UNIVERSIDADE CATÓLICA PORTUGUESA

Aplicação de Marketing Analítico

Farmácia Gaia Jardim

Trabalho Final na modalidade de Projecto
apresentado à Universidade Católica Portuguesa
para obtenção do grau de mestre em Gestão com especialização em Serviços

por

Ana Rita Miguel Saraiva

sob orientação de
Professora Auxiliar Convidada Vera Miguéis

Católica Porto Business School
Fevereiro de 2018

Agradecimentos

Este trabalho não teria sido possível sem a ajuda incansável de algumas pessoas.

Em primeiro lugar, gostaria de agradecer à Professora Vera Miguéis por toda a paciência para todas as minhas dúvidas ao longo de todos estes meses e à Professora Conceição Portela pela oportunidade de realizar este projecto que me ensinou tanto.

Aos colaboradores da Farmácia Gaia Jardim, em especial à Dr. Luísa Venda, à Susana e ao Rui por toda a disponibilidade para me fornecerem os dados necessários.

Agradeço a todos os meus amigos por todas as vezes que me motivaram e incentivaram a continuar este projecto quando a vontade diminuía, em especial à Joana e ao António por sempre estarem presentes.

Por fim, queria agradecer aos meus pais, por todas as ínumeras oportunidades que sempre me deram e pela presença constante. A paciência e o carinho deles foram essenciais para a conclusão deste trabalho.

Resumo

O projecto apresentado surgiu no âmbito de uma proposta de investigação da Católica Porto Business School. Cada vez mais, os sistemas de informação fazem parte do desenvolvimento das empresas e a Farmácia Gaia Jardim procurou evoluir nesse sentido e utilizar a sua base de dados para desenvolver a sua relação com os clientes. Assim, neste trabalho pretendeu-se estabelecer estratégias de *cross-selling* através do uso de ferramentas de marketing analítico, nomeadamente *data mining*. De um modo geral, esta análise aos dados da farmácia permitiu um maior conhecimento sobre os seus clientes. Mais concretamente, este trabalho permitiu a identificação dos segmentos de clientes com base na frequência de visitas à fármacia e o valor médio gasto e a criação de regras de associação entre produtos para estabelecer as estratégias de *cross-selling* mais adequadas para cada segemento de clientes da farmácia.

Palavras-chave: *cross-selling*, marketing analítico, *customer relationship management*, *clustering*

Abstract

The presented project is developed in the scope of a research proposal from Católica Porto Business School. Information systems are becoming of increased importance to the development of companies. Gaia Jardim Pharmacy is taking this fact into consideration and using its data base to develop their relationship with their customers. Therefore, in this paper we aimed at creating cross-selling strategies through of the use of analytic marketing tools, as data mining. In general, this analysis enabled to obtain a more detailed knowledge about the customers of the pharmacy. uMore specifically, this study enabled the identification of segments of customers and finally on the identification of association rules between the products which allows the pharmacy to definethe most adequate cross-selling strategies for the segments of customers identified.

Keywords: cross-selling, analytical marketing, customer relationship management

Índice

Índice de Figuras

Índice de Tabelas

Capítulo 1
1. Introdução

Actualmente, é impossível não observar a importância evidente das tecnologias em qualquer actividade do quotidiano. Cada vez mais, os sistemas de informação têm um papel relevante e imprescindível no desenvolvimento das empresas e das organizações. Foi neste sentido que surgiu este trabalho final de semestre.

A Farmácia Gaia Jardim foi criada em Setembro de 2014 e localiza-se no concelho de Vila Nova de Gaia, distrito do Porto, com mais de 300.000 habitantes. A referida farmácia encontra-se dentro do Centro Comercial Gaia Jardim sendo a entrada de acesso à farmácia feita pelo exterior. Tem acesso rápido e um parque com 300 lugares de estacionamento. Esta farmácia é considerada a maior de Portugal com 800m^2 e situa-se perto do Hospital da Arrábida, Hospital Privado de Gaia, Arrábida Shopping, Conforama, escolas e colégios e da loja do cidadão. Esta localização permite que haja uma grande diversidade de clientes e de procura de produtos, uma vez que as farmácias continuam a ser a única forma de aviar prescrições médicas.

Tendo em consideração que esta é uma farmácia relativamente recente, a direcção desta organização percebeu que poderia desenvolver a sua relação com os actuais clientes. Assim, foi implementado um cartão de fidelização em dezembro de 2016 de forma a fidelizar clientes e a optimizar a sua rentabilidade. Actualmente, o cartão permite a acumulação de pontos aquando das vendas aos clientes, permite suportar a definição de descontos aos seus clientes e, esporadicamente, permite a utilização da informação obtida sobre os seus clientes para promoções muito específicas. Este último processo é sempre desenvolvido de uma forma manual e trabalhosa. Este projecto permitirá à

farmácia estabelecer estratégias de *cross-selling* na farmácia através dos dados obtidos a partir do referido cartão. É então, neste contexto que este trabalho ganha relevância, na medida em que, auxilia a farmácia para que esta consiga optimizar o seu investimento no cartão de fidelização, e no respectivo programa informático. Saber quem são os consumidores, com que frequência vão à farmácia, o que compram, quando compram e a evolução das compras ao longo do tempo para que se direccione campanhas especificas são alguns dos exemplos de informações que podem ser extraídas do cartão de fidelização e são aspetos que foram visados ao longo da realização deste projecto.

De forma a desenvolver os modelos pretendidos, foi necessária a extracção dos dados referentes às vendas do *data warehouse* da farmácia para serem posteriormente analisados. Assim sendo, a análise foi feita com dados recolhidos durante os meses de Janeiro, Fevereiro e Março de 2017 recolhidos a partir do cartão de fidelização. O programa informático tinha até Março de 2017 cerca de 10.000 utilizadores registados. Numa primeira fase, considerou-se importante fazer uma descrição dos clientes que frequentam a farmácia. Dos 10.000 só cerca de 4.000 é que têm a informação socio-demográfica deviamente preenchida, sendo que foi essa a amostra considerada para descrever os clientes da Farmácia Gaia Jardim. Para garantir a confidencialidade dos dados, os dados obtidos para a contextualização dos clientes da farmácia não continham nomes, mas sim números identificativos de cada cartão. Para além disso os dados continham a localidade, a idade, o género e o total comprado nos três meses referidos. Posteriormente, obteve-se os dados referentes às transacções efectuadas, isto é, os produtos que foram comprados, quem é que os comprou e quando é que foram comprados.

Para atingir o objetivo do trabalho, este será suportado por ferramentas de *data mining* que permitem automatizar os processos e adquirir conhecimentos sobre

os clientes e as suas características comportamentais que não eram conhecidos até então.

Importa referir que a presente investigação se enquadra na modalidade de projecto que surgiu no âmbito da proposta de investigação pelo *Service Management Lab* da Católica Porto Business School da Universidade Católica Portuguesa.

Capítulo 2
2. Marketing

Este capítulo tem como principal objectivo apresentar os fundamentos e os conceitos teóricos de marketing referentes ao trabalho desenvolvido, para uma melhor compreensão do mesmo. Este capítulo está dividido em quatro partes principais: a primeira é referente ao marketing relacional, a segunda ao *customer relationship management* (CRM), a terceira à segmentação de clientes das empresas e a última refere-se a *cross-selling*.

2.1 Marketing Relacional

O conceito inicial de marketing consistia no processo através do qual, os bens e os serviços circulavam dos fornecedores para os consumidores. Este emergiu da transacção de uma sociedade baseada na subsistência, em que famílias produziam os seus próprios bens, para formas de civilização mais especializadas. Por exemplo, o simples acto de trocar uma ferramenta por um cereal implicava aspectos de marketing. O termo marketing deriva da palavra mercado, que significa um grupo de vendedores e compradores que cooperam na troca de bens e serviços.

Alguns investigadores dividiram a história do marketing em quatro eras distintas, correspondendo a diferentes práticas e focos (Hollander et al., 2005; Boone and Kurtz, 2008): a era da produção, a era das vendas, a era do marketing e a era da relação.

A era da produção é anterior à Segunda Guerra Mundial. Nessa altura, o maior objectivo do marketing era fabricar um produto satisfatório, sem grandes

esforços e apresentá-lo aos potenciais clientes, através de catálogos e publicidade. A era das vendas decorreu entre os anos 30 e os 50 e promoveu o conceito de marketing transaccional. Devido ao excesso de oferta sobre a procura, as empresas reconheceram a necessidade de existirem vendedores para vender os produtos. O objectivo do marketing tornou-se no desenvolvimento de argumentos persuasivos para incentivar os clientes a comprar os produtos. Na década de 50 e até à década de 60, este contexto evoluiu para a era do marketing, quando as empresas começaram a adoptar uma orientação para o cliente e tomaram consciência da importância de seguir as preferências e motivações do consumidor. Finalmente, nos anos 80 começou a era da relação em que o conceito de marketing se tornou como ele é conhecido hoje em dia. Este foi um período tecnológico e científico que resultou em produções em escala. Esse facto juntamente com o desenvolvimento dos sistemas de transporte e dos meios de comunicação, nomeadamente do rádio, promoveu a gestão das vendas de bens e criou uma separação entre as empresas e os seus clientes, já que já não era mais viável para as empresas personalizarem os seus produtos. Assim, as empresas já não eram capazes de conhecer pessoalmente os seus clientes e já não havia interacção entre eles. O conceito de marketing relacional, proposto por Berry (1983) surgiu como uma tentativa de minimizar esse *gap* entre as empresas e os seus clientes. O marketing relacional não é focado em transacções simples, mas sim, em manter os clientes e construir relações mais longas e mais complexas com eles. O conceito de Berry de marketing relacional assemelha-se às ideias de outros investigadores de marketing de serviços, como Levitt (1981), Gronroos (1982) e Gummesson (1987).

Actualmente, existem várias definições do conceito de marketing relacional. Kotler (1967) define o conceito como a análise, organização, planeamento e controlo dos recursos, políticas e actividades de uma empresa orientados para o cliente, tentando satisfazer as necessidades e desejos de um certo mercado *target*

de uma forma lucrativa. De acordo com Grönroos (1990), o marketing relacional consiste em identificar, estabelecer, manter e melhorar as relações de longo-prazo com os clientes e as outras partes interessadas com lucro, para que os objectivos das partes envolvidas sejam atingidos. A forma de isto ser feito é através da troca mútua e do cumprimento de promessas. Gronroos (1990) afirma também que o marketing relacional deve considerar a extinção das relações com os clientes quando é conveniente. Coviello et al. (1997) define o marketing relacional como uma actividade integrativa que envolve funções de toda a organização, com ênfase em facilitar, construir e manter relações ao longo do tempo (ver Lamberti & Noci (2010) para mais informações). Peppers and Rogers (2011) defendem que pode ser 10 vezes mais caro ganhar um novo cliente do que manter um cliente já existente, e que, o custo de arranjar um novo cliente até ao nível de rentabilidade que o perdido tem, é cerca de 16 vezes superior, daí a relevância do marketing relacional.

2.2 Customer Relationship Management

Customer relationship management (CRM) é baseado no marketing relacional e é focado na tecnologia subjacente para a gestão dos clientes. O CRM teve origem no desejo de combinar o balcão de ajuda (*help desk*), o apoio ao cliente, o *Enterprise Resource Planning* (ERP) e o *data mining* (Peel, 2002).

Payne and Frow (2005) definem CRM como uma abordagem estratégica que consiste em criar valor para os *shareholders* através do desenvolvimento de relações apropriadas com clientes e segmentos-chave. O CRM unifica o potencial das estratégias de marketing relacional e das tecnologias de informação para criar relações rentáveis e longas com os clientes e *stakeholders*. O CRM fornece oportunidades aprimoradas para usar dados e informações para compreender os

clientes e para criar valor para eles. Para isto, é necessária uma integração multifuncional de processos, pessoas, operações e conhecimentos de marketing.

As primeiras iniciativas do CRM foram lançadas no início dos anos 90 e foram maioritariamente focadas nas actividades de *call center* (Roya Rahimi, 2007). O promissor aparecimento do CRM foi influenciado pelos avanços das tecnologias de informação, sistemas de gestão de dados, análises e comunicações melhoradas, integração de sistemas e adopção da internet (Greenberg, 2001). Actualmente, relativamente às tecnologias de informação, o CRM significa a integração de tecnologias, tais como: *data warehouse*, *website*, intranet/extranet, balcão de ajuda (*help desk*), vendas, contabilidade, ERP e *data mining*. A tecnologia de informação que é capaz de coleccionar dados é integrada para fornecer a informação necessária para criar uma interacção mais pessoal/próxima com os clientes (Bose, 2002).

Nairn (2002) aprofunda mais o conceito de CRM e define-o como uma filosofia de negócio de longo prazo que se foca na recolha e na compreensão da informação do cliente, tratando de forma diferente os clientes diferentes, fornecendo um nível de serviço mais alto para os melhores clientes, e usando isto, para aumentar a lealdade e rentabilidade dos clientes. Esta ideia é apoiada por Buttle (2003) que afirma que o CRM é uma estratégia de negócios *core* que combina processos internos e funções com redes externas para criar e entregar valor para clientes rentáveis. Este autor também destaca a importância de usar dados de clientes de elevada qualidade. Outras definições de CRM podem ser encontradas na literatura (ver Payne e Frow, 2005; Ngai, 2005, para uma visão geral).

Relativamente aos benefícios do CRM, apesar de serem diferentes em cada área de negócio, alguns são comuns a todos (Swift, 2000). Estes benefícios geralmente são: baixos custos de aquisição de clientes, melhoria do serviço ao cliente, retenção dos clientes e aumento da sua lealdade, aumento da

rentabilidade dos clientes, facilidade na identificação de clientes rentáveis e aumento da produtividade da empresa (Alhaiou, 2011). O custo da aquisição dos clientes diminui devido à possibilidade de poupar em marketing, comunicação, contacto, acompanhamento, serviços de atendimento, etc. (Swift, 2000; Romano e Fjermestad, 2003; Curry e Kkolou, 2004). O serviço ao cliente melhora devido à análise dos processos promovidos pelo CRM. A integração de dados e a partilha de conhecimento com todos os negociadores incentivam no *design* de processos personalizados, estimulando o aumento dos níveis de serviço (Fjermestad et al., 2006). Como consequência da melhoria do serviço ao cliente, a satisfação e a presença dos clientes aumenta. Adicionalmente, a lealdade aumenta porque as empresas podem usar o conhecimento sobre os clientes para desenvolver programas de fidelidade (Crosby, 2002; Swift, 2000; Curry e Kkolou, 2004). Relativamente à rentabilidade dos clientes, esta aumenta devido ao aumento das vendas e às vendas cruzadas (*cross-selling*) (Bull, 2003; Curry e Kkolou, 2004). As empresas são capazes de saber quais são os clientes rentáveis, quais os que serão rentáveis no futuro e quais os que nunca serão rentáveis pela análise dos dados dos clientes (Kotler, 1999, Swift, 2000, Curry e Kkolou, 2004). O CRM também promove o aumento da produtividade das empresas, uma vez que, permite a integração de todos os departamentos das empresas, como tecnologia de informação, finanças e recursos humanos (Romano e Fjermestad, 2003; Crosby, 2002; Kracklauer et al., 2001).

De acordo com Dych (2001), as tecnologias de CRM podem ser divididas em três componentes: operacionais, colaborativas e analíticas. A primeira é referente à componente que ajuda a melhorar a eficiência das operações dos clientes do dia-a-dia (Peppers and Rogers, 2011), automatizando os processos dos negócios (Ngai et al., 2009). O CRM colaborativo pode ser visto como um centro de comunicação que fornece a conexão entre as empresas e os seus clientes, fornecedores e parceiros de negócio. Permite que os clientes, funcionários,

vendedores e parceiros consigam aceder, distribuir e partilhar informação. No passado, as empresas concentravam-se nas ferramentas operacionais e colaborativas, mas esta tendência parece estar a mudar (Reynolds, 2002). Os decisores aperceberam-se que as ferramentas analíticas são necessárias para conduzir decisões estratégicas, relacionadas com a identificação, atracção, desenvolvimento e retenção dos consumidores. O CRM analítico refere-se à análise das características e comportamentos do consumidor, sendo maioritariamente focado na análise de dados recolhidos e armazenados, para criar interacções mais significativas e lucrativas com os clientes. Para atingir este objectivo, os dados são processados, interpretados e reportados através de várias ferramentas (Greenberg, 2004). Os dados analisados fazem parte de um grande reservatório de informações, isto é, um *data warehouse*, que contém dados de fontes externas e internas, obtidos através de ferramentas operacionais. A informação ganha valor quando o conhecimento extraído se começa a usar. De acordo com Reynolds (2002), o componente de CRM mais crítico é a analítica. As soluções analíticas de CRM permitem gerir eficazmente a relação com os clientes. Só através da análise dos dados dos clientes é que as empresas podem compreender os comportamentos, identificar padrões e tendências de compra e descobrir relações causais. As informações obtidas a partir dos dados ajudam a modelar e a prever a satisfação e o comportamento dos clientes, de forma mais precisa, e podem constituir uma base quantificada para a tomada de decisões estratégicas. O CRM operacional, o CRM colaborativo e o CRM analítico trabalham juntos para criar valor para o negócio.

O trabalho realizado diz respeito maioritariamente ao CRM analítico, uma vez que foram utilizadas técnicas de *data mining* para explorar e melhorar a relação entre a empresa em estudo e os seus clientes. Assim, descreve-se de seguida, com mais detalhe, o CRM analítico, nomeadamente as suas dimensões e aplicações.

Swift (2000), Parvatiyar e Sheth (2001), Kracklauer et al. (2004), Ngai et al. (2009) categorizam o CRM analítico em quatro dimensões: (1) identificação de clientes, (2) atracção de clientes, (3) desenvolvimento de clientes e (4) retenção de clientes. Estas quatro dimensões podem ser vistas como um ciclo fechado do sistema de gestão de clientes (Au et al., 2003; Kracklauer et al., 2004; Ling and Yen, 2001).

Estas quatro dimensões podem ser descritas da seguinte forma:

1) Identificação de clientes: o CRM começa com esta dimensão, também chamada de aquisição de clientes (Kracklauer et al., 2004). A identificação de clientes inclui principalmente segmentação de clientes e análise de clientes-alvo. A segmentação dos clientes implica a subdivisão do conjunto de todos os clientes em segmentos mais pequenos, incluindo clientes com características semelhantes (Woo et al., 2005). A análise do cliente-alvo envolve a definição dos segmentos mais atractivos para a empresa, com base nas características dos clientes. A seleção dos grupos-alvo requer a recolha de dados quantitativos e qualitativos sobre esses grupos.
2) Atracção de clientes: ao identificar os grupos-alvo, as empresas concentram esforços e alocam recursos para atrair esses segmentos. As vantagens competitivas, como o preço e outras características de diferenciação, podem ser condutores da atracção de clientes. Outro condutor de atracção do cliente é o marketing directo. Este é um elemento do marketing mix da empresa que motiva os clientes a fazerem uma compra imediatamente (Cheung et al., 2003; Liao e Chen, 2004; He et al., 2005; Prinzie e Van den Poel, 2005). Por exemplo, a distribuição de correspondência ou de cupões são exemplos típicos de marketing directo. A atracção do cliente envolve o uso de um método apropriado de

comunicação e a eliminação de qualquer tipo de esforço desperdiçado (Kracklauer et al., 2004).

3) Desenvolvimento de clientes: o objectivo principal desta dimensão é aumentar a intensidade da transacção, o valor da transacção e a rentabilidade individual dos clientes. Os principais elementos do desenvolvimento dos clientes são: a análise do valor ao longo da vida do cliente e o *up/cross selling*. O valor da vida útil do cliente é o rendimento líquido total que uma empresa pode esperar de um cliente (Drew et al., 2001; Rosset et al., 2003). O *up/cross selling* são as actividades promocionais que procuram aumentar o número de serviços/produtos associados ou relacionados que um cliente usa dentro de uma empresa (Prinzie e Van den Poel, 2006). O *design* de tais actividades promocionais geralmente é apoiado pela análise de *market basket* que permite a identificação de padrões subjacentes ao comportamento do cliente (Aggarwal et al., 2002; Giraud-Carrier e Povel, 2003; Kubat et al., 2003).
4) Retenção de clientes: esta dimensão é uma das principais preocupações do CRM. De acordo com Kracklauer et al. (2004), a satisfação do cliente é a principal questão da retenção de clientes. A satisfação do cliente pode ser definida como a comparação das expectativas dos clientes (resultante do padrão pessoal, imagem da empresa, conhecimento de alternativas, etc.) com as percepções (resultantes da experiência real, impressão subjetiva do desempenho do produto, adequação do produto/serviço, etc.). A percepção do cliente sobre o valor oferecido pela empresa leva à retenção do cliente. Para além disso, uma experiência de compra de elevada qualidade leva a um sentimento emocional positivo, que permite que a empresa atinja a fidelização do cliente desejada. Os elementos desta dimensão do CRM incluem programas de marketing *one-to-one*, de fidelização, de prémios e de gestão de reclamações. O marketing *one-to-*

one envolve campanhas de marketing personalizadas suportadas pela análise, detecção e previsão de mudanças no comportamento dos clientes (Chen et al., 2005a, Jiang e Tuzhilin, 2006). Os programas de fidelização e de prémios envolvem campanhas ou actividades de apoio que procuram manter uma relação de longo prazo com os clientes. Exemplos de programas de fidelização incluem a pontuação de crédito, qualidade de serviço ou de satisfação e análise de abandono, ou seja, análise se um cliente é susceptível de sair da empresa para um concorrente (Ngai et al., 2009).

Apesar da sua aparente contribuição para a sustentabilidade e crescimento das empresas, o CRM analítico ainda não foi sistematicamente aplicado. De facto, as pesquisas sobre o CRM analítico são bastante limitadas (Anderson et al., 2007).

Relativamente à dimensão de identificação de clientes, Han et al. (2012) segmentou clientes de um operador de telecomunicações na China, através da consideração de valor do cliente como uma derivação do valor histórico, valor actual, valor a longo prazo, fidelização e crédito. Kim et al. (2006) propôs um modelo para analisar o valor do cliente e para segmentar clientes com base no seu valor. Este artigo utilizou como caso de estudo uma empresa de telecomunicações sem fios. Bae et al. (2003) propôs um seleccionador de anúncios da internet para fornecedores de *e-newspapers* que personaliza mensagens publicitárias. Para isto, os clientes foram divididos em diferentes segmentos, com base nas suas preferências e interesses. Woo et al. (2005) sugeriu um método de segmentação de clientes, baseado num mapa de visualização de clientes que descreve a distribuição de valor entre as necessidades e as características dos clientes. O mapa destes clientes foi aplicado a uma empresa coreana de cartões de crédito. Chen et al. (2003) construiu um modelo para uma empresa de turismo que prevê em quais rotas turísticas um novo cliente estará interessado. Este

modelo utiliza informações sobre os perfis de clientes e informações recolhidas anteriormente pelas rotas feitas.

Relativamente à atracção de clientes, Baesens et al. (2002) abordou a modelagem da incidência de compra para uma grande empresa europeia de correio. Foi avaliado se um cliente compraria ou não, tendo em consideração diferentes preditores de perfis de clientes. Buckinx et al. (2004) propôs um modelo que faz previsões sobre o resgate de cupões distribuídos por um retalhista *fast-moving* de bens de consumo. Este modelo tem em consideração o comportamento histórico e a demografia dos clientes. Ahn et al. (2006) introduziram um algoritmo optimizado que classifica os clientes em grupos de compradores ou de não compradores. Para validar o algoritmo proposto, este estudo usou dados de um site de alimentação online na Coreia. Este site contém todos os tipos de serviços para alimentação online, como fornecimento de informações, serviços comunitários e um centro comercial. O algoritmo também foi testado usando dados de outra loja online. Kim e Street (2004) sugeriram um modelo que permite identificar grupos alvo óptimos da campanha, com base na probabilidade de cada pessoa responder positivamente às mensagens da campanha. O modelo foi testado num contexto de seguros de veículos, usando dados de muitas famílias europeias. Chiu (2002) propôs um modelo para identificar os clientes que têm maior probabilidade de comprar seguros de vida. Este modelo de comportamento de compra foi desenvolvido através do uso de casos reais fornecidos por uma empresa de seguros de marketing directo de todo o mundo.

Relativamente ao desenvolvimento do cliente, Baesens et al. (2004) focou-se em prever se um novo cliente aumentaria ou diminuiria os seus gastos no futuro, tendo em consideração as informações iniciais de compra. Esta pesquisa foi realizada com *scanner data* de uma grande cadeia de retalho *do-it-yourself* Belga. Rosset et al. (2003) utilizaram modelos analíticos para estimar o efeito de várias

actividades de marketing no valor de vida dos clientes. Este estudo foi desenvolvido no contexto da indústria de telecomunicações. Brijs et al. (2004) abordaram o problema da análise de variedade de produtos e introduziu um modelo microeconómico de programação de integração para a seleção de produtos, considerando os conjuntos de produtos que geralmente são comprados em simultaneo. O estudo empírico foi baseado em dados de uma loja de conveniência totalmente automatizada. Chen et al. (2005b) introduziram um método para descobrir os padrões de compras dos clientes das bases de dados transaccionais das lojas, identificando produtos que geralmente eram comprados em conjunto. Prinzie e Van den Poel (2006) analisaram sequências de compras para identificar os padrões de *cross-selling*, que poderiam ser usadas para descobrir oportunidades de *cross-selling* no contexto de serviços financeiros.

A retenção de clientes mereceu especial atenção na literatura de CRM. Larivire e Van den Poel (2005) analisaram o impacto de um vasto conjunto de variáveis explicativas sobre a probabilidade de abandono. Este conjunto de variáveis incluiu o comportamento passado do cliente, a observação da heterogeneidade dos clientes e algumas variáveis típicas relacionadas com os intermediários em três medidas do resultado do cliente: próxima compra, abandono parcial e evolução da rentabilidade do cliente. Esta análise utilizou uma grande empresa europeia de serviços financeiros como caso de estudo. Hung et al. (2006) estimaram a previsão de abandono em telecomunicações móveis através do uso de dados demográficos de clientes, informações de cobranças, estados de contrato/serviço, registos de chamadas e registos de mudanças de serviço. Chen et al. (2005a) integraram variáveis comportamentais do cliente, variáveis demográficas, e uma base de dados transaccional para estabelecer um método para identificar mudanças no comportamento do cliente. Esta abordagem foi desenvolvida usando dados de uma loja de retalho. Ha et al. (2006) propuseram um sistema de recomendação de conteúdo que sugere conteúdo na *web*, neste

caso artigos de notícias. Este estudo considera as preferências dos utilizadores observadas quando estes estão a visualizar um site de notícias na internet. O estudo desenvolvido por Cho et al. (2005) propôs uma nova metodologia para a recomendação de produtos que usa sequências de compras do cliente. A metodologia proposta foi aplicada a uma grande *department store* na Coreia. Chang et al. (2006) usaram os padrões de navegação online dos clientes para auxiliar a pesquisa dos utilizadores de itens de interesse. Este estudo realizou uma análise empírica desenhada para o caso de uma loja de comércio digital que vendia câmaras digitais.

Este trabalho incidirá essencilamente nas dimensões relativas à identificação dos clientes, através da segmentação, e ao desenvolvimento de clientes, através da definição de estratégias de cross-selling.

2.3 Segmentação

Relacionado com a identificação dos cientes, a segmentação do mercado tornou-se um dos principais conceitos de marketing desde que foi apresentado por Wendell Smith. Wendell Smith (1956) propôs a segmentação do mercado como uma alternativa a técnicas de desenvolvimento de mercado para produtos diferenciados num mercado competitivo imperfeito. Os mercados tendem a ser mais rentáveis à medida que se definem produtos da perspetiva do consumidor, em vez das próprias necessidades da empresa (Wong and Saunders 1993; Day 1994).

A definição geral de segmentação é agrupar itens que são similares. Estes itens podem ser pessoas, espécies de plantas, partes ou sinais.

Kotler (1992) afirmou que, de forma a garantir segmentos de mercado diferentes, estão envolvidas três fases: de pesquisa, de análise e de perfis. Kolter apontou que uma segmentação de mercado eficiente deve possuir as seguintes cinco características: mensurabilidade, substancialidade, acessibilidade, diferenciação e capacidade de acção.

Kotler e Armstrong (2011) definiram segmentação como uma divisão do mercado em grupos mais pequenos de compradores com diferentes necessidades, características ou comportamentos que possam requerer estratégias de marketing diferentes. Os mercados não sendo homogéneos, há subgrupos de clientes com características específicas e necessidades diferenciadas. A segmentação pode ser feita segundo vários critérios como o tamanho dos segmentos, critérios geográficos, demográficos, psicográficos, etc.

2.4 *Cross-selling*

Destacando a dimensão de desenvolvimento de clientes, importa enfatizar o conceito de *cross-selling*. De acordo com Ansell et al. (2007) *cross-selling* é uma forma de criar laços mais fortes na relação de uma empresa com os clientes, isto é, é a estratégia de vender outros produtos a um cliente que já tenha comprado um produto ao vendedor, de forma a aumentar a confiança na empresa e diminuir a probabilidade de mudar para outro fornecedor. Para além disso, também tem o objectivo de exercer uma influência que geralmente é positiva na relação com o cliente, tornando-a mais forte (Kamakura et al., 2003). Consequentemente, argumenta-se que o *cross-selling* aumenta o valor do cliente ao longo do tempo (Kamakura et al., 2003). De acordo com Felvey (1982), é mais fácil uma empresa crescer desta forma do que crescer a tentar atrair novos clientes.

As bases de dados de marketing fornecem às empresas uma alternativa para olhar para a dinâmica dos seus negócios. A informação em massa gerada pelas bases de dados permite às empresas revolucionarem as vendas e os processos de marketing para atingirem uma forma mais eficaz de vender os seus processos e serviços. Estas bases de dados permitem obter informação que seja capaz de caracterizar os diferentes perfis de comportamento de consumidor, sendo uma boa base para tomar decisões mais sustentadas. Segundo Ansell et al. (2007), as campanhas de marketing diferenciadas fomentam um aumento da probabilidade de recompra.

Em geral, há duas abordagens de *cross-selling*: modo activo ou passivo. O modo activo refere-se a esforços de vendas proactivos iniciados pela empresa para identificar as perspectivas adequadas para os seus produtos e serviços. O *product-based campaign* consiste em encontrar o cliente certo para um produto particular, enquanto que o *customer-based campaign* consiste em vender o produto certo a um cliente particular de acordo com o seu perfil. O modo passivo refere-se à descoberta de oportunidades de *cross-selling* que surjam por própria iniciativa dos clientes.

Por exemplo, o sector bancário é utilizador do modo activo, obtendo retornos positivos. No entanto, têm surgido uma nova tendência (neste modo de *cross-selling)* para responder ao mercado volátil: os bancos utilizam as técnicas de *data mining* aplicadas sobre os dados das suas bases de dados para compreenderem melhor os seus clientes, e assim, tornar o serviço oferecido mais próximo. As vendas activas podem ser direccionadas para um grupo particular de clientes (que tenham as mesmas características num ou em mais aspectos) para que os produtos sejam desenhados para esse grupo baseado nas suas preferências, necessidades e capacidades financeiras.

Não há dúvida de que as campanhas direccionadas para um *target* são mais propensas a ter uma maior taxa de resposta e a incorrer em *running costs* mais

baixos, no entanto, a maior mudança de paradigma nas vendas é causada pela disponibilidade de informações de clientes na área de venda passiva. (Liu, 2004)

Porém, há desvantagens no *cross-selling*. Quando as ações de cross-selling são muito frequentes, estas podem incomodar os consumidores e fazer com que estes sejam menos receptivos, enfraquecendo a relação. Consequentemente, o uso desta táctica requer que o objectivo básico seja cumprido: vender o produto certo à pessoa certa.

Capítulo 3
3. *Data Mining*

Este capítulo refere-se a *data mining*, a abordagem metodológica utilizada para a elaboração deste projecto, e é dividido em três partes. Inicialmente, este capítulo define *data mining* e introduz as suas principais ferramentas. Seguidamente, mostra a relação entre *data mining* e o CRM analítico, e finalmente, refere as técnicas mais usadas, especialmente as utilizadas neste estudo.

As técnicas de *data mining* têm sido utilizadas para analisar os dados resultantes da actividade dos clientes armazenados nas bases de dados fornecidas por cartões de fidelização. Estes dados podem ser usados para detectar padrões e regras significativas subjacentes ao comportamento do consumidor.

Turban, Aronson, Liang, and Sharda (2007) definem *data mining* como o processo que utiliza estatística, matemática, inteligência artificial e técnicas informáticas para extrair e identificar informação útil e, consequentemente, obter informação de grandes bases de dados.

3.1 Principais tarefas de *Data Mining*

As ferramentas de *data mining* são formas populares de analisar a informação dos consumidores. Várias organizações coleccionam e armazenam a informação sobre os seus consumidores, no entanto, não têm a habilidade de descobrir informação de valor que lhe permite transformá-la em conhecimento (Berson et al., 2000).

O "knowledge discovery in databases" é uma ferramenta do CRM analítico. Como já foi referido, o desenvolvimento das tecnologias de informação permite o armazenamento e a análise de grandes volumes de dados, criando uma boa oportunidade para obter conhecimento. No entanto, a transformação dos dados em conhecimento útil é um processo lento e, por vezes, difícil. A abordagem sistemática que junta a fase de pré-processamento dos dados e a fase de pós-processamento é designada por "knowledge discovery in databases".

O "knowledge discovery in databases" é um processo complexo para descobrir as relações e outros padrões nos dados. Inclui um conjunto bem definido de fases, desde a preparação dos dados até à extracção da informação dos mesmos. As fases deste processo usado para extrair o conhecimento das bases de dados são: selecção dos dados, tratamento dos dados, pré-processamento dos dados, *data mining* e interpretação dos resultados. É um processo interactivo pois há várias decisões que têm de ser tomadas durante o processo e, se necessário, pode-se voltar a uma fase anterior para fazer alterações e depois continuar com o processo.

3.2 *Data Mining* e CRM analítico

As aplicações de *data mining* têm emergido de uma variedade de áreas, incluindo marketing, sector bancário, finanças, manufactura e saúde (Brachman et al., 1996). Além disso, o *data mining* também tem sido aplicado noutras áreas como na espacial, nas telecomunicações, na internet e na multimédia. De acordo com a área e o tipo de informação disponível, a técnica mais apropriada de *data mining* pode variar.

A maioria das técnicas de *data mining* podem ser vistas como conjuntos de algumas técnicas básicas e de princípios. Mais do que uma técnica pode ser

aplicada para a execução de uma tarefa específica. O desempenho de cada uma depende da tarefa a ser realizada e da qualidade dos dados disponíveis. De acordo com Ngai et al. (2009), a associação, classificação, *clustering*, previsão, regressão, *sequence discovery* e a visualização são as principais tarefas de *data mining*. Podem-se resumir estes grupos de tarefas de *data mining* da seguinte forma:

- A associação tem intenção de determinar relações entre atributos em bases de dados (Mitra et al., 2002; Ahmed, 2004; Jiao et al., 2006). O enfoque é a derivação de correlações de vários atributos, com respectivo suporte e confiança. Exemplos de resultados de associação são regras de associação. Por exemplo, estas regras podem ser usadas para descrever quais são os produtos que geralmente são comprados ao mesmo tempo que outros produtos num hipermercado.
- A classificação tem como objectivo classificar um item de dados numa das várias classes categóricas predefinidas (Berson et al., 1999; Mitra et al., 2002; Chen et al., 2003; Ahmed, 2004). Por exemplo, um modelo de classificação pode ser usado para identificar candidatos a empréstimos com riscos de crédito baixos, médios ou altos.
- O *clustering*, tal como a classificação, visa mapear os itens de dados numa das várias classes categóricas (ou *clusters*). Ao contrário da classificação em que as classes são predefinidas, no *clustering* são determinadas a partir dos dados. Os *clusters* são definidos pela descoberta de grupos naturais de itens de dados, com base em métricas de similaridade ou modelos de densidade de probabilidade (Berry e Linoff, 2004; Mitra et al., 2002; Giraud-Carrier e Povel, 2003; Ahmed, 2004). Por exemplo, um modelo de *clustering* pode ser usado para agrupar clientes que geralmente compram o mesmo grupo de produtos.

- A previsão estima o valor futuro de um determinado atributo, com base nos padrões dos registos. Esta lida com os resultados medidos como variáveis contínuas (Ahmed, 2004; Berry e Linoff, 2004). Os elementos centrais da análise de previsão são os preditores, ou seja, os atributos medidos para cada item, a fim de prever o comportamento futuro.
- A regressão mapeia um item de dados para uma variável de predição de valor real (Mitra et al., 2002; Giraud-Carrier e Povel, 2003). A adaptação das curvas, a modelagem das relações causais, a previsão e o teste de hipóteses sobre relações entre variáveis são aplicações frequentes de regressão.
- A *sequence discovery* pretende identificar relações entre itens ao longo do tempo (Berson et al., 1999; Mitra et al., 2002; Giraud-Carrier e Povel, 2003). Pode essencialmente ser pensada como uma descoberta de associação em bases de dados temporais. Por exemplo, a análise de sequências pode ser desenvolvida para determinar que se os clientes que se inscreveram para o plano A, se irão inscrever num próximo plano e em que prazo.
- A visualização é usada para apresentar os dados para que os utilizadores sejam capazes de observar padrões complexos (Shaw, 2001). Geralmente, esta técnica é utilizada em conjunto com outras de *data mining* para fornecer uma compreensão mais clara dos padrões ou das relações descobertas (Turban et al., 2010). Exemplos de aplicações de visualização incluem *mindmaps*.

As técnicas de *data mining* podem também ser categorizadas em aprendizagem supervisionada e aprendizagem não supervisionada. A primeira requer que o *dataset* contenha *targets* pré-definidos que representam as classes de itens dos dados ou os comportamentos que vão ser previstos. Por exemplo, um modelo supervisionado pode ser preparado para identificar padrões que permitam classificar clientes do banco como potenciais cumpridores ou não cumpridores

de crédito. Por outro lado, as técnicas não supervisionadas não requerem a preliminar definição das variáveis *target* (ex. classes de itens de dados) (Bose and Chen, 2009). Um modelo não supervisionado pode ser preparado para agrupar clientes em grupos desconhecidos semelhantes. A maior parte das técnicas de *data mining* são supervisionadas. De entre as não supervisionadas, as que são relativas ao *clustering* são as mais populares.

O uso de técnicas de *data mining* para extrair informação útil de dados é muito promissor. De facto, várias empresas têm coleccionado e armazenado resultados da interacção com os clientes, fornecedores e parceiros de negócio. No entanto, de acordo com Berson et al. (1999), a falta de capacidade de encontrar informação de valor nos dados tem impedido empresas de converterem esses dados em conhecimento útil e de valor. Em particular, dentro da dimensão analítica de CRM, as técnicas de *data mining* têm-se tornado formas populares de analisar os dados dos clientes. De facto, o uso de *data mining* para apoiar a dimensão analítica de CRM é visto como uma tendência emergente (Ngai et al., 2009). As técnicas de *data mining* podem ser usadas para apoiar estratégias competitivas de marketing, analisando e compreendendo os comportamentos e as características dos clientes, de forma a adquirir e manter clientes e maximizar o seu valor.

A selecção da técnica de *data mining* mais apropriada que permite extrair conhecimento útil da larga base de dados dos clientes é extremamente importante. De acordo com Berson et. Al (1999), quando são selecionadas cuidadosamente, as técnicas de *data mining* são uma das melhores ferramentas de suporte de decisões do CRM.

Ngai et al. (2009) propôs uma *framework* que retrata a relação entre as tarefas de *data mining* e o CRM analítico. Esta ilustração resulta de uma revisão da literatura das técnicas de *data mining* no CRM e é maioritariamente baseada na pesquisa conduzida pelo Swift (2000), Parvatiyar and Sheth (2002) and Kracklauer et al. (2004).

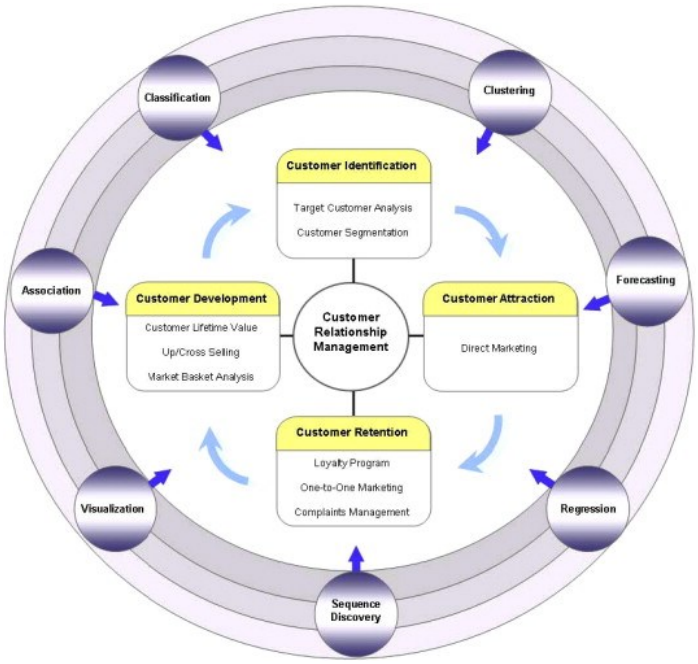


Figura 1: Quadro de classificação das técnicas de *data mining* no CRM. (Ngai et l., 2009)

De acordo com a revisão de literatura desenvolvida por Ngai et al. (2009), para o desenvolvimento do cliente através de *cross-selling*, uma das técnicas mais adequadas é a associação. Porém, se a dimensão for a identificação de clientes, as técnicas de classificação e *clustering* são as mais usadas, enquanto que se o objectivo for para manter clientes, as que se usam mais frequentemente são a associação e a classificação. Apesar disto, sabe-se que a combinação de técnicas de *data mining* é requerida muitas vezes para suportar cada dimensão analítica de CRM (Ngai et al., 2009).

Para uma melhor compreensão das tarefas de *data mining* que foram usadas neste caso de estudo, estas são descritas em seguida.

3.3 Técnicas mais utilizadas e aplicadas ao caso

3.3.1 *Clustering*

As técnicas de *clustering* são muitos úteis para adquirir conhecimento de um *dataset*. O *clustering* analisa os dados sem considerar uma classe pré-definida. Em geral, as classes não estão presentes nos dados, uma vez que não são conhecidas. Os itens são divididos em *clusters* de acordo com o princípio da maximização da similaridade dentro dos *clusters* e a minimização da similaridade entre os *clusters*. Isto significa que os *clusters* são formados para que os itens dentro de um *cluster* tenham grande similaridade, mas que sejam diferentes dos itens dos outros *clusters*. A figura 2 apresenta um exemplo ilustrativo de uma divisão de itens em *clusters*.

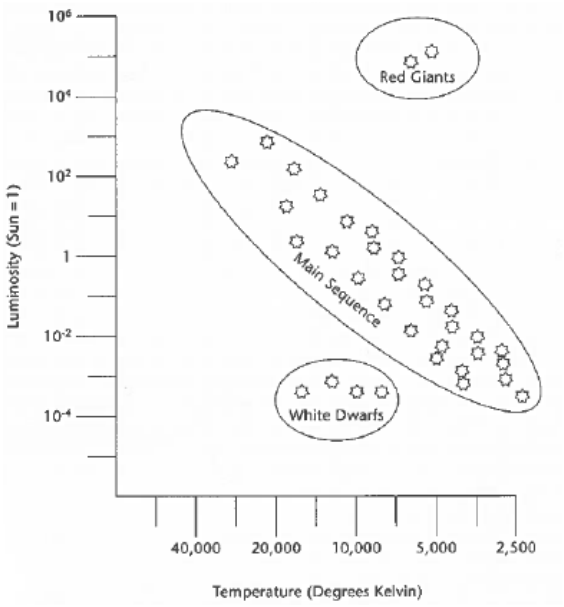


Figura 2: Exemplo de resultados de *clusters*. (Berry and Linoff, 2004)

Na maior parte dos algoritmos de *clustering*, o número de *clusters* a obter é definido previamente pelo analista (Thilagamani and Shanthi, 2010). De forma a apoiar a escolha do número de *clusters*, há diferentes métricas que visam avaliar a qualidade do resultado do *clustering* (ver Tibshirani et al., 2001, para mais informações). Por exemplo, o índice de Davies-Bouldin desenvolvido pelo Davies and Bouldin (1979) é é baseado na razão entre a soma da dispersão interna dos *clusters* e a distância entre os *clusters*. Um bom valor para o número de *clusters* é um baixo valor do índice de Davies-Bouldin. O critério do *elbow*,

proposto por Aldenderfer and Blashfield (1984), é baseado num típico *plot* de uma medida de erro (a dispersão de *cluster* definida tipicamente como a soma dos quadrados das distâncias entre todos os itens e o centróide do *cluster* correspondente, dividido pelo número de *clusters*) em relação ao número de *clusters* (*k)*. Se o número de *clusters* aumentar, a medida de erro diminui e a partir de algum *k*, a diminuição decresce significativamente. O termo e*lbow* é vulgarmente assumido para indicar o número de *clusters* apropriado. O método do *elbow* nem sempre funciona bem por não existir uma mudança na tendência de descréscimo do erro muito pronunciada, sendo por vezes aconselhável usar um método diferente para determinar o *k* (nº de *clusters* óptimo) como o Davies-Bouldin.

As técnicas de *clustering* mais conhecidas podem ser classificadas nas seguintes categorias: métodos particionados, métodos baseados em modelos e métodos hierárquicos (Yau e Holmes, 2011). Os métodos particionais são os mais utilizados na literatura, sendo apresentados de seguida.

3.3.1.1 Métodos particionados ou não hierárquicos

Métodos particionados (ou métodos não hierárquicos) criam *clusters* através da optimização de um critério objectivo, como a minimização da distância aos centroides. Dada uma base de dados de *n* itens, um método particionado constrói *k* partições de dados, onde cada uma representa um *cluster* e $k \leq n$. Cada grupo deve conter pelo menos um item e cada item deve pertencer exactamente a um grupo. Este segundo requisito pode não ser tão rígido em algumas técnicas de partição.

O método de partição mais comum é o *k-means* (Witten et al., 2001; Huang et al., 2007).

3.3.1.1.1 *k-means*

O algoritmo *k-means*, introduzido por Forgy (1965) e mais tarde desenvolvido por MacQueen (1967), atribui um conjunto de itens *n* a *clusters k.* O número de *clusters* é pré-definido pelo analista. De acordo com a definição do algoritmo de partição, *k-means* visa alcançar uma grande semelhança *intracluster* e uma baixa semelhança *intercluster*. Assim, cada item é atribuído ao *cluster* mais próximo, baseada na distância mínima entre o item e o *cluster* médio. Este algoritmo requer a definição das *initial seeds* (itens iniciais definidos como os centróides dos *clusters)* na primeira iteração do algoritmo. Depois de classificar um novo item, é calculado uma nova média para o *cluster* correspondente, determinando o novo centroide, e o processo continua. Este algoritmo envolve várias iterações com várias *initial seeds*. O processo está acabado quando a função de critério de partição, na maior parte das vezes a distâncias aos centroides, converge.

As vantagens do uso do *k-means* que foram frequentemente mais referidas são a simplicidade do conceito, a facilidade da implementação e a alta velocidade do processo computacional, o que o torna adequada para grandes *datasets* (Huang, 1998; Fred e Jain, 2002).

Por outro lado, o algoritmo *k-means* tem as desvantagens de exigir a especificação anterior do número de grupos (comum a todas as técnicas de partição) e de depender fortemente das *initial seeds*. A selecção de *initial seeds* diferentes pode gerar diferenças nos resultados de *clustering*, especialmente se o *target dataset* contiver muitos *outliers*. Este algoritmo não tem nenhum mecanismo para escolher as *initial seeds* apropriadas. Além disso, o *cluster mean* pode não ser o ponto mais representativo do *cluster*, e para *clusters* não convexos, esta técnica de *clustering* dará maus resultados (Looney, 2002; Tan et al., 2006). Portanto, apesar do k-means não ser adequado para todos os tipos de dados, apesar de requerer a definição dos nº de *clusters,* de ser susceptível a *outliers* e só lidar com

clusters com distribuição de pontos simétricos esféricos, o algoritmo é simples e consegue ser muito eficiente e usar grandes bases de dados.

Para fazer o *clustering* dos dados da organização em estudo, o método que se utilizou foi o *k-means*.

3.3.2 Classificação

O objectivo das técnicas de classificação é a previsão dos atributos de uma classe. O processo da classificação dos dados incluí dois passos. No primeiro, a fase de aprendizagem, uma técnica de classificação constrói o classificador através da análise de um conjunto de dados de treino, incluindo o vector de atributos e o valor da classe. No segundo passo, a função de aprendizagem obtida anteriormente é usada para prever a classe do restante conjunto do *dataset*.

Há várias técnicas de classificação na literatura tais como regressão logística, árvores de decisão, *random forest*, *etc*. No entanto, neste estudo a técnica de classificação utilizada foi as árvores de decisão, uma vez que permitem inferir regras subjacentes à classificação dos itens em cada classe. Assim, é importante explicar o seu conceito para uma melhor compreensão.

3.3.2.1 Árvores de decisão

Uma árvore de decisão é uma estrutura em forma de árvore que representa conjuntos de decisões em relação aos dados que permite gerar regras para a classificação de um conjunto de dados. As árvores de decisão são construídas através da repetida divisão dos dados de treino usando um critério de divisão,

até que todos os registos ou uma grande parte dos registos de uma partição pertençam à mesma classe. Partindo da raiz, que representa a totalidade dos dados de treino, os dados são divididos em dois ou mais subconjuntos. Esta divisão é baseada nos valores de um atributo escolhido de acordo com um critério de selecção de atributos, isto é, um critério que identifica o atributo que "melhor" separa um determinado conjunto de dados de itens em classes individuais. Para cada subconjunto, um nó filho é criado e os dados do subconjunto são incluídos. O processo é posteriormente repetido até que um critério de término seja cumprido. A árvore de classificação resultante torna-se um ponto de partida para a tomada de decisão.

De acordo com Friedl e Brodley (1997), as vantagens das árvores de decisão passam pela capacidade de lidar com dados medidos em diferentes escalas, pela falta de quaisquer pressupostos relativos à distribuição de frequência dos dados em cada classe, pela flexibilidade e pela capacidade de lidar com as relações não-lineares entre atributos e classes. Para além disso, as árvores de decisão são rápidas e fáceis de formar (Teeuwsen et al., 2004). As árvores também podem ser utilizadas para a selecção/redução de atributos (Borak e Strahler, 1999). Por último, o analista pode interpretar facilmente uma árvore de decisão.

Uma das principais desvantagens das árvores de decisão é a instabilidade, isto é, diferentes conjuntos de dados de treino de um determinado domínio problemático podem produzir árvores muito diferentes (Breiman, 1996).

3.3.3 Associação

As regras de associação são a aplicação do *data mining* para extrair combinações frequentes de atributos. A extracção de conjuntos de itens frequentes (ou *itemset*) é um dos tópicos mais populares na descoberta de conhecimento em bases de dados e dos mais utilizados na indústria e comércio.

O trabalho pioneiro nesta área foi a análise de carrinhos de compras (*market-basket analysis*), mais especificamente, a análise de dados transaccionais que descrevem o comportamento de consumo dos clientes (Agrawal et al.,1993). O objectivo desta análise consiste em descobrir grupos de produtos que são frequentemente comprados em conjunto e, com base nesses grupos, inferir os produtos que são comprados, dado que foram comprados também outros produtos. Para resolver este problema, foi desenvolvido um grande número de algoritmos, que têm sido aplicados aos mais diversos problemas. Existem então algoritmos para extracção de conjunto de itens frequentes, regras de associação, assim como extensões para sequências de itens, onde é importante considerar a ordem pela qual os itens de um conjunto aparecem.

A partir de conjuntos frequentes, é possível derivar regras de associação. As regras de associação têm a forma de regras: "se antecedente então consequente", em que ambos os antecedentes e consequentes são *itemsets*. Por exemplo, se o cliente compra pão então também compra manteiga. Estas regras são calculadas a partir dos dados e têm natureza probabilística. O grau de incerteza de uma regra é dado pela *confiança* da regra, que pode ser definida como a relação entre o número de transacções que incluem todos os itens no conjunto (consequente e antecedente), e o número de transacções que incluem todos os itens do antecedente.

Desde a sua introdução por Agrawal et al. (1993), os problemas de aprendizagem de *itemsets* frequentes e regras de associação receberam uma grande atenção. Na década de 1990, foram publicados centenas de artigos propondo novos algoritmos, ou melhorias nos algoritmos existentes, para resolver estes problemas de aprendizagem de uma forma mais eficiente.

Os resultados da aplicação de algoritmos de associação poderão resultar num conjunto de acções que vão desde a promoção conjunta de artigos até alterações da sua localização no supermercado.

No entanto, os resultados das regras de associação devem ser interpretados cuidadosamente. Implicam uma forte relação de coocorrência entre os itens, mas não uma relação causa-efeito.

O método utilizado para ajudar a criar as regras de associação foi o algoritmo FP Growth. Este algoritmo utiliza estratégia de procura em profundida e árvores de sufixo que são empregues na maioria dos algoritmos de extracção de padrões frequentes aplicados a dados de fluxo contínuo (Chi et al., 2004). Para a construção de regras de associação, este algoritmo abrange duas fases. Na primeira fase constrói uma estrutura de dados, a FP-tree, percorrendo a base de dados duas vezes. Na primeira vez, encontra o conjunto de itens frequentes (1-*itemsets*) e os respectivos suportes. A FP-tree é então usada para encontrar as regras de associação. A ideia base consiste em percorrer o espaço de procura em profundidade, conforme a figura.

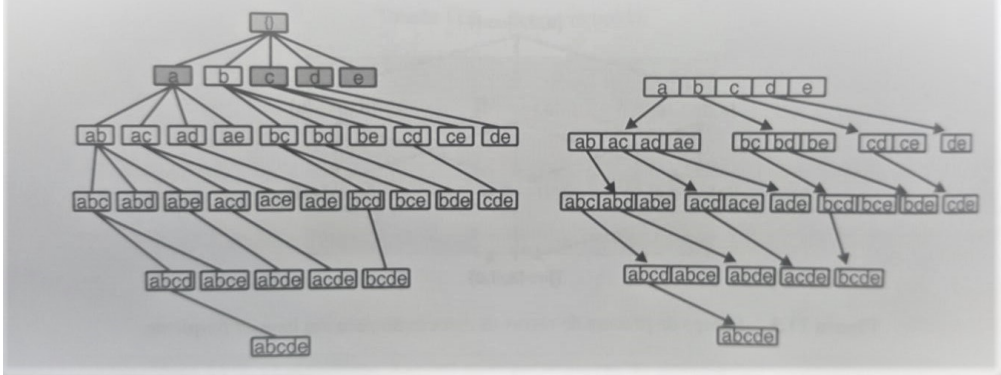


Figura 3: O espaço de procura em profundidade e a árvore de prefixos correspondente para 5 itens. (Gama, Carvalho, Faceli, Lorena, & Oliveira, 2017)

O algoritmo atravessa duas vezes a base de dados. Na primeira encontra o conjunto de itens frequentes (1-*itemsets*) e os respectivos suportes. O conjunto de itens é ordenado por ordem decrescente do seu suporte é armazenado numa matriz. No caso da base de transacções da figura 3 o item mais frequente é o *a*, seguido de *b*, *c*, *d* e *e*. O algoritmo percorre uma segunda vez a base de

transacções, construindo a árvore de padrões frequentes, FP-tree, como se segue. Primeiro, cria-se o nó raiz da árvore, rotulado com *null*. Para cada transacção presente na base de dados, os itens são processados em ordem decrescente de suporte. Por exemplo, a primeira transacção (*a,b*), gera o percurso: $null \rightarrow a \rightarrow b$. Cada nó tem associado um contador de frequência com o valor 1. A segunda transacção gera um novo conjunto de nós, correspondente ao percurso $null \rightarrow b \rightarrow c \rightarrow d$. Este percurso é disjunto do primeiro porque as transacções não compartilham um prefixo comum. Como *b* aparece nos dois percursos, é estabelecida uma ligação que possibilita o cálculo da frequência de *b*. A terceira transacção partilha um prefixo comum (o item *a,* com a primeira transacção). Por esse motivo, o percurso $null \rightarrow a \rightarrow c \rightarrow d \rightarrow e$ sobrepõe-se ao primeiro percurso, e a frequência de *a* é incrementada. Este processo continua até todas as transacções terem sido processadas.

A razão para a primeira leitura da base de dados e para o processamento de transacções ser feita por ordem decrescente do suporte está relacionada com a disposição dos itens na FP-tree. Quanto mais frequentes, mais próximo da raiz aparecem, o que favorece a possibilidade de serem compartilhados. Desta forma, a apresentação da FP-tree da base de dados é mantida tão pequena quanto possível. Além disso, é possível eliminar da análise os itens com suporte inferior ao suporte mínimo.

Os algoritmos para as regras de associação tendem a gerar um número elevado de regras. Nos últimos anos, têm sido propostas várias medidas para extrair padrões interessantes a partir de grandes bases de dados. A ideia consiste em selecionar um subconjunto de padrões ou regras que de alguma forma sejam mais relevantes.

O Coeficiente de Interesse, ou *Lift*, reflecte a noção de estatística de independência entre duas variáveis aleatórias. Foi abordado por muitos autores, como uma medida para avaliar os níveis de associação. Esta métrica é definida

pelo quociente entre a probabilidade conjunta de duas variáveis em relação à sua probabilidade, pressupondo a hipótese de independência. O *lift* não é mais do que o quociente entre a confiança e o valor esperado para a confiança. A confiança esperada é o número de transacções que incluem o consequente dividido pelo número total de de transacções. Um valor de *lift* igual a 1 indica que A e B são independentes. Valores de *lift* inferiores a 1 indicam que A e B são negativamente correlacionados, aparecendo em conjunto menos vezes que o esperado e indicando que a ocorrência de A tem um efeito negativo sobre a ocorrência de B. Valores superiores a 1 indicam uma correlação positiva, indicando que A e B aparecem mais frequentemente juntos que o esperado.

Capítulo 4
4. Metodologia e Dados

4.1 Metodologia

Ao longo deste capítulo pretende-se explicar a metodologia usada para responder ao problema identificado. Primeiramente, apresenta-se o método de obtenção e tratamento dos dados, e depois, o método utilizado para o desenvolvimento da segmentação de clientes e do *cross-selling*.

Como foi referido anteriormente, o objectivo desta investigação é auxiliar a Farmácia Gaia Jardim na caracterização dos seus clientes e no desenvolvimento de estratégias de *cross-selling* para melhorar a sua relação com os clientes. Para isto recorreu-se a ferramentas de marketing analítico. Espera-se que a metodologia proposta possa vir a ser utilizada no futuro por outra organização com objectivos similares.

Primeiramente, o projecto envolveu uma análise da organização em questão e do seu ambiente de operação. Esta foi feita através de visitas à farmácia e implicou a exploração do software informático associado ao cartão de fidelização. Os dados referentes à utilização do cartão de fidelização são armazenados a partir de um sistema chamado "Yclient", implementado no final de 2016.

Conhecida a realidade de operação da farmácia e do seu sistema informático, procedeu-se à recolha dos dados a utilizar. Estes dados foram exportados do sistema e armazenados num ficheiro Excel, incluindo a identificação de cada cliente, através de um número de cartão, alguns factores descritivos dos clientes, das transacções efectuadas (produtos comprados por cada cartão) e a listagem dos produtos existentes na farmácia. É de realçar que os produtos em questão

não são alvo de prescrição médica, uma vez que não faria sentido promovê-los através de *cross-selling*.

Após uma análise preliminar da base de dados, em primeiro lugar, foi feito o pré-processamento dos dados. De seguida, foi feita a segmentação dos clientes tendo em consideração a frequência (isto é, visitas à farmácia) e o valor despendido pelos clientes nas suas visitas. Depois da segmentação, procedeu-se à caracterização dos segmentos obtidos, e finalmente, foram identificadas as regras de associação entre produtos presentes para apoiar a definição deestratégias de *cross-selling*. Note-se que a extração de regras de associação foi feita para cada segmento de clientes identificado, partindo do pressuposto que estes poderiam ter comportamentos de compra diferentes, no que respeita aos artigos comprados em conjunto.

De seguida, dar-se-á mais detalhe às fases referidas anteriormente.

Pré-processamento

Relativamente ao pré-processamento dos dados, este foi feito em duas fases. Em primeiro lugar, efectuou-se uma identificação dos clientes a considerar na análise e, posteriormente, seleccionou-se as transacções efectuadas por estes clientes. Esta identificação e selecção resultou da eliminação de dados incompletos e/ou duplicados. De seguida, associou-se a descrição do produto presente na tabela dos "Produtos" à tabela de "Transacções", para se obter a "Subsubfamília" dos produtos, de forma a caracterizar o produto pela sua categoria para uma melhor compreensão. Posteriormente, foi necessário eliminar as transacções referentes aos produtos sem "Subsubfamílias", uma vez que estes respeitavam a produtos com prescrição médica, e as transações referentes aos produtos que não constam da lista de produtos.

Por forma a realizar a análise de *clustering*, houve a necessidade de se estimar as variáveis a utilizar na caracterização dos clientes. As variáveis que se

considerou mais relevantes neste contexto foram o valor médio gasto por transação e a frequência de idas à farmácia. Para este efeito, recorreu-se a uma *pivot table* que permite fazer a agregação destes dois indicadores para cada cliente.

Depois de feito o pré-processamento dos dados, foi usado o RapidMiner como ferramenta para se realizar os próximos passos: segmentação de clientes, descrição dos mesmos e obtenção de regras de associação, relativas à compra de produtos em simultâneo.

Segmentação dos clientes

Relativamente à segmentação dos clientes, utilizou-se a ferramenta de *clustering* descrita anteriormente, nomeadamente o algoritmo *k-means*. Como mencionado anteriormente, esta segmentação foi feita com base na frequência (visitas à farmácia) e no valor médio gasto por visita à farmácia. Foi necessário avaliar o resultado do *clustering*, para diferentes valores de *k* (em que *k* é o número de *clusters*), para se definir o número de segmentos mais adequado. Para tal, através do *elbow method* obteve-se um gráfico com a soma dos erros quadrados (SSE) para os diferentes valores de *k*. Uma vez que nem sempre o *elbow method* funciona perfeitamente (especialmente se os *clusters* dos dados não forem naturalmente bem definidos), usou-se também um método alternativo, o Davies Bouldin.

Note-se que para que a análise de *clustering* fosse válida, foi necessário normalizar os valores das variáveis usadas para apoio à segmentação.

Descrição dos segmentos

Em seguida, foi necessário proceder à caracterização dos segmentos obtidos, e para tal, recorreu-se ao uso das técnicas de classificação, nomeadamente às árvores de decisão. Assim, construiu-se um modelo de classificação, suportado

por uma árvore de decisão, em que a variável dependente era o *cluster* obtido e as variáveis preditivas a frequência e valor gasto por transação.

Note-se que a obtenção de regras de classificação, através das árvores de decisão, permitirão uma alocação de futuros clientes aos segmentos estabelecidos através deste trabalho, sem que se volte a desenvolver uma nova análise de *clustering*.

Obtenção de regras de associação

Depois dos *clusters* obtidos, foi necessário acrescentar ao documento em excel com os dados, os *clusters* associados a cada número de cartão. Este documento serviu de base para uma das últimas fases necessárias para atingir o objectivo proposto. Agrupou-se os dados transacionais por segmento de clientes, a fim de se obter regras de associação distintas para grupos de clientes com comportamentos diferentes.

Feito isto, recorreu-se novamente ao RapidMiner, para obter as regras de associação através algoritmo FP-Growth, descrito no capítulo 3.

Findo o processo analítico, e obtidas as regras de associação de produtos para cada segmento, procedeu-se à elaboração de sugestões de *cross-selling*.

4.2 Dados

Os dados que foram recolhidos dizem respeito a três meses de operação na farmácia Gaia Jardim: Janeiro, Fevereiro e Março de 2017.

Os dados recolhidos eram de carácter secundário, uma vez que foram recolhidos com um próposito que não o da análise descrita neste trabalho. Estes

dados incluíam quatro tipos de informação diferentes. Em primeiro lugar, a identificação e caracterização de clientes através de um número de cartão identificativo (cada cartão é um individuo), com a localidade, idade, género e total gasto de cada um. Em segundo lugar, uma lista de todas as transacções registadas na farmácia entre Janeiro e Março de 2017, ordenadas por ordem cronológica, com a indicação do dia e da hora em que ocorreu e com um código de transacção associado a cada ocorrência. Para completar a informação anterior, também foi cedida uma lista com a associação desses códigos de transacção aos produtos em questão. Por último, foi recolhida uma lista com a "subsubfamilia" de cada produto anteriormente referido para se classificar a categoria de cada produto.

Relativamente à primeira parte da informação disponibilizada, isto é, referente aos clientes registados na farmácia, o software do sistema informático à data da recolha dos dados, em Abril de 2017, tinha cerca de 10.000 clientes registados. No entanto, nem todos os cartões tinham a informação completa sobre os consumidores, sendo que para se conseguir analisar e ter um visão o mais completa possível do perfil dos clientes, foram usados os cerca de 4.000 clientes que tinham todas as suas informações devidamente preenchidas.

Capítulo 5
5. Resultados da pesquisa

Neste capítulo, é apresentada a informação que foi possível obter através dos dados dos cartões dos clientes da Farmácia Gaia Jardim para se definir as estratégias de *cross-selling* mais adequadas.

A exposição dos resultados está dividida da seguinte forma: inicia-se pela análise do contexto da farmácia, de seguida são apresentados os resultados do pré-processamento necessário às fases seguintes. Posteriormente, é apresentada a análise de segmentação dos seus clientes, ao que se segue a caracterização desses segmentos e as regras de associação de produtos obtidas para cada segmento de clientes. Por fim, são apresentadas as estratégias de *cross-selling* que podem ser aplicadas pela Farmácia Gaia Jardim.

5.1 Análise do contexto da farmácia

Primeiramente, para uma melhor compreensão do contexto da farmácia onde esta investigação se insere, foi feita uma análise dos clientes sobre os quais a farmácia em estudo dispunha de toda a informação que permitia caracterizar os mesmos. Assim, como já foi referido no capítulo da metodologia, foram analisados cerca de 4000 clientes. Relativamente a esta amostra, tal como se pode verificar na figura 4, a maioria dos clientes são crianças até aos 3 anos e jovens adultos entre os 30-40 anos, maioritariamente de género feminino (ver Figura 4). Considera-se que no caso dos cartões associados a perfis de crianças, os pais optam por fazer esta associação, uma vez que, eventualmente a maior parte das compras feitas na farmácia têm como alvo a criança. Este aspeto pode desde logo

revelar uma limitação do formulário de adesão ao cartão de fidelização. Deveria ser possível que um adulto ficasse associado ao cartão, disponibilizando ainda informação sobre qual ou quais os principais visados pelas compras na farmácia.

A maior parte dos clientes são de freguesias de Vila Nova de Gaia, embora haja uma percentagem minimamente significativa que pertence a freguesias do Porto. Em baixo, está apresentada a figura 5 onde estão representadas as percentagens dos concelhos dos clientes da farmácia com os dados devidamente preenchidos.

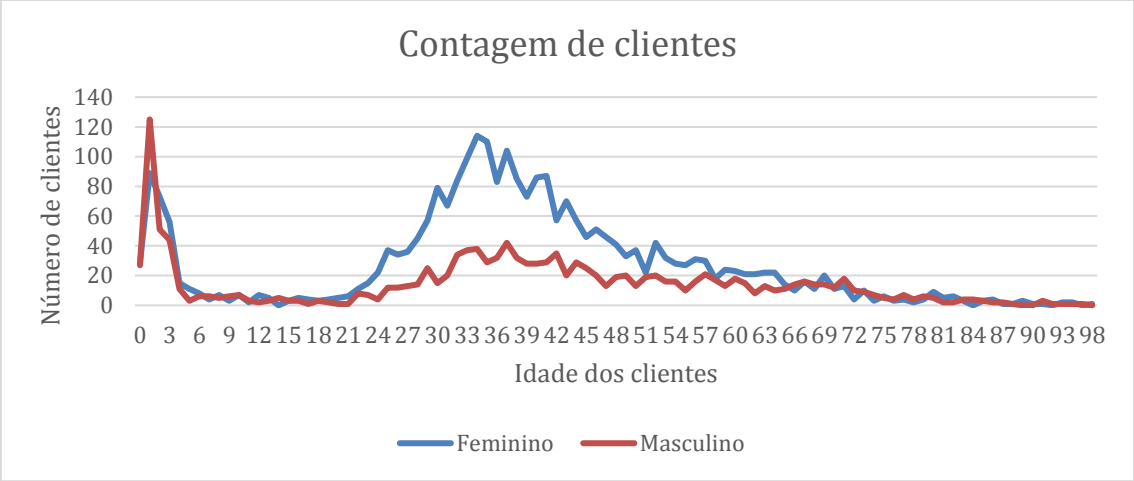


Figura 4: Gráfico representativo da idade e sexo dos clientes

O valor médio total gasto observado é de 174€ por pessoa durante estes três meses em análise. No entanto, a figura 6 apresentada também permite observar que há uma parte significativa de clientes que gastou valores mais altos ao longo dos três meses.

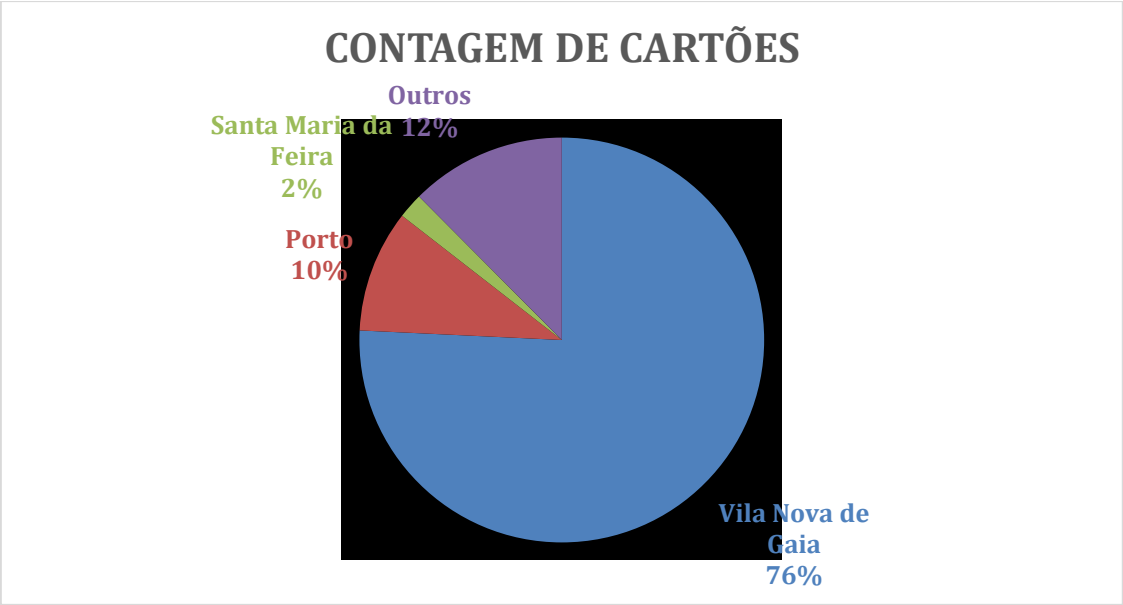


Figura 5: Gráfico representativo das localidades dos clientes

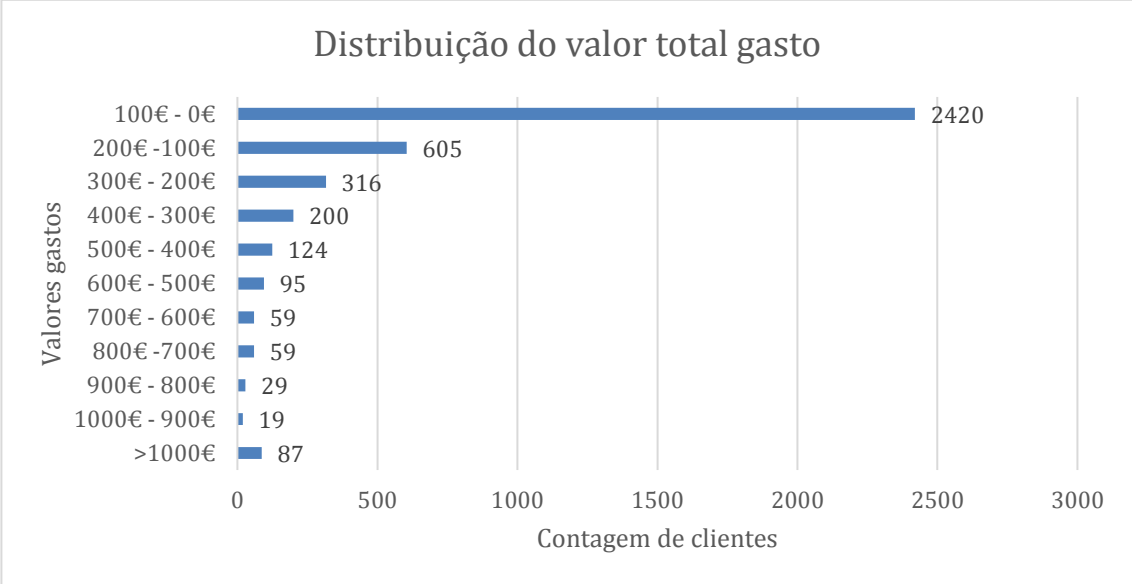


Figura 6: Gráfico representativo dos valores gastos.

Ainda com a informação disponibilizada pela farmácia foi possível construir a figura 7 em que avalia o número de produtos vendidos ao longo dos três meses por categoria. Estas categorias foram desenhadas pela farmácia para os seus próprios sistemas. É facilmente observável que a farmácia vende essencialmente produtos associadas à gravidez, aos bebés e às mães. É importante referir que nas "Ajudas técnicas" estão inseridas fraldas para bebés e nos "Suplementos

Alimentares" há vários produtos para gravidez e amamentação. Este facto ainda reforça mais as vendas da farmácia no que toca a produtos relacionados com a maternidade.

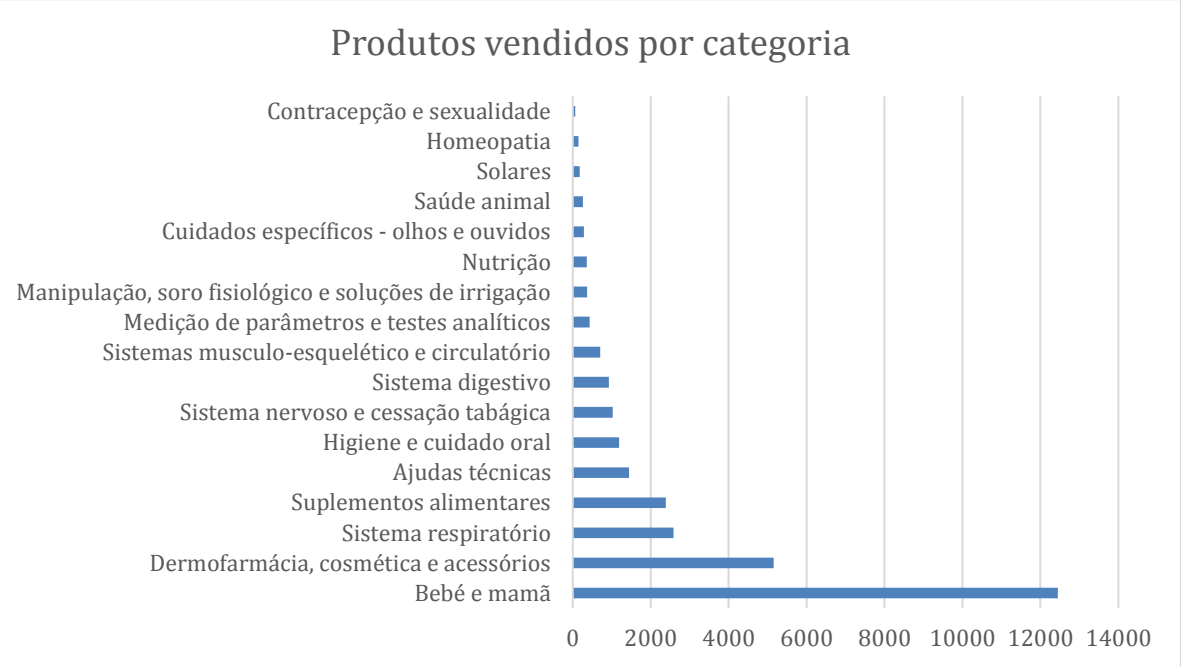
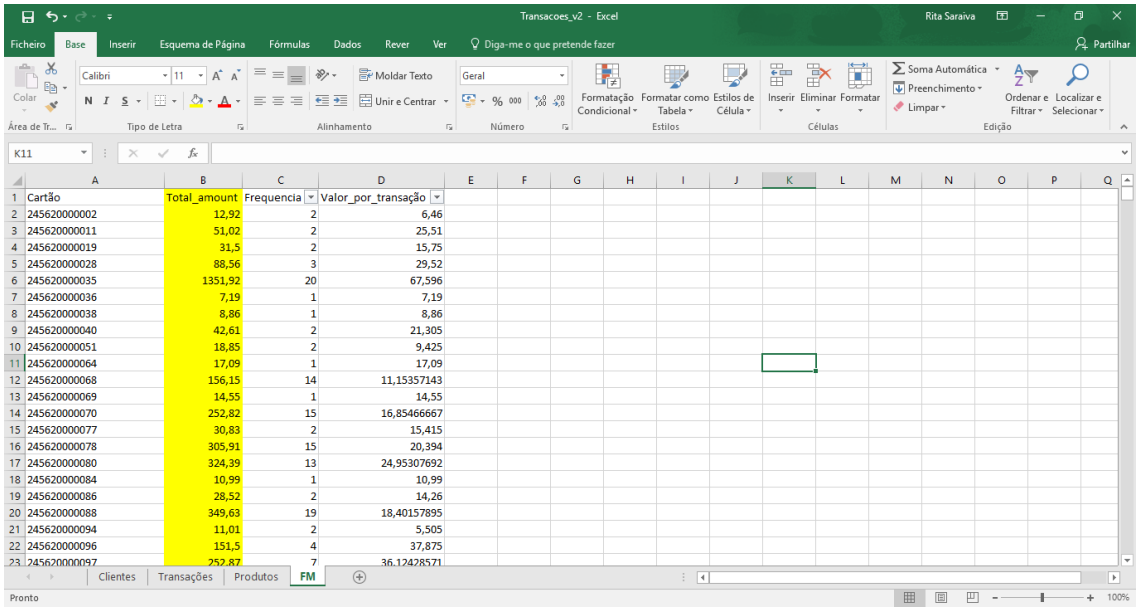


Figura 7: Gráfico representativo das categorias dos produtos vendidos

5.2 Pré-processamento

Depois de uma caracterização genérica dos clientes da farmácia, foram utilizados os dados pré-processados referentes às transacções na farmácia durante os três meses em análise para se desenvolver a análise de *clustering* já referida. Este pré-processamento permitiu a correcta utilização dos dados no RapidMiner, que como já foi referido, foi o software utilizado para se realizar as próximas fases desta pesquisa. Assim, depois deste pré-processamento, explicado no capítulo da metodologia, obteve-se um total de 6521 cartões, 15670 transaccões e 4209 produtos transaccionados, que foram utilizados parar analisar o comportamento dos consumidores na farmácia. Seguidamente, para cada

cartão diferente (considerado como representando um cliente diferente) fez-se as operações necessárias para obter o valor médio gasto em cada transacção e o número de transaccões diferentes (isto é, visitas à farmácia). O resultado deste pré-processamento foi um documento em excel, ilustrado na figura 8.



	A	B	C	D
1	Cartão	Total_amount	Frequencia	Valor_por_transação
2	245620000002	12,92	2	6,46
3	245620000011	51,02	2	25,51
4	245620000019	31,5	2	15,75
5	245620000028	88,56	3	29,52
6	245620000035	1351,92	20	67,596
7	245620000036	7,19	1	7,19
8	245620000038	8,86	1	8,86
9	245620000040	42,61	2	21,305
10	245620000051	18,85	2	9,425
11	245620000064	17,09	1	17,09
12	245620000068	156,15	14	11,15357143
13	245620000069	14,55	1	14,55
14	245620000070	252,82	15	16,85466667
15	245620000077	30,83	2	15,415
16	245620000078	305,91	15	20,394
17	245620000080	324,39	13	24,95307692
18	245620000084	10,99	1	10,99
19	245620000086	28,52	2	14,26
20	245620000088	349,63	19	18,40157895
21	245620000094	11,01	2	5,505
22	245620000096	151,5	4	37,875
23	245620000097	252,87	7	36,12428571

Figura 8: Amostra do resultado do pré-processamento

5.3 Segmentação

Seguidamente, utilizaram-se os dados da figura 8, para se fazer a segmentação com base nas variáveis já referidas, isto é, frequência e valores gastos, com o auxílio do módulo de *clustering* do RapidMiner.

Para se fazer a segmentação, foi necessário definir o número de *clusters* adequados ao caso, e para isso, utilizaram-se diferentes valores de *k* para se usar o *Elbow Method*, obtendo-se o seguinte gráfico apresentado na figura 9 com a soma dos erros quadrados (SSE):

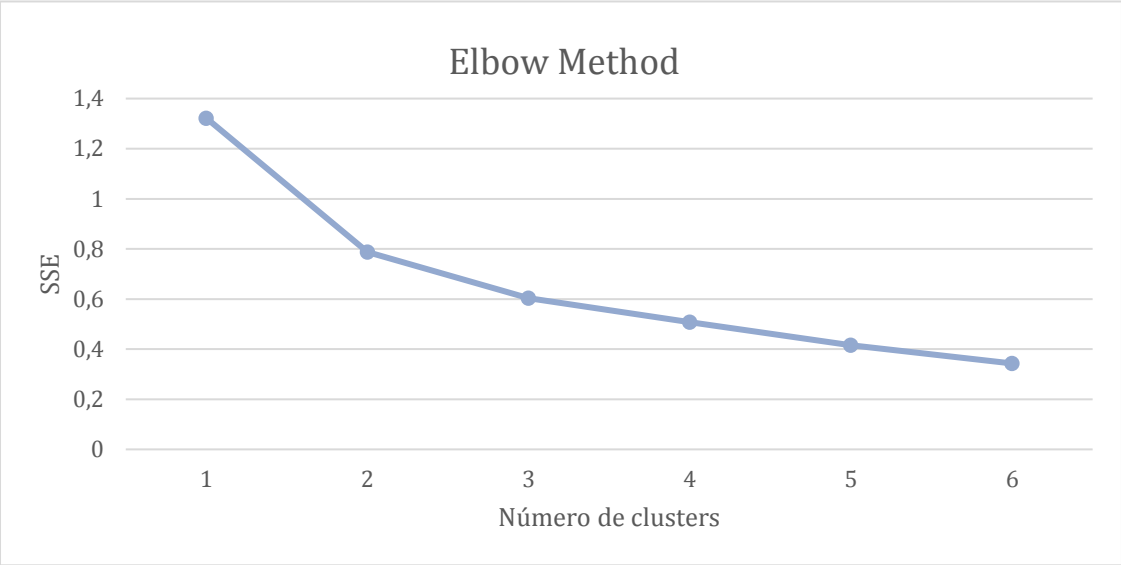


Figura 9: Elbow Method

Uma vez que, com o uso deste método não foi possível detectar um *elbow* pronunciado, optou-se pela utilização de um método alternativo, o Davies Bouldin. Assim, obtiveram-se os seguintes valores apresentados na tabela 1:

Tabela 1: Valores Davies Bouldin

Número de *clusters* (k)	Davies Bouldin
2	1,157
3	0,710
4	0,770
5	0,861
6	**0,686**
7	0,733

Através da utilização deste método, foi possível concluir que o número de segmentos/*clusters* adequados para a caracterização desta amostra de dados é de 6.

Depois de definido o número de *clusters* adequados, foram extraídos do programa dois documentos: a figura 10 com a representação gráfica dos *clusters*, tendo em conta as variáveis de segmentação, e uma listagem para excel mostrada na figura 11 a que foi dado o nome de "Cartão e *Cluster*" em que se associa cada cartão ao respectivo *cluster*. No anexo I, é possível ver a forma como se utilizou o software RapidMiner para se obter os *clusters*.

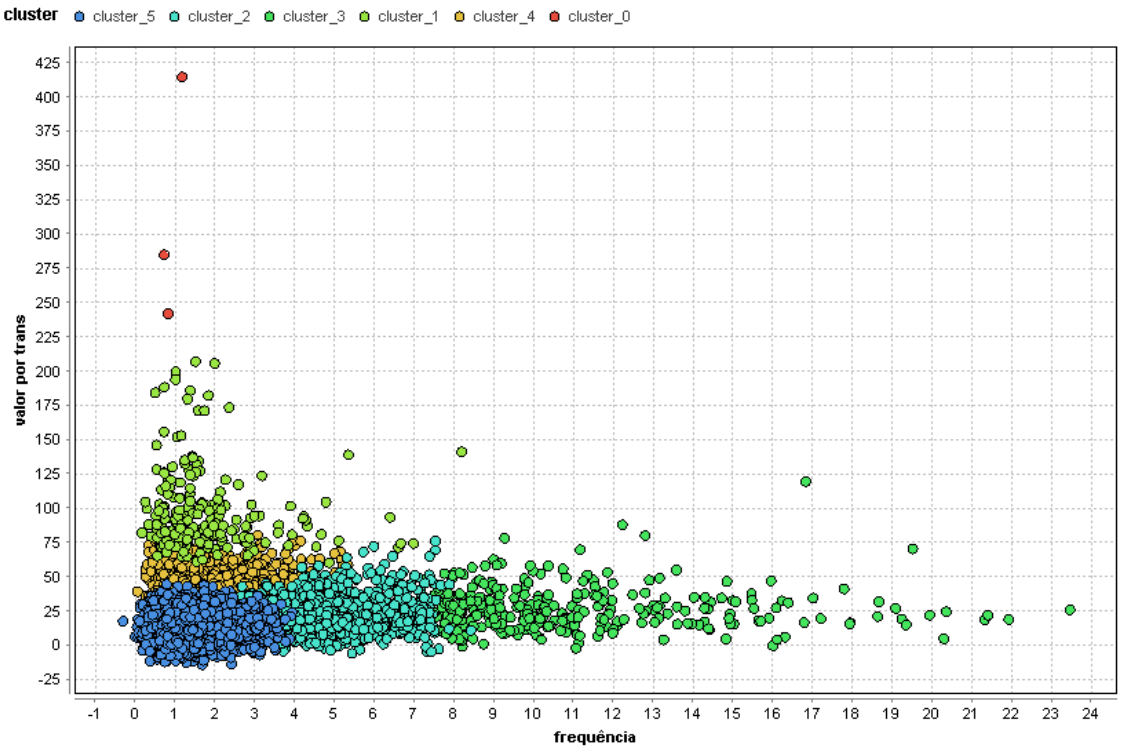


Figura 10: Gráfico representativo dos *clusters* obtidos

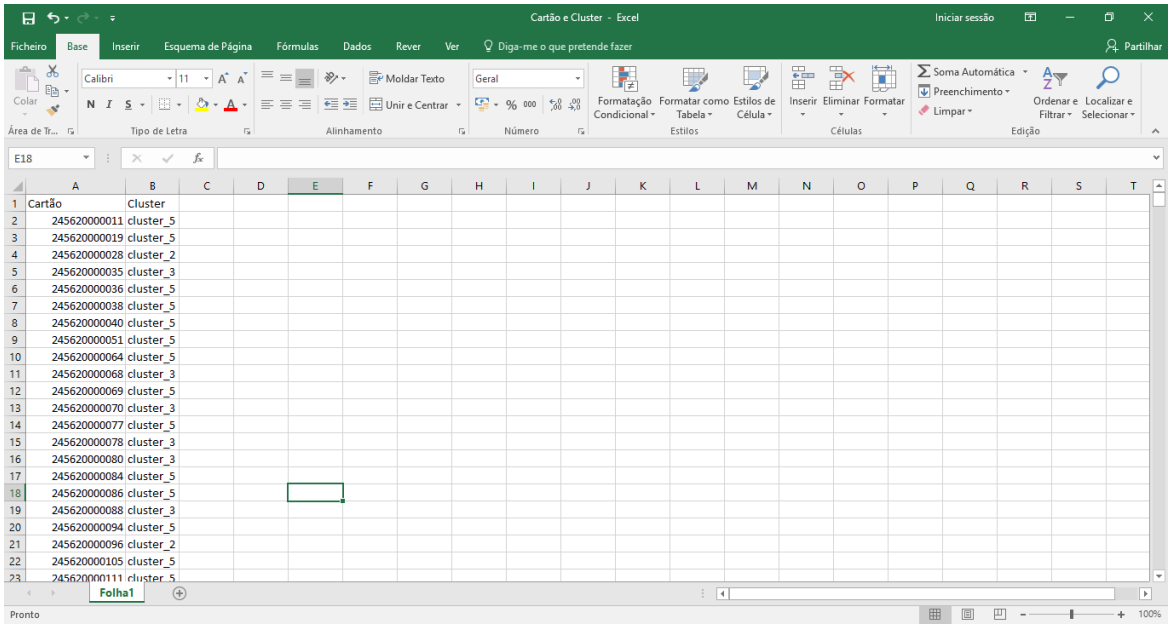
No gráfico obtido através do RapidMiner consegue-se verificar os seis *clusters* obtidos. O *cluster* 0 (vermelho) representa clientes com comportamento muito distinto dos restantes, pelo que não se desenvolveu qualquer análise referente a estes. É possível identificar o *cluster* 5 (azul escuro) que caracteriza o conjunto de clientes que menos visita a farmácia e que menos dinheiro gasta comparativamente com os outros grupos. Este grupo de clientes parece ser o mais representativo. Parece haver outros dois grupos que despendem relativamente pouco por transacção: o *cluster* 2 (azul turquesa) e o *cluster* 3 (verde forte). Porém, estes dois *clusters* visitam mais vezes a farmácia do que o *cluster* 5. Por outro lado,

parece que o *cluster* 4 (amarelo) e o *cluster* 1 (verde claro) se referem a clientes que visitam relativamente menos vezes a farmácia (tal como o cluster 5), mas gastam mais dinheiro por visita. Note-se que os valores apresentados não estão normalizados, o que permite uma clara perceção dos valores reais que foram extraídos da base de dados da farmácia.

Como se pode verificar na tabela 2, o *cluster* que mais clientes abrange é o 5 com 60% dos clientes da farmácia. Assim como, o *cluster* 4 tem uma percentagem de 19%. Por outro lado, o *cluster* 1 e 3 são os que menos clientes têm.

Tabela 2: Proporção dos clientes por *cluster*

Cluster	**1**	**2**	**3**	**4**	**5**
Clientes	205	825	317	1221	3857
Proporção	3,2%	12,9%	4,9%	19,0%	60,0%

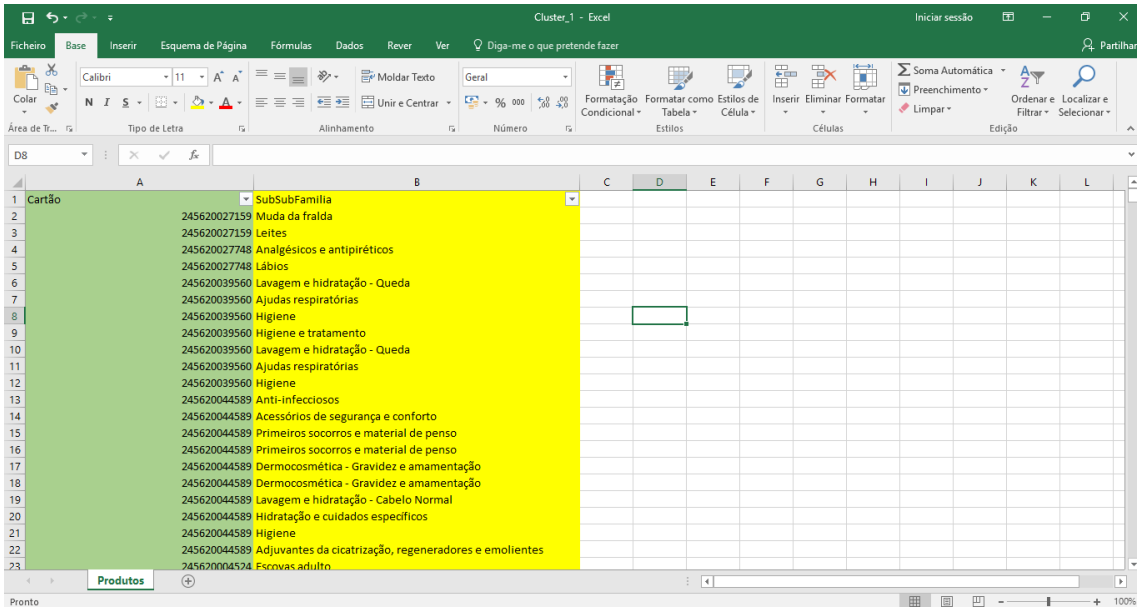


	A	B
1	Cartão	Cluster
2	245620000011	cluster_5
3	245620000019	cluster_5
4	245620000028	cluster_2
5	245620000035	cluster_3
6	245620000036	cluster_5
7	245620000038	cluster_5
8	245620000040	cluster_5
9	245620000051	cluster_5
10	245620000064	cluster_5
11	245620000068	cluster_3
12	245620000069	cluster_5
13	245620000070	cluster_3
14	245620000077	cluster_5
15	245620000078	cluster_3
16	245620000080	cluster_3
17	245620000084	cluster_5
18	245620000086	cluster_5
19	245620000088	cluster_3
20	245620000094	cluster_5
21	245620000096	cluster_2
22	245620000105	cluster_5
23	245620000111	cluster_5

Figura 11: Amostra dos resultados obtidos com a associação do cartão ao *cluster*

De seguida, foi feita uma última agregação de informação. Foi associada cada transacção a cada número de cartão, e consequentemente a cada *cluster*, e por fim, filtrou-se por *clusters* para se obter cada *cluster* separadamente. Como exemplo

temos o resultado do *Cluster_1,* na figura 12, sendo que o mesmo foi feito para os restantes *clusters*.



Cartão	SubSubFamilia
245620027159	Muda da fralda
245620027159	Leites
245620027748	Analgésicos e antipiréticos
245620027748	Lábios
245620039560	Lavagem e hidratação - Queda
245620039560	Ajudas respiratórias
245620039560	Higiene
245620039560	Higiene e tratamento
245620039560	Lavagem e hidratação - Queda
245620039560	Ajudas respiratórias
245620039560	Higiene
245620044589	Anti-infecciosos
245620044589	Acessórios de segurança e conforto
245620044589	Primeiros socorros e material de penso
245620044589	Primeiros socorros e material de penso
245620044589	Dermocosmética - Gravidez e amamentação
245620044589	Dermocosmética - Gravidez e amamentação
245620044589	Lavagem e hidratação - Cabelo Normal
245620044589	Hidratação e cuidados específicos
245620044589	Higiene
245620044589	Adjuvantes da cicatrização, regeneradores e emolientes
245620004524	Escovas adulto

Figura 12: Amostra do cluster 1 com subsubfamília

Isto permitiu mostrar as compras de cada cartão relativamente à "subsubfamília" do produto. O mesmo foi feito para associar os cartões às transacções, mas em vez de ser à "subsubfamília" do que foi comprado, foi associado ao produto em específico da transacção. Como se pode observar na figura 13 que exemplifica o *Cluster*1_Produtos:

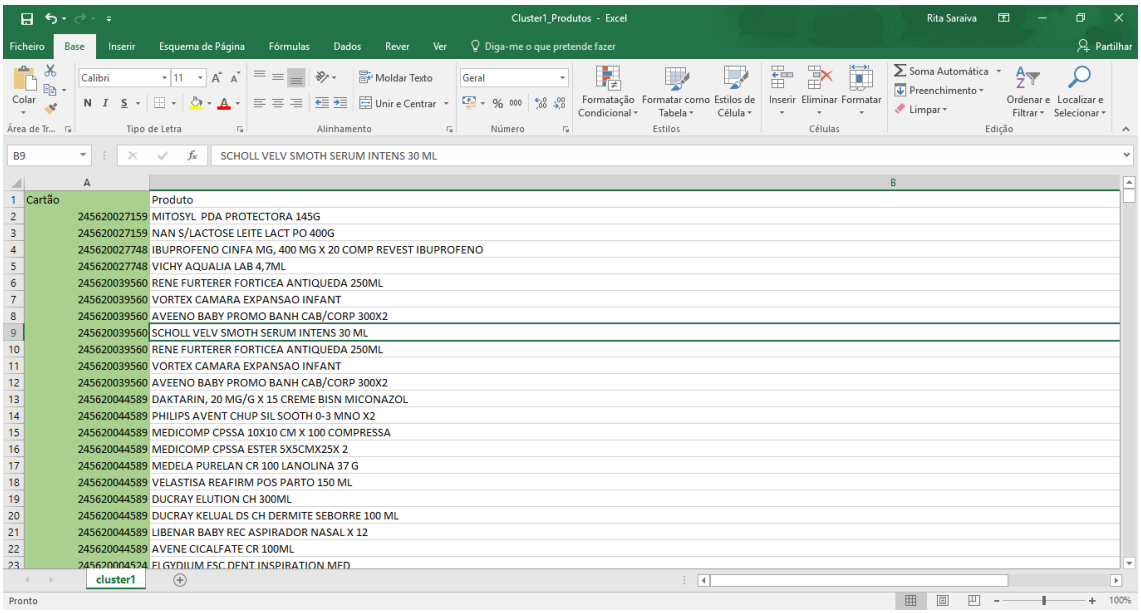


Figura 13: Amostra do *cluster* 1 com os produtos

Este processo permitiu segmentar e identificar cada *cluster*. O resultado foram 5 documentos como o exemplo mostrado anteriormente com a identificação das transacções de cada *cluster*.

5.4 Classificação

Posteriormente, foi feita a caracterização dos *clusters* ou segmentos de clientes, e para tal, recorreu-se ao uso de uma árvore de decisão em que a variável dependente era o *cluster* obtido e as variáveis preditivas a frequência e valor médio gasto por transação. Obteve-se o seguinte modelo representado na figura 14 em se pode identificar os *clusters* da árvore pelas cores na figura 10 apresentada anteriormente. Importa referir que também para este caso, os valores apresentados na árvore de decisão já não são valores normalizados.

Primeiramente, pode-se observar que o *cluster* 3 é caracterizado por mais de 7,5 visitas à farmácia nos três meses em análise e que os restantes *clusters*

representam clientes que foram 7,5 ou menos de 7,5 vezes à farmácia. O *cluster* que representa os clientes que gastaram mais dinheiro por transação, isto é, mais de 71€ na farmácia, é o 1. No *cluster* 5 estão os clientes que visitaram a farmácia menos de 3,5 vezes e gastaram igual ou menos de 28,8€ por cada transacção. O *cluster* 4 representa os clientes que gastam entre 71,79€ e 28,88€ por transacção e que visitam a farmácia igual ou menos de 4,5 vezes. A árvore não foi capaz de discriminar o *cluster* 0 que representa os *outliers*, uma vez que são relativamente poucos, e por assumir que estes *podem* ser incluídos noutros *clusters*, sem grande penalização no erro de classificação dos cartões. No anexo II, é possível ver a forma como se utilizou o software RapidMiner para se obter a árvore de decisão.

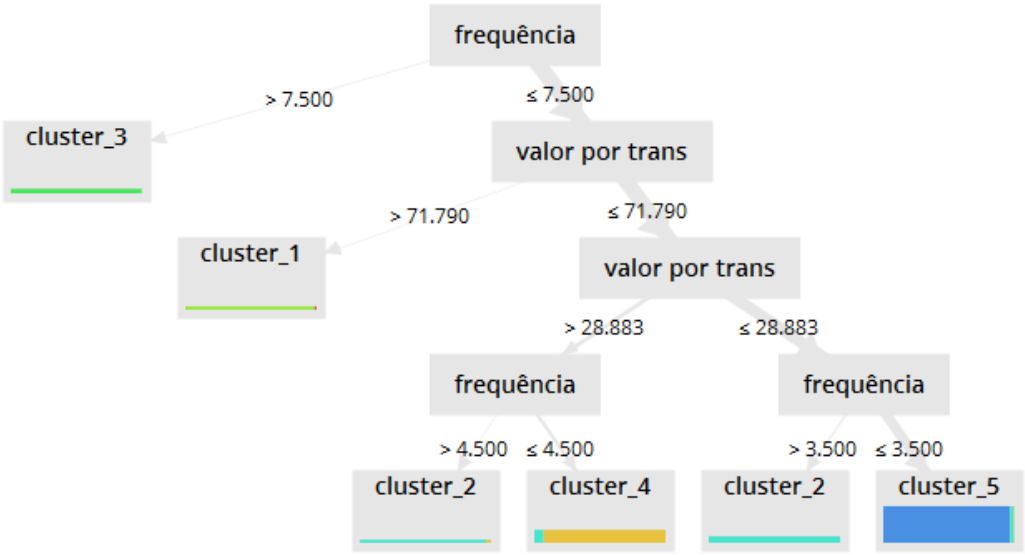


Figura 14: Árvore de decisão

É possível observar que, naturalmente, os *clusters* não são totalmente purificados. O erro de classificação para esta árvore é de 3,71% como podemos ver na tabela 2 em baixo. Este erro é bastante baixo o que aumenta a confiança nesta árvore. O número de clientes corretamente e incorretamente classificados apresenta-se na tabela abaixo.

Tabela 3: Medida de erro

accuracy: 96.29%

	true cluster_5	true cluster_2	true cluster_3	true cluster_1	true cluster_4	true cluster_0	class precision
pred. cluster_5	3857	120	0	0	19	0	96.52%
pred. cluster_2	0	705	0	0	6	0	99.16%
pred. cluster_3	0	0	317	1	0	0	99.69%
pred. cluster_1	0	0	0	204	0	3	98.55%
pred. cluster_4	0	93	0	0	1196	0	92.79%
pred. cluster_0	0	0	0	0	0	0	0.00%
class recall	100.00%	76.80%	100.00%	99.51%	97.95%	0.00%	

5.5 Regras de associação

A fase seguinte desta investigação consistiu na utilização do software RapidMiner para gerar as regras de associação de produtos por cada *cluster* anteriormente referido. Assim, através do algoritmo FP-Growth, explicado no capítulo 3, foi possível correr o programa para cada segmento de clientes e obter as regras de associação para cada um. A fim de se obter regras o mais específicas possível, foram usados os documentos em que as transacções estavam descritas com os produtos em especifico (por exemplo "Farinha 8 Cereais 4 Frutas Nutribén"). Porém, achou-se também interessante correr o programa com as listas das "Subsubfamília" para complementar os dados (por exemplo "Papas e Cereiais").

Foi necessário fazer um conjunto de testes para diferentes valores de suporte para se conseguir obter regras. As regras apresentadas neste documento são as 10 que tinham o lift superior.

Foi possível obter várias regras que apesar de terem um suporte relativamente baixo, o lift é alto, revelando a pertinência da regra. O facto de analisar as regras

por clusters foi relevante, na medida em que foi possível verificar a diferença de regras de segmento para segmento, isto é, as regras identificadas para o cluster 1 são diferentes das regras identificadas para o *cluster* 2. Assim, conseguiu-se obter regras mais específicas para cada cluster, que em princípio virão a se traduzir em ações diferenciadas de marketing mais efetivas. Quanto ao facto de se ter identificado regras com as "subsubfamilia", isto parece não ajudar directamente na definição das estratégias pois são categorias demasiado vagas, acrescentando apenas uma visão geral dos produtos associados.

As regras de associação mais relevantes que se conseguiu obter em cada *cluster* são apresentadas de seguida. No anexo III, é possível ver a forma como se utilizou o software RapidMiner para se obter as regras apresentadas.

5.5.1 *Cluster* 1

Tabela 4: Regras identificadas do *cluster* 1 com subsubfamilias

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Acessórios de segurança e conforto	Higiene, Acessórios de alimentação	0,1073	0,5366	3,7931
2	Higiene, Acessórios de alimentação	Acessórios de segurança e conforto	0,1073	0,7586	3,7931
3	Acessórios de segurança e conforto	Higiene, Fraldas	0,0927	0,4634	3,5185
4	Higiene, Fraldas	Acessórios de segurança e conforto	0,0927	0,7037	3,5185
5	Acessórios de alimentação	Higiene, Acessórios de segurança e conforto	0,1073	0,5238	3,3557
6	Higiene, Acessórios de segurança e conforto	Acessórios de alimentação	0,1073	0,6875	3,3557
7	Acessórios de segurança e conforto	Higiene, Muda da fralda	0,0976	0,4878	3,1250
8	Higiene, Muda da fralda	Acessórios de segurança e conforto	0,0976	0,6250	3,1250
9	Acessórios de alimentação	Acessórios de segurança e conforto	0,1268	0,6190	3,0952
10	Acessórios de segurança e conforto	Acessórios de alimentação	0,1268	0,6341	3,0952

Tabela 5: Regras identificadas do *cluster* 1 com produtos

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Termómetro rectal Chicco	Compressa esterilizada Medicomp, Discos absorv. amamentação Medela	0,0244	1,0000	22,7778
2	Compressa esterilizada Medicomp, Discos absorv. amamentação Medela	Termómetro rectal Chicco	0,0244	0,5556	22,7778
3	Termómetro rectal Chicco	Discos absorv. amamentação Medela	0,0244	1,0000	17,0833
4	Compressa Medicomp, Termómetro rectal Chicco	Discos absorv. amamentação Medela	0,0244	1,0000	17,0833
5	Discos absorv. amamentação Medela	Termómetro rectal Chicco	0,0244	0,4167	17,0833
6	Discos absorv. amamentação Medela	Compressa esterilizada Medicomp, Termómetro rectal Chicco	0,0244	0,4167	17,0833
7	Álcool 70° 250 ml AGA	Compressa esterilizada Medicomp	0,0293	1,0000	13,6667
8	Termómetro rectal Chicco	Compressa esterilizada Medicomp	0,0244	1,0000	13,6667
9	Compressa Medicomp, Discos absorv. Amamentação Medela	Compressa esterilizada Medicomp	0,0293	1,0000	13,6667
10	Compressa esterilizada Medicomp	Álcool 70° 250 ml AGA	0,0293	0,4000	13,6667

5.5.2 *Cluster* 2

Tabela 6: Regras identificadas do *cluster* 2 para subsubfamilias

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Acessórios de segurança e conforto	Acessórios de alimentação	0,1009	0,4381	2,3503
2	Acessórios de alimentação	Acessórios de segurança e conforto	0,1009	0,5412	2,3503
3	Papas e cereais	Leites	0,1140	0,6933	2,1219
4	Leites	Acessórios de alimentação	0,1217	0,3725	1,9983
5	Acessórios de alimentação	Leites	0,1217	0,6529	1,9983
6	Acessórios de segurança e conforto	Muda da fralda	0,0976	0,4238	1,9821
7	Muda da fralda	Acessórios de segurança e conforto	0,0976	0,4564	1,9821
8	Higiene	Muda da fralda	0,1162	0,4109	1,9215
9	Muda da fralda	Higiene	0,1162	0,5436	1,9215
10	Leites	Acessórios de segurança e conforto	0,1239	0,3792	1,6468
11	Acessórios de segurança e conforto	Leites	0,1239	0,5381	1,6468

Tabela 7: Regras identificadas do *cluster* 2 com produtos

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Farinha Multicereais e mel Bolacha Maria Nutribén	Farinha 8 Cereais Mel Nutribén	0,0132	0,6667	20,2667
2	Farinha 8 Cereais Mel Nutribén	Farinha Multicereais e mel Bolacha Maria Nutribén	0,0132	0,4000	20,2667
3	Farinha 8 Cereais Mel Nutribén	Farinha 8 Cereais 4 Frutas Nutribén	0,0186	0,5667	16,6710
4	Farinha 8 Cereais 4 Frutas Nutribén	Farinha 8 Cereais Mel Nutribén	0,0186	0,5484	16,6710
5	Farinha Multicereais e mel Bolacha Maria Nutribén	Farinha 8 Cereais 4 Frutas Nutribén	0,0110	0,5556	16,3441
6	Farinha 8 Cereais 4 Frutas Nutribén	Farinha Multicereais e mel Bolacha Maria Nutribén	0,0110	0,3226	16,3441
7	Fraldas Comfort Prematuro Libero	Fraldas Baby Soft Newborn Libero	0,0121	0,6875	12,5400
8	Fraldas Baby Soft Newborn Libero	Fraldas Comfort Prematuro Libero	0,0121	0,2200	12,5400
9	Creme protector de mamilos Medela	Fraldas Baby Soft Newborn Libero	0,0154	0,6667	12,1600
10	Fraldas Baby Soft Newborn Libero	Creme protector de mamilos Medela	0,0154	0,2800	12,1600

5.5.3 *Cluster* 3

Tabela 8: Regras identificadas do *cluster* 3 com subsubfamilias

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Acessórios de segurança e conforto	Higiene, Acessórios de alimentação	0,2145	0,4564	1,6075
2	Higiene, Acessórios de alimentação	Acessórios de segurança e conforto	0,2145	0,7556	1,6075
3	Acessórios de alimentação	Leites, Acessórios de segurança e conforto	0,2524	0,5714	1,5482
4	Leites, Acessórios de segurança e conforto	Acessórios de alimentação	0,2524	0,6838	1,5482
5	Acessórios de alimentação	Higiene, Acessórios de segurança e conforto	0,2145	0,4857	1,5245
6	Higiene, Acessórios de segurança e conforto	Acessórios de alimentação	0,2145	0,6733	1,5245

7	Muda da fralda	Fraldas	0,2366	0,4967	1,4715
8	Fraldas	Muda da fralda	0,2366	0,7009	1,4715
9	Acessórios de alimentação	Leites, Muda da fralda	0,2334	0,5286	1,4570
10	Leites, Muda da fralda	Acessórios de alimentação	0,2334	0,6435	1,4570

Tabela 9: Regras identificadas do *cluster* 3 com produtos

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Água purificada Alvita 1L	AirProject Aerossol de Ultra-sons Pic	0,0203	0,6667	19,7000
2	AirProject Aerossol de Ultra-sons Pic	Água purificada Alvita 1L	0,0203	0,6000	19,7000
3	Fraldas Comfort Fit Tamanho 3 Libero	Fraldas Comfort Fit Tamanho 4 Libero	0,0254	0,3571	4,3076
4	Fraldas Comfort Fit Tamanho 4 Libero	Fraldas Comfort Fit Tamanho 3 Libero	0,0254	0,3061	4,3076
5	Gotas Cólicas BioGaia	Gotas Cólicas Aero-OM	0,0203	0,4138	4,2164
6	Gotas Cólicas Aero-OM	Gotas Cólicas BioGaia	0,0203	0,2069	4,2164
7	Creme bebé Uriage	Compressa Medicomp	0,0203	0,6316	3,9709
8	Discos absorv. amamentação Medela	Compressa Medicomp	0,0203	0,5455	3,4294
9	Compressa Esterilizada Medicomp	Compressa Medicomp	0,0406	0,5333	3,3532
10	Recargas descartáveis limpeza nasal bebé Rhinomer	Compressa Medicomp	0,0271	0,5333	3,3532

5.5.4 *Cluster* 4

Tabela 10: Regras identificadas do *cluster* 4 com subsubfamilias

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Leites	Papas e cereais	0,0345	0,2857	5,4375
2	Papas e cereais	Leites	0,0345	0,6563	5,4375
3	Fraldas	Primeiros socorros e material de penso	0,0230	0,2917	4,2801
4	Primeiros socorros e material de penso	Fraldas	0,0230	0,3373	4,2801
5	Muda da fralda	Fraldas	0,0320	0,3305	4,1933
6	Fraldas	Muda da fralda	0,0320	0,4063	4,1933
7	Acessórios de segurança e conforto	Acessórios de alimentação	0,0304	0,2960	3,7555

8	Acessórios de alimentação	Acessórios de segurança e conforto	0,0304	0,3854	3,7555
9	Rosto	Anti-envelhecimento	0,0205	0,2717	3,6775
10	Anti-envelhecimento	Rosto	0,0205	0,2778	3,6775

Tabela 11: Regras identificadas do *cluster* 4 com produtos

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Compressa Esterilizada Medicomp	Compressa Medicomp	0,012315271	0,53571429	13,3163

5.5.5 *Cluster* 5

Tabela 12: Regras identificadas do *cluster* 5 com subsubfamilias

Regra	Antecedente	Consequente	Suporte	Confiança	Lift
1	Gripe e constipações	Dor de garganta e rouquidão	0,0101	0,1418	2,4661
2	Dor de garganta e rouquidão	Gripe e constipações	0,0101	0,1765	2,4661
3	Gripe e constipações	Analgésicos e antipiréticos	0,0109	0,1527	1,2127
4	Expectorantes	Analgésicos e antipiréticos	0,0109	0,1505	1,1953

Não foram encontradas regras de associação neste *cluster* com os produtos especificos.

A análise das regras de associação apresentadas revela que os produtos e as categorias de produtos que dominam as regras são referentes a bebés e à maternidade. Isto prende-se como facto desta categoria de produtos estar presente em muitos cabazes de compra dos clientes da farmácia em causa.

Importa destacar que os resultados das regras de associação devem ser interpretados com cuidado. Estas regras não implicam causalidade, mas apenas implicam uma forte relação de coocorrência entre os produtos.

5.6 Estratégias de *cross-selling*

As regras de associação de produtos que se identificaram podem ser utilizadas na definição de estratégias de *cross-selling* por parte da farmácia em estudo. É importante referir que as estratégias de *cross-selling* podem variar de segmento para segmento.

A título de exemplo, as estratégias de *cross-selling* a adoptar pela farmácia podem passar pelas seguintes estratégias:

1. Relativamente à regra nº 7 e 8 do *cluster* 2, seria interessante elaborar pacotes complementares para o *cluster* em questão em que a venda de fraldas Libero de recém-nascido "soft" venha acompanhada de uma amostra de fraldas da mesma marca "comfort";
2. Utilizar a mesma estratégia referida em 1 para as regras nº 1 a 6 do *cluster* 2 em que se poderia criar pacotes com os diferentes tipos de farinha de bebé juntos;
3. Para a regra nº 3 e 4 do *cluster* 3, poderia ser relevante elaborar pacotes complementares em que a venda de fraldas de tamanho 3 e 4 sejam compradas em simultâneo, isto é, ao comprar fraldas de tamanho 3, teria uma percentagem de desconto na compra de fraldas de tamanho 4;
4. Ainda referente às regras referidas no ponto 3 (referentes ao *cluster* 3), para se promover a recompra, através de comunicação por email ou mensagem, após determinado tempo da compra de fraldas tamanho 3, alertar para uma promoção de fraldas tamanho 4;
5. Utilizando a informação das várias regras do *cluster* 1 em que são associados os discos de amamentação, os termómetros rectais para bebés e cremes de protecção de mamilos, poderia ser promissor expor estes produtos juntos uns dos outros;

6. Utilizar a mesma estratégia referida em 1, isto é, criar pacotes conjuntos para as gotas das duas marcas para as cólicas dos bebés referentes às regras nº 5 e 6 do *cluster* 3;
7. Relativamente às regras nº 1 e 2 do *cluster* 3, seria interessante criar pacotes de três águas purificadas para os aerossóis para se vender mais que um quando se vender o aparelho de aerossóis;
8. De um modo geral, para aumentar as vendas de outras áreas, como por exemplo, produtos de cosmética, colocar alguns produtos mais pequenos de cosmética para mulheres, ao pé dos produtos de bebé.
9. Relativamente à regra do *cluster* 4, seria interessante verificar quem comprou as Compressas Esterilizadas da Medicomp (antecedente) e não comprou as Compressas sem ser esterilizadas da Medicomp (consequente) e promover este último através do envio de vouchers.

Capítulo 6
6. Conclusão

Neste último capitulo são apresentadas as conclusões obtidas com o estudo deste trabalho de projecto aplicado à Farmácia Gaia Jardim. Esta conclusão está dividida em três pontos: a síntese das conclusões desta investigação, as limitações ao longo da elaboração deste TFM e as sugestões para possíveis investigações futuras.

Actualmente, o marketing tem vindo a evoluir no sentido de analisar toda a cadeia de interacções das organizações com os clientes. Nomeadamente, através da análise dos dados gerados por essas mesmas interacções que são armazenadas em bases de dados.

De um modo geral, ao longo da realização deste trabalho final de semestre foi possível perceber a importância e a pertinência dos papéis dos modelos analíticos para o apoio às estratégias de marketing. Esta importância foi observada em várias fases do projecto. Em primeiro lugar, a aplicação de algoritmos de *clustering* permitiu a obtenção de segmentos de clientes com comportamentos semelhantes, e assim, facilitou a oportunidade para a farmácia desenvolver estratégias e acções de marketing diferenciadas mais orientadas a cada segmento para melhorar os seus resultados. De seguida, os algoritmos de classificação permitem à farmácia fazer uma afectação dos seus novos clientes a cada segmento, sem haver necessidade de voltar a desenvolver a análise de *clustering*. E por último, os algoritmos de associação permitiram identificar as regras de associação entre os produtos da farmácia que podem sustentar a implementação das estratégias de marketing para um dos segmentos previamente identificados.

Durante a realização deste trabalho encontraram-se algumas limitações que o condicionaram. O facto de os dados utilizados serem de apenas três meses é uma

limitação, uma vez que este período de tempo é relativamente curto para analisar comportamentos recorrentes dos clientes. Acredita-se quequanto maior fosse o intervalo de tempo dos dados recolhidos, mais certezas haveria nos resultados encontrados. Ainda relativamente aos dados recolhidos, a estrutura hierárquica dos produtos apresenta algumas limitações, na medida em que a forma como os produtos estão categorizados (as "famílias" e "subsubfamilias") é muito genérica para se obter regras concretas que apoiem uma concepção de acções especificas de marketing. E para além disto, por vezes essa categorização não está dividida de uma forma intuitiva. Por último nas limitações referentes aos dados recolhidos, estes tinham informação em falta referente à descrição dos clientes, o que fez com que se tivesse que eliminar certos dados, e assim, fazer uma caracterização dos clientes menos completa. A última limitação encontrada é referente ao facto de as regras de associação estarem a serem contaminadas pelos produtos mais frequentes. A desconsideração desses produtos poderia vir a expor outras relações entre produtos que não são visíveis neste trabalho.

Assim, em investigações futuras espera-se, em primeiro lugar que seja possível a minimização das limitações encontradas neste trabalho. De seguida, poderia ser interessante identificar regras relativas a outros produtos menos frequentes. A aplicação e avaliação do impacto das estratégias de *cross-selling* definidas também poderia ser uma sugestão de uma investigação futura, assim como a aplicação do que foi feito neste trabalho a qualquer retalhista interessado.

7. Bibliografia

Aggarwal, C. C., Procopiuc, C., & Yu, P. S. (2002). Finding localized associations in market basket data. *IEEE Transactions on Knowledge and Data Engineering*, *14*(1), 51–62.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association in Large Databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 207–216.

Ahmed, S. R. (2004). Applications of data mining in retail business. In *International Conference on Information Technology: Coding Computing, ITCC* (Vol. 2, pp. 455–459).

Ahn, H., Kim, K., & Han, I. (2006). Hybrid genetic algorithms and case-based reasoning systems for customer classification. *Expert Systems*, *23*(3), 127–144.

Alhaiou, T. A. (2011). A Study on the Relationship between E-CRM Features and E-Loyalty: the case in UK. *PhD Thesis Brunel University*, (April), 238.

Anderson, J. L., Jolly, L. D., & Fairhurst, A. E. (2007). Customer relationship management in retailing: A content analysis of retail trade journals. *Journal of Retailing and Consumer Services*, *14*(6), 394–399.

Ansell, J., Harrison, T., & Archibald, T. (2007). Identifying cross-selling opportunities, using lifestyle segmentation and survival analysis. *Marketing Intelligence & Planning*, *25*(4), 394–410.

Au, W. H., Chan, C. C., & Yao, X. (2003). A novel evolutionary data mining algorithm with applications to churn prediction. *IEEE Transactions on Evolutionary Computation*, *7*(6), 532–544.

Bae, S. M., Park, S. C., & Ha, S. H. (2003). Fuzzy Web Ad Selector Based on Web Usage Mining. *IEEE Intelligent Systems*.

Baesens, B., Verstraeten, G., Van den Poel, D., Egmont-Petersen, M., Van Kenhove, P., & Vanthienen, J. (2004). Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers. *European Journal of Operational Research*, *156*(2), 508–523.

Baesens, B., Viaene, S., Van Den Poel, D., Vanthienen, J., & Dedene, G. (2002). Bayesian neural network learning for repeat purchase modelling in direct marketing. *European Journal of Operational Research*, *138*(1), 191–211.

Bailey, C. R., Baines, P. R., Wilson, H., & Clark, M. (2009). Segmentation and customer insight in contemporary services marketing practice: why grouping customers is no longer enough. *Journal of Marketing Management*, *25*(3), 227–252.

Berry, L. L. (2002). Relationship Marketing of Services Perspectives from 1983 and 2000. *Journal of Relationship Marketing*, *1*(1), 59–77.

Borak, J. S., & Strahler, A. H. (1999). Feature selection and land cover classification of a modis-like data set for a semiarid environment. *International Journal of Remote Sensing*, *20*(5), 919–938.

Bose, I., & Chen, X. (2009). Hybrid Models Using Unsupervised Clustering for Prediction of Customer Churn. *Journal of Organizational Computing and Electronic Commerce*, *19*(2), 133–151.

Bose, R. (2002). Customer relationship management: key components for IT success. *Industrial Management & Data Systems*, *102*(2), 89–97.

Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. *Communications of the ACM*, *39*(11), 42–48.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brijs, T., Swinnen, G., Vanhoof, K., & Wets, G. (2004). Building an association rules framework to improve product assortment decisions. *Data Mining and Knowledge Discovery*, *8*(1), 7–23.

Buckinx, W., Moons, E., Van den Poel, D., & Wets, G. (2004). Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, *26*(4), 509–518.

Bull, C. (2003). Strategic issues in customer relationship management (CRM) implementation. *Business Process Management Journal*, *9*(5), 592–602.

Chang, S. E., Changchien, S. W., & Huang, R. H. (2006). Assessing users' product-specific knowledge for personalization in electronic commerce. *Expert Systems with Applications*, *30*(4), 682–693.

Chen, M. C., Chiu, A. L., & Chang, H. H. (2005). Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*.

Chen, Y. L., Hsu, C. L., & Chou, S. C. (2003). Constructing a multi-valued and multi-labeled decision tree. *Expert Systems with Applications*.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, *40*(2), 339–354.

Cheung, K. W., Kwok, J. T., Law, M. H., & Tsui, K. C. (2003). Mining customer product ratings for personalized marketing. *Decision Support Systems*, *35*(2), 231–243.

Chiu, C. (2002). A case-based customer classification approach for direct marketing. *Expert Systems with Applications*, *22*(2), 163–168.

Cho, Y. Bin, Cho, Y. H., & Kim, S. H. (2005). Mining changes in customer buying behavior for collaborative recommendations. *Expert Systems with Applications*, *28*(2), 359–369.

Coviello, N. E., Brodie, R. J., & Munro, H. J. (1997). Understanding Contemporary Marketing: Development of a Classification Scheme. *Journal of Marketing Management*, *13*(1990), 501–522.

Curry, A., & Kkolou, E. (2004). Evaluating CRM to contribute to TQM improvement – a cross-case comparison. *The TQM Magazine*, *16*(5), 314–324.

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227.

Drew, J. H., Mani, D. R., Betz, A. L., & Datta, P. (2001). With Statistical and Data-Mining Techniques. *Journal of Service Research*, *3*(3), 205–219.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, *21*(3), 768–769.

Fred, a. L. N., & Jain, a. K. (2002). Data clustering using evidence accumulation. *Object Recognition Supported by User Interaction for Service Robots*, *4*, 276–280 vol.4.

Friedl, M. a. M. A., & Brodley, C. E. C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(3), 399–409.

Gama, J., Carvalho, A. P. de L., Faceli, K., Lorena, A. C., & Oliveira, M. (2017). *Extração de Conhecimento de Dados* (3ª). Edições Sílabo.

Giraud-Carrier, C., & Povel, O. (2003). Characterising Data Mining Software. *Intelligent Data Analysis*, *7*(3), 181–192.

Gronroons, C. (1982). "Strategic Management and Marketing in The Service Sector. *Swidish School of Economics and Business Administration*, *17*(19), 15.

Gronroos, C. (1990). Relationship approach to marketing in service contexts: The marketing and organizational behavior interface. *Journal of Business Research*, *20*(1), 3–11.

Gummesson, E. (1987). The New Marketing - Developing Long-Term Interactive Relationships. *Long Range Planning*, *20*(4), 10–20.

Ha, S. H. (2006). Digital Content Recommender on the Internet. *IEEE Intelligent Systems*, *21*(2), 70–77.

Han, S. H., Lu, S. X., & Leung, S. C. H. (2012). Segmentation of telecom customers based on customer value by decision tree model. *Expert Systems with Applications*, *39*(4), 3964–3973.

Harker, M. J. (1999). Relationship marketing defined? An examination of current relationship marketing definitions. *Marketing Intelligence & Planning*, *17*(1), 13–20.

He, Z., Xu, X., Deng, S., & Ma, R. (2005). Mining action rules from scratch. *Expert Systems with Applications*, *29*(3), 691–699.

Hollander, S. C., Rassuli, K. M., Jones, D. G. B., & Dix, L. F. (2005). Periodization in Marketing History. *Journal of Macromarketing*, *25*(1), 32–41.

Huang, J. J., Tzeng, G. H., & Ong, C. S. (2007). Marketing segmentation using support vector clustering. *Expert Systems with Applications*, *32*(2), 313–317.

Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, *2*(3), 283–304.

Hung, S.-Y., Yen, D. C., & Wang, H. (2006). Applying data mining to telecom churn. *Expert Systems with Applications*, *31*, 515–524.

Jiang, T., & Tuzhilin, A. (2006). Segmenting customers from population to individuals: Does 1-to-1 keep your customers forever? *IEEE Transactions on Knowledge and Data Engineering*, *18*(10), 1297–1311.

Jiao, J., Zhang, Y., & Helander, M. (2006). A Kansei mining system for affective design. *Expert Systems with Applications*, *30*(4), 658–673.

Kamakura, W. A., Wedel, M., de Rosa, F., & Mazzon, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing*, *20*(1), 45–65.

Kim, Y. S., & Street, W. N. (2004). An intelligent system for customer targeting: A data mining approach. *Decision Support Systems*, *37*(2), 215–228.

Kracklauer, A., Passenheim, O., & Seifert, D. (2001). Mutual customer approach: how industry and trade are executing collaborative customer relationship management. *International Journal of Retail & Distribution Management*, *29*(12), 515–519.

Kubat, M., Hafez, A., Raghavan, V. V., Lekkala, J. R., & Chen, W. K. (2003). Itemset Trees for Targeted Association Querying. *IEEE Transactions on Knowledge and Data Engineering*, *15*(6), 1522–1534.

Kumar, V., & Shah, D. (2004). Building and sustaining profitable customer loyalty for the 21st century, *80*, 317–330.

Lamberti, L., & Noci, G. (2010). Marketing strategy and marketing performance measurement system: Exploring the relationship. *European Management Journal*, *28*(2), 139–152.

Larivière, B., & Van Den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Systems with Applications*, *29*(2), 472–484.

Lejeune, M. A. P. M. (2001). Measuring the impact of data mining on churn management. *Internet Research*, *11*(5), 375–387.

Levitt, T. (1981). Marketing Intangible Products and Product Intangibles. *Cornell Hotel and Restaurant Administration Quarterly*, *22*(2), 37–44.

Li, S., Sun, B., & Montgomery, A. L. (2011). Cross-Selling the Right Product to the Right Customer at the Right Time. *Journal of Marketing Research*, *48*(4), 683–700.

Liao, S. H., & Chen, Y. J. (2004). Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications*, *27*(4), 521–532.

Ling, R., & Yen, D. D. C. (2001). Customer relationship management: An analysis framework and implementation strategies. *Journal of Computer Information Systems*, *41*(3), 82–97.

Linoff, G. S., Berry, M. J., & Michael J. A. Bery, G. S. L. (2004). *Data Mining Techniques for Marketing, Sales and Customer Relationship Management*. *Management*.

Liu, C. (2004). A database approach to cross selling in the banking industry: Practices, strategies, *11*, 216–234.

Looney, C. G. (2002). Interactive clustering and merging with a new fuzzy expected value. *Pattern Recognition*, *35*(11), 2413–2423.

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297).

Martins, D. A. P. (2012). Relatório de estágio, 1–22.

Miguéis, V. L., Camanho, A. S., & Falcão, J. (2012). Expert Systems with Applications Customer data mining for lifestyle segmentation, *39*, 9359–9366.

Mitra, S., Pal, S. K., & Mitra, P. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14.

Nairn, A. (2002). CRM: Helpful or full of hype? *Journal of Database Marketing*, *9*(4), 376–382.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2 PART 2), 2592–2602.

Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2 PART 2), 2592–2602.

Ngai, E. W. T. (2005). Customer relationship management research (1992-2002): An academic literature review and classification. *Marketing Intelligence & Planning*, *23*(6), 582–605.

Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. *Library of Congress*.

Payne, A., & Frow, P. (2005). A strategic framework for customer relationship management. *Journal of Marketing*, *69*(4), 167–176.

Payne, A., & Frow, P. (2005). A Strategic Framework for Customer Relationship Management. *The Journal of Marketing*, *69*(4), 167–176.

Peppers, D., & Rogers, M. (2011). *Managing Customer Relationships: A Strategic Framwork*. *John Wiley & Sons*.

Prinzie, A., & Van Den Poel, D. (2006). Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. *European Journal of Operational Research*, *170*(3), 710–734.

Prinzie, A., & Van Den Poel, D. (2005). Constrained optimization of data-mining problems to improve model performance: A direct-marketing application. *Expert Systems with Applications*, *29*(3), 630–640.

Prinzie, A., & Van den Poel, D. (2006). Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM. *Decision Support Systems*, *42*(2), 508–526.

Romano, N. C., & Fjermestad, J. (2003). Electronic commerce customer relationship management: A research agenda. *Information Technology and Management*, *4*, 233–258.

Rosset, S., Neumann, E., Eick, U. R. I., & Vatnik, N. (2003). Customer Lifetime Value Models for Decision Support. *Data Mining and Knowledge Discovery*, *7*(3), 321–339.

Ryals, L., & Payne, A. (2001). Customer relationship management in financial services: towards information-enabled relationship marketing. *Journal of Strategic Marketing*, *9*(1), 3–27.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems*, *31*(1), 127–137.

Sheth, J. N., & Parvatiyar, A. (2001). Customer Relationship Management: Emerging Practice, Process, and Discipline. *Journal of Economic and Social Research*, 1–34.

Tibshirani, R., Walther, G., & Hastie, T. (2000). Estimating the Number of Clusters in a Dataset via the Gap Statistic. *Journal of the Royal Statistical Society, Series B*.

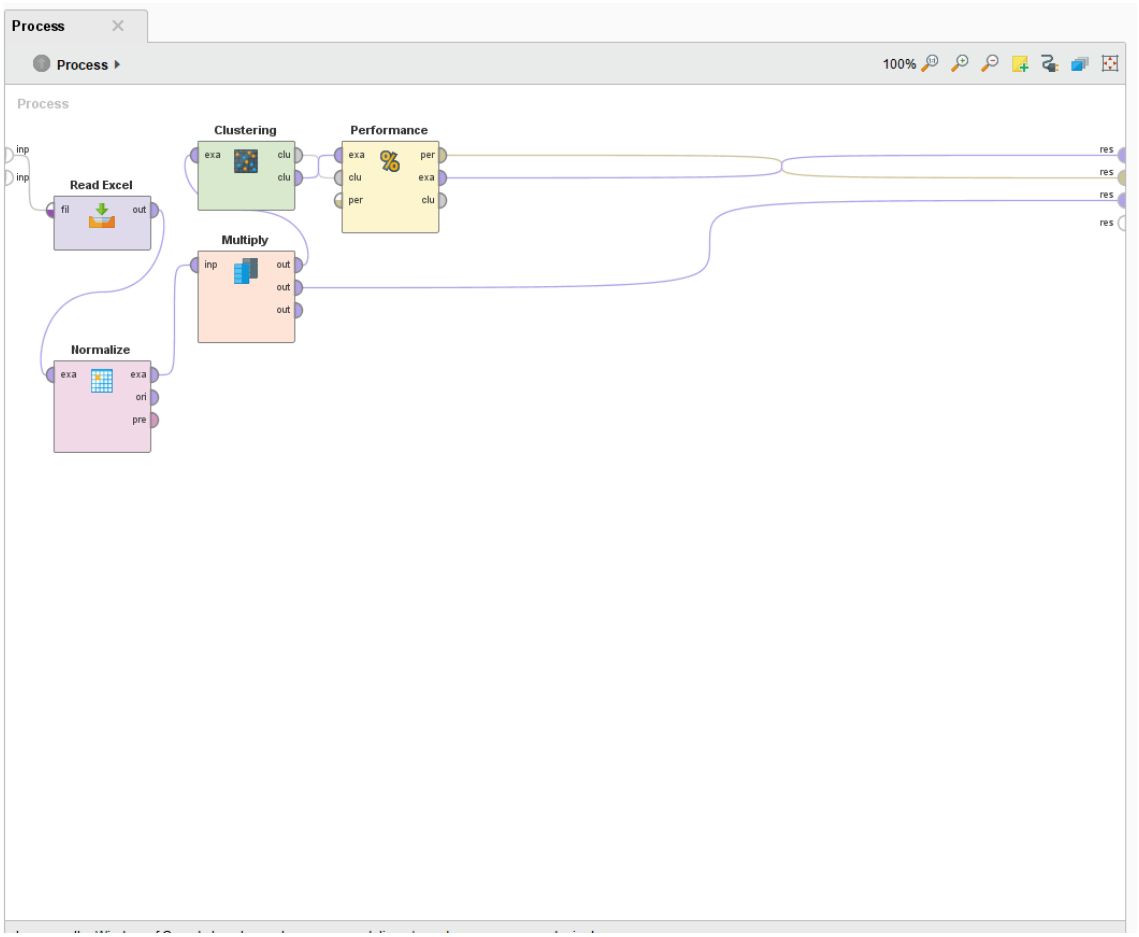
Vera, L. (2012). Management in Retailing Supported by Data Mining Techniques.

Woo, J. Y., Bae, S. M., & Park, S. C. (2005). Visualization method for customer targeting using customer map. *Expert Systems with Applications*.

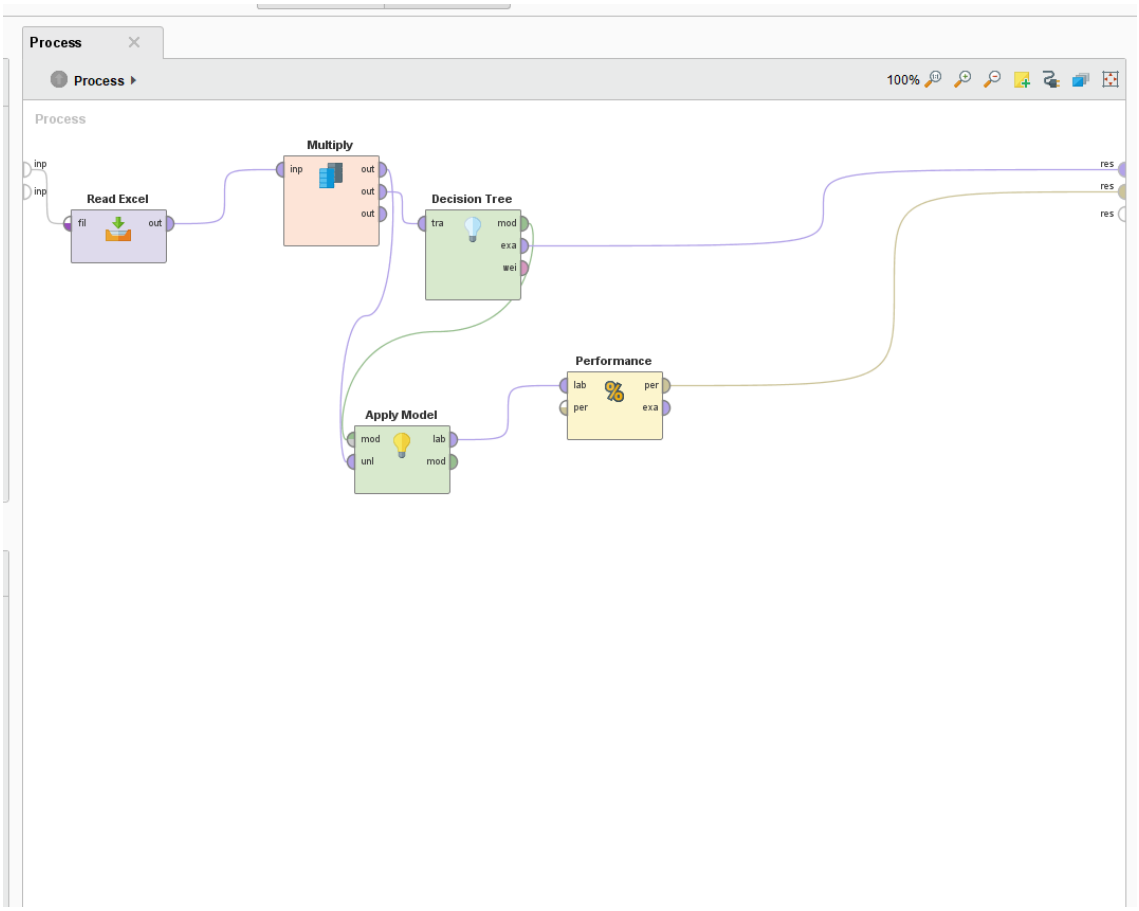
Yau, C., & Holmes, C. (2011). Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. *Bayesian Analysis*, *6*(2), 329–352.

Anexos

Anexo I – *Clustering* no RapidMiner



Anexo II – Classificação no RapidMiner



Anexo III – Associação no RapidMiner