30	0.14	0.67
descricao_distrito_ocorrenciaPortalegre	0.86	0.46	1.63
descricao_distrito_ocorrenciaPorto	0.93	0.53	1.63
descricao_distrito_ocorrenciaSantarém	1.05	0.59	1.85
descricao_distrito_ocorrenciaSetúbal	1.10	0.63	1.92
descricao_distrito_ocorrenciaViana_do_Castelo	0.73	0.41	1.29
descricao_distrito_ocorrenciaVila_Real	0.63	0.35	1.14
descricao_distrito_ocorrenciaViseu	0.78	0.44	1.37
historia_de_febre_ou_calafrios1	1.51	1.42	1.59
coriza1	0.55	0.48	0.63
cefaleia1	0.61	0.54	0.68
tosse_seca_ou_produtiva1	0.80	0.76	0.85
taquicardia1	1.65	1.38	1.97
coma1	3.77	2.38	5.97
odinofagia1	0.67	0.58	0.78
irritabilidade_confusao1	1.37	1.16	1.61
fraqueza_geral_ou_astenia1	1.13	1.06	1.21
dor_no_peito1	0.81	0.71	0.92
pneumonia1	1.20	1.07	1.35
nauseas_vomitos1	0.87	0.77	0.97
auscultacao_pulmonar_anomala1	0.88	0.78	0.99
artralgia1	1.24	1.02	1.52

I.2.2 Regressão Logística com interações - Dados Desequilibrados

Tabela I.14: Regressão Logistica com interações com o método do *stepwise*, modelo final.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.6728	1.7008	-6.28	0.0000
idade_utente_a_data_validacao	0.0982	0.0213	4.60	0.0000
dispneia	3.5067	0.2058	17.04	0.0000
sexo_utenteM	0.3604	0.1522	2.37	0.0179
radiografia_pulmonar_com_alteracoes	2.4019	0.2634	9.12	0.0000
mialgias	-1.5477	0.2660	-5.82	0.0000
descricao_distrito_ocorrenciaAveiro	-0.1928	1.7234	-0.11	0.9109
descricao_distrito_ocorrenciaBeja	0.9488	1.8351	0.52	0.6051
descricao_distrito_ocorrenciaBraga	-1.3925	1.7199	-0.81	0.4182
descricao_distrito_ocorrenciaBragança	0.0875	1.7927	0.05	0.9611
descricao_distrito_ocorrenciaCastelo Branco	-1.0593	1.8336	-0.58	0.5634
descricao_distrito_ocorrenciaCoimbra	-0.0537	1.7322	-0.03	0.9753
descricao_distrito_ocorrenciaÉvora	-0.2522	1.8235	-0.14	0.8900
descricao_distrito_ocorrenciaFaro	0.1172	1.7393	0.07	0.9463
descricao_distrito_ocorrenciaGuarda	-0.1989	1.8410	-0.11	0.9140
descricao_distrito_ocorrenciaLeiria	0.2563	1.7422	0.15	0.8830
descricao_distrito_ocorrenciaLisboa	-0.3414	1.7010	-0.20	0.8409
descricao_distrito_ocorrenciaMadeira	-3.4127	2.7178	-1.26	0.2092
descricao_distrito_ocorrenciaPortalegre	-1.0069	2.0144	-0.50	0.6172
descricao_distrito_ocorrenciaPorto	-1.1879	1.7080	-0.70	0.4867
descricao_distrito_ocorrenciaSantarém	0.7597	1.7392	0.44	0.6623
descricao_distrito_ocorrenciaSetúbal	-0.3199	1.7103	-0.19	0.8516
descricao_distrito_ocorrenciaViana do Castelo	-0.8152	1.7813	-0.46	0.6472
descricao_distrito_ocorrenciaVila Real	-0.5983	1.8153	-0.33	0.7417
descricao_distrito_ocorrenciaViseu	-0.7396	1.7479	-0.42	0.6722
historia_de_febre_ou_calafrios	0.5567	0.0432	12.88	0.0000
coriza	-2.6924	0.4759	-5.66	0.0000
cefaleia	-1.2713	0.3198	-3.98	0.0001
tosse_seca_ou_produtiva	-0.7429	0.1898	-3.91	0.0001
taquicardia	1.3562	0.5349	2.54	0.0112
coma	1.2550	0.2266	5.54	0.0000
odinofagia	-0.5546	0.1087	-5.10	0.0000
irritabilidade_confusao	2.9832	0.6497	4.59	0.0000
dor_no_peito	-0.6325	0.4364	-1.45	0.1472
pneumonia	1.2330	0.1195	10.32	0.0000
nauseas_vomitos	-0.1426	0.0841	-1.70	0.0900
fraqueza_geral_ou_astenia	0.0832	0.0359	2.32	0.0205
artralgia	-2.3067	0.8069	-2.86	0.0043
idade_utente_a_data_validacao:dispneia	-0.0258	0.0025	-10.36	0.0000
dispneia:historia_de_febre_ou_calafrios	-0.4541	0.0598	-7.60	0.0000
dispneia:radiografia_pulmonar_com_alteracoes	-0.3755	0.0775	-4.84	0.0000
idade_utente_a_data_validacao: radiografia_pulmonar_com_alteracoes	-0.0174	0.0031	-5.61	0.0000
idade_utente_a_data_validacao:mialgias	0.0129	0.0034	3.75	0.0002
idade_utente_a_data_validacao:coriza	0.0249	0.0059	4.18	0.0000
radiografia_pulmonar_com_alteracoes:cefaleia	0.4625	0.1338	3.46	0.0005
cefaleia:odinofagia	0.5892	0.1752	3.36	0.0008
idade_utente_a_data_validacao:irritabilidade_confusao	-0.0312	0.0077	-4.08	0.0000
historia_de_febre_ou_calafrios:tosse_seca_ou_produtiva	0.1981	0.0572	3.46	0.0005
radiografia_pulmonar_com_alteracoes: historia_de_febre_ou_calafrios	-0.1370	0.0734	-1.86	0.0622
radiografia_pulmonar_com_alteracoes:irritabilidade_confusao	-0.3995	0.1582	-2.52	0.0116
radiografia_pulmonar_com_alteracoes:pneumonia	-0.7561	0.1084	-6.98	0.0000
dispneia:pneumonia	-0.5174	0.1081	-4.78	0.0000
tosse_seca_ou_produtiva:taquicardia	0.4032	0.1718	2.35	0.0189
coriza:cefaleia	0.4583	0.1761	2.60	0.0093
idade_utente_a_data_validacao:tosse_seca_ou_produtiva	0.0050	0.0023	2.15	0.0317
historia_de_febre_ou_calafrios:pneumonia	-0.2656	0.1004	-2.65	0.0081
tosse_seca_ou_produtiva:irritabilidade_confusao	0.4284	0.1569	2.73	0.0063
radiografia_pulmonar_com_alteracoes:dor_no_peito	-0.2693	0.1441	-1.87	0.0616
radiografia_pulmonar_com_alteracoes:odinofagia	0.5401	0.1977	2.73	0.0063
coriza:fraqueza_geral_ou_astenia	0.3291	0.1425	2.31	0.0209
idade_utente_a_data_validacao:taquicardia	-0.0137	0.0065	-2.10	0.0353
idade_utente_a_data_validacao:sexo_utenteM	0.0039	0.0019	2.09	0.0368
historia_de_febre_ou_calafrios:nauseas_vomitos	-0.2415	0.1153	-2.09	0.0362
cefaleia:nauseas_vomitos	0.3544	0.1589	2.23	0.0257
dispneia:dor_no_peito	-0.2953	0.1462	-2.02	0.0434
idade_utente_a_data_validacao:cefaleia	0.0069	0.0042	1.65	0.0983
cefaleia:irritabilidade_confusao	-0.7238	0.3611	-2.00	0.0450
irritabilidade_confusao:nauseas_vomitos	0.5055	0.2455	2.06	0.0395
irritabilidade_confusao:dor_no_peito	-0.7223	0.3976	-1.82	0.0693

Tabela I.15: Regressão Logística com interações com o método do *stepwise*, modelo final.

	Estimate	Std. Error	z value	Pr(>|z|)
mialgias:artralgia	1.1578	0.2714	4.27	0.0000
taquicardia:artralgia	1.1998	0.4827	2.49	0.0129
idade_utente_a_data_validacao:artralgia	0.0204	0.0095	2.16	0.0309
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaAveiro	0.0005	0.0216	0.02	0.9832
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBeja	-0.0106	0.0231	-0.46	0.6462
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBraga	0.0131	0.0216	0.61	0.5448
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBragança	-0.0020	0.0223	-0.09	0.9286
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCastelo Branco	0.0081	0.0228	0.36	0.7209
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCoimbra	0.0011	0.0217	0.05	0.9594
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaÉvora	0.0066	0.0228	0.29	0.7730
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaFaro	-0.0030	0.0218	-0.14	0.8902
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaGuarda	0.0010	0.0229	0.04	0.9660
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLeiria	-0.0002	0.0218	-0.01	0.9913
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLisboa	0.0046	0.0214	0.22	0.8293
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaMadeira	0.0288	0.0338	0.85	0.3937
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPortalegre	0.0109	0.0249	0.44	0.6621
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPorto	0.0142	0.0214	0.66	0.5089
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSantarém	-0.0085	0.0218	-0.39	0.6960
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSetúbal	0.0053	0.0215	0.25	0.8043
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViana do Castelo	0.0065	0.0223	0.29	0.7702
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaVila Real	0.0023	0.0227	0.10	0.9179
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViseu	0.0064	0.0219	0.29	0.7692
pneumonia:fraqueza_geral_ou_astenia	-0.1750	0.0980	-1.79	0.0741
idade_utente_a_data_validacao:dor_no_peito	0.0084	0.0055	1.52	0.1274
cefaleia:pneumonia	0.3021	0.1963	1.54	0.1237
dispneia:odinofagia	-0.2505	0.1736	-1.44	0.1492
AIC		58936.97		

Tabela I.16: Regressão Logística com interações modelo final, base de dados restrita dos sintomas.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.6491	1.6976	-6.27	0.0000
idade_utente_a_data_validacao	0.0979	0.0213	4.59	0.0000
dispneia	3.4413	0.2033	16.92	0.0000
sexo_utenteM	0.3617	0.1521	2.38	0.0174
radiografia_pulmonar_com_alteracoes	2.4156	0.2631	9.18	0.0000
mialgias	-1.5594	0.2659	-5.86	0.0000
historia_de_febre_ou_calafrios	0.5624	0.0431	13.05	0.0000
descricao_distrito_ocorrenciaAveiro	-0.2218	1.7203	-0.13	0.8974
descricao_distrito_ocorrenciaBeja	0.9209	1.8317	0.50	0.6152
descricao_distrito_ocorrenciaBraga	-1.4109	1.7168	-0.82	0.4112
descricao_distrito_ocorrenciaBragança	0.0690	1.7896	0.04	0.9692
descricao_distrito_ocorrenciaCastelo Branco	-1.0823	1.8306	-0.59	0.5544
descricao_distrito_ocorrenciaCoimbra	-0.0700	1.7290	-0.04	0.9677
descricao_distrito_ocorrenciaÉvora	-0.2695	1.8203	-0.15	0.8823
descricao_distrito_ocorrenciaFaro	0.0944	1.7362	0.05	0.9567
descricao_distrito_ocorrenciaGuarda	-0.2062	1.8377	-0.11	0.9107
descricao_distrito_ocorrenciaLeiria	0.2282	1.7390	0.13	0.8956
descricao_distrito_ocorrenciaLisboa	-0.3648	1.6978	-0.21	0.8299
descricao_distrito_ocorrenciaMadeira	-3.4411	2.7132	-1.27	0.2047
descricao_distrito_ocorrenciaPortalegre	-1.0283	2.0116	-0.51	0.6092
descricao_distrito_ocorrenciaPorto	-1.2109	1.7048	-0.71	0.4776
descricao_distrito_ocorrenciaSantarém	0.7331	1.7361	0.42	0.6728
descricao_distrito_ocorrenciaSetúbal	-0.3428	1.7072	-0.20	0.8409
descricao_distrito_ocorrenciaViana do Castelo	-0.8377	1.7782	-0.47	0.6376
descricao_distrito_ocorrenciaVila Real	-0.6194	1.8124	-0.34	0.7325
descricao_distrito_ocorrenciaViseu	-0.7622	1.7447	-0.44	0.6622
coriza	-2.7075	0.4762	-5.69	0.0000
cefaleia	-1.2804	0.3191	-4.01	0.0001
tosse_seca_ou_produtiva	-0.7496	0.1893	-3.96	0.0001
taquicardia	1.3813	0.5344	2.58	0.0098
coma	1.2548	0.2266	5.54	0.0000
odinofagia	-0.6177	0.1011	-6.11	0.0000
irritabilidade_confusao	2.9799	0.6510	4.58	0.0000
pneumonia	1.2113	0.1184	10.23	0.0000
nauseas_vomitos	-0.1393	0.0841	-1.66	0.0977
fraqueza_geral_ou_astenia	0.0623	0.0339	1.84	0.0663
artralgia	-2.3236	0.8062	-2.88	0.0040
idade_utente_a_data_validacao:dispneia	-0.0250	0.0025	-10.16	0.0000
dispneia:historia_de_febre_ou_calafrios	-0.4589	0.0597	-7.68	0.0000
dispneia:radiografia_pulmonar_com_alteracoes	-0.3775	0.0774	-4.88	0.0000
idade_utente_a_data_validacao:radiografia_pulmonar_com_alteracoes	-0.0175	0.0031	-5.65	0.0000
idade_utente_a_data_validacao:mialgias	0.0131	0.0034	3.80	0.0001
idade_utente_a_data_validacao:coriza	0.0250	0.0059	4.21	0.0000
radiografia_pulmonar_com_alteracoes:cefaleia	0.5297	0.1235	4.29	0.0000
cefaleia:odinofagia	0.5731	0.1751	3.27	0.0011
idade_utente_a_data_validacao:irritabilidade_confusao	-0.0312	0.0077	-4.07	0.0000
historia_de_febre_ou_calafrios:tosse_seca_ou_produtiva	0.1964	0.0572	3.43	0.0006
radiografia_pulmonar_com_alteracoes:historia_de_febre_ou_calafrios	-0.1406	0.0734	-1.92	0.0553
radiografia_pulmonar_com_alteracoes:irritabilidade_confusao	-0.4291	0.1577	-2.72	0.0065
radiografia_pulmonar_com_alteracoes:pneumonia	-0.7794	0.1066	-7.31	0.0000
dispneia:pneumonia	-0.5267	0.1082	-4.87	0.0000
tosse_seca_ou_produtiva:taquicardia	0.3967	0.1717	2.31	0.0209
coriza:cefaleia	0.4483	0.1762	2.54	0.0110
idade_utente_a_data_validacao:tosse_seca_ou_produtiva	0.0051	0.0023	2.21	0.0274
historia_de_febre_ou_calafrios:pneumonia	-0.2608	0.1003	-2.60	0.0093
tosse_seca_ou_produtiva:irritabilidade_confusao	0.4190	0.1571	2.67	0.0077
radiografia_pulmonar_com_alteracoes:dor_no_peito	-0.2732	0.1381	-1.98	0.0478
radiografia_pulmonar_com_alteracoes:odinofagia	0.4385	0.1832	2.39	0.0167
coriza:fraqueza_geral_ou_astenia	0.3305	0.1426	2.32	0.0205
idade_utente_a_data_validacao:taquicardia	-0.0140	0.0065	-2.16	0.0305
idade_utente_a_data_validacao:sexo_utenteM	0.0039	0.0019	2.08	0.0373
historia_de_febre_ou_calafrios:nauseas_vomitos	-0.2468	0.1153	-2.14	0.0323
cefaleia:nauseas_vomitos	0.3434	0.1588	2.16	0.0305
dispneia:dor_no_peito	-0.2964	0.0966	-3.07	0.0021
idade_utente_a_data_validacao:cefaleia	0.0072	0.0042	1.73	0.0841
cefaleia:irritabilidade_confusao	-0.6506	0.3557	-1.83	0.0674
irritabilidade_confusao:nauseas_vomitos	0.4965	0.2458	2.02	0.0434
irritabilidade_confusao:dor_no_peito	-0.6698	0.3929	-1.70	0.0883
mialgias:artralgia	1.1502	0.2716	4.24	0.0000

Tabela I.17: Regressão Logística com interações modelo final, base de dados restrita dos sintomas.

	Estimate	Std. Error	z value	Pr(>|z|)
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaAveiro	0.0008	0.0216	0.04	0.9709
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBeja	-0.0103	0.0230	-0.45	0.6558
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBraga	0.0133	0.0216	0.62	0.5381
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaBragança	-0.0018	0.0223	-0.08	0.9358
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCastelo Branco	0.0084	0.0228	0.37	0.7128
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaCoimbra	0.0013	0.0217	0.06	0.9529
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaÉvora	0.0068	0.0227	0.30	0.7657
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaFaro	-0.0027	0.0218	-0.13	0.8997
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaGuarda	0.0011	0.0229	0.05	0.9627
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLeiria	0.0001	0.0218	0.00	0.9968
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaLisboa	0.0049	0.0213	0.23	0.8192
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaMadeira	0.0292	0.0338	0.86	0.3870
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPortalegre	0.0111	0.0249	0.45	0.6545
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaPorto	0.0144	0.0214	0.67	0.5005
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSantarém	-0.0082	0.0218	-0.38	0.7064
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaSetúbal	0.0056	0.0214	0.26	0.7946
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViana do Castelo	0.0068	0.0222	0.30	0.7611
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaVila Real	0.0026	0.0227	0.11	0.9092
idade_utente_a_data_validacao:descricao_distrito_ocorrenciaViseu	0.0067	0.0219	0.31	0.7598
taquicardia:artralgia	1.1582	0.4835	2.40	0.0166
idade_utente_a_data_validacao:artralgia	0.0207	0.0094	2.19	0.0286
AIC		58936.49		

I.2.3 LASSO de Grupo (*Group* LASSO)

Tabela I.18: Estimativas LASSO de Grupo, base de dados balanceada.

Variáveis Explicativas	Estimativas
Intercept	-8.6341
idade_utente_a_data_validacao	0.1108
sexo_utenteM	0.6834
descricao_distrito_ocorrenciaAveiro	0.5281
descricao_distrito_ocorrenciaBeja	0.4448
descricao_distrito_ocorrenciaBraga	0.4654
descricao_distrito_ocorrenciaBragança	0.8076
descricao_distrito_ocorrenciaCastelo Branco	0.4679
descricao_distrito_ocorrenciaCoimbra	0.5389
descricao_distrito_ocorrenciaÉvora	0.8439
descricao_distrito_ocorrenciaFaro	0.5478
descricao_distrito_ocorrenciaGuarda	0.7870
descricao_distrito_ocorrenciaLeiria	0.8016
descricao_distrito_ocorrenciaLisboa	0.6886
descricao_distrito_ocorrenciaMadeira	0.1929
descricao_distrito_ocorrenciaPortalegre	0.2999
descricao_distrito_ocorrenciaPorto	0.6147
descricao_distrito_ocorrenciaSantarém	0.8439
descricao_distrito_ocorrenciaSetúbal	0.7227
descricao_distrito_ocorrenciaViana do Castelo	0.3768
descricao_distrito_ocorrenciaVila Real	0.32475
descricao_distrito_ocorrenciaViseu	0.6160
data_confirmado1	0
historia_de_febre_ou_calafrios1	0.5091
pneumonia1	0.42030
tosse_seca_ou_produtiva1	-0.2141
dispneia1	1.3329
coriza1	-0.7536
odinofagia1	-0.4375
cefaleia1	-0.4225
dor_abdominal1	0.1144
dor_no_peito1	-0.3790
artralgia1	0.0538
mialgias1	-0.5144
nauseas_vomitos1	0
convulsoes1	0
irritabilidade_confusao1	0.1953
fraqueza_geral_ou_astenia1	0.1720
auscultacao_pulmonar_anomala1	0
radiografia_pulmonar_com_alteracoes1	0.7402
coma1	0.2164
taquicardia1	0.7450

I.2.4 Modelo GAM

Tabela I.19: Modelo inicial GAM ajustado à base de dados restrita dos sintomas.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.93	0.30	-36.62	0.00
idade_utente_a_data_validacao	0.10	0.00	107.37	0.00
sexo_utenteM	0.70	0.02	30.06	0.00
descricao_distrito_ocorrenciaAveiro	-0.24	0.29	-0.81	0.42
descricao_distrito_ocorrenciaBeja	0.08	0.31	0.24	0.81
descricao_distrito_ocorrenciaBraga	-0.43	0.29	-1.49	0.14
descricao_distrito_ocorrenciaBragança	-0.20	0.29	-0.67	0.50
descricao_distrito_ocorrenciaCastelo_Branco	-0.50	0.30	-1.68	0.09
descricao_distrito_ocorrenciaCoimbra	-0.08	0.29	-0.28	0.78
descricao_distrito_ocorrenciaÉvora	0.19	0.30	0.64	0.52
descricao_distrito_ocorrenciaFaro	-0.19	0.29	-0.64	0.52
descricao_distrito_ocorrenciaGuarda	-0.23	0.30	-0.75	0.45
descricao_distrito_ocorrenciaLeiria	0.13	0.29	0.46	0.65
descricao_distrito_ocorrenciaLisboa	-0.03	0.29	-0.10	0.92
descricao_distrito_ocorrenciaMadeira	-1.14	0.41	-2.80	0.00
descricao_distrito_ocorrenciaPortalegre	-0.24	0.32	-0.74	0.46
descricao_distrito_ocorrenciaPorto	-0.13	0.29	-0.46	0.65
descricao_distrito_ocorrenciaSantarém	-0.01	0.29	-0.05	0.96
descricao_distrito_ocorrenciaSetúbal	0.03	0.29	0.10	0.92
descricao_distrito_ocorrenciaViana_do_Castelo	-0.39	0.29	-1.34	0.18
descricao_distrito_ocorrenciaVila_Real	-0.55	0.30	-1.82	0.07
descricao_distrito_ocorrenciaViseu	-0.34	0.29	-1.15	0.25
s(*data_confirmado1*, df = 3)	0.00	0.00	1.38	0.17
historia_de_febre_ou_calafrios	0.42	0.03	14.46	0.00
pneumonia	0.23	0.06	3.92	0.00
tosse_seca_ou_produtiva	-0.22	0.03	-7.78	0.00
dispneia	1.17	0.03	38.71	0.00
coriza	-0.60	0.07	-8.71	0.00
odinofagia	-0.39	0.08	-5.03	0.00
cefaleia	-0.50	0.06	-8.96	0.00
dor_no_peito	-0.22	0.07	-3.21	0.00
artralgia	0.22	0.10	2.17	0.03
mialgias	-0.53	0.04	-12.22	0.00
nauseas_vomitos	-0.14	0.06	-2.37	0.02
irritabilidade_confusao	0.34	0.08	4.08	0.00
fraqueza_geral_ou_astenia	0.12	0.03	3.63	0.00
auscultacao_pulmonar_anomala	-0.11	0.06	-1.79	0.07
radiografia_pulmonar_com_alteracoes	0.62	0.04	15.53	0.00
coma	1.35	0.23	5.75	0.00
taquicardia	0.52	0.09	5.62	0.00
AIC		59233.26		

Tabela I.20: Modelo final GAM ajustado à base de dados balanceada dos sintomas.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.67	0.61	-14.16	0.00
idade_utente_a_data_validacao	0.11	0.00	63.00	0.00
sexo_utenteM	0.73	0.05	14.77	0.00
descricao_distrito_ocorrenciaAveiro	0.31	0.60	0.51	0.61
descricao_distrito_ocorrenciaBeja	0.45	0.65	0.69	0.49
descricao_distrito_ocorrenciaBraga	0.12	0.60	0.21	0.84
descricao_distrito_ocorrenciaBragança	0.66	0.62	1.07	0.28
descricao_distrito_ocorrenciaCastelo_Branco	0.07	0.62	0.11	0.92
descricao_distrito_ocorrenciaCoimbra	0.31	0.61	0.52	0.61
descricao_distrito_ocorrenciaÉvora	0.73	0.63	1.15	0.25
descricao_distrito_ocorrenciaFaro	0.35	0.61	0.58	0.56
descricao_distrito_ocorrenciaGuarda	0.31	0.64	0.49	0.63
descricao_distrito_ocorrenciaLeiria	0.65	0.61	1.07	0.29
descricao_distrito_ocorrenciaLisboa	0.55	0.60	0.92	0.36
descricao_distrito_ocorrenciaMadeira	-0.28	0.78	-0.36	0.72
descricao_distrito_ocorrenciaPortalegre	0.00	0.66	0.00	1.00
descricao_distrito_ocorrenciaPorto	0.40	0.60	0.67	0.50
descricao_distrito_ocorrenciaSantarém	0.70	0.61	1.14	0.25
descricao_distrito_ocorrenciaSetúbal	0.54	0.60	0.91	0.36
descricao_distrito_ocorrenciaViana_do_Castelo	0.09	0.61	0.15	0.88
descricao_distrito_ocorrenciaVila_Real	-0.04	0.62	-0.06	0.95
descricao_distrito_ocorrenciaViseu	0.47	0.61	0.77	0.44
s(*data_confirmado1*, df = 3)	-0.00	0.00	-0.17	0.87
historia_de_febre_ou_calafrios	0.56	0.06	8.72	0.00
pneumonia	0.50	0.17	2.96	0.00
tosse_seca_ou_produtiva	-0.25	0.06	-4.12	0.00
dispneia	1.31	0.08	17.24	0.00
coriza	-0.77	0.12	-6.66	0.00
odinofagia	-0.46	0.12	-3.79	0.00
cefaleia	-0.46	0.10	-4.85	0.00
dor_no_peito	-0.41	0.14	-2.87	0.00
artralgia	0.45	0.21	2.15	0.03
mialgias	-0.61	0.08	-7.51	0.00
nauseas_vomitos	-0.01	0.12	-0.05	0.96
irritabilidade_confusao	0.52	0.28	1.82	0.07
fraqueza_geral_ou_astenia	0.20	0.08	2.59	0.01
auscultacao_pulmonar_anomala	-0.11	0.18	-0.64	0.52
radiografia_pulmonar_com_alteracoes	0.88	0.11	7.84	0.00
coma	0.78	0.74	1.06	0.29
taquicardia	0.94	0.27	3.44	0.00
AIC		11449		

Tabela I.21: Modelo final GAM ajustado à base de dados restrita dos sintomas.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.92	0.30	-36.66	0.00
idade_utente_a_data_validacao	0.10	0.00	107.37	0.00
sexo_utenteM	0.70	0.02	30.08	0.00
descricao_distrito_ocorrenciaAveiro	-0.25	0.29	-0.87	0.38
descricao_distrito_ocorrenciaBeja	0.06	0.31	0.20	0.84
descricao_distrito_ocorrenciaBraga	-0.45	0.29	-1.55	0.12
descricao_distrito_ocorrenciaBragança	-0.21	0.29	-0.73	0.47
descricao_distrito_ocorrenciaCastelo_Branco	-0.52	0.30	-1.73	0.08
descricao_distrito_ocorrenciaCoimbra	-0.09	0.29	-0.32	0.75
descricao_distrito_ocorrenciaÉvora	0.18	0.30	0.60	0.55
descricao_distrito_ocorrenciaFaro	-0.20	0.29	-0.69	0.49
descricao_distrito_ocorrenciaGuarda	-0.24	0.30	-0.81	0.42
descricao_distrito_ocorrenciaLeiria	0.12	0.29	0.41	0.68
descricao_distrito_ocorrenciaLisboa	-0.04	0.29	-0.14	0.89
descricao_distrito_ocorrenciaMadeira	-1.15	0.41	-2.84	0.00
descricao_distrito_ocorrenciaPortalegre	-0.25	0.32	-0.78	0.43
descricao_distrito_ocorrenciaPorto	-0.14	0.29	-0.50	0.61
descricao_distrito_ocorrenciaSantarém	-0.03	0.29	-0.10	0.92
descricao_distrito_ocorrenciaSetúbal	0.01	0.29	0.05	0.96
descricao_distrito_ocorrenciaViana_do_Castelo	-0.41	0.29	-1.40	0.16
descricao_distrito_ocorrenciaVila_Real	-0.56	0.30	-1.87	0.06
descricao_distrito_ocorrenciaViseu	-0.35	0.29	-1.19	0.24
s(*data_confirmado1*, df = 3)	0.00	0.00	1.64	0.10
historia_de_febre_ou_calafrios	0.42	0.03	14.46	0.00
pneumonia	0.20	0.06	3.52	0.00
tosse_seca_ou_produtiva	-0.22	0.03	-7.83	0.00
dispneia	1.17	0.03	38.67	0.00
coriza	-0.60	0.07	-8.72	0.00
odinofagia	-0.38	0.08	-5.03	0.00
cefaleia	-0.50	0.06	-8.95	0.00
dor_no_peito	-0.22	0.07	-3.21	0.00
artralgia	0.22	0.10	2.17	0.03
mialgias	-0.53	0.04	-12.22	0.00
nauseas_vomitos	-0.14	0.06	-2.36	0.02
irritabilidade_confusao	0.33	0.08	3.92	0.00
fraqueza_geral_ou_astenia	0.12	0.03	3.62	0.00
radiografia_pulmonar_com_alteracoes	0.61	0.04	15.43	0.00
coma	1.33	0.23	5.69	0.00
taquicardia	0.49	0.09	5.43	0.00
AIC		59234.23		

Tabela I.22: *OR* variáveis explicativas dos sintomas Modelo GAM.

		Intervalo Confiança	
Variável	***Odds-Ratio***	**2.5%**	**97.5%**
idade_utente_a_data_validacao	1.11	1.11	1.11
sexo_utenteM	2.01	1.92	2.10
descricao_distrito_ocorrenciaAveiro	0.78	0.44	1.37
descricao_distrito_ocorrenciaBeja	1.06	0.58	1.96
descricao_distrito_ocorrenciaBraga	0.64	0.36	1.13
descricao_distrito_ocorrenciaBragança	0.81	0.45	1.44
descricao_distrito_ocorrenciaCastelo_Branco	0.60	0.33	1.07
descricao_distrito_ocorrenciaCoimbra	0.91	0.52	1.61
descricao_distrito_ocorrenciaÉvora	1.20	0.66	2.16
descricao_distrito_ocorrenciaFaro	0.82	0.46	1.45
descricao_distrito_ocorrenciaGuarda	0.78	0.44	1.41
descricao_distrito_ocorrenciaLeiria	1.13	0.64	2.00
descricao_distrito_ocorrenciaLisboa	0.96	0.55	1.68
descricao_distrito_ocorrenciaMadeira	0.32	0.14	0.70
descricao_distrito_ocorrenciaPortalegre	0.78	0.41	1.47
descricao_distrito_ocorrenciaPorto	0.87	0.49	1.52
descricao_distrito_ocorrenciaSantarém	0.97	0.55	1.72
descricao_distrito_ocorrenciaSetúbal	1.01	0.58	1.78
descricao_distrito_ocorrenciaViana_do_Castelo	0.66	0.37	1.18
descricao_distrito_ocorrenciaVila_Real	0.57	0.32	1.03
descricao_distrito_ocorrenciaViseu	0.71	0.40	1.25
historia_de_febre_ou_calafrios1	1.52	1.43	1.60
pneumonia1	1.22	1.09	1.36
tosse_seca_ou_produtiva1	0.80	0.76	0.85
dispneia1	3.21	3.03	3.41
coriza1	0.55	0.48	0.63
odinofagia1	0.68	0.59	0.79
cefaleia1	0.61	0.55	0.68
dor_no_peito1	0.80	0.70	0.92
artralgia1	1.25	1.02	1.53
mialgias1	0.59	0.54	0.64
nauseas_vomitos1	0.87	0.78	0.98
irritabilidade_confusao1	1.39	1.18	1.64
fraqueza_geral_ou_astenia1	1.13	1.06	1.21
radiografia_pulmonar_com_alteracoes1	1.84	1.71	1.99
coma1	3.79	2.40	6.00
taquicardia1	1.64	1.37	1.96

I.2.5 Árvore de Classificação-Dados Balanceados

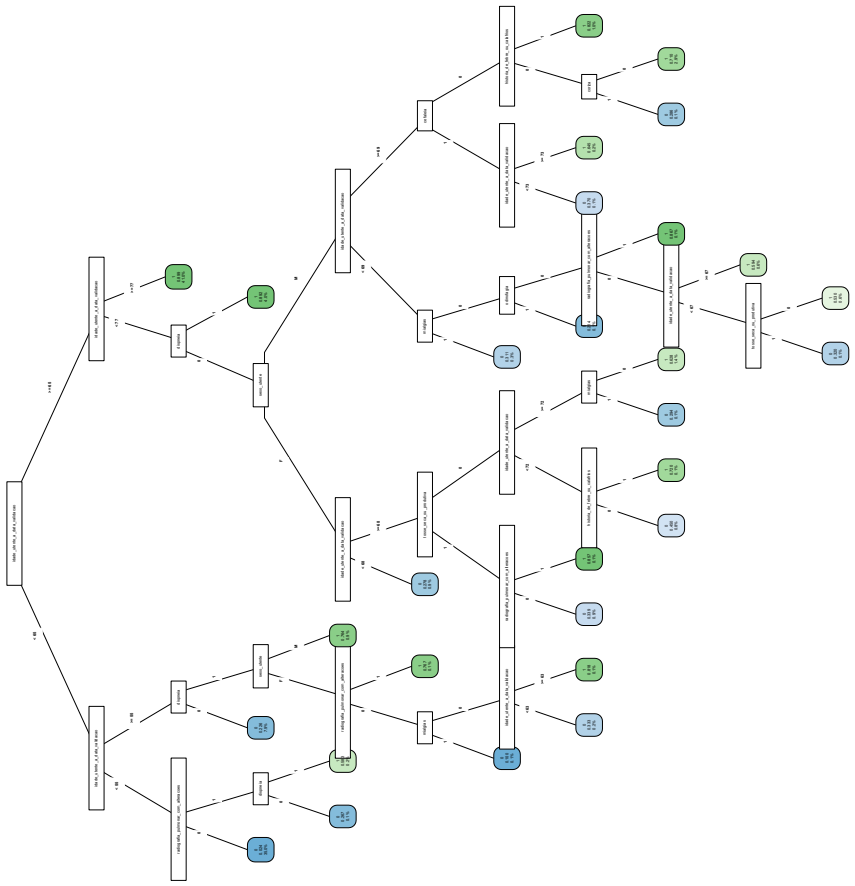


Figura I.2: Árvore de Decisão Final.

I.2.5.1 Descrição dos resultados da Árvore de Classificação

Através da observação da Árvore de Classificação representada na Figura I.2, pode-se observar que a variável explicativa *idade_utente_a_data_validacao* é das variáveis explicativas mais vezes utilizadas para o processo da divisão binária de um nó, pelo que, se pode afirmar que esta é uma variável bastante importante para explicar o falecimento de um indivíduo que tenha testado positivo à COVID-19, como se virá de seguida. Através da análise da Figura I.2, pode-se observar que:

- Se um indivíduo tiver mais de 77 anos, então a Árvore de Classificação, indica-nos que estes acabam por falecer;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos e se apresentar o sintoma dispneia, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino e tem menos de 68 anos, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino, se tiver 68 ou mais anos, apresenta o sintoma tosse seca ou produtiva e não apresente o sintoma radiografia pulmonar com alterações, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino, se tiver 68 ou mais anos, apresenta o sintoma tosse seca ou produtiva e apresente o sintoma radiografia pulmonar com alterações, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino, se tiver 68 ou mais anos, não apresenta o sintoma tosse seca ou produtiva, se tiver menos de 72 anos e não apresente o sintoma historia de febre ou calafrios, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino, se tiver 68 ou mais anos, não apresenta o sintoma tosse seca ou produtiva, se tiver menos de 72 anos e apresente o sintoma historia de febre ou calafrios, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino, se tiver 68 ou mais anos, não apresenta o sintoma tosse seca ou produtiva, se tiver 72 ou mais anos e não apresente o sintoma mialgias, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é feminino, se tiver 68 ou mais anos, não apresenta o sintoma tosse seca ou produtiva, se tiver 72 ou mais anos e apresente o sintoma mialgias, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver menos de 69 e apresentar o sintoma mialgias, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver menos de 69, se não apresentar o sintoma mialgias e se apresentar o sintoma odinofagia, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver menos de 69, se não apresentar o sintoma mialgias, se não apresentar o sintoma odinofagia e apresentar o sintoma radiografia pulmonar com alterações, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver menos de 69, se não apresentar o sintoma mialgias, se não apresentar o sintoma odinofagia, se não apresentar o sintoma radiografia pulmonar com alterações e se tiver 67 ou mais anos, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver menos de 69, se não apresentar o sintoma mialgias, se não apresentar o sintoma odinofagia, se não apresentar o sintoma radiografia pulmonar com alterações, se tiver menos 67 e se não apresentar o sintoma tosse seca ou produtiva, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver menos de 69, se não apresentar o sintoma mialgias, se não apresentar o sintoma odinofagia, se não apresentar o sintoma radiografia pulmonar com alterações, se tiver menos 67 e se apresentar o sintoma tosse seca ou produtiva, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver 69 ou mais anos, se apresentar o sintoma cefaleia e se tiver menos de 73 anos, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver 69 ou mais anos, se apresentar o sintoma cefaleia e se tiver 73 ou mais anos, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver 69 ou mais anos, se não apresentar o sintoma cefaleia e se apresentar o sintoma historia de febre ou calafrios, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver 69 ou mais anos, se não apresentar o sintoma cefaleia, se não apresentar o sintoma historia de febre ou calafrios e se apresentar o sintoma coriza, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver 65 anos ou mais e menos de 77 anos, se não apresentar o sintoma dispneia, se o seu sexo é masculino, se tiver 69 ou mais anos, se não apresentar o sintoma cefaleia, se não apresentar o sintoma historia de febre ou calafrios e se apresentar o sintoma coriza, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos de 65 anos, menos de 55 anos e não apresentar o sintoma radiografia pulmonar com alterações, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos de 65 anos, menos de 55 anos, se apresentar o sintoma radiografia pulmonar com alterações e se não apresentar o sintoma dispneia, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos de 65 anos, menos de 55 anos, se apresentar o sintoma radiografia pulmonar com alterações e se apresentar o sintoma dispneia, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos de 65 anos, se tiver 55 ou mais anos anos e se não apresentar o sintoma dispneia, então a Árvore de Classificação indica-nos prognóstico sobreviveu;

- Se um indivíduo tiver menos de 65 anos, se tiver 55 ou mais anos anos, se apresentar o sintoma dispneia e se o seu sexo é o masculino, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos de 65 anos, se tiver 55 ou mais anos anos, se apresentar o sintoma dispneia, se o seu sexo é o feminino e se apresentar o sintoma radiografia

pulmonar com alterações, então a Árvore de Classificação indica-nos prognóstico morte;

- Se um indivíduo tiver menos de 65 anos, se tiver 55 ou mais anos anos, se apresentar o sintoma dispneia, se o seu sexo é o feminino, se não apresentar o sintoma radiografia pulmonar com alterações e se apresentar o sintoma mialgias, então a Árvore de Classificação indica-nos prognóstico sobreviveu;
- Se um indivíduo tiver menos de 65 anos, se tiver 55 ou mais anos anos, se apresentar o sintoma dispneia, se o seu sexo é o feminino, se não apresentar o sintoma radiografia pulmonar com alterações, se não apresentar o sintoma mialgias e se a sua idade for de 63 ou mais anos, então a Árvore de Classificação indica-nos prognóstico morte;
- Se um indivíduo tiver menos de 65 anos, se tiver 55 ou mais anos anos, se apresentar o sintoma dispneia, se o seu sexo é o feminino, se não apresentar o sintoma radiografia pulmonar com alterações, se não apresentar o sintoma mialgias e se a sua idade for de inferior a 63, então a Árvore de Classificação indica-nos prognóstico sobreviveu;