NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

DATA MINING NO TURISMO EM PORTUGAL

Análise Preditiva no Suporte à Tomada de Decisão

Pedro Filipe Soares Linheiro Galinha

Dissertação 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

LOMBADA MGI

2017

Título: Data Mining no Turismo em Portugal
Subtítulo: Análise Preditiva no Suporte à Tomada de Decisão

Pedro Filipe Soares Linheiro Galinha

MGI

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

DATA MINING NO TURISMO EM PORTUGAL:

ANÁLISE PREDITIVA NO SUPORTE À TOMADA DE DECISÃO

por

Pedro Filipe Soares Linheiro Galinha

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Gestão de Informação, Especialização em Gestão do Conhecimento e Business Intelligence

Orientador: Professor Doutor Roberto Henriques

Novembro 2017

DEDICATÓRIA

Para o Sasha,

Por ter sido quem nos une em todos os momentos.

AGRADECIMENTOS

Com o decorrer deste trabalho, muitas foram as pessoas que me apoiaram e às quais gostaria de agradecer.

Pela força, inspiração e incessante contribuição, um agradecimento especial para a Joana.

Aos meus amigos João Oliveira, Miguel Candeias, João Gavela, Sandra Coutinho, Pedro Pinheiro, João Farinha, Daniel Teófilo, Filipe Silva, Augusto Tosta e Lígia, pelo apoio incondicional que me deram na realização deste projeto.

Ao meu orientador, Professor Roberto Henriques, que me acompanhou ao longo deste trabalho.

Aos Serviços Académicos da NOVAIMS, particularmente à Rita, Ângela, Gisela, Raquel e Ana por serem os melhores serviços académicos do mundo e por me ajudarem sempre.

À minha família.

Muito obrigado a todos

RESUMO

Registou-se na última década um aumento significativo na procura de Portugal como destino turístico. A crescente angariação de dados dos consumidores por parte dos agentes turísticos representa uma oportunidade para extrair conhecimento e valor. A intenção por parte das políticas públicas do turismo está contemplada no plano estratégico para o setor até 2027 e as suas medidas visam promover a integração de políticas setoriais que influenciam a atividade do turismo e que assegurem estabilidade.

Sendo o setor do Turismo um dos que mais contribui para o desenvolvimento económico em Portugal e sendo a utilização de ferramentas analíticas e preditivas um fator decisivo na capacidade de potenciar esse desenvolvimento, pretende-se que este estudo forneça especificidade na relação entre gestão de informação e a sua mais-valia no contexto do turismo em Portugal.

Recorrendo às aplicações Google Scholar, Web of Science e NOVA Discovery, foram inseridas combinações de *keywords* com os termos (mineração de dados, turismo, análise preditiva) e foi possível verificar que a literatura específica disponível que aborde a relação entre métodos de análise preditiva, mineração de dados e conhecimento do consumidor para a área do turismo é ainda pouco significativa em Portugal. Desta forma é percetível a necessidade em promover a criação de literatura específica sobre a relação entre análise preditiva, estudo do comportamento e a sua aplicação ao setor do turismo em Portugal como fator decisivo na criação de inovação e competitividade, através do suporte que presta à tomada de decisão.

Neste sentido, a presente dissertação pretende apresentar uma prova de conceito que contribua para um maior conhecimento sobre a aplicação de técnicas de *Data Mining* e modelação preditiva para dados do turismo.

PALAVRAS-CHAVE

Data Mining; Análise Preditiva; Turismo de Portugal; Estratégia para o Turismo; Big Data

ABSTRACT

In the last decade there has been a significant increase in the demand for Portugal as a tourist destination. The growing collection of consumer data by tourism agents represents an opportunity to extract knowledge and value. The intention of the public policies of tourism is contemplated in the strategic plan for the sector until 2027 and its measures are aimed at promoting the integration of sectoral policies that influence the activity of tourism and that ensure stability.

Since the tourism sector is one of the main contributors to economic development in Portugal and the use of analytical and predictive tools is a decisive factor in the capacity to promote this development, it is intended that this study provides specificity in the relationship between information management and It´s added value in the context of tourism in Portugal.

Using combinations of keywords with the terms (data mining, tourism, predictive analysis) and using the Google Scholar, Web of Science and NOVA Discovery applications, it was possible to verify that the specific literature available that addresses the relationship between methods of predictive analysis, data mining and consumer knowledge for tourism is still not very significant in Portugal. This urges the need to promote the creation of specific literature on the relationship between predictive analysis, behavioral study and its application to the tourism sector in Portugal as a decisive factor in the creation of innovation and competitiveness through the support it provides to the outlet of decision-making.

In this sense, the present dissertation intends to present a proof of concept that contributes to a greater knowledge on the application of data mining techniques and predictive modeling for tourism data.

KEYWORDS

Data Mining; Predictive Analysis; Tourism in Portugal; Strategies for Tourism; Big Data

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

KDD	Knowledge Discovery in Databases (Descoberta de Conhecimento em Bases de Dados)
DM	Data Mining (Mineração de Dados)
BI	Business Intelligence (Inteligência de Negócios)
INE	Instituto Nacional de Estatística
ACP	Análise de Componentes Principais
ANN	Artificial Neural Networks (Redes Neuronais Artificiais)
EUA	Estados Unidos da América
TCA	Teoria dos Conjuntos Aproximados
SVM	Support Vector Machines (Máquina de Vetores de Suporte)
SOM	Self Organizing Maps (Mapas de Kohonen)

1. INTRODUÇÃO

1.1. ENQUADRAMENTO DO TEMA

Big Data é hoje um dos mais populares e mais frequentes termos utilizados para descrever o aumento exponencial e disponibilidade dos dados na era moderna, sendo espectável inclusivamente que num futuro próximo mantenha ou inclusivamente acelere o seu ritmo de crescimento (Hassani & Silva, 2015). Este termo refere-se a bases de dados cujo tamanho é tão grande em tamanho e complexidade que se torna inadequado para as ferramentas atuais capturar e processar dados num período de tempo aceitável (Snijders, Matzat, U., & Reips, U.-D., & Reips, 2012). Em *Big Data* existem constrangimentos no que diz respeito à análise, captura, procura, partilha, armazenamento, transferência, visualização e privacidade da informação e estes problemas necessitam novas tecnologias que permitam descobrir "valores escondidos" em bases de dados que são complexas e massivas (Hashem, Yaqoob, Anuar, Mokhtar, Gani, & Ullah Khan, 2015). A análise de bases de dados vastas possibilita criar novas oportunidades na sociedade moderna (Fan, Han, & Liu, 2014) uma vez que, estes novos repositórios de informação são tão vastos que providenciam aos investigadores, gestores e legisladores, as informações necessárias de forma a que consigam tomar decisões baseadas em números e análises, em vez de intuição ou experiência adquirida (Frederiksen, 2012), permitindo assim análises e tomadas de decisão com maior confiança e melhor eficiência operacional bem como redução de custos e riscos (De Mauro, Greco, & Grimaldi, 2015).

Segundo Benckendorff, Sheldon, & Fesenmaier (2014), a indústria do turismo caracteriza-se como sendo altamente competitiva e próspera em informações. Atualmente, essa competição intensa entre empresas, bem como as caraterísticas do mercado de turismo, suscitam que estas tenham a necessidade de se diferenciar. O uso de novas tecnologias de informação surge como uma forma de acompanhar a evolução do mercado, bem como, acompanhar a evolução do comportamento de compra dos seus clientes.

Considera-se como sendo atividade turística a deslocação de pessoas para outras regiões ou países com a propósito de vivenciar momentos de lazer, conhecer outras culturas ou visitar lugares específicos. A escolha de destinos de viagem por parte dos consumidores pode ser influenciada pela oferta e tendências existentes no momento da compra e sendo as tendências passageiras, surge a necessidade imperativa de permanente inovação e adaptação, de forma a garantir o crescimento.

Com base nesta necessidade, o investimento no turismo é assumido como uma prioridade governamental sendo reconhecido como um dos setores com maior crescimento nas últimas

décadas. Em Portugal, o setor do Turismo é o maior contribuidor para o crescimento económico e para a criação de emprego (Turismo de Portugal I.P. (TdP), 2017).

Torna-se assim imperativo conhecer melhor o turista, as suas vontades e preferências e qual a oferta disponível no mercado, sendo para esse efeito determinante a busca constante de ferramentas que ajudem na gestão da informação no setor.

Recorrendo à base tecnológica para responder ao desafio do conhecimento da atividade turística e da gestão dos recursos existentes em Portugal, é necessária a implementação de *Business Intelligence* orientada para o Turismo, aliando a Mineração de Dados e criação de modelos de Análise Preditiva como processo analítico projetado de forma a explorar grandes quantidades de dados, na detecção de padrões consistentes e/ou relacionamentos sistemáticos entre variáveis. A validação dos padrões detectados a novos subconjuntos de dados automaticamente, permitirão assim a simplificação do processo de tomada de decisão e a geração de novos "*insights"* que promovam a inovação, crescimento exponencial do setor e aumento da riqueza.

1.2. OBJETIVOS

Considerando o crescimento significativo da afluência turística em Portugal, torna-se cada vez mais importante identificar o perfil e prever o comportamento do consumidor turístico.

De forma a atingir os objetivos do estudo, foram definidos os seguintes objetivos específicos:

i) Conhecer a literatura aplicável referente à utilização de técnicas de *Data Mining* para o turismo;

ii) Identificar quais as variáveis com maior contribuição no comportamento do turista, nomeadamente no que diz respeito à disponibilidade do turista em adquirir viagens para Portugal;

iii) Exemplificar o uso de técnicas de *Data Mining* (modelação preditiva) para variáveis do turismo, de forma a prever o valor despendido por estadia;

iv) Avaliar a eficácia do modelo preditivo criado.

2. REVISÃO DA LITERATURA

O presente capítulo pretende fazer um enquadramento, evidenciando o estado da arte dos temas abordados ao longo da presente tese.

Uma das tendências mais marcantes da sociedade em que vivemos relaciona-se com os desenvolvimentos no campo da computação. Durante os últimos 50 anos estes desenvolvimentos têm alterado, de forma radical, muitas das atividades quotidianas. Esta verdadeira revolução, baseada numa evolução muito rápida das capacidades computacionais e da indústria do *software*, afecta todos os domínios da vida e do conhecimento humano. Uma das vertentes desta revolução diz respeito ao progresso nas tecnologias de recolha, organização e armazenamento de informação digital, que têm vindo a promover o aparecimento de bases de dados de grandes dimensões em todos os contextos da atividade e do conhecimento humano. É por isso natural que as organizações e investigadores se tenham deixado seduzir pela ideia de extrair informação relevante e de valor a partir destes repositórios de dados. Podemos mesmo supor que a resposta a muitos dos problemas da atualidade se pode encontrar "enterrado" nestas bases de dados. No entanto, para transformar estes dados em informação e conhecimento são necessárias metodologias e ferramentas apropriadas.

A presente revisão literária abordará os temas sobre a utilização de técnicas de *Data Mining* e modelação preditiva como geradoras da criação de "*insights*" que contribuam para a dinamização e desenvolvimento do setor turístico em Portugal.

2.1. O TURISMO EM PORTUGAL

O turismo em Portugal é hoje o principal motor da economia do país. O ano de 2016 ficou marcado por resultados históricos para o turismo nacional nos principais indicadores: dormidas, receitas, hóspedes, emprego e exportações, sendo mesmo considerado a maior atividade económica exportadora do país, com 16,7% das exportações. Entre 2005 e 2015 as receitas turísticas cresceram a uma taxa média anual de 6,3% (Figura 2.1). Estes resultados evidenciam que turismo tem capacidade para ser uma atividade sustentável ao longo do ano e para acrescentar valor, sendo para isso essencial a definição das metas que se querem atingir e o desenvolvimento das ações necessárias para tal. Neste sentido, a Estratégia Turismo 2027 foi desenhada para tornar Portugal num destino cada vez mais competitivo numa atividade em contínuo crescimento, atenta às mudanças internacionais e ao ambiente tecnológico.

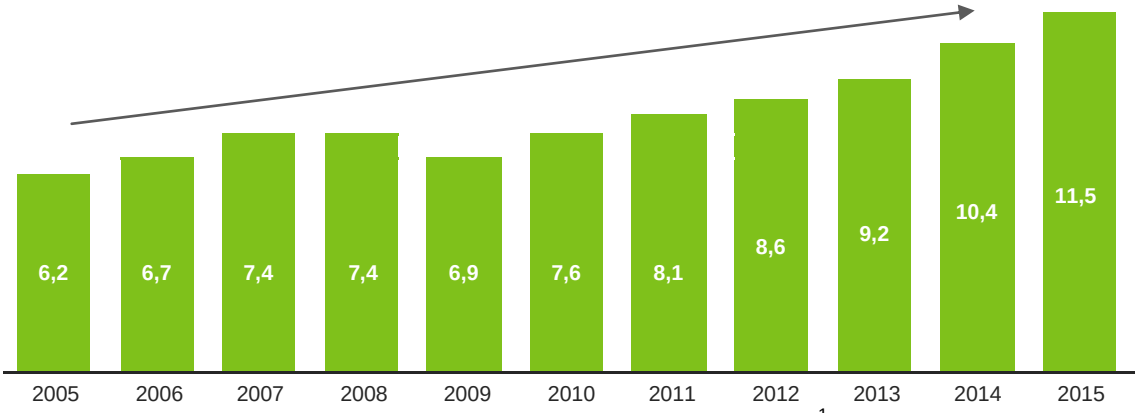

Figura 2.1 - Receitas Turísticas em valor e em % do PIB[1] (INE, 2016)

A implementação de medidas estratégicas e o investimento forte no setor, contribuíram para potenciar Portugal como um dos destinos internacionais preferidos. Entre 2005 e 2015, Portugal registou um crescimento médio anual superior ao dos concorrentes, sendo o segundo país com melhor desempenho na evolução das receitas turísticas (Tabela 2.1).

RECEITAS TURÍSTICAS INTERNACIONAIS	2005	2010	2015	TVMA 2005-2015
Malta	0,6	0,8	1,2	+7,4%
Portugal	**6,2**	**7,6**	**11,5**	**+6.3%**
Turquia	15,4	17,0	24,0	+4,5%
Marrocos	3,7	5,1	5,3	+3,7%
Croácia	5,9	6,1	8,0	+3,1%
Grécia	10,7	9,6	14,1	+2,8%
Espanha	40,00	41,2	50,9	+2,4%
Itália	28,5	29,3	35,6	+2,2%
França	35,4	35,5	41,4	+1,6%
Egito	5,5	9,4	5,5	+0,0%
Tunísia	1,7	2,0	1,2	-3,2%

Tabela 2.1 - Crescimento das receitas[2] (UNWTO, 2017)

O impacto do Turismo na economia nacional, no desenvolvimento de novas infraestruturas e serviços tem verificado uma evolução significativa nos últimos anos (Tabela 2.2), no entanto é vital para a sustentabilidade do setor no país, identificar as principais fragilidades, oportunidades e potencialidades para a próxima década, tanto no contexto interno como no ambiente externo.

[1] Valores em mil milhões de euros
[2] Comparativo internacional (valores em mil milhões de euros)

	2015	2016 PO	Variação
Dormidas	48,9 milhões	53,5 milhões	+ 4,6 milhões | + 9,4%
Receitas	11,5 mil milhões	12,7 mil milhões	+ 1,2 mil milhões | + 10,4%
Hóspedes	17,4 milhões	19,1 milhões	+1,7 mil milhões | + 9,7%
Exportações	15,4% do total de Exportações de bens e serviços do País	16,7% do total de Exportações de bens e serviços do País	+ 1,3%
Saldo da Balança Turística	7,8 mil milhões €	8,8 mil milhões €	+ mil milhões | + 12,8%
Emprego	280 mil	328 mil	+ 48 mil | + 17,1%

Tabela 2.2 – Variação das Receitas do Turismo em Portugal[3] (INE, 2015)

No contexto interno foi possível averiguar as seguintes fragilidades:

- Défice de informação sobre a oferta;
- Falta de conhecimento e de informação sobre a atividade turística;
- Falta de estruturação do produto.

No que diz respeito ao ambiente externo foram identificadas oportunidades na dinamização do setor relacionadas com:

- Alteração dos padrões de consumo e motivações;
- Crescimento do volume de informação recolhida sobre o consumidor.

Segundo a estratégia para o turismo em Portugal, as novas tendências internacionais terão impacto no setor e perspetivam acentuadas mudanças nos próximos anos. As mais significativas dizem respeito à crescente importância das Tecnologias de Informação e Comunicação como veículo condutor da Nova Economia. A pertinência da utilização de técnicas que possibilitem o acesso a informação e conhecimento é suportada pela necessidade de fazer face às exigências que se avizinham, suprimindo as fragilidades e potenciando as oportunidades.

[3] Emprego compreende alojamento, restauração, similares, agências de viagens/operadores turísticos e outros serviços de reservas.

2.2. *DATA MINING* E ANÁLISE PREDITIVA

O fenómeno da descoberta e tomada de decisão baseado em dados é hoje mais do que uma tendência. A mineração de dados refere-se à extração de conhecimento de um grande conjuntos de dados observados, de forma a descobrir o relacionamento insuspeito e os padrões escondidos nos dados, apresentado os mesmos de maneiras inovadoras, compreensíveis e úteis para os usuários (Adeniyi, Wei, & Yongquan, 2016).

A mineração de dados é também o processo de exploração e análise, por meios automáticos ou semiautomáticos, de grandes quantidades de dados e que possibilita descobrir padrões e regras significativas. Sendo um subcampo interdisciplinar da ciência da computação, envolve um processo computacional de descoberta de padrões em grandes conjuntos de dados. O objetivo deste processo avançado de análise recai na forma de extrair informações de um conjunto de dados e transformá-lo numa estrutura compreensível para uso posterior (Jain & Srivastava, 2013).

Tradicionalmente utilizado em áreas onde os dados abundam, a tarefa principal de mineração de dados consiste em identificar padrões dentro dos dados, com o propósito de extrair conhecimento. Para este fim, os métodos de mineração de dados utilizados como a análise de *Clusters*, *Link Analysis*, classificação e regressão, visam normalmente reduzir a quantidade de informações (ou dados) facilitando o reconhecimento de padrões (Zhu & Davidson, 2007).

Conway (2011), divulgou na sua pesquisa uma poderosa declaração sobre *Data Mining*: "A capacidade de adquirir dados - poder compreendê-los, processá-los, extrair valor deles, visualizá-los, comunicá-los - será uma técnica extremamente importante nas próximas décadas".

Segundo Bação (2009), mais importante do que estabelecer uma única definição de *Data Mining*, interessa reter uma noção geral do que se entende por *Data Mining*, compreender os contornos das suas fronteiras enquanto área de investigação e a forma como interage com outras áreas do conhecimento. A primeira observação a fazer é a de que o termo *Data Mining* é utilizado de forma muito diferente na literatura. Por vezes refere-se a todo o processo de "Descoberta de Conhecimento em Bases de Dados", onde se incluem todos os procedimentos que organização e preparação dos dados; outras apenas à fase específica de aplicação dos algoritmos. É de salientar que o termo "Descoberta de Conhecimento em Bases de Dados" (do inglês *Knowledge Discovery in Databases* - *KDD*) tem vindo a ganhar cada vez maior aceitação, especialmente na área académica, como forma de designar todo o processo que medeia entre o acesso aos dados digitais até à aplicação concreta e prática do conhecimento gerado no processo. No entanto, apesar destas ténues distinções, o facto é que "Descoberta de Conhecimento" e *Data Mining* são utilizados, por grande

parte dos autores, como sinónimos. São apresentadas, de seguida, várias definições de diferentes autores sobre *Data Mining*:

• *"DM is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"* – Fayyad et. all (1996)

• *"Data Mining is used to discover patterns and relationships in data, with emphasis on large observational databases"* – Friedman (1997)

• *"Data Mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner"* – Hand et. all (2001)

- *"Data Mining is the process of automatically discovering useful information in large data repositories"* – Tan, Steinbach & Kumar (2006)

É possível verificar uma grande sobreposição entre as diferentes definições apresentadas. Tipicamente, as relações e resumos extraídos por via do *Data Mining* são normalmente designados modelos ou padrões. Esta poderosa tecnologia com enorme potencial de crescimento, procura traduzir dados em informação e informação em conhecimento, que por sua vez proporciona a oportunidade de agir, sobre o real, racionalmente e com propriedade. A capacidade de prever é possível através da análise preditiva que é o ramo da mineração de dados que prevê tendências e comportamentos futuros, permitindo decisões pró-ativas e guiadas pelo conhecimento. O impacto da mineração de dados e da análise preditiva na sociedade tem potenciado também o recurso à automação e à necessidade de recorrer a máquinas que tenham a capacidade de aprender. *Machine Learning* constitui de uma forma geral, um conjunto de ferramentas que, em termos gerais, permitem "ensinar" os computadores a realizar tarefas, fornecendo exemplos de como elas devem ser feitas (Hertzmann & Fleet, 2012).

Galliers (1987) reforça a importância decisiva que o acesso à informação tem no processo de tomada de decisão. A necessidade de recorrer a ferramentas que possibilitem a criação de *insights* e conhecimento é salientada por Rascão (2004) que afirma que as Tecnologias de informação e de comunicação são definidas como o conjunto de conhecimentos, de meios materiais (infraestruturas) e de "*know how*", necessários à produção, comercialização e ou utilização de bens ou serviços relacionados com o armazenamento temporário ou permanente da informação, bem como o processamento e a comunicação da mesma. A capacidade de recolher grandes quantidades de informação e a consequente necessidade de agir, faz com que a capacidade analítica seja determinante na criação de valor e no suporte à tomada de decisão.

O'Brian (2004) afirmou que, para atender de forma eficiente a crescente demanda por informações de qualidade, os sistemas tiveram que evoluir de uma fase primária onde os processos eram apenas informatizados, para uma nova fase com um papel relevante no auxílio da tomada de decisão por meios preditivos. Os meios preditivos podem assim ser determinantes na tomada de decisão, assumindo particular destaque a capacidade de antecipar comportamentos e escolhas. Eric Siegel (2013) refere que através da análise preditiva, um computador analisando dados, literalmente aprende a prever o comportamento futuro de indivíduos.

Markus Hofmann (2013) realça que à medida que o fluxo de informação continua a aumentar, a necessidade de recorrer e dominar a mineração de dados e análise preditiva nunca foi maior. Afirma o autor que as técnicas e ferramentas providenciam insights dos dados sem precedentes, permitindo melhores decisões e capacidade de previsão e a derradeira solução na resolução de problemas cada vez mais complexos. A demanda por essa capacidade de prever nasceu da frustração com os sistemas BI (*Business Intelligence*), que ajudavam os executivos apenas a entender o que aconteceu enquanto eles necessitavam de ferramentas que conseguissem prever o que iria acontecer com os seus negócios (Monk & Wagner, 2013).

As empresas tomavam as suas decisões baseando-se no conhecimento e nas experiências de especialistas, o que acabava por influenciar as operações quotidianas. Algumas décadas atrás, uma série de técnicas estatísticas surgiu com a intenção de descobrir padrões de dados invisíveis ao olho humano. E visto que os dados são capturados num volume cada vez maior, estas técnicas tornaram-se indispensáveis para extrair valor a partir destes dados. É através da analítica que se torna possível produzir estatísticas e previsões confiáveis (Guazelli, 2012).

2.2.1. MÉTODOS UTILIZADOS EM *DATA MINING*

As técnicas de *Data Mining* são aplicadas em diversos campos devido a serem consideradas bastante adequadas ao desenvolvimento de métodos e modelos de análise.

Segundo (Berry & Linoff, 2004), as atividades mais comuns em *Data Mining* dividem-se em dois tipos de métodos: o preditivo e o descritivo. Nas atividades preditivas, são realizadas inferências nos dados existentes na base de dados de forma a prever valores relevantes, como por exemplo, utilizar padrões de consumo para prever correlações entre produtos e direcionar campanhas de marketing. No que diz respeito a atividades descritivas, os dados que constam na base de dados são classificados através da caracterização das suas propriedades gerais, descobrindo padrões e descrevendo os dados de forma a serem interpretados pelos utilizadores (Sondwale, 2015).

A modelação descritiva recorre principalmente a técnicas de *Data Mining* como Clustering, Regras de Associação, *Link Analysis* e Visualização. No que diz respeito à modelação preditiva, a Classificação e a Regressão são as técnicas mais frequentemente utilizadas. Desta forma, o processo de descoberta de conhecimento em bases de dados (*KDD*), consiste na utilização de métodos descritivos e preditivos do tipo *machine learning* que assentam na análise dos dados utilizando algoritmos que aprendem interactivamente a partir destes, fazendo com que os computadores encontrem conhecimento "escondido" sem serem explicitamente programados para encontrarem algo.

O processo *KDD* é iterativo e interativo, sendo constituído por etapas que se iniciam com a definição de objetivos e que evoluem posteriormente para a criação de um *target* no *Dataset,* limpeza e pré-processamento dos dados, transformação dos dados, escolha da técnica de *Data Mining,* dos algoritmos e interpretação dos padrões resultantes da mineração através do conhecimento gerado (Figura 2.2).

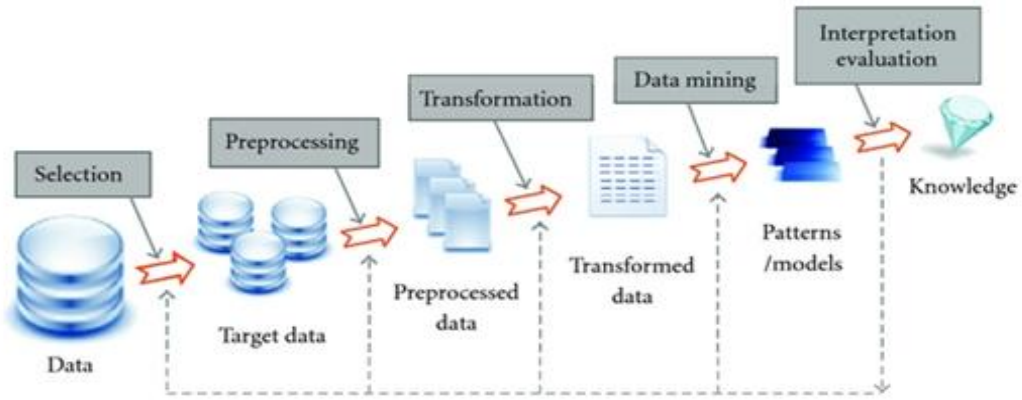

Figura 2.2 - Processo de Descoberta de Conhecimento, adaptado de Fayyad et al. (1996)

2.2.2. *MACHINE LEARNING*

A aprendizagem máquina ou *Machine Learning* pode realizar-se de forma supervisionada ou não supervisionada (*Supervised vs Unsupervised*). *Supervised Machine Learning* refere-se à procura de algoritmos que aprendam através de instâncias externas de forma a produzirem hipóteses, que posteriormente fazem previsões sobre instâncias futuras (Kotsiantis, 2007).

Neste tipo de aprendizagem existe um "agente externo" que avalia a resposta da rede ao padrão atual de *inputs*. As alterações dos pesos são calculadas de forma a que a resposta da rede tenda a coincidir com a do "agente externo". Pode ser dividida em Classificação se o atributo de classe for discreto ou em Regressão quando o atributo de classe é contínuo. Exemplos disso são as árvores de

decisão, classificador *Naive Bayes,* classificador *K-nearest neighbor,* classificação e métodos de regressão (linear e logística) (Figura 2.3). No método *Unsupervised learning* não existe um "agente externo". A rede necessita descobrir por si mesma as relações, padrões, regularidades ou categorias nos dados que lhe vão sendo apresentados e codificá-las nas saídas (Rojas, 1996).

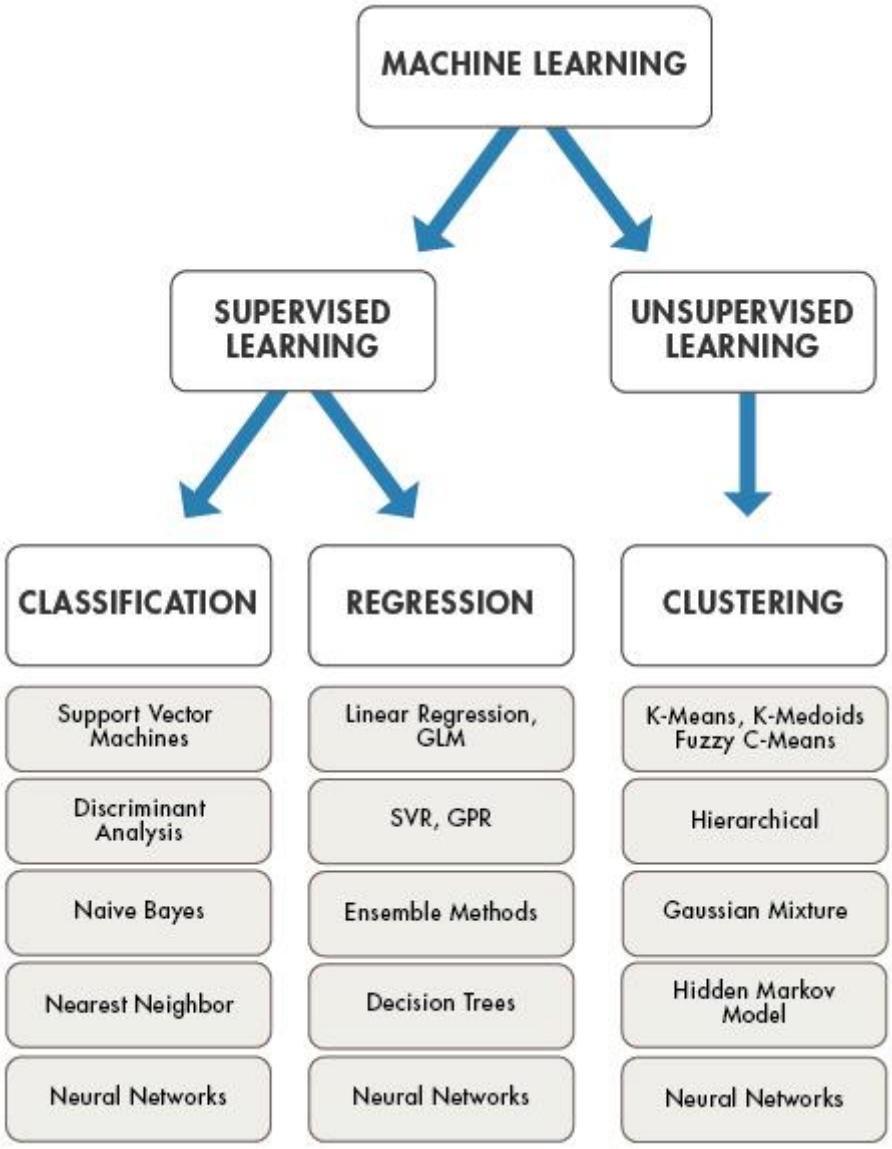

Figura 2.3 – Modelos e Algoritmos utilizados em *Machine Learning* (MathWorks, 2016)

2.3. *DATA MINING* APLICADO AO TURISMO

Diariamente, milhões de pessoas viajam por todo o mundo para negócios, férias, passeios ou outras razões. A quantidade de dinheiro gasto em ingressos, acomodações, comida, transporte e entretenimento traduz-se em valores exorbitantes. De acordo com o Conselho Mundial de Viagens e

Turismo, as viagens e o turismo representam atualmente aproximadamente 11% do Produto Interno Bruto (PIB) mundial (Werthner & Ricci, 2004).

O turismo é um negócio baseado na informação, onde existem dois tipos de fluxos de informação. Um desses fluxos de informação é direcionado para os consumidores ou turistas, sendo esta informação sobre bens que os turistas consomem, tais como ingressos, quartos de hotel, entretenimentos entre outros. O outro fluxo de informações segue uma direção inversa e consiste em informações agregadas sobre turistas para prestadores de serviços. Quando os dados agregados sobre os turistas são apresentados da maneira correta, analisados pelo algoritmo correto e colocados nas mãos certas, podem ser traduzidos em informações significativas para tomar decisões vitais por parte dos prestadores de serviços turísticos para aumentar a receita e lucros. A mineração de dados pode ser uma ferramenta muito útil para analisar dados relacionados ao turismo.

A indústria do turismo é hoje uma das principais utilizadoras das tecnologias de informação (Sheldon, 1997) . Os avanços tecnológicos afetam decisivamente os serviços e instalações desenvolvidas bem como a forma como elas são disponibilizadas e promovidas. Esta nova realidade influencia a estrutura organizacional e as relações entre clientes e prestadores de serviços (Olsen & Connolly, 1999).

Existe atualmente por parte dos legisladores, executivos, diretores de empresas e organizações governamentais, a necessidade em conhecer a relação entre a atividade turística e as preferências dos turistas. Este conhecimento visa promover o desenvolvimento de novas infraestruturas que possibilitem a obtenção de uma maior dinamização da atividade turística. Devido ao crescente volume de dados recolhidos, torna-se preponderante obter a capacidade de efetuar análises detalhadas de forma a auxiliar a tomada de decisões operacionais, táticas e estratégicas (Bose, 2009).

Devido a esta necessidade analítica, técnicas estatísticas formais foram progressivamente introduzidas no turismo. Verificou-se no entanto, que as técnicas estatísticas sofriam com a desvantagem dos vários pressupostos sobre distribuições dos dados que tinham que ser efectuadas, antes que qualquer análise pudesse ser realizada, sabendo que, se esses pressupostos fossem violados, não existiria garantia de que os resultados fossem válidos. Esta limitação nos métodos estatísticos, levou a que os investigadores atualmente necessitem de ferramentas de *Data Mining* que recorram a *Machine Learning*, de forma a tornar possível a análise de dados no turismo.

Segundo Buhalis (2002), o conhecimento integrado das características turísticas, imagens, atitudes e atributos de destino preferidos, devem ser usados para comercializar destinos turísticos com maior

facilidade. As cadeias hoteleiras podem assim recorrer à mineração de dados para criar campanhas de marketing, planear promoções sazonais, planear o tempo e a colocação das campanhas publicitárias, criar propaganda personalizada e definir quais os segmentos de mercado que estão a crescer mais rapidamente (Pyo , Uysal , & Chang, 2002).

Outro factor a ter em conta diz respeito à explosão do volume de dados existentes relativamente a viagens e ao turismo. A criação de sistemas centralizados de reservas e gestão de propriedades resultou na acumulação de uma enorme quantidade de dados no setor turístico e ao mesmo tempo, em maior acessibilidade a mais dados (Magnini, E.D., & Hodge, 2003).

Abordando o fenómeno do imparável crescimento do volume de dados, Sharma (2016) conclui que o *Data Mining* e a sua utilidade no turismo é preponderante, uma vez que permite tomar as melhores decisões de negócios. Salienta o autor que estas técnicas são decisivas a detectar tendências de vendas, desenvolver campanhas de marketing mais inteligentes e com maior precisão, bem como promover a fidelização de clientes. A capacidade de minerar dados pode também prever o valor possível de cada cliente e produzir informações que possibilitem uma melhor gestão da relação com o visitante (Kasavana & Knutson , 1999).

Desta forma, diferentes tipos de técnicas de *Machine Learning* (*Supervised Learning* ou *Unsupervised Learning*) podem ser usadas para analisar dados relacionados ao turismo. São descritas de seguida as diferentes técnicas de *Machine Learning* mais comuns na mineração de dados de turismo:

- Redes neurais artificiais (ANN) [4].

- Algoritmos de Kohonen (da sigla *SOM*, do inglês *Self Organizing Maps*) - Referem-se a algoritmos baseados em *ANN´s* que permitem a detecção de agrupamentos (*Clusters*). Este processo consiste em dividir objetos em grupos cujos membros são semelhantes de alguma forma (Han & Kamber, 2001). Embora existam muitos algoritmos de agrupamento, os mais utilizados são: "*exclusive clustering*" e "*distanced-based clustering"*. No algoritmo de "*exclusive clustering*" se um determinado dado pertence a um cluster definido, não pode ser incluído noutro *cluster*. Por sua vez no algoritmo "*distanced-based clustering"*, se dois ou mais objetos estiverem "fechados" de acordo com uma determinada distância, eles são agrupados no mesmo cluster.

- Teoria dos Conjuntos Aproximados (*TCA*) - A *TCA* é proposta para abordar o problema da incerteza e da imprecisão na classificação de objetos (Slowinski & Vanderpooten, 2000).

[4] Descrição de uma (ANN) disponível na página 19 deste documento.

Baseia-se na hipótese de que cada objeto está associado a algumas informações, e os objetos associados à mesma informação são semelhantes e pertencem à mesma classe. O primeiro passo na *TCA* é a discretização de atributos independentes onde os atributos numéricos são convertidos em atributos categóricos. O segundo passo é a formação de reduções que proporcionam a mesma qualidade de classificação que o conjunto original de atributos. O último passo é a classificação de dados desconhecidos com base em regras de decisão e reduções.

- Máquina de Vetores de Suporte (do inglês *Support Vector Machine* - *SVM*) - O SVM classifica um vetor de entrada em classes de saída conhecidas. Começa com vários pontos de dados de duas classes e obtém o *hiperplano* ideal que maximiza a separação das duas classes. Para dados não linearmente separáveis, utiliza o método *kernel* para transformar o espaço de entrada num espaço de recursos de alta dimensão, onde um *hiperplano* ideal linearmente separável pode ser construído (Figura 2.4). Exemplos de funções *kernel* incluem a função linear, função polinomial, função de base radial e função sigmoid (Chang & Lin, 2001).

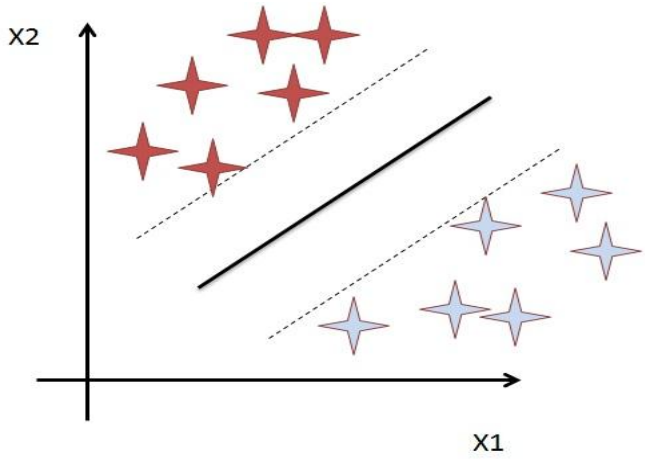

Figura 2.4 – Representação gráfica de *SVM* (Shouval, 2012)

A mineração de dados do turismo através da utilização de algoritmos específicos, possibilita capacitar o setor turístico com técnicas que permitem obter maior eficácia e conhecimento na tomada de decisões e consequentemente, maior crescimento económico.

2.4. PROCESSO DE PREVISÃO

O processo de previsão aplicado a qualquer área em estudo através da utilização de técnicas de *Data Mining* consiste num conjunto de tarefas que têm como principal função garantir a qualidade do

modelo selecionado. A aplicação de técnicas de previsão é utilizada quando se conhece a variável *target* (aprendizagem supervisionada ou *supervised learning*) através da criação de modelos paramétricos (Berry & Linoff, 2004).

A estratégia de investigação deste trabalho baseou-se no modelo preditivo SEMMA que possibilita o ajuste fino do modelo aos dados esparsos através de ciclos redundantes.

Este modelo propicia a aplicação da análise exploratória dos dados e técnicas de visualização, seleção e transformação das variáveis mais significativas, permitindo modelar as variáveis de modo a prever resultados e confirmar a precisão do modelo utilizado. O acrónimo SEMMA - Sample, Explore, Modify, Model, Assess é descrito da seguinte forma:

- *Sample* (amostra) – amostra dos dados representativa da população, é normalmente particionada nos conjuntos de treino, validação e teste;
- *Explore* (exploração) - ajuda a redefinir todo o processo de descoberta de conhecimento pela procura de tendências e anomalias nos dados através de técnicas estatísticas é o pré-processamento dos dados;
- *Modify* (modificar) - permite selecionar e transformar as variáveis tendo em vista o tipo de modelo utilizado, baseia-se na fase exploratória para manipular a informação. (converter variáveis nominais em numéricas);
- *Model* (modelo) - através de modelos de *data mining*, procura combinações na informação que melhor preveja o resultado esperado com o modelo;
- *Assess* (avaliação) - avalia os resultados obtidos através da medição da *performance* do processo de *data mining* permitindo optimizar os resultados pelo ajuste do modelo.

O modelo SEMMA permite fazer o ajustamento dos dados de acordo com vários métodos de regressão ou classificação simultaneamente, sendo que o modelo com melhor performance em termos de ajuste do erro quadrático médio será selecionado para fazer a previsão dos dados.

2.4.1. *Data Set*

Em *Data Mining* as bases de dados são grandes não apenas pelo número de registos, elas são grandes também pelo número de variáveis. No processo de constituição do *dataset*, deve-se atender à identificação das variáveis a constituir a base de dados que sejam representativas da caracterização do problema em estudo e à quantidade de exemplos necessários para a obtenção das regras do

modelo com qualidade. Atualmente é usual ter mais do que 1000 variáveis a caracterizar cada um dos registos armazenados. Com este nível de dimensionalidade (número de variáveis) é extremamente difícil encontrar conjuntos de registos ou de indivíduos que partilhem algum tipo de semelhança, este efeito é normalmente designado por *Curse of Dimensionality* (Bação, 2009).

Neste contexto, o espaço de *input*, definido pelo hiperespaço formado por todas as variáveis, torna-se massivo e qualquer tentativa de exploração encontra grandes dificuldades. Uma das necessidades mais prementes dos departamentos de *Data Mining* das empresas prende-se com o domínio de metodologias "inteligentes" para a redução do espaço de *input*. A natureza secundária dos dados também levanta questões quanto à qualidade e adequação dos dados para os objetivos a atingir. É frequente que as bases de dados utilizadas possuam erros de medição, valores extremos (*outliers*) e valores omissos. Todos estes constituem problemas que o analista tem que defrontar antes de poder iniciar a aplicação dos algoritmos. A não-estacionaridade, o enviesamento na seleção das amostras, a dependência entre observações, são apenas mais alguns dos problemas com os quais há que lidar (Bação, 2009).

A redução do espaço de *input* é assim crucial sendo a redução do número de variáveis possível através da seleção das que têm maior capacidade discriminativa preterindo as que possuem uma relação espúria entre *input/output*, identificação das variáveis correlacionadas, redundantes e irrelevantes, aplicação de médias ou de outras medidas estatísticas, efetuando análise das componentes principais (ACP), etc. A normalização dos dados para uma representação espacial em escalas equivalentes, assim como, a criação de novas variáveis (criação de indicadores através de rácios com variáveis existentes) que resultam do *know-how* do analista sobre o problema em estudo, contribuem igualmente para garantir um adequado *dataset* a utilizar no processo de *Data Mining* (Hand, 1998).

2.4.2. Criação do Modelo Preditivo

A modelação preditiva engloba as três primeiras classes da tipologia de Berry e Linhof (1997) (classificação, estimação e predição). Na modelação preditiva o objetivo é "aprender" um critério de decisão que nos permita atribuir valores a exemplos novos e desconhecidos. A modelação preditiva engloba a classificação e a regressão. A diferença entre classificação e regressão reside no facto de a primeira produzir valores discretos (sim/não, 0/1) e a segunda valores contínuos. De uma forma genérica é possível afirmar que a modelação preditiva procura desenvolver um modelo que permita prever resultados de um fenómeno de interesse. Este modelo servirá então para analisar as

características de um novo elemento e associá-lo a uma, de entre um conjunto de classes pré-definidas (classificação), ou a atribuir-lhe um determinado valor contínuo (regressão). Assim, o processo decorre em duas fases: a fase de aprendizagem e a fase de predição (Figura 2.5). Durante a fase de aprendizagem o algoritmo escolhido extrai conhecimento a partir dos exemplos de treino. Na fase de predição o conhecimento extraído é utilizado para classificar os novos exemplos (o termo classificar é usado no sentido lato que engloba a atribuição de uma classe ou valor contínuo). Na fase de aprendizagem, os exemplos de treino são registos para os quais se possui não só o valor das variáveis independentes como também o verdadeiro valor da variável dependente.

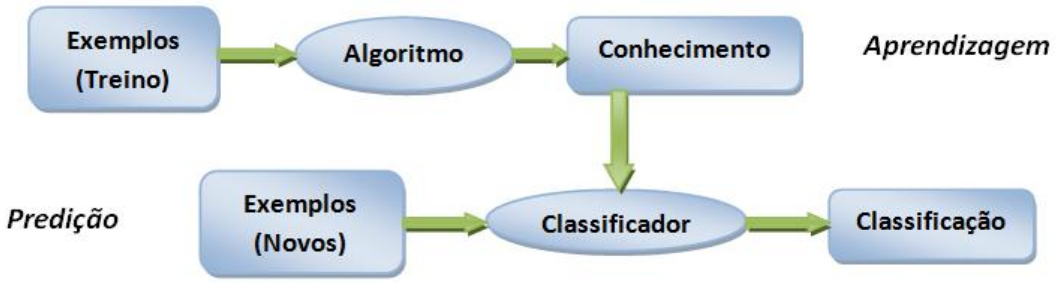

Figura 2.5 - Processo de modelação preditiva, *adaptado de* Bação (2009)

Segundo Bação (2009), a forma de avaliação dos modelos preditivos depende do tipo específico de modelação. Assim, se estamos a desenvolver um modelo de classificação então a forma mais intuitiva de medir a qualidade do modelo consiste em contabilizar o número de vezes em que o modelo se enganou na sua previsão. No caso de um modelo preditivo de regressão a avaliação da qualidade do modelo deverá contemplar outro tipo de medidas, uma vez que o interesse não está na contabilização do número de falhas, mas sim na proximidade da previsão ao verdadeiro valor. Assim, uma das formas mais comuns de avaliar a qualidade de modelos preditivos de regressão consiste em considerar a média dos quadrados dos desvios das previsões do modelo em relação ao verdadeiro valor de cada indivíduo (Figura 2.6).

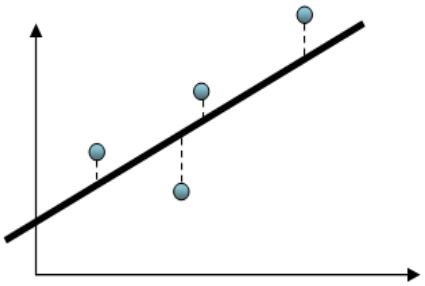

Figura 2.6 - Desvios das previsões face aos verdadeiros valores do modelo de regressão, *adaptado de* Bação (2009)

Existem inúmeras ferramentas que podem ser utilizadas na modelação preditiva, sendo as mais populares a regressão logística, as árvores de decisão e as redes neuronais.

2.4.3. Algoritmos utilizados na Análise Preditiva

Um algoritmo de mineração de dados é um conjunto de heurísticas e cálculos que criam um modelo de mineração a partir de dados. Para criar um modelo, o algoritmo analisa primeiro os dados fornecidos, o que é necessário para tipos específicos de padrões ou tendências. O algoritmo usa os resultados dessa análise para definir os parâmetros óptimos para a criação do modelo de mineração, aplicando posteriormente esses parâmetros em todo o conjunto de dados de forma a extrair padrões e estatísticas detalhadas. Várias técnicas básicas usadas na mineração de dados descrevem o tipo de operação de mineração e recuperação de dados:

• Os algoritmos de classificação efetuam predições numa ou mais variáveis discretas, com base nos outros atributos do conjunto de dados.

• Os algoritmos de regressão efetuam predições numa ou mais variáveis contínuas, como lucro ou perda, com base noutros atributos do conjunto de dados.

• Os algoritmos de associação encontram correlações entre diferentes atributos num conjunto de dados. A aplicação mais comum deste tipo de algoritmo remete para a criação de regras de associação, que podem ser usadas em análises *basket market*.

Algoritmos de associação

A associação (ou relação) é provavelmente a técnica de mineração de dados mais conhecida e também a mais direta.

Existe neste caso, uma correlação simples entre dois ou mais items, muitas vezes do mesmo tipo, de forma a possibilitar a identificação de padrões. Um modelo de associação consiste numa série de conjuntos de items e as regras que descrevem como esses items são agrupados em conjunto dentro dos casos. As regras que o algoritmo identifica podem ser usadas de forma a efetuar predições.

A regra de associação pode ser utilizada em análises *basket market* em hotéis, companhias aéreas e outros serviços, potenciando a criação de parcerias e alianças (Dev, Klein, & Fisher, 1996).

Algoritmos de classificação

A classificação é o processo de encontrar um modelo (ou função) que descreva e faça distinções em classes de dados ou conceitos. O modelo é derivado com base na análise de um conjunto de dados de treino (Os objetos de dados para os quais a classe rótulos são conhecidos). O modelo é usado para prever o rótulo da classe dos objetos para os quais o rótulo da classe não é conhecido. O modelo derivado pode ser representado em várias formas, tais como Regras de Classificação (ou seja, *IF ...THEN rules*), árvores de decisão, fórmulas matemáticas ou redes neuronais. A classificação pode ser usada para construir um perfil do tipo de cliente, item ou objeto, descrevendo vários atributos para identificar uma classe específica. A classificação prevê rótulos categóricos (discretos, não ordenados), enquanto que a análise de regressão se foca numa metodologia estatística mais utilizada para predição numérica.

- **Árvores de decisão**

O algoritmo de árvores de decisão é um algoritmo de classificação e regressão. As Árvores de Decisão são um método não paramétrico que divide consecutivamente um conjunto alargado de dados em subconjuntos aplicando regras simples que promovem a homogeneidade destes atendendo a determinada variável *target* (Berry & Linoff, 2004).

As Árvores de Decisão consistem numa abordagem de *top-down* para seleção dos atributos que constituem as regras do modelo, com sucessiva divisão do *dataset* de treino até formar grupos com características homogéneas (Hand D. J., 1998).

Numa representação gráfica (Figura 2.7), este procedimento consiste em “nós” que representam testes aos atributos e em “ramos” que representam as respostas aos testes, formando desta forma uma árvore enraizada, o que significa uma árvore direcionada com um nó chamado "raiz" que não possui entradas. Todos os outros “nós” possuem somente uma entrada. Um “nó” com “ramos” de saída é chamado de nó interno ou de teste. Todos os outros “nós” são chamados de folhas (também conhecidos como nós terminais ou de decisão) e representam um conjunto homogéneo de dados (Hand D. J., 1998). Numa árvore de decisão, cada nó interno divide o espaço da instância em dois ou mais subespaços de acordo com uma determinada função discreta dos valores dos atributos de entrada. No caso mais simples e mais frequente, cada teste considera um único atributo, de modo a que o espaço da instância seja particionado de acordo com o valor do atributo. No caso de atributos numéricos, a condição refere-se a um intervalo (Ben-Gal, Maimon e Rokach, 2005).

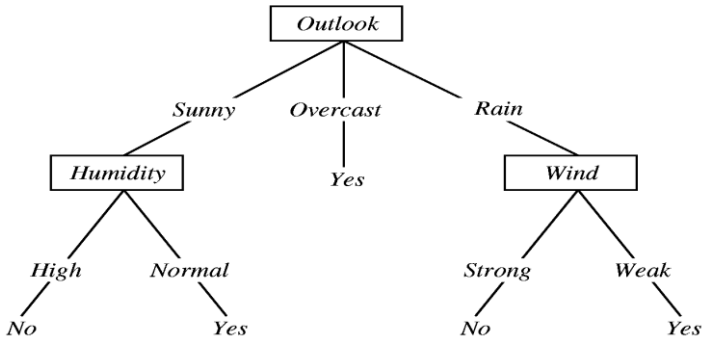

Figura 2.7 - Representação gráfica de uma *Decision Tree, adaptado de* (Hand,1998)

Este processo de divisão implica a utilização de medidas adequadas à seleção dos atributos com maior capacidade de descriminação. (Hand D. J., 1998) refere que da aplicação das medidas de seleção de atributos adequada, resulta o critério de divisão de cada nó da árvore em novos ramos (*subsets*) que permite identificar o melhor atributo a partir do qual deve ser feita a divisão em cada nó e os ramos que dele devem crescer. A divisão sucessiva dos vários *subsets* acaba quando os terminais do grafo apresentarem grupos de classes o mais homogéneos possível. O grau de pureza ideal será alcançado quando existir somente uma única classe nos *subsets* finais.

- **Neural Networks**

Uma rede neural é uma representação artificial do cérebro humano que tenta simular o seu processo de aprendizagem. Uma rede neural artificial (ANN) é conhecida como rede neural. Uma rede neural

artificial é um sistema de informação eficiente cujas características se assemelham a uma rede neural biológica. Nas ANN´s, o comportamento coletivo é caracterizado pela sua capacidade de aprender, de lembrar e pelo seu padrão de treino semelhante ao de um cérebro humano (Nanda, Tripathy, Nayak, & Mohapatra , 2013).

As ANN´s são modelos preditivos não-lineares que aprendem através do treino (Jain & Srivastava, 2013). São compostos por neurónios interconectados. Cada neurónio recebe um conjunto de entradas. Cada entrada é multiplicada por um peso. A soma de todas as entradas ponderadas determina o nível de ativação (Figura 2.8). Um algoritmo muito poderoso usado no treinamento de ANN´s tem o nome de "*back-propagation*". Neste caso específico, os pesos da conexão são ajustados iterativamente de forma a minimizar o erro com base na diferença entre as saídas desejadas e reais.

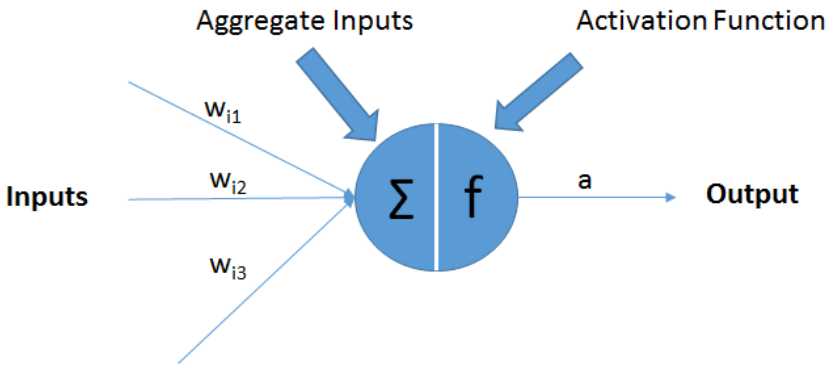

Figura 2.8 – Modelo de funcionamento das Redes Neuronais, *adaptado de* Berry & Linoff (2004)

Algoritmos de análise de sequência

O *Sequential Pattern Mining* permite encontrar padrões sequenciais em bases de dados. Descobre frequentemente subsequências como padrões de uma base de dados de sequências. Com enormes quantidades de dados continuamente coletados e armazenados, muitas indústrias estão interessadas em minerar padrões sequenciais de seu banco de dados. O padrão sequencial de mineração é um dos métodos mais conhecidos e possui aplicações abrangentes, incluindo análise de web log´s e análise de padrões de compra dos clientes. Esta análise permite entender os interesses dos clientes, satisfazer os seus interesses e acima de tudo prever as suas necessidades.

3. METODOLOGIA

Neste capítulo procede-se à apresentação da metodologia utilizada na presente dissertação. Em qualquer trabalho académico é fundamental utilizar-se uma metodologia coerente e adequada e que cumpra os objetivos propostos no trabalho.

Assim, a metodologia utilizada neste projeto é baseada no Design Quantitativo uma vez que esta metodologia tem como objetivo a compreensão do fenómeno baseado em evidências empíricas. Segundo Given, (2008), Design Quantitativo é definido como um processo empírico sistemático de investigação do fenómeno observável através de técnicas estatísticas, matemáticas ou computacionais. Tem como objetivo desenvolver modelos matemáticos e estatísticos que permitam uma melhor compreensão do fenómeno.

O modelo preditivo foi desenvolvido recorrendo ao *software SAS Enterprise Miner 14.1.*

Pretendeu-se com a seguinte metodologia, exemplificar o uso de técnicas de Data Mining e análise preditiva para variáveis do turismo, de forma a prever o gasto médio dos visitantes.

3.1. FONTE DE DADOS

Este trabalho utilizou uma amostra composta por dados estatísticos do turismo (INE, 2016) e (PORDATA, 2016). Foram consideradas as seguintes variáveis: (idade, sexo, rendimento anual, estado civil, nível de instrução, data da estadia, meio de transporte utilizado, número de viagens por ano, número de filhos, despesa média por viagem, despesa média diária, tipo de estabelecimento e nacionalidade). A base de dados foi criada no formato *Excel (Microsoft Office)*.

3.2. DESCRIÇÃO DOS DADOS

A base de dados é composta pelos campos abaixo descritos (Tabela 3.1):

Variável	Conteúdo
Custid	ID do cliente
DepVar	Variável Dependente
Despesa_media_diaria	Quantia média gasta diariamente
Despesa_media_Viagem	Despesa média por Viagem
Data_Registo	Data de registo do Cliente
Estado_Civil	Estado Civil do Cliente

Filhos	Número de filhos do agregado
Idade	Ano de Nascimento do Cliente
Meio_Transporte_Utilizado	Meio de Transporte Utilizado
Nacionalidade	Nacionalidade do Cliente
Nivel_Instrução	Nível de Escolaridade do Cliente
NumViagensAno	Número de Viagens efectuadas (últimos 12 meses)
Vencimento_Anual	Rendimento anual do agregado do Cliente
Dura_o_Estadia	Duração da Estadia do Cliente
Motivo_Viagem	Motivo da Viagem do Cliente
Sexo	Género do Cliente
Tipo_Estabelecimento	Tipo de Estabelecimento escolhido pelo Cliente

Tabela 3.1 – Variáveis do Dataset

3.3. CARACTERIZAÇÃO DAS VARIÁVEIS

A partir do nó inicial que contém a Base de Dados foi possível identificar o papel e o nível de cada uma das variáveis (Tabela 3.2):

Name	***Role***	***Level***	***Report***	***Drop***
Custid	*ID*	*Nominal*	*No*	*No*
DepVar	*Target*	*Binary*	*No*	*No*
Despesa_media_diaria	*Input*	*Interval*	*No*	*No*
Despesa_media_Viagem	*Input*	*Interval*	*No*	*No*
Data_Registo	*Time ID*	*Interval*	*No*	*No*
Estado_Civil	*Input*	*Nominal*	*No*	*No*
Filhos	*Input*	*Interval*	*No*	*No*
Ano_Nascimento	*Input*	*Interval*	*No*	*No*
Meio_Transporte_Utilizado	*Input*	*Nominal*	*No*	*No*
Nacionalidade	*Input*	*Nominal*	*No*	*No*
Nivel_Instrução	*Input*	*Nominal*	*No*	*No*
NumViagensAno	*Input*	*Interval*	*No*	*No*
Vencimento_Anual	*Input*	*Interval*	*No*	*No*
Dura_o_Estadia	*Input*	*Interval*	*No*	*No*
Motivo_Viagem	*Input*	*Nominal*	*No*	*No*
Sexo	*Input*	*Nominal*	*No*	*No*
Tipo_Estabelecimento	*Input*	*Nominal*	*No*	*No*

Tabela 3.2 – Papel e Nível da Variáveis

3.4. ESTATÍSTICAS DAS VARIÁVEIS ORIGINAIS

A análise das estatísticas gerais das variáveis foi efetuada através do nó "StatExplore". Estes valores, combinados com os dados gráficos constantes nos nós "MultiPlot" e "Variable Clustering", possibilitaram retirar informações importantes utilizadas nas fases de Preparação e Pré-Processamento de Dados.

A base de dados é constituída por 13 variáveis de Input, sendo que 7 delas são intervalares e as restantes de classe. Assim sendo, foi necessário analisar as estatísticas de cada grupo de variáveis separadamente.

O facto de se tratar de uma modelação preditiva implica que a análise das variáveis independentes seja efetuada tendo também consideração os resultados da variável dependente.

3.4.1. Variáveis Intervalares

A análise das variáveis intervalares (Tabela 3.3) demonstrou que não existem valores omissos em nenhuma delas. A análise das estatísticas gerais permitiu tirar algumas ilações acerca de potenciais *outliers*. Por exemplo, os valores máximos das variáveis "Despesa_media_diaria" e "NumViagensAno" são, respectivamente, 199 e 16. Estes valores devido ao seu afastamento da média levarão a que sejam tratados como *outliers*.

Variáveis Intervalares	Target (VarDep)	N Omissos	Média	Desvio	Mediana	Mínimo	Máximo
Despesa_media_diaria	0	0	26.2666	38.7897	9	0	199
Despesa_media_diaria	1	0	36.1219	43.2929	22	0	198
Vencimento_Anual	0	0	52650.69	24251.3	51873	1532	162934
Vencimento_Anual	1	0	59674.94	24163.8	63848	7500	102692
Filhos	0	0	0.4555	0.5581	0	0	3
Filhos	1	0	0.4108	0.5261	0	0	2
NumViagensAno	0	0	2.3600	2.1437	2	0	16
NumViagensAno	1	0	2.4986	2.2451	2	0	13
Despesa_media_Viagem	0	0	1259.890	1000.073	990	19	5000
Despesa_media_Viagem	1	0	1286.198	1012.08	960	230	5000
Dura_o_Estadia	0	0	19563.43	199.5432	19570	19203	19903
Dura_o_Estadia	1	0	19518.50	199.8804	19502	19205	19901
Ano_Nascimento	0	0	1968.733	11.9232	1970	1941	1996
Ano_Nascimento	1	0	1969.586	13.1923	1972	1942	1993

Tabela 3.3 – Variáveis Intervalares

3.4.2. Variáveis de Classe

Através da análise das variáveis de classe consideradas (Tabela 3.4), foi possível verificar que não existem valores omissos.

Relativamente ao total dos clientes em análise e ainda sem ter em conta a variável target, foi possível chegar a algumas conclusões através da análise das frequências relativas das variáveis:

- 59% são mulheres;
- 43% são casados;
- 43% viajaram para Portugal por motivos de Lazer;
- 42% preferiram hotéis;
- 37% utilizaram o automóvel como meio de transporte.

Variável	Tipo	Frequência	%
Estado_Civil	Casado	929	43
Estado_Civil	Casado	113	32
Estado_Civil	Divorciado	187	9
Estado_Civil	Divorciado	46	13
Estado_Civil	Solteiro	394	18
Estado_Civil	Solteiro	104	29
Estado_Civil	União de Facto	567	26
Estado_Civil	União de Facto	73	21
Estado_Civil	Viúvo	70	3
Estado_Civil	Viúvo	17	5
Meio_Transporte_Utilizado	Autocarro	269	13
Meio_Transporte_Utilizado	Autocarro	43	12
Meio_Transporte_Utilizado	Automóvel	811	38
Meio_Transporte_Utilizado	Automóvel	127	36
Meio_Transporte_Utilizado	Avião	545	25
Meio_Transporte_Utilizado	Avião	81	23
Meio_Transporte_Utilizado	Barco	88	4
Meio_Transporte_Utilizado	Barco	16	5
Meio_Transporte_Utilizado	Comboio	433	20
Meio_Transporte_Utilizado	Comboio	86	24
Motivo_Viagem	Lazer	892	42
Motivo_Viagem	Lazer	139	39
Motivo_Viagem	Profissional	365	17

Motivo_Viagem	Profissional	62	18
Motivo_Viagem	Religião	85	4
Motivo_Viagem	Religião	19	5
Motivo_Viagem	Saúde	173	8
Motivo_Viagem	Saúde	34	10
Motivo_Viagem	Visita Familiar	632	29
Motivo_Viagem	Visita_Familiar	99	28
Nacionalidade	Alemanha	173	8
Nacionalidade	Alemanha	20	6
Nacionalidade	Angola	55	3
Nacionalidade	Angola	9	3
Nacionalidade	Argentina	55	3
Nacionalidade	Argentina	9	3
Nacionalidade	Austrália	57	3
Nacionalidade	Austrália	7	2
Nacionalidade	Brasil	53	2
Nacionalidade	Brasil	11	3
Nacionalidade	Bélgica	50	2
Nacionalidade	Bélgica	14	4
Nacionalidade	Canadá	54	3
Nacionalidade	Canadá	10	3
Nacionalidade	China	115	5
Nacionalidade	China	13	4
Nacionalidade	Colômbia	52	2
Nacionalidade	Colômbia	12	3
Nacionalidade	Dinamarca	56	3
Nacionalidade	Dinamarca	8	2
Nacionalidade	EUA	116	5
Nacionalidade	EUA	12	3
Nacionalidade	Espanha	105	5
Nacionalidade	Espanha	23	7
Nacionalidade	França	275	13
Nacionalidade	França	46	13
Nacionalidade	Grécia	54	3
Nacionalidade	Grécia	10	3
Nacionalidade	Holanda	107	5
Nacionalidade	Holanda	21	6
Nacionalidade	Itália	58	3
Nacionalidade	Itália	6	2
Nacionalidade	Japão	109	5
Nacionalidade	Japão	20	6

Nacionalidade	Luxemburgo	57	3
Nacionalidade	Luxemburgo	7	2
Nacionalidade	México	54	3
Nacionalidade	México	10	3
Nacionalidade	Noruega	108	5
Nacionalidade	Noruega	20	6
Nacionalidade	Reino Unido	112	5
Nacionalidade	Reino Unido	17	5
Nacionalidade	República Checa	53	2
Nacionalidade	República Checa	11	3
Nacionalidade	Rússia	55	3
Nacionalidade	Rússia	9	3
Nacionalidade	Suíça	58	3
Nacionalidade	Suíça	6	2
Nacionalidade	Suécia	56	3
Nacionalidade	Suécia	8	2
Nacionalidade	África do Sul	50	2
Nacionalidade	África do Sul	14	4
Nivel_Instrução	12º Ano	1037	48
Nivel_Instrução	12º Ano	180	50
Nivel_Instrução	Doutoramento	475	22
Nivel_Instrução	Doutoramento	67	19
Nivel_Instrução	Ensino Básico	58	3
Nivel_Instrução	Ensino Básico	4	1
Nivel_Instrução	Licenciado	218	10
Nivel_Instrução	Licenciado	36	10
Nivel_Instrução	Mestrado	359	17
Nivel_Instrução	Mestrado	66	19
Sexo	Homem	872	41
Sexo	Homem	143	41
Sexo	Mulher	1275	59
Sexo	Mulher	210	59
Tipo_Estabelecimento	Alojamento Local	645	30
Tipo_Estabelecimento	Alojamento Local	99	28
Tipo_Estabelecimento	Campismo	306	14
Tipo_Estabelecimento	Campismo	44	12
Tipo_Estabelecimento	Hostel	301	14
Tipo_Estabelecimento	Hostel	52	15
Tipo_Estabelecimento	Hotel	895	42
Tipo_Estabelecimento	Hotel	158	45

Tabela 3.4 – Variáveis de Classe

3.4.3. Valor das Variáveis

O gráfico "Variable Worth" (Figura 3.9) permite compreender qual o valor que cada variável assume na definição da variável dependente " DepVar". Neste caso específico, as variáveis que têm a maior contribuição e capacidade discriminativa são "Vencimento_Anual" e "Despesa_media_diaria".

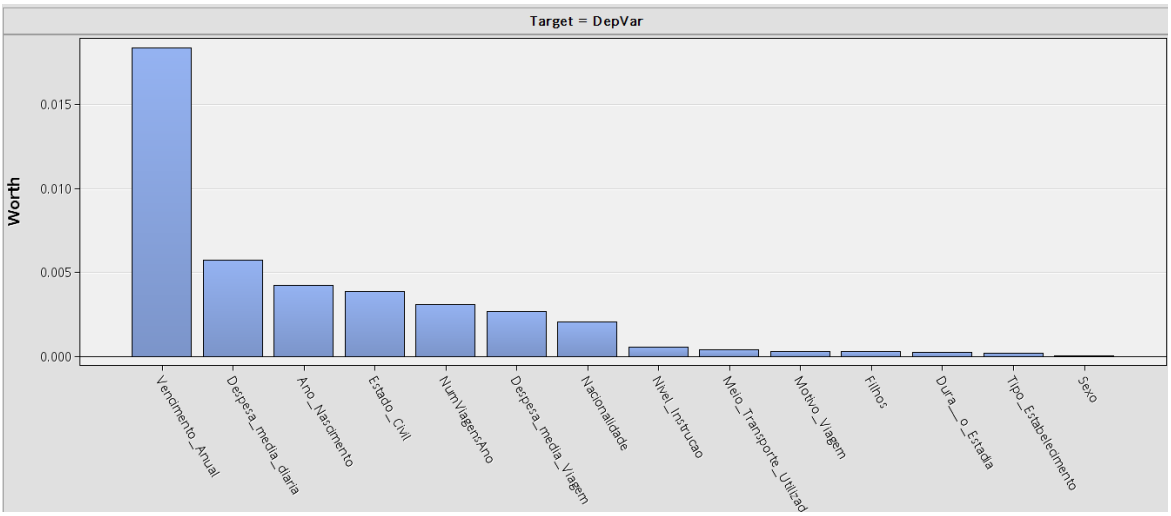

Figura 3.9 - Análise das variáveis a integrar o modelo preditivo através do critério *Worth*

Por sua vez as variáveis que têm possuem uma menor capacidade discriminativa da variável target são "Sexo", "Tipo_Estabelecimento", "Filhos", "Duração_Estadia", "Motivo_Viagem", "Meio_Transporte_Utilizado" e "Nivel_instrução".

Importance	***Variable***	***Worth***
1	Vencimento_Anual	0,0184
2	Despesa_media_diaria	0,0058
3	Ano_Nascimento	0,0039
4	Estado_Civil	0,0038
5	NumViagensAno	0,0031
6	Despesa_media_Viagem	0,0026
7	Nacionalidade	0,0020
8	Nivel_instrução	0,0005
9	Meio_Transporte_Utilizado	0,0004
10	Motivo_Viagem	0,0003
11	Duração_Estadia	0,0003
12	Filhos	0,0003
13	Tipo_Estabelecimento	0,0002
14	Sexo	0,0001

Tabela 3.4 – Importância das Variáveis

3.5. PREPARAÇÃO DE DADOS

A preparação dos dados consiste na "limpeza" dos dados originais. Inclui procedimentos como a remoção ou não de dados inconsistentes ou *outliers*, o tratamento de valores omissos e a conversão de *inputs* não numéricos em formato numérico.

3.5.1. Remoção de Outliers

Os *outliers* definem-se como pontos que se situam fora da região normal de interesse do espaço de *input*. Podem representar situações fora do vulgar mas que estão corretas, no entanto podem igualmente corresponder a medições incorretas que prejudicam a performance do modelo. São valores extremos que se situam nas "caudas" dos histogramas, sendo detectáveis pela distância e frequência relativa dos seus valores em relação à distribuição.

A análise dos histogramas das variáveis através do nó "*MultiPlot*", permitiu a detecção de *outliers* em "Vencimento_Anual", "Despesa_media_Viagem", "Despesa_media_diaria" e "NumViagensAno". Estas variáveis apresentam alguns valores bastante afastados na distribuição que conduzem ao enviesamento das suas medidas de tendência central. Foi selecionado o método manual na remoção dos *outliers* ("*user specified*") e foram também alterados os limites inferiores negativos das variáveis intervalares para 0.

Variável	Mínimo	Máximo	Método de Filtro
Vencimento_Anual	0	105000	Manual
Despesa_media_Viagem	0	3500	Manual
Despesa_media_diaria	0	100	Manual
NumViagensAno	0	10	Manual

Tabela 3.5 – Variáveis e valores de Filtro

Foram retirados 60 registos relativos a *outliers*, os quais representam **2,51%** da população total. Este número de registos filtrados é aceitável de forma a evitar o enviesamento na modelação preditiva.

A análise da Base de Dados não revelou a existência de dados inconsistentes ou incorretos nas variáveis, não sendo assim necessário o tratamento de dados no nó "*Transform Variables*", de forma a transformar as variáveis existentes.

3.5.2. Pré-Processamento de Dados

A fase do Pré-processamento de dados tem como objetivo facilitar e simplificar o problema sem excluir ou danificar informação relevante para a modelação e para o entendimento do problema. Trata-se de uma etapa revestida de alguma complexidade, já que por um lado se pretende reduzir a informação a utilizar, por outro lado, não se pretende eliminar a informação relevante que ajude a compreender o problema.

3.5.3. Redução do Espaço de Input

A dimensão do espaço de *input* aumenta exponencialmente relativamente à dimensionalidade do problema. Se o número de variáveis de *input* for demasiado elevado comparativamente com os dados disponíveis para proceder à modelação, será difícil fazê-lo de uma forma precisa. Deste modo, deverá ser equacionada uma redução do espaço de *input* para melhorar a performance dos modelos. A redução de dimensionalidade permite de igual forma reduzir a complexidade do problema, não considerando variáveis que não trazem qualquer valor acrescentado para a sua resolução. Para compreender quais as variáveis mais relevantes na modelação, é necessário ter em consideração dois conceitos fundamentais:

Relevância – Importância para resolver o problema em questão. Capacidade que a variável tem para determinar a variável target;

Redundância – Duas variáveis podem trazer a mesma informação para o problema, estando portanto correlacionadas.

Para analisar a relevância das variáveis foi novamente analisado o nó "*StatExplore*", nomeadamente para o gráfico "*Variable Worth*" após as alterações efectuadas anteriormente.

Importance	***Variable***	***Worth***
1	Vencimento_Anual	0,0179
2	Despesa_media_diaria	0,0062
3	Ano_Nascimento	0,0034
4	Nacionalidade	0,0031
5	Despesa_media_Viagem	0,0029
6	NumViagensAno	0,0028
7	Estado_Civil	0,0028
8	Motivo_Viagem	0,0007

9	Duração_Estadia	0,0005
10	Tipo_Estabelecimento	0,0004
11	Filhos	0,0002
12	Meio_Transporte_Utilizado	0,0001
13	Nível_Instrução	0,0001
14	Sexo	0,0001

Tabela 3.6 – Importância das Variáveis

Apesar de pequenas alterações no "*Worth*" das variáveis e consequentemente na sua posição em termos de importância, aquelas que possuem um maior poder discriminativo da variável dependente continuaram a ser as variáveis "Vencimento_Anual" e "Despesa_media_diaria". As variáveis "Tipo_Estabelecimento" e "Sexo" continuaram a estar entre aquelas que menor contribuição tiveram para a resolução do problema. Assim sendo, por se tratarem das dimensões com menor relevância não foram consideradas na modelação.

Para fazer a análise de redundância das variáveis intervalares foi utilizada a análise da matriz de correlação *Spearman* (Figura 3.7). Foi utilizado o coeficiente de correlação de *Spearman* na análise, uma vez que é menos sensível a valores muito distantes do que o coeficiente de *Pearson*, não requerendo a suposição de que a relação entre as variáveis é linear.

Considerou-se que existe redundância entre variáveis quando o coeficiente de correlação absoluto for superior ou igual a 0,8.

NAME	Ano_Nas...	Data_Reg...	Despesa...	Despesa...	Dura__o_...	Filhos	NumViag...	Vencime...	Observati...
Ano_Nascimento	1	0.003088	-0.00315	-0.03759	-0.03257	0.212192	-0.07609	-0.18709	0.004475
Data_Registo	0.003088	1	0.033812	-0.07967	-0.02224	0.011803	-0.19124	0.024211	-0.00396
Despesa_media_Viagem	-0.00315	0.033812	1	0.026491	0.020839	-0.03075	-0.02773	-0.00683	0.014098
Despesa_media_diaria	-0.03759	-0.07967	0.026491	1	-0.01008	-0.36584	-0.16547	0.392186	-0.02776
Dura__o_Estadia	-0.03257	-0.02224	0.020839	-0.01008	1	-0.02264	-0.00666	0.02145	-0.00274
Filhos	0.212192	0.011803	-0.03075	-0.36584	-0.02264	1	0.248172	-0.43863	-0.01804
NumViagensAno	-0.07609	-0.19124	-0.02773	-0.16547	-0.00666	0.248172	1	-0.16163	-0.0053
Vencimento_Anual	-0.18709	0.024211	-0.00683	0.392186	0.02145	-0.43863	-0.16163	1	0.057355
dataobs	0.004475	-0.00396	0.014098	-0.02776	-0.00274	-0.01804	-0.0053	0.057355	1

Tabela 3.7 – Correlação Spearman

Através da matriz de correlação foi possível verificar que a correlação mais relevante é a existente entre as variáveis "Vencimento_Anual" e "Despesa_media_diaria" (0,39). A variável "Vencimento_Anual" apresenta um "*Worth*" superior ao da variável "Despesa_media_diaria", sendo

portanto mais relevante para a definição da variável target. No entanto, a variável "Despesa_media_diaria" não foi excluída uma vez que o coeficiente de correlação absoluto é inferior a 0,8 não existindo desta forma redundância entre as variáveis em causa.

Uma outra forma de avaliar a redundância das variáveis passa por recorrer ao nó "*Variable Clustering*", no qual as variáveis de *input* são agrupadas em clusters e que pode servir como uma base para a redução da dimensionalidade do problema (Figura 3.10).

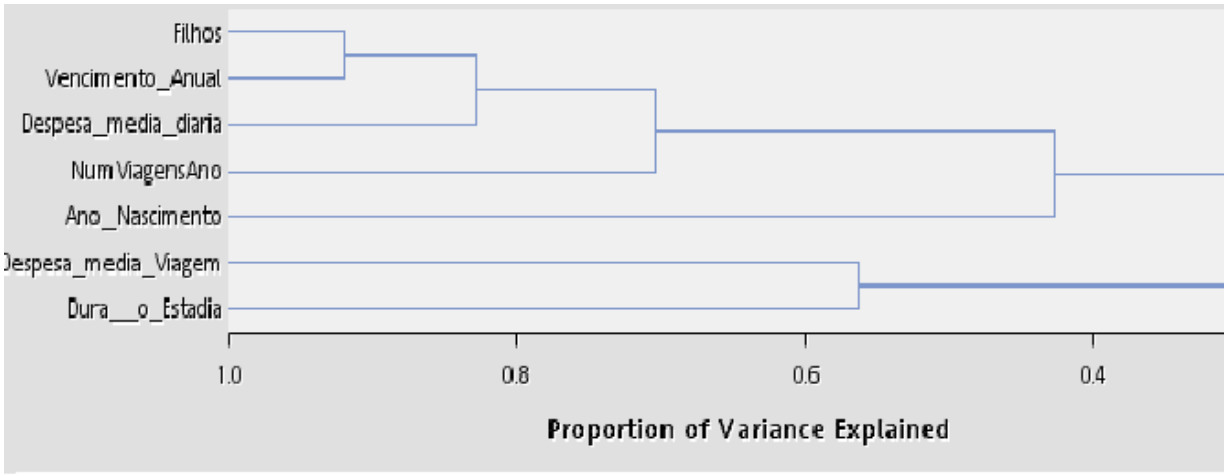

Figura 3.10 - Dendograma (*Variable Clustering*)

Utilizando "*Correlation*" como "*Clustering Source*", os clusters vão sendo definidos através do agrupamento sucessivo das variáveis mais correlacionadas. Para compreender a relação de proximidade entre *Clusters* é possível também recorrer à sua análise gráfica (Figura 3.11), procurando identificar onde se situam os grandes focos de informação.

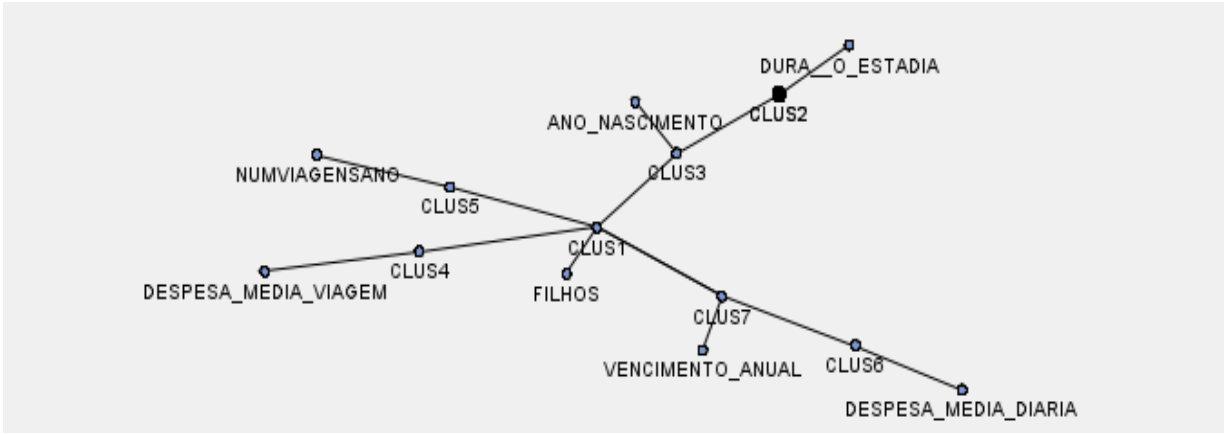

Figura 3.11 - *Cluster Plot (Variable Clustering)*

3.6. PARTIÇÃO DOS DADOS

Antes de efetuar a modelação preditiva foi necessário dividir a base de dados em três conjuntos distintos no nó "*Data Partition*":

· Conjunto de Treino – Treina os dados e ajusta o modelo. Quanto maior for este conjunto, melhor é o classificador e maior é a sua experiência;

· Conjunto de Validação – Controla o processo de treino e monitoriza o erro. Determina quando o modelo é suficientemente complexo e quando o treino deve parar. Quanto maior for este conjunto, mais facilmente é possível saber quando o modelo já é satisfatório e o treino deve ser interrompido (evitar o "*overfitting*");

· Conjunto de Teste – Estima a qualidade do modelo e a sua precisão quando aplicado a novos dados. Quanto maior este conjunto de dados, melhor é a performance do modelo face aos novos dados apresentados.

Tipicamente, para uma base de dados com uma dimensão entre 1.000 e 2.000 indivíduos, a partição mais correta é a seguinte (Figura 3.12):

Conjunto	%
Treino	70%
Validação	30%
Teste	0%

Summary Statistics for Class Targets

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
DepVar	0	0	2147	85.88	DepVar
DepVar	1	1	353	14.12	DepVar

Data=TRAIN

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
DepVar	0	0	1502	85.9268	DepVar
DepVar	1	1	246	14.0732	DepVar

Data=VALIDATE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
DepVar	0	0	645	85.7713	DepVar
DepVar	1	1	107	14.2287	DepVar

Figura 3.12 - Partição dos Dados

Uma vez que o conjunto de dados é relativamente baixo, não existiu necessidade de utilizar o conjunto de teste. A divisão dos dados pelos vários conjuntos de dados é feita de uma forma estratificada com base na variável dependente.

3.7. METADATA

No nó "Metadata" são alteradas as definições das variáveis que se pretendem excluir para que não sejam consideradas na modelação preditiva.

Name	***Role***	***Level***
Custid	*ID*	*Nominal*
DepVar	*Target*	*Binary*
Despesa_media_diaria	*Input*	*Interval*
Despesa_media_Viagem	*Input*	*Interval*
Motivo_Viagem	*Rejected*	*Nominal*
Duração_Estadia	*Rejected*	*Interval*
Data_Registo	*Input*	*Interval*
Estado_Civil	*Input*	*Nominal*
Filhos	*Rejected*	*Interval*
Ano_Nascimento	*Input*	*Interval*
Meio_Transporte_Utilizado	*Rejected*	*Nominal*
Nacionalidade	*Input*	*Nominal*
Nivel_Instrução	*Rejected*	*Nominal*
NumViagensAno	*Input*	*Interval*
Vencimento_Anual	*Input*	*Interval*
Sexo	*Rejected*	*Nominal*
Tipo_Estabelecimento	*Rejected*	*Nominal*

Tabela 3.6 – Papel e Nível das Variáveis

7 variáveis foram rejeitadas:

*"*Tipo_Estabelecimento*", "*Sexo*", "*Nivel_Instrução*","*Meio_Transporte_Utilizado*", "*Motivo_Viagem*", "*Duração_Viagem*". Tratam-se de variáveis com um "*Worth*" baixo e portanto com um menor poder discriminativo da variável dependente;

Assim sendo, na modelação foram consideradas 8 variáveis de Input**:** 6 intervalares e 2 de classe. As variáveis selecionadas são aquelas que se considerou serem simultaneamente relevantes e não

redundantes. A partir deste ponto, somente foram utilizadas as variáveis que possuem maior capacidade discriminativa.

3.8. MODELAÇÃO PREDITIVA

Após a preparação e o pré-processamento dos dados, bem como a sua partição, procede-se à modelação. Para tal são utilizados diversos modelos de modo a compreender qual aquele que apresenta os melhores resultados.

3.8.1. Modelos

Os modelos utilizados nesta fase foram os seguintes:

Modelo	Observações
Neural Network	*Multilayer Perceptron – 1 Hidden Unit*
Neural Network 2	*Multilayer Perceptron – 2 Hidden Unit*
Neural Network 3	*Multilayer Perceptron – 3 Hidden Unit*
Neural Network 4	*Multilayer Perceptron – 4 Hidden Unit*
Decision Tree	*Ordinal Criterion - Entropy*
Regression	*Type – Logistic / Model Selection - Stepwise*
Ensemble	*Class Target - Voting*

Tabela 3.7 – Modelos Utilizados

Foram utilizadas Redes Neuronais assentes no modelo do perceptrão multicamadas, cada uma com diferentes unidades na camada escondida (entre 1 e 4). A ideia subjacente às Redes Neuronais passa por conseguir que estas sejam capazes de "aprender" a chegar sozinhas às conclusões. Após a informação ser disponibilizada, pretende-se que cada Rede Neuronal seja capaz de desenvolver uma "aprendizagem" automática.

No modelo do perceptrão multicamadas, a Rede Neuronal tem uma ou mais camadas escondidas. A camada escondida é composta por um ou mais neurónios, sendo que quanto maior for o seu número, maior é a complexidade da própria rede.

Desta forma, fez sentido começar com uma camada com menos neurónios e ir fazendo crescer a rede para analisar os resultados obtidos. A partir de certa altura é natural que o modelo dê piores resultados, à medida que a rede vá aumentando a sua complexidade. Outra ferramenta de modelação preditiva utilizada foi a Árvore de Decisão (Figura 3.13). Este tipo de modelo é de fácil

interpretação e as variáveis mais importantes para a solução final são aquelas que ficam no topo da árvore.

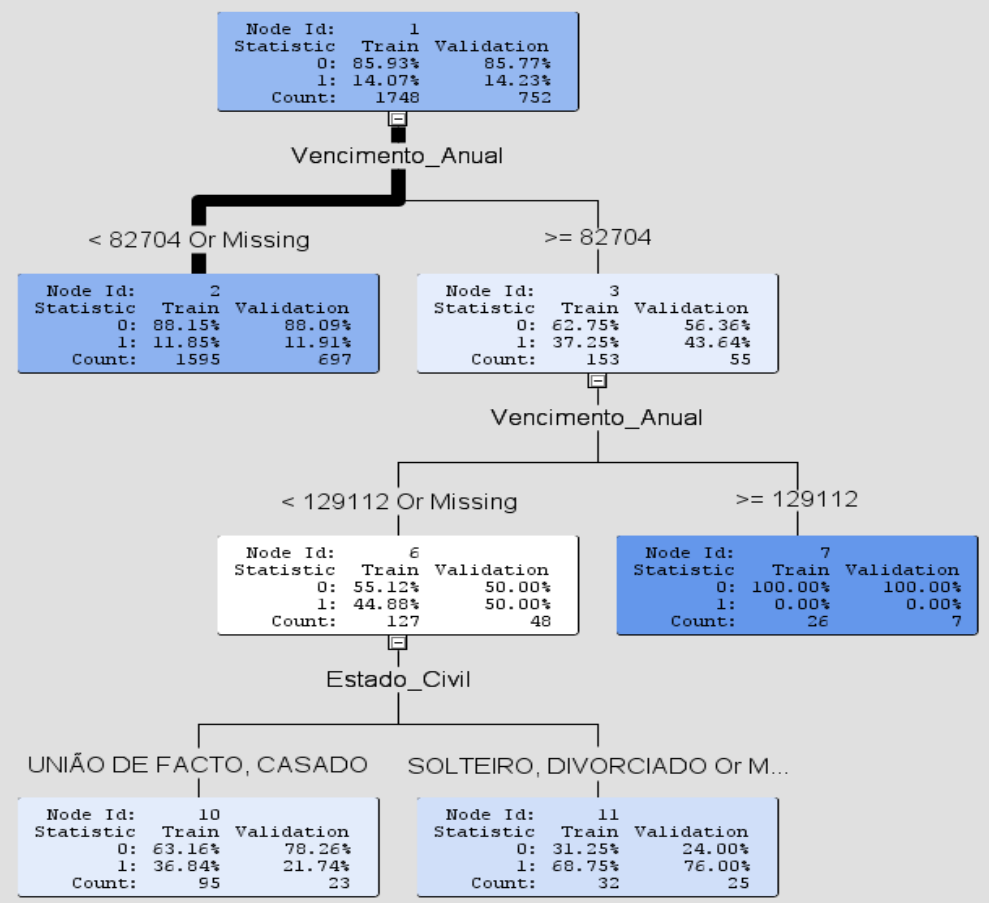

Figura 3.13 – Árvore de Decisão

Foi utilizada também a Regressão logística para determinar a variável target (Figura 3.14).

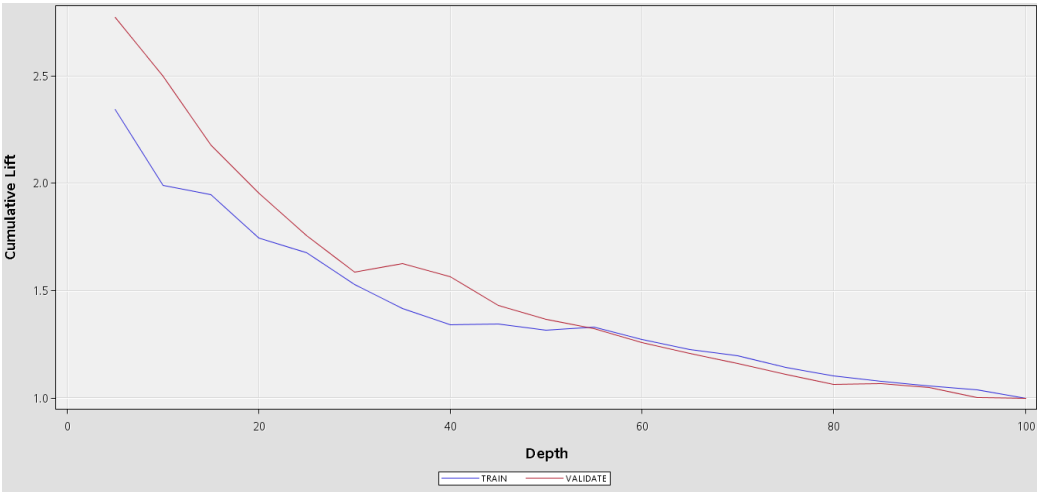

Figura 3.14 – *Cumulative Lift* (Regressão)

Por fim, foi utilizado o modelo Ensemble (Figura 3.15) que não trabalha especificamente nos dados, mas sim nos próprios modelos.

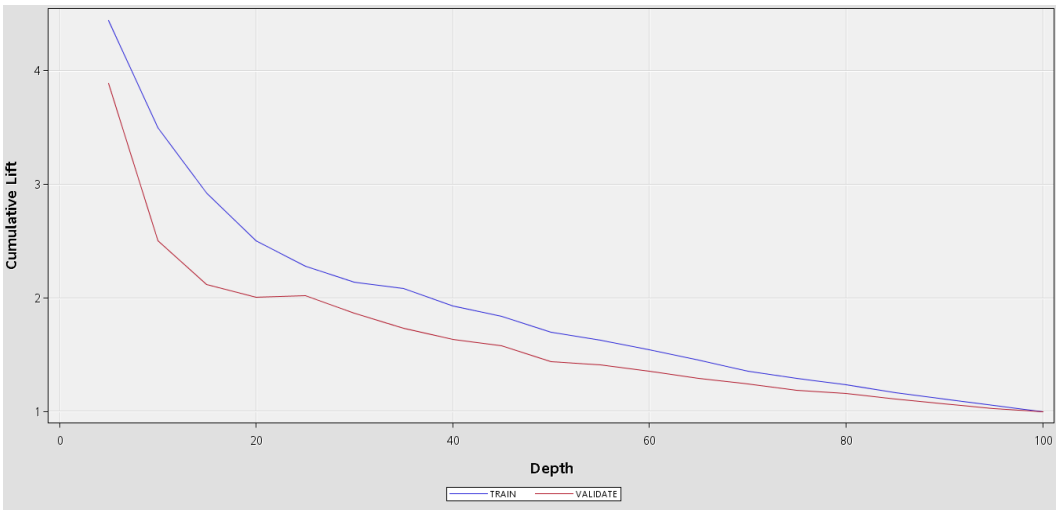

Figura 3.15 – *Cumulative Lift (Ensemble)*

A ideia inerente ao Ensemble passa pela utilização dos diferentes modelos, os quais efetuam uma "votação" que conduz ao melhoramento da capacidade preditiva. Trata-se portanto de um modelo de segunda linha que combina modelos de primeira linha.

4. RESULTADOS E DISCUSSÃO

Depois de ter sido efectuada a modelação, foi necessário escolher o modelo que apresenta os melhores resultados. Esta comparação foi efectuada recorrendo ao nó "*Model Comparison*".

O Gráfico ROC permite a análise comparativa dos vários modelos tendo em conta dois aspectos:

· "*Sensitivity*" – Velocidade com que o modelo captura os "1"

· "*Specificity*" – Velocidade com que o modelo captura os "0"

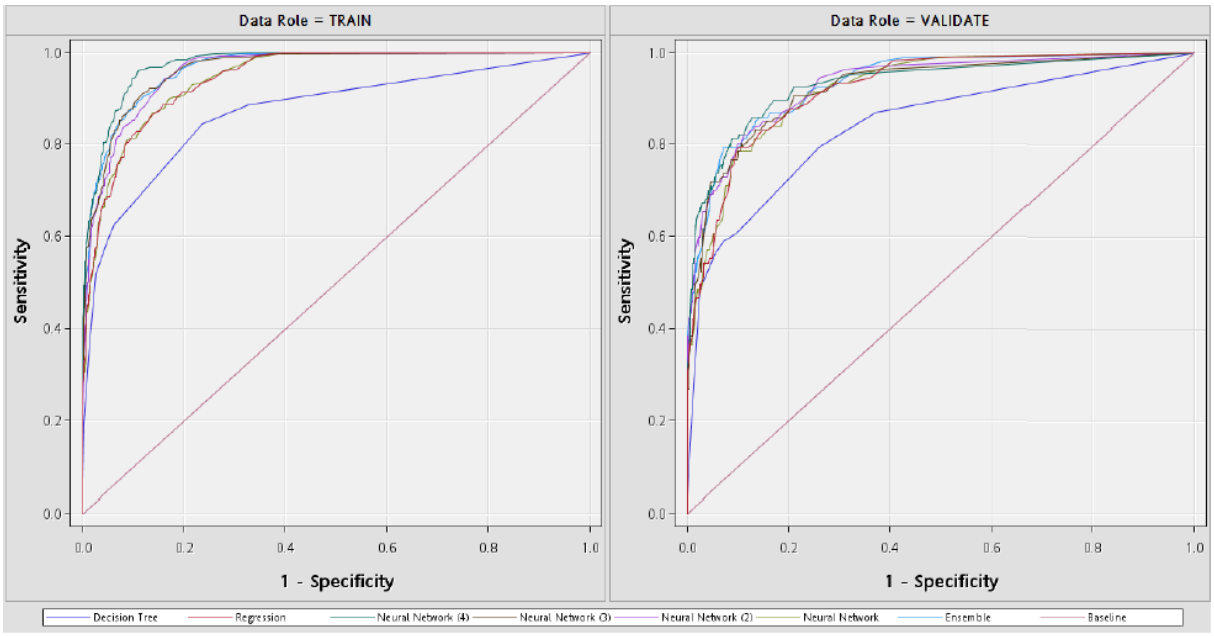

Figura 3.16 – *ROC Curve*

Uma vez que podem ser utilizados diversos modelos preditivos, a análise meramente gráfica torna-se mais complexa. A alternativa consiste em recorrer directamente ao "*ROC Index*" que traduz a área do gráfico abaixo de cada curva. Deste modo, o modelo que apresentar um índice mais elevado será aquele com melhor performance.

Modelo	*Train: ROC Index*	*Valid: ROC Index*
Neural Network	*0.946*	*0.924*
Neural Network 2	*0.961*	*0.936*
Neural Network 3	*0.965*	*0.931*
Neural Network 4	*0.975*	*0.940*
Decision Tree	*0.871*	*0.841*
Regression	*0.945*	*0.924*
Ensemble	*0.966*	*0.936*

Tabela 3.8 – Performance dos Modelos Utilizados

O modelo *Neural Network 4* foi aquele que apresentou um "*ROC Index*" mais elevado no conjunto de treino e validação, pelo que será aquele que conduzirá aos resultados mais satisfatórios.

No comparativo entre modelos, foi possível verificar que o Neural Network 4 é aquele que permite melhor performance. Este facto já era esperado, uma vez previamente já se tinha concluído que este era o modelo que demonstrava melhor capacidade preditiva neste caso específico.

Modelo	*Cumulative % Response*	*Depth %*
Neural Network 4	*70*	*20*
Neural Network 3	*69*	*15*
Ensemble	*59*	*15*
Neural Network 2	*58*	*20*
Regression	*57*	*20*
Neural Network	*57*	*20*
Decision Tree	*57*	*15*

Tabela 3.9 – Comparativo dos valores *Depth%*

O gráfico seguinte permite deduzir que o modelo *Neural Network 4* obteve os melhores resultados, no que diz respeito à Base de Dados analisada.

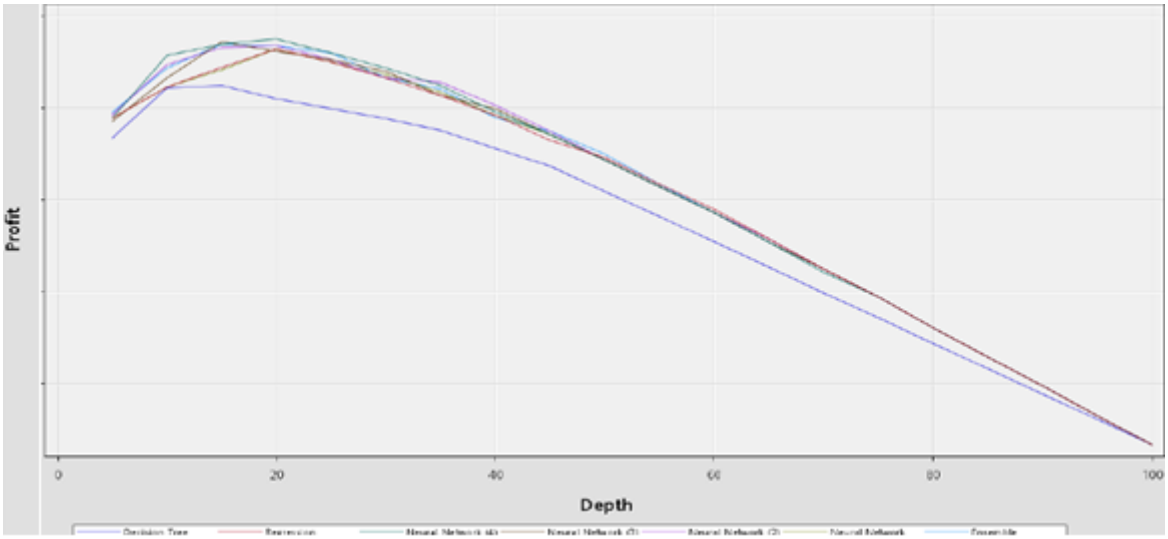

Figura 3.17 – Curva de lucro

O valor óptimo esperado com a utilização do Neural Network 4 será de 1552 € para um "Depth" de 20%.

5. CONCLUSÕES

O turismo é uma indústria crescente e uma das maiores fontes de rendimento para vários países. No entanto, existe a necessidade em concentrar atenções nos dados disponíveis da indústria do turismo, a fim de fornecer conhecimento e desenvolver sistemas que possam ajudar as empresas de turismo prosperar.

O presente trabalho consistiu no desenvolvimento de diversos modelos preditivos que pudessem auxiliar na identificação do perfil do turista e da sua recetividade em adquirir serviços ou produtos relacionados com o setor turístico em Portugal. A Base de Dados recolhida foi fundamental para obter um maior nível de conhecimento acerca dos visitantes, nomeadamente no que diz respeito ao seu poder de compra e hábitos de consumo.

Foi possível concluir que o perfil de cliente com maior probabilidade de visitar Portugal é caracterizado por indivíduos casados, que utilizam preferencialmente o automóvel e que procuram o país por motivos de lazer. A informação recolhida permitiu também concluir que as estadias até 7 dias em infraestruturas hoteleiras são as que representam a preferência da maioria dos clientes, com um gasto médio previsto em 1552 €.

O recurso às técnicas de *Data Mining* referidas mostrou ter capacidade preditiva utilizando uma amostra de todo o universo dos indivíduos, possibilitando posteriormente estabelecer perfis e prever comportamentos. Esta capacidade aplicada aos dados do turismo terá aplicação direta na criação de campanhas e infraestruturas. Este conhecimento permitirá agir e inovar, diminuindo os custos e aumentando os lucros envolvidos.

Do ponto de vista da análise técnica é possível concluir que os resultados do modelo preditivo foram bons. O algoritmo utilizado (Neural Network 4) obteve 70% de previsões corretas. Existiu alguma dificuldade do modelo em evitar correlações espúrias e como tal foi necessário tratar a configuração dos parâmetros de forma a evitar introduzir "ruido" que de alguma forma pudesse prejudicar os resultados obtidos. Foi possível constatar que, pequenas alterações na forma como o conjunto de treino e conjunto de validação eram divididos, produziam erros quadráticos médios totalmente distintos. Após vários ajustes no modelo, verificou-se que a rede neuronal com 4 camadas escondidas foi o método que lidou melhor com valores elevados de *input* e também o que apresentava um erro médio menor, sendo por isso o mais adequado nas previsões.
No entanto, a restrição inicial associada à criação de uma amostra da realidade e ao reduzido número de variáveis disponíveis, suscitou uma eficácia do modelo inferior a 80% traduzindo-se numa

performance que não atingiu uma capacidade preditiva ótima.

Em suma, os resultados alcançados permitiram dar resposta aos objetivos inicialmente propostos. Através da literatura recolhida para a elaboração deste estudo e da utilização de vários algoritmos de classificação com diferentes parametrizações, foi possível identificar as variáveis que mais contribuíram para identificar perfis e comportamentos do turista em Portugal. Apesar da performance do modelo preditivo desenvolvido não ter sido ótima, a capacidade preditiva do mesmo revelou ser eficaz na previsão do valor gasto por estadia.

O trabalho realizado pretende ser um contributo para os investigadores da área de *Data Mining* e também para os agentes do turismo.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA TRABALHOS FUTUROS

A mineração de dados pode ser definida como o processo de análise de grandes conjuntos de dados que possibilita pesquisar e descobrir padrões anteriormente desconhecidos, tendências e relacionamentos de forma a gerar informações que auxiliem a tomada de decisões. Trata-se de uma tecnologia e metodologia muito dominante e positiva para a indústria do turismo, a mineração de dados pode contribuir decisivamente para a capacidade que as empresas de viagens e turismo tenham em se desenvolver e em criar vantagens competitivas.

No entanto, a mineração de dados também tem limitações. Algumas das mais recorrentes consistem nas seguintes:

- A qualidade dos resultados e da utilização de ferramentas de mineração de dados depende da disponibilidade e qualidade dos dados (Chopoorian , Witherell, Khalil , & Ahmed , 2001). Além disso, os dados necessários para a mineração geralmente existem em configurações e sistemas distintos, necessitando assim ser adquiridos e integrados antes que a mineração de dados possa ser realizada. Problemas como dados omissos, dados corrompidos, dados inconsistentes, etc. devem ser identificados e tratados antes da mineração ser concluída. Estima-se que a preparação de dados utilize cerca de 75% dos recursos necessários num projeto de mineração de dados.

- A mineração de dados pode não identificar alguns padrões que sejam produto de flutuações aleatórias (Hand D. J., 1998). Este constrangimento verifica-se geralmente em grandes conjuntos de dados com muitas variáveis. A utilização de Mineração de dados de forma mecanizada não garante resultados ou sucesso, uma vez que existirá sempre a necessidade de intervenção humana, interpretação e julgamento (Pyo , Uysal , & Chang, 2002).

- A aplicação bem sucedida da mineração de dados no Turismo está dependente do conhecimento que o utilizador tem do setor, bem como, das tecnologias e ferramentas de mineração de dados. O conhecimento do setor é importante porque permite identificar os problemas inerentes ao negócio e a forma mais adequada de desenvolver as aplicações de mineração de dados. Possibilita também desenvolver modelos apropriados e uma correta interpretação dos resultados (McQueen & Thorley , 1999).

Finalmente, as organizações que desenvolvem aplicações de mineração de dados necessitam de investimentos substanciais em recursos. Um projeto de mineração de dados pode falhar por uma variedade de razões, como por exemplo, a falta de apoio à gestão e ao compromisso organizacional, as expectativas não realistas dos utilizadores, a má gestão do projeto, a mineração de dados inadequada, etc (Gillespie , 2000). Independentemente das limitações mencionadas, não existem dúvidas que a mineração de dados irá desempenhar um papel decisivo no setor turístico. É somente necessário que a indústria turística consiga entender e capitalizar os potenciais benefícios e a utilidade da mineração de dados.

7. BIBLIOGRAFIA

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics, 12(1)* , 90–108.

Buhalis, D. (2002). eTurism: Information technologies for strategic tourism management. *Financial Times Pretice Hall* .

Bação, F. (2009). *Introdução ao Data Mining, Apontamentos Mestrado.* Lisboa: NOVA IMS.

Benckendorff, P. J., Sheldon, P. J., & Fesenmaier, D. R. (2014). *Tourism Information Technology.* Wallingford: Cab International.

Berry, M., & Linoff, G. (2004). *Data Mining Techniques, for sales, and customer relationship management, 2nd Edition.* Wiley Publishing, Inc.

Bose, R. (2009). "Advanced analytics: opportunities and challenges". *Industrial Management & Data Systems , 109* (2), pp. 155-172.

Chopoorian , J. A., Witherell, R., Khalil , O. M., & Ahmed , M. (2001). *SAM Advanced Management Journal, 66(2)* , 45-51.

Conway, D. (2011). Data Science in the U.S. Intelligence Community. *IQT Quarterly Spring , Vol 2*, 24-27.

De Mauro, A., Greco, M., & Grimaldi, M. (2015). What is big data? A consensual definition and a review of key research topics. *AIP Conference Proceedings, 1644(1)* , 97-104.

Dev, C. S., Klein, S., & Fisher, R. A. (1996). *Journal of Travel Re-search (35)1* , 11-17.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth , S. (1996). Advances in Knowledge Discovery and Data Mining, American Association for Artificial Intelligence. *AI Magazine* .

Fan, J., Han, F., & Liu, H. (2014). Challenges of big data analysis. *National Science Review, 1(2)* , 293-314.

Frederiksen, L. (2012). Big data. *Public Services Quarterly, 8(4)* , 345-349.

Friedman, J. H. (1997). Data Mining and Statistics: What's the Connection? *Proceedings of the 29th Symposium on the Interface: Computing Science and Statistics* .

Guazzelli, A. *Predicting the future, Part 1: What is predictive analytics? .* 2012.

Gillespie , G. (2000). Health Data Management, 8(11). 40-52.

INE. (2015). *Estatísticas do Turismo-2015. .* Instituto Nacional de Estatística.

INE. (2016). *Estatísticas do Turismo-2016. .* Instituto Nacional de Estatística.

Han, J., & Kamber, M. (2001). *Data Mining – Concepts and Techniques* . San Francisco, California: Morgan Kaufmann.

Hand , D. J. (1998). *The American Statistician, 52(2)* , 112-118.

Hand, D., Manila, H., & Smyth, P. (2001). *Principles of Data Mining.* Cambridge: MIT Press.

Hashem, I. A., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Ullah Khan, S. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems,* , 98-115.

Hassani, H., & Silva, E. (2015). Forecasting with big data: A review. *Annals of Data Science, 2(1)* , 5-19.

Hertzmann, A., & Fleet, D. (2012). *Machine Learning and Data Mining Lecture Notes.* Lecture Notes.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data Mining Use Cases and Business Analytics Applications (Chapman & Hall/CRC DM & Knowledge Discovery Series)".* CRC Press.

Jain, N., & Srivastava, V. (2013). Data Mining Techniques: a Survey Paper. *IJRET: International Journal of Research in Engineering and Technology, 2(11)* , 116–119.

Kasavana , M. L., & Knutson , B. J. (1999). *Journal of Hospitality and Leisure Marketing, 6(1)* , 83-86.

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

Nanda, S. K., Tripathy, D. P., Nayak, S. K., & Mohapatra , S. (2013). Prediction of Rainfall in India using Artificial Neural Network (ANN) Models. *International Journal of Intelligent Systems and Applications, 5(12)* , 1-22.

Magnini, V. P., E.D., H. J., & Hodge, S. K. (2003). *Cornell Hotel and Restaurant Administration Quarterly, 44(2)* , 94-105.

Mainon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook, 2nd Edition.* Outlier detection.

McQueen , G., & Thorley , S. (1999). Financial Analysts Journal, 55(2) . 61-72.

Monk, E. F., & Wagner, B. J. (2013). *Concepts in Enterprise Resource Planning.* USA: Cengage Learning.

O'BRIEN, J. A. (2004). *Sistemas de informação e as decisões gerenciais na era da Internet* (Vol. 2. ed.). São Paulo: Saraiva.

Olsen, M., & Connolly, D. (1999). Tourism Analysis, 4(1). 29-46.

Pyo , S., Uysal , M., & Chang, H. (2002). *Journal of Travel Re-search, 40(4)* , 396-403.

Siegel, E. (2013). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.* Wley & Sons, Inc.

Sheldon, P. J. (1997). *The Tourism Information Technology.* Wallingford: CAB International.

Snijders, C., Matzat, U., & Reips, U.-D., U., & Reips, U. D. (2012). Big data: Big gaps of knowledge in the field of. *International Journal of Internet Science, 7(1)* , 1-5.

Sondwale, P. P. (2015). Overview of Predictive and Descriptive Data Mining Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering, 5(4)* , 262-265.

Rascão, J. (2004). *Sistemas de Informação para as Organizações – A informação Chave para a Tomada de Decisão.* Edições Sílabo.

Rojas, R. (1996). Neural networks: a systematic introduction. *Neural Networks, 502* .

Tan, Steinbach, & Kumar. (2006). *Introduction Data Mining.* Pearson Addison-Wesley.

Turismo de Portugal I.P. (TdP). (2017). *Estratégia Turismo 2027 .*

UNWTO. (2017). *Barómetro do Turismo Mundial.* World Tourism Organization.

Werthner, H., & Ricci, F. (2004). E-commerce and tourism. *Communications of the ACM , 42* (12), pp. 101-105.

Zhu, X., & Davidson, I. (2007). Knowledge Discovery in Biomedical Data Facililtated by Domain Ontologies. *Knowledge Discovery and Data Mining: Challenges and Realities* , 189–201.

8. ANEXOS

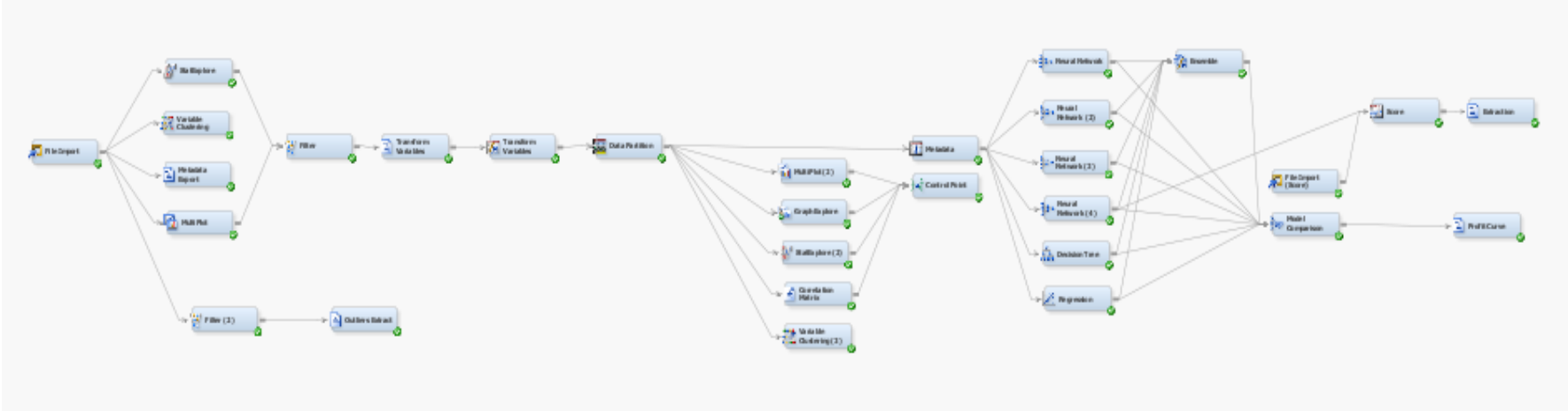

Figura 8.18 – Diagrama Final do Projeto