Universidade do Minho
Escola de Engenharia

Rita Daniela Gomes Ferreira

Avaliação de Perfis de Agentes Comerciais em Sistemas de Retalho Especializado

Tese de Mestrado

Mestrado em Engenharia de Sistemas

Trabalho efetuado sob a orientação do

Professor Doutor Orlando Manuel de Oliveira Belo

Outubro de 2019

AGRADECIMENTOS

Quando se conquista uma nova etapa, seja académica ou pessoal, existem pessoas que sem as quais seria bem mais complicado alcançar essa conquista, e a quem se deve agradecer de forma especial, tais como:

Aos meus pais, por todo o amor, carinho e apoio que deram ao longo destes cinco anos académicos que agora terminam, mas também ao longo de toda a minha vida.

Ao meu namorado, por ser o meu pilar em todos os momentos, e pelo apoio e força que me deu para que este percurso terminasse com sucesso.

Á minha avó Ermelinda, que não me vê terminar o meu percurso académico, mas que muito contribuiu para que me tornasse na pessoa que sou hoje.

Ao professor Orlando Belo, pela disponibilidade demonstrada, ao longo de toda a dissertação, pela forma como sempre mostrou qual o melhor caminho a seguir, por todo o conhecimento transmitido, que muito contribuiu para o sucesso desta dissertação, e pela oportunidade concebida de puder trabalhar sob a sua tutoria.

À empresa *F3M Information Systems S.A.* pela oportunidade de puder realizar o meu estágio curricular, em especial ao diretor do departamento de desenvolvimento, Engenheiro Manuel Pereira, pelo acompanhamento prestado ao longo do estágio e a todos os colaboradores pela forma como me integraram e por toda a ajuda prestada.

Por fim, a todos os familiares e amigos, que direta ou indiretamente, contribuíram para que este objetivo fosse alcançado.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

RESUMO

Nos últimos anos, registou-se um crescimento abrupto do volume de dados armazenados pelas empresas, o que fez com que, hoje, a aplicação de ferramentas que auxiliem a análise desses dados seja imprescindível. Depois de analisados, esses dados podem ser um valioso auxílio à tomada de novas decisões estratégicas. Atualmente, a informação é a chave para tudo. Por esse motivo, é preciso optar-se por sistemas que sejam iterativos com os utilizadores, automatizando a forma de analisar os dados e obtendo informação importante para o negócio. Devido a este aumento significativo de dados, a forma de extrair o conhecimento sofreu uma enorme evolução, aumento de forma extraordinária o número de aplicações da mineração de dados. Através do recurso das técnicas de mineração de dados é possível encontrar informação pertinente para perfilar clientes ou agentes comercias, entre outros, permitindo aos agentes de decisão tomar medidas especificas nos diversos casos que vão surgindo no seu quotidiano profissional.

Neste trabalho de dissertação foram aplicadas diferentes técnicas de mineração de dados, nomeadamente, classificação, segmentação e associação, e de modelação de modelos de mineração de dados, com o objetivo de extrair conhecimento válido na construção de perfis de agentes comerciais e optometristas. Os resultados obtidos através das referidas técnicas, serão avaliados, e caso obedeçam aos critérios de sucesso de negócio definidos, transformar-se-ão em conhecimento que apoie a tomada de novas decisões estratégicas.

PALAVRAS-CHAVE

Mineração de Dados, Classificação, Segmentação, Associação, *Profiling* de Agentes Comerciais e Optometristas.

Abstract

In recent years, there has been an abrupt increase in the volume of data stored by companies, which has made the application of tools to assist in the analysis of this data today essential. After the introduction of tools for data analysis and storage, the data can be valuable aid in strategic decisions. Nowadays, information is the key to everything. For that reason, you must opt for systems that are user interactive, automating the way you analyze data and getting important information for your business. Due to this significant data increment, the way to extract the knowledge has undergone an enormous evolution, increase extraordinary the number of applications of data mining. Using these techniques of data mining, it is possible to find the information relevant to profile clients or commercial agents, among others, allowing the decision agents to take specific steps and measures in the various cases that arise in your day by day professional.

In this dissertation, different data mining techniques were applied, namely, classification, segmentation and association, and modelling data mining models, with the objective of extracting the knowledge for the construction of profiles of commercial agents and optometrists. The results obtained by techniques referred of data mining are evaluated, and if they obey with the business success criteria, they will be transformed in knowledge that support the new strategic decisions making.

Keywords

Data Mining, Classification, Clustering, Association, Profiling of Commercial Agents and Optometrists.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

Lista de Abreviaturas, Siglas e Acrónimos

EM – *Expectation-Maximization*

CRISP-DM - *Cross Industry Standard Process for Data Mining*

KDD - *Knowledge Discovery from Data*

CAPÍTULO 1

1. INTRODUÇÃO

1.1 Contextualização

Nos últimos anos, as tecnologias de informação padeceram de uma grande evolução, que influenciou várias das suas áreas de trabalho. Pode-se, inclusive, afirmar-se que uma empresa, independentemente do setor em que opera, não sobrevive sem o uso das tecnologias de informação, seja para organizar as tarefas por si desenvolvidas, seja para que os seus serviços atinjam algumas vantagens competitivas relativamente aos seus concorrentes (Pacheco *et al.*, 2000).

Como consequência do aumento do volume de dados, a um ritmo dramático, é necessário criar técnicas e ferramentas computacionais, que possibilitem a partir desses dados a extração de conhecimento benéfico (Fayyad *et al.*, 1996). A abundância de dados e a necessidade de encontrar uma ferramenta eficaz na análise de dados, pode ser descrita com uma situação *"a data rich but information poor"* (Han *et al.*, 2011). Uma das técnicas que tem emergido com o objetivo de extrair padrões significantes das bases de dados é a Mineração de Dados. Esta técnica combina a pesquisa em diferentes áreas, tais como, *machine learning*, estatística, sistemas de bases de dados, entre outras (Zaki *et al.*, 1997).

No domínio da mineração de dados podemos encontrar seis técnicas correntes de mineração, como seja a descrição, a estimação, a previsão, a classificação, a segmentação e a associação - nesta dissertação apenas serão abordadas as últimas três técnicas referidas. Na técnica de classificação, existe uma variável alvo categórica que pode ser dividida em diferentes classes. O modelo de mineração de dados analisa o conjunto de dados, onde cada registo de dados contém informação relativa à variável destino, mas também um conjunto de variáveis de entrada. A técnica de segmentação agrupa os registos em *clusters*. Os registos que pertencem a um *cluster* devem ser semelhantes entre si e dissemelhantes quando comparados aos de outro *cluster*. Esta técnica diferencia-se da anterior, pois não classifica,

estima ou prevê uma variável de destino, visto que esta não existe. A última técnica abordada é a associação que é conhecida como a análise de afinidade e tem como objetivo descobrir regras de modo a estabelecer qual a relação existente entre dois ou mais atributos, sendo do tipo *"Se antecedente, em seguida, consequente"*, apoiadas por medidas de suporte e confiança relativas a cada regra gerada (Larose & Larose, 2014).

Entre as metodologias existentes para a mineração de dados, na presente dissertação adotou-se a metodologia CRISP-DM. Esta metodologia fornece um plano completo para a realização de um processo de mineração de dados, abordando com detalhe cada uma das suas etapas. Na CRISP-DM, um projeto é dividido em seis fases, nomeadamente: análise de negócio, compreensão de dados, preparação de dados, modelação, avaliação e implementação. De realçar que apesar da divisão em fases distintas, o fluxo deste ciclo pode sofrer um retorno de uma para outra fase sempre que necessário (Shearer, 2000) (Camilo & Silva, 2009).

Atualmente, tendo em conta o crescimento exponencial do volume de informação disponível, torna-se indispensável utilizar esta informação como recurso, cujo valor é inestimável, de forma a automatizar a análise de dados, obtendo assim a informação presente nos dados e que classifica os diferentes utilizadores. Este processo é conhecido por *profiling* (Gauch *et al.*, 2007). Consequente à evolução do armazenamento de dados, e com um mercado cada vez mais competitivo, é imprescindível para qualquer marca ou empresa conhecer o perfil do seu agente comercial, bem como do seu cliente. Para algumas empresas o crescente volume de informação disponível na internet, pode revelar-se um problema, visto que, caso esta não ofereça o produto preterido pelo cliente, este facilmente procura e encontra o tal produto, optando pela concorrência. De modo a combater esta fragilidade, é importante reunir e tratar informações acerca dos seus utilizadores, oferecendo assim informações personalizadas, como por exemplo, exibir produtos que outros utilizadores compraram aquando da compra do produto que o utilizador está a visualizar, ou tendo em conta as suas preferências pessoais (Wiedmann *et al.*, 2002). No entanto, a construção de um perfil vai um pouco para além destas informações, sendo necessário reunir outro tipo de informação, para que seja possível fazer a construção de um perfil bem mais real. Nos dias que correm, a internet fornece um conjunto de técnicas capazes de coletar mais dados sobre os utilizadores, que depois de processados, cuidadosamente, permitem construir o seu perfil, que depois deverá ser capaz de derivar informações personalizadas (Fleuren, 2012). Pode assim concluir-se, que a informação é importante, e que quando bem interpretada e analisada, pode solucionar problemas e criar novas estratégias de negócio.

1.2 Motivação e Objetivos

Na secção anterior foram expostas as principais áreas que serão abordadas ao longo desta dissertação. Contudo, o ponto central focou-se na utilização de algoritmos de mineração de dados, que fossem capazes de classificar, agrupar e gerar regras de associação sobre um dado conjunto de dados. Num primeiro passo, foi necessário coletar todos os dados necessários para a análise, de forma a que fosse possível perceber quais os objetivos do negócio, e verificar também a qualidade dos dados presentes na base de dados, na nossa fonte de dados. Depois, modelou-se os modelos de mineração consoante os objetivos de negócio. O último passo permitiu avaliar os resultados obtidos através destes modelos para perceber o nível de confiabilidade dos mesmos.

Assim, os objetivos da presente dissertação foram os seguintes:

1. Definir o conjunto de dados a utilizar na análise, atendendo aos objetivos de negócio.
2. Compreender e preparar os dados, procedendo à sua limpeza e possível transformação para a aplicação dos algoritmos.
3. Aplicar os algoritmos de classificação, segmentação e associação ao conjunto de dados.
4. Avaliar os resultados obtidos e o desempenho dos modelos obtidos através dos algoritmos.
5. Conferir entre os modelos gerados, os que apresentam melhor desempenho, tendo em conta os objetivos do problema.

1.3 Organização da Dissertação

Para além do presente capítulo, esta dissertação encontra-se organizada em mais 6 capítulos, nomeadamente:

- **Capítulo 2 – Processos de Análise em Ambiente Empresarial** – Neste capítulo são apresentados processos de análise em ambiente empresarial, o conceito de mineração de dados, as técnicas de mineração que serão aplicadas no caso de estudo, classificação, segmentação e associação, que são explicadas detalhadamente, e, é exposta a metodologia de mineração de dados adotada, CRISP-DM. É também abordado o tema *profiling*, explicando-o e mostrando os seus benefícios, de seguida, esclarece-se qual o tipo de conteúdo que deve constar nos perfis, clarificando, por fim, a forma como se deve obter o conteúdo presente nos perfis.

- **Capítulo 3 – Caso de Estudo** – Neste capítulo é descrito o processo de análise e preparação de dados. Inicialmente, explica-se a análise de negócio efetuada para o agente comercial e para o optometrista. Depois, é também abordada a análise efetuada aos dados que serão utilizados, bem como a preparação que os mesmos sofreram para que se pudesse transitar para a fase de construção dos modelos de mineração pretendidos.
- **Capítulo 4 –Modelação dos Modelos de Mineração de Dados** – Aqui são apresentados os modelos de mineração de dados que foram elaborados. Apresenta-se o processo de modelação de cada modelo, bem como os atributos utilizados em cada modelo.
- **Capítulo 5 – Análise e Avaliação dos Modelos de Mineração de Dados** – No capítulo anterior foram gerados os modelos de mineração de dados, enquanto que neste capítulo efetua-se a análise e a avaliação dos resultados obtidos.
- **Capítulo 6 – Conclusões e Trabalho Futuro** – Neste último capítulo retiram-se conclusões de todo o trabalho desenvolvido ao longo da dissertação, e apresentam-se considerações a ter em conta, no futuro, de modo a dar seguimento ao trabalho realizado.

CAPÍTULO 2

2. PROCESSOS DE ANÁLISE EM AMBIENTE EMPRESARIAL

2.1 Processos de Análise Empresariais

Com um mercado cada vez mais competitivo e complexo, hoje é quase obrigatório aplicar métodos e tecnologias de análise de dados, para que qualquer empresa possa interpretar os sinais que o mercado dá, de modo a preparar estratégias que possibilitem uma rápida reação as constantes transformações. Atualmente, as empresas lidam com um problema inédito, visto que nunca se tinham armazenado quantidades tão grandes de dados como se faz agora, sendo necessário desenvolver mecanismos que realizem uma análise de dados precisa e efetiva.

Todas as decisões estratégicas assertivas são dotadas de grande confiabilidade e agilidade, requisitos esses que são alcançados quando a análise de dados é bem concretizada e se consegue compreender e explorar todos os cenários que resultam da referida análise. Daí a necessidade, a nível interno, de se optar por ter uma equipa com profissionais especializados, capazes de desenvolver metodologias consolidadas e soluções que apoiem todas as estratégias de negócio.
Existem técnicas, como por exemplo a estatística, que ajudam na análise de informações para a tomada de decisões, a nível profissional ou pessoal, num futuro próximo. No entanto, nos últimos anos a mineração de dados é a técnica que mais tem emergido e sofrido uma evolução muito significativa, tornando-se numa das mais importantes "ferramentas" na análise de dados.

2.2 Mineração de Dados

A origem da mineração de dados está relacionada com o início do armazenamento de dados e com a procura contínua de melhorar o acesso a esses dados. Na primeira fase da mineração de dados, fase

denominada por coleta de dados, os dados eram recolhidos através do endereço eletrónico, e era possível utilizar esses dados para fazer cálculos simples, tais como somatórios e médias. Estes dados permitiam dar resposta a questões comerciais, como, qual a receita total ou a receita total média durante um período de tempo. Com o evoluir da mineração de dados, tem-se uma nova fase designada por acesso aos dados. Nesta fase, os dados encontram-se armazenados em bancos de dados estruturados, formulam-se políticas para coleta de dados, e pode-se, inclusive, produzir relatórios de gestão de informações. Durante as últimas décadas, a mineração de dados tem-se desenvolvido muito através da pesquisa em áreas como a estatística, a inteligência artificial ou a aprendizagem supervisionada, levando a que nos dias de hoje, as várias tecnologias desenvolvidas sejam capazes, de quando emparelhadas com sistemas de bases de dados, gerem um ambiente de negócio capaz de capitalizar o conhecimento, anteriormente desconhecido (Zuber *et al.*, 2013).

Com o aumento da informação disponível, torna-se imprescindível recorrer a técnicas que possibilitem a análise aos dados presentes em grandes bases de dados. Por vezes, a mineração de dados é referenciada como sinónimo de descoberta de conhecimento a partir dos dados, também conhecido como, *Knowledge Discovery from Data* (KDD), ou é considerada como uma etapa importante no processo de descoberta de conhecimento. Este processo, divide-se em 7 passos (Figura 2.1), que são seguidos como uma sequência interativa, sendo que os primeiros 4 passos dizem respeito ao processamento de dados, onde os dados devem ser preparados para a mineração, sendo depois aplicados modelos de mineração através dos quais são descobertos padrões que devem ser avaliados e, por fim, considerados como conhecimento (Han *et al.*, 2011).

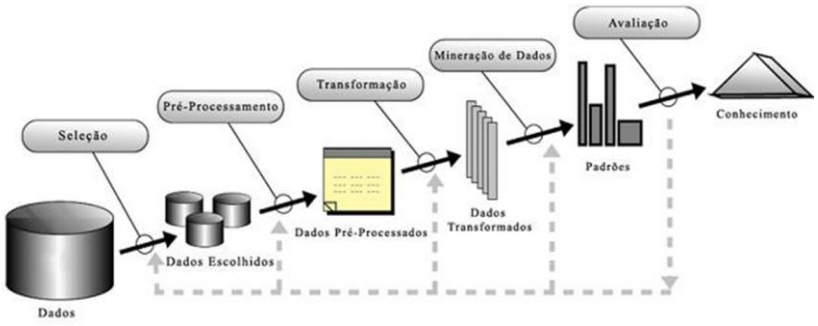


Figura 2.1 - Processo de descoberta de conhecimento a partir dos dados - KDD (figura extraída de Zuber *et al.*, 2013, p. 667)

Várias são as definições que hoje podemos encontrar para mineração de dados. Veja-se, por exemplo as seguintes:

- *"Data Mining is a step in the KDD process consisting of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns over the data"* (Fayyad *et al.*, 1996);
- *"Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically"* (Han *et al.,* 2011);
- *"Data mining is the process of extracting hidden knowledge from large volumes or raw data. This process includes collection and mining of records on a continuous basis of large volumes of transactions. The goal of this technique is to find pattern that were previously unknown. Once these pattern are found they can further used to make better decisions for their businesses"* (Singh, 2012).

A mineração de dados é uma solução cada vez mais procurada para assegurar a segurança interna, quer por empresas do setor público ou privado, uma vez que que pode ser aplicada de modo a detetar fraudes, avaliar riscos e retalhistas de produtos, através do uso de ferramentas de análise de dados, nomeadamente algoritmos estatísticos, que têm a capacidade de descobrir padrões e relações em conjuntos de dados volumosos, anteriormente desconhecidos. A mineração de dados utiliza diversas técnicas de análise como os modelos bayesianos, as árvores de decisão, as redes neurais artificiais, associação de regras ou algoritmos genéticos (Zuber *et al.*, 2013). Usualmente, durante um processo de mineração de dados, existem três etapas essenciais:

- **Exploração**, na qual se efetua o tratamento dos dados, e com base no problema, determina-se a natureza dos dados.
- **Identificação de padrões,** de entre um conjunto de padrões obtidos, escolhe-se os que fazem uma melhor previsão;
- **Implementação,** tendo em conta o objetivo final, implementa-se os padrões que dão a melhor resposta.

Apesar das vantagens da mineração de dados, esta apresenta algumas limitações, nomeadamente, o facto de ajudar a revelar padrões e relacionamentos desconhecidos nos dados, não explica qual a importância dos padrões descobertos, ou que, apesar de identificar ligações entre comportamentos e variáveis, não permite identificar claramente uma relação causal. Para que a mineração de dados tenha

sucesso, é necessário o apoio de especialistas analíticos capazes de estruturar a análise e interpretar o resultado obtido pela mineração de dados, de forma a que as conclusões obtidas possam ser utilizadas no desenvolvimento de novas tomadas de decisão (Zuber *et al.*, 2013). É necessário especificar os tipos de padrões que se espera encontrar nas tarefas de mineração de dados, sendo assim utilizadas funcionalidades de mineração de dados. Estas podem ser classificadas em duas diferentes classes: descritiva, na qual se carateriza as propriedades dos dados num conjunto de dados de destino, e preditiva, em que se induzem os dados para fazer previsões para o futuro (Han *et al.,* 2011).

Devido a importância da mineração de dados, esta tem sido utilizada em diversas áreas tais como, saúde, marketing, engenharia, análise criminal, previsão especializada, entre outras. De seguida, destacam-se algumas das áreas nas quais a aplicação da mineração de dados tem um papel relevante na resolução de problemas ou na criação de novas estratégias (Zuber *et al.*, 2013):

- Na área do retalho, os retalhistas através dos cartões de crédito e dos sistemas de ponto de venda com a marca da loja obtêm transações de cada compra de forma detalhada, possibilitando um melhor conhecimento dos diferentes segmentos de clientes. Podem assim aplicar técnicas de mineração de dados para realizar a análise de cesta, isto é, quais os produtos que o cliente tende a comprar em conjunto, fazer previsão de vendas, permitindo que o retalhista assegure que terá em stock a quantidade de produto necessária para fazer face à procura, entre outros aspetos que beneficiarão o negócio.
- No setor bancário, a mineração de dados pode ser útil na deteção de fraude através de análises efetuadas sobre transações históricas que se concluíram como fraudulentas, sendo possível identificar padrões sobre as transações, ou aplicar técnicas de mineração em outros problemas detetados.
- No ramo das telecomunicações, com o objetivo de fidelizar o cliente, os analistas usam a mineração de dados para traçar o perfil dos clientes que se manterão fidelizados apesar das propostas aliciantes da concorrência, direcionando maiores investimentos para esses clientes, que consequentemente são os que geram mais lucro para a empresa.

Existem outras aplicações, que também fazem uso da mineração de dados para se sobressaírem em relação a aplicações concorrentes, segmentando os diferentes clientes e oferecendo benefícios diferenciados aos segmentos de clientes detetados.

2.3 Tarefas de Mineração de Dados

Como referido, existem, basicamente, seis técnicas de mineração de dados. Na presente dissertação serão abordadas apenas as seguintes: classificação, segmentação e associação.

2.3.1 Classificação

De todas as técnicas de mineração de dados, a classificação é a técnica mais vulgar. Esta emprega um conjunto de exemplos pré-classificados, de forma a que seja desenvolvido um modelo capaz de obter registos de toda a população. Pode ser descrita como uma análise realizada sobre os dados capaz de extrair modelos, ou classificadores, que descrevem as classes categóricas mais importantes dos dados (Han *et al.*, 2011). Para este tipo de análise, as aplicações de deteção de fraude e risco de crédito são as mais vulgares. Por exemplo, as empresas de marketing utilizam a classificação para que as tomadas de decisão evitem, automaticamente, as fraudes e os riscos (Singh, 2012).

A classificação pode ser equiparada à estimação, existindo uma diferença na variável de destino, visto que esta é categórica ao invés de ser numérica. O modelo de mineração de dados analisa um conjunto de dados, no qual cada registo contém informação sobre a variável de destino, mas também um conjunto de variáveis de entrada, também chamadas de previsão (Larose & Larose, 2014). Geralmente, a técnica de classificação aplica algoritmos de classificação por indução de árvores de decisão ou de redes neurais. O processo de classificação envolve duas diferentes fases: aprendizagem e classificação. Na fase da aprendizagem, também designada por fase de treino, é a fase em que o modelo de classificação é construído e os dados de treino são analisados pelo algoritmo de classificação. A regra de classificação é criada pelo algoritmo de classificação através da análise de um conjunto preestabelecido de classes, designado de conjunto de treino. Na fase de classificação, o modelo é utilizado para prever classes de dados e estima-se a precisão das regras de classificação através dos dados de teste, com a condição de só aplicar as regras de classificação, caso a precisão seja aceitável. A precisão de uma regra de classificação é a percentagem de classes do conjunto de testes que são classificadas corretamente (Han *et al.*, 2011) (Zuber *et al.*, 2013).

Uma técnica de classificação pode ser aplicada de diferentes formas, tais como (Larose & Larose, 2014):

- avaliar se existe fraude numa determinada transação com cartão de crédito;
- determinar qual o ano em que um aluno com necessidades especiais deve ser colocado;
- dado um pedido de hipoteca, avaliar se o risco de crédito é prejudicial ou não;

- compreender se o testamento de uma pessoa foi escrito pela própria, ou se alguém cometeu fraude;
- determinar se um dado comportamento (financeiro ou pessoal) deve ou não ser considerado criminoso.

2.3.2 Segmentação

A segmentação, também conhecida por agrupamento ou *clustering*, é um processo de particionamento de um conjunto de dados em subconjuntos, efetuado por um algoritmo de *clustering*, isto é, um processo que agrupa um conjunto de objetos de dados em vários grupos, designados por *clusters*, cujo objetivo é os objetos que sejam altamente semelhantes num mesmo grupo, mas que sejam diferentes de objetos que estão em outros grupos. O conjunto de grupos, ou *clusters*, originados pelo processo de análise pode ser referenciado como *clustering*. No entanto, os clusters obtidos podem variar dependendo do método aplicado (Han *et al.*, 2011). Como tipos de métodos de clusters, podemos ter os: métodos de particionamento, métodos hierárquicos aglomerativos, métodos baseados em densidades, métodos baseados em rede ou métodos baseados em modelos (Singh, 2012) (Zuber *et al.*, 2013).

Através do uso de técnicas de segmentação, é possível identificar regiões dispersas e densas no espaço de objetos de dados, descobrindo o padrão mais comum da distribuição e a correlação entre os diferentes atributos de dados. A classificação, técnica abordada anteriormente, pode também ser usada para diferenciar grupos de objetos de dados, no entanto, é dispendiosa, sendo que a segmentação pode ser utilizada como pré-processamento para seleção e classificação dos subconjuntos de dados (Zuber *et al.*, 2013). A segmentação é conhecida por aprendizagem não supervisionada, visto que esta não aprende por exemplos, como a classificação, mas por observação (Han *et al.*, 2011). A técnica de segmentação difere da de classificação, no que se refere à variável de destino, uma vez que esta não existe. Por isso, contrariamente à classificação, esta técnica não tenta classificar, estimar ou prever o valor dessa variável. O objetivo do algoritmo de *clustering* é segmentar o conjunto de dados em subconjuntos homogéneos, maximizando a similaridade dos objetos do mesmo cluster e minimizando a similaridade com outros clusters (Larose & Larose, 2014). A avaliação das similaridades e dissimilaridades é baseada nos valores dos atributos, mas pode também envolver frequentemente algumas medidas de distância (Han *et al.*, 2011).

A segmentação é aplicada, recorrentemente, em diversas áreas, tais como a *business intelligence*, o reconhecimento de padrões de imagem, a biologia, entre outras, com o objetivo de reconhecer um padrão de variabilidade entre os indivíduos ou objetos que são o alvo do estudo. Este facto deve-se à necessidade crescente de transformar a informação contida em grandes quantidades de dados em informação útil, capaz de sustentar decisões (Han *et al.*, 2011). Algumas aplicações da tarefa de segmentação em negócios, abrangem (Larose & Larose, 2014):

- segmentar o marketing de um produto de nicho para uma pequena empresa que não possui grande orçamento de marketing;
- segmentar o comportamento financeiro em diferentes categorias, para fins de auditorias de contabilidade;
- reduzir a dimensão de um conjunto de dados com muitos atributos.

2.3.3 Associação

Devido ao elevado número de dados que são coletados e armazenados, cada vez mais as diferentes áreas de negócio têm interesse em extrair padrões dos seus dados. Assim, ao descobrirem-se relações de correlação interessantes entre o conjunto de dados, estas transformam-se em conhecimento que ajuda as empresas na tomada de novas decisões. Se se considerar o universo como um conjunto de itens disponíveis numa dada loja, então cada item tem associado uma variável booleana que confirma a presença ou a inexistência desse item. Assim, uma cesta de produtos pode ser representada por um vetor booleano com os valores que foram atribuídos às variáveis. O objetivo desses vetores booleanos é serem analisados por padrões de compra que mostram quando é que os itens são comprados em conjunto ou se são frequentemente associados. Esses padrões são apresentados na forma de regras de associação (Han *et al.*, 2011).

A técnica de associação é também conhecida como análise de afinidade ou análise da cesta do mercado, e tem como objetivo encontrar regras que quantifiquem a relação entre dois ou mais atributos, comprovadas por uma medida de suporte e de confiança associada à regra (Larose & Larose, 2014). Comummente, a associação e a correlação são ideais para encontrar conjuntos de itens frequentes entre os grandes conjuntos de dados transacionais ou relacionais (Han *et al.*, 2011). Um algoritmo de associação deve ser capaz de gerar regras cujo valor de confiança seja menor que um. Geralmente, esse algoritmo retorna um elevado número de regras de associação. Grande parte das mesmas têm pouco ou nenhum valor. Os tipos de regras de associação que existem são de: associação multinível, associação

multidimensional, ou associação quantitativa (Singh, 2012) (Zuber *et al.*, 2013). Para se considerar uma regra de associação interessante, deve-se definir um limite mínimo de suporte e um limite de confiança mínimo. Adicionalmente, pode-se realizar uma análise que evidencie correlações estatísticas interessantes entre itens associados (Han *et al.*, 2011).

Como exemplos de aplicação da técnica de associação, tem-se, por exemplo (Larose & Larose, 2014):

- prever a degradação de redes de telecomunicações;
- num contexto de retalho, descobrir os itens que são frequentemente comprados em conjunto e os que nunca foram;
- determinar a proporção de casos em que uma nova droga apresenta efeitos colaterais perigosos.

2.4 Técnicas de Mineração de Dados – Algoritmos Selecionados

2.4.1 Classificação – *Decision Trees*

O algoritmo de classificação que mais é aplicado ao nível da mineração de dados é o de árvores de decisão. Este facto é explicado por este algoritmo conseguir dar suporte a diferentes tipos de caraterísticas, categóricas ou numéricas, bem como a facilidade de compreender o conhecimento que as árvores de decisão oferecem ao utilizador, e, ainda, devido ao processo de aprendizagem e aquisição de conhecimento ocorrer rapidamente, quando comparado a outros algoritmos existentes.

As árvores de decisão são modelos de classificação e a sua estrutura compreende um conjunto de nós e ramificações. A estrutura de uma árvore de decisão é semelhante a um fluxograma, na qual cada nó interno, designado também por nó probabilidade, representa um teste sobre um determinado atributo, exibindo a respetiva probabilidade, cada ramificação exibe o resultado desse teste e cada nó terminal, conhecido também por nó folha, apresenta o resultado final de um determinado caminho de decisões ou classes de classificação. O nó inicial, que está mais a esquerda, é designado por nó raiz (Han *et al.,* 2011).

O primeiro passo quando se inicia a construção de uma árvore de decisão é escolher o atributo que será utilizado como nó de raiz. Um dos critérios a considerar nessa escolha, é que cada divisão na árvore deve ser feita da forma mais pura possível, ou seja, grande parte das instâncias alocadas a um subconjunto devem pertencer a uma única classe (Carvalho, 2014). A árvore de decisão deve ser lida da

esquerda para a direita, devendo-se dar particular atenção a todos os nós nos quais se obtém satisfação aos testes que são realizados até se alcançar um nó terminal, no qual se encontra a nova classificação da instância. Quando se tem como objetivo classificar um item específico de dados, a análise deve iniciar no nó raiz e deve seguir as ramificações até se chegar a um nó terminal, permitindo assim suportar uma decisão. No entanto, as árvores de decisão podem ser interpretadas como um conjunto especial de regras, caraterizadas pela hierarquia existente nessas regras (Gupta *et al.,* 2011).

Depois de construída a árvore de decisão, é possível que algumas ramificações sejam afetadas por ruído, *outliers* que estão presentes nos dados de treino. O modelo construído pode ser muito específico, isto é, pode possuir uma folha para cada classe do subconjunto de dados de treino, o que revela uma enorme precisão do modelo, mas torna-se ineficiente quando recebe instâncias do conjunto de dados de teste. Através de métodos de remoção que utilizam medidas estatísticas para remover ramos menos confiáveis, consegue-se resolver este problema sem sobrecarregar os dados. Geralmente, aplicam-se duas abordagens para se realizar o processo de poda: a pré-poda e a pós-poda. Depois de podadas, as árvores têm tendência a serem menos complexas e menores, facilitando assim a sua compreensão (Han *et al.*, 2011) (Carvalho, 2014).

Na primeira abordagem, pré-poda, a poda da árvore é feita durante a sua construção através da aplicação de um determinado critério. Contudo, é necessário ter em atenção o critério escolhido, visto que um critério fraco pode podar pouco a árvore, enquanto que um critério forte pode gerar árvores simplificadas em demasia, prejudicando, consequentemente, a precisão do modelo. A segunda abordagem é a pós-poda. Esta é a mais comum. Basicamente, esta consiste em construir a árvore de decisão e depois substituir determinadas subárvores por um nó folha que deve ser a classe mais comum da subárvore que se está a substituir. Como critério de decisão sobre quais subárvores devem ser excluídas, calcula-se a estimativa de erro num determinado nó que deve ser o nó pai (N) da subárvore, comparando à estimativa de erro da subárvore, e caso o erro de N seja menor então a subárvore é podada. Na Figura 2.2, apresenta-se um exemplo de uma árvore de decisão completa e a sua versão após se efetuar a poda (Han *et al.*, 2011).

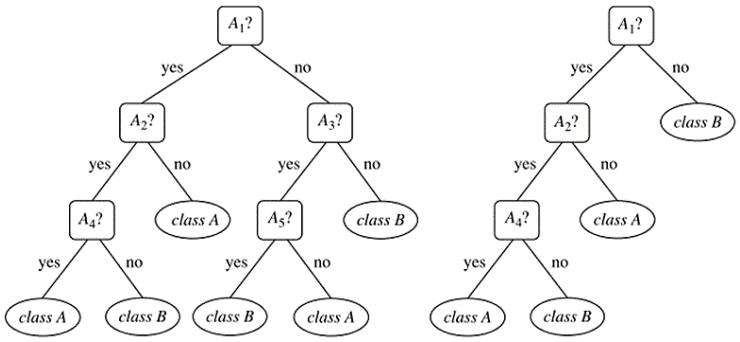


Figura 2.2 - Árvore completa e respetiva versão após se efetuar a poda (figura extraída de Han *et al.,* 2011, p.345)

A entropia é uma medida que calcula a "impureza" num determinado subconjunto, sendo assim possível classificar uma determinada instância através do mínimo de informação necessária. Toma o valor zero quando todos os objetos de S são do mesmo valor, e o valor de 1 quando S tem igual quantidade de objetos negativos e positivos. A expressão que define a entropia de S, dado um conjunto de dados S com caraterísticas negativas e positivas de uma determinada instância é:

$$Entropia\ (S) = \ -p_+ \log_2 p_+ - \ p_- \log_2 p_-$$

em que p_- é a proporção objetos negativos num conjunto de S e p_+ é a proporção objetos positivos num conjunto de S.

Existem quatro termos adicionais que são aplicados no cálculo de muitas medidas de avaliação e que ajudam a compreender o significado das várias medidas, são eles:

- **Verdadeiros Positivos (TP)** – número de verdadeiros positivos, que são tuplos positivos que foram corretamente classificadas pelo classificador.
- **Negativos Positivos (TN)** – número de negativos positivos, que são tuplos negativos que o classificador classificou de forma correta.
- **Falsos Positivos (FP)** – número de falsos positivos, que são tuplos negativos que foram classificados como positivos incorretamente.
- **Falsos Negativos (FN)** – número de falsos negativos, que são tuplos positivos classificados incorretamente como negativos.

Estes termos são apresentados na matriz de confusão, representada na Figura 2.3. A matriz de confusão, também conhecida por matriz de classificação, analisa a forma como o classificador está a reconhecer

os tuplos de classes diferentes. Os dois primeiros termos indicam que o classificador está a classificar os diferentes tuplos corretamente, isto é, as predições estão corretas, e os dois últimos termos indicam o contrário, quando as predições estão erradas (Han *et al.*, 2011).

		Predicted class		
		yes	*no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	P'	N'	$P + N$

Figura 2.3 - Matriz de confusão (figura extraída de Han *et al.*, 2011, p.366)

Uma matriz de confusão é do tamanho $m \times m$, em que m terá sempre de ser, obrigatoriamente, igual ou maior a dois. Cada entrada i, j indica, nas m linhas e m colunas o número de tuplos que pertencem à classe i que o classificador considerou como sendo da classe j. Idealmente, a maioria dos tuplos deveria estar representada na diagonal e as restantes entradas deveriam assumir valor zero ou muito próximo de zero. Como medidas de avaliação, podemos ter a precisão, a taxa de erro de classificação, a sensibilidade e a especificidade (Han *et al.*, 2011). A precisão do classificador é a percentagem de pares do conjunto de teste que o classificador classifica corretamente, reflete quanto o classificador conhece os tuplos das diferentes classes. Quando se analisa uma matriz de confusão, é fácil de compreender se o classificador está a classificar corretamente ou se está a confundir as duas classes. A precisão é calculada pela seguinte fórmula:

$$Precisão = \frac{TP + TN}{P + N} = \frac{Número\ de\ predições\ corretas}{Número\ total\ de\ predições}$$

A taxa de erro do classificador é a percentagem de tuplos que são classificados incorretamente pelo classificador, ou seja, este cálculo é igual a:

$$Taxa\ de\ Erro = \frac{FP + FN}{P + N} = \frac{Número\ de\ predições\ erradas}{Número\ total\ de\ predições} = 1 - Precisão$$

A sensibilidade calcula a taxa positiva verdadeira, isto é, a proporção de tuplos positivos que estão classificados corretamente. A especificidade representa a taxa negativa verdadeira, ou seja, a proporção de tuplos negativos que estão corretamente identificados. Estas medidas são definidas por:

$$Sensibilidade = \frac{TP}{P}$$

$$Especificidade = \frac{TN}{N}$$

A precisão pode ser calculada em função da sensibilidade e especificidade, tal como mostra a seguinte fórmula:

$$Precisão = Sensibilidade\ \frac{P}{(P+N)} + Especificidade\ \frac{N}{(P+N)}$$

Além das medidas baseadas em precisão, os classificadores podem ser comparados em relação a outros aspetos, tais como:

- **Velocidade**, os custos computacionais relacionados com a geração e uso do classificador.
- **Robustez**, a capacidade de o classificador realizar previsões corretas. É avaliada através de uma série de conjuntos de dados sintéticos com o intuito de representar o crescimento de dados ruidosos ou valores ausentes.
- **Escalabilidade**, a capacidade de construir um classificador de forma eficiente em grandes volumes de dados. Avalia-se através de conjuntos de dados que vão aumentado gradualmente.
- **Interpretabilidade**, o nível de compreensão fornecido pelo classificador. Apesar de as árvores de decisão serem de fácil interpretação, esta é menor quanto mais complexas forem as árvores.

2.4.2 Segmentação – *Clustering*

Na literatura, existe um vasto conjunto de algoritmos de segmentação, de tal modo que é complicado categorizar nitidamente os métodos de *clustering*, visto que essas mesmas categorias se podem sobrepor e que um método pode ter recursos de várias categorias. Ou seja, um algoritmo de *clustering* pode integrar várias ideias de diferentes métodos de segmentação. Por esse motivo, não é fácil classificar um dado algoritmo como pertencente a uma dada categoria de um método de clustering unicamente. Além disso, existem aplicações cujos critérios de segmentação exigem a integração de várias técnicas. Na Tabela 2-1, apresentam-se resumidamente os diferentes métodos e as suas caraterísticas gerais, notando que alguns destes combinam técnicas dos diferentes métodos (Han *et al.*, 2011).

Tabela 2-1 - Resumo dos diferentes métodos de segmentação (tabela adaptada de Han *et al.*, 2001, p. 450)

Método	**Caraterísticas Gerais**

Método de Particionamento	- Encontra mutuamente exclusivos de forma esférica; - Baseado em distância; - Pode usar média ou medóide (entre outros) para representar o centro do *cluster*; - Eficaz para conjuntos de dados pequenos e médios.
Método Hierárquico	- *Clustering* é uma decomposição hierárquica (isto é, vários níveis); - Não é possível corrigir fusões ou separações incorretas; - Pode incorporar outras técnicas como *microclustering* ou considerar "ligações" de objetos.
Métodos Baseados em Densidade	- Pode descobrir clusters arbitrariamente; - Os *clusters* são regiões densas de objetos no espaço, separadas por regiões de baixa densidade; - Densidade do *cluster*: cada ponto deve ter um número mínimo de pontos dentro da sua "vizinhança"; - Pode filtrar calores extremos.
Métodos Baseados em Grade	- Usa uma estrutura de dados em grade com várias soluções; - O tempo de processamento é rápido (tipicamente, é independente do número de objetos de dados, mas dependente do tamanho da grade).

Nesta dissertação, foram aplicados dois algoritmos de *clustering* diferentes: *k-means* e *expectation-maximization*. Veja-se cada um deles em particular.

O Algoritmo K-means

Considerando um conjunto de dados, D , que contém no seu espaço euclidiano n objetos. Os métodos de particionamento separam os objetos do conjunto D em k *clusters*, C_1, ..., C_k, o que quer dizer que $C_i \subset D$ e que $C_i \cap C_j = \emptyset$, sempre que $i \geq 1, j \leq k$. Para avaliar a qualidade do particionamento utiliza-se uma função objetivo. Esta tem como propósito assegurar que os objetos dentro do mesmo *cluster* sejam semelhantes, e quando comparados a objetos de outros clusters sejam diferentes, ou seja, deve existir alta similaridade intracluster e baixa similaridade intercluster. A técnica de particionamento pode ser baseada em centroides, em que cada *cluster* (C_i) utiliza o seu centroide (ponto central do cluster) para se representar. Define-se o centroide como a média ou *medoid* dos objetos aglomerados, que é medido pela distância euclidiana atribuída ao *cluster*, representada por $dist(p, c_i)$, em que p é um ponto que pertence ao cluster C_i, e c_i o respetivo centroide desse cluster. A qualidade de um cluster pode ser avaliada pela variação dentro do cluster, esta é a soma do erro quadrado entre todos os objetos no cluster C_i e o centroide c_i, tal como:

$$E = \sum_{i=1}^{k} \sum_{p \in C_I} dist(p, c_i)^2$$

em que E é a soma do erro quadrado de todos os objetos no conjunto de dados, p é o ponto que representa um determinado objeto e c_i é o centroide do *cluster* C_i. A distância do objeto ao centro do

cluster é quadrada e as distâncias são somadas, para cada objeto em cada cluster. Otimizar a variação que existe dentro do cluster é um desafio computacional. O algoritmo *k-means* (Figura 2.4) é simples e é frequentemente usado para combater essa variação. Este define o centroide como sendo o valor médio dos pontos que pertencem a um cluster. Basicamente, o algoritmo atua da seguinte maneira (Han *et al.,* 2011):

1º Aleatoriamente, seleciona-se k dos objetos contidos em D , e cada objeto representa a média ou centro do *cluster*.

2º Para os restantes, cada objeto é atribuído ao *cluster* ao qual mais se assemelha, baseando-se na distância euclidiana entre o objeto e a média desse *cluster*.

3º De forma iterativa, o algoritmo *k-means* melhora a variação dentro do *cluster*. Para cada *cluster*, calcula a nova média utilizando os objetos que foram atribuídos a esse *cluster* na iteração anterior.

4º Todos os objetos são redistribuídos, à medida que o novo *cluster* é centralizado.

5º As iterações só terminam quando a atribuição é estável, ou seja, quando os *clusters* criados na iteração são iguais aos da iteração anterior.

Algoritmo: *k-means*. O algoritmo *k-means* para particionamento, onde o centro de cada *cluster* é representado pelo valor médio dos objetos do *cluster*.
Entrada:
- k: número de *clusters,*
- D : conjunto de dados contendo n objetos.

Saída: Conjunto de k *clusters*
Método:
(1) escolhe arbitrariamente k objetos de D como os centros de *cluster* iniciais;
(2) **repetir**
(3) (re) atribuir cada objeto ao *cluster* no qual os objetos são mais semelhantes, com base no valor médio dos objetos de cada *cluster*;
(4) atualizar os meios do *cluster*, isto é, calcular o valor médio dos objetos para cada *cluster*;
(5) **até** que não ocorra mudança;

Figura 2.4 - Algoritmo de particionamento k-means (figura adaptada de Han *et al.,* 2011, p. 452)

O Algoritmo Expectation-Maximization

Como foi explicado, o algoritmo *k-means* faz iterações até que exista estabilidade nos *clusters*, isto é, até que não seja possível fazer melhorias nos dados. Cada uma dessas iterações pode ser descrita em duas etapas: expectativa e maximização. Na primeira etapa (*E-step*), o objeto é atribuído ao *cluster* cujo centro

seja ao mais próximo do objeto. Na segunda etapa (*M-step*), tendo em conta a atribuição de cada *cluster*, o algoritmo ajusta o centro do *cluster* de modo a minimizar a soma das distâncias dos objetos atribuídos comparado ao novo centro do *cluster*, maximizando a semelhança dos objetos do mesmo *cluster*. Este método composto por duas etapas pode ser generalizado para ser aplicado em *fuzzy clustering* (*clustering* difuso) e em *clustering* baseado em modelos probabilísticos. O algoritmo expectation-maximization (EM) aproxima-se de estimativas de máxima verossimilhança ou estimativas máximas à posteriori de parâmetros de modelos estatísticos. Este algoritmo inicia a sua execução com um conjunto de parâmetros e faz iterações até que não ocorram novas melhorias no *clustering*. As iterações correspondem a duas etapas (Han *et al.,* 2011):

- **Etapa da expectativa, ou *E-step*.** Nesta etapa, os objetos são atribuídos tendo em conta o seu agrupamento difuso atual ou com base em parâmetros de *clusters* probabilísticos.
- **Etapa da maximização, ou *M-step*.** Aqui, o algoritmo maximiza a soma dos erros quadrados no agrupamento difuso ou a probabilidade esperada no caso de o *cluster* ter por base modelos probabilísticos.

As iterações são repetidas até que os centros dos *clusters* convirjam ou até que a alteração seja suficiente pequena, para isso, deve ser definido um limite.

2.4.3 Associação – *Association Rules*

Regras e suas Propriedades

Assumamos que $I = \{I_1, I_2, \dots, I_m\}$ é um conjunto de itens e D um conjunto de transações, em que cada transação T é constituída por um conjunto de itens, tal que $T \subseteq I$. Assim, uma regra de associação é uma implicação na forma $A \Rightarrow B$, onde $A \subset I, B \subset I, A \neq \emptyset, B \neq \emptyset \; e \; A \cap B = \emptyset$ (Srikant *et al.,* 1997). Todas as regras que satisfaçam limites mínimo de suporte (*min_supp*) e mínimo de confiança (*min_conf*) são consideradas fortes (Agrawal & Srikant, 1994).

Vejamos, com um pouco mais de detalhe, algumas das principais propriedades de uma regra. Uma regra tem uma medida de suporte no conjunto de transações D, que representa a percentagem de transações em D que contêm $A \cup B$, ou seja, os conjuntos A e B juntos, sendo considerada a $P(A \cup B)$. A fórmula que calcula o suporte da regra é:

$$supp(A \Rightarrow B) = P(A \cup B)$$

Além do suporte temos a confiança. A confiança da regra no conjunto de transações D, é a percentagem de transações de D em que está contido A mas também B, ou seja, tem-se a probabilidade condicional $P(B|A)$. A confiança é calculada por:

$$conf(A \Rightarrow B) = \frac{supp(A \Rightarrow B)}{supp(A)} = \frac{supp(A \cup B)}{supp(A)} = P(B|A)$$

Depois, além do suporte e da confiança, temos a Lift. A medida de interesse *lift*, conhecida também por *interest*, é das mais aplicadas para avaliar dependências. Numa regra de associação $A \Rightarrow B$, indica com que frequência ocorre B quando aconteceu A. Esta medida toma valores entre zero e infinito, e quanto maior for o valor de *lift,* significa que A aumentou B numa maior taxa, concluindo-se que a regra é mais interessante. Se o valor de $lift(A \Rightarrow B)$ for igual a 1, significa que A e B são independentes, se for menor que 1 são negativamente dependentes, se for maior que 1 são positivamente dependentes (Gonçalves, 2005). O valor de *lift* é calculado:

$$lift(A \Rightarrow B) = \frac{conf(A \Rightarrow B)}{supp(B)}$$

De seguida, temos a convicção de uma regra. Tal como a medida de interesse anterior, a convicção compara a independência ou dependência entre os acontecimentos contidos numa regra, no entanto, esta é sensível à direção da regra, isto é $conv(A \Rightarrow B) \neq conv(B \Rightarrow A)$ (Azevedo & Jorge, 2007). A medida de interesse convicção tem o objetivo de avaliar uma regra de associação como uma verdadeira implicação, e apresenta algumas caraterísticas, tais como, considerar o suporte do antecedente como o suporte do consequente. O valor da convicção é igual a 1, se existir completa independência entre os dois elementos da regra, e terá valor infinito sempre que o existam regras em que o antecedente aparece sempre com o consequente (Gonçalves, 2005). O valor da convicção é obtido por:

$$conv(A \Rightarrow B) = \frac{1 - supp(B)}{1 - conf(A \Rightarrow B)}$$

Além das propriedades já referidas, temos o fator de certeza. Este mede forma como varia a probabilidade de B estar presente numa transação num conjunto em que todas as transações têm B presente (Shortliffe & Buchanan, 1990). Quando $-1 \leq CF(A \Rightarrow B) < 0$ existe correlação ou dependência negativa, se $0 < CF(A \Rightarrow B) \leq 1$ então existe correlação ou dependência positiva, se o valor for igual a 0 existe dependência entre A e B. Para calcular o fator de certeza, mediante o valor de B, pode-se aplicar uma de três fórmulas:

$$CF(A \Rightarrow B) = \begin{cases} \dfrac{conf(A \Rightarrow B) - supp(B)}{1 - supp(B)}, & conf(A \Rightarrow B) > supp(B) \\ \dfrac{conf(A \Rightarrow B) - supp(B)}{supp(B)}, & conf(A \Rightarrow B) < supp(B) \\ 0, & conf(A \Rightarrow B) = supp(B) \end{cases}$$

O Algoritmo *Apriori*

O algoritmo *Apriori* utiliza o conhecimento adquirido anteriormente no conjunto de itens sobre as propriedades presentes e que são mais frequentes, daí resulta o seu nome. A abordagem deste algoritmo é iterativa e reconhecida como uma pesquisa de nível, isto é, k conjuntos de itens são utilizados para analisar $k + 1$ conjuntos de itens (Han *et al.,* 2011).

Na primeira iteração do algoritmo, este faz a contagem de ocorrências do item para determinar o primeiro conjunto de itens. A função *apriori-gen* tem como argumento L_{k-1}, isto é, o conjunto de todos os grandes conjuntos de itens $(k - 1)$, retornando um superconjunto formado por todos os conjuntos de itens k. Inicialmente, os conjuntos de itens, L_{k-1}, encontrados no passo $k - 1$, são utilizados para gerar os conjuntos de itens candidatos C_k. De modo a tornar a contagem realizada pelo algoritmo mais rápida, é necessário determinar os candidatos em C_k contidos numa dada transação t, eficientemente (Agrawal & Srikant, 1994). Assim, em cada iteração do algoritmo, são somente considerados os grandes candidatos da iteração anterior para gerar um novo conjunto de candidatos que é contado na iteração atual, utilizando estruturas de dados especializadas de modo a tornar a contagem mais rápida (Zaki *et al.*, 1997). Seja L_k, os k conjuntos de itens, e C_k, os k conjuntos de itens candidatos, verifica-se que $C_k = L_{k-1} \times L_{k-1}$. Na Figura 2.5 pode-se consultar a estrutura geral do algoritmo *Apriori*.

```
1)  L_1 = {large 1-itemsets};
2)  for ( k = 2; L_{k-1} ≠ ∅; k++ ) do begin
3)     C_k = apriori-gen(L_{k-1}); // New candidates
4)     forall transactions t ∈ D do begin
5)        C_t = subset(C_k, t); // Candidates contained in t
6)        forall candidates c ∈ C_t do
7)           c.count++;
8)     end
9)     L_k = {c ∈ C_k | c.count ≥ minsup}
10) end
11) Answer = ⋃_k L_k;
```

Figura 2.5 - Algoritmo *Apriori* (figura extraída de Agrawal & Srikant, 1994, p. 3)

2.5 A Metodologia CRISP-DM

A metodologia que se adotou para desenvolver este trabalho, o CRISP-DM, é inclusiva da mineração de dados e fornece um modelo de processo com um plano completo que ajuda a aplicar as melhores práticas, de forma a realizar eficazmente e obter rapidamente os resultados, de modo a que um projeto de mineração de dados seja bem-sucedido. Assim, esta metodologia divide um projeto de mineração em seis fases: análise de negócio, análise de dados, preparação de dados, modelação, avaliação e implementação (Shearer, 2000).

O modelo de processo referido, dá uma visão geral do ciclo de vida de qualquer projeto de mineração de dados, mostrando as suas fases, as tarefas de cada fase e qual a relação entre cada tarefa, considerando que estes relacionamentos podem existir entre qualquer tarefa, sendo dependente das metas e interesses do utilizador, e também dos dados. Após o término de cada fase, através do resultado obtido decide-se qual a fase seguinte. Na figura 2.6 encontra-se representada a metodologia CRISP-DM, na qual as setas entre as diferentes fases indicam as dependências mais frequentes e importantes entre as mesmas, sendo que a sequência entre fases não é rígida, e as setas circulares indicam a sequência natural deste processo (Chapman *et al.*, 2000).

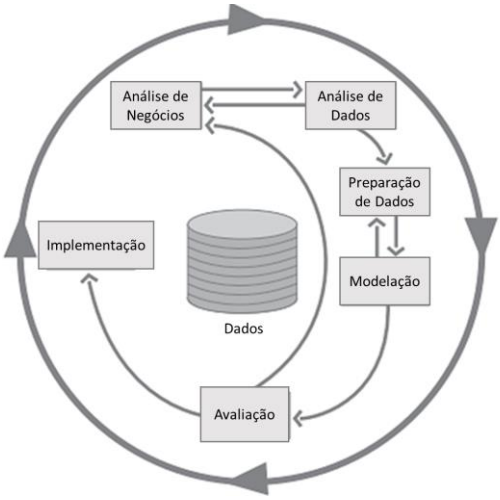


Figura 2.6 – Ciclo de vida da metodologia CRISP-DM (figura adaptada de Chapman *et al*, 2000, p. 10)

De seguida, de uma forma breve, explica-se cada uma das fases da metodologia CRISP-DM (Chapman *et al*, 2000):

1) Análise de negócios. A primeira fase tem como propósito entender os objetivos e requisitos do projeto de negócio, converter esses objetivos num problema de mineração de dados e definir um plano para atingir os objetivos.

2) Análise de dados. A segunda fase tem como tarefas a coleta de dados inicial, a realização de análises estatísticas para compreender os dados e a identificação de problemas relacionados com a qualidade dos dados.

3) Preparação dos dados. A terceira fase é considerada como a mais demorosa. Esta fase abrange todas as etapas necessárias à construção do conjunto final de dados, desde seleção, limpeza, construção, integração e formatação dos dados. Geralmente, estas tarefas são executadas várias vezes e não obedecem a qualquer ordem.

4) Modelação. Na quarta fase, selecionam-se e aplicam-se as técnicas de modelação eleitas, afinando os parâmetros para valores ótimos. Como não existe apenas uma técnica para um problema de mineração de dados, muitas vezes é necessário voltar a fase anterior, para cumprir os requisitos específicos, de cada técnica, na forma de dados. É nesta fase que é construído o modelo de mineração de dados.

5) Avaliação. A quinta fase é responsável pela avaliação e revisão das etapas do modelo de mineração de dados construído na fase de modelação, de modo a perceber se o modelo atinge os objetivos de negócio. Na fase de avaliação, decide-se qual o próximo passo, se o modelo está

adequado aos objetivos, transita-se para a sua implementação, caso contrário é necessário regressar à fase inicial.

6) Implementação. O projeto de mineração termina com a sua implementação. Independentemente de o objetivo do modelo ser apenas o aumento de conhecimento dos dados, é necessário organizar e apresentar o conhecimento ao cliente. Assim este pode ser exibido de duas formas através de um relatório ou da implementação do processo de mineração de dados.

Na Figura 2.7 estão apresentadas todas as fases da metodologia CRISP-DM, bem como as suas respetivas tarefas e *outputs* do processo de mineração de dados fornecido.

Análise de Negócios	Análise de Dados	Preparação de Dados	Modelação	Avaliação	Implementação
Determinar objetivos de Negócio Background Objetivos de Negócio Critério de Sucesso de Negócio **Avaliar a Situação** Inventário de Recursos Requerimentos, Premissas e Restrições Riscos e Contingências Terminologia Custos e Benefícios **Determinar as Metas de Mineração de Dados** Metas de Mineração de Dados Critérios de Sucesso de Mineração de Dados **Elaborar Plano de Projeto** Plano do Projeto Avaliação Inicial de Ferramentas e Técnicas	**Coleta Inicial de Dados** Relatório Inicial da Coleta de Dados **Descrever os Dados** Relatório da Descrição dos Dados **Explorar os Dados** Relatório da Exploração dos Dados **Verificar a Qualidade dos Dados** Relatório da Qualidade dos Dados	**Selecionar os Dados** Justificar a Inclusão/Exclusão **Limpar os Dados** Relatório da Limpeza de Dados **Construir os Dados** Atributos Derivados Registos Gerados **Integrar os Dados** Misturar os Dados **Formatar os Dados** Dados Reformatados **Conjunto de Dados** Descrição do Conjunto de Dados	**Selecionar as Técnicas de Modelação** Técnicas de Modelação Premissas de Modelação **Gerar o Design do Teste** Testar Design **Contruir o Modelo** Configurar os Parâmetros do Modelo Descrição do Modelo **Avaliar o Modelo** Avaliação do Modelo Configuração dos Parâmetros Revisados	**Avaliar os Resultados** Avaliação dos Resultados da Mineração de Dados em conformidade com o Critério de Sucesso de Negócio Aprovação dos Modelos **Rever o Processo** Revisão do Processo **Determinar os Próximos Passos** Lista de Possíveis Ações de Decisão	**Planear a Implementação** Plano de Implementação Planear a Monitorização e Manutenção Plano de Monitorização e Manutenção Elaborar um Relatório Final Relatório Final Apresentação Final Rever o Projeto Experiência Documentada

Figura 2.7- Fases, tarefas e outputs do processo de mineração de dados (figura adaptada de Chapman *et al.*, 2000, p. 12)

2.6 Definição de Perfis

Na atualidade, com o aumento de informação disponível, aumenta também a procura por técnicas que personalizem o acesso à informação. Assim, vários sistemas personalizados têm surgido, laborando com o problema do excesso de informação, singularizando-a para os utilizadores individuais. Esta personalização pode atuar de duas formas: filtrando informações relevantes ou identificando informações adicionais que sejam do interesse do utilizador.

Como áreas que se destacam na área da personalização da informação tem-se inteligência artificial, a mineração de dados, entre outras (Gauch *et al.*, 2007). Existem, porém, outras áreas, tais como sistemas de recomendação e aplicações de comércio eletrónico, em que é essencial criar o perfil dos seus utilizadores, obtendo dessa forma conhecimento sobre os mesmos (Schiaffino & Amandi, 2009). Segundo Scridon (Scridon, 2008) podemos definir a definição de perfis como:

> *"Profiling is exactly what it implies: the act of using data to describe or profile a group of customers or prospects. It can be performed on an entire database or distinct sections of the database. The distinct sections are known as segments. Typically they are mutually exclusive, which means no one can be a member of more than one segment".*

Uma outra definição para o perfil do utilizador é ver tal ação com um processo cujo objetivo é identificar os dados relativos a um domínio do seu interesse. A informação que se obtém pode ser empregue de modo a melhorar a satisfação do utilizador, visto que um dos aspetos mais importantes sobre o conhecimento do seu perfil é recomendar itens que correspondam aos seus gostos (Kanoje *et al.*, 2014). Para traçar o perfil completo do utilizador, é necessário dividir este processo em dois tipos de perfis: factual e comportamental. No primeiro tipo constam informações obtidas dos dados factuais dos utilizadores, tais como nome, género, data de nascimento, entre outras, podendo também conter informações derivadas dos dados transacionais. No perfil comportamental fazem parte informações sobre as suas ações e usualmente, são derivadas dos dados transacionais (Adomavicius & Tuzhilin, 2001). Assim, um perfil deve conter informação importante tal como o nome, a idade, a localização, entre outras, mas quando o contexto é de aplicações de *software* o perfil deve ser algo que vai além da informação pessoal, visto que toda a pesquisa efetuada na internet pode ser seguida, isto é, todos os detalhes da pesquiza efetuada são guardados, por exemplo, o que foi pesquisado, o que permite que possam ser combinados com os dados referentes ao tempo gasto ao analisar as informações que foram retornadas da pesquisa, obtendo assim mais informação. Através da combinação desses dados com as informações pessoais do utilizador, constrói-se um perfil mais próximo do real (Wiedmann *et al.,* 2002). Pode-se afirmar que, *"A profile is a description of someone containing the most important or interesting facts about him or her"* (Schiaffino & Amandi, 2009).

O profiling emerge como um método capaz de criar um retrato do utilizador, prestando ajuda às decisões que as empresas têm de tomar, centradas nos seus utilizadores, que afetam os seus negócios. Nos dias

de hoje, para que uma qualquer área de negócio seja bem-sucedida, compreender o seu utilizador é fundamental, inclusive, é mais fácil manter um utilizador antigo ao invés de angariar um novo (Upadhyay *et al.,* 2016). Para compor um perfil de um utilizador, recorre-se à representação fictícia do utilizador ideal, sendo esta designada de "Persona", que é criada tendo por base as experiências de vendas e a relação que têm com a venda dos produtos ou serviços prestados, sustentadas por dados históricos que retratam o comportamento dos seus utilizadores, bem como pelos dados já referidos anteriormente. De forma a tornar a "Persona" o mais próximo possível dos utilizadores reais, é importante aceder a base de dados histórica dos utilizadores, de uma marca ou de uma empresa, ou a um segmento de mercado o mais parecido possível, caso seja novo no mercado, de forma a obter dados de consumo e sazonalidade com que o utilizador vende o produto.

A criação de perfis de cliente acarreta vantagens e desvantagens para uma marca ou uma empresa. As vantagens podem ser vistas em dois aspetos: primeiro, quanto maior for o grau de conhecimento do utilizador, mais facilmente se sabe o tipo de produtos que é mais vendido e os que têm menos chance de serem vendidos. Isso faz surgir novas oportunidades de negócio, como por exemplo, fazer promoções dos produtos que apresentam menor número de vendas, ou criar campanhas especiais nos meses em que um dado produto tem maior volume de vendas. Um outro benefício, é o facto de permitir também personalizar os grupos, tendo em conta os interesses do utilizador, isto é, sugerindo produtos para um grupo específico que apresente os mesmos gostos. Por outro lado, existe uma limitação da criação de perfis, visto que, por vezes, não podem ser aplicados em situações mais amplas, no entanto, este facto não pode ser entendido como uma desvantagem, uma vez que não faz muito sentido recolher informações sobre os utilizadores quando se atua num mercado global, já que nesse caso o seu público-alvo é toda a população e não um segmento da mesma.

De todas as áreas que beneficiam com a criação de perfis, pode-se afirmar que os sistemas de recomendação são os maiores beneficiários, devido à importância que os perfis têm nas recomendações geradas. No entanto, nos dias de hoje, os perfis já não são usados unicamente pelos sistemas de recomendação, mas também por outras aplicações, tais como a pesquisa personalizada, websites adaptados, gestão da relação com o cliente, entre outras. Assim, de seguida são expostas quatro aplicações em que a criação de perfis é útil para resolver alguns problemas, nomeadamente (Kanoje *et al.*, 2014).

- **Recomendação de trabalhos de pesquisa** – O sistema desenvolvido por Jae Tang, designado por ArnetMiner, divide a tarefa de criar perfis em três: extração de perfis, integração e descoberta de interesses (Jie *et al.*, 2010). Existe uma outra abordagem idêntica, em que é criada uma etapa extra, cujo objetivo é visualizar o perfil de modo a representar o perfil gerado pelo sistema cuja abordagem adotada era ontológica (Middleton *et al.,* 2004).
- **Turismo** – Outra aplicação beneficiada pelo perfil do utilizador é um site alicerçado em e-Tourism. Este é capaz de fornecer informações aos seus utilizadores com base na sua localização. O negócio do turismo é dependente do seu posicionamento geográfico. Este sistema gera recomendações tendo em conta o local onde o utilizador se encontra e os pontos turísticos localizados nessa zona (Ouanain *et al.*, 2010).
- **Gestão de energia** – Nos dias de hoje, gerir a energia é de extrema importância, é, inclusive, um desafio que algumas empresas enfrentam, gerir de forma eficiente e otimizada a energia. Assim, foi desenvolvido um sistema que utiliza o perfil do utilizador e micro contabilidade para gerir a energia de forma inteligente (Caruso *et al.*, 2013).
- **Recomendação de emprego** – A tarefa de encontrar um emprego é algo que o utilizador tem de fazer na sua vida, nem que seja somente uma vez. Assim, foi desenvolvido um sistema denominado de CASPER (*Case-Based Profiling for Electronic Recruitment*, em português, *Profiling* Baseado em Casos de Recrutamento Eletrónico) capaz de recomendar ao utilizador, de forma automática, empregos de acordo com as informações que constam no seu perfil, as suas qualificações e experiência profissional (Bradley *et al.*, 2000).

2.7 Conteúdo dos Perfis

Cada utilizador apresenta informação distinta dos demais. Porém, existem informações comuns entre eles, como por exemplo, os seus interesses, o seu conhecimento, histórico e habilidades, as suas metas, o seu comportamento, as suas preferências, as suas caraterísticas individuais, a sua informação contextual e, ainda, os perfis de grupo. Estes diferentes tipos de informação devem constar no perfil de utilizador devido aos benefícios que têm. Assim, o perfil de utilizador pode ser baseado em informação heterogénea associada a um utilizador individual ou a um grupo de utilizadores expondo os seus interesses semelhantes (Gauch *et al.*, 2007). O perfil do utilizador é representado por informações

individuais essenciais para que se possa desenvolver uma aplicação inteligente (Schiaffino & Amandi, 2009).

2.7.1 Interesses do Utilizador

Os interesses do utilizador correspondem ao tipo de informação mais importante a ser coletada para construir o perfil do utilizador, e podem, inclusive, ser considerados os únicos em sistemas de recuperação e de filtragem de informações, sistemas de recomendação, agentes de interface e sistemas adaptativos orientados por informações, como por exemplo, guias de museus e notícias (Brusilovsky & Millán, 2007). Para conseguir obter informações sobre os interesses dos utilizadores, existem diferentes fontes, nomeadamente: pedir aos utilizadores obtendo a informação explicitamente, inferindo-os das ações e feedback dos utilizadores ou através da interação com outros agentes (Maes, 1994).

Através de hierarquias de tópicos é possível realizar uma representação mais poderosa dos interesses dos utilizadores (Godoy *et al.*, 2004). Assim, cada nó da hierarquia representa um interesse do utilizador, que por sua vez é caraterizado por um conjunto de palavras representativas. Geralmente, utiliza-se esta hierarquia para representar não só os interesses gerais de um utilizador, mas também os subtópicos dos seus interesses, que podem ser importantes para um dado utilizador. Por vezes, é possível classificar os interesses dos utilizadores em interesses de curto ou longo prazo. Por exemplo, um utilizador com interesse em futebol pode ser considerado de curto prazo, no caso de apenas ler ou ouvir notícias durante o Mundial, ou de longo prazo, caso mantenha o interesse ao longo do ano (Schiaffino & Amandi, 2009).

2.7.2 Conhecimento, Histórico e Habilidades

Nas diferentes áreas, o conhecimento que o utilizador possui sobre o domínio de uma aplicação, o seu histórico e as suas habilidades, são caraterísticas importantes para construir o seu perfil. A primeira das três caraterísticas referidas, o conhecimento, pode ser representado de formas distintas, como por exemplo, através de um modelo que monitoriza o conhecimento que um aluno tem sobre um dado assunto, ou representando o seu conhecimento por meio de erros ou equívocos. Uma outra forma, é modelando o que o utilizador não tem conhecimento. Existem sistemas que categorizam os utilizadores como iniciantes, intermédios e especialistas, tendo em conta o conhecimento que apresentam sobre um dado domínio (Schiaffino & Amandi, 2009).

O histórico de um utilizador refere-se a caraterísticas do utilizador que não estão relacionadas diretamente com o domínio da aplicação (Schiaffino & Amandi, 2009). Usualmente, o histórico do utilizador é obtido explicitamente, quer seja pelo próprio utilizador ou por alguém superior, um professor de faculdade ou um administrador de uma instituição, visto que o histórico é uma caraterística que não muda durante o trabalho e é impossível inferir sobre o mesmo observando o trabalho desempenhado pelo utilizador (Brusilovsky & Millán, 2007).

As habilidades dos utilizadores são fundamentais em áreas como a gestão do conhecimento. Os sistemas de gestão de habilidades auxiliam como plataformas técnicas nos mercados corporativos internos, sendo construídos sobre bases de dados com perfis de funcionários e candidatos, onde constam diferentes valores paras as diversas habilidades do utilizador, estes valores podem ser representados através de vetores (Schiaffino & Amandi, 2009).

2.7.3 Metas

Apesar de não constituir uma informação trivial na construção do perfil do utilizador, é importante saber quais as metas do utilizador que devem ser consideradas na construção do seu perfil, visto que estas representam o objetivo que o utilizador tem. As metas podem ser consideradas como tarefas que estão no foco de atenção do utilizador (Horvitz *et al*, 1998). O utilizador pode querer alcançar diferentes metas, como por exemplo, se está a realizar uma pesquisa na web, o seu objetivo é de conseguir informações significantes, se está num sistema de *e-learning*, o objetivo do utilizador é adquirir conhecimento sobre um dado domínio do seu interesse (Schiaffino & Amandi, 2009).

2.7.4 Comportamento

O comportamento do utilizador é um fator relevante na construção do perfil. No entanto, é necessário que o comportamento do utilizador verifique um determinado padrão, que seja repetitivo, de forma a poder ser aplicado a um sistema adaptativo, a um site por um agente inteligente ou para auxiliar o utilizador conforme o comportamento aprendido. Usualmente, o comportamento dos utilizadores é rotineiro, ou seja, as suas ações têm uma certa regularidade ou repetem-se de forma sazonal (Schiaffino & Amandi, 2009). A informação acerca do comportamento do utilizador, deve ser recolhida de forma

implícita, através da observação das atitudes que tomam perante as distintas situações, por parte de agentes.

2.7.5 Preferências

As preferências e os hábitos têm também um grande contributo na construção do perfil do utilizador. As informações coletadas sobre as suas preferências são armazenadas no perfil de interação do utilizador e são acionadas quando o utilizador precisa de uma sugestão para lidar com um problema ou de um aviso sobre um dado problema. Normalmente, uma preferência do utilizador manifesta a ação que o agente deve tomar, preferencialmente, nas diferentes situações. Por exemplo, no caso de o utilizador receber um aviso sobre um problema com uma dada reunião, alguns utilizadores podem preferir um aviso simples, enquanto outros preferem que nesse aviso seja sugerida uma data alternativa para a reunião que conjugue as preferências e prioridades do participante. Assim, o agente deve ser capaz de conhecer quando o utilizador prefere cada ação (Schiaffino & Amandi, 2006).

2.7.6 Caraterísticas Individuais

As informações pessoais do utilizador também devem ser consideradas na construção do perfil. Devem estar incluídas informações demográficas, tais como género, idade, cidade, país, número de filhos, estado civil, entre outras (Schiaffino & Amandi, 2009). Uma caraterística a ter em conta é a personalidade do utilizador. Um dos modelos de personalidade mais conhecido é o OCEAN que compreende cinco dimensões da personalidade: abertura à experiência, consciência, extroversão, amabilidade e neuroticismo (Goldberg, 1993). Para identificar a personalidade do utilizador são usados alguns métodos, cujo resultado é guardado no perfil do utilizador.

2.7.7 Informação Contextual

Mais recente, surgiu um novo recurso no perfil do utilizador, o seu contexto. Dependendo do domínio da aplicação, a definição de contexto pode variar. No entanto, o contexto pode ser explicado como qualquer informação que pode ser usada para caraterizar a situação de uma entidade, que pode ser uma pessoa, um local ou um objeto computacional (Dey & Abwod, 2000). O contexto do utilizador está dividido em três: contexto pessoal, social e espaço temporal. O primeiro divide-se em dois contextos: o fisiológico que

contém informações relativas à saúde do utilizador, tais como pressão arterial, peso, nível de glicose, entre outras, e o mental que tem informações como o humor, irritabilidade e stress. O contexto social relata os aspetos sociais atuais do utilizador. Por último, tem-se o contexto espaciotemporal do utilizador que refere os aspetos do utilizador relacionados com o tempo e à extensão espacial do seu contexto (Schiaffino & Amandi, 2009).

2.7.8 Perfis de Grupo

Contrariamente aos perfis de utilizadores individuais, os perfis de grupo têm como objetivo combinar diversos perfis individuais, formando assim um grupo. Estes perfis de grupo são de extrema importância quando é preciso fazer recomendações para um grupo de utilizadores, e não apenas a um só utilizador. Geralmente, o feedback do utilizador do grupo é usado para criar recomendações para um utilizador individual ou para um grupo de utilizadores (Schiaffino & Amandi, 2009). Como exemplo, os autores do artigo *"TV Program Recommendation for Multiple Viewers Based on user Profile Merging"* propõem um sistema de recomendação que junte os perfis dos utilizadores individuais com o objetivo de se criar um perfil do utilizador comum e que gere recomendações de acordo com o perfil criado (Yu *et al.*, 2006).

2.8 Obtenção de Perfis

Quando se pretende criar um perfil de utilizador, as informações necessárias podem ser obtidas de duas formas: explicitamente, ou seja, é o próprio utilizador que fornece as informações diretamente, ou implicitamente, através da observação das ações do utilizador (Schiaffino & Amandi, 2009). Estas duas formas de coletar informação podem ser combinadas, obtendo melhores resultados.

Inicialmente, os sistemas apostavam na obtenção dos dados necessários diretamente do utilizador, isto é, perguntavam explicitamente o que era necessário coletar. Contudo, esta abordagem não era eficaz, visto que nem sempre o utilizador mostrava interesse em fornecer as suas informações. Assim, hoje em dia, procura-se criar o perfil do utilizador obtendo os dados implicitamente, optando por observar algumas atitudes do utilizador (Kanoje *et al*, 2014). Também é possível apostar numa abordagem híbrida, por exemplo, através de um componente automatizado que cria o perfil de um utilizador, conjugando-se as observações do utilizador e um mecanismo de feedback de relevância explícita capaz de ajustar os perfis aos seus interesses individuais (Papazoglou, 2001).

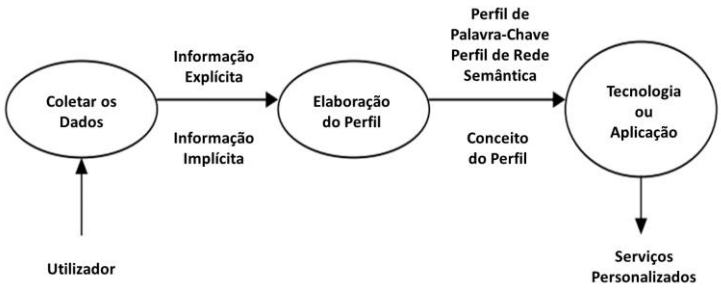


Figura 2.8- Visão geral da construção de um perfil baseado em personalização (figura adaptada de Gauch *et al.*, 2007, p. 56)

Podem identificar-se três diferentes abordagens para a obtenção de perfis (Figura 2.8): explícita, implícita e híbrida. Como Schiaffino e Amandi (Schiaffino & Amandi, 2009) referem, a forma mais simples de conseguir informações relativas ao utilizador, é através do preenchimento de formulários ou outros meios criados para esse propósito. No entanto, nem sempre os utilizadores estão dispostos a preencher esses formulários onde fornecem as suas informações pessoais, tais como idade, género, profissão, data de nascimento, estado civil e passatempos. Além deste tipo de informação, os utilizadores podem opinar sobre um dado assunto atribuindo um valor de um intervalo.

O perfil que é criado a partir de informações obtidas explicitamente, é também conhecido como perfil estático. Este tipo de perfil é conhecido pelo processo que analisa as caraterísticas estáticas e previsíveis do utilizador. É através deste perfil que se obtém informação relativa aos interesses do utilizador. No entanto, o perfil estático apresenta algumas desvantagens, nomeadamente o facto deste tipo de perfil ser válido apenas por um determinado período de tempo, devido a possível mudança de interesse do utilizador, que consequentemente faz com que o perfil se torne mais impreciso ao longo do tempo. Outra desvantagem é o facto de o utilizador puder não precisar de forma objetiva o seu interesse, o que pode influenciar a forma como se infere sobre os interesses de outros utilizadores que sejam semelhantes (Poo *et al.*, 2003) (Kanoje *et al.*, 2014).

Além das desvantagens referidas, ao recolher as informações explicitamente o utilizador fica sobrecarregado, e por esse motivo, juntamente com a preocupação de fornecer as suas informações devido a questões de privacidade, o utilizador pode optar por não participar. De realçar, que quando os utilizadores não consentem em ceder as suas informações pessoais, não é possível criar o seu perfil. No

entanto, existem sempre utilizadores dispostos a fornecer as suas informações pessoais e a partilhar os seus interesses (Gauch *et al.*, 2007).

O perfil que resulta da coleta de informações de forma implícita, é também conhecido por perfil dinâmico, e consiste em analisar as atividades ou ações do utilizador para conhecer os seus interesses. Apesar das atividades e interesses dos utilizadores serem registados em tempo real, existem interesses que não podem ser rastreados (Poo *et al.*, 2003) (Kanoje *et al.*, 2014). A principal vantagem de coletar as informações implicitamente, é que não é necessário que o utilizador intervenha durante o processo de construção do perfil (Gauch *et al.*, 2007).

Como referido anteriormente, quando as informações são coletadas explicitamente podem ocorrer alguns problemas, particularmente, o facto de os utilizadores não se mostrarem disponíveis para fornecer as suas informações, por vezes, não dizem a verdade sobre eles mesmos quando preenchem formulários, ou ainda, não saberem expressar o que pretendem nem os seus interesses. Por estes motivos, usualmente, o método de recolha de informação aplicado é observar as ações dos utilizadores, anotando essas mesmas ações e descobrindo padrões através de técnicas de *machine learning* ou *data mining*. Assim, estas ações devem ser repetitivas, tal como foi explicado na subcapítulo anterior (Schiaffino & Amandi, 2009).

Kelly e Teevan (Kelly & Teevan, 2003), explicaram as vantagens e desvantagens de algumas técnicas utilizadas para coletar informação de forma implícita. Uma das técnicas é a extração de informação dos históricos de navegação, sendo que podem ser coletadas de duas formas: através da partilha periódica do histórico pelo utilizador ou instalando um servidor proxy que guarda todo o histórico gerado pelo utilizador (Barrettet *et al.*, 1997) (Trajkova & Gauch, 2004). No entanto esta técnica acarreta algumas desvantagens, nomeadamente, é possível apurar o número de visitas a um determinado site durante alguns períodos de tempo, no entanto, o tempo gasto em cada página web, bem como o tempo entre cliques consecutivos pode ser inferido com erro. Outra desvantagem é o facto de apenas coletar o histórico do utilizador apenas num computador, no entanto, é possível que o utilizador forneça os diferentes históricos deixados nos equipamentos que usou. A grande vantagem desta técnica é que evita a sobrecarga do utilizador, apesar de se obterem menos informações do que de forma explícita visto que apenas são rastreadas as atividades na página web, estas coletam-se por meio de cookies, logins ou ids de sessão (Gauch *et al.*, 2007).

A abordagem híbrida combina as vantagens das duas abordagens anteriores: explícita e implícita, isto é, considera caraterísticas estáticas, mas também caraterísticas dinâmicas do utilizador, criando perfis mais eficientes e precisos, uma vez que a informação é atualizada temporalmente (Kanoje *et al.*, 2014). Na abordagem híbrida, o utilizador pode fornecer informação explicitamente, através da avaliação de um agente preenchendo um formulário elaborado com esse propósito, ou implicitamente, quando o agente observa as ações do utilizador após o auxiliar a detetar alguma avaliação implícita da sua assistência. Quando o utilizador fornece informação de forma explícita, esta pode ser efetuada de maneira simples, através da avaliação à assistência do agente que pode ser feita numa escala quantitativa ou qualitativa, ou complexa, quando o utilizador cede grandes quantidades de informação em diversas etapas (Schiaffino & Amandi, 2014).

Um estudo desenvolvido por Waern (Waern, 2004), compara a eficácia de perfis de utilizadores que foram parcialmente ou totalmente construídos por meios automáticos. Foram ainda construídos dois perfis, considerados mais leves, o primeiro a partir do histórico de pesquisas e o segundo através de uma lista dos domínios que foram visitados durante a navegação. Conclui-se que quanto mais rica for a informação disponível, melhor será o perfil construído, e que os perfis construídos através da informação implícita são superiores aos perfis criados com informação explícita. De notar, que um estudo anterior tinha concluído o contrário, ou seja, que os perfis gerados com informação explícita eram melhores, permitindo assim concluir que com a experiência adquirida ao longo do tempo, a forma de coletar informação implicitamente melhorou, e consequente, os perfis construídos por este método melhoraram também (Gauch e*t al.*, 2007).

Assim, pode concluir-se que, o objetivo do perfil do utilizador é coletar informações sobre os assuntos nos quais está interessado e quantificar durante quanto tempo apresentou o mesmo interesse, com o propósito de melhorar a qualidade da informação disponível (Gauch *et al.*, 2007).

CAPÍTULO 3

3. UM CASO DE ESTUDO

3.1 Apresentação do Caso

No presente caso de estudo utilizámos uma base de dados relativa a um setor de retalho especializado, que foi cedida por uma empresa. Tal como já foi explicado anteriormente, neste trabalho de dissertação pretendia-se fazer a aplicação de algoritmos de classificação, segmentação e associação, a um conjunto de dados previamente selecionado, de forma a poder estabelecer os perfis de agentes comerciais e optometrista. O conjunto de dados construído teve em conta os objetivos de negócio estabelecidos para a criação dos referidos perfis, isto é, foram considerados todos os atributos que tinham relevância para a análise em questão, detetados e resolvidos os problemas existentes no conjunto de dados, de forma a que estes estivessem preparados para se proceder à aplicação dos algoritmos.

Para a realização deste trabalho de análise utilizámos as seguintes ferramentas: o *SQL Server Management Studio*, o *software* estatístico R e o *Microsoft Visual Studio*. A primeira destas ferramentas foi utilizada para explorar a fonte de dados e para criar o conjunto de dados para a análise, a segunda para calcular as estatísticas básicas e a terceira, em conjunção com a primeira, para suportar a aplicação dos algoritmos de mineração de dados.

3.2 Análise do Negócio

A análise de negócio inclui várias tarefas importantes, tais como: determinar os objetivos de negócio, avaliar a situação, determinar as metas da mineração de dados e elaborar um plano do projeto (Shearer, 2000). Neste processo analisam-se dois perfis: o agente comercial e o optometrista. O objetivo foi determinar quais os modelos que apresentam melhor desempenho para classificar, segmentar e associar

os dados do conjunto de dados. A fase da análise do negócio é fundamental para compreender o desempenho e a forma como desenrola o negócio do setor ótico.

3.2.1 O Agente Comercial

Inicialmente, o negócio das óticas em Portugal era tipicamente familiar e conservador. No entanto, na década de 80, com a entrada de Portugal na União Europeia, ocorreram várias mudanças, refira-se, por exemplo, a entrada de grupos e marcas estrangeiras, e consequentemente, a introdução de *franchising* no país. Como resultado destas alterações, os grupos óticos que atuavam no mercado de forma independente agruparam-se, de modo a conquistarem mais valor perante o consumidor final. Mas, independentemente do modelo de negócio adotado (ótica independente, grupo ou *franchising*), a diferença mais evidente entre as lojas óticas era o preço praticado, uma vez que as marcas oferecidas ao cliente eram praticamente iguais em todos estes estabelecimentos, devido ao facto de, geralmente, se cingirem as marcas que dominavam o mercado ótico mundialmente. Neste contexto, um dos aspetos que pode ser diferenciador é o atendimento e o acompanhamento prestado, bem como a personalização do serviço, fatores que podem levar a uma consequente fidelização do cliente. Assim, as técnicas de mineração de dados aliadas à construção do perfil do agente comercial e do optometrista, tornam-se essenciais para avaliar o desempenho dos mesmos.

Segundo um estudo publicado, em 2012, pela consultora D&B que visava o setor ótico na Península Ibérica, concluiu que *"a nível ibérico, Catalunha, Lisboa e Madrid são as zonas onde se localiza o maior número de operadores, reunindo, respetivamente, 29%, 25% e 15% do total"*, evidenciando-se *também "uma diferença significativa relativamente à dimensão média das empresas em Espanha e Portugal. Assim, enquanto no mercado espanhol o número médio de trabalhadores por empresa chega aos 43, no português situa-se nos 29"*, e ainda que *"os cinco principais grupos que operam neste sector são responsáveis por cerca de 45% das vendas totais na Península Ibérica"* (Marques, 2013). Um outro estudo, realizado pela empresa Informa DBK S.A., em junho de 2015, verifica um abrandamento nas vendas de lentes de contacto, enquanto que as lentes oftálmicas continuam a ser o produto mais utilizado a nível europeu. Esta diferença é explicada pela tendência crescente de óculos como complemento de moda e também pelo aumento da cirurgia refrativa.

O setor ótico carateriza-se por ser especializado em fornecer produtos e serviços óticos ao consumidor final, nomeadamente ao nível da venda de lentes oftálmicas e de contacto, armações, óculos de sol, e

produtos oculares. A função do agente comercial é vender estes mesmos produtos, rececionando o receituário de refração que foi prescrita pelo oftalmologista ou optometrista ao cliente, indicando a diferente gama de produtos que tem para oferecer. No caso de a venda a realizar for de lentes oftálmicas, deve expor os diferentes tipos, fatores como a espessura das lentes e os respetivos preços, e no caso de o cliente desejar comprar a armação, indicar também qual a que mais se adequa, sugestão essa que deve ter em conta a idade e os gostos pessoais do cliente. Por último, é também da sua competência proceder a montagem das lentes oftálmicas na respetiva armação, efetuando os ajustes necessários para que o cliente se sinta confortável e satisfeito com a sua escolha. No entanto, com o aumento de informação disponível na internet e em outros meios de comunicação, torna-se imprescindível apostar na contínua melhoria da qualidade dos agentes comerciais, visto que o cliente tem muito mais facilidade de fazer comparação de preços entre as diferentes lojas, bem como o facto do cliente ser muito mais informado.

3.2.2 O Optometrista

Segundo o *World Council of Optometry* e o *European Council of Optometry and Optics*, o optometrista *é "o especialista dos cuidados de saúde primários à visão, que pratica Optometria e que fornece cuidados extensivos em visão e sistema visual, que inclui refração e prescrição, deteção/diagnóstico e acompanhamento/tratamento de doenças oculares e a reabilitação/tratamento de condições do sistema visual"* (APLO, 2006). A profissão do optometrista é essencial para a prevenção da saúde visual, prestando os cuidados primários de saúde visual e dedicando-se a cuidar da visão, tendo funções como a refração e prescrição, diagnóstico e tratamento de doenças oculares. É o optometrista quem tem a responsabilidade de detetar, analisar e tratar problemas visuais que afetam os seus pacientes, sejam eles de natureza refrativa, funcional, binocular, entre outras, através da prescrição de lentes, ou outros produtos, adequados ao problema diagnosticado.

No presente caso de estudo, assumiu-se que o optometrista tem como função detetar quais as necessidades oftálmicas do paciente. O optometrista presta consultas de um determinado tipo de serviço e, no fim, prescreve um receituário com o artigo que corresponde ao que foi detetado durante o serviço. Podem ocorrer situações em que a consulta era de um determinado serviço, mas o problema detetado não se encaixar nesse serviço, receitando o artigo que corresponde ao que é detetado. Durante o diagnóstico feito ao paciente, é possível que o optometrista considere pertinente prescrever um produto

ocular para combater algum problema que o paciente apresente. É comum atribuir-se um optometrista a uma ou mais lojas de óticas, permitindo assim perceber as necessidades dos pacientes que frequentam as diferentes lojas.

Com o propósito de se atingirem os objetivos de um processo de mineração de dados, foram estudados vários dos algoritmos disponíveis no *Microsoft Visual Studio*, em particular, o algoritmo *Decision Tree*, o algoritmo de *Clustering* e o algoritmo de *Association*. Quando se inicia a aplicação de qualquer um dos referidos algoritmos, é essencial escolher um atributo chave, um atributo previsível e atributos de entrada, sendo necessário ter em atenção também os parâmetros de cada algoritmo, de forma a que estes estejam ajustados para obter os melhores resultados. Para que se possa concluir, se os resultados obtidos são bons para a análise, definiram-se critérios de sucesso dependendo da técnica aplicada, tais como o modelo deve apresentar uma taxa de erro inferior a 50%, no caso da classificação, o número de casos em cada *cluster* não deve diferir de forma exagerada e que o número de *clusters* deve ser igual ou superior a três, no caso da segmentação, e por último que o valor de confiança de uma regra de associação deve ser inferior a um, no caso de associação.

3.3 Análise dos Dados

Nesta fase do processo, pretende-se que o analista se familiarize com os dados, que sejam identificados problemas relativos à qualidade dos dados e que se detetem subconjuntos interessantes para formular hipóteses sobre informações desconhecidas (Shearer, 2000). A base de dados mencionada anteriormente, e que suportou este processo de análise, pertence a um grupo ótico composto por três diferentes lojas - esta base de dados foi cedida pela empresa que presta serviços informáticos a ótica, com o consentimento do grupo ótico referido. Por questões de privacidade, os nomes do agente, do optometrista e do cliente, nunca serão mencionados ao longo desta dissertação, mas também devido ao novo Regulamento Geral de Proteção de Dados (RGPD), que entrou em vigor no dia 25 de maio de 2018.

3.3.1 O Agente Comercial

Após uma análise cuidadosa ao caso em estudo, analisaram-se todos os atributos que se consideram importantes para dar uma resposta positiva ao caso em estudo. Nesta fase, foram consultadas todas as tabelas da base de dados, com o objetivo de compreender as relações existentes entre elas, bem como para perceber onde estavam os atributos necessários para a análise. Assim, tendo como premissa, reunir

o máximo de atributos que enriquecessem o estudo que foi feito, elaborou-se um esquema relacional, Figura 3.1, onde é possível observar todas as tabelas consultadas que tinham relevância para o caso em estudo, bem como os relacionamentos que existem entres estas tabelas.

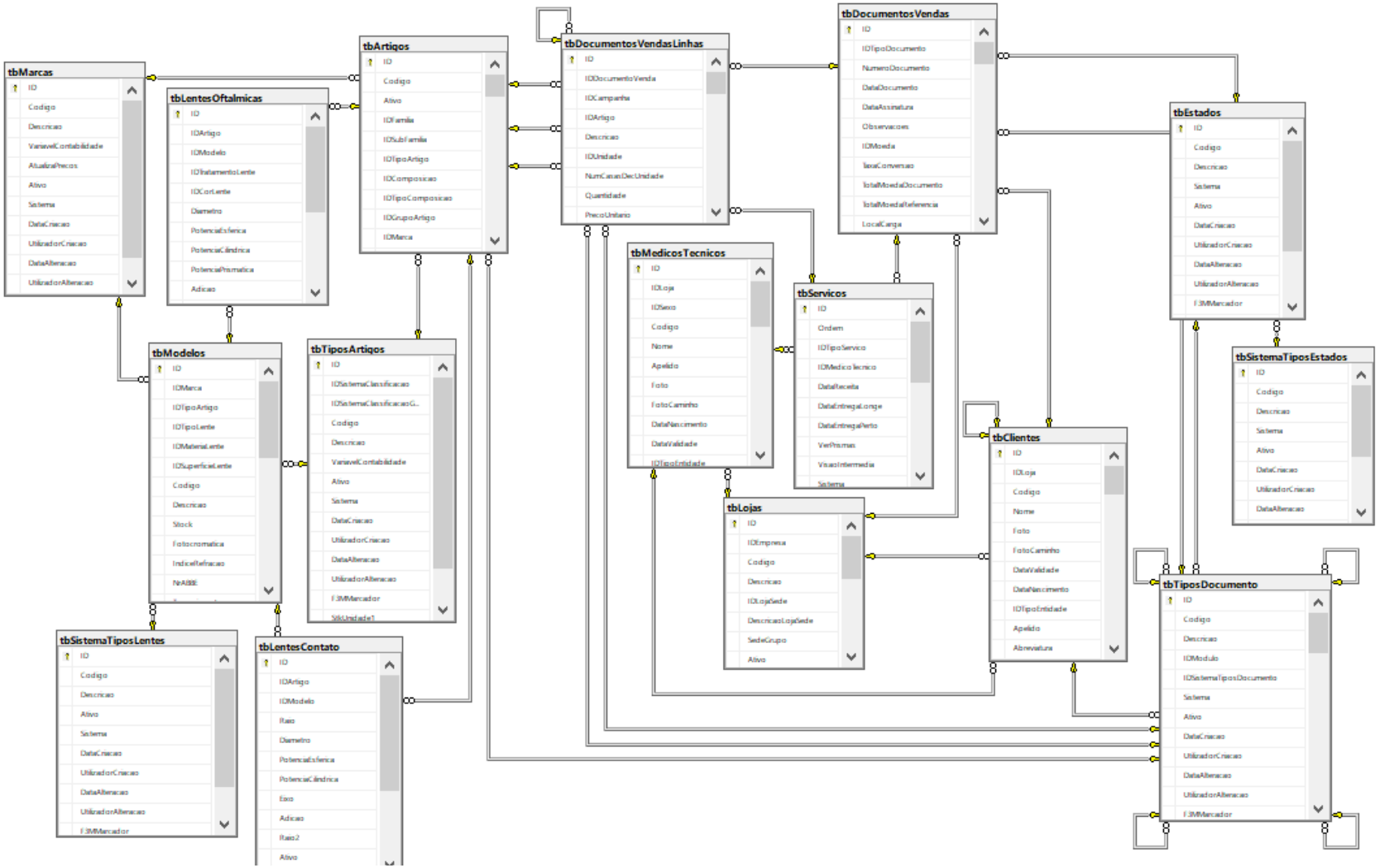


Figura 3.1 - Esquema relacional utilizado na construção da vista para o perfil do vendedor.

Depois de consultadas as tabelas presentes no esquema utilizado (Figura 3.1), procedeu-se a uma análise dos atributos, considerando-se apenas aqueles que eram relevantes para a análise pretendida. Na Tabela 3-1 estão apresentados todos os atributos selecionados, bem como a tabela a que pertencem.

Tabela 3-1 - Descrição dos atributos que serão utilizados na elaboração da vista do perfil de vendedor.

Campo	Descrição
tbMedicosTecnicos.IDTipoEntidade	Identificador que identifica cada entidade presente no conjunto de dados. Neste caso, sempre que IDTipoEntidade for igual a 14, refere-se à entidade agente comercial.
tbMedicosTecnicos.Codigo	Código que identifica cada agente comercial.
tbMedicosTecnicos.Nome	Nome do agente comercial.
tbDocumentosVendas.ID	Identificador único de cada documento de venda emitido.
tbTiposDocumento.Codigo	Código que indica de que tipo é o documento emitido.
tbTiposDocumento.Descricao	Descrição do tipo de documento.
tbDocumentosVendas.DataDocumento	Identifica a data e a hora em que o documento é emitido.
tbClientes.Codigo	Código que identifica cada cliente.
tbDocumentosVendas.NomeFiscal	Nome do cliente.
tbDocumentosVendas.Idade	Indica a idade do cliente, no momento da compra.
tbArtigos.Codigo	Código do artigo.
tbArtigos.Descricao	Descrição do artigo que está presente no documento.
tbTiposArtigos.Codigo	Código do tipo de artigo.
tbTiposArtigos.Descricao	Descrição do tipo de artigo.
tbModelos.IDTipoLente	Identifica unicamente cada tipo de lente.
tbSistemaTiposLentes.Descricao	Descreve o tipo de lente (Unifocal, Bifocal, Progressiva, Ocupacional, Descartáveis, Hidrófilas, Rígidas, Semirrígidas, Híbridas).
tbMarcas.Descricao	Descrição da marca a que os artigos do tipo LO ou LC pertencem.
tbDocumentosVendasLinhas.PrecoUnitarioSemIVA	Preço unitário de cada artigo sem a aplicação da respetiva taxa de IVA.
tbDocumentosVendasLinhas.ValorDescontoEfetivoSemIVA	Valor do desconto aplicado sobre o preço unitário do artigo.
tbSistemaTiposEstados.Codigo	Código do tipo de estado em que está o documento (EFT, ANL, RSC).
tbSistemaTiposEstados.Descricao	Descrição do tipo de estado (Efetivo, Anulado, Rascunho).

Numa análise preliminar dos dados, retiraram-se três observações que se podem revelar importantes na fase de preparação de dados. Em particular, assume-se que, os atributos referentes à idade do cliente e à descrição do tipo de lente, são importantes para a análise e que têm grande impacto para a mesma. No entanto, estes dados contêm um número considerável de registos nulos. Se o objetivo é criar o perfil do agente comercial, então considerar o atributo que o identifica é inevitável. Contudo, em alguns registos este campo aparece com valor nulo, o que significa que, por algum motivo, o mesmo não foi preenchido.

De todos os agentes comerciais identificados, alguns apresentam um baixo número de vendas. Existem casos com apenas uma venda, o que fez com que fosse necessário decidir se estes deviam ou não ser considerados no processo análise.

Depois de se terem escolhido os atributos a utilizar no estudo, realizaram-se várias estatísticas exploratórias, tais como mínimo, máximo, mediana, média, primeiro e terceiro quartil, desvio padrão, variância e distância interquartil, com o objetivo de nos familiarizar com os dados e fazer a sua descrição. Estas estatísticas foram realizadas com o *software* estatístico R (Landeiro, 2011). Os dados utilizados no processo de elaboração dessas estatísticas foram extraídos de uma base de dados *SQL Server* para um ficheiro *Microsoft Excel*. Depois realizou-se o carregamento deste ficheiro para o *software* estatístico R (Figura 3.2).

```
> library(readxl)
> PVendedor <- read_excel("C:/Users/ritaf/Desktop/PVendedor.xlsx")
> View(PVendedor)
> attach(PVendedor)
```

Figura 3.2 - Execução do carregamento do ficheiro Excel no software estatístico R.

Dos vinte e um atributos em análise, apenas em cinco se conseguem obter as estatísticas que eram pretendidas. Este facto acontece porque os restantes atributos são constituídos por carateres e não por números, pelo que não é possível obter os valores que eram desejados, tal como se pode constatar na análise (Figura 3.3).

```
> summary.matrix(PVendedor)
 Tecnico.Tecnico CodigoTecnico.CodigoTecnico NomeTecnico.NomeTecnico
 Min.    :0      Length:319                  Length:319
 1st Qu.:0      Class :character            Class :character
 Median :0      Mode  :character            Mode  :character
 Mean   :0
 3rd Qu.:0
 Max.   :0
 DocumentoVenda.DocumentoVenda CodigoTipoDocumento.CodigoTipoDocumento
 Min.    :  11.0000           Length:319
 1st Qu.: 431.5000            Class :character
 Median : 873.0000            Mode  :character
 Mean   : 929.0063
 3rd Qu.:1480.0000
 Max.    :1854.0000
 DescricaoTipoDocumento.DescricaoTipoDocumento DataVenda.DataVenda
 Length:319                                    Length:319
 Class :character                              Class :character
 Mode  :character                              Mode  :character


 CodigoCliente.CodigoCliente NomeCliente.NomeCliente IdadeCliente.IdadeCliente
 Min.    :   2.000           Length:319              Min.    : 6.00000
 1st Qu.:2906.000            Class :character        1st Qu.:33.50000
 Median :5423.000            Mode  :character        Median :50.00000
 Mean   :5239.947                                    Mean   :49.88715
 3rd Qu.:8058.000                                    3rd Qu.:69.50000
 Max.    :8220.000                                   Max.   :91.00000
 CodigoArtigo.CodigoArtigo DescricaoArtigo.DescricaoArtigo
 Length:319                Length:319
 Class :character          Class :character
 Mode  :character          Mode  :character


 CodigoTipoArtigo.CodigoTipoArtigo DescricaoTipoArtigo.DescricaoTipoArtigo
 Length:319                        Length:319
 Class :character                  Class :character
 Mode  :character                  Mode  :character


 IDTipoLente.IDTipoLente DescricaoTipoLente.DescricaoTipoLente
 Min.    : 1.000000      Length:319
 1st Qu.: 1.000000       Class :character
 Median : 1.000000       Mode  :character
 Mean   : 1.680251
 3rd Qu.: 3.000000
 Max.    :11.000000
 DescricaoMarca.DescricaoMarca PrecoUnitarioSemIVA.PrecoUnitarioSemIVA
 Length:319                    Min.    :  0.00000
 Class :character              1st Qu.: 18.37735
 Mode  :character              Median : 38.58490
                               Mean   : 84.20154
                               3rd Qu.:149.25470
                               Max.    :296.71700
 ValorDescontoSemIVA.ValorDescontoSemIVA CodigoTipoEstado.CodigoTipoEstado
 Min.    :  0.00000                      Length:319
 1st Qu.:  7.08000                       Class :character
 Median : 12.56000                       Mode  :character
 Mean   : 23.33759
 3rd Qu.: 37.74000
 Max.    :107.30000
 DescricaoTipoEstado.DescricaoTipoEstado
 Length:319
 Class :character
 Mode  :character
```

Figura 3.3 - Análise exploratória efetuada sobre os dados com recurso ao software R.

Relativamente aos atributos código do documento de venda e código do cliente, as estatísticas revelaram a sua irrelevância para o processo de análise, visto que estes atributos servem apenas para identificar o

documento e o cliente da venda efetuada. No que diz respeito à idade do cliente, verificou-se que esta se situa entre os seis e os noventa e um anos, sendo a média da idade dos clientes de 49,89 e a mediana de 50,00 anos. O preço unitário dos artigos vendidos varia entre 0,00 e 296,72 euros, situando-se o preço médio em 84,20 euros e a mediana é de 38,58 euros. O valor de desconto aplicado a esses mesmos artigos tem um intervalo de preços muito abaixo, pois varia entre 0,00 e 107,30 euros, e a sua média é 12,56 euros e a mediana deste atributo é 12,56 euros. De referir que os valores do preço unitário e o valor de desconto não tinham aplicada a respetiva taxa de IVA.

Tendo em conta o que foi referido, apenas se calculou o desvio padrão, a variância e a distância interquartil para os seguintes atributos: idade do cliente, preço unitário e valor de desconto. Os cálculos desses valores podem ser consultados na Figura 3.4.

```
> desviopadrao<-c(sd(IdadeCliente), sd(PrecoUnitarioSemIVA), sd(ValorDescontoSemIVA))
> desviopadrao
[1] 21.80257 84.56208 24.48242
> variancia<- c(var(IdadeCliente), var(PrecoUnitarioSemIVA), var(ValorDescontoSemIVA))
> variancia
[1]  475.3520 7150.7455  599.3889
> aiq<-c(IQR(IdadeCliente), IQR(PrecoUnitarioSemIVA), IQR(ValorDescontoSemIVA))
> aiq
[1]  36.0000 130.8774  30.6600
```

Figura 3.4 - Cálculo do desvio padrão, variância e distância interquartil no *software* estatístico R.

3.3.2 O Optometrista

O processo de análise do optometrista difere do anterior. Tendo em vista a análise do caso do optometrista, elaborou-se um novo esquema relacional (Figura 3.5), que contém as tabelas consultadas para este novo caso, bem com os relacionamentos estabelecidos entre elas.

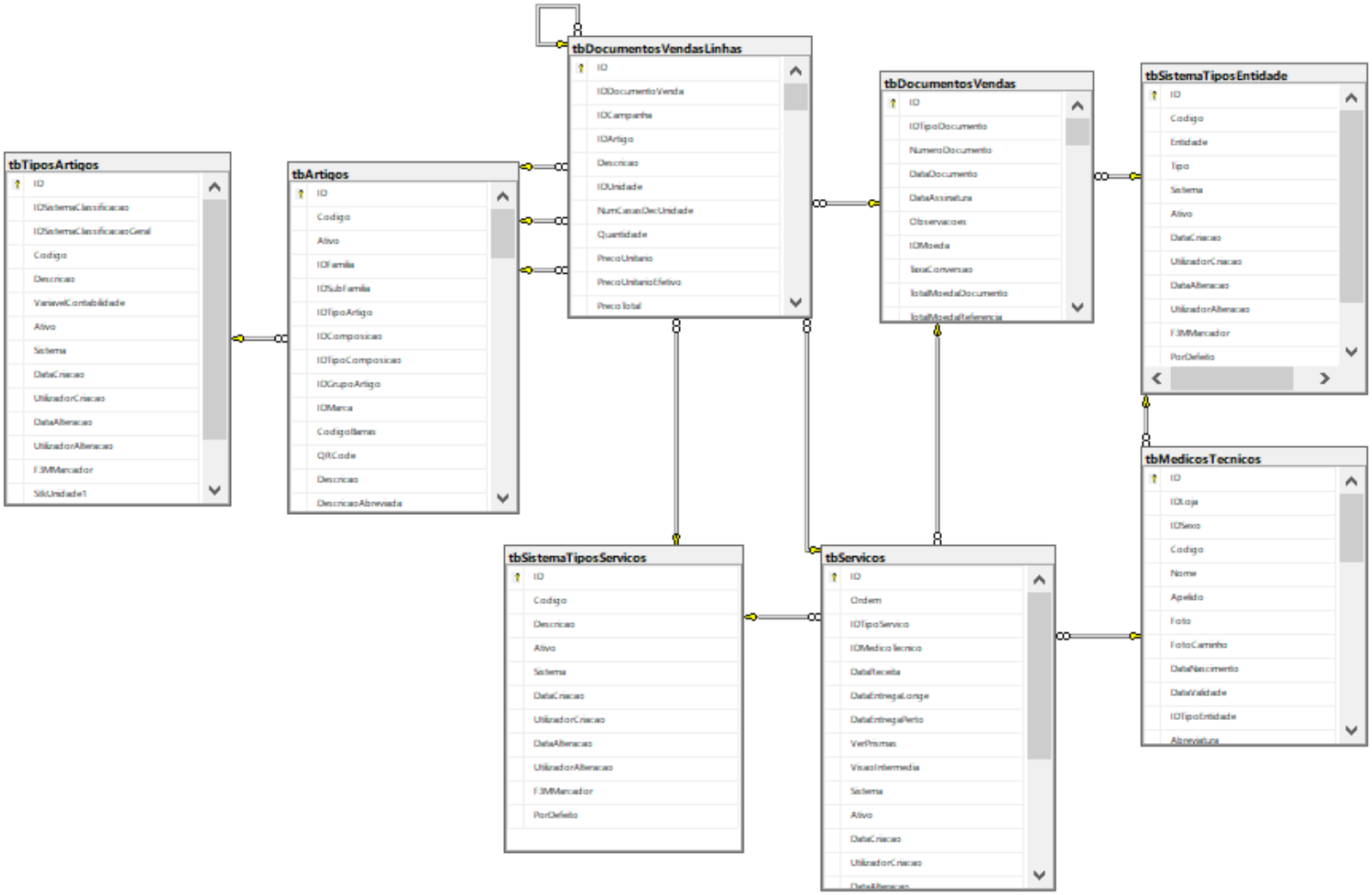


Figura 3.5 - Esquema relacional utilizado na construção da vista para a definição do perfil do optometrista.

Após uma análise sobre os dados disponíveis, verificou-se que em alguns registos o campo correspondente à identificação do optometrista não estava preenchido. Tendo como objetivo traçar o perfil do optometrista é necessário identificar sempre o responsável pela consulta, sendo por isso preciso tomar uma decisão relativa aos registos nos quais o campo do optometrista não está completo. Uma outra decisão diz respeito à consideração, ou não, dos optometristas que apresentam poucas consultas - em alguns casos apenas de verificaram uma única consulta.

Tal como no caso anterior, também para a análise do optometrista, foi analisado o esquema relacional construído (Figura 3.5), tendo como objetivo selecionar todos os atributos importantes e que devem ser considerados, reunindo assim dados necessários à análise. Na Tabela 3-2 são expostos todos os atributos considerados, bem como a tabela a que pertencem.

Tabela 3-2 - Descrição dos atributos considerados para a construção do perfil do optometrista.

Campo	**Descrição**
tbMedicosTecnicos.IDTipoEntidade	Identificador que identifica cada entidade presente no conjunto de dados. Neste caso, sempre que IDTipoEntidade for igual a 13 refere-se à entidade optometrista.
tbMedicosTecnicos.Codigo	Código que identifica cada optometrista.
tbMedicosTecnicos.Nome	Nome do optometrista.
tbServicos.IDTipoServico	Identificador que identifica unicamente o tipo de serviço prestado pelo optometrista.
tbSistemaTiposServicos.Codigo	Código do tipo de serviço prestado.
tbSistemaTiposServicos.Descricao	Descrição do tipo de serviço prestado.
tbServicos.IDDocumentoVenda	Identificador único de cada documento de venda emitido.
tbDocumentosVendas.IDEntidade	Identifica o código associado a cada cliente.
tbDocumentosVendas.DescricaoConselhoFiscal	Identifica o local onde reside o cliente.
tbServicos.DataReceita	Data em que o optometrista prescreveu a receita.
tbDocumentosVendas.DataDocumento	Identifica a data e a hora em que o documento é emitido.
tbLojas.Codigo	Código que identifica cada loja.
tbLojas.Descricao	Descreve a loja em que foi efetuada a venda.
tbTiposArtigos.Codigo	Código do tipo de artigo.
tbTiposArtigos.Descricao	Descrição do tipo de artigo.

Contrariamente ao que aconteceu no caso anterior, neste caso não foi possível efetuar análises exploratórias aos dados, uma vez que todos os atributos considerados eram constituídos por carateres.

3.4 Preparação dos Dados

Na fase de preparação de dados realizaram-se várias tarefas, como a seleção de tabelas, atributos e registos, bem como se efetuou a transformação e limpeza de dados, de forma a que fosse possível aplicar posteriormente as ferramentas de modelação. Esta fase desenvolveu-se em cinco etapas: seleção de dados, limpeza de dados, construção de dados, integração de dados e formatação de dados (Shearer, 2000).

3.4.1 O Agente Comercial

Como o objetivo era o de elaborar o perfil do agente comercial, foi necessário aceder às informações relativas ao mesmo. Assim, inicialmente, consultou-se a tabela designada por "MedicosTecnicos" para obter o código e o nome do agente comercial responsável por cada venda e para confirmar se este era efetivamente um agente comercial, já que, na tabela consultada, existe informação relativa aos optometristas e aos agentes comerciais, sendo estes diferenciados pelo atributo "IDTipoEntidade" que é distinto para cada um deles - 13 para o optometrista e 14 para o agente comercial.

Para se conhecer o tipo de vendas de cada agente comercial, acedeu-se à tabela "DocumentosVendas" de modo a obter a identificação desta tabela, a data do documento em que o documento foi emitido, o nome fiscal, isto é, o nome do cliente, e a respetiva idade. Adicionalmente, acedeu-se também à tabela "TiposDocumento", filtrando o código e a descrição do tipo de documento, e à tabela "Clientes", para aceder ao código do cliente em questão.

De modo a saber-se o valor dos artigos vendidos e o desconto aplicado pelo agente comercial, consultou-se a tabela "DocumentosVendasLinhas" para se obter os atributos "PrecoUnitarioEfetivoSemIVA", que é o preço de cada unidade vendida ainda sem a aplicação da respetiva taxa de IVA, e "ValorDescontoEfetivoSemIVA", para se poder caraterizar o agente em relação aos descontos que este faz. De forma, a conhecer-se qual o estado em que se encontram os documentos, acedeu-se à tabela "SistemaTiposEstados" para se obter o seu código e descrição.

Para se caraterizar o agente comercial quanto ao tipo de artigos que este vende, consultou-se a tabela "TiposArtigos" para se obter o código e a descrição do tipo de artigo, mas antes foi necessário aceder à tabela "Artigos" de modo a obter o número exato de artigos vendidos, pois cada artigo vendido tem um código único e uma descrição associada, isto é, a tabela "TiposArtigos" era a forma de confirmar se o

artigo vendido era uma 'Armação' ou uma 'Lente Oftálmica', por exemplo. Complementarmente, para o tipo de artigo denominado 'Lentes Oftálmicas' averiguou-se o seu tipo de lente ('Unifocal', 'Oftálmica', 'Longe', 'Perto'). Assim, através da tabela "Modelos" obteve-se o seu "ID", identificador único de cada tipo de lente que existe na loja ótica, e pela tabela "SistemaTiposLentes" a respetiva descrição. Para que o perfil do agente comercial pudesse ser traçado da forma mais real possível, através da tabela "Marcas" obteve-se a marca dos artigos vendidos e detetou-se a preferência, ou não, do agente comercial por uma dada marca.

Para agrupar os dados obtidos construíram-se vistas específicas, que fornecessem os dados considerados importantes para a análise efetuada. Numa primeira fase, os dados são integrados numa vista e, depois, essa vista é transformada numa tabela materializada. A decisão de criar essa tabela deveu-se à necessidade da transformação dos dados, assim todos os dados nulos detetados no conjunto de dados poderiam ser eliminados ou corrigidos, sem que estas alterações afetassem toda a base de dados, mas se cingisse apenas à tabela materializada criada.

Como foi referido anteriormente, na fase de análise dos dados, existiam atributos com grande importância para a análise, mas que continham um número considerável de registos nulos. Assim, e para que a análise não fosse inflacionada por esses valores, decidiu-se retirar todos os registos que continham valores nulos, o que fez reduzir substancialmente o número de registos disponíveis para análise. Outra decisão que foi tomada, foi a remoção dos registos dos agentes comerciais que possuíam apenas três ou menos registos de vendas, e agrupar os restantes que tinham mais do que três desses registos, mas cujo o número era muito inferior ao número de registos dos três agentes comerciais com mais registos. Depois de ter sido feita a limpeza e transformação dos dados, o conjunto de dados ficou apenas com 319 registos, associados a seis agentes comerciais.

3.4.2 O Optometrista

Consultando a tabela "MedicosTecnicos" acede-se à informação relativa aos optometristas e aos técnicos, sendo estes diferenciados pelo atributo "IDTipoEntidade" que é diferente para cada um deles. Tal como foi explicado anteriormente, pretende-se obter também o código e o nome do optometrista responsável pela consulta que depois originou uma venda. Através da tabela "Servicos" consegue-se o a identificação do tipo de serviço prestado, a identificação do documento de venda e a data da receita. De

modo a conhecer-se qual o serviço prestado, consulta-se a tabela "SistemaTiposServicos" para se obter o código e a descrição do tipo de serviço prestado. Adicionalmente, através da tabela "DocumentosVendas" acede-se à data de venda do documento. Para que sejam conhecidos os artigos que foram vendidos, utilizamos a informação contida na tabela "TiposArtigos", para ter acesso aos tipos de artigos que constam no documento de venda gerado por um certo serviço.

Tal como aconteceu no caso do agente comercial, neste caso optou-se também por construir uma vista para agregar os dados necessários ao processo de análise pretendido. Os problemas relacionados com o não preenchimento do campo correspondente à identificação do optometrista, que foram detetados na fase anterior, solucionaram-se através da remoção dos registos que continham esse campo nulo. Relativamente aos optometristas que tinham um baixo número de consultas associada -, três no máximo -, decidiu-se retirá-los e agrupar os optometristas que tinham poucas consultas. Devido à necessidade de transformar alguns dos dados existentes, tomou-se a decisão de os armazenar numa tabela materializada, visto que as alterações efetuadas implicavam mudanças em toda a base de dados (não era desejável que tal acontecesse). Depois de realizadas todas as alterações, obtiveram-se 266 registos, divididos por seis optometristas.

CAPÍTULO 4

4. MODELAÇÃO DE MINERAÇÃO DE DADOS

Na quarta fase do processo CRISP-DM - a fase da modelação - faz-se a seleção das técnicas de modelação a aplicar, cria-se os modelos que parecem mais adequados ao problema e transita-se depois para a sua avaliação (Shearer, 2000). Neste capítulo, explica-se a modelação necessária para cada técnica de mineração de dados, nomeadamente, a resolução adotada para combater os problemas que poderiam surgir, explicando também como foram elaborados os modelos e qual o objetivo de cada um, isto é, o que se pretendia que os resultados dessem a conhecer.

4.1 Classificação – Modelos DT1, DT2, DT3, DT4, DT5 e DT6

Na tarefa de classificação foi utilizado um conjunto de dados de treino contendo atributos numéricos e categóricos, também chamados de atributos preditores, e atributos que indicam a que classe a que o registo pertence. O objetivo foi extrair do conjunto de dados de treino um modelo que descrevesse cada classe, a partir dos atributos preditores. Assim, o modelo gerado deveria ser capaz de prever a classe de registos dos quais se desconhece a respetiva classe.

A utilização de árvores de decisão é uma das técnicas possíveis para realizar a tarefa de classificação. A construção de uma árvore de decisão inicia-se a partir de um conjunto de treino, cujas classes dos exemplos que contém são previamente conhecidas. No processo de construção de uma árvore de decisão, alguns ramos podem conter anomalias causadas, principalmente, por "ruído" contido nos dados de treino. Este problema, designado de *overfitting*, provoca uma classificação bastante específica, o que não é desejável. Assim, a poda da árvore é a solução para resolver o problema detetado. Isso faz com que o processo de classificação seja mais rápido e bem como melhora a classificação dos dados. A pré-

poda é efetuada durante o processo de treino e interrompe a divisão do nó em função da avaliação de um dado conjunto de medidas.

O principal objetivo é fazer a geração de uma árvore de decisão que tenha uma elevada taxa de precisão. Porém, para que isso seja conseguido, é necessário escolher corretamente os atributos, de modo a que seja gerada uma árvore com o menor número possível de subconjuntos e que cada folha contenha um número significativo de casos.

O primeiro passo para criar estes modelos, é definir o atributo chave e o atributo previsível. Através desta escolha é possível ver qual a correlação que os restantes atributos têm com o atributo previsível. Existem, também, outros parâmetros que devem ser alterados para se obterem melhores resultados, tais como, o parâmetro que especifica a partir de que valor deve ser efetuada uma divisão na árvore de decisão ou qual o método que deve ser aplicado, entre outros. Na tabela 4-1, estão descritos os parâmetros do algoritmo *Microsoft Decision Tree*.

Tabela 4-1 – Descrição dos parâmetros do algoritmo *Microsoft Decision Tree* (tabela adaptada de Microsoft 1, 2011).

Parâmetros	Descrição
Complexity_Penalty	Controla o crescimento da árvore de decisão, quanto mais alto for o valor aplicado, menos divisões ocorrem e quanto menor for mais divisões acontecem. Se o número de atributos variar entre 1 e 9, o valor a aplicar é 0,5, se varia entre 10 e 99, aplica-se 0,9 e se o número de atributos for superior a 100, o valor é de 0,99.
Force_Regressor	Este parâmetro só é utilizado em árvores de decisão cujo atributo previsível é contínuo. A sua função é forçar o algoritmo a usar colunas especificadas como regressores, independentemente da importância dessas colunas.
Maximum_Input_Attributes	Define o número de atributos de entrada que o algoritmo pode manipular antes da seleção de recursos. O valor padrão é 255, quando se pretende desativar a seleção de recursos, este valor deve ser definido como 0.
Maximum_Output_Attributes	Tem a mesma função que o parâmetro anterior, *Maximum_Input_Attributes*, com a diferença de manipular os atributos de saída ao invés de manipular os de entrada. Os valores a aplicar são iguais aos do parâmetro anterior.
Minimum_Support	Determina qual o número mínimo de casos numa folha para que ocorra divisão na árvore de decisão. O valor padrão é 10, mas se o conjunto de dados for grande, este valor deve ser aumentado para evitar o problema de *overffiting*.
Score_Method	Define qual o método que deve ser usado para calcular a pontuação da divisão. Os métodos são 1- Entropia, 2- *Bayesian with K2 Priori* e 3- *Bayesian Dirichlet Equivalent* (BDE). O valor padrão é 3.
Split_Method	Determina o método aplicado na divisão do nó. Existe o método 1-Binário, 2- Completo e o 3- Both, sendo este último o valor padrão.

Assim, com recurso a esta técnica, foram criados seis modelos de classificação: três para o agente comercial e três para o optometrista. De seguida, explica-se cada um dos modelos elaborados, os

cenários formulados e a forma como se alteraram os parâmetros de cada modelo de modo a obterem-se modelos significativos - a avaliação do desempenho dos modelos é exposta de seguida no capítulo 5.

Para cada um dos modelos que são explicados de seguida, o número de atributos de entrada varia entre dois e três, o método que calcula a pontuação da divisão selecionado em todos os cenários foi a entropia, e, de modo a controlar o crescimento da árvore de decisão o valor aplicado no parâmetro *Complexity_Penalty* é 0,5, pois o número de atributos considerados é inferior a 10.

4.1.1 O Agente Comercial

Tendo em vista a construção do perfil do agente comercial, foram elaboradas várias árvores de decisão considerando o atributo "CodigoCliente" como chave e o atributo "CodigoTecnico" como previsível. Relativamente aos atributos de entrada (*inputs*), estes foram selecionados tendo em conta os valores de correlação apresentados na Figura 4.1. Estes valores são calculados pelo *Microsoft Visual Studio,* automaticamente.

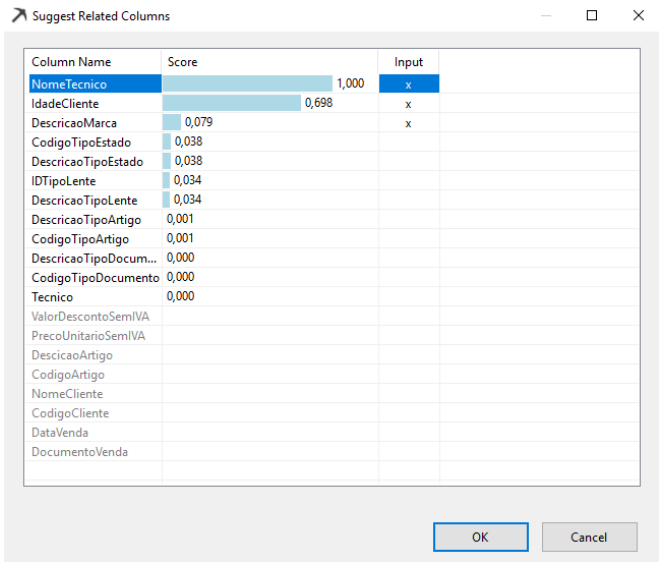


Figura 4.1 - Valores de correlação para os modelos do agente comercial.

Tal como foi referido, foram elaborados três diferentes cenários para a construção das árvores de decisão. O objetivo foi clarificar ao máximo a árvore de decisão, obtendo resultados de fácil interpretação, ao invés de colocar todos os atributos, alcançando resultados confusos, com menos informação e de difícil interpretação. De realçar, que no caso do agente comercial, quando foi elaborado o primeiro

modelo, foi necessário regressar à fase de preparação de dados, devido à existência de valores nulos nos atributos "DescricaoTipoLente" e "DescricaoMarca" (Secção 3.4).

Modelo DT1

No primeiro modelo, consideraram-se como atributos de entrada a "DescricaoTipoLente" e a "DescricaoMarca". Assim, pretende-se compreender se existe alguma relação entre o tipo de lente e a sua marca, isto é, se o agente comercial perante um tipo de lente tende sempre a recomendar uma dada marca, e se esta é uma caraterística apenas sua, ou do conjunto de agentes comerciais. Porém, pode também tratar-se de uma política da loja. A propriedade *HoldoutMaxPercent*, que não foi referenciada anteriormente, especifica o número de casos a serem incluídos no conjunto de testes - uma percentagem do conjunto de dados. À medida que este número aumenta, o número de casos em análise diminui. Se se diminui o número de casos aumenta. O valor padrão é 30. Neste modelo, definiu-se este valor como 20, depois de se variar o valor, percebeu-se que na árvore de decisão, variava o número de casos. Porém, a caraterística que tinha maior número de casos era sempre a mesma. Além disso, a taxa de precisão mudava e verifica-se que com este valor de *HoldoutMaxPercent* obtinha-se a taxa de precisão mais alta para este modelo e, consequentemente, a menor taxa de erro. O valor mínimo de casos que cada nó deveria ter foi definido em 5 casos. A árvore de decisão obtida (Figura 4.2) analisa 123 dos 319 casos presentes no conjunto de dados.

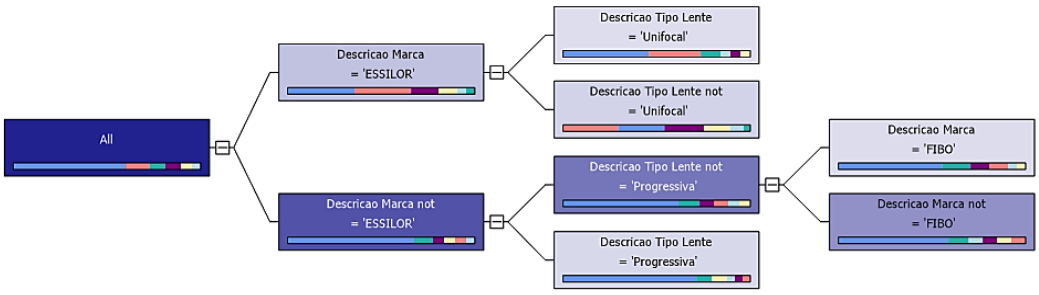


Figura 4.2 – A árvore de decisão referente ao modelo DT1.

Modelo DT2

No segundo modelo, o cenário formulado considerou os atributos "DescricaoTipoLente" e "IdadeCliente" como *inputs*. O objetivo foi perceber se existe um tipo de lente que seja mais recomendado a uma faixa etária e qual o motivo de isso ter acontecido. O atributo "IdadeCliente" foi discretizado e dividido em 5 diferentes faixas etárias. O valor mínimo para que ocorresse uma divisão na árvore de decisão foi fixado em 3. Neste modelo, o valor de percentagem *HoldoutMaxPercent* com o qual se obtém o melhor modelo

foi de 40. Assim, dos 319 casos do conjunto de dados, a árvore de decisão analisa 92 casos (Figura 4.3).

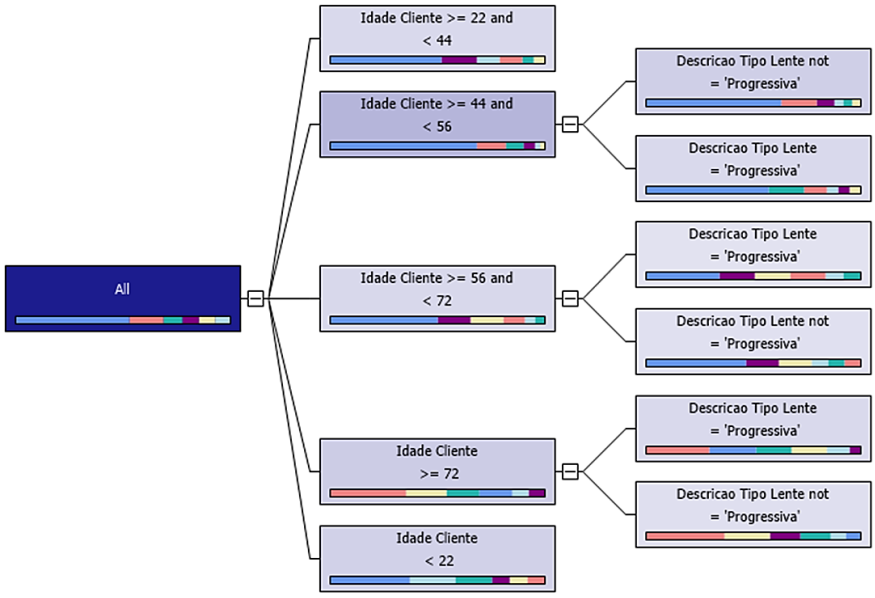


Figura 4.3- A árvore de decisão referente ao modelo DT2.

Modelo DT3

No último modelo concebido para o agente comercial, foram considerados três atributos de entrada: "DescricaoTipoLente", "IdadeCliente" e "PrecoUnitarioSemIVA". O objetivo deste modelo foi relacionar o tipo de lente com a idade do cliente, tal como foi feito no modelo anterior, mas agora adicionando o preço unitário da lente, com o propósito de compreender se existe uma relação entre a idade e o preço da lente vendida. Para que ocorresse uma divisão na árvore, foi preciso que a folha tivesse no mínimo 3 casos. O valor da propriedade *HoldoutMaxPercent* fixou-se em 25, o que permitiu a análise de 115 dos 319 casos do conjunto de dados (Figura 4.4).

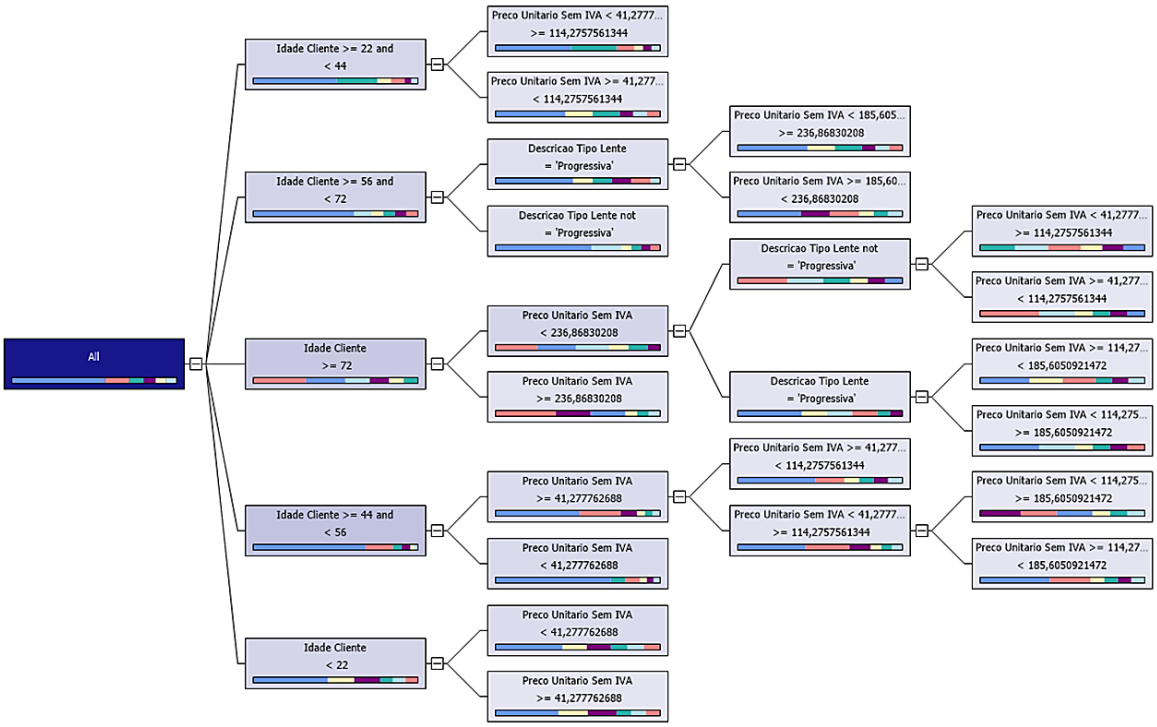


Figura 4.4 – A árvore de decisão referente ao modelo DT3.

4.1.2 O Optometrista

Na construção dos modelos do optometrista, à semelhança do que sucedeu no agente comercial, foi preciso definir o atributo chave, os atributos de entrada e o atributo previsível. Neste caso, o atributo chave foi novamente o "CodigoCliente" e o atributo previsível o "CodigoMedico". Os atributos de entrada foram selecionados considerando-se a correlação destes atributos com o previsível. Os valores de correlação estão apresentados na Figura 4.5.

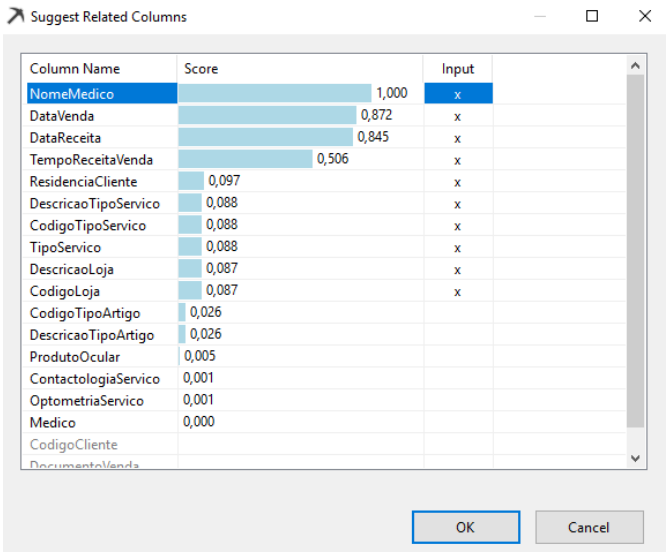


Figura 4.5 - Valores de correlação para os modelos do optometrista.

Para o caso do optometrista, foram elaborados três modelos. Os atributos de entrada foram selecionados tendo em conta o valor da correlação e os atributos que, juntos, poderiam retornar informação útil.

Modelo DT4

Neste modelo, consideraram-se como *inputs* os atributos "DescricaoLoja" e "DescricaoTipoServico". Aqui o objetivo foi compreender se, tendo em conta a localização da loja, existia um tipo de serviço que tivesse maior probabilidade de ocorrer. O valor mínimo que cada nó tinha de conter para que ocorresse nova divisão é 3. Relativamente à propriedade *HoldoutMaxPercent*, o valor foi fixado em 20, uma vez que este valor foi o que gerou o melhor modelo em termos de precisão. Assim, a árvore de decisão que resultou deste modelo (Figura 4.6) permite analisar 108 registos dos 266 existentes no conjunto de dados.

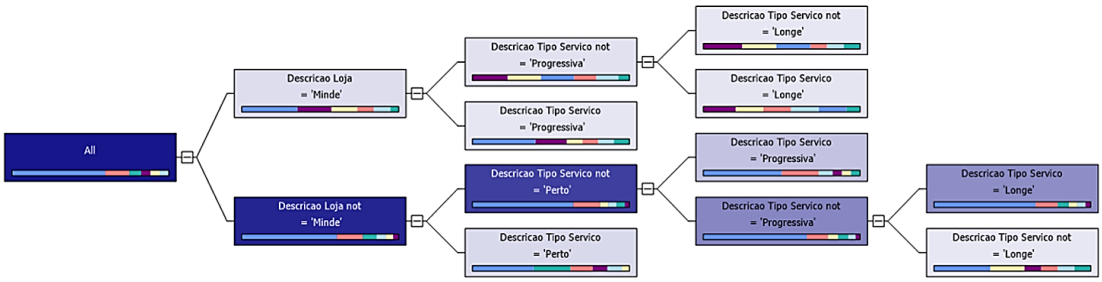


Figura 4.6 – A árvore de decisão referente ao modelo DT4.

Modelo DT5

Neste modelo, escolheram-se os atributos "CodigoLoja" e "ResidenciaCliente" como *inputs*. O objetivo do quinto modelo foi comprovar que existia uma relação entre a localização da loja e a residência do cliente. O valor de casos considerado mínimo para que ocorresse uma nova divisão foi de 3, e a percentagem utilizada de 25. Portanto, dos 266 casos contidos no conjunto de dados, a árvore de decisão permite analisar 101 casos (Figura 4.7).

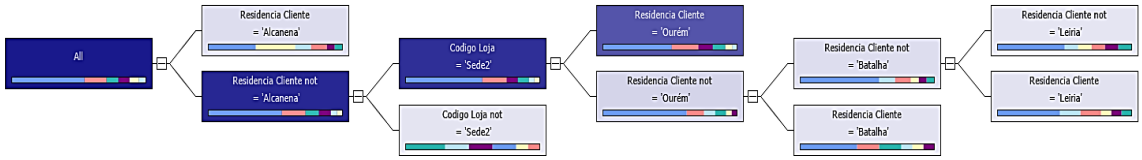


Figura 4.7 – A árvore de decisão referente ao modelo DT5.

Modelo DT6

No último modelo do optometrista, e também de classificação, optou-se pela seleção de três atributos de entrada: o "CodigoLoja", a "ResidenciaCliente" e a "DescricaoTipoServico". O objetivo deste modelo foi entender se existia alguma relação entre estes três atributos. Para que ocorresse uma divisão na árvore de decisão, o nó devia conter pelo menos três casos. Definiu-se a propriedade *HoldoutMaxPercent* com o valor de 25. A árvore de decisão apresentada na Figura 4.8 permite analisar 101 casos do conjunto de dados disponíveis para análise.

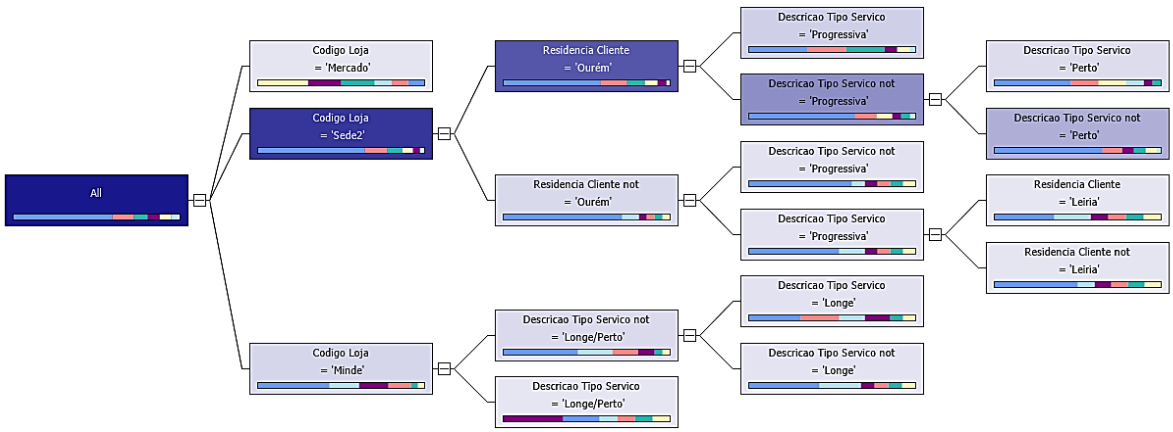


Figura 4.8 – A árvore de decisão referente ao modelo DT6.

4.2 Segmentação – Modelos C1, C2, C3 e C4

Com a utilização de *clustering* teve-se como objetivo identificar de forma automática comportamentos similares num dado conjunto de dados, segmentando a informação em *clusters*, isto é, agrupando objetos semelhantes em conjuntos. Contrariamente, ao que acontece em classificação, na segmentação não existem classes pré-rotuladas. As instâncias de dados são agrupadas maximizando a similaridade intracluster e minimizando a similaridade intercluster. Após o término do processo de segmentação é o analista quem deve estudar os padrões identificados e concluir se podem ou não ser transformados em conhecimento estratégico.

O sistema de clustering da *Microsoft* disponibiliza dois algoritmos de segmentação: o *k-means* e o *expectation-maximization*. O primeiro tem a premissa de que cada ponto de dados apenas pode pertencer a um cluster e que é calculada uma única probabilidade para a associação de cada ponto de dados nesse cluster. O segundo é mais flexível que o primeiro, pois cada ponto de dados pode pertencer a mais que um cluster, sendo assim calculada uma probabilidade para cada combinação do ponto de dados com o cluster. Tal como o algoritmo anterior, este também possui parâmetros que devem ser ajustados de modo a obterem-se modelos mais esclarecedores (Tabela 4-2).

Tabela 4-2 - Descrição dos parâmetros do algoritmo *Microsoft Clustering* (tabela adaptada da Microsoft 2, 2011)

Parâmetro	Descrição
Clustering_Method	Define qual o método de segmentação que o algoritmo deve aplicar. Existem quatro métodos disponíveis: 1- EM escalonável, 2- EM não escalonável, 3- *k-means* escalonável e 4- *k-means* não escalonável. O padrão é 1.
Cluster_Count	Determina o número aproximado de *clusters* que devem ser formados. Caso não possa construir o número especificado a partir do conjunto de dados, o algoritmo forma o número mais próximo de *clusters.*
Cluster_Seed	Especifica o número semente que é utilizado para gerar, de forma aleatória, os *clusters* iniciais. Quando se altera o número semente inicial, a forma como foram gerados os *clusters* pode ser modificada, devendo-se depois comparar os *clusters* gerados por diferentes sementes. Quando a semente é alterada e não ocorrem mudanças significativas nos *clusters*, o modelo é, minimamente, estável. O valor padrão é 1.
Minimum_Support	Determina o número mínimo de casos para se criar um *cluster*. Deve-se ter cuidado na escolha do número, visto que se for maior que o número de casos no *cluster* este é considerado vazio e é descartado, se for definido com um número elevado, é possível que se percam *clusters* válidos. O padrão é 1.
Modelling_Cardinality	Define o número de modelos construídos durante o processo de segmentação. Reduzindo este número, o desempenho do algoritmo é melhor, no entanto, pode-se perder alguns bons modelos. O valor padrão é 10.
Stopping_Tolerance	Valor que é utilizado para determinar quando é atingida a convergência e o algoritmo termina de construir o modelo. A convergência é atingida quando as variações nas probabilidades do *cluster* são menores que a proporção do parâmetro a dividir pelo tamanho do modelo. O valor padrão é 10.
Sample_Size	Quando o método aplicado é escalonável, este parâmetro define o número de casos usados em cada passagem. Se este valor for definido como 0, é realizada apenas uma passagem com todos os casos, e caso o conjunto de dados seja grande podem ocorrer problemas de memória ou desempenho. O valor padrão é 50000.

Maximum_Input_Attributes	Define o número de atributo de entrada, *inputs*, que se podem manipular antes de fazer a seleção de recursos. Especificar este valor como 0 significa que não existe número máximo de atributos, aumentar este número pode degradar o desempenho do algoritmo. O valor padrão é 255.
Maximum_States	Número máximo de estados que o algoritmo suporta. Caso um atributo possua mais estados que o máximo, o algoritmo considera os mais populares e ignora os restantes.

Quando se inicia o processo de segmentação é necessário considerar um atributo chave e atributos de entrada. Um atributo definido como *predict* é utlizado como atributo de entrada. Porém se este atributo estiver como *predict only*, então não é considerado na criação dos *clusters*. Assim, construíram-se quatro modelos de segmentação, dois para o agente comercial e dois para o optometrista.

4.2.1 O Agente Comercial

Com o objetivo de se aplicar o algoritmo de *clustering* foi necessário selecionar o atributo chave e os atributos de entrada, sendo que o atributo chave considerado foi o "CodigoCliente" e os de entrada são diferentes nos dois modelos. Os parâmetros do modelo foram alterados de modo a obter melhores conjuntos de *clusters*. O algoritmo aplicado nestes dois modelos foi o *k-means* escalonável.

Modelo C1

Neste primeiro modelo de *clustering*, consideraram-se como *inputs* os atributos "CodigoTecnico", "DescricaoTipoLente" e "IdadeCliente". O objetivo foi perceber quais os clusters que seriam criados tendo em conta estes atributos, de forma a poder verificar se existia uma maior probabilidade de uma dada faixa etária adquirir mais um determinado tipo de lente e se é mais provável que seja vendida por um dado agente comercial. De modo a obterem-se *clusters* com um número considerável de casos, definiu-se que deveriam ser criados cinco clusters e que o número mínimo de casos em cada cluster deveria ser de 17. Neste modelo, a propriedade *HoldoutMaxPercent* foi fixada em 20. Para este caso o diagrama de clustering obtido (Figura 4.9) analisou um total de 123 casos.

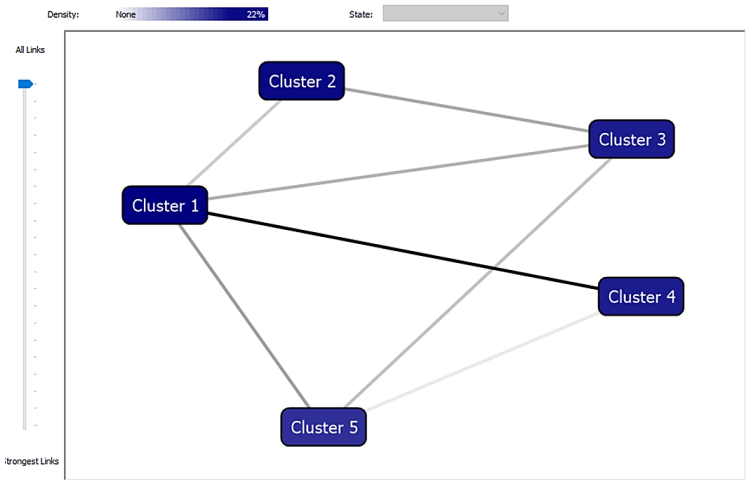


Figura 4.9 - Diagrama de *clustering* referente ao modelo C1.

Neste modelo a densidade é de 22%, dado que permite concluir que os cinco clusters têm aproximadamente o mesmo número de casos, isto é, quando comparado o número de casos entre os clusters, não existe valores muito dispares, mas sim muito semelhantes.

Modelo C2

No segundo modelo de clustering, consideraram-se como *inputs* os atributos "CodigoTecnico" e "PrecoUnitarioSemIVA". O número máximo de *clusters* foi definido em três e o valor mínimo de casos em cada *cluster* de 20. A percentagem de casos utilizados no processo de segmentação foi de 30 e os casos analisados 108. Na Figura 4.10 pode-se ver o correspondente diagrama de *clustering*.

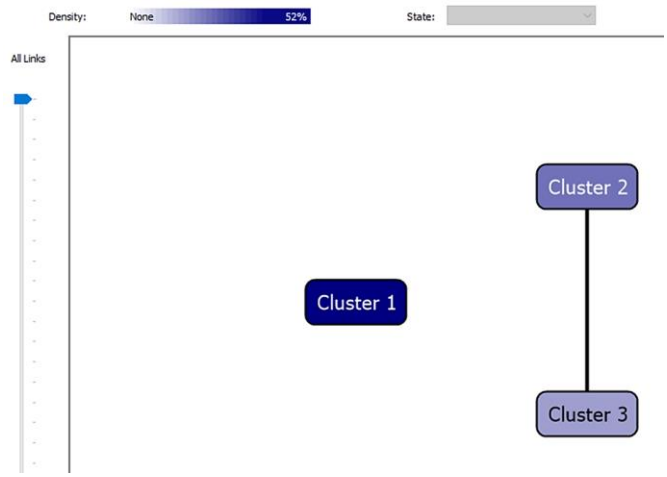


Figura 4.10 - Diagrama de *clustering* referente ao modelo C2.

Como se pode observar pela Figura 4.10, o *cluster* 1 apresenta 52% dos casos presentes no conjunto de dados. Este facto pode ser explicado devido ao elevado número de venda de lentes oftálmicas unifocais, que geralmente apresentam preços unitários mais baixos.

4.2.2 O Optometrista

Tal como no agente comercial, também neste caso foi preciso escolher o atributo chave e os atributos de entrada. O atributo chave utilizado nestes modelos foi "CodigoCliente", os atributos de entrada variam nos três modelos e os parâmetros foram alterados de modo a gerar modelos de *clustering* estáveis.

Modelo C3

No primeiro modelo de clustering do optometrista, os atributos considerados *inputs* foram "CodigoLoja", "CodigoMedico" e "DescricaoTipoServico". Aqui definiu-se que o número máximo de *clusters* deveria ser 5 e que cada *cluster* deveria ter no mínimo 15 casos. O algoritmoaplicado neste modelo foi o EM escalonável, tendo sido utilizada uma percentagem de 20 em *HoldoutMaxPercent*. No diagrama de clustering foram considerados 108 casos (Figura 4.11), sendo possível observar que o número de casos em cada *cluster* não se diferencia muito. A densidade é de 32%.

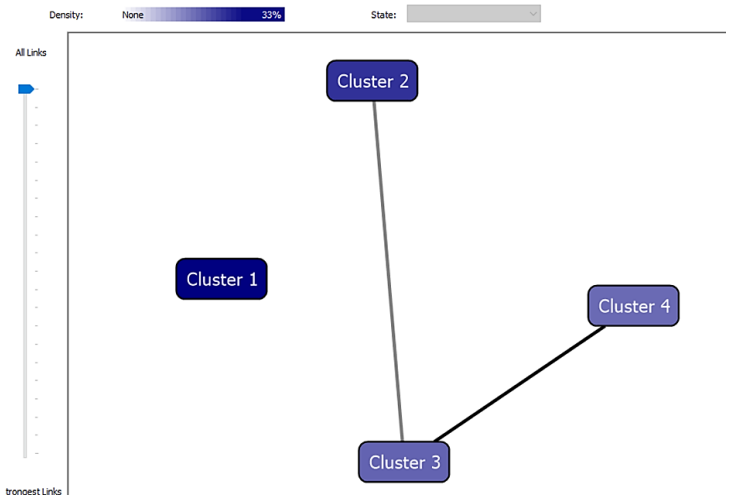


Figura 4.11 - Diagrama de *clustering* referente ao modelo C3.

Modelo C4

No último modelo de *clustering*, os *inputs* selecionados foram "CodigoMedico" e "DescricaoTipoServico". Neste modelo definiu-se que o número máximo de *clusters* seria de 3 e que o número mínimo de casos, em cada *cluster*, de 3. O algoritmo aplicado foi o EM escalonável, empregando-se uma percentagem de 20 na propriedade *HoldoutMaxPercent*. Na Figura 4.12 pode-se ver o diagrama de *clustering* resultante, que inclui 216 casos, com uma densidade de 38%.

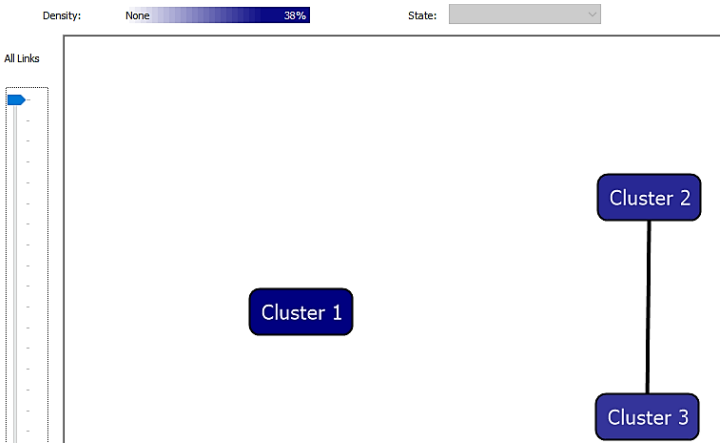


Figura 4.12 - Diagrama de *clustering* referente ao modelo C4.

4.3 Associação – Modelos A1, A2, A3, A4

As regras de associação encontram-se entre um dos mais importantes tipos de conhecimento que podem ser descobertos em bases de dados. O resultado da aplicação desta técnica permite descobrir padrões de relacionamento entre itens de uma base de dados. Basicamente, o objetivo é encontrar tendências que ajudem a compreender tais padrões, descobrindo eventuais relações entre os itens e verificar os eventos que ocorrem simultaneamente, o que proporciona a elaboração de novos modelos e, como tal, melhores resultados. Uma das aplicações mais conhecidas de regras de associação é a análise de transações de compras, vulgarmente reconhecido como *Market Basket Analysis*, cujo objetivo é identificar os padrões de compra de consumidores e detetar quais os produtos que costumam comprar em conjunto.

Os modelos de associação são construídos com base em conjuntos de dados que contêm identificadores de casos individuais e de itens contidos em casos. A um grupo de itens num determinado caso designa-se conjunto de itens. Um algoritmo de associação costuma utilizar dois parâmetros, o suporte e a probabilidade (ou confiança), para descrever o conjunto de itens e as regras que gera. O suporte

representa o número de casos contidos na combinação de itens e a confiança representa a fração de casos do conjunto de dados que contêm A mas também B, ou seja, enquanto o suporte indica o número de casos, a confiança calcula a probabilidade que existe de ocorrer essa combinação de itens.

À semelhança do que acontece nos modelos de classificação e segmentação, para se alcançar bons resultados nos modelos de associação, estes existem parâmetros que podem ser alterados de forma a alcançar melhores resultados. Esses parâmetros encontram-se descritos na Tabela 4-3.

Tabela 4-3 - Descrição dos parâmetros do algoritmo *Microsoft Association Rules* (tabela adaptada da Microsoft 3, 2018)

Parâmetro	Descrição
Maximum_Itemset_Count	Define o número máximo de conjunto de itens a serem formados. Quando não se define um valor, é utilizado o valor padrão que é 200000.
Maximum_Itemset_Size	Determina o número máximo de itens num conjunto de itens. Sempre que o valor é definido como 0, então não existe limite no tamanho do conjunto de itens. O valor padrão é 3.
Maximum_Support	Restringe o número máximo de casos que suporta um conjunto de itens. Este parâmetro pode eliminar itens muito frequentes, que potencialmente têm pouco significado. Pode ser expresso em número decimal menor que um ou em número inteiro. O valor padrão é 1.
Minimum_Itemset_Size	Este parâmetro tem o mesmo objetivo que *Maximum_Itemset_Count,* isto é, especifica o número mínimo de itens num conjunto de itens. Quando se pretende ignorar conjuntos de itens que apenas contém um item, este valor deve ser aumentado. O valor de referência é 1.
Minimum_Probability	Define a probabilidade mínima para que uma regra seja considerada verdadeira. O valor por defeito é 0,4.
Minimum_Support	Determina o número mínimo de casos que deve constar no conjunto de itens, ainda antes do algoritmo gerar uma regra. Este valor pode ser aumentado pelo algoritmo, caso exista restrição de espaço de memória. O valor padrão é 0,03.
Optimized_Prediction_Count	Número de itens que deve ser armazenado para otimizar a previsão. O valor padrão é 0, e quando aplicado gera o número de previsões pedidas na consulta. Definir um valor pode melhorar o desempenho da previsão.

Como aconteceu nos algoritmos das técnicas anteriores, quando se inicia um modelo de mineração de dados com um algoritmo de associação é necessário definir um atributo chave, atributos de entrada e um atributo previsível.

4.3.1 O Agente Comercial

No caso do agente comercial o atributo definido como chave foi o "CodigoCliente" e o atributo previsível o "CodigoTecnico", enquanto que os atributos de entrada variam nos diferentes modelos. Os parâmetros foram afinados de diferentes formas de modo a obter as melhores regras de associação.

Modelo A1

No primeiro modelo de associação pretendeu-se testar a associação entre a marca e o tipo de lente. Nesse sentido, definiu-se como atributos de entrada os atributos "DescricaoMarca" e a "DescricaoTipoLente". Ao parâmetro Minimum_Importance atribui-se o valor 0,05 e à propriedade HoldoutMaxPercent o valor de 35. O algoritmo de associação utilizado, *Association Rules*, foi aplicado ao conjunto de dados, gerando sete regras de associação, com uma probabilidade mínima de 0,50 e uma importância mínima de 0,07 (Figura 4.13). Numa primeira observação, pode-se ver que todas as regras obedecem ao critério de sucesso definido, ou seja, o valor de confiança é inferior a um.

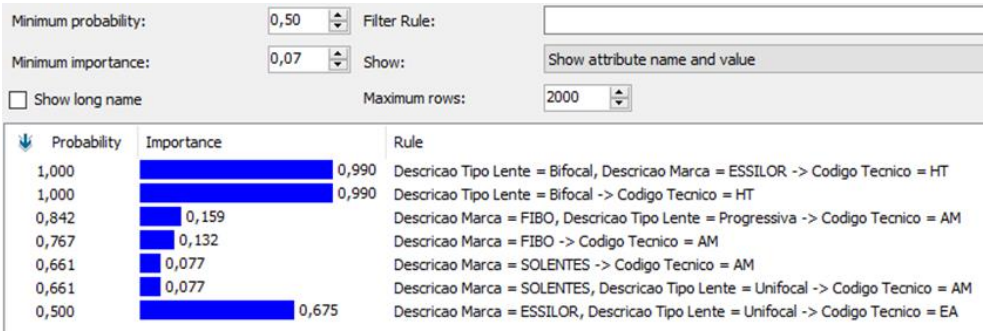


Figura 4.13 - Regras de associação referentes ao modelo A1.

Modelo A2

No segundo modelo de associação, consideraram-se os atributos "DescricaoTipoLente" e "IdadeCliente" como *inputs*. Definiu-se, também, que o valor de *Minimum_Importanc*e seria de 0,05 e o mínimo de casos num conjunto de itens, *Minimum_Support*, de 2. À propriedade HoldoutMaxPercent atribui-se o valor 20. O modelo gerou seis regras de associação, cuja probabilidade varia entre 0,40 e 0,75 e o valor de importância mínimo entre 0,06 e 0,87 (Figura 4.14).

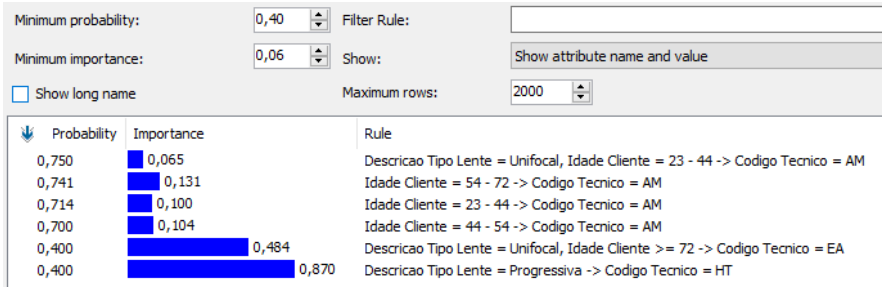


Figura 4.14 - Regras de associação referentes ao modelo A2.

Observando as regras de associação apresentadas na Figura 4.14, constata-se que em todas elas o valor de importância é inferior a um. Por esse motivo, as regras cumprem o critério de sucesso definido.

4.3.2 O Optometrista

Os modelos gerados para o caso do optometrista consideraram o atributo "CodigoCliente" como chave e o atributo "CodigoMedico" como previsível. Os atributos de entrada são diferentes nos dois modelos elaborados e os parâmetros, mais uma vez, modificaram-se de modo a se obterem melhores resultados.

Modelo A3

Este modelo de associação considerou como *inputs* os atributos "DescricaoLoja" e "DescricaoTipoServico". Definiu-se que o valor mínimo de confiança seria de 0,05 e o mínimo de casos presentes num conjunto de 2. À propriedade *HoldoutMaxPercent* foi atribuído o valor de 20. Após a aplicação do algoritmo foram geradas oito regras de associação (Figura 4.15), cuja probabilidade varia entre 0,40 e 0,67, e a confiança entre 0,08 e 0,906.

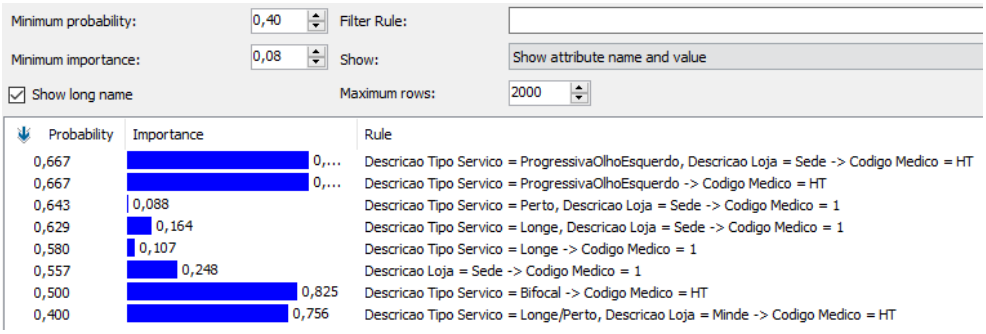


Figura 4.15 - Regras de associação referentes ao modelo A3.

Todas as regras de associação que foram descobertas apresentam valores de confiança inferiores a um, pelo que respeitam o critério de sucesso definido.

Modelo A4

No quarto modelo de associação pretendeu-se testar a associação entre o tipo de serviço e o tipo de artigo para um dado optometrista. Os *inputs* definidos foram os atributos "DescricaoTipoArtigo" e "DescricaoTipoServico". O valor de mínima importância foi de 0,05, o suporte mínimo de casos de 2, e a propriedade *HoldoutMaxPercent* o valor de 20. O modelo gerou 8 regras de associação (Figura 4.16) com valores de probabilidade compreendidos entre 0,50 e 1,00 e de confiança entre 0,06 e 1,03.

Figura 4.16 - Regras de associação referentes ao modelo A4.

Numa primeira análise, pode-se concluir que das oito regras, uma não cumpre o critério de sucesso definido, pois o seu valor de confiança é superior a um.

CAPÍTULO 5

5. ANÁLISE E AVALIAÇÃO DOS MODELOS DE MINERAÇÃO

Na fase da avaliação, é fundamental compreender se existe alguma questão comercial que não foi considerada suficientemente. No final desta fase, o analista deve decidir como usar os resultados de mineração de dados. As principais etapas desta fase são a avaliação dos resultados, a revisão do processo e a definição do próximo passo a dar, isto é, se é necessário refazer todo o processo de mineração ou se é possível avançar para a fase de implementação (Shearer, 2000).

5.1 Modelos DT1, DT2, DT3, DT4, DT5 e DT6

Vamos agora analisar cada um dos modelos de árvores de decisão utilizados, evidenciando os nós com maior importância e, no fim, apresentar os valores de precisão e taxa de erro, indicando quais os modelos que obedecem ao critério de sucesso que foi definido.

Nos modelos gerados para adquirir conhecimento sobre o agente comercial, retiraram-se as seguintes observações:

- No modelo **DT1**, no nó raiz, existem 123 casos em análise, dos quais 74 são da responsabilidade do agente comercial 'AM', 10 do agente 'AC' e do 'MM', 16 do agente EA, 5 do FC e 8 do 'HT'. Este nó divide-se em dois outros nós: no primeiro, em 34 dos casos a marca vendida é a 'ESSILOR', senda esta vendida maioritariamente pelos comerciais 'AM' e 'EA', 13 e 11 unidades, respetivamente. Após esse nó existe uma outra divisão, que resulta em dois nós folha. Na primeira divisão, registam-se 12 casos, sendo que a decisão da descrição do tipo de lente ser 'Unifocal', é tomada em 7 casos pelo agente comercial 'AM' e 4 pelo 'EA'. No seguinte nó, quando a descrição do tipo de lente não é 'Unifocal', ocorrem 22 casos, divididos por cinco dos seis técnicos. Na segunda divisão, ou seja, quando a descrição da marca não é 'ESSILOR',

tem-se 79 casos divididos da seguinte forma: 61 para o agente 'AM', 5 para os comerciais 'AC' e 'EA', 4 para o 'FC' e por último, 4 para o agente comercial 'FC'. Neste nó, tem-se duas divisões e 72 casos ocorrem quando a descrição do tipo de lente é diferente de 'Progressiva' - o agente comercial 'AM' apresenta um maior número de casos. Este nó divide-se, tendo como decisões se a marca é 'FIBO' ou não. A decisão com mais casos é quando a marca não é 'FIBO'. Quando a descrição do tipo de lente é 'Progressiva' registam-se 17 casos.

- No modelo **DT2**, os 92 casos analisados encontram-se distribuídos pelos 6 agentes comerciais da seguinte forma: 8 da responsabilidade do agente 'AC', 49 do comercial 'AM', 15 do 'EA', 7 casos do agente comercial 'HT' e 'MM', e por último, 6 casos para o 'FC'. Da primeira divisão da árvore de decisão resultam cinco nós, referentes à idade. Quando a idade do cliente é igual ou superior a 22, mas inferior a 44, ocorrem 12 casos, 8 deles da responsabilidade do agente comercial 'AM', e na faixa etária inferior a 22 anos, dos 17 casos que se verificam, 7 pertencem ao agente 'AM'. Nas restantes faixas etárias ocorreu uma divisão do nó, originando dois nós folha para cada. Se a idade do cliente for igual ou superior a 44 e inferior a 56, ocorrem 30 casos, que em 23 deles o agente comercial 'AM' é o responsável. Neste nó, há uma divisão, e em 18 casos verifica-se que a descrição do tipo de lente não é 'Progressiva', e em 12 casos acontece o contrário. No entanto, tanto na primeira como na segunda decisão, o agente 'AM' é o responsável por mais vendas. Se a idade do cliente for igual ou superior a 56 e inferior a 72, contabilizam-se 13 casos. Destes casos, em 6 a decisão foi a descrição do tipo de lente ser do tipo 'Progressiva', e nos restantes, vendeu-se outro tipo de lente. O agente maioritariamente responsável pelas vendas é o 'AM'. Se a idade for igual ou superior a 72, verificam-se 20 casos, sendo que em 12 a venda é de lentes do tipo 'Progressiva'. O agente comercial com mais vendas é o 'EA'.
- No modelo **DT3**, dos 115 casos analisados, mais de cinquenta por cento dos casos, foram da responsabilidade do agente 'AM', 66 casos, 17 do agente 'EA', 10 do 'MM', 8 do 'AC', e, por fim, 7 no caso dos agentes 'FC' e 'HT'. A árvore de decisão divide-se em cinco nós relativos a cada classe da idade dos clientes. Se a idade do cliente é igual ou superior a 56 e inferior a 72, geralmente o agente comercial opta por lentes do tipo 'Progressivas' em 11 casos, e o preço situa-se entre 185,61 e 236,87 euros, em 6 desses casos. Estes casos ocorrem maioritariamente com o agente comercial 'AC'. Se a idade do cliente for igual ou superior a 72, o preço do artigo é inferior a 236,87 euros e em 12 casos verifica-se que as lentes não são do tipo 'Progressivas', e existe uma outra restrição no preço, este é inferior a 41,28 e superior a

114,28. O agente comercial responsável por mais vendas nestas condições é o 'EA', seguido pelo 'HT'. Se a idade do cliente é igual ou superior a 44 e inferior a 56, o que acontece em 30 casos dos 115 analisados, o mais provável, 16 casos, é que o preço do artigo esteja abaixo dos 41,28 euros. No entanto, em 14 casos, o preço é igual ou superior a 41,28, mas em cinco desses casos, o valor é também inferior a 114,28, enquanto que os restantes são superiores a esse valor. Dos restantes nove casos, em três o valor pode ser igual ou superior a 185,61 euros, e 6 têm preço igual ou superior a 114,28 euros e inferior a 185,61 euros.

No que concerne aos valores de precisão e taxa de erro destes modelos, estes estão apresentados na Tabela 5-1.

Tabela 5-1 - Valores da precisão e taxa de erro dos modelos do agente comercial.

Modelo	Precisão	Taxa de Erro
DT1	0,5333	0,4667=46,67%
DT2	0,6393	0,3607=36,07%
DT3	0,6842	0,3158=31,58%

Pode-se, assim, concluir que todos os modelos obedecem ao critério de sucesso definido, ou seja, todos apresentam uma taxa de erro inferior a 50%. Como a precisão de todos os modelos se situa entre 0,3 e 0,7, também se pode concluir que estes estão sujeitos a sofrer alterações, com risco moderado, quando receberem informações novas.

Através da análise aos modelos construídos é possível retirar algumas conclusões sobre o perfil dos agentes comerciais, tais como:

- Quando a marca da lente é 'ESSILOR' é provável que o tipo de lente vendido não seja 'Unifocal' e o agente comercial mais provável é 'AM'.
- Se a idade do cliente é igual ou superior a 72, o tipo de lente mais vendido é a 'Progressiva' e o agente comercial responsável é 'EA'.
- Enquanto que se a idade for igual ou superior a 56 e inferior a 72, as lentes mais vendidas é o tipo 'Progressiva' pelo agente comercial 'AC'. Estas lentes progressivas apresentam preços mais elevados (185,61 a 236,87 euros).
- Já quando as vendas são da responsabilidade dos agentes comercias 'EA' e 'HT' e a idade do cliente é igual ou superior a 72, a venda do tipo de lente é diferente de progressiva, com valor entre os 41,28 e 114,28euros.

Relativamente aos modelos que visam informação relativa ao optometrista, podem-se retirar as seguintes observações:

- No modelo **DT4**, dos 108 casos analisados, 65 são da responsabilidade do optometrista identificado com o código '1', '17' do 'DE', '8' do 'Dr. JM', e os optometristas 'DR' 'AC', 'Dr.HL' e 'HT' têm 6 casos a seu cargo. Após o nó raiz, a árvore de decisão divide-se em dois nós. No primeiro a descrição da loja é 'Minde', e ocorrem 12 casos. Dos casos em análise em 4 a descrição do tipo de serviço prestado é 'Progressiva', e dos restantes 5 são do tipo 'Longe'. Conclui-se que dos serviços prestados em 'Minde', grande parte é prestado pelo médico '1'. No segundo nó, isto é, quando a loja não é 'Minde', desses 96 casos, 15 deles ocorrem quando o tipo de serviço é 'Perto', 28 quando não é 'Progressiva' e 50 são do tipo 'Longe'. O optometrista que apresenta maior taxa de todos estes serviços é o '1'.
- No modelo **DT5**, foram analisados 101 casos, sendo que em 55 a consulta foi realizada pelo optometrista '1', 17 pelo 'DE', 9 pelo 'Dr. JM', 7 pelo 'HT', 8 pelo 'DR AC' e 5 pelo 'Dr. HL'. A seguir ao nó raiz, a árvore de decisão divide-se em dois nós. No primeiro os casos analisados referem-se ao grupo de clientes cuja residência é 'Alcanena'. No segundo, em 90 casos, os clientes não residem em 'Alcanena', e 85 desses clientes frequentam a loja 'Sede2' e a maior parte desses, 49 casos, são examinados pelo optometrista '1', e os 5 que frequentam uma outra loja, são consultados pelo optometrista 'Dr. JM'. Destes clientes, 68 residem em 'Ourém', 6 na 'Batalha' e 7 em 'Leiria'.
- No modelo **DT6**, dos 101 casos em análise, 61 dos casos são da competência do optometrista '1', 13 pelo 'DE', 8 pelo 'DR AC', 7 casos pela optometrista 'Dr. JM' e 'HT', e 5 pelo 'Dr.HL'. Depois, existe uma divisão, da qual resultam três nós. Quando a loja onde é realizada a consulta é 'Mercado', ocorrem 4 consultas, realizadas pelos optometristas 'Dr. JM', 2 casos, e um caso pelo 'DR AC' e pelo 'HT'. Se a loja é 'Sede2', então ocorrem 81 casos e, destes, 53 são realizados pelo optometrista '1'. Quanto à residência desses clientes, 67 residem em 'Ourém', e o serviço que mais lhes é prestado é do tipo 'Progressiva' e 'Perto'. Relativamente, aos restantes 14 casos, em 7 casos o serviço mais prescrito é 'Progressiva', e desses casos, 3 clientes residem em 'Leiria'.

Tabela 5-2 - Valores de precisão e taxa de erro nos modelos para o optometrista.

Modelo	Precisão	Taxa de Erro
DT4	0,5769	0,4231=42,31%
DT5	0,7576	0,2424=24,24%
DT6	0,6061	0,3939=39,39%

Na Tabela 5-2 apresentam-se os valores de precisão e da taxa de erro dos modelos DT4, DT5 e DT6. Com base nesses valores pode-se concluir que todos os modelos cumprem o critério de sucesso especificado. O modelo DT5 apresenta a menor taxa de erro dos três modelos. Como a sua precisão é superior a 0,7 então tem baixo risco de sofrer alterações quando o conjunto de dados receber informações novas.

Após a análise das diferentes árvores de decisão, verifica-se uma grande prevalência do optometrista '1', retirando-se, basicamente, duas conclusões:

- Quando a loja na qual decorre a consulta é 'Minde', o serviço mais prescrito é 'Longe', e o optometrista que mais prescreve, considerando estes factos, é o '1'.
- Já quando a consulta se efetua na loja 'Sede2', o serviço mais prescrito é do tipo 'Progressiva' ou 'Perto' e é também o optometrista '1' que o presta.

5.2 Modelos C1, C2, C3 e C4

Na secção 4.2, analisou-se cada um dos modelos de *clustering*, compreendendo a forma como foram segmentados os casos presentes no conjunto de dados e avaliando se os mesmos cumprem os critérios de sucesso definidos, anteriormente. Assim, relativamente ao primeiro modelo de *clustering* **C1**, com informação sobre o agente comercial, é possível concluir que:

- No *cluster* 1 constam 27 casos, destacando o domínio do agente comercial 'AM', do tipo de lente 'Unifocal' e as idades dos clientes dividem-se pelas faixas etárias 44-56 e 56-72 anos.
- No *cluster* 2, que conta com 26 casos, existe superioridade do agente comercial 'AM', novamente, sendo o tipo de lente mais vendido é 'Progressiva' e as faixas etárias que mais sobressaem são dos 44 aos 56 anos e dos 56 aos 72 anos.
- No *cluster* 3, que analisa 24 casos, a maioria dos casos é da responsabilidade do agente comercial 'EA', a descrição do tipo de lente é, essencialmente, 'Unifocal' e 'Progressiva', e a idade dos clientes é igual ou superior a 72 anos.

- No *cluster* 4, analisam-se 24 casos, o agente comercial que mais se evidencia é 'AM', a descrição do tipo de lente é 'Unifocal' e a idade dos clientes é igual ou superior a 22 e inferior a 44.
- No *cluster* 5, dos 22 casos analisados, existe uma prevalência do agente comercial 'AM', a descrição do tipo de lente é 'Unifocal' e a idade dos clientes é inferior a 22 anos.

Assim, é possível concluir que este *clustering* obedece aos critérios estabelecidos, visto que existem mais de quatro *clusters* e o número de casos analisados não varia muito de *cluster* para *cluster*. Na Figura 5.1 está apresentado o perfil de cada *cluster* deste agrupamento.

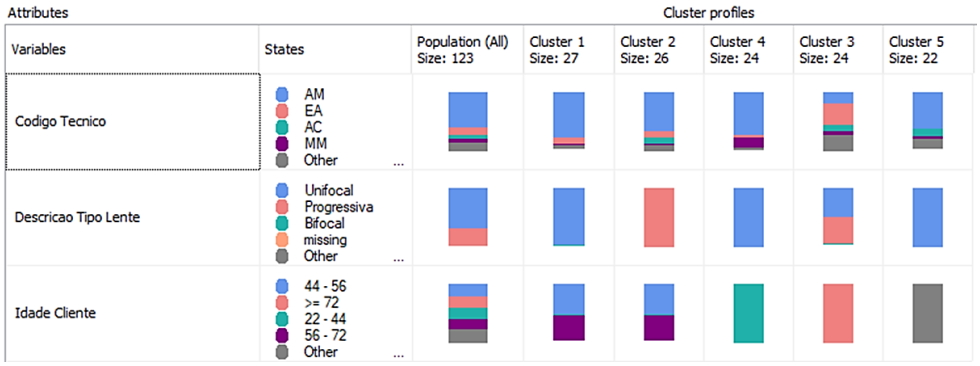


Figura 5.1 - Perfil dos *clusters* do modelo C1.

No que respeita ao modelo **C2**, também este relativo ao agente comercial, o mesmo não obedece aos critérios instituídos, visto que o número de casos nos diferentes *clusters* diferencia-se muito. Inclusive, o *cluster* 1 tem quase metade dos casos analisados neste agrupamento (Figura 5.2).

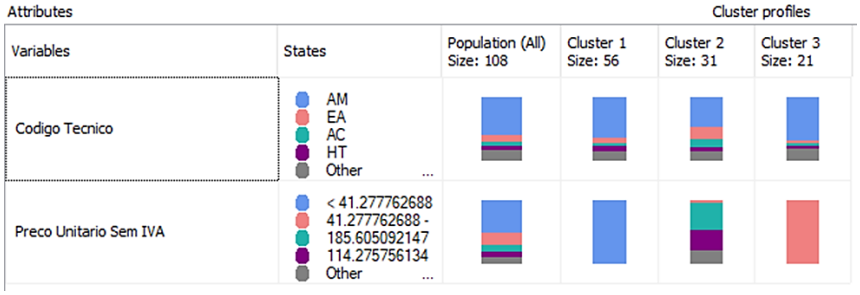


Figura 5.2 - Perfil dos *clusters* do modelo C2.

No entanto é possível concluir, que os três *clusters* diferem no preço unitário, existindo maior prevalência de o preço aplicado ser inferior a 41,28 no *cluster* 1, estar entre 41,28 e 114,28 no *cluster* 3 e os restantes preços no *cluster* 2. Assim, através dos dois agrupamentos elaborados para o agente comercial, pode concluir-se que:

- O agente comercial 'AM' efetua vendas de lentes dos tipos 'Unifocal' e 'Progressiva' a clientes com idade entre os 44 e 56 anos e 56 e 72 anos, e o tipo de lente 'Unifocal' quando a idade do cliente é inferior a 22 anos ou igual ou superior a 22 e inferior a 44 anos.
- Quando a idade do cliente é igual ou superior a 72 anos, o agente comercial 'EA' faz registos de vendas do tipo de lentes 'Unifocal' e 'Progressiva'.
- Geralmente, o preço aplicado as lentes que o agente comercial comercializa é inferior a 41,28 euros e igual ou superior a 41,28 euros e inferior a 114,28 euros.

O terceiro modelo de *clustering*, **C3**, referente ao optometrista, é formado por 4 *clusters*, que se diferenciam, da seguinte forma:

- No *cluster* 1, analisam-se 36 casos, maioritariamente realizados na loja 'Sede2', o optometrista com mais casos é o '1' e o tipo de serviço prescrito é 'Longe'.
- No *cluster* 2, composto por 29 casos, a maioria dos serviços ocorrem na loja 'Sede2', é mais provável que o serviço seja realizado pelo optometrista '1' e os serviços mais prováveis são o tipo 'Progressiva' e o 'Perto'.
- No *cluster* 3, existem 22 casos, que ocorrem, na generalidade, na loja 'Sede2', o serviço é realizado pelo optometrista 'DE', e os serviços mais prováveis é o tipo 'Longe' e 'Progressiva'.
- No *cluster* 4, formado por 21 casos, a consulta ocorre, com maior probabilidade, na loja 'Sede2' e 'Minde'. Quanto ao optometrista destaca-se o 'Dr. JM', com uma diferença mínima de probabilidade. Os serviços que mais se efetuam são o 'Longe' e 'Perto'.

Este agrupamento, cumpre os critérios de sucesso delineados, apesar de se verificar uma discrepância considerável de casos entre o *cluster* 1 e 4, mas não significativa (Figura 5.3).

Figura 5.3 - Perfil dos *clusters* do modelo C3.

Por último, tem-se o modelo **C4**, alusivo ao optometrista, que é formado por 3 *clusters*. Na Figura 5.4 está apresentado o perfil dos *clusters* referenciados.

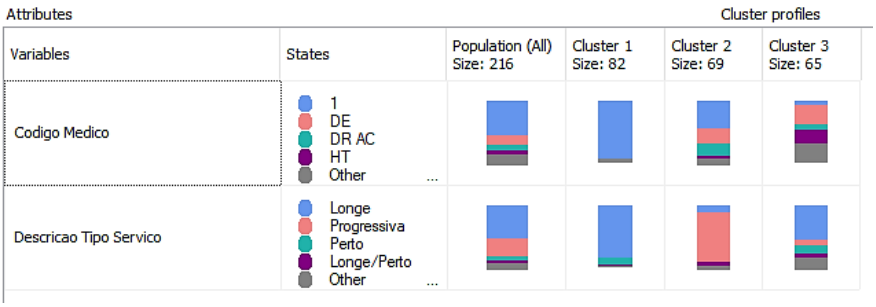


Figura 5.4 – Perfil dos *clusters* do modelo C4.

As caraterísticas de cada *cluster*, diferenciam-se do seguinte modo:

- No *cluster* 1, o optometrista que mais efetua consultas é o '1' e o tipo de serviço mais prescrito é o 'Longe'.
- No *cluster* 2, o optometrista com mais consultas é o '1', seguido pelos optometristas 'DE' e 'DR AC', cujo serviço mais prescrito é 'Progressiva'.
- No *cluster* 3, o optometrista que se destaca com mais consultas é o 'DE' e o tipo de serviço mais prescrito é 'Longe'.

Analisando os agrupamentos referidos acima, é possível retirar algumas conclusões acerca do perfil do optometrista, nomeadamente:

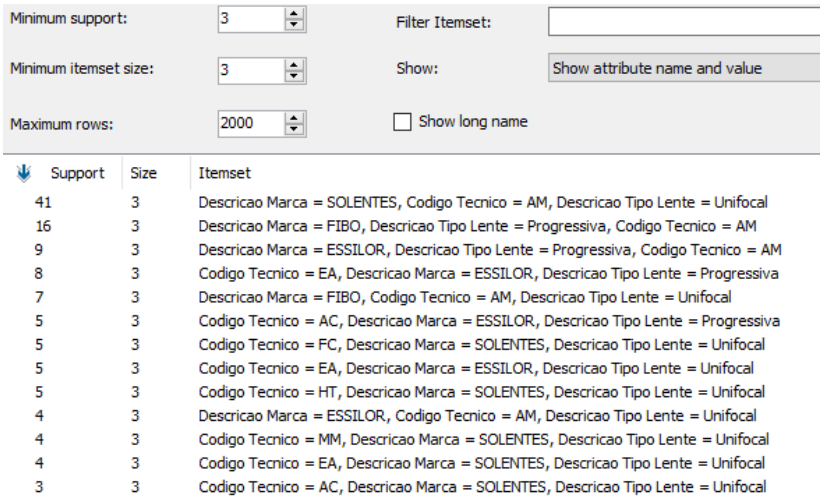
- O optometrista '1' e 'DE' efetuam consultas na loja 'Sede2' de serviços 'Progressiva' e 'Longe'.
- O optometrista 'Dr. JM' realiza consultas na loja 'Sede2' e 'Minde' e prescreve serviços do tipo 'Longe' e 'Perto'.
- Na loja 'Sede2', o optometrista 'DR AC' efetua consultas do tipo 'Longe' e 'Progressiva'.
- O optometrista 'HT' prescreve consultas que correspondem ao serviço 'Longe' na loja 'Minde'.
- Os optometristas '1', 'DE' e 'DR AC' o serviço prescrito é 'Progressiva' e os optometristas 'DE', 'HT' e 'Dr. JM' o serviço é 'Longe'.

5.3 Modelos A1, A2, A3, A4

Por último, realiza-se a análise aos modelos de associação elaborados. Como é referido na secção 5.3, as regras geradas cumprem o critério de avaliação para estes modelos é cumprido à exceção de uma

regra. Resta analisar os conjuntos de itens formados e os valores de suporte que cada conjunto formado apresenta.

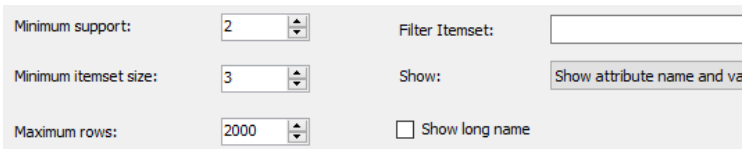
No que diz respeito ao primeiro modelo de associação, **A1**, foram formados treze conjuntos de itens, apresentando todos eles três atributos na sua constituição. O valor de suporte máximo, 41, ocorre quando o agente comercial 'AM' realiza vendas de lentes 'Unifocal' da marca 'SOLENTES' e o mínimo que é 3, acontece quando são vendidas lentes também do tipo 'Unifocal' da mesma marca, mas pelo agente 'AC'. Todos os conjuntos de itens gerados estão apresentados na Figura 5.5.



Support	Size	Itemset
41	3	Descricao Marca = SOLENTES, Codigo Tecnico = AM, Descricao Tipo Lente = Unifocal
16	3	Descricao Marca = FIBO, Descricao Tipo Lente = Progressiva, Codigo Tecnico = AM
9	3	Descricao Marca = ESSILOR, Descricao Tipo Lente = Progressiva, Codigo Tecnico = AM
8	3	Codigo Tecnico = EA, Descricao Marca = ESSILOR, Descricao Tipo Lente = Progressiva
7	3	Descricao Marca = FIBO, Codigo Tecnico = AM, Descricao Tipo Lente = Unifocal
5	3	Codigo Tecnico = AC, Descricao Marca = ESSILOR, Descricao Tipo Lente = Progressiva
5	3	Codigo Tecnico = FC, Descricao Marca = SOLENTES, Descricao Tipo Lente = Unifocal
5	3	Codigo Tecnico = EA, Descricao Marca = ESSILOR, Descricao Tipo Lente = Unifocal
5	3	Codigo Tecnico = HT, Descricao Marca = SOLENTES, Descricao Tipo Lente = Unifocal
4	3	Descricao Marca = ESSILOR, Codigo Tecnico = AM, Descricao Tipo Lente = Unifocal
4	3	Codigo Tecnico = MM, Descricao Marca = SOLENTES, Descricao Tipo Lente = Unifocal
4	3	Codigo Tecnico = EA, Descricao Marca = SOLENTES, Descricao Tipo Lente = Unifocal
3	3	Codigo Tecnico = AC, Descricao Marca = SOLENTES, Descricao Tipo Lente = Unifocal

Figura 5.5 - Conjunto de itens referente ao modelo A1.

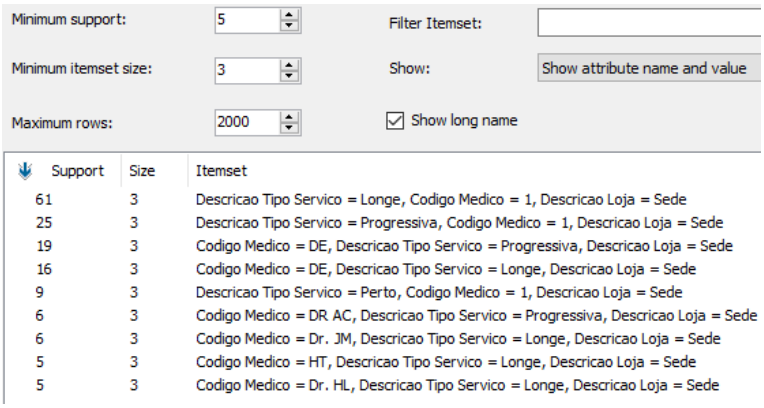
Em relação ao modelo **A2**, o algoritmo formou cinco conjuntos de itens, com três atributos cada um. Os valores de suporte são baixos, sendo o conjunto formado pelo tipo de lente 'Unifocal', com clientes com idade entre os 23 e 44 e o agente comercial 'AM' responsável pela venda o que apresenta maior valor, 3. Estes conjuntos de itens estão expostos na Figura 5.6.



Support	Size	Itemset
3	3	Descricao Tipo Lente = Unifocal, Idade Cliente = 23 - 44, Codigo Tecnico = AM
2	3	Descricao Tipo Lente = Unifocal, Codigo Tecnico = EA, Idade Cliente >= 72
2	3	Descricao Tipo Lente = Unifocal, Idade Cliente >= 72, Codigo Tecnico = AM
2	3	Descricao Tipo Lente = Unifocal, Idade Cliente = 54 - 72, Codigo Tecnico = AM
2	3	Descricao Tipo Lente = Unifocal, Idade Cliente < 23, Codigo Tecnico = AM

Figura 5.6 - Conjunto de itens referentes ao modelo A2.

No modelo **A3**, formaram-se nove conjuntos de itens, cujos valor de suporte variam entre 5 e 61, todos eles compostos por três atributos. O conjunto de itens com maior suporte é aquele em que a loja onde decorre o serviço 'Longe' é a 'Sede' e o optometrista responsável é '1'. Na Figura 5.7 estão apresentados todos os conjuntos de itens.

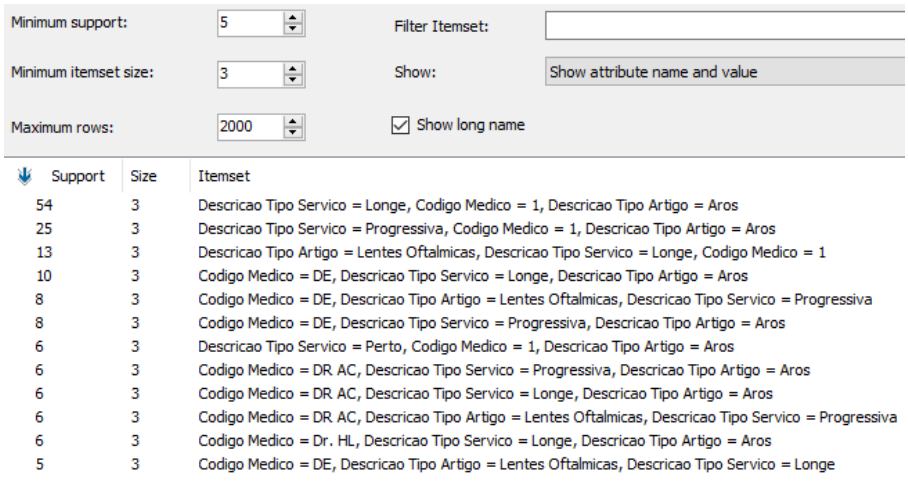


Minimum support: 5 | Filter Itemset:
Minimum itemset size: 3 | Show: Show attribute name and value
Maximum rows: 2000 | Show long name

Support	Size	Itemset
61	3	Descricao Tipo Servico = Longe, Codigo Medico = 1, Descricao Loja = Sede
25	3	Descricao Tipo Servico = Progressiva, Codigo Medico = 1, Descricao Loja = Sede
19	3	Codigo Medico = DE, Descricao Tipo Servico = Progressiva, Descricao Loja = Sede
16	3	Codigo Medico = DE, Descricao Tipo Servico = Longe, Descricao Loja = Sede
9	3	Descricao Tipo Servico = Perto, Codigo Medico = 1, Descricao Loja = Sede
6	3	Codigo Medico = DR AC, Descricao Tipo Servico = Progressiva, Descricao Loja = Sede
6	3	Codigo Medico = Dr. JM, Descricao Tipo Servico = Longe, Descricao Loja = Sede
5	3	Codigo Medico = HT, Descricao Tipo Servico = Longe, Descricao Loja = Sede
5	3	Codigo Medico = Dr. HL, Descricao Tipo Servico = Longe, Descricao Loja = Sede

Figura 5.7 - Conjunto de itens referentes ao modelo A3.

Por fim, tem-se o modelo **A4** formado por 12 conjuntos de itens. Cada um deles apresenta um total de três atributos, com valores de suporte que variam entre 54 e 5, tal como pode ser consultado na Figura 5.8. O conjunto de itens que apresenta maior suporte é aquele em que o optometrista '1' realiza serviços do tipo 'Longe' e o artigo vendido é 'Aros'.



Minimum support: 5 | Filter Itemset:
Minimum itemset size: 3 | Show: Show attribute name and value
Maximum rows: 2000 | Show long name

Support	Size	Itemset
54	3	Descricao Tipo Servico = Longe, Codigo Medico = 1, Descricao Tipo Artigo = Aros
25	3	Descricao Tipo Servico = Progressiva, Codigo Medico = 1, Descricao Tipo Artigo = Aros
13	3	Descricao Tipo Artigo = Lentes Oftalmicas, Descricao Tipo Servico = Longe, Codigo Medico = 1
10	3	Codigo Medico = DE, Descricao Tipo Servico = Longe, Descricao Tipo Artigo = Aros
8	3	Codigo Medico = DE, Descricao Tipo Artigo = Lentes Oftalmicas, Descricao Tipo Servico = Progressiva
8	3	Codigo Medico = DE, Descricao Tipo Servico = Progressiva, Descricao Tipo Artigo = Aros
6	3	Descricao Tipo Servico = Perto, Codigo Medico = 1, Descricao Tipo Artigo = Aros
6	3	Codigo Medico = DR AC, Descricao Tipo Servico = Progressiva, Descricao Tipo Artigo = Aros
6	3	Codigo Medico = DR AC, Descricao Tipo Servico = Longe, Descricao Tipo Artigo = Aros
6	3	Codigo Medico = DR AC, Descricao Tipo Artigo = Lentes Oftalmicas, Descricao Tipo Servico = Progressiva
6	3	Codigo Medico = Dr. HL, Descricao Tipo Servico = Longe, Descricao Tipo Artigo = Aros
5	3	Codigo Medico = DE, Descricao Tipo Artigo = Lentes Oftalmicas, Descricao Tipo Servico = Longe

Figura 5.8 - Conjunto de itens referentes ao modelo A4.

Conclui-se que os modelos de associação construídos, apesar de cumprirem o critério de sucesso definido, não são modelos satisfatórios. Isto porque não se retiraram associações entre valores de atributos suficientemente fortes ou válidas, existindo mesmo algumas associações que são duvidosas. Este facto pode dever-se, em grande parte, ao conjunto de dados utilizado, isto é, o número de casos apresentado é bastante baixo, o que não permite efetuar uma análise com valores de confiança e suporte significativos para retirar boas conclusões

CAPÍTULO 6

6. CONCLUSÕES E TRABALHO FUTURO

6.1 Conclusões

Atualmente, o avanço tecnológico proporciona às empresas grande capacidade de armazenar todos os tipos de dados existentes. Com estas facilidades, o armazenamento e análise dos dados começa a ser uma atividade frequentemente utilizada pelas empresas, sendo divulgada com a ideia de que esta atividade de obtenção de informação que se transforma em conhecimento, aspetos estes que, quando bem aproveitados, têm importância para qualquer empresa que ambicione obter elevados índices de competitividade, independentemente do setor de mercado no qual está inserida. No entanto, este aumento do volume de dados, cria um problema: como é possível extrair conhecimento de grandes bases de dados? De modo a solucionar este problema, torna-se essencial a aplicação de técnicas que possibilitem a extração de conhecimento. Assim, a solução encontrada passa pela implementação de técnicas de mineração de dados capazes de extrair o conhecimento contido nos dados das grandes bases de dados.

Ao longo da presente dissertação, teve-se como ponto principal a construção de perfis de agentes comerciais e do optometrista através da aplicação de técnicas de mineração de dados, com o objetivo de perceber qual a venda ou a prescrição que fazem, perante as diferentes situações, como por exemplo, para o agente comercial pretende-se conhecer o tipo de lente que vende a um cliente de determinada faixa etária, qual a marca mais provável para um certo tipo de lente, e qual a gama de preços dos diferentes tipos de lentes, enquanto que para o optometrista é desejável saber o tipo de consulta que mais presta numa certa loja, se o tipo de artigo vendido condiz com o tipo de serviço ocorrido na consulta.

Com o objetivo de dar resposta aos diferentes cenários colocados, aplicaram-se três diferentes técnicas de mineração de dados, a classificação, a segmentação e a associação.

Na primeira técnica aplicada, a classificação, empregou-se o algoritmo *decision trees* da *Microsoft*. As árvores de decisão são modelos estatísticos utilizados em problemas em que um conjunto de dados é utilizado para prever o valor de um atributo de saída (ou previsível). Esse conjunto de dados contém valores de entrada e resultados. Aplicando esta técnica, construíram-se seis modelos de classificação: três modelos para o agente comercial e três modelos para o optometrista. No caso do agente comercial, conclui-se que: o agente comercial EA, perante um cliente com idade igual ou superior a 72 anos, vende lentes progressivas, ocorrendo à mesma decisão em casos do agente comercial HT, o agente comercial AM quando receciona o pedido de um cliente com idade entre os 56, inclusive, e os 72 anos, tende a vender lentes progressivas, o que também acontece no caso do agente comercial AC. É possível concluir também que os agentes comerciais AM e MM praticam preços mais baixos quando comparados com os agentes EA, AC e HT. Para o perfil do optometrista, observa-se que o optometrista '1' é responsável por casos que ocorrem na loja 'Minde' se este for do tipo 'Progressiva', e se a loja não é 'Minde' o optometrista '1' realiza serviços do tipo 'Perto', 'Longe' e 'Progressiva', e também foi possível aferir que o optometrista '1' realiza consultas na loja 'Sede2' cujo tipo de serviço efetuado é 'Progressiva' e 'Perto' e a maioria dos seus clientes são de 'Ourém'. O optometrista 'Dr. HL' presta serviços do tipo 'Perto' na loja 'Minde' ao par que o optometrista 'HT' prescreve serviços do tipo 'Longe/Perto'. Os optometristas 'Dr. JM', 'HT' e 'DR AC' realizam serviços na loja 'Mercado'.

A segunda técnica de mineração aplicada foi a segmentação, através da aplicação do algoritmo *clustering* da Microsoft, no qual se tem de especificar qual o método de segmentação selecionado: o *k-means* e o *expectation-maximization*. O primeiro separa os objetos tendo em conta a distância do objeto do conjunto ao seu centroide. O segundo faz a separação dos objetos tendo em conta a probabilidade que estes apresentam de pertencer a um dado *cluster*. Foram elaborados quatro modelos, dois para o agente comercial e dois para o optometrista. A partir dos seus resultados retiraram-se algumas conclusões. Quando o agrupamento é feito sobre o agente comercial, o agente 'AM', em clientes com faixa etária entre os 44 e 72 anos vende lentes do tipo 'Progressiva' ou 'Unifocal', o agente comercial 'EA' em cliente com idade igual ou superior a 72 anos, vende o tipo de lente 'Progressiva' ou 'Unifocal', os agentes comerciais 'AM', 'AC', 'FC' e 'HT' vendem lentes do tipo 'Unifocal', se a idade é inferior a 22 anos, e se está compreendida entre os 22 e os 44 anos o agente 'AM' e 'MM' tomam a mesma decisão, conclui-se também que os agentes 'AM',' HT' e 'MM' praticam, com maior probabilidade preços, mais

acessíveis, isto é, inferiores a 114,28 euros e os agentes 'EA', 'AC' e 'HT' praticam preços mais altos, superiores a 114,28 euros. Relativamente ao optometrista, conclui-se que o optometrista '1' presta serviço do tipo 'Longe' na loja 'Sede'2, os optometrista '1' e 'Dr. HL' na loja 'Sede2' e 'Minde' realizam serviços do tipo 'Progressiva' e 'Perto', os optometristas 'DE', 'Dr. JM', 'DR AC' e 'HT' prestam consultas na loja 'Sede2' do tipo 'Longe', 'Progressiva' e 'Perto', e o optometrista 'Dr. JM' nas lojas 'Sede2' e 'Minde' presta consultas relativas ao tipo 'Longe' e 'Perto', conclui-se também que a nível do tipo de serviço prestado, independentemente da loja onde é realizado o mesmo, o optometrista 'DE' prescreve três serviços, 'Progressiva', 'Longe' ou 'Perto', o optometrista '1' prestas consultas do tipo 'Longe' ou 'Progressiva', e os optometristas 'Dr. HL', 'Dr. JM' e 'DE' do tipo 'Longe' ou 'Perto'.

A última técnica de mineração aplicada aos dados foi a associação através do algoritmo *Association Rules* da Microsoft. As regras de associação são padrões descritivos que representam a probabilidade de um dado conjunto de itens constar numa transação tendo em conta que outro conjunto de itens está presente, isto é, associa valores de atributos que estão presentes numa mesma transação. Como foi concluído na secção 6.3, apesar de obedeceram ao critério de sucesso definido, nestes modelos a regra gerada não permitiram tirar boas conclusões para este caso de estudo.

6.2 Trabalho Futuro

Após o término deste trabalho de dissertação, verificou-se que algumas das coisas que foram realizadas poderiam ser concretizados de uma outra forma, de modo a obter, possivelmente, melhores resultados. Entre estes, um dos aspetos que deve ser melhorado no futuro é a quantidade de dados disponíveis. Como se constatou ao longo da dissertação, foi necessário proceder à limpeza de dados, facto que levou a uma redução significativa do conjunto de dados.

No futuro, seria interessente criar um sistema iterativo através do qual o utilizador determina o tipo de análise que deseja, definindo quais os atributos a considerar e obtém os resultados para essa análise. Sendo o mercado ótico direcionado ao cliente e objetivando um aumento do número de clientes, a criação de um sistema de recomendação tendo em conta os gostos do cliente, que poderiam ser analisados no sistema iterativo referido, seria também um bom trabalho a desenvolver num futuro próximo. A nível das lojas, seria interessante prever o stock necessário para satisfazer a procura através da inclusão de técnicas direcionadas para a previsão no sistema a desenvolver.

Este trabalho deve ser a base de novos projetos, devendo ser atualizado constantemente, visto que com a entrada de novas informações na base de dados, é possível que ocorra mudanças nos modelos

elaborados, e se este não sofrer atualizações, em consequência, estes modelos são atuais no presente, mas com grande probabilidade, de serem obsoletos num futuro próximo.

REFERÊNCIAS BIBLIOGRÁFICAS

Adomavicius, G., Tuzhilin, A. 2001. *Using data mining methods to build customer profiles. Computer,* Vol.34, pp. 74-81. Disponível em: https://pdfs.semanticscholar.org/b298/e06b9ee4b3056c68a023035f228527a891a2.pdf [Acedido em 22 de janeiro de 2019]

Agrawal, R. & Srikant, R., 1994. *Fast Algorithms for Mining Association Rules.* Disponível em: http://www.cse.msu.edu/~cse960/Papers/MiningAssoc-AgrawalAS-VLDB94.pdf [Acedido em 27 de maio de 2019]

APLO, 2006. Que faz um Optometrista. Disponível em: https://www.aplo.pt/SobreaOptometria/QuefazumOptometrista.aspx [Acedido em 28 de maio de 2019]

Azevedo, P.J. & Jorge, A.M., 2007. *Comparing Rule Measures for Predictive Association Rules. In Proceedings of the 18 th European Conference on Machine Learning*. ECML '07. Berlin, Heidelberg: Springer-Verlag, pp. 510–517. Disponível em: https://www.researchgate.net/publication/221112649_Comparing_Rule_Measures_for_Predictive_Association_Rules [Acedido em 17 de junho de 2019]

Barrett, R., Maglio, P. & Kellem, D.C., 1997. *How to Personalize the Web. In Proceedings of the SIGCHI conference on Human factors in computing systems,* Atlanta, March 22-27, pp. 75-82.

Bradley, K., Rafter, R. & Smyth, B., 2000. *Case-based user profiling for content personalisation. In Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer Berlin Heidelberg, pp. 62-72.

Brusilovsky, P. & Millán, E., 2007. *User Models for Adaptive Hypermedia and Adaptive Educational Systems.* Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.5703&rep=rep1&type=pdf [Acedido em 31 de maio de 2019]

Camilo, C. & Silva, J., 2009. *Mineração de Dados: Conceitos, tarefas, métodos e ferramentas*. Disponível em: http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf [Acedido em 9 de maio de 2019]

Caruso, M., Mecella, M., Baldoni, R. & Querzoni, L., 2013. *User profiling and micro-accounting for smart energy management. In Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems,* Rome, Italy, ACM, p. 42.

Carvalho, H., 2014. *Aprendizado de Máquina voltado para Mineração de Dados: Árvores de Decisão,* Bachelor's Thesis, Universidade de Brasília, Brasília. Disponível em: https://fga.unb.br/articles/0000/5556/TCC_Hialo_Muniz.pdf [Acedido em 31 de outubro de 2018]

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., 2000. *CRISP-DM 1.0 Step-by-step data mining guide*. Disponível em: https://www.the-modeling-agency.com/crisp-dm.pdf [Acedido em 22 de fevereiro de 2019]

Dey, A. & Abwod, G., 2000. T*owards a better understanding of context and context-awareness. In Proceedings of the PrCHI 2000 Workshop on the What, Who, Where, When and How of Context-Awareness*. Disponível em: https://www.researchgate.net/publication/274074382_Towards_a_Better_Understanding_of_Context_and_Context-Awareness [Acedido em 3 de junho de 2019]

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. *Knowledge Discovery and Data Mining: Towards a Unifying Framework.* Disponível em: https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf [Acedido em 24 de maio de 2019]

Fleuren, M. 2012, *User Profiling Techniques: A comparative study in the context of e-commerce websites*, *Bachelor's Thesis*, Utrecht University, Utrecht. Disponível em: http://igitur-archive.library.uu.nl/student-theses/2012-0801-200525/UUindex.html [Acedido em 4 de marlo de 2019]

Gauch, S., Speretta, M., Chandramouli, A. & Micarelli, A., 2007. *User profiles for personalized information access. The adaptive web*, Vol. 4321, pp. 54-89. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-72079-9_2 [Acedido em 29 de janeiro de 2019]

Goldberg, L.R., 1993. *The structure of phenotypic personality traits. American Psychologist*, Vol. 48, pp. 26–34. Disponível em: http://psych.colorado.edu/~carey/courses/psyc5112/readings/psnstructure_goldberg.pdf [Acedido em 3 de junho de 2019]

Gonçalvez, E., 2005. *Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. INFOCOMP–Journal of Computer Science.* Disponível em: https://www.researchgate.net/publication/301504294%0ARegras [Acedido em 6 de maio de 2019]

Gupta, S., Kumar, D. & Sharma, A., 2011. *DATA MINING CLASSIFICATION TECHNIQUES APPLIED FOR BREAST CANCER DIAGNOSIS AND PROGNOSIS. Indian Journal of Computer Science and Engineering (IJCSE),* Vol.2, pp. 188-195. Disponível em: http://www.ijcse.com/docs/IJCSE11-02-02-53.pdf [Acedido em 27 de maio de 2019]

Han, J., Kamber, M. & Pei, J., 2011. *Data Mining. Concepts and Techniques* [Online]. Waltham: Morgan Kaufmann Disponível em: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf [Acedido em 18 de fevereiro de 2018]

Horvitz, E., Breese, J., Heckerman, D., Hovel, D. & Rommelse, K., 1998. *The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, UAI*, pp. 256–265.

Kanoje, S., Girase, S. & Mukhopadhyay, D., 2014. *User Profiling Trends, Techniques and Applications. International Journal of Advance Foundation and Research in Computer (IJAFRC),* Vol. 1, pp. 119-124. Disponível em: https://arxiv.org/ftp/arxiv/papers/1503/1503.07474.pdf [Acedido em 4 de março de 2019]

Kelly, D. & Teevan, J., 2003. *Implicit feedback for inferring user preference: a bibliography. ACM SIGIR Forum*, Vol. 37, pp. 18-28.

Landeiro, V., 2011. Introdução ao uso do programa R. Disponível em: https://cran.r-project.org/doc/contrib/Landeiro-Introducao.pdf [Acedido em abril de 2019]

Larose, D. & Larose, C. 2014. *DISCOVERING KNOWLEDGE IN DATA* [Online]. New Jersey: John Wiley & Sons, Inc. Disponível em: https://doc.lagout.org [Acedido em 28 de maio de 2019]

Maes, P., 1994. *Agents that reduce work and information overload.* Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.368.2096&rep=rep1&type=pdf [Acedido em 31 de maio de 2019]

Marques, R., 2013. *ANÁLISE AO MERCADO DAS ÓPTICAS EM PORTUGAL*, Meios & Publicidade. Disponível em: http://www.meiosepublicidade.pt/2013/01/analise-ao-mercado-das-opticas-em-portugal/ [Acedido em 27 de maio de 2019]

Microsoft 1, 2011. *Microsoft Decision Trees Algorithm Technical Reference.* Disponível em: https://docs.microsoft.com/en-us/previous-versions/sql/sql-server-2008-r2/cc645868%28v%3dsql.105%29 [Acedido em março de 2019]

Microsoft 2, 2011. *Microsoft Clustering Algorithm Technical Reference.* Disponível em: https://docs.microsoft.com/en-us/previous-versions/sql/sql-server-2008-r2/cc280445%28v%3dsql.105%29 [Acedido em abril de 2019]

Microsoft 3, 2018. *Microsoft Association Algorithm Technical Reference.* Disponível em: https://docs.microsoft.com/pt-br/sql/analysis-services/data-mining/microsoft-association-algorithm-technical-reference?view=sql-server-2017 [Acedido em abril de 2019]

Middleton, S., Shadbolt, N. & Roure, D., 2004. *Ontological user profiling in recommender systems. ACM Transactions on Information Systems (TOIS),* Vol. 22, pp. 54-88. Disponível em: https://eprints.soton.ac.uk/258926/1/tois2004.pdf [Acedido em 30 de maio de 2019]

Ouanaim, M., Harroud, H., Berrado, A. & Boulmalf, M., 2010. *Dynamic user profiling approach for services discovery in mobile environments. In Proceedings of the 6th IWC MC Conference,* New York, USA, ACM, pp. 550-554.

Pacheco, R., Fatima, T. & Tait, T., 2000. *Tecnologia de Informação: evolução e aplicações. Teoria e Evidência Económica*, Vol. 8, pp. 97-113. Disponível em: https://www.researchgate.net/publication/228462814_Tecnologia_de_Informacao_evolucao_e_aplicacoes [Acedido em 27 de maio de 2019]

Papazoglou, M., 2001. *Agent-oriented technology in support of e-business. Communications of the ACM*, Vol. 44, pp. 71-77. Disponível em: https://www.researchgate.net/publication/27290027_Agent-Oriented_Technology_in_Support_of_E-_Business [Acedido em 3 de junho de 2019]

Poo, D., Chng, B. & Goh, J., 2003. *A hybrid approach for user profiling. In System Sciences, In Proceedings of the 36th Annual Hawaii International Conference on*, IEEE, pp. 9-13. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.96.7457&rep=rep1&type=pdf [Acedido em 31 de maio de 2019]

Schiaffino, S. & Amandi, A., 2009. *Intelligent user profiling. Artificial Intelligence An International Perspective*, pp. 193–216. Disponível em: https://link.springer.com/chapter/10.1007/978-3-642-03226-4_11 [Acedido em 29 de janeiro de 2019]

Shearer, C., 2000. *The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing*, Vol.5, pp. 13-21. Disponível em: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf [Acedido em 21 de fevereiro]

Shortliffe, E.H. & Buchanan, B.G., 1990. *A model of inexact reasoning in medicine. In G. Shafer & J. Pearl, eds. Readings in uncertain reasoning.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 259–275. Disponível em: https://dl.acm.org/citation.cfm?id=85330 [Acedido em 17 de junho]

Singh, H., 2012. *Implementation Benefit to Business Intelligence using Data Mining Techniques. International Journal of Computing & Business Research.* Disponível em: http://www.researchmanuscripts.com/isociety2012/63.pdf [Acedido em 18 de fevereiro de 2019]

Srikant, R., Agrawal, R., & Vu, Q., 1997. *Mining Association Rules with Item Constraints. In Proceeding KDD'97 Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pp. 67-73. Disponível em: www.aaai.org [Acedido em 15 de fevereiro de 2019]

Tang, J., Yao, L., Zhang, D. & Zhang, J., 2010. *A Combination Approach to Web User Profiling. ACM Transactions on Information Systems (TOIS)*, Vol. 5, pp. 1-38. Disponível em: http://keg.cs.tsinghua.edu.cn/jietang/publications/TKDD11-Tang-et-al-web-user-profiling.pdf [Acedido em 30 de maio de 2019]

Trajkova, J. & Gauch, S., 2004 *Improving Ontology-Based User Profiles. In Proceedings of RIAO 2004*, University of Avignon (Vaucluse), France, April 26-28, pp. 380-389.

Upadhyay, T., Vidhani, A. & Dadhich, V. 2016, *Customer Profiling and Segmentation using Data Mining Techniques. IJCSC*, Vol.7, pp. 65-67. Disponível em: http://csjournals.com/IJCSC/PDF7-2/10. Tejpal.pdf [Acedido em 23 de janeiro de 2019]

Wærn, A., 2004. *User Involvement in Automatic Filtering: An Experimental Study. User Modeling and User-Adaptive Interaction*, Vol. 14, pp. 201-237. Disponível em: https://www.researchgate.net/publication/220116278_User_Involvement_in_Automatic_Filtering_An_Experimental_Study [Acedido em 4 de junho de 2019]

Wiedmann, K. P., Buxel, H., Walsh, G. 2002, *Customer profiling in e-commerce: Methodological aspects and challenges. Journal of Database Marketing & Customer Strategy Management*, Vol.9, pp. 170-184. Disponível em: http://link.springer.com/10.1057/palgrave.jdm.3240073 [Acedido em 18 de janeiro de 2019]

Yu, Z., Zhou, X., Hao, Y., Gu, J., 2006. *TV Program Recommendation for Multiple Viewers Based on user Profile Merging. User Modeling and User-Adapted Interaction*, Vol. 16, pp. 63– 82. Disponível em: http://www.ccm.media.kyoto-u.ac.jp/~yu/UMUAI_Zhiwen%20Yu.pdf [Acedido em 3 de junho de 2019]

Zaki, M.J., Parthasarathy, S., Li, W. & Ogihara, M., 1997. *Evaluation of Sampling for Data Mining of Association Rules. In Proceedings of the IEEE International Workshop on Research Issues in Data Engineering*, RIDE'97 - Birmingham, UK, IEEE, pp. 42-50. Disponível em: http://www.cs.rpi.edu/~zaki/PaperDir/RIDE97.pdf [Acedido em 6 de maio de 2019]

Zuber, M., Suman, N., Pasha, M.G. & Adam, M., 2013. *A STUDY ON DATA MINING APPROACHES. International Journal of Emerging Trends in Engineering and Development*, Vol. 1, pp. 676-683. Disponível em: https://rspublication.com/ijeted/jan13/68.pdf [Acedido em 18 de fevereiro de 2019]