NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

ANÁLISE PREDITIVA NO AGENDAMENTO DE CONSULTAS MÉDICAS

O FENÓMENO NO-SHOW

Cláudia Filipa de Oliveira Pereira

Trabalho de Projeto apresentada(o) como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ANÁLISE PREDITIVA NO AGENDAMENTO DE CONSULTAS MÉDICAS

O FENÓMENO NO-SHOW

por

Cláudia Filipa de Oliveira Pereira

Trabalho de Projeto apresentado como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Business Intelligence e Gestão de Conhecimento

Orientador: Professor Doutor Roberto Henriques

Coorientador: Mestre Hugo Miguel Ferrão Casal da Veiga

Novembro 2020

DECLARAÇÃO DE ORIGINALIDADE

Declaro que o trabalho descrito neste documento é da minha autoria e não de outra pessoa. Toda a assistência que recebi de outras pessoas é devidamente reconhecida e todas as fontes (publicadas ou não) são referenciadas.

Este trabalho não foi avaliado ou enviado anteriormente à NOVA Information Management School ou a qualquer outro lugar.

Lisboa, 30 de Novembro de 2020

Cláudia Filipa de Oliveira Pereira

AGRADECIMENTOS

Foi longo o período de elaboração deste trabalho e foram muitas as vezes que pensei que terminar não seria possível ou não estava ao meu alcance. Foi nessas alturas que o suporte de amigos, família e professores foi importante para continuar por isso não poderia terminar esta etapa sem começar por agradecer a todos os que de alguma forma me motivaram e encorajaram a não desistir.

Ao meu orientador e co-orientador obrigada por me guiarem e apoiarem no desenvolvimento deste trabalho.

Aos meus amigos e família, em especial à Soraia e ao Pedro, obrigada pela força, ajuda e motivação. Por não me deixarem desistir e desmascararem as minhas desculpas.

Aos professores que me acompanharam no percurso na Nova IMS, em especial à Maria, obrigada por me mostrarem que a educação tem o valor que nós lhe damos e por me desafiarem a conseguir mais e melhor.

Por último, quero agradecer à minha avó Laura que mesmo sem saber ler ou escrever sempre me incentivou a ser melhor, a lutar pelos meus objetivos e me formou enquanto pessoa.

RESUMO

Os não comparecimentos às consultas agendadas (no-shows) são um fenómeno cada vez problemático no sector da saúde tendo implicações nos custos, planeamento e prestação atempada de cuidados de saúde. Torna-se imperativo aos estabelecimentos de prestação de cuidados de saúde conseguir prever este fenómeno.

Neste trabalho foram aplicados métodos de aprendizagem automática por forma a conseguir estudar fatores que influenciam os no-shows e identificar o modelo com melhor desempenho neste tipo de problemas.

Foram aplicados modelos de regressão logística, árvores de decisão, random forests e métodos ensemble a uma base de dados de consultas médicas do serviço público de saúde da cidade de Vitória no estado de Espírito Santo no Brasil entre abril de 2016 e junho de 2016.

Observou-se uma percentagem maior de no-shows quando o doente não apresenta nenhum condicionamento à sua saúde (Ex: diabetes, handicap, etc.), que as mulheres apresentam 65% de probabilidade de não comparecer as consultas face aos 35% dos homens.

Os algoritmos aplicados produziram resultados entre os 71% e 78% de AUC, tendo sido as random forests o modelo que mostrou uma capacidade discriminativa superior com 78,1% de AUC e uma curva de ROC superior às curvas dos restantes modelos. Este modelo identifica corretamente cerca de 80% das vezes os no-shows, para cerca de 17% da totalidade dos mesmos identificados.

PALAVRAS-CHAVE

No-Shows; Data Mining; Modelos Preditivos;

ABSTRACT

No shows are an increasing problem in the health sector having implications on the costs, planning and the timely delivery and effectiveness of medical care. Being able to predict such phenomenon is imperative for medical clinics.

In the present work, machine learning methods were applied in order to study factors that affect no shows and to identify the best performing model for this kind of problem subject.

Logistic regression, decision trees, random forests and ensemble methods were applied to a medical appointments database from public sector of Vitória, a city of the Espírito Santo state of Brazil between April and June of 2016.

A higher percentage of no shows were observed when patients had no health conditions (e.g. diabetes, handicaps, etc.) and women were more prone to fail appointments than men (65% and 35% respectively).

The applied algorithms presented an AUC between 71% and 78%, with random forests being the model with the higher discriminative power with 78,1% of AUC and a superior ROC curve in relation with the remaining algorithms. This model also identifies accurately around 80% no shows and identifies correctly around 17% of the total of no shows.

KEYWORDS

No-Shows; Data Mining; Predictive Modelling;

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

DM	Data *Mining*
ML	*Machine Learning*
No-Show	Não comparecimento às consultas médicas
Show-Up	Comparecimento às consultas médicas
ROC	*Receiver Operating Characteristic*
AUC	*Area under the curve*

1. INTRODUÇÃO

1.1. ENQUADRAMENTO

O sector da saúde está em constante mudança por forma a encontrar alternativas para reduzir custos, ser mais competitivo e melhorar o serviço de prestação de cuidados. O *data mining* permite às organizações de saúde a utilização das mais recentes tecnologias de modo a criar novas formas de conhecimento para tomadas de decisão mais conscientes por forma a atingirem os seus objetivos de negócio.

Um dos grandes objetivos das organizações de saúde atualmente é a redução do fenómeno *no-show* que se traduz num problema para os serviços dos mais variados países ao acarretar custos devido a dificuldades operacionais e de agendamento bem como redução da produtividade.

Um *no-show*, ou não comparecimento, é um fenómeno que acontece quando um paciente falha a sua consulta médica sem aviso prévio ou cancelamento da mesma. O custo relativo à média de não comparecimentos numa clínica norte americana, ronda os 720 dólares diários e, anualmente, custa mais de 100 mil milhões de dólares a clínicas detidas por grupos (Rawl, 2018).

Pacientes com doenças crónicas tendem a falhar mais às consultas médicas devido aos desafios associados às suas condições de saúde. Este grupo de pacientes é também um dos que mais perde com o não comparecimento devido aos riscos implícitos às suas doenças. Os pacientes com seguro médico são outro grupo com taxas de não comparecimento acima da média. Tal pode ser explicado por razões socioeconómicas como necessidade de usar transportes ou residir em zonas rurais longe da localização do consultório médico ou clínica (Crosschx blog, 2017).

Em 2015, um estudo da *Medical Group Management Association* revelou que até clínicas bem geridas apresentam uma média diária de 12% de não comparecimentos e cancelamentos em cima da hora. O estudo revelou, ainda, que algumas clínicas apresentam uma taxa gritante de 50%.

Dantas *et al.* (2016) fez a revisão sistemática da literatura sobre o fenómeno do *no-show* em agendamento de consultas e identificou de entre as conclusões de um total de 724 trabalhos, que a maior taxa de não comparecimento às consultas agendadas relativamente aos diferentes continentes, foi observada no continente africano (43%) enquanto que a menor foi observada na Oceânia (13,2%).

Relativamente ao tipo de clínicas, Dantas identificou que as menores taxas de não comparecimento foram observadas em clínicas de exames e clínicas pediátricas (17,4% e 18,8%, respetivamente) enquanto as clínicas de fisioterapia e de cardiologia contam com as maiores taxas observadas (50,8% e 30,4%, respetivamente).

1.2. MOTIVAÇÃO

O conhecimento dos fatores que levam os pacientes a faltar às consultas agendadas pode auxiliar todas as partes integrantes dos sistemas de saúde, desde administradores de áreas de saúde, provedores a investigadores, a identificar medidas que possam mitigar este fenómeno. Tal poderá levar a melhorias na produtividade, práticas de gestão, técnicas de overbooking e políticas de agendamento.

1.3. OBJETIVOS

Os objetivos deste trabalho são:

- Encontrar padrões nos dados que permitam compreender quais os fatores mais importantes para o efeito *no-show*;
- Analisar os modelos implementados por forma a identificar o modelo com melhor desempenho neste tipo de problemas.

Pretende-se realçar a possibilidade de criação de valor para as clínicas através da obtenção de conhecimento sobre os fatores que levam os seus pacientes a não comparecer às consultas marcadas. Com este conhecimento, as clínicas podem adotar técnicas de *overbooking* que permitam tirar o melhor partido dos seus prestadores de cuidados em horário de trabalho ou adaptar os horários dos prestadores de cuidados conforme as horas/dias da semana com mais probabilidades de não comparecimentos.

1.4. ESTRUTURA

O presente trabalho está organizado em 6 secções de informação. No bloco 1, foi feita uma introdução ao problema em análise e apresentados os objetivos propostos e motivos para a escolha da temática.

O bloco 2 pretende apresentar uma revisão do estado da arte relativo ao tema do problema bem como apresentar e discutir os conceitos subjacentes às técnicas referenciadas ao longo do presente trabalho.

No bloco 3 será exposta a metodologia adotada em todas as fases do trabalho, no bloco 4 serão apresentados e discutidos os resultados obtidos e no bloco 5 serão apresentadas as conclusões finais do trabalho.

Por último, o bloco 6 contém as limitações ocorridas ao longo do projeto e possíveis recomendações para trabalhos futuros.

2. REVISÃO DA LITERATURA

1.5. ESTADO DA ARTE

Neal et al. (2001) analisaram informação recolhida em 4 clínicas inglesas com o objetivo de perceber, através de análises estatísticas, se o género, idade e zonas demográficas desfavorecidas diferem entre os doentes que faltaram a consultas bem como analisar diferenças entre as clínicas no que toca a não comparecimentos às consultas. O estudo concluiu que jovens adultos, que vivam em zonas desfavorecidas têm uma maior probabilidade de faltar a consultas. No entanto, apesar de haver uma inclinação para o género ser um fator contributivo para a perda de consultas, os resultados não foram conclusivos entre as 4 clínicas.

Em 2004 Husain et al. inquiriu os profissionais de saúde de diversas clínicas sobre as consultas que geraram não comparecimentos. Os mesmos culpabilizavam os doentes pela disrupção dos processos de trabalho das receções e o aumento de problemas na gestão de doentes.

Onze anos mais tarde, Alaeddini (2015) desenvolveu um modelo probabilístico híbrido baseado numa regressão logística multinominal e inferência Bayesiana para prever com precisão a probabilidade de cancelamentos e não comparecimentos em tempo real. Este estudo teve em conta o comportamento individual de cada doente enriquecido de informação da base de dados dos doentes para obter estimativas probabilísticas reais. Alaeddini (2015) chegou à conclusão que a eficácia de qualquer sistema primário de agendamentos depende da própria habilidade para prever e gerir diferentes tipos de disrupções e incertezas.

Y. Huang (2014) propôs uma abordagem para overbooking utilizando a probabilidade de no-show dos pacientes, baseando-se em informação histórica e nas características de cada indivíduo. O objetivo do seu estudo era determinar se seria possível prever com uma eficácia razoável quando um paciente não iria comparecer. Os resultados mostraram que a abordagem tomada trazia vantagens em relação ao sistema standard de overbooking em termos de eficácia na redução de custos e tempos de espera menores para os pacientes.

Kurasawa (2016) criou um modelo que prevê com elevada precisão a probabilidade de um doente diabético faltar a uma consulta, podendo resultar na descontinuação do tratamento. Os autores utilizaram a regressão logística com regularização L2 (L2-*norm regularization*), para evitar o problema de *overfitting*, e a validação cruzada (10-fold *cross validation*).

No entanto, Dantas (2016) concluiu, através de uma revisão sistemática da literatura sobre o tema do no-show no agendamento de consultas, que este fenómeno é mais comum entre pacientes jovens. Adicionalmente, Dantas concluiu sobre quais os fatores que afetam significativamente o fenómeno no-show:

- Tempo de espera entre data de agendamento e a data de consulta (lead time);
- Histórico de agendamento perdido - falta a consultas prévias;
- Baixo nível sócio-económico;
- Distância entre a residência e o local da consulta;

- Inexistência de seguro de saúde e beneficiário do sistema nacional de saúde;
- Atendimento por médicos com menos experiência;
- Pacientes com diagnóstico psiquiátrico que necessitam de medicamentos;
- Pacientes que usam tabaco, drogas e/ou álcool.

Os fatores identificados como os mais fortes e que carecem de mais atenção foram o lead time e o historial de no-show de cada paciente.

Em 2017, Devasahay et al. usaram *decision trees* e regressão logística para prever no-shows e o tipo de doentes que não compareciam às consultas num hospital de Singapura. Os resultados foram inconclusivos a partir do conjunto de dados utilizados pelo que os melhores resultados foram encontrados quando aplicado um *cutoff* às arvores de decisão.

Adicionalmente, este estudo identificou que os homens apresentam, em média, uma taxa de não comparecimentos superiores em 2% em relação às mulheres. E que, ao contrário do que tinha sido identificado por Dantas em 2016 e contrariamente às pressuposições do hospital em estudo, não foi encontrada nenhuma tendência quanto à contribuição dos pacientes jovens para o fenómeno no-show. Devasahay et al. concluíram ainda que doentes que usufruem de algum tipo de subsídio ou comparticipação para a saúde tendem a faltar mais às consultas do que doentes não subsidiados e que a hora do dia tem um efeito na taxa de consultas perdidas. Tipicamente das 8h às 9h da manhã e das 12h às 14h da tarde há uma diminuição significativa da taxa de consultas perdidas.

De uma forma geral, a maioria dos estudos apresentados nesta secção mostraram-se inconclusivos em relação ao modelo que melhor prevê o fenómeno no-show. O trabalho aqui proposto, pretende identificar o modelo com melhor desempenho neste tipo de problemas e, em adição aos estudos apresentados nesta secção, pretende analisar de que forma os dias da semana e hora das consultas podem contribuir para explicar o fenómeno de não comparecimentos às consultas médicas.

1.6. DATA MINING

Data Mining (DM) pode ser definido como um processo de descoberta de padrões e tendências em conjuntos de dados por forma a explicar e inferir sobre os mesmos (Witten, Frank, & Hall, 2011). Tendo em consideração que os dados crescem exponencialmente com o passar do tempo, torna-se crucial a criação de técnicas para análise de tamanha informação. Deste modo, as ferramentas de DM permitem prever tendências e acontecimentos através da identificação de padrões nas bases de dados (Silltow, 2006).

O DM é um campo interdisciplinar da ciência da computação que combina várias ferramentas, entre as quais, aprendizagem automática (Hand, Mannila, & Smyth, 2001). Muitas das técnicas automáticas usadas em DM têm origem na aprendizagem automática, no entanto tal não pode ser visto apenas como um subconjunto do *data mining* (Kononenko & Matjaz, 2007).

A aprendizagem automática (doravante denominada ML, do inglês *Machine Learning*) pode ter diversas definições conforme a sua área de aplicação. De um modo geral, pode ser descrita como um campo focado no desenvolvimento de teorias computacionais da aprendizagem e construção de sistemas de aprendizagem (Michalski, Carbonell, & Mitchell, 1986). Um dos objetivos do ML é o

estudo, construção e melhoria de modelos matemáticos que possam ser treinados com dados reais para inferir o futuro e tomar decisões sem o conhecimento total de todos os fatores possíveis (Bonaccorso, 2017).

2.1.1. *Data Mining* na saúde

Atualmente, as técnicas de DM são amplamente estudadas e aplicadas a diversas áreas como educação, banca e seguros (deteção de fraude), marketing (*customer relationship management*), retalho (análise carrinho de compras) e saúde. Nesta última área, o potencial é enorme uma vez que os algoritmos de DM podem facilitar aos sistemas de saúde a identificação de ineficiências e a aplicação de melhores práticas que reduzam os custos e melhorem o nível de cuidados prestados (Archer blog, 2020).

Nos tempos que correm, o *data mining* na saúde é largamente utilizado na previsão de doenças, diagnóstico e auxílio da tomada de decisão por parte dos médicos e técnicos de saúde (Jothi et al., 2015). Um exemplo da aplicabilidade do *data mining* neste sector é a utilização de algoritmos de *deep learning* no processamento de imagens de exames de diagnóstico (radiografias, TACs, ressonâncias, etc.) para auxiliar no diagnóstico e encontrar o tratamento mais eficaz (Dutta, 2020).

2.1.2. Aprendizagem Supervisionada vs Não Supervisionada

Os algoritmos de aprendizagem automática podem ser primeiramente divididos em aprendizagem supervisionada e aprendizagem não supervisionada. A aprendizagem supervisionada baseia-se em ter conhecimento *à priori* do resultado a que os dados devem chegar. O objetivo desta aprendizagem é aprender, a partir de um conjunto de dados de input e a partir do output desejado, uma função que se aproxime o melhor que consiga da relação entre os dados estudados. A principal desvantagem deste tipo de aprendizagem é o facto do *dataset* ter de ser classificado manualmente previamente o que se torna frequentemente um processo exigente.

Por forma a fazer face a esta desvantagem, a aprendizagem semi-supervisionada foi introduzida. Neste tipo de aprendizagem existe uma combinação entre um conjunto pequeno de dados classificados e um conjunto maior de dados não classificados. O segundo conjunto vai servir para agrupar dados similares usando aprendizagem não supervisionada e usar posteriormente o conjunto de dados classificados para os classificar. Uma das aplicações deste tipo de abordagem é a classificação de sequências de ADN dada a grande dimensão deste tipo de dados. A aprendizagem não supervisionada tem como objetivo descobrir padrões num conjunto de dados e agrupá-los partir de semelhanças entre si e reduzir o número de atributos. Por outro lado, a aprendizagem por reforço é uma área da aprendizagem automática preocupada com a forma como os programas (“agentes”) interagem com o ambiente e tomam uma sequência de decisões. Neste tipo de aprendizagem, os agentes utilizam uma abordagem de tentativa e erro para encontrar solução para um problema enquanto são atribuídas recompensas ou penalidades pelas ações que executam. Apesar das recompensas e penalidades serem definidas pelo utilizador, é a máquina que tem de descobrir como executar as tarefas para alcançar a recompensa. Trata-se de tomar ações adequadas para maximizar a recompensa em determinados cenários e aprender através da experiência. Este tipo de aprendizagem é muito comum em áreas de robótica e jogos.

Problemas de classificação e regressão são tarefas comumente suportadas por aprendizagem supervisionada (Kantardzic, 2011). O presente trabalho consiste num problema de classificação e sendo um dos objetivos, prever o não comparecimento em consultas médicas, podemos dizer que será usada uma aprendizagem supervisionada.

2.1.3. Modelos de Aprendizagem Supervisionada

Árvores de Decisão

As árvores de decisão são um dos algoritmos de classificação mais populares usado em ML por forma a criar estruturas de conhecimento que guiem o processo de tomada de decisão (Abdelhalim & Traore, 2009). MI

As árvores de decisão classificam instâncias através de uma divisão baseada nos valores dos atributos. Uma árvore de decisão é composta por nós, arcos e folhas. Cada nó numa árvore de decisão representa um atributo numa instância a ser classificada e cada ramo representa o valor que cada nó pode assumir. As instâncias são classificadas a partir da raiz da árvore - o primeiro nó – e divididas pelos valores dos seus atributos até chegar à folha (Kotsiantis, S. 2007).

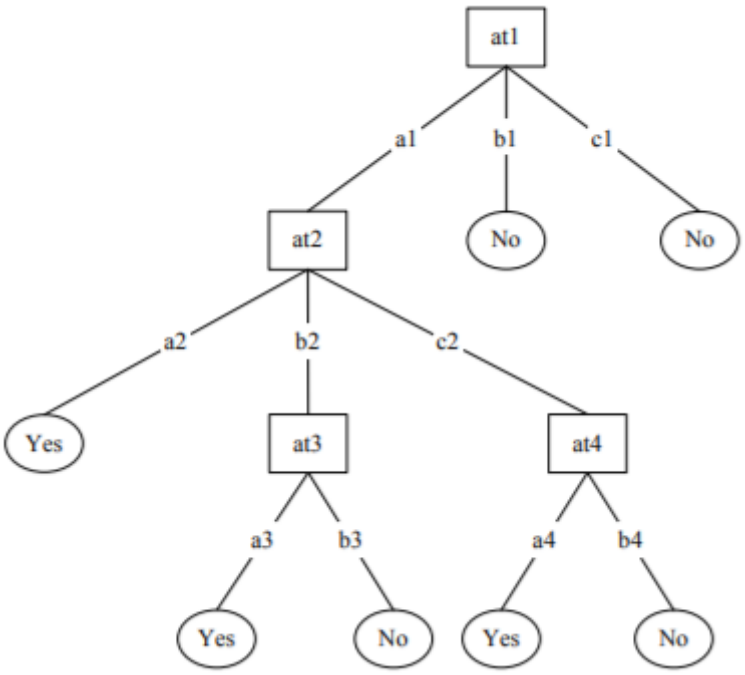

Figura 1 - Estrutura de uma árvore de decisão (Fonte: Kotsiantis, S. 2007)

Pode dizer-se que as árvores de decisão utilizam uma estratégia de "dividir para conquistar" (Kotsiantis, S. 2007) decompondo o problema formulado no nó inicial em problemas cada vez mais simples nos nós seguintes até encontrar a folha na qual é atribuída uma classe.

De uma forma geral, Mitchell (1997) definiu as árvores de decisão como "disjunções de conjunções de restrições" uma vez que cada caminho entre a raiz e a folha corresponde a uma conjunção de testes a cada atributo e a árvore em si corresponde a uma disjunção destas conjunções. Todos os nós apresentam regras mutuamente exclusivas o que resulta em uma observação do conjunto de dados existe num nó apenas (Ville, 2006).

Os algoritmos de árvores de decisão mais utilizados são o CART e o C4.5. O algoritmo C4.5 – como sucessor do algoritmo ID3 proposto por Ross Quinlan (1993) – permite gerar árvores de decisão a partir de um conjunto de dados aplicando uma abordagem *top-down* ambiciosa ao testar cada atributo em todos os nós da árvore. É iniciado com todas as amostras de treino na raiz da árvore. Em cada iteração, vai passando por cada atributo não utilizado do conjunto de dados e calcula a entropia para cada um selecionando o atributo com valor menor de entropia. Seguidamente, o conjunto de dados é dividido pelo atributo escolhido criando subconjuntos de dados. O algoritmo continua a iterar em cada subconjunto considerando apenas atributos nunca selecionados.

Por sua vez, o algoritmo CART - Classification And Regression Trees – gera árvores binárias nas quais cada nó é dividido em dois nós seguintes. O coeficiente de Gini é aplicado por forma a selecionar a variável para o primeiro nó da árvore. Este algoritmo confere a capacidade de gerar árvores cujas folhas preveem um número em vez de uma classe, ou seja, gerar árvores de regressões (Maimon & Rokack, 2005).

Regressões lineares

Uma regressão linear é uma equação que descreve a linha que melhor se ajusta à relação entre as variáveis de *input* e *output* através do cálculo de pesos específicos (coeficientes) para as variáveis de entrada (Kononenko & Matjaz, 2007).

As regressões lineares são talvez um dos algoritmos mais conhecidos e compreendidos tanto em estatística como em ML o que pode ser benéfico em decisões de negócio. São fáceis de modelar e práticos quando as relações entre variáveis dependentes e independentes não são complexas ou se os conjuntos de dados para modelar não forem extensos.

Contudo, quando lidamos com conjuntos de dados mais complexos, as regressões lineares apresentam algumas desvantagens. Este tipo de algoritmo é bastante sensível a *outliers* podendo levar a resultados enganadores e a modelos com fraca capacidade de explicação. Adicionalmente, as regressões lineares são suscetíveis ao problema da sobreaprendizagem dos dados (*overfitting)*. O *overfitting* consiste no excessivo ajustamento da regressão ao conjunto de dados (Hand, Mannila, & Smyth, 2001). Em problemas nos quais existem demasiados parâmetros, por exemplo, a regressão tende a modelar o erro aleatório dos dados em vez da sua relação apenas.

Regressões logísticas

As regressões logísticas são regressões lineares generalizadas que usam a função *logit*. Estas distinguem-se pelo facto de permitirem a previsão de variáveis dependentes binárias (Kononenko & Matjaz, 2007). Para o presente trabalho irá ser aplicada uma regressão logística uma vez que se pretende prever se o doente comparece ou não às consultas médicas.

Métodos Ensemble

Métodos *ensemble* é uma técnica de ML que combina diversos algoritmos por forma a produzir um melhor modelo preditivo. Esta técnica procura a vantagem individual de cada algoritmo, combinando-a de forma a encontrar a melhor solução. Normalmente, um ensemble é construído em 2 passos, ou seja, criam-se os algoritmos individuais e depois combinam-se os mesmos (Zhou, 2012).

Dois dos métodos *ensemble* existentes e mais populares irão ser aplicados no presente trabalho: o método *voting* e *random forests*.

Majority Voting

Voting é um dos métodos de combinação mais populares para *outputs* nominais. Existem diversos algoritmos que usam este método, sendo o mais conhecido o *majority voting*, ou seja, votação por maioria (Zhou, 2012).

Neste processo, cada classificador vota numa classe e a classe final do output é aquela que recebe mais de metade dos votos. Caso nenhuma classe receba mais de vontade dos votos, o classificador *ensemble* não classifica o output (Zhou, 2012). Esta técnica foi aplicada no presente trabalho.

Random Forests

Random forests é um dos métodos ensemble mais conhecidos. Baseado no método *Bagging* é utilizado tanto em problemas de classificação como regressão. Este método consiste em produzir múltiplas árvores de decisão e combiná-las por forma a obter um classificador estável e preciso.

O método *random forests* permite que todas as variáveis sejam utilizadas em oposição às árvores de decisão que permitem usar apenas as mais relevantes (Finlay,2014).

2.1.4. Validação e escolha do melhor modelo

A seleção do melhor modelo pode ser uma tarefa complicada. Existem diversas medidas que podem ser utilizadas para avaliar o desempenho dos modelos tendo em conta as características dos algoritmos aplicados.

A medida mais comumente utlizada é a *accuracy* que traduz o rácio entre o número de previsões corretas e o total de previsões. Esta medida tem um melhor desempenho quando as classes a prever têm pesos distribuídos no conjunto de dados por ser uma medida arbitrária. Em dados com classes

não balanceadas, uma *accuracy* elevada pode significar que o modelo está apenas a prever corretamente uma das classes.

Deste modo, existem outras medidas para avaliar problemas de classificação não balanceados como a *Precision* e *Recall*, bem como a área debaixo da curva ROC (AUC – *area under the curve*).

Curva ROC e AUC

A curva ROC (*Receiver Operating Characteristics*) é a representação gráfica entre a taxa de verdadeiros positivos em função da taxa de falsos positivos. Um teste com uma classificação perfeita tem uma curva que passa o canto superior esquerdo de coordenadas (0,1) representado uma taxa nula de falsos positivos e 100% de verdadeiros positivos. A Figura 2 pretende ilustrar a curva ROC.

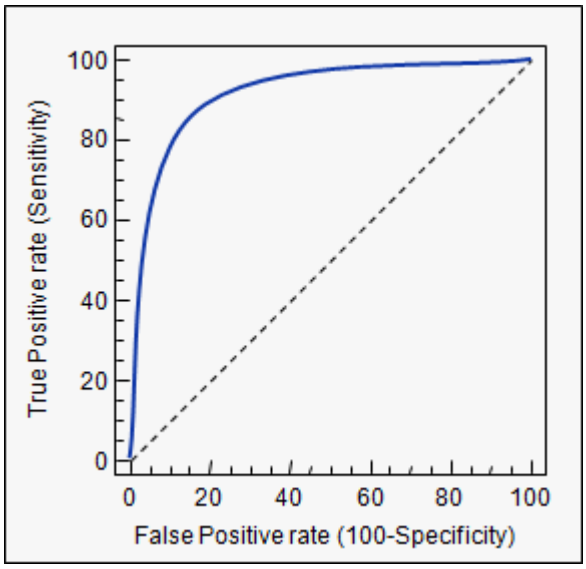

Figura 2 - Curva ROC
(Fonte: https://www.medcalc.org/manual/roc-curves.php)

A área abaixo da curva (AUC) mede a capacidade discriminativa do modelo, ou seja, quão bom é o modelo a prever a variável dependente. Desta forma, quanto maior for a área abaixo da curva ROC, melhor será o modelo.

Um AUC de valor 0 caracteriza um modelo cujas previsões estão 100% erradas enquanto um AUC de valor 1 caracteriza um modelo cujas previsões estão 100% corretas. Modelos cujo AUC seja 0,5 caem na linha diagonal do gráfico (representado acima na Figura 2) indicando que o modelo tem a capacidade discriminativa equivalente a virar uma moeda no ar, ou seja, o modelo não consegue diferenciar entre as duas classificações possíveis para a variável dependente. Usando o objetivo do presente trabalho, um AUC de 0,5 indicaria que o modelo seria aleatório quanto à classificação de *no-show* vs. *show-up*.

Recall e Precision

A *recall*, também conhecida como *sensitivy*, pode ser definida como a percentagem de resultados positivos corretamente identificados enquanto a *Precision* pode ser definida como a percentagem de resultados que é relevante.

Quando usadas para avaliar um algoritmo, há que ter em conta que existe um *trade-off* entre ambas as medidas. O aumento da *Recall* de um modelo tende a baixar a *Precision* do mesmo. Descomplicando os conceitos com um exemplo simples, podemos entender que para nos lembrarmos de todos as vezes que choveu nos últimos 6 meses (*Recall*), vamos de certeza errar algumas vezes e ter de usar mais tentativas para acertar corretamente em todos os eventos de chuva (*Precision*). Desta forma, não é possível maximizar ambas as medidas ao mesmo tempo.

As curvas *recall-precision* (RPC) ajudam a visualizar o *trade-off* entre as medidas para *thresholds* diferentes. Um bom modelo apresenta uma RPC mais chegada ao canto superior direito do gráfico possível (100% *recall* e 100% *precision*) como representado na Figura 3.

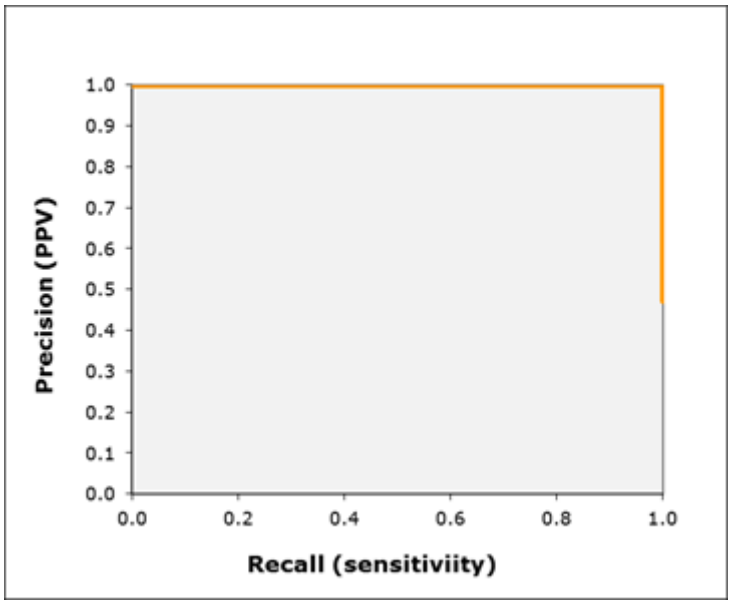

Figura 3 - Curva precision-recall para um teste perfeito
(Fonte: https://acutecaretesting.org/)

3. METODOLOGIA

Nos capítulos anteriores foi efetuado um enquadramento do tema em estudo. Os capítulos seguintes pretendem apresentar a organização do processo de investigação. No âmbito deste estudo, será seguida uma abordagem em linha com a revisão de literatura efetuada. Numa primeira fase procedeu-se à recolha dos dados. Seguidamente, foi efetuado o processamento e análises dos mesmos. Posteriormente, foram aplicados modelos de regressão logística, árvores de decisão e *random forests* por forma a prever o não comparecimento nas consultas médicas. A Figura 4 pretende ilustrar o processo descrito.

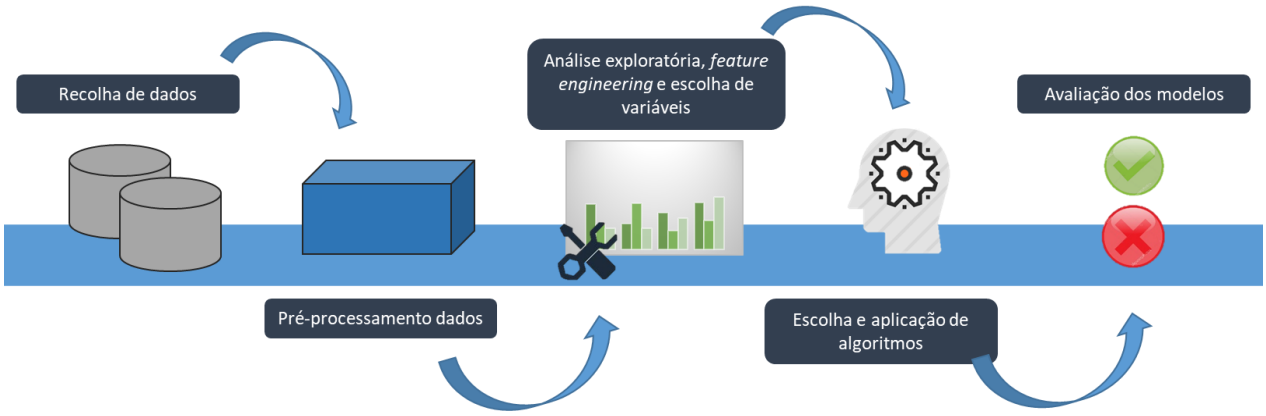

Figura 4 - Metodologia utilizada

1.7. RECOLHA DE DADOS

Os dados utilizados no âmbito deste estudo foram obtidos através do repositório de *datasets* da página *Kaggle*[1]. O *dataset* obtido compreende à volta de 110 mil registos de consultas médicas do serviço público de saúde da cidade de Vitória no estado de Espírito Santo no Brasil. Os dados mostram o registo das consultas e data das mesmas entre abril de 2016 e junho de 2016. A tabela seguinte pretende ilustrar as variáveis que compõem o conjunto de dados.

Variável	**Descrição**
Age	Idade
Gender	Género
ScheduledDay	Data de registo/marcação da consulta
AppointmentDay	Data da consulta
AppointmentID	Identificador único do registo de consulta
PatientID	Identificador único do doente
No-show	Identifica se o doente compareceu à consulta

[1] https://www.kaggle.com/datasets

Variável	Descrição
Diabetic	Identifica se o doente tem diabetes;
Alchoholic	Identifica se o doente tem problemas de alcoolismo;
Hypertension	Identifica se o doente tem hipertensão;
Handicap	Identifica se o doente tem alguma deficiência
Scholarship	Indica se o doente é beneficiário de uma "bolsa família" – comparticipado pelo Estado Brasileiro
Sms_received	Identifica se foi enviado um lembrete da consulta por sms
Neighbourhood	Identifica o local onde a consulta ocorreu

Tabela 1- Lista de variáveis obtidas e respetiva descrição

As variáveis presentes no *dataset* podem ser divididas em 3 grupos:

- Estado – variável dependente "No-Show" que pode assumir os valores "No" ou "Yes" consoante o doente compareça ou não à consulta médica;
- Consultas - variáveis que caracterizam detalhes sobre a consulta em termos de tempos de espera, registo de datas e respetivos lembretes;
- Doentes – variáveis que pretendem caracterizar o doente em termos de género, idade e condições de saúde bem como sócio-económicas.

A linguagem *Python*[2] (através da plataforma Anaconda[3] com o editor *Spyder*) foi utilizada na implementação deste projeto, desde o pré-processamento dos dados até à avaliação dos modelos preditivos.

1.8. PRÉ-PROCESSAMENTO DOS DADOS

A fase de pré-processamento dos dados pretende identificar e corrigir problemas com a qualidade dos dados e prepará-los para a modelação. Em alguns casos, novas variáveis podem ser adicionadas por forma a potencializar o conhecimento extraído dos dados.

Numa primeira fase fez-se uma renomeação das variáveis extraídas no dataset para corrigir erros de datilografia bem como tornar os nomes mais percetíveis de acordo com os dados. E em seguida verificou-se a existência de valores nulos ou *missing* para todas as variáveis, bem como se computou estatísticas descritivas para antecipar alguma anomalia ou valores fora do normal.

3.1.1. Estado

A variável dependente, renomeada para *Status*, consiste na indicação se o doente compareceu ("*Show-Up*") ou não à consulta médica ("*No-Show*"). Verificou-se que 30% dos registos diz respeito a

[2] https://www.python.org/
[3] https://www.anaconda.com/enterprise/

não comparecimentos. A tabela abaixo pretende ilustrar a distribuição da variável *target* no *dataset* em estudo.

Status	Distribuição	Percentagem
No-Show	22 319	20,2%
Show-Up	88 208	79,8%

Tabela 2- Distribuição variável target

Para posterior aplicação dos modelos de previsão, esta variável "Status" foi codificada de acordo com o seguinte:

- Assume o valor 0 quando apresenta valores *Show-Up*;
- Assume o valor 1 quando apresenta valores *No-Show*.

3.1.2. Consultas

As variáveis que caracterizam as consultas são a data de registo da consulta, data efetiva da consulta, o lembrete da consulta via SMS e a zona/bairro (localidade) onde a consulta ocorre.

Às variáveis de data ("*AppointmentDay*" e "*ScheduledDay*") foi retirada a componente tempo (formato original *datetime*) para capturar apenas a componente data das mesmas.

A variável "*Neighbourhood*" identifica a localidade onde a consulta ocorreu. Após análise da mesma, identificaram-se mais de 80 localidades diferentes, que posteriormente teriam de ser codificadas para aplicação dos algoritmos. Dada a dimensão da variável, e não tendo informação adicional para poder trabalhar a mesma (e.g. criar grupos de localidades), decidiu-se excluir a mesma.

3.1.3. Doentes

Das 7 variáveis que caracterizam o doente, 5 são categóricas e identificam condições de saúde (Ex: portador de Diabetes). Estas apresentam-se como variáveis binárias, apresentando valor 1 quando o doente sofre ou é portador de determinada condição de saúde e o valor 0 caso contrário.

A variável "*Handicap*" apresentava valores entre 0 e 4 dependendo do número de deficiências que o doente apresente. Desta forma, e devido não só ao facto de mais classes neste caso não acrescentarem relevância na informação que passam, mas também devido à baixa expressão das classes 1 a 4, optou-se por agrupar as mesmas e transformar a variável em binária.

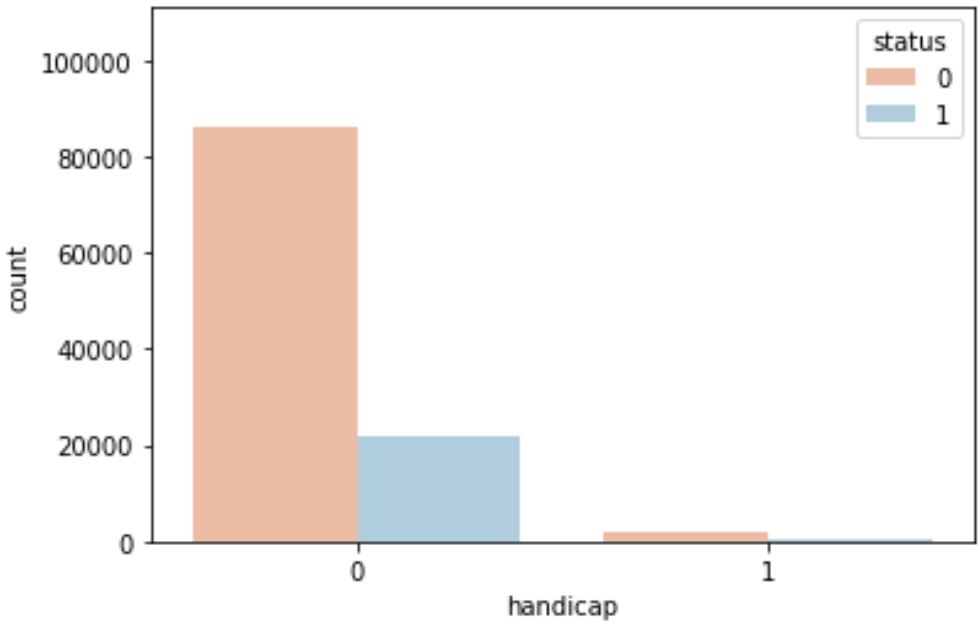

Figura 5 - Distribuição do status por handicap (0 – sem deficiência, 1 – com deficiência)

A variável "*Scholarship*" é também uma variável binária que apresenta o valor 1 quando os doentes são beneficiários de um programa intitulado de "bolsa família". Este programa social é atribuído pelo governo brasileiro por forma a prover as famílias menos favorecidas financeiramente. Com esta medida, é pretendido que as famílias com crianças no seu agregado familiar consigam assegurar que as mesmas têm acesso a educação e vacinação.

As restantes 2 variáveis que caracterizam o doente dizem respeito à idade e ao género do mesmo. A idade apresenta-se como variável contínua medindo os anos completos.

Observou-se que a variável idade continha valores abaixo de zero. Uma vez que não podem existir idades com valor negativo, os registos que apresentavam tais dados foram tratados como *outliers* tendo sido eliminados consequentemente (1 registo). A figura seguinte pretendem ilustrar a variação da variável após o tratamento da mesma.

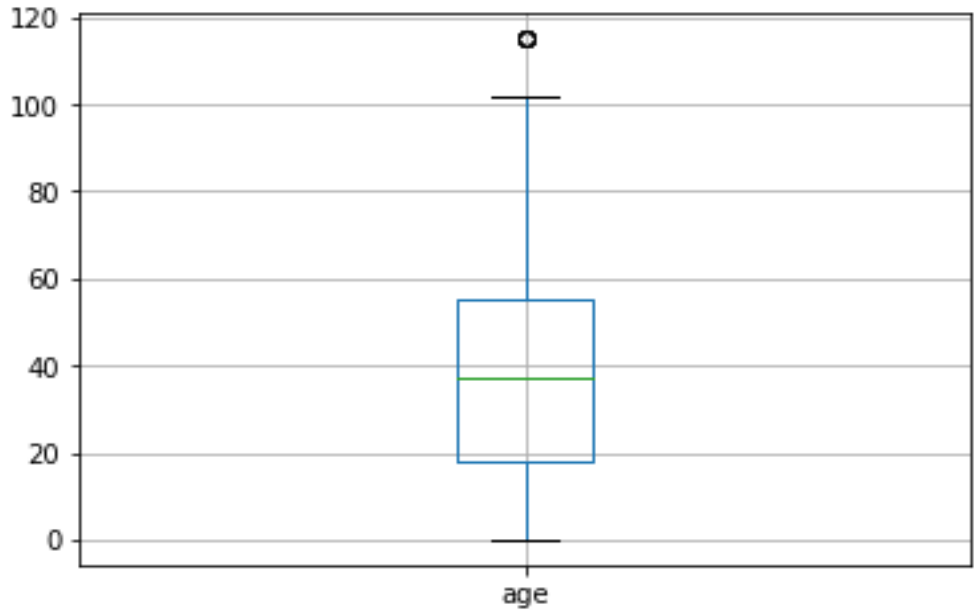

Figura 6 - Representação dos valores da idade depois do tratamento

1.9. ANÁLISE EXPLORATÓRIA DOS DADOS

Num primeiro momento, foi efetuada uma análise descritiva (univariada e multivariada) por forma a identificar possíveis relações entre as mesmas e a variável dependente.

Através da análise do género distribuído pela variável dependente é possível observar que as mulheres (Gender=0) são o grupo mais presente no conjunto de dados, tendo um total de 71 836 registos (65 % do total de registos), ou seja, são o grupo que mais consultas médicas tem marcadas no período analisado. É também neste grupo que observamos a maior percentagem de não comparecimentos (65 % em doentes do género feminino e 35% do género masculino, em relação ao universo de *no-shows*). A Figura 3 mostra a distribuição da variável género por "*Status*".

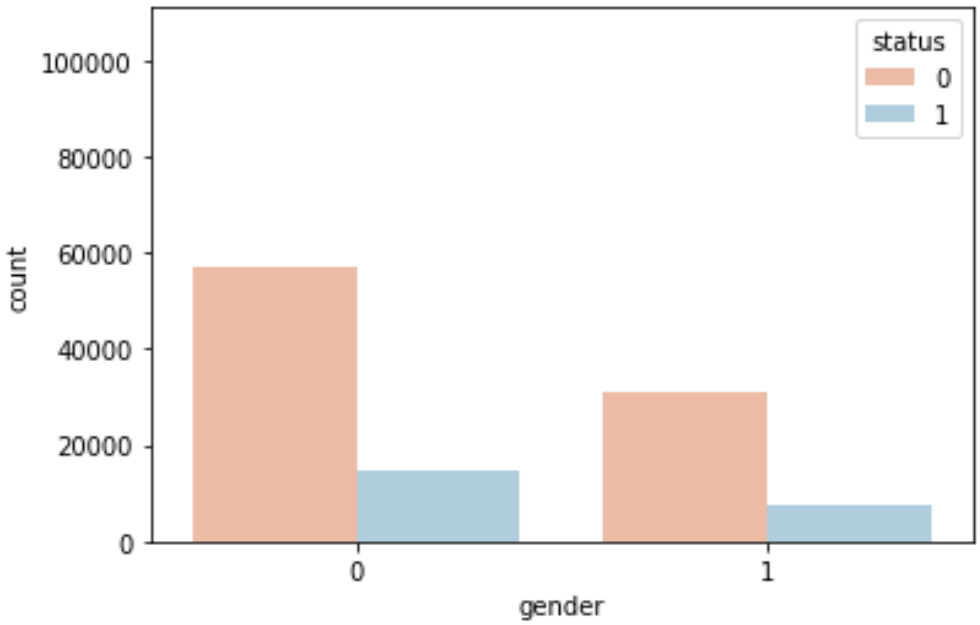

Figura 7 - Distribuição do status por género (valor 0 – feminino, valor 1 – masculino)

Foi ainda possível observar que há uma maior percentagem de não comparecimentos às consultas médicas quando o doente não tem condicionamentos de saúde. Esta tendência foi observada em todas as variáveis que caracterizam as condições de saúde do doente. As figuras que se seguem ilustram esta tendência.

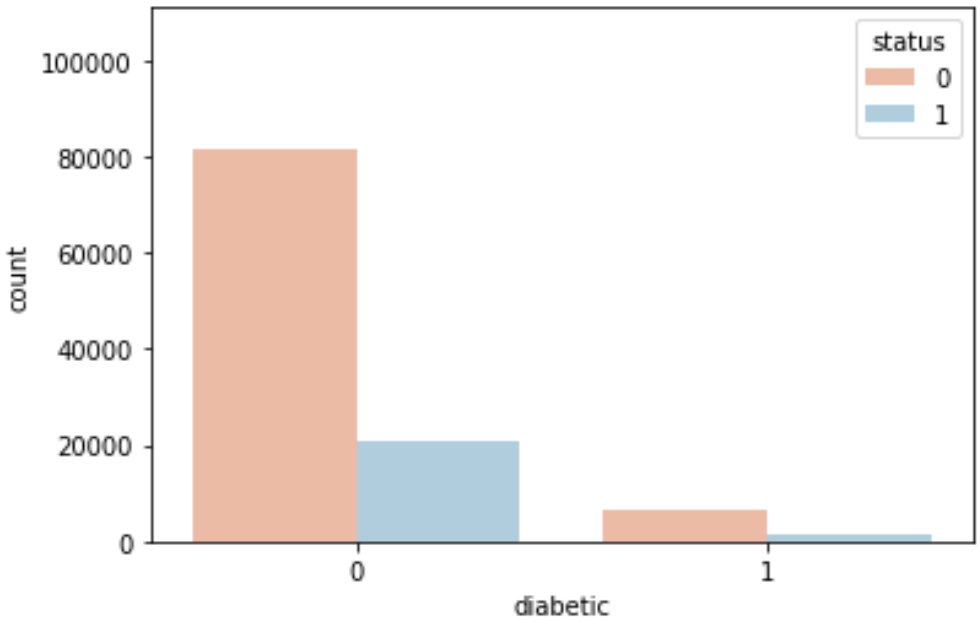

Figura 8 - Distribuição do status por Diabetes

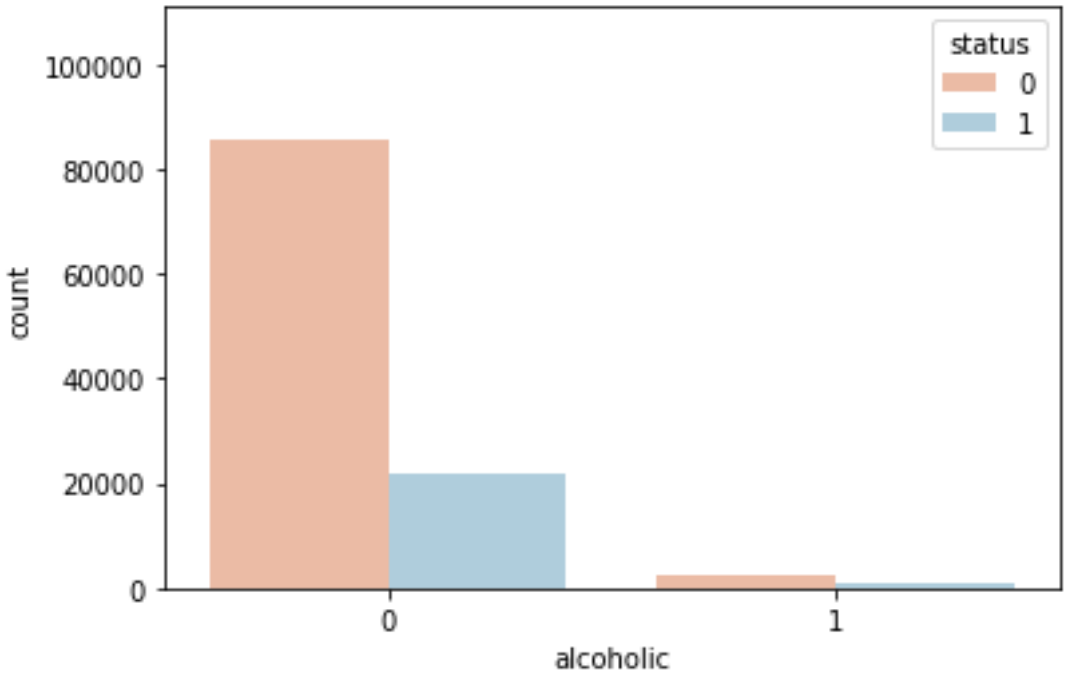

Figura 9 - Distribuição do status por Alcoolismo

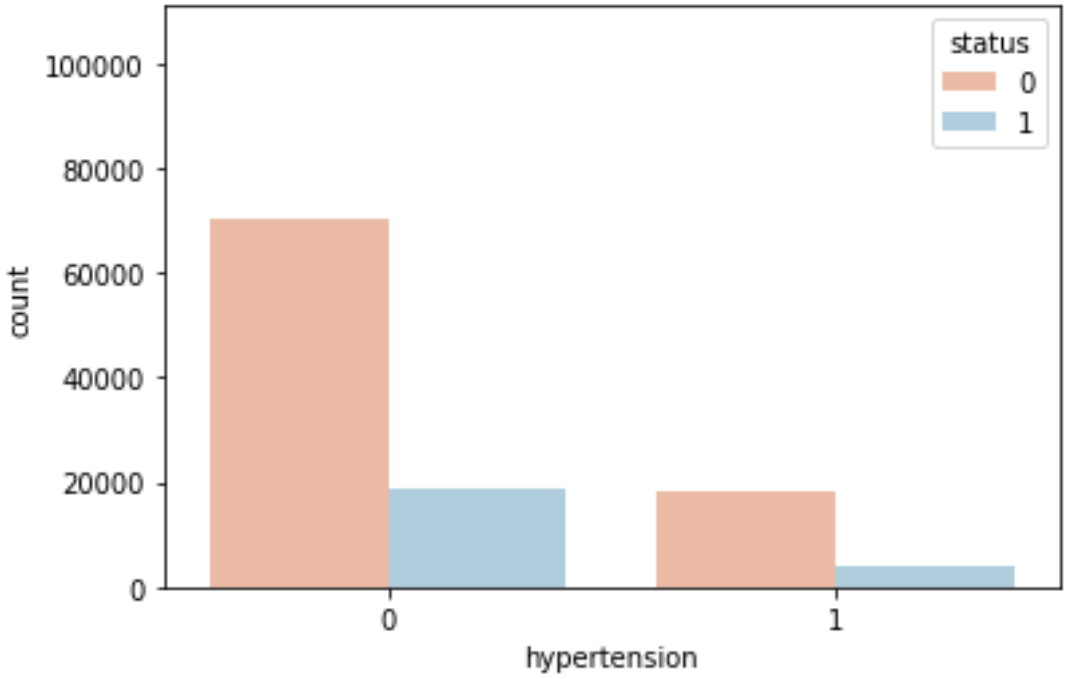

Figura 10 - Distribuição do status por Hipertensão

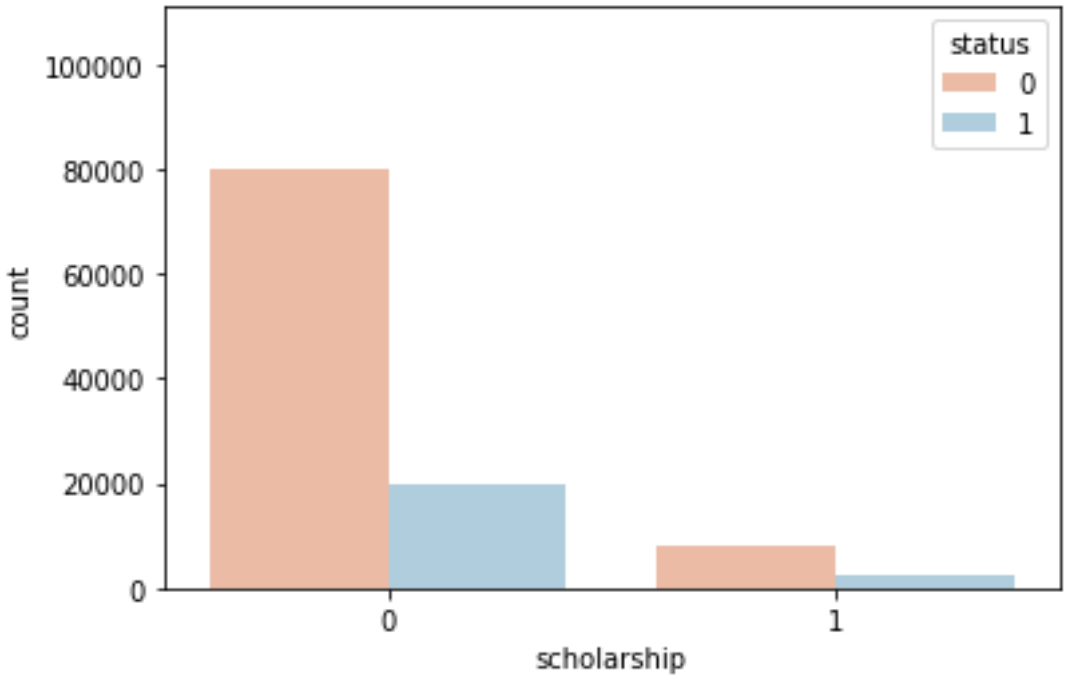

Figura 11 - Distribuição do status por “bolsa família”

3.1.4. Outliers

Outliers são valores extremos que fogem à tendência dos restantes e podem enviesar o processo de treino dos algoritmos e levar a resultados não corretos (Larose, 2015).

Algoritmos que utilizam árvores, sejam árvores de decisão ou *random forests* são indiferentes à existência de *outliers*, não impactando os seus pontos de corte. No entanto, para algoritmos como regressões logísticas existe um impacto significativo quanto à presença de valores extremos. Desta forma, no decorrer da análise exploratória, houve uma preocupação em identificar possíveis outliers nos dados. No *dataset* utilizado não foi detetada a presença de *outliers*.

1.10. CRIAÇÃO DE NOVAS VARIÁVEIS (FEATURE ENGINEERING)

No decorrer das análises foram identificadas possíveis novas variáveis que poderiam contribuir para prever melhor a variável dependente. Desta forma, a Tabela 3 ilustra as varáveis criadas através das já existentes.

Variável	Descrição
ScheduledDay_Hours	Hora do dia relativa ao dia de agendamento
Day_Difference	Diferença em dias entre a data da consulta e a data de agendamento
Appointment_Weekday	Dia da semana relativo à data da consulta
Scheduled_Weekday	Dia da semana relativo à data de agendamento

Variável	Descrição
Prior_appointments	Existência de consultas anteriores para o mesmo doente
Prior_no_shows	Existência de não comparecimentos anteriores para o mesmo doente
Total_conditions	Indica de entre as condições de saúde, se o doente tem pelo menos uma (diabetes, hipertenso, etc.)

Tabela 3 - Lista de variáveis criadas via feature engineering

A variável "*ScheduledDay_Hours*" representa a hora do dia relativa ao dia de agendamento da consulta. Esta variável foi criada com intuito de perceber o efeito das horas de marcação no efeito de *no-show*. A hora do dia relativa ao dia da consulta em si, não foi possível obter pois a variável "*AppointmentDay*" apresentava a componente hora a zeros (formato *datetime*).

A variável "*Day_Difference*" representa a diferença, em dias, entre a data de agendamento e a data efetiva da consulta. Esta permite avaliar se a incidência de no-shows pode ser influenciada pelo facto de ter decorrido mais ou menos dias entre a data de marcação e a data da consulta.

As variáveis "*Appointment_Weekday*" e "*Scheduled_Weekday*" representam os dias da semana relativos às datas de agendamento e data da consulta. Estas variáveis pretendem perceber o efeito do dia da semana para o fenómeno dos *no-shows*.

Foi possível observar que não ocorrem consultas aos domingos (dia 7) nem existem agendamentos. Adicionalmente, identificou-se que as marcações bem como as consultas são residuais ao sábado (dia 6), o que pode indicar que se trata de algumas exceções ou que o horário das clínicas é mais reduzido neste dia.

Verificou-se que, em relação às datas das consultas, os três primeiros dias da semana (segunda a quarta-feira) são os dias com mais incidência de *no-shows*. Importa referir, que considerando as datas de agendamento, esta tendência também se verifica, isto é, há uma maior incidência de *no-shows* quando as consultas são agendadas entre segunda e quarta-feira.

As figuras abaixo pretendem ilustrar este comportamento para ambas as variáveis referidas.

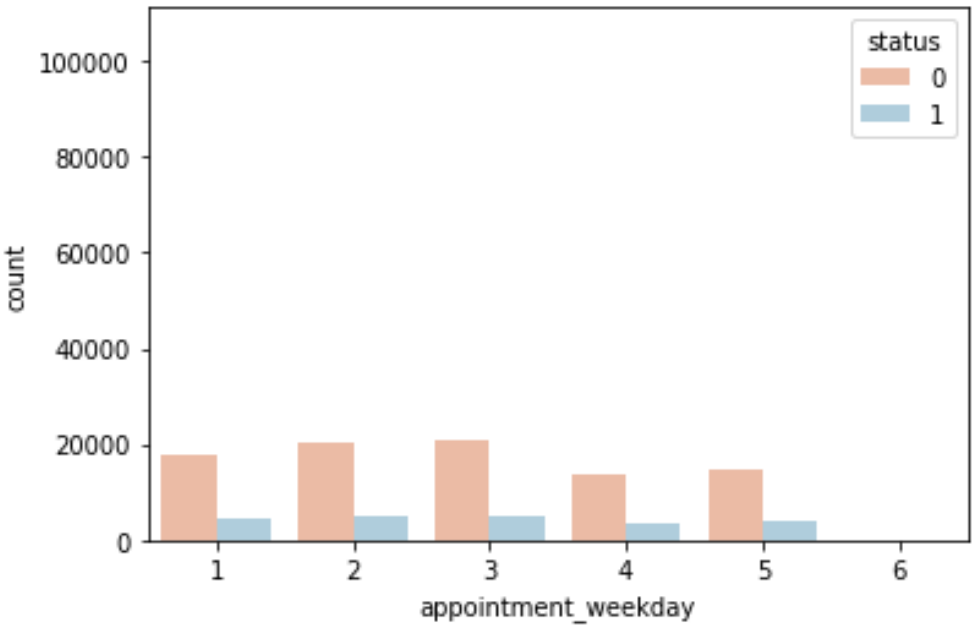

Figura 12 - Distribuição do status por dia da semana da data da consulta

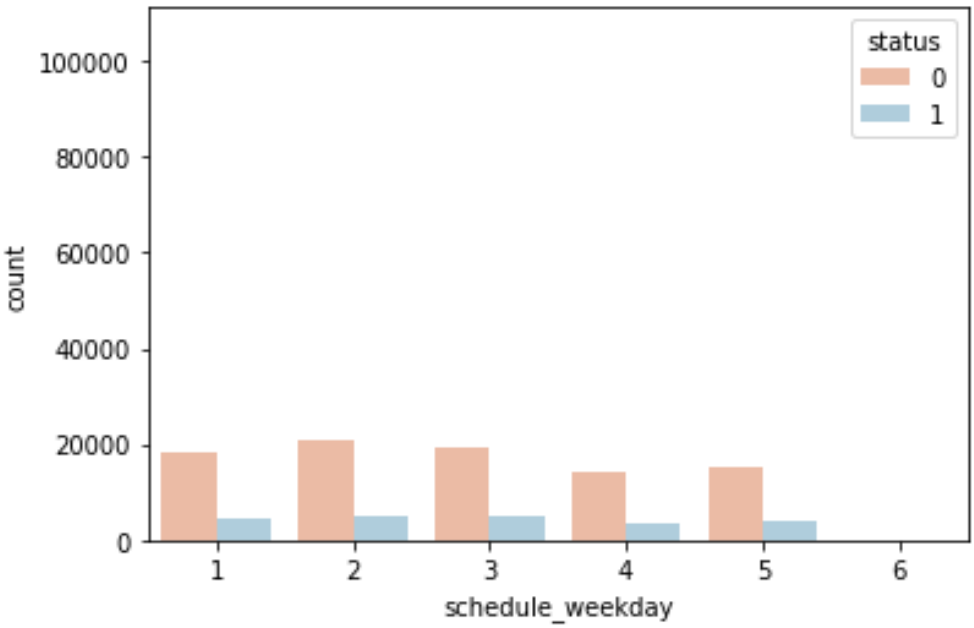

Figura 13 - Distribuição do status por dia da semana da data de agendamento

As variáveis "*prior_appointments*" e "*prior_no_shows*" representam o número total de consultas e não comparecimentos por cada doente. A criação destas variáveis pretende verificar o efeito do histórico do doente em termos de consultas marcadas e *no-shows*.

A variável "*total_conditions*" identifica se o doente tem pelo menos uma condição de saúde de entre as referidas em **Error! Reference source not found.,** ou seja, se o doente sofre de hipertensão, alcoolismo, diabetes ou deficiência física.

1.11. ESCOLHA DE VARIÁVEIS

Nesta fase o *dataset* conta com 20 variáveis, podendo algumas não ser totalmente interessantes para considerar na fase de modelação. Desta forma, é necessário escolher as variáveis mais relevantes por forma a maximizar o desempenho do modelo preditivo e potenciar a precisão dos resultados.

As variáveis "*appointment_id*" e "*patient_id*" foram excluídas pelo facto de se tratar de identificadores que não trazem nenhuma informação adicional para a fase da modelação.

É pretendido um modelo que consiga prever em dados desconhecidos pelo que se o treinarmos com dados históricos, como as datas da consulta e agendamento, o seu desempenho será pobre a prever eventos futuros. Assim, foram também excluídas as variáveis de data ("*AppointmentDay*" e "*ScheduledDay*") e serão utilizadas as variáveis construídas a partir destas, como o dia da semana e a diferença em dias entre as datas.

Foi efetuada uma análise multivariada das variáveis por forma a identificar possíveis correlações entre as variáveis. Uma matriz de correlação foi obtida tendo em conta as variáveis remanescentes. A Figura 9 ilustra a matriz de correlação obtida.

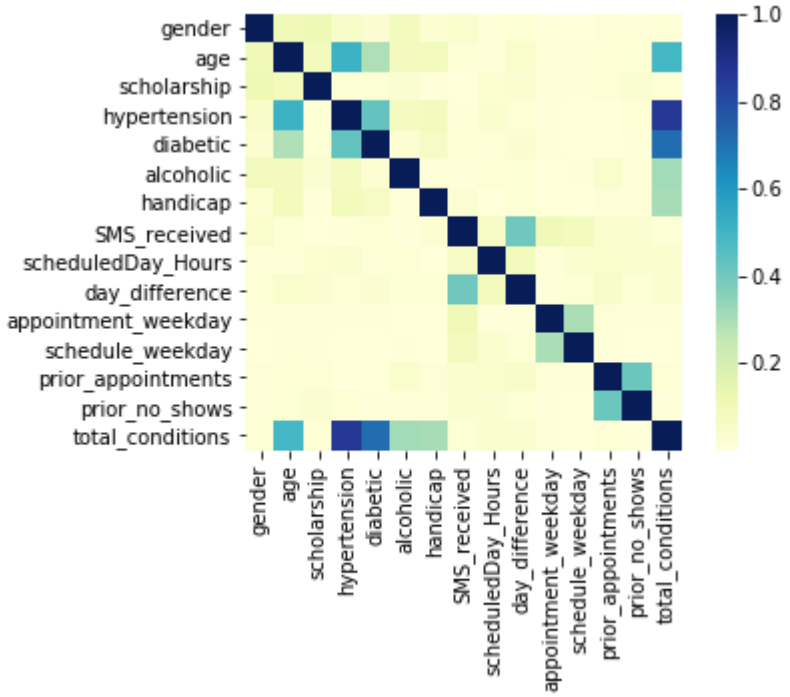

Figura 14 - Matriz de correlação entre as variáveis (coeficiente correlação de Pearson)

Através da matriz acima é percetível a alta correlação entre a variável que representa o número total de condições de saúde ("*total_conditions*") e as variáveis que representam a existência de diabetes e hipertensão, com 71% e 86%, respetivamente. Também é visível, ainda que em menores valores, a correlação entre a "*total_conditions*" e a idade, existência de dependência de álcool e existência de deficiência física ("age", "*alcoholic*" e "*handicap*" com 49%, 31% e 30%, respetivamente).

É ainda possível observar que a idade ("*Age*"), como esperado, está correlacionada com a existência de doenças como hipertensão e diabetes com valores de 50% ("*hypertension*") e 30% ("*diabetic*"), respetivamente. Importa referir também que a própria existência de doenças como hipertensão e diabetes estão também correlacionadas entre si, apresentando um valor de 43%.

Adicionalmente, é possível observar que a diferença em dias entre as datas de agendamento e de consulta ("*day_difference*") apresenta um valor de 40% de correlação com a variável que representa a existência de um lembrete da data da consulta via SMS ("*SMS_received*"). Este comportamento seria também de esperar.

Além das correlações positivas descritas acima em relação à variável "*total_conditions"*, não foram observadas correlações superiores a 60%, pelo que todas as variáveis foram mantidas à exceção desta.

1.12. ESCOLHA E APLICAÇÃO DOS ALGORITMOS

Uma vez que os dados foram recolhidos do repositório de *datasets* da plataforma *Kaggle*, e dado que a mesma permite a partilha de conhecimentos na exploração dos *datasets*, decidiu-se analisar os trabalhos feitos sobre estes dados. Desta forma, foram selecionados os 3 trabalhos mais votados[4] pela comunidade dentro da plataforma e identificados os algoritmos mais escolhidos.

Nos trabalhos analisados, os autores aplicaram *random forests*, *Multinomial Naive Bayes* e árvores de decisão. Dos modelos analisados, o que demonstrou melhor performance foi a utilização de *random forests* com uma *accuracy* de 80%. A Tabela 5 apresenta os principais resultados dos 3 trabalhos analisados.

Modelos	**1º lugar**	**2º lugar**	**3º lugar**
Random forests	-	80%	66%
Árvores decisão	-	72%	-
Naive Bayes	71%	-	-

Tabela 4- Accuracy dos modelos obtidos nos 3 trabalhos mais votados na plataforma.

Dos trabalhos analisados foi possível perceber que as *features* mais relevantes no *dataset* são a idade seguida do género.

[4] Para este efeito considerou-se apenas os trabalhos submetidos na mesma linguagem (*Python*) e cujo objetivo era também a previsão de *no-shows* (https://www.kaggle.com/joniarroba/noshowappointments/notebooks?datasetId=792&language=Python)

Desta forma, a escolha dos algoritmos a aplicar no presente trabalho está em linha com a revisão da literatura e com os trabalhos mais votados da plataforma *kaggle*. Assim, optou-se por aplicar árvores de decisão e *random forests*. Adicionalmente, optou-se pela construção de um modelo de regressão logística por forma a perceber qual o seu desempenho neste tipo de problemas.

Por forma a diferenciar dos trabalhos revistos na plataforma *kaggle*, optou-se por aplicar um algoritmo *ensemble*, usando o método *voting* em busca de um modelo com melhor capacidade de generalização.

Todos os algoritmos foram acompanhados de validação cruzada (método *k-fold*) e os dados foram particionados em subconjuntos.

Adicionalmente, sempre que possível os hiperparâmetros foram calibrados em busca da combinação com melhor desempenho.

3.1.5. Partição de dados

Após a seleção dos algoritmos, foi feita a partição dos dados do *dataset* para aplicação dos mesmos. A partição de dados, uma das abordagens mais defendidas na aplicação de algoritmos de *machine learning*, permite simular a capacidade de um modelo prever certo acontecimento em dados futuros. Para tal, o conjunto de dados é dividido em subconjuntos com propósitos diferentes dependendo da metodologia.

O mais comum é dividir os dados em dois subconjuntos, o conjunto de treino e o conjunto de teste. O conjunto de treino serve para treinar o modelo, ou seja, para o modelo aprender sobre o conjunto de dados em questão. Enquanto o conjunto de teste serve para avaliar a performance do modelo em relação a dados diferentes dos do treino, isto é, serve para avaliar a capacidade preditiva do modelo aplicado.

No âmbito do presente trabalho, para as experiências iniciais, o *dataset* foi dividido inicialmente em 2 conjuntos. O conjunto de treino composto por 80% das observações e o conjunto de teste composto pelos 20% de observações remanescentes.

Posteriormente, para todos os algoritmos, foi feita a procura dos melhores parâmetros usando validação cruzada com procura aleatória, e com os melhores parâmetros encontrados foi feito o *refit* a todo o conjunto de treino e consequente predição no conjunto de teste.

3.1.6. Validação cruzada

A validação cruzada é uma técnica amplamente usada em ML para avaliar a capacidade de generalização de um modelo. O conceito principal desta técnica é a divisão do conjunto de dados em subconjuntos mutuamente exclusivos usando alguns para treino e outros para teste do modelo.

Existem vários métodos para a aplicação da validação cruzada, sendo os três mais conhecidos o método *holdout*, o método *k-fold* e o *leave-one-out*. No âmbito do presente trabalho iremos aplicar o método *k-fold*.

K-fold

O método *k-fold* consiste em dividir o conjunto total de dados em *k* subconjuntos de dados do mesmo tamanho sendo um conjunto usado para teste e os restantes *k-1* conjuntos usados para treino. Este processo é realizado *k* vezes rodando o conjunto de teste para que no final todos os conjuntos tenham servido de conjunto de teste. Além de permitir avaliar a estabilidade do modelo, esta técnica permite evitar o problema de sobre aprendizagem.

3.1.7. Procura dos hiperparâmetros

Muitas vezes o desempenho de um algoritmo de ML pode ser melhorado através da otimização dos hiperparâmetros do mesmo, isto é, os parâmetros que definem a sua arquitetura. Estes parâmetros não estão diretamente relacionados aos dados para treino nem ao processo de treino em si. São variáveis de configuração do algoritmo, que uma vez calibradas, permitem encontrar a melhor combinação para o problema em estudo.

Para encontrar a melhor combinação, primeiramente define-se quais as variáveis a calibrar e a faixa de valores possíveis para as mesmas e posteriormente define-se a técnica a utilizar para a calibração.

Existem diversas técnicas de procura de hiperparâmetros, nomeadamente procura exaustiva do espaço, procura aleatória ou aplicação de modelos substitutos (e.g. *Bayesian optimization*). No presente trabalho, foi feita uma procura exaustiva do espaço de procura usando *Random search*.

Random Search

Random Search é uma estratégia de calibração de parâmetros onde combinações aleatórias de hiperparâmetros são usados para obter o melhor resultado para o modelo. Esta técnica é acompanhada de validação cruzada (ver secção **Error! Reference source not found.**) por forma a treinar e testar as diversas combinações selecionadas aleatoriamente.

1.13. VALIDAÇÃO

Após aplicação dos algoritmos selecionados em **Error! Reference source not found.**, foram usadas métricas de avaliação dos modelos criados por forma a perceber qual o modelo com melhor desempenho. As medidas selecionadas para avaliação dos modelos obtidos foram *accuracy*, AUC, *precision* e *recall*. Na secção **Error! Reference source not found.** foram apresentadas com mais detalhe estas medidas para a comparação e seleção de modelos.

Os resultados podem ser consultados na secção seguinte.

4. RESULTADOS E DISCUSSÃO

No capítulo anterior foi descrita a metodologia aplicada no desenvolvimento do presente trabalho. Neste capítulo será efetuada uma apresentação e discussão dos resultados obtidos.

Com recurso à linguagem *python* e tendo definido as variáveis a usar na modelação preditiva de acordo com a secção 3.2.4.2, foram criados os modelos usando os algoritmos regressão logística, árvores de decisão, *random forests e voting*.

Um dos objetivos deste trabalho passa pela aplicação dos algoritmos e identificação do melhor modelo de acordo com as suas características. Nesta secção são apresentados os resultados obtidos tendo em conta as medidas referidas em **Error! Reference source not found.** para comparação e seleção de modelos.

1.14. TREINO

De acordo com a metodologia apresentada anteriormente, na Tabela 6 apresentamos os resultados dos algoritmos utilizados após aplicação ao conjunto de treino.

Modelo	AUC
Regressão Logística	56,9
Árvores de decisão	58,2
Random forests	57,8
Voting	57,9

Tabela 5- Resumo dos resultados para cada técnica e algoritmo aplicados ao conjunto de treino

De um modo geral, todos os modelos apresentam cerca de 60% de AUC, sendo o algoritmo *ensemble* com método *voting* o que apresenta o valor mais elevado (57,9%). O algoritmo *random forest* segue-se com o valor de AUC de 57,8%. A regressão logística é o algoritmo que apresenta um valor menor de AUC de entre os 4 modelos.

Na secção seguinte são apresentados os resultados da aplicação ao conjunto de teste.

1.15. TESTE

Após o treino, aplicaram-se os mesmos modelos ao conjunto de teste por forma a verificar como os mesmos se comportam com novos dados. Na Tabela 7 são apresentados os resultados da aplicação dos algoritmos ao conjunto de validação.

Modelo	***Accuracy***	**AUC**	***Precision***	***Recall***
Regressão logística	81,7	71,3	71,5	15,4
Árvores de decisão	**82,4**	77,9	79,7	**17,5**
Random forests	**82,4**	**78,1**	80,3	16,7
Voting	**82,5**	77,2	**81,9**	16,8

Tabela 6- Resumo dos resultados para cada algoritmo aplicado ao conjunto de validação

De um modo geral todos os algoritmos apresentam valores similares para todas as medidas apresentadas. Em termos de AUC, os modelos apresentam entre 71% e 78%, o que em relação aos resultados do treino, é uma melhoria de cerca de 17 pontos percentuais, em média.

É possível observar que o método *voting* foi o algoritmo que obteve melhor valor de *accuracy* (82,5%) de entre os 4 algoritmos aplicados a melhor *precision* (81,9%) e identificou cerca de 17% de *no-shows* (*recall* de 16,8%). Já a regressão logística é o algoritmo que apresenta os valores mais baixos para todas as medidas.

As *random forests* são o algoritmo que apresenta o valor de AUC mais elevado, com 78%. Este algoritmo consegue identificar corretamente os *no-shows* 80% das vezes, identificando efetivamente cerca de 17% dos mesmos (*recall* de 16,7%).

Nas figuras 18 e 19 é possível visualizar as curvas de ROC e as curvas de *Precision-Recall* dos 4 modelos para uma melhor comparação da performance dos mesmos.

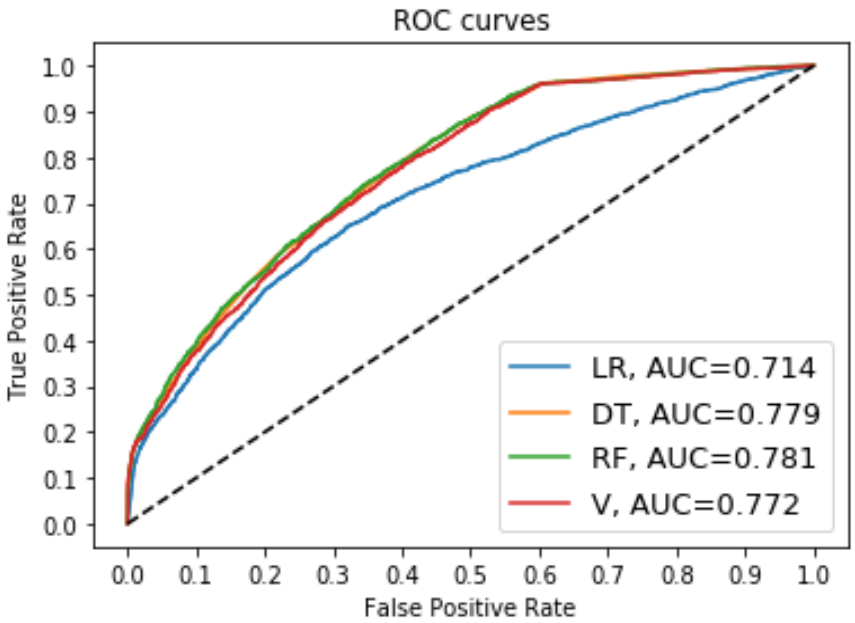

Figura 15 - Representação gráfica das curvas de ROC para todos os modelos e respetivas áreas debaixo da curva (AUC)

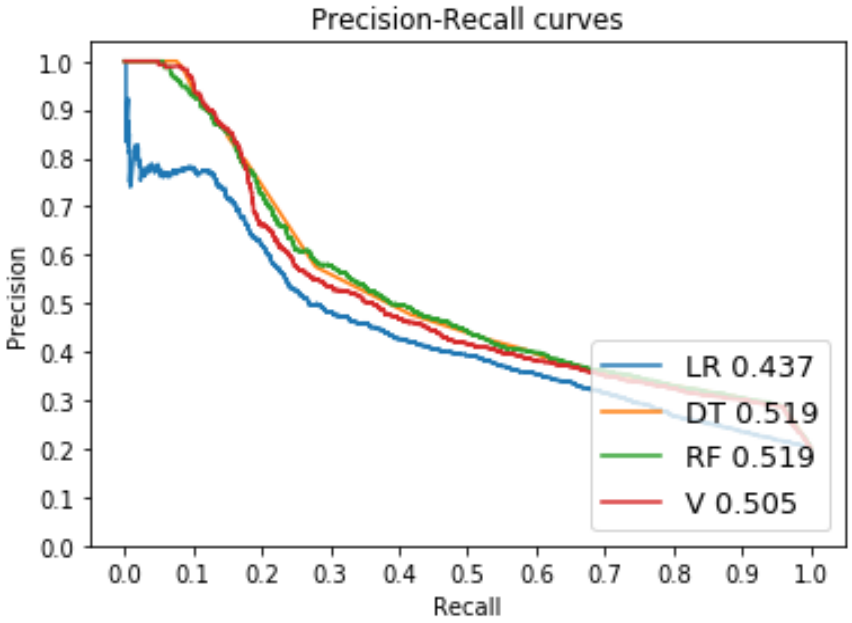

Figura 16 - Representação gráfica das curvas de Precision-Recall para todos os modelos e respetivas áreas debaixo da curva (AUC)

A partir da figura 19 podemos verificar que as curvas de ROC se aproximam bastante para todos os modelos aplicados, à exceção da regressão logística que mostra uma curva mais perto da *baseline*, o que indica um modelo com menos capacidade discriminativa.

As *random forests* são o algoritmo que mostra um AUC e curva ROC superiores aos outros e mais perto do canto superior esquerdo do gráfico, representando uma taxa de positivos verdadeiros ligeiramente superior e uma taxa de falsos positivos ligeiramente inferior aos demais algoritmos.

Quando comparadas as curvas *precision-recall*, é possível verificar que o número de no-shows corretamente identificados decresce a partir dos 10% de *recall* para todos os modelos apresentados. No entanto o modelo com as *random forests* e árvores de decisão, mostra um melhor balanço entre as duas medidas, com um AUC relativo às curvas de *precision* e *recall* de aproximadamente 52%.

Revisitando os resultados apresentados anteriormente para as medidas *precision* e *recall*, podemos ver que as *random forests* comportam-se ligeiramente melhor, identificando corretamente cerca de 80% das vezes os não comparecimentos para um total de 16,7% de eventos de no-show identificados.

5. CONCLUSÕES

Este projeto tinha como objetivos compreender os fatores mais relevantes aquando do não comparecimento a consultas médicas e encontrar o melhor algoritmo para prever *no-shows*.

Por forma a atingir os objetivos propostos foi efetuada uma análise aos dados com particular foco nos eventos de *no-show* por forma a identificar padrões nas diferentes dimensões associadas – estado, consultas e caraterísticas dos doentes. Foi possível observar que há uma maior percentagem de *no-shows* quando o doente não apresenta qualquer condicionamento de saúde (ex: diabetes, handicap, etc.) e que as mulheres apresentam maior probabilidade de falhar as consultas relativamente aos homens (65% e 35% respetivamente).

Por forma a encontrar o melhor algoritmo para prever *no-shows*, foram aplicados os algoritmos regressão logística, árvores de decisão, *random forests* e *voting* como técnicas de ensemble.

Os algoritmos apresentam valores na mesma ordem de grandezas para todas as medidas apresentadas. Comparando os resultados obtidos com os resultados dos modelos da plataforma *kaggle*, apresentados na secção **Error! Reference source not found.**, o método *Voting* é o que tem um melhor valor de *accuracy*, de cerca de 83%.

No entanto, como os dados utilizados apresentam classes não balanceadas (80% *shows* vs 20% *no-shows*), a *accuracy* não é a melhor medida a considerar para a escolha do melhor modelo. Deste modo, iremos utilizar o valor de AUC e a curva de ROC para identificar o modelo com maior capacidade discriminativa de entre os 4 algoritmos aplicados.

As *random forests* apresentam o valor mais elevado de AUC, com cerca de 78% de capacidade discriminativa entre as classes de *no-shows* e *shows* e uma curva de ROC superior às curvas dos restantes modelos. Adicionalmente, este algoritmo identifica corretamente cerca de 80% das vezes os no-shows, identificado cerca de 17% dos mesmos.

Na

Tabela 7 estão representadas as variáveis mais importantes a explicar o fenómeno *no-show* para o modelo *random forests*.

Variáveis	*Importâncias*
day_difference	0,446
prior_no_shows	0,365
age	0,056
SMS_received	0,051
prior_appointments	0,032

scheduledDay_Hours	0,021
hypertension	0,009
schedule_weekday	0,009
appointment_weekday	0,006
scholarship	0,002
alcoholic	0,001
gender	0,001
diabetic	0,001
handicap	0,000

Tabela 7 - Valores relativos à importância das variáveis para o modelo random forests

Podemos observar que a diferença em dias entre o agendamento da consulta e a data da mesma é a variável que melhor ajuda a explicar o não comparecimento às consultas médicas. Segue-se a existência prévia de eventos de não comparecimentos, depois a idade do doente, a existência de um aviso da data da consulta por sms.

Das condições de saúde do doente, a hipertensão é a condição que melhor ajuda a explicar o fenómeno de no-shows, enquanto a existência de deficiências motoras é a que menos contribui.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Em futuros trabalhos seria interessante conseguir incrementar o *dataset* com informação relativa a marcação de consultas anteriores e balancear mais o mesmo com informação sobre não comparecimentos.

Adicionalmente, seria interessante aplicar algoritmos de redes neuronais como por exemplo *deep learning*, por forma a verificar o seu comportamento com o *dataset* em questão.

Variáveis adicionais que caracterizem a região, a meteorologia e características sócio-económicas dos doentes também seriam um incremento interessante a explorar futuramente.

7. BIBLIOGRAFIA

Abdelhalim, A., Traore, I. (2009). Converting Declarative Rules into Decision Trees. *Proceedings of the World Congress on Engineering and Computer Science*, Vol. I.

Alaeddini, A., et al (2015). A hybrid prediction model for no-shows and cancellations of outpatient appointments. IIE Transactions on Healthcare Systems Engineering, Issue 5:1, 14-32, DOI:10.1080/19488300.2014.993006

Bajaj, P. Reinforcement learning. GeeksforGeeks, A computer science portal for geeks. Disponível em https://www.geeksforgeeks.org/what-is-reinforcement-learning/. Acedido em 2020.

Bissuel, A., (2019). Hyper-parameter optimization algorithms: a short review. Criteo tech blog, Medium. Disponível em: https://medium.com/criteo-labs/hyper-parameter-optimization-algorithms-2fe447525903. Acedido em 2020.

Bonaccorso, G., (2017). Machine Learning Algorithms. A reference guide to popular algorithms for data science and machine learning. Birmingham, UK: Packt Publishing Ltd.

deVille, B, (2006). Decision trees for business intelligence and data mining: Using SAS® Enterprise Miner™. Cary, NC: SAS Institute Inc.

Dantas, L., Oliveira, F. (Orientador), Hamacher, S. (Co-orientador) (2016). Revisão sistemática da literatura sobre no-show em agendamento de consultas. Dissertação de Mestrado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

Deep Learning Book. Data Science Academy. Disponível em http://deeplearningbook.com.br/o-que-e-aprendizagem-por-reforco/. Acedido em 2020.

Devasahay, S., *et al* (2017). Predicting appointment misses in hospitals using data analytics. *mHealth*, Vol. 3, Nº 4.

Dutta S., (2020). A 2020 Guide to Deep Learning for Medical Imaging and the Healthcare Industry. Nanonets blog. Disponível em https://nanonets.com/blog/deep-learning-for-medical-imaging/#deep-learning-for-medical-imaging. Acedido em 2020.

Elrazek, A. (2017). How Can Data Mining Improve Health Care?. *Applied Mathematics & Information Sciences*, 11, No. 2, 585-588.

Finlay, S., (2014). Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods. Palgrave Macmillan, 978-1-137-37927-6, https://doi.org/10.1057/9781137379283.

Gupta, A. ML | Semi-Supervised Learning. *GeeksforGeeks, A computer science portal for geeks*. Disponível em https://www.geeksforgeeks.org/ml-semi-supervised-learning/. Acedido em 2020.

Hand, D., et al (2001). Principles of Data Mining. The MIT Press, 026208290x.

How does data mining help healthcare? Archer blog. Disponível em https://archer-soft.com/blog/how-does-data-mining-help-healthcare#data_mining_in_healthcare. Acedido em 2020.

Huang, Y., Hanauer, D.A. (2014). Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach. *Applied Clinical Informatics*, Issue 5, 836-860.

Hussain, M., *et al* (2004). Missed appointments in primary care: questionnaire and focus group study of health professionals. *British Journal of General Practice*, Issue 54, 108-113.

Jothi, N., *et al* (2015). Data Mining in Healthcare – A Review. *Procedia Computer Science*, 72, 306-313.

Kantardzic, M (2011). Data mining: Concepts, models, methods, and algorithms, Second Edition.

Kaplan-Lewis, E., Percac-Lima, S., (2013). No-Show to primary care appointments: Why patients do not come. *Journal of Primary Care & Community Health*, 4(4) 251–255.

Kotsiantis, S. B., (2007). Supervised machine learning: A review of classification techniques. Informatica, 31, 249-268.

Kononenko, I., & Matjaz, K. (2007). MACHINE LEARNING AND DATA MINING: Introduction to Principles and Algorithms. Communications of the ACM. West Sussex, UK: Horwood Publishing Limited. Acedido em https://nemor.cz/data/mac.pdf.

Kurasawa, H., *et al* (2016). Machine-Learning-Based Prediction of a Missed Scheduled Clinical Appointment by Patients with Diabetes. *Journal of Diabetes Science and Technology*, Vol. 10(3), 730–736.

Maimon, O., Rokack, L., (2005). Data Mining and Knowledge Discovery Handbook, Second Edition, 978-0-387-09822-7.

Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (1986). Machine Learning. An Artificial Intelligence Approach. California: Morgan Kaufmann Publishers, Inc.

Mishra, A., (2018). Metrics to Evaluate your Machine Learning Algorithm. *Towards Data science – Medium.* Disponível em https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234. Acedido em 2020.

Mitchell, T. M. (1997). Machine learning. McGraw-Hill Science/Engineering/Math, 0070428077.

Neal, Richard D., *et al* (2001). Missed appointments in general practice: retrospective data analysis from four practices. *British Journal of General Practice*, Issue 51, 830-832.

Quinlan, J.R., (1993). C4.5 Programs for Machine Learning. 1-55860-238-0.

Rawl, A., (2018). Lyft, Hitch Health transportation pilot reduces no-shows by 27%. *South Carolina Alliance of Health plans*. Disponível em http://www.scalliance.org/lyft-hitch-health-transportation-pilot-reduces-no-shows-by-27/. Acedido em 2020.

Silltow, J., (2006). Data mining 101: Tools and Techniques. Disponível em https://iaonline.theiia.org/data-mining-101-tools-and-techniques. Acedido em 2020.

The Cost of No Shows. Defining the problem, understanding the impact, and reviewing the solutions. Crosschx blog. Disponível em http://www.crosschx.com/blog. Obtido em Jan 2018.

Witten, I. H., *et al* (2011). Data Mining: Practical Machine Learning Tools and Techniques. Third Edition. Morgan Kaufmanne.

Zhou, Z. (2012). *Ensemble methods: Foundations and algorithms.*