NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Técnicas de Data mining: análise ao cesto de compras e segmentação de clientes de um supermercado

Ana Catarina Pereira Claudino Santos

Trabalho de Projeto apresentado como requisito parcial para obtenção do grau de Mestre em Gestão de Informação

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

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

TÉCNICAS DE DATA MINING: ANÁLISE AO CESTO DE COMPRAS E SEGMENTAÇÃO DE CLIENTES DE UM SUPERMERCADO

por

Ana Catarina Pereira Claudino Santos

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

Orientador/Coorientador: Frederico Miguel Campos Cruz Ribeiro de Jesus

Agosto de 2020

AGRADECIMENTOS

O meu profundo agradecimento a todos os meus familiares e amigos por todo o afeto, motivação e carinho que me deram durante a realização deste projeto e sonho.

Em primeiro lugar gostaria de agradecer ao meu orientador Professor Frederico Cruz Jesus, pelos seus conselhos, sempre fundamentados em conhecimento, pela sua paciência e rigor profissional.

Ao meu marido José Santos gostaria de agradecer o seu apoio incondicional, a sua dedicação e compreensão. A sua motivação e força foram sempre constantes nesta etapa da minha vida.

Por último, quero agradecer aos meus colegas de turma e colegas de trabalho pelas partilhas de informação e conhecimento que muito me ajudaram na concretização deste projeto.

RESUMO

Os utilizadores estão mais móveis, e utilizam cada vez mais dispositivos móveis para aceder a aplicações. Essa nova realidade contribuiu para o crescimento e uso massivo de aplicações móveis. O uso exponencial de aplicações levou a que muitos negócios investissem nessa ferramenta, não só para vender os seus produtos, mas também para garantir uma maior proximidade entre a empresa e os seus consumidores. A ideia é estar presente em qualquer altura e em qualquer lugar.

O presente estudo utilizou técnicas de *data mining*, tais como, regras de associação e de *clustering* para descobrir os tipos de clientes e os hábitos de consumo dos consumidores de uma aplicação móvel de supermercado. Para o estudo das regras de associação, realizaram-se dois modelos. O primeiro serviu para descobrir as associações entre artigos, e o segundo para identificar as associações entre departamentos. Na análise de *clustering*, realizaram-se dois tipos distintos de segmentação. O primeiro foi realizado com o objetivo de compreender o comportamento dos clientes em termos de consumo, através da medição de quanto eles gastaram por departamento. O segundo foi construído com o objetivo de conhecer o valor dos clientes para a empresa.

Os resultados das associações entre departamentos serviram para elaborar uma proposta de reordenação dos corredores e lineares da *app*, enquanto que os resultados da análise de segmentação foram úteis para redefinir algumas estratégias de comunicação. Outras ideias e recomendações fundamentadas em conhecimento adquirido foram referenciadas neste estudo, por forma a que a empresa possa tomar decisões de negócio mais fundamentadas.

PALAVRAS-CHAVE

Data mining, regras de associação, clustering, cesto de compras, SAS Enterprise Miner, aplicações móveis

ABSTRACT

Users are more mobile and use mobile devices more than ever to access applications. This new reality has contributed to the growth and massive use of mobile applications. The exponential use of applications made many businesses invest in this tool, not only to sell their products, but also to guarantee greater proximity between the company and the consumer. The idea is to be present anytime, anywhere.

This project used data mining techniques, such as association rules and clustering to discover types of clients, and their consumption habits in a supermarket mobile application. Two models were done, regarding the association rules. The first to find the associations between articles, and the second to identify the associations between departments. In the clustering analysis, two types of segmentation were performed. The first had the purpose of understanding the behavior of consumers, regarding their amount spent in various departments. The second was built in order to get to know the value of each type of customer to the company.

The results of the associative models were used for the elaboration of a restructure plan for the supermarket corridors and aisles in the mobile app, while the results of the segmentation analysis were useful to redefine communication strategies towards customers. Other ideas and suggestions founded in new knowledge were reference in this project, so that the company may take better informed business decisions.

KEYWORDS

Data mining, association rules, clustering, market basket analysis, SAS Enterprise Miner, mobile applications

INDEX

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Nos últimos anos, o *data mining* tem vindo a ganhar grande popularidade no ambiente organizacional. A sua importância deve-se sobretudo ao desenvolvimento e crescimento tecnológico, ao aumento explosivo da produção de informação e à necessidade de armazenar, tratar e compreender esses grandes volumes de dados. Neste caso em particular, estes são frequentemente transações de vendas, stock de produtos, descrição de produtos, campanhas de Marketing, promoções de produtos, dados comerciais, comentários de clientes, etc.

Data mining é um conjunto de técnicas analíticas que serve para transformar dados em informação (Linoff & Berry, 2011). Além disso, é um processo empresarial utilizado para explorar grandes volumes de dados, com o objetivo de descobrir padrões e regras relevantes. Assim, para a tomada de decisões de negócio inteligentes, torna-se fundamental que as empresas consigam extrair conhecimento relevante através do tratamento de dados, para posteriormente justificar decisões de negócio tais como: a penetração em novos mercados, lançamento de novos produtos ou serviços, otimização de processos, ou concorrer em mercados cada vez mais competitivos e orientados por decisões fundamentadas em dados. O processo de *Knowledge Discovery in Databases* (KDD), ou extração do conhecimento de grandes bases de dados, é uma das metodologias mais utilizadas e populares. O KDD é um processo simples que inclui as seguintes fases: seleção, pré-processamento, transformação, *data mining*, interpretação e avaliação de dados. Portanto, o *data mining* pertence a uma das etapas do processo de KDD e tem como funcionalidade analisar volumes extensos de dados recorrendo à matemática, à probabilidade e à identificação de padrões, para no final obter informações sobre novos padrões, tendências e associações. O *data mining* utiliza modelos descritivos, ou não supervisionados, e preditivos, ou supervisionados, para fundamentar decisões de negócio inteligentes. Este último modelo utiliza um conjunto de dados, ou registos, que são categorizados de acordo com uma variável pré-definida. As regras de associação podem ser utilizadas em ambos os modelos (Pimenta, Ribeiro, Sá, & Belfo, 2018).

Neste projeto, serão abordadas as regras de associação através da análise de dados de um cesto de compras. Estas têm como função avaliar se a um conjunto de itens numa base de dados está associada a presença de um outro conjunto de itens (Agrawal & Srikant, 1994). Além disso, o carácter simplista das regras de associação torna-as muito apreciada pelos retalhistas, não só porque explicam o grau de frequência com que os produtos estão em conjunto, mas também porque preveem quando voltarão a estar novamente em conjunto. A sua finalidade é analisar os padrões de comportamento de consumo dos clientes, e identificar quais os artigos ou conjuntos de produtos frequentemente comprados (Raeder & Chawla, 2011). Os retalhistas e fabricantes de bens de consumo utilizam a análise do cesto de compras, principalmente para estudar o seu tipo de conteúdo, ou seja, para conhecer quando e quais os artigos que os clientes compram (Linoff & Berry, 2011).

Assim, este estudo visa utilizar técnicas de *data mining*, nomeadamente regras de associação através da análise do cesto de compras, para analisar o comportamento de consumo dos clientes de uma aplicação de supermercado. Visa também retirar conclusões sobre os padrões de compra dos mesmos.

Atualmente, neste supermercado existem inúmeras plataformas digitais que permitem a recolha de grandes volumes de dados dos clientes. Há muita informação disponível para tratar, e analisar. A empresa tem consciência que o tratamento e análise de dados são fundamentais para desenvolver

estratégias e táticas, para identificar tendências e fazer previsões de negócio. Assim, a experiência profissional na área do retalho alimentar, aliada ao conhecimento teórico adquirido ao longo do mestrado será um forte contributo quer para o aumento de informação nesta área de estudo específica, quer para o enriquecimento do conhecimento da empresa.

1.1. OBJETIVOS

Neste sentido, este projeto tem como primeira finalidade estudar as regras de associação entre produtos, por forma a conhecer quais os artigos que estão correlacionados em cada encomenda. Em segundo lugar, pretende identificar os tipos de departamentos, e as subcategorias de produtos mais frequentes numa transação de supermercado. Uma outra finalidade é chegar a conclusões acerca dos padrões e hábitos de consumo dos clientes que possam trazer vantagens competitivas para o negócio e para o aumento das vendas. Pretende-se também identificar se a presença de um conjunto de artigos resulta na existência de outro conjunto distinto de artigos, através da aplicação das regras de associação. O último objetivo deste projeto é também a identificação dos diferentes tipos de clientes até agora desconhecidos.

Neste trabalho pretende-se também responder às seguintes questões: "Existe algum registo de informação sobre os hábitos de consumo dos clientes da a*pp* do supermercado?", "Quais são os artigos que são comprados em conjunto?", "Quais os produtos com maior relação de proximidade e semelhança entre si?" e "Há produtos que podem ser recomendados ao cliente com base no seu histórico de compras, ou com base no critério de similaridade entre artigos?"

Convém referir que o estudo das regras de associação cai sobre uma base de dados de transações (pedidos de compras finalizados) de março de 2018 até outubro de 2019 dos clientes da aplicação *mobile* do supermercado. Cada pedido corresponde a um cliente e a um centro de expedição. Este é constituído por um código, por uma referência de departamento, por uma referência interna do artigo, pela descrição, pela quantidade, e pelo valor unitário do artigo. As transações que estão registadas na base de dados correspondem a todos os artigos vendidos na *app*. Os produtos estão alocados a categorias que se dividem em subcategorias. No total, existem 16 departamentos: talho, peixaria, charcutaria, frutaria e legumes, bebidas, lácteos e ovos, drogaria e limpeza, perfumaria, mercearia, congelados, doces e pequeno-almoço, padaria e pastelaria, vinhos e licores, e animais.

Por último, é importante realçar algumas conclusões sobre o estado da arte e como elas se aplicam a este tópico. A utilização das regras de associação permitirá perceber quais os produtos que mais se relacionam entre si, e que refletem um tipo de comportamento de compra padronizado do cliente. Esta informação sobre a relação entre artigos e o tipo de padrão de consumo ajudará a identificar produtos potenciais para estratégias comerciais de *cross-selling* e *up-selling*, ou para destacar produtos em corredores estratégicos. A otimização de espaços publicitários junto das marcas com base nesse estudo servirá de fundamento para a tomada de decisão dos fornecedores. Além disso, a segmentação de clientes permitirá redefinir e personalizar o tipo de conteúdo informativo e adequá-lo às características de consumo de cada tipo de segmento de clientes.

1.2. A APLICAÇÃO MÓVEL DE SUPERMERCADO

O supermercado em estudo tem lojas físicas e online. Relativamente à distribuição geográfica dos seus supermercados em Portugal, existem dois supermercados de grande dimensão e seis de menor dimensão. Os supermercados estão localizados nos grandes centros urbanos do país. A sua filosofia

caracteriza-se por ser uma loja de proximidade e as suas principais prioridades são: a satisfação do cliente; o atendimento personalizado capaz de proporcionar experiências únicas e memoráveis; a oferta de variedade, qualidade, segurança e confiança no sortido de produtos; horários amplos, de modo a satisfazer as necessidades de um mercado cada vez mais ativo.

Recentemente a empresa aproveitou o crescimento exponencial do comércio eletrónico mundial para dar também o salto tecnológico. A sua presença nas plataformas de *e-commerce* tem sido uma forte aposta do grupo, de maneira que em 2004 foi lançado o *website*, e em 2015 a sua aplicação *mobile*.

A *app* do supermercado foi desenhada na horizontal para *smartphone* e *tablet*, e é um espelho do modelo do supermercado físico. O cliente pode passear de forma cómoda e rápida pelos corredores do supermercado, encontrar os artigos que pretende nos lineares, e colocá-los no cesto, tal como faz no supermercado físico, sem sair do conforto do seu lar. Os serviços de entrega ao domicílio e *recolha em loja* permitem a escolha de quando e onde o consumidor quer receber o seu pedido.

Graças ao conceito de omnicanalidade, o cliente consegue ter a mesma experiência de compra em qualquer canal de venda. Isso significa que o cliente consegue ter a mesma experiência na *app*, no *site* e na loja física. Com a implementação de um registo único, o sistema informático consegue localizar o mesmo cliente nos vários sistemas e dispositivos. Assim, um cliente pode iniciar um pedido na *app*, e terminá-lo na loja.

2. REVISÃO DA LITERATURA

A literatura pesquisada incide sobre livros técnicos, casos de estudo, revistas profissionais e dissertações académicas. Com o processo de revisão bibliográfica, pretende-se estudar, conhecer e desenvolver os principais conceitos, etapas e abordagens dos principais autores e teóricos. Este processo iniciou-se com uma descrição detalhada e pormenorizada dos conceitos de *data mining*, regras de associação, análise do cesto de compra e segmentação.

A metodologia utilizada foi a investigação aplicada com o objetivo descritivo. Esta pesquisa bibliográfica foi utilizada para a fundamentação teórica do projeto, utilizando o método indutivo, pois parte de informação particular e específica. A pesquisa foi feita em várias etapas.

2.1. DATA MINING

Data mining é uma área de estudo jovem e em forte crescimento, e que sofreu uma enorme evolução, sobretudo nas últimas décadas. A primeira fase aconteceu antes dos anos sessenta, e caraterizou-se como um sistema muito primitivo de recolha, criação, processamento e arquivo de dados (Han, Kamber, & Pei, 2011).

A segunda fase ocorre nos anos 70 e 80 e destaca-se como um sistema sofisticado de gestão de base de dados (*Database Management System*). Nesta fase, há uma evolução na pesquisa e no desenvolvimento de sistemas de bases de dados, *i.e.,* há uma transição de um sistema de bases de dados hierárquico e em rede para um sistema relacional. Também há evoluções nas ferramentas. Estas evoluíram para métodos de indexação e acesso. Sucederam-se também outras evoluções. Os utilizadores passaram a ter mais acessibilidade e flexibilidade aos dados, através do uso de linguagens baseadas em consultas, interface do utilizador, bem como a otimização de consultas e gestão de transações. Outro método que contribuiu para a evolução da tecnologia relacional e que a tornou numa das principais ferramentas para o armazenamento, recuperação e gestão de grandes volumes de base de dados foi a OLTP, que é um método de processamento de transações *online*, onde uma consulta é visualizada como uma transação apenas de leitura.

A terceira fase aconteceu na década de 90 e ficou conhecida como Sistema de Base de Dados Avançado (*Advanced Database System*). Este sistema inclui novos e poderosos modelos de dados, como por exemplo relações estendidas, orientação a objetos, objetos relacionais e modelos dedutivos. Esta fase também ficou conhecida pelo florescimento de aplicações orientadas para bases de dados, (espaciais, temporais, multimédia, *streaming*, científicas, engenharia, de conhecimento e de informação empresarial) e pelo estudo intensivo da distribuição, diversificação e partilha de dados.

O final dos anos 80 ficou conhecido como uma altura de grandes progressos na área de tecnologia computacional de *hardware*. Os computadores tornaram-se máquinas poderosas e essenciais para a recolha e armazenamento de extensos volumes de dados. Além disso, foram essenciais para impulsionar as indústrias de base de dados e de gestão de informação, e contribuíram para o aumento de repositórios de bases de dados de informação. Graças ao desenvolvimento tecnológico, a informação passou a ficar arquivada em diversos centros de dados e repositórios de informação. Nessa altura emergiram os grandes repositórios ou os armazéns de dados (*data warehouse*s). Estes armazéns são depósitos de informações heterogéneas e provenientes de diversas fontes. A sua organização é feita sob um sistema unificado, localizado num único site, por forma a facilitar a gestão e tomadas de

decisões dos utilizadores. A tecnologia de armazenamento de dados ou as técnicas OLAP incluem a limpeza de dados, a sua integração, e o seu processamento analítico *online*. As principais funcionalidades dessas técnicas de análise passam por resumir, consolidar, agregar e visualizar informações de diferentes ângulos. No entanto, estas ferramentas, por si só, não são suficientes, e precisam de ferramentas de tratamento de dados para análises mais complexas, tais como, a classificação, segmentação, deteção de *outliers*, identificação de anomalias, e verificação da alteração dos dados ao longo do tempo.

A última, corresponde ao século XXI, é caracterizada como sendo a era da informação e do uso de técnicas de *data mining*. A sociedade é caracterizada como sendo a sociedade da informação, porque as empresas, as instituições de ensino, saúde, justiça e outras mais produzem e consomem diariamente elevados volumes de informação e de dados, como nunca registado. Na era atual, a heterogeneidade de dados leva ao aparecimento de novas abordagens, tais como: o *text mining*, que consiste em descobrir informação através de mensagens de texto; o *web mining* que extrai informação a partir da análise da estrutura de um *website*, do histórico de navegação, e dos conteúdos pesquisados na internet; o *ubiquitous data mining*, que consiste no processo de extração do fluxo de informação continuo que é independente do lugar físico onde é gerado; o *multimedia data mining* que procura descobrir padrões de informação em conteúdos digitais, tais como, imagens, vídeos, áudio, e outros; o *spatial data mining* que consiste em descobrir e extrair conhecimentos de dados espaciais; e o *spatiotemporal data mining* que pretende encontrar padrões através do estudo do tempo (meteorologia) e do espaço (história do mundo). Essas abordagens, são apenas alguns exemplos recentes, no entanto, outras mais vanguardistas surgirão. Bases de informações globais assentes na *internet*, e outros tipos de bancos de dados heterogéneos interligados surgirão. O futuro é promissor e gerir, recuperar, tratar e analisar distintos tipos de informação é uma tarefa desafiante e benéfica para a evolução e crescimento da sociedade da informação. (Han, Kamber, & Pei, 2011)

2.2. MÉTODOS DE PROJETO DE DATA MINING

O *data mining* utiliza diferentes tipos de técnicas para analisar problemas. No entanto, aplica tipicamente uma metodologia comum. As metodologias mais conhecidas são: SEMMA (*Sample*, *Explore*, *Modify*, *Model* and *Assess*) (Vijaylaxmi, Batra, & Alam, 2012), *Knowledge Discovery and Data Mining* (KDD) (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), e *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Shearer, 2000).

Segundo o SAS, a metodologia SEMMA é um processo que serve para o tratamento de dados. Este está dividido em cinco etapas: *Sample* (Amostragem), *Explore* (Exploração), *Modify* (Modificação), *Model* (Modelação), *Assess* (Avaliação). Esta metodologia torna o processo de tratamento de dados mais simples para os negócios, pois seleciona e transforma as variáveis mais importantes, cria modelos para prever novos resultados e confirma a credibilidade do modelo. Alguns benefícios desta metodologia são o facto de agrupar clientes de acordo com as tendências de consumo, identificar os clientes mais rentáveis da empresa, perceber quais os fatores que afetam os seus padrões de compra, pagamento e tempos de resposta, adquirir novos clientes, e direcionar publicidade para clientes alvo com maior interesse de compra (Vijaylaxmi, Batra, & Alam, 2012).

1. *Sample* (Amostragem): Inclui uma amostra de uma base de dados extensa. A amostra é usada para criar tabelas de treino, de validação e de teste de modelos.

2. *Explore* (Exploração): Esta etapa consiste em explorar os dados para fins de análise e visualização. Além disso, a exploração serve também para encontrar nos dados associações, previsões, e tendências improváveis. A exploração também classifica os atributos em três tipos de categorias: atributos chave, atributos identificadores, e confidenciais.
3. *Modify* (Modificação): Algumas empresas precisam de manipular alguns dados mais sensíveis ou confidenciais. Através da transformação de valores, de atributos, e da normalização de atributos privados é possível modificar dados mais sensíveis.
4. *Model* (Modelação): Depois de transformar os dados e deixá-los prontos para serem analisados é preciso inferir um tipo de modelo. Algumas técnicas para a criação de modelos de tratamento de dados são: árvores de decisão, redes neuronais, análises estatísticas, modelos logísticos e máquina de vetores de suporte.
5. *Assess* (Avaliação): Nesta última etapa os analistas criticam os resultados dos modelos e aplicam os modelos em casos reais.

A metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM) foca-se no objetivo e nas melhores práticas para cada tipo de negócio. Ajuda as empresas a encontrar a estrutura ideal que facilita o tratamento de dados, por forma a obter resultados rápidos e relevantes (Shearer, 2000). O processo CRISP-DM é constituído por seis fases:

1. Compreensão do negócio: É a etapa mais importante do processo, e consiste em compreender os objetivos do projeto de acordo com a visão do negócio. Compreender um determinado tipo negócio exige determinar primeiramente os seus objetivos, em segundo avaliar o problema, de seguida definir os objetivos da tarefa de tratamento de dados, e, por último, criar um plano de projeto.
2. Compreensão dos dados: Começa com a recolha e exploração de dados. Nesta etapa é fundamental a familiarização com os dados para assim verificar a sua qualidade e que informações importantes podem transparecer. Esta etapa envolve quatro fases: A recolha inicial de dados, a sua descrição, exploração e a confirmação da qualidade dos dados.
3. Preparação dos dados: Os dados são provenientes de diversas fontes, pelo que é sempre necessário recorrer à organização ou preparação dos dados, que consiste na seleção, limpeza, construção, integração e formatação dos mesmos. No caso de estes serem de fraca qualidade, então é necessário recorrer à sua limpeza, através da sua filtragem, combinação e preenchimento de valores em falta.
4. Modelação: Nesta etapa, procede-se à seleção de um modelo específico, como por exemplo as árvores de decisão, ou as redes neuronais. De seguida, deve testar-se a sua qualidade, criar um ou mais modelos a partir do conjunto de dados já trabalhados, e, por fim, avaliar o modelo.
5. Avaliação: Nesta etapa, as fases mais importantes são a avaliação dos resultados, a revisão do processo e o planeamento das próximas etapas. No passo da avaliação dos resultados, revê-se a precisão e amplitude do modelo. Na revisão do processo, os dados são revistos por forma a não negligenciar nenhum fator ou tarefa importante. No planeamento das próximas etapas cabe ao gestor de projeto decidir descartar o projeto ou avançar com a sua implementação.
6. Distribuição: Aqui, os dados reais e o conhecimento extraído são organizados e apresentados de forma a que o cliente possa utilizá-los.

Data mining é uma disciplina que pode ser definida de diversas formas. Por um lado, pode definir-se como o processo de encontrar conhecimento nos dados. Por outro, também se pode definir como o

processo de encontrar um conjunto de pequenas partículas preciosas a partir de uma quantidade enorme de matéria-prima. Outros autores definem *data mining* como sendo um sinónimo de outro termo igualmente popular, o KDD, isto é, a descoberta de conhecimento a partir de dados (Han, Kamber, & Pei, 2011). O processo KDD é realizado através de uma sequência iterativa de sete etapas:

1. Limpeza de dados: Remoção de ruídos e de dados inconsistentes.
2. Integração de dados: Diversas fontes de dados podem ser associadas.
3. Seleção da informação: Dados relevantes para a análise são extraídos da base de dados.
4. Transformação de dados: Os dados são transformados e consolidados em formas apropriadas para o seu tratamento para posteriormente resumir e tratar essas operações.
5. *Data mining*: Métodos inteligentes são utilizados para extrair padrões na informação.
6. Avaliação de padrões: Identificar padrões verdadeiros e interessantes, que representam o conhecimento baseado em medidas inteligentes.
7. Apresentação do conhecimento: Técnicas de visualização e representação são utilizadas para dar a conhecer a informação aos utilizadores.

KDD é classificado como o processo de descobrir conhecimento relevante através dos dados, enquanto que *data mining* se refere a uma etapa específica desse processo. *Data mining* consiste na aplicação de algoritmos específicos cuja finalidade consiste em extrair padrões de uma determinada base de dados recorrendo à matemática, à probabilidade e à identificação de padrões, para no final obter informações sobre novos padrões, tendências e associações. O objetivo final é obter altos níveis de conhecimento a partir de baixos níveis de informações no contexto de grandes bases de dados (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

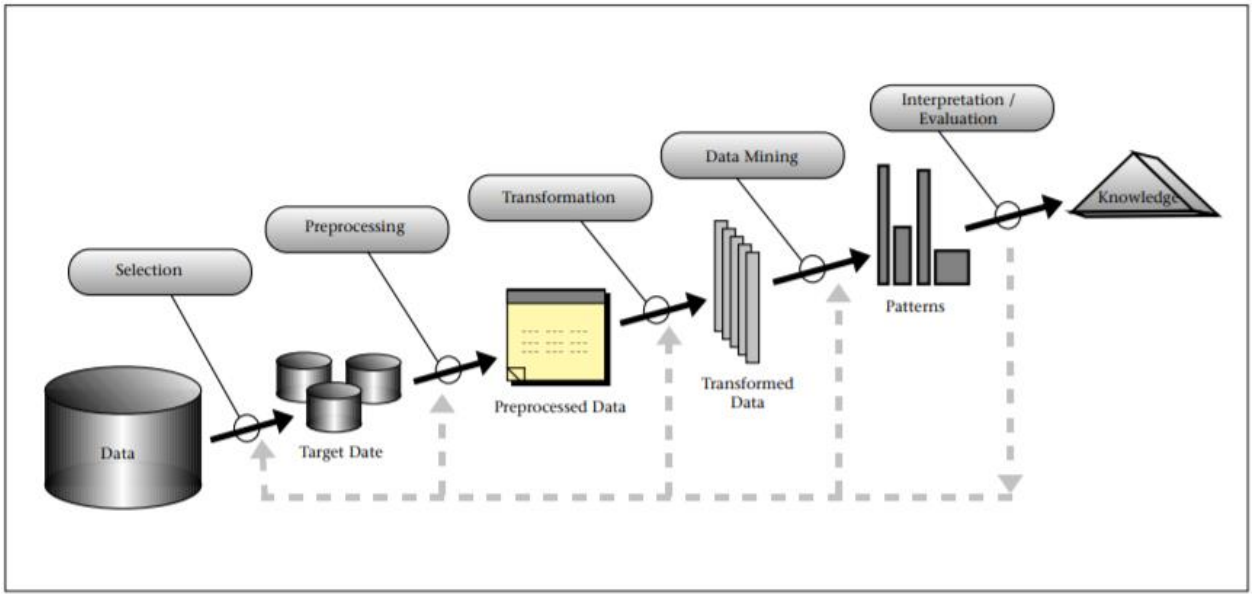

Figura 1 – Metodologia KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

Para os mesmos autores, o processo de KDD é composto por várias etapas. Antes de iniciar o processo é preciso conhecer qual é o objetivo a alcançar, e ter uma ideia clara dos dados a analisar. A primeira etapa chama-se seleção, e corresponde à escolha dos dados para analisar. A segunda etapa corresponde ao pré-processamento, ou limpeza de dados, e aqui é necessário criar uma tabela de dados a ser analisada posteriormente. A limpeza dos dados inconsistentes e imprecisos da base de

dados original é fundamental, pois previne o aparecimento de *outliers*, ou seja, valores que se afastam dos valores da amostra. A terceira etapa corresponde à transformação de dados. Nesta etapa aplicam-se algumas técnicas sobre as variáveis de modo a diminuir a dimensionalidade da amostra. Os dados sofrem algumas transformações de modo a poderem ser processados por algoritmos de *data mining*. A redução de atributos menos relevantes também é uma forma de transformar os dados. A quinta etapa consiste no *data mining*, e tem como objetivo analisar extensos volumes de dados para no final descobrir e extrair informações relevantes e valiosas. A sexta etapa consiste na interpretação e visualização de padrões e modelos dos dados tratados. Por último, o conhecimento extraído servirá para ser aplicado diretamente ou ficar documentado e ser reportado às entidades interessadas.

O *data mining* utiliza modelos descritivos e preditivos para fundamentar decisões de negócio inteligentes. As tarefas de agrupamento e associação pertencem aos modelos descritivos e as tarefas de previsão e regressão aos modelos preditivos. Este último modelo utiliza um conjunto de dados que são categorizados de acordo com uma variável pré-definida. As regras de associação podem ser utilizadas em ambos os modelos (Gama, Carvalho, Faceli, Lorena, & Oliveria, 2015).

O *data mining* tem também inúmeras funcionalidades. Destacam-se a caracterização e a discriminação, tratamento de padrões frequentes, associações e correlações, classificação e regressão, análise de segmentos e de *outliers*. As suas funcionalidades servem para especificar os tipos de padrões que são encontrados nas tarefas do *data mining* (Han, Kamber, & Pei, 2011). Os autores classificam estas tarefas em duas categorias principais: a descritiva e a preditiva. A primeira tem como função caracterizar as propriedades dos dados numa determinada base de dados, enquanto que a segunda executa deduções sobre os dados para no final fazer previsões. Sobre as funcionalidades, e nomeadamente a caracterização, a mesma consiste em resumir as características ou funcionalidades de uma classe que existe num conjunto de dados. Existem várias formas de sintetizar e caracterizar dados, tais como a utilização de medidas estatísticas, diagramas, etc. Por sua vez, a discriminação é uma comparação dos traços gerais de objetos da classe alvo contra as características gerais de objetos de um ou mais segmentos contrastantes. As formas de representação dos dados discriminatórios são iguais às formas de caracterização. No entanto, há uma pequena diferença: têm medidas comparativas que ajudam a distinguir o objeto alvo e as classes contrastantes.

Em relação, às associações e correlações, estas são padrões que ocorrem com frequência num conjunto de dados. Existem muitos tipos de associações, tais como os itens frequentes, os padrões sequenciais e as subestruturas frequentes. Os itens frequentes são conjuntos de itens que geralmente aparecem juntos numa transação ou num conjunto de dados. Um exemplo disso é o conjunto de leite e pão no caso das compras de supermercado, que são itens comprados em conjunto frequentemente. Os padrões de compra sequencias são itens adquiridos sequencialmente de forma frequente, como o caso de clientes que compram primeiro um portátil, de seguida uma câmara digital, e depois um cartão de memória. Uma subestrutura refere-se a diferentes tipos de formas estruturais que podem ser combinadas com grupos de itens e/ou sequências. Se uma subestrutura ocorre frequentemente, então tem o nome de estrutura de padrões frequentes. O tratamento de padrões frequentes num conjunto de dados leva à descoberta de associações e correlações muito interessantes.

A classificação e regressão são usadas para criar modelos de análise preditiva. A classificação é o processo de encontrar um modelo que descreva e diferencie classes ou segmentos. O modelo deriva da análise de um conjunto de dados previamente treinados e tratados. O modelo serve para classificar um cliente, para depois colocá-lo numa determinada classe. Há vários métodos de classificação, mas

os mais conhecidos e utilizados são as árvores de decisão, as redes neuronais artificiais, e as fórmulas matemáticas. O modelo de classificação analisa conjuntos de dados já classificados, ou seja, classes de dados já tipificadas.

A análise de agrupamentos ou *cluster analysis* é um método de aprendizagem descritivo, ou não-supervisionado. Estuda os objetos ignorando a etiquetação dos elementos. A segmentação pode ser utilizada para tipificar classes num conjunto de dados. Os objetos são aglomerados com base nos critérios de maximização de semelhanças *inter-cluster* e minimização de semelhanças *intra-cluster*. Por outras palavras, elementos homogéneos são colocados no mesmo grupo e elementos diferentes são agrupados em grupos distintos. Este método utiliza um leque abrangente de algoritmos, como por exemplo o *k-means*, *k-medoids*, *Ward's method*, *centroid method*, *k-nearest-neighbor*, *single-linkage method*, entre outros.

Por último, existe a análise de *outliers*. Os *outliers* são objetos que se distanciam dos outros objetos no conjunto de dados, ou de um modelo específico de um determinado conjunto de dados. Os *outliers* são vistos por muitos métodos de *data mining* como exceções ou dados ruidosos. Porém, em alguns casos singulares (deteção de fraude), os *outliers* são considerados eventos raros e bem mais importantes que os dados considerados normais.

2.2.1. Regras de associação

As regras de associação são uma forma de descobrir padrões associativos através dos dados dos cestos de compras. Os avanços tecnológicos nos equipamentos de registo nos pontos de venda, como o código de barras, possibilitaram que as empresas conseguissem recolher e armazenar enormes volumes de dados dos cestos de compras. Esses dados são constituídos por transações, que por sua vez são conjuntos de artigos adquiridos por clientes (Agrawal & Srikant, 1994).

Esta técnica de *market basket analysis* é muito utilizada pelos supermercados para criar vantagens competitivas nos seus negócios, tais como criar sistemas de fidelização, dos quais os cartões de fidelização são um exemplo. Estes servem para decidir a localização de produtos numa loja, para criar campanhas de *marketing* personalizadas, e até para definir a validade de produtos e promoções. Efetivamente, os cartões de fidelização têm sido uma estratégia de negócio muito apreciada pelas empresas, pois são benéficos tanto para os clientes como para as empresas. Por um lado, oferecem descontos aos clientes e por outro permitem que as empresas consigam conhecer os hábitos de consumo dos seus clientes, através dos registos das operações de compra.

A análise dos dados dos cestos de compras é uma forma de análise que incide sobre o comportamento de consumo de um cliente num determinado supermercado. A análise é realizada através de meios que identificam associações e ligações entre os artigos colocados numa lista de compras. Além disso, a análise visa não só identificar a frequência de compra de determinados artigos, mas também quais os artigos adquiridos simultaneamente. As informações recolhidas permitem ajudar a empresa a criar estratégias de negócio mais inteligentes, *e.g.* colocar produtos com maior frequência de compra num local específico e estratégico (Kurniawan, Umayah, Hammad, Nugroho, & Hariadi, 2018).

A análise de associações é utilizada para conhecer os artigos que são comprados em conjunto. Han, Kamber and Pei, explicam as regras de associação a partir do exemplo dos dados transacionais de uma loja de eletrónica (Han, Kamber, & Pei, 2011). Na regra de associação

$$Compra\ (X.\text{“}Computador\text{”}) \Rightarrow Compra\ (X.\text{“}Software\text{”})\ [suporte = 1\%, confiança = 50\%]$$

X é a variável que representa um tipo de cliente. A percentagem de 50% refere-se à confiança ou certeza, de que quando um cliente compra um computador, há 50% de possibilidade de vir a comprar um *software*. O suporte de 1% demonstra que em todas as transações analisadas, 1% das compras têm os artigos computador e software adquiridos em conjunto. Podemos concluir que regras de associação que envolvem apenas um atributo ou predicado são referidas como sendo regras de associação unidimensionais. Porém, também existem as regras de associação multidimensionais, que são mais complexas que a anterior, pois envolvem mais do que um atributo e/ou predicado. A regra multidimensional é explicada através da seguinte fórmula

$$Idade\ (X, \text{“}20..29\text{”}) \wedge Rendimento(X, \text{“}40K..49K\text{”}) \Rightarrow Compra(X, \text{“}Computador\text{”})\ [suporte = 2\%, confiança = 60\%]$$

Esta regra indica que 2% do total dos clientes estudados encontram-se na faixa etária dos 20 a 29 anos, com rendimentos entre os 40 e 49 mil dólares, e compraram um computador na loja de eletrónica. Esta regra indica também que há 60% de probabilidade de clientes com idade e rendimento dentro dos valores referidos comprarem um computador. Regras de associação que contêm fracos valores de suporte e confiança são consideradas irrelevantes.

A técnica do *market basket analysis* procura entender um conjunto de problemas de negócios relacionados com os dados transacionais dos pontos de venda. A sua utilização tem vindo a expandir-se para diferentes domínios, como por exemplo, nos negócios *online* para compreender o comportamento de navegação dos utilizadores no seu *website*. Além disso, são também frequentemente utilizadas em estudos de diversas áreas, tais como no *marketing*, bioinformática, banca, seguradoras, educação, saúde, etc. (Linoff & Berry, 2011).

Stilou, Bamidis and Maglaveras realizaram um estudo na área da saúde onde utilizou as regras de associação para a realização de um processo inteligente de diagnósticos para pacientes diabéticos. Eles avaliaram os parâmetros dos pacientes diabéticos de modo a conhecer os padrões da doença para posteriormente prescrever tratamentos de acordo com o seu historial clínico. Neste estudo o uso das regras de associação permitiu realizar diagnósticos inteligentes, extrair informações valiosas e construir bases de dados com conhecimentos relevantes de forma rápida e automática (Stilou, Bamidis, & Maglaveras, 2001).

Gonçalo Barros organizou um estudo onde utilizou as regras de associação para reorganizar a loja de parafarmácia dos grandes armazéns do El Corte Inglés. Para isso recorreu, em primeiro lugar às regras de associação através da análise dos dados do cesto de compras dos clientes. A partir da análise dos seus talões de venda e dos resultados das regras de associação, conseguiu conhecer os padrões de consumo dos clientes, quais as famílias e marcas de produtos compradas em grupo, bem como as suas associações. Por exemplo, verificou-se que os talões de compra que contêm produtos da família de materiais de cura também contêm artigos da família de podologia, portanto, existe uma enorme probabilidade de um cliente passar no corredor material de cura e comprar um artigo de podologia. Foi com base nesses resultados que os corredores da parafarmácia foram reorganizados. Portanto, as famílias e marcas de produtos com maior relação entre si foram organizadas por forma a estarem mais próximas, com a finalidade de aumentar a probabilidade de compra (Barros, 2018).

As regras de associação são os melhores modelos para o estudo de tratamento de dados. Sánchez, Vila, Cerda and Serrano realizaram um estudo sobre transações de cartões de crédito fraudulentas no maior grupo de retalhistas do Chile. A sua proposta basou-se na extração de conhecimento por forma a obter padrões de comportamento a partir das transações de cartões de crédito fraudulentas. Nas regras de associação, o objeto de interesse é o item que aparece numa determinada transação. Neste estudo, as transações dos clientes são as compras pagas com cartão de crédito, e os itens são os artigos e as pessoas que fazem a compra. Para além das regras de associação, foram utilizadas as regras *fuzzy*, que permitem analisar dados imprecisos e incompletos por forma a obter informação útil e não explícita. Após o tratamento e exploração dos dados transacionais, os autores aplicaram um *software FuzzyQuery 2+*, e chegaram a várias conclusões. Mostraram que a preconceção de que o género feminino é alvo de fraude mais frequentemente é falsa. Também provaram que os jovens do género masculino são os mais afetados pelas fraudes bancárias. Por fim, chegaram à conclusão que as regras devolvidas pelo *software* não só são as mesmas que os analistas de risco usam para a deteção de fraude, como oferece conhecimento relevante que não é explícito, mas que contribui para o seu trabalho (Sánchez, Vila, Cerda, & Serrano, 2009).

As regras de associação conseguem identificar padrões de consumo através de um *dataset* de dados transacionais, e, posteriormente, recomendar produtos ou serviços com base nas preferências de consumo dos clientes. Lucas, Luz, Moreno, Anacleto, Figueiredo, and Martins levaram a cabo um estudo com uma abordagem híbrida de recomendações para o sector do turismo. Nesse estudo, os autores utilizaram dois métodos. O primeiro consistiu em criar e agrupar os utilizadores de acordo com as suas semelhanças, preferências e características através de um algoritmo de *clustering*. Este algoritmo usou atributos demográficos (idade, grau escolar e código-postal) e itens adquiridos pelos utilizadores para criar os *clusters*. O segundo método utilizado foi a classificação associativa, através do algoritmo *CBA-fuzzy*, que serviu para recomendar pontos turísticos aos utilizadores. O objetivo de utilizar uma abordagem híbrida de sistemas de recomendação permite não só considerar apenas a ocorrência de itens e utilizadores, mas também os atributos que descrevem os utilizadores e itens do sistema. As conclusões a que chegaram foram que o método proposto tem como principais objetivos ajudar outros sistemas de recomendação a evitar problemas comuns, tais como a escalabilidade, escassez, primeiros avaliadores e ovelhas negras (clientes com exigências de consumo muito particulares), e também ajudar as empresas do sector turístico a recomendar produtos e serviços de acordo com os critérios de interesse de cada utilizador. Efetivamente, diversos tipos de empresas utilizam sistemas e técnicas de recomendação para melhorar a experiência de compra e para facilitar a pesquisa de produtos. Além disso, as técnicas de recomendação também servem para potencializar positivamente os resultados dos seus negócios, a partir de estratégias de sistemas de fidelização, estratégias comerciais mais agressivas de *cross-selling* e *up-selling*, aumentar o valor gasto por talão de venda, transformar clientes potenciais em clientes reais (Lucas, et al., 2013).

O comércio eletrónico, ou *e-commerce*, tem vindo a apoiar-se nos sistemas de recomendação, não só para sugerir produtos, mas também para facultar informação importante sobre as suas caraterísticas, por forma a facilitar o processo de decisão de compra ao consumidor. A recomendação de produtos aos clientes pode ser feita recorrendo a estratégias como o *top* de produtos mais vendidos, aspetos sociodemográficos, ou previsão de futuras compras recorrendo à análise do histórico de compras. As recomendações podem ser: sugestões de produtos aos clientes, informação detalhada sobre o produto, resumos de opiniões de outros utilizadores ou críticas de outros utilizadores. Resumidamente, as plataformas de comércio eletrónico têm vindo a utilizar os sistemas de

recomendação para oferecer uma experiência de compra cada vez mais personalizada ao consumidor. O objetivo é demonstrar ao cliente que as lojas virtuais e produtos recomendados foram criados exclusivamente para eles. Um exemplo perfeito de empresas que utilizam os sistemas de recomendação para oferecer serviços e produtos personalizados aos seus clientes são a Amazon, HBO e Netflix. Estas empresas recorrem ao histórico de pesquisa e de compra dos seus utilizadores para posteriormente recomendar os seus produtos. Também utilizam técnicas de recomendação como a filtragem social e filtragem de conteúdos. A primeira permite que o sistema aglomere informação sobre os hábitos de consumo e preferências dos utilizadores para posteriormente fazer recomendações a outros utilizadores recorrendo sempre ao critério de semelhança entre eles. A segunda, funciona com a premissa que as recomendações são feitas com base na análise pormenorizada das características de todos os produtos classificados.

2.2.2. Clustering

O *clustering* consiste em agrupar dados, observações e casos, e torná-los em grupos de elementos semelhantes. Um *cluster* é uma coleção de amostras que são semelhantes entre si, e diferentes entre coleções de *clusters* distintos. O método de *clustering* difere do modelo de associação, na medida que não precisa de uma variável alvo (*target variable*) para construir *clusters*. Ao contrário do modelo de associação, que consiste em classificar, estimar ou prever uma variável alvo, o modelo de *clustering* procura agrupar um conjunto de dados em subgrupos de elementos homogéneos, ou segmentos, onde o critério de semelhança das amostras aumenta dentro dos *clusters* e diminui fora deles (Larose, 2014).

A análise de *clusters* consiste em dividir os dados em grupos que sejam significativos e úteis. Para fazer essa divisão, o algoritmo *k-means* é o mais popular, o mais direto, e também o mais efetivo para esse efeito (Larose, 2014). Esse algoritmo é explicado através dos seguintes passos:

1. Coloca-se a pergunta de em quantos *clusters* (o valor de k) se quer dividir o total da amostra.
2. Definem-se aleatoriamente k sementes, como sendo a localização central inicial dos *clusters*.
3. Para cada amostra o algoritmo procura encontrar o centro do *cluster* mais próximo. Deste modo cada centro de *cluster* está associado a um subconjunto de amostras.
4. Para cada *cluster* ele procura o seu novo centroide e atualiza a localização do seu centro para o seu novo centroide.
5. Repetem-se os passos 3 a 5, até o algoritmo convergir, ou o número máximo de iterações for atingido.

O *k-means* traz como vantagens a formação automática dos elementos em grupos, a simplicidade e rapidez computacional. Para uma primeira abordagem à segmentação, este algoritmo é realmente muito eficiente. No entanto, tem também algumas limitações, tais como a necessidade de definir um número inicial de *clusters a priori*, e ser muito sensível à posição inicial das sementes e *outliers*. É um algoritmo que funciona muito bem com *clusters* de formato esférico, porém o contrário não acontece com outros formatos.

3. METODOLOGIA

"Uma pesquisa é um processo formal e sistemático de desenvolvimento do método científico" (Gil, 2008). O seu objetivo consiste na aplicação de métodos científicos por forma a encontrar respostas para problemas.

A presente metodologia tem por base a análise descritiva, pois permite analisar e descrever detalhadamente as características e os hábitos de consumo dos clientes da aplicação de um supermercado.

A análise descritiva é divida em duas etapas. A primeira etapa está relacionada com a análise do cesto de compras e com a construção de modelos de associação. A segunda etapa com realização e descrição de *clusters* através de dois fluxos de segmentação: um de consumo e outro de valor.

Agrawal, Imielinski and Swami explicam as regras de associação através da seguinte declaração: 90% dos talões de compras que detêm os artigos *pão* e *manteiga* também incluem o artigo *leite* (Agrawal, Imielinski, & Swami, 1993). Os artigos *pão* e *manteiga* correspondem ao antecedente e o artigo *leite* ao consequente, e a percentagem 90% corresponde ao fator confiança, isto é, à percentagem que contém o antecedente e o consequente. Por outras palavras, as regras de associação servem para explicar a relação entre o antecedente e o consequente, ou seja, se o cliente comprou os artigos *pão* e *manteiga* em conjunto qual é a probabilidade de também comprar o artigo *leite*. No que diz respeito ao suporte, este contabiliza conjuntos de itens numa transação, isto é, na seguinte regra genérica $\{X\} \Rightarrow \{Y\}$, o suporte da regra contabiliza o número total de transações que têm os grupos de itens X e Y. Por fim, o suporte vai medir a repetição de um determinado conjunto de itens numa base de dados transacionais.

Efetivamente, para esses autores, as medidas de confiança e de suporte são muito importantes para avaliar as regras de associação. Porém, há outros autores que defendem que a medida *lift* é a mais relevante de todas. Esta medida avalia a ocorrência do consequente face ao antecedente. Resumidamente, o *lift* é um aumento da confiança do consequente sabendo que ocorreu o antecedente. O valor do *lift* é obtido através da divisão da confiança pela confiança esperada, onde a segunda corresponde ao número de ocorrências do consequente em todas as transações.

3.1. DADOS E FERRAMENTAS

Para a análise do cesto de compras, foi necessário construir os modelos de associação no *software SAS Enterprise Miner 14.2*. Relativamente à obtenção dos dados, foi utilizada uma base de dados fornecida pela empresa, referente às compras *online* de supermercado de março de 2018 até outubro de 2019. Estes foram tratados no *Microsoft Excel 2016*, por forma a serem usados mais facilmente no *SAS Enterprise Miner*. Os parâmetros escolhidos para a seleção dos dados foram:

- Dados sociodemográficos dos clientes:
    - data de nascimento
    - código postal
    - género
- Dados transacionais das compras:
    - endereço eletrónico

- código do cliente
- código de pedidos
- código de produtos
- código de departamentos
- código dos centros

Cada código de pedido identifica uma venda, das quais foram contabilizadas um total de 11383. Entre estas, contabilizaram-se 22520 referências de artigos distintas.

3.2. PRÉ-PROCESSAMENTO

Após obter o *dataset* da empresa, procedeu-se à etapa do pré-processamento, ou seja, à transformação dos dados brutos. A tabela ABT (*Analytical Base Table*), a tabela de artigos e tabela de departamentos resultaram dessa transformação. Os objetivos da elaboração das tabelas são, por um lado, a construção das regras de associação, e, por outro, a elaboração de um modelo de segmentação.

Na tabela ABT, reuniram-se os dados da tabela original, e adicionaram-se novos dados, tais como a recência, frequência e o valor monetário para avaliar o valor dos clientes para a empresa. Estas variáveis vêm na sequência da possibilidade de fazer análise RFM (*Recency, Frequency, Monetary*). A recência é uma medida que indica que os clientes que fizeram uma compra recentemente são mais prováveis de voltar a comprar num futuro próximo. A frequência indica que clientes que costumavam fazer compras no passado são também muito prováveis de realizar uma compra num futuro próximo, e o valor monetário demonstra que clientes que gastaram muito dinheiro no passado têm uma elevada probabilidade de gastar uma quantidade igual no presente. Nas tabelas Tabela 1 a Tabela 5, é possível visualizar o exemplo de alguns clientes representados na tabela ABT, mostrando todas as variáveis associadas a cada um.

CUC Cliente	Género	Ano de Nascimento	Recência (semanas)	Frequência (semanas)	Monetário
116490327	Feminino	1970	0	50	6081.83 €
116528142	Feminino	1969	37	1	49.05 €
118187947	Feminino	1972	19	15	4900.39 €
118201839	Feminino	1976	3	16	2574.34 €
118202035	Masculino	1967	8	2	260.17 €
118202274	Feminino	1981	78	1	167.92 €
118202795	Feminino	1969	61	3	326.40 €
118203199	Feminino	1979	2	1	102.60 €
118203520	Feminino	1979	1	16	1961.69 €

Tabela 1 – Tabela ABT de clientes (1/5)

CUC Cliente	Total Pedidos	Núm. médio de artigos por pedido	Total de artigos	Valor médio por pedido	Núm. de departamentos
116490327	85	30	2510	71.55	15
116528142	1	33	33	49.05	9
118187947	17	78	1326	288.26	16
118201839	16	63	1008	160.9	14
118202035	2	56	113	130.09	11
118202274	1	68	68	167.92	11
118202795	3	31	93	108.8	12
118203199	1	52	52	102.6	5
118203520	16	51	817	122.61	13

Tabela 2 – Tabela ABT de clientes (2/5)

CUC Cliente	Monetário Bebidas	Monetário Talho	Monetário Charcutaria	Monetário Congelados	Monetário Dietéticos	Monetário Drogaria
116490327	127.41 €	322.63 €	974.55 €	109.40 €	247.05 €	988.71 €
116528142	3.16 €	0.00 €	3.98 €	4.78 €	0.00 €	2.39 €
118187947	448.90 €	272.85 €	155.08 €	182.56 €	70.73 €	663.29 €
118201839	300.48 €	19.54 €	244.71 €	231.92 €	96.81 €	254.22 €
118202035	20.19 €	21.22 €	21.81 €	39.87 €	0.00 €	25.90 €
118202274	12.61 €	0.00 €	11.52 €	4.99 €	7.73 €	11.38 €
118202795	9.69 €	107.14 €	40.12 €	0.00 €	0.00 €	33.07 €
118203199	0.00 €	0.00 €	31.48 €	0.00 €	9.52 €	8.48 €
118203520	347.39 €	6.03 €	67.46 €	66.06 €	91.56 €	129.11 €

Tabela 3 – Tabela ABT de clientes (3/5)

CUC Cliente	Monetário Doces e Pequeno-almoço	Monetário Frutaria	Monetário Legumes e Conservas	Monetário Animais	Monetário Pastelaria
116490327	394.26 €	505.48 €	108.24 €	0.00 €	5.54 €
116528142	9.87 €	2.99 €	8.22 €	0.00 €	0.00 €
118187947	502.25 €	229.75 €	43.96 €	35.88 €	22.58 €
118201839	316.53 €	6.12 €	410.58 €	0.00 €	0.00 €
118202035	5.57 €	0.00 €	32.86 €	0.00 €	3.90 €
118202274	28.67 €	19.29 €	7.13 €	0.00 €	0.00 €
118202795	13.31 €	24.35 €	0.00 €	0.00 €	3.56 €
118203199	0.00 €	0.00 €	0.00 €	0.00 €	0.00 €
118203520	336.80 €	75.68 €	85.10 €	0.00 €	0.00 €

Tabela 4 – Tabela ABT de clientes (4/5)

CUC Cliente	Monetário Perfumaria	Monetário Peixaria	Monetário Refrigerados e Lácteos	Monetário Mercearia	Monetário Vinhos
116490327	210.07 €	352.54 €	782.76 €	682.21 €	270.98 €
116528142	0.00 €	0.00 €	2.99 €	10.67 €	0.00 €
118187947	175.52 €	45.56 €	518.39 €	252.61 €	1280.48 €
118201839	242.84 €	3.39 €	278.84 €	103.79 €	64.57 €
118202035	0.00 €	0.00 €	40.35 €	5.72 €	42.78 €
118202274	32.24 €	0.00 €	25.35 €	7.01 €	0.00 €
118202795	8.98 €	29.25 €	35.77 €	5.41 €	15.75 €
118203199	0.00 €	0.00 €	40.40 €	12.72 €	0.00 €
118203520	125.88 €	0.00 €	225.76 €	285.21 €	119.65 €

Tabela 5 – Tabela ABT de clientes (5/5)

Cód. do pedido	Cód. do artigo	Núm. de artigos
2018001320167	181301000045	1
2018001320167	186300002808	1
2018001320167	181117000532	1
2018001320167	181101001120	1
2018001320167	181120000197	1
2018001320167	188641001136	1
2018001320167	280224001708	4
2018001320167	209451006142	1

Tabela 6 – Tabela de artigos por pedido para o modelo associativo

Relativamente à tabela dos artigos (Tabela 6 – Tabela de artigos por pedido para o modelo associativo), a mesma é constituída unicamente por dados transacionais. Esta tabela contém o código de pedido, pela referência dos artigos, e as suas quantidades. Por sua vez, a Tabela 7 – Tabela de departamentos por pedido para o modelo associativo, relativa aos departamentos, é composta pelos códigos dos pedidos, pelos nomes dos departamentos e respetivas quantidades.

Cód. do pedido	Nome do departamento	Núm. de artigos
2018000224972	Bebidas	4
2018000224972	Talho	1
2018000224972	Charcutaria	2
2018000224972	Drogaria e Limpeza	1
2018000224972	Frutaria	9
2018000224972	Legumes	1
2018000224972	Perfumaria	5
2018000224972	Refrigerados e Lácteos	1

Tabela 7 – Tabela de departamentos por pedido para o modelo associativo

3.3. CRIAÇÃO DE MODELOS

3.3.1. Modelos associativos

Após a transformação dos dados, seguiu-se a fase de construção do modelo de regras associação no *software SAS Enterprise Miner.* Este sistema é conhecido como sendo um dos melhores no suporte ao processo de tratamento de dados, e pela sua elevada capacidade de agregar, armazenar e tratar grandes volumes de dados. para que se possam descobrir padrões previamente desconhecidos, que podem ser úteis para criar vantagens competitivas. O acesso rápido e intuitivo a detalhes e a relações de dados também são características únicas que o distingue dos demais.

As funções do *SAS* são organizadas num processo chamado SEMMA: *Sample* (Amostra), *Explore* (Exploração), *Modify* (Modificação), *Model* (Modelação) e *Assess* (Avaliação). O primeiro passo corresponde a uma amostra dos dados que é feita através da extração e preparação dos mesmos para a criação do modelo. O passo seguinte – Exploração – consiste na exploração dos dados para obtenção de soluções e ideias. A Modificação representa o processo de transformação das variáveis. A Modelação diz respeito ao uso de técnicas analíticas com a finalidade de encontrar os resultados desejados. Por último, a Avaliação traduz-se na credibilidade e confiança dos resultados obtidos.

No *SAS,* foi necessário construir dois modelos de associação: um para a associação de artigos e o outro para a associação dos departamentos. Para a construção do primeiro, em primeiro lugar, utilizou-se o nó *File Import*, que está localizado no separador *Sample*, na barra de ferramentas do *SAS*. Este nó serve para importar ficheiros, *datasets*, folhas de cálculo, etc., para que o sistema os consiga tratar e processar. Além disso, no nó *File Import* foi possível configurar os dados e especificar as variáveis a importar, bem como os seus papéis. Para a variável "código do pedido" definiu-se o papel ID e para a variável "referência do artigo" o papel *target*.

Em segundo lugar, criou-se o nó *Association* que pertence à categoria *Explore*. Este nó é utilizado para descobrir associações e sequências. Além disso, permite identificar quais os artigos que ocorrem em conjunto num determinado evento ou registo. Nesta fase, definiram-se os valores mínimos para os critérios de suporte (5%) e confiança (10%), bem como o número máximo de itens por associação (2).

Para a construção do segundo modelo, procedeu-se à mesma metodologia. A exceção é a fonte dos dados, que passou a ser a tabela dos departamentos. Nas figuras Figura 2 e Figura 3, podem observar-se os diagramas para associação de artigos e de departamentos.

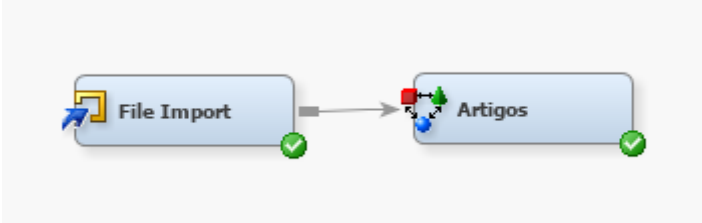

Figura 2 – Fluxo do diagrama SAS para a associação de artigos

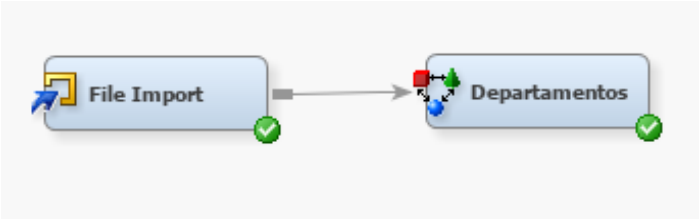

Figura 3 – Fluxo do diagrama SAS para a associação de departamentos

3.3.2. Modelos descritivos

O processo de segmentação consiste em agrupar os clientes de acordo com as suas semelhanças e variáveis de interesse. Tipicamente, o que se pretende na segmentação é obter *clusters* o mais homogéneos possível dentro de si, o mais heterogéneos possível entre si.

Relativamente ao modelo de segmentação de clientes, ou análise de *clusters*, o mesmo também foi desenvolvido no programa *SAS Enterprise Miner*, e tal como no modelo de associação, também se utilizou o modelo SEMMA. A base de dados utilizada para a segmentação de clientes foi a tabela ABT referida acima**Error! Reference source not found.**.

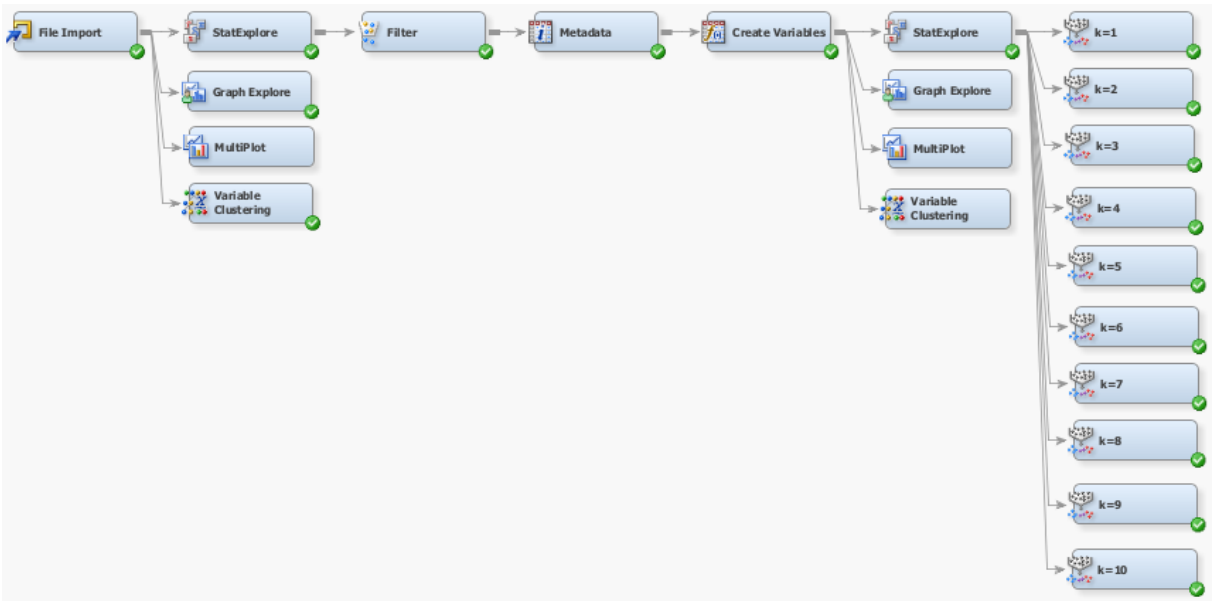

Figura 4 – Fluxo do diagrama SAS para a segmentação

Para o *clustering* de clientes, realizaram-se dois tipos distintos de segmentos. O primeiro foi realizado com a intenção de compreender o comportamento de consumo, através da medição de quanto os clientes gastaram por departamento. O segundo foi construído com a finalidade de conhecer o valor dos clientes para a empresa de acordo com os parâmetros de recência (tempo passado desde a data da última compra), frequência (o número de vezes que o cliente realizou compras), e valor monetário (valor total gasto em compras). Estes parâmetros foram complementados com outras métricas, tais como o número total de artigos, ou o número total de pedidos.

Em relação à base de dados original, esta foi transformada no *Microsoft Excel*, de modo a poder extrair dados e variáveis úteis e relevantes para os algoritmos de segmentação. Este passo foi essencial para reduzir o esforço de transformação dos dados dentro do *SAS*. A partir dos dados em bruto, procedeu-se ao cálculo das seguintes variáveis:

- Recência: foi calculada através do tempo que passou da última compra até à data dos últimos dados. Valor medido em semanas.
- Frequência: é o número de semanas diferentes em que um cliente fez compras.
- Monetário: valor total gasto por cliente.
- Total de pedidos: somaram-se todos os pedidos distintos de cada cliente.
- Número médio de artigos por pedido: total de artigos por pedido, em média, de cada cliente.
- Número total de artigos: número total de artigos que o cliente comprou.
- Valor médio por pedido: valor gasto por pedido, em média, de cada cliente.
- Número total de departamentos: total de departamentos distintos por cliente.
- Valor gasto por departamento: valor total gasto por cliente por departamento.

Para além dessas variáveis calculadas, também se adicionaram variáveis demográficas para auxiliar no processo de segmentação de clientes. A variáveis foram extraídas de um sistema de base de dados de clientes e estão na tabela ABT como:

- Género
- Data de nascimento
- Código postal
- Concelho de habitação

Esta preparação dos dados corresponde ao passo *Sample* do modelo SEMMA. Este passo foi levado a cabo fora do *SAS,* e posteriormente importado para o mesmo através do nó *File Import*.

No segundo passo deste modelo – *Explore* – exploraram-se os dados. Para isso usaram-se vários nós, tais como: o *Stat Explorer* para informações estatísticas (exemplificadas na Tabela 8), o *Graph Explorer* para visualizar os dados através dos histogramas e *scatter plots*, e, por último, o *Variable Clustering*, onde se verificaram as correlações entre as variáveis.

Variável	**Mínimo**	**Média**	**Máximo**
Ano de Nascimento	1926	1976	2004
Valor Médio por pedido	3.99 €	148.74 €	1255.08 €
Recência	0	33	87
Frequência	1	5	69
Monetário	3.99 €	869.74 €	18790.32 €
Total de Pedidos	1	5.73	123
Mercearia	0.00 €	61.61 €	2184.26 €
Drogaria	0.00 €	122.02 €	4816.85 €
Frutaria	0.00 €	83.74 €	4148.17 €

Tabela 8 – Dados estatísticos de algumas das variáveis antes de serem transformadas

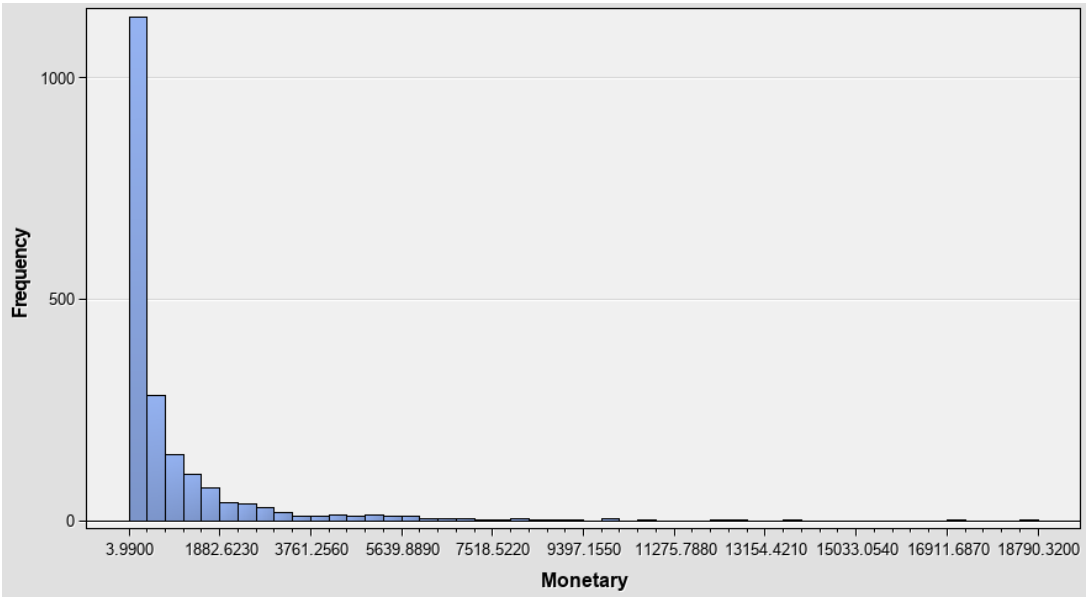

Figura 5 – Histograma da variável "Monetário" antes de ser transformada

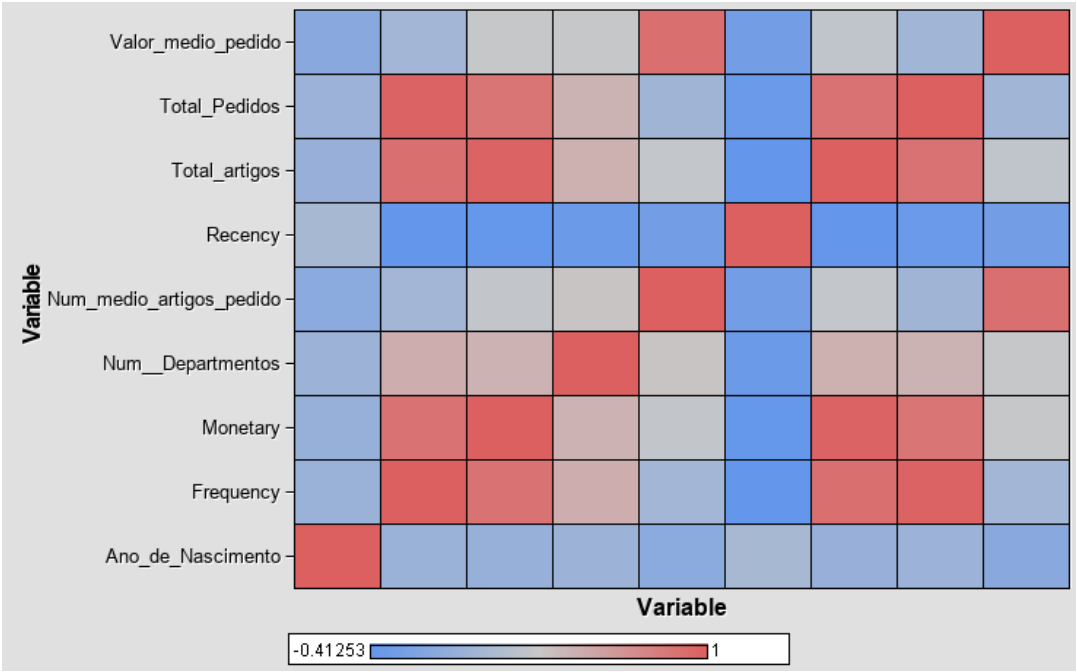

Figura 6 – Correlações entre variáveis usadas na segmentação por valor

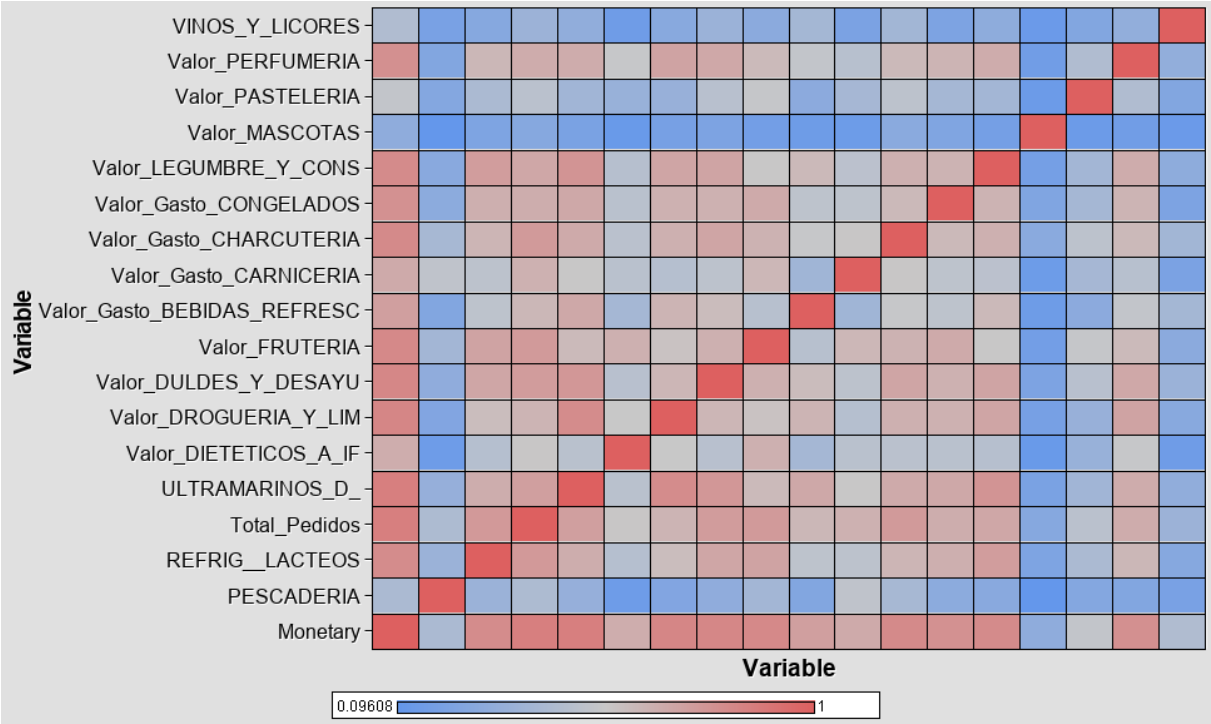

Figura 7 – Correlações entre variáveis usadas na segmentação por consumo

Sobre o passo da modificação, o mesmo serviu para criar variáveis de maneira a obter melhores resultados para o *clustering*. Na segmentação por valor, criou-se a variável "idade" com base no ano de nascimento. Já na segmentação por consumo, criou-se uma variável nova por departamento, para representar o valor médio gasto nesse departamento por cliente e por visita, ao invés do seu valor importado da tabela ABT, que representa o valor acumulado gasto em todas as vistas por cliente.

Para cada uma das segmentações, aplicaram-se modelos com vários números diferentes de *clusters,* designados por k. Isto foi feito com a finalidade de encontrar o valor de k que melhor agrupa os clientes. Para o encontrar, foi necessário obter e analisar a *Elbow Curve* resultante das várias segmentações.

A *Elbow Curve* representa a variância da segmentação explicada em função do número de *clusters*. É conhecida como uma ferramenta útil para a descoberta do melhor número de *clusters* através da análise do melhor compromisso entre a variância explicada e o respetivo número de *clusters*.

Para ambas as segmentações, os valores da variância (R^2) foram obtidos para o número de *clusters* k entre 1 e 10. Para cada k, as sementes foram espalhadas de acordo com o método *Princomp* (*Principal Component Method*).

4. RESULTADOS E DISCUSSÃO

Com os resultados obtidos, pode proceder-se à realização de diferentes estratégias de marketing, comunicar quais as categorias de artigos a colocar em promoção, criar talões de desconto, selecionar e ordenar produtos nos lineares por forma a torná-los mais apelativos ao olhar do consumidor, ou mesmo enviar mensagens personalizadas através do sistema *push notification* e mensagens *in-app*.

As regras de associação explicam a relação entre um acontecimento X (antecedente) e um acontecimento Y (consequente), bem como a probabilidade de ambas ocorrerem simultaneamente. As regras de associação são avaliadas e posteriormente explicadas através das métricas das tabelas Tabela 9 e Tabela 10. A tabela Tabela 9 é referente a pares de artigos e a tabela Tabela 10 a pares de departamentos. As regras de associação mais relevantes serão enumeradas consoante os valores obtidos do seu *lift*, suporte, confiança e confiança esperada.

4.1. ASSOCIAÇÃO DE ARTIGOS

A tabela Tabela 9 contém 12 regras distintas e as quatro métricas típicas: a confiança esperada, a confiança, o suporte e o *lift*. Uma vez que o *SAS* usou referências de artigos para criar regras de associação foi necessário substituir a referência do artigo pelo nome do artigo, por forma a tornar os resultados mais legíveis.

Regra	*Lift*	Confiança esperada	Confiança	Suporte
Lindahls Iogurte Líquido Proteico de Morango sem Lactose (330ml) ⇒ YOGHOUR ESPECIALIDADES	66.80	1.27%	84.51%	1.05%
Compal Essencial de Pêssego (3×110ml) ⇒ Compal Essencial de Manga (3×110ml)	30.23	2.21%	66.67%	1.02%
Royal Gelatina de Ananás (4×100g) ⇒ Royal Gelatina de Morango Pack (4×100g)	28.62	1.98%	56.58%	0.76%
Compal Essencial de Manga (3×110ml) ⇒ Compal Essencial de Maçã (3×110ml)	22.68	2.21%	50.00%	0.75%
Compal Essencial de Manga (3×110ml) ⇒ Compal Essencial de Pêra (3×110ml)	20.53	2.21%	45.26%	0.76%
Couve ao peso ⇒ Couve Flor ao peso (≈ 1.2kg/unid.)	7.47	3.95%	29.51%	0.90%
Couve Coração ao peso (≈ 410g/unid.) ⇒ Couve Flor ao peso (≈ 1.2kg/unid.)	7.19	3.95%	28.42%	0.95%
Monchique Água Mineral Natural (5L) ⇒ Monchique Água Mineral Natural (6×1.5L)	7.08	3.51%	24.86%	0.81%
Papaia Frutana de Avião ao peso (≈ 430g/unid.) ⇒ Manga Palmer Indaia ao peso (≈ 450g/unid.)	6.94	4.09%	28.42%	0.95%
Branca de Neve Farinha Fina para Bolos (1kg) ⇒ Sidul Açúcar (1kg)	6.57	4.24%	27.87%	0.90%
Abóbora ao peso (≈ 900g/unid.) ⇒ Nabo ao peso (≈ 225g/unid.)	5.95	5.39%	32.06%	0.81%

Couve Flor ao peso (≈ 1.2kg/unid.) ⇒ Nabo ao peso (≈ 225g/unid.)	5.19	4.09%	21.24%	0.84%

Tabela 9 – Regras de associação entre artigos

Em relação à tabela Tabela 9 e aos valores da primeira regra de associação – *Yoghour Especialidades ⇒ Lindahls iogurte líquido proteico de morango* – verifica-se que o valor da confiança da regra é 83.33%. Isso explica que 83.33% das transações que têm o artigo *Yoghour* também contêm o produto *Lindahls*. Por sua vez, a confiança esperada é de 1.25%, o que corresponde à percentagem de vezes que o produto *Lindahls iogurte* aparece no total das transações.

O valor do suporte da primeira regra é de 1.05%, e indica a percentagem de transações onde se verifica a combinação do antecedente e consequente, isto é, a repetição com que essa associação se apresenta no total das transações.

Por último, o *lift* desta primeira regra tem um valor superior a 1 tornando dependentes o antecedente e o consequente. O seu valor é 66.80 e isto quer dizer que a possibilidade de um cliente comprar o produto *Lindahls* aumenta aproximadamente 67 vezes mais, sempre que o produto *Yoghour Especialidades* for vendido. O mesmo acontece quando a regra é inversa.

A segunda regra mais relevante é a combinação entre *Compal Essencial De Pêssego Pack 3 embalagem 110 ml ⇒ Compal Essencial de Manga Pack 3 embalagem 110 ml*. O valor da confiança é de 66.67%, o que indica que em todos os talões com o antecedente *Compal Essencial De Pêssego* também têm o artigo *Compal Essencial de Manga*. O valor da confiança esperada mostra que o artigo consequente *Compal Essencial de Manga* surge em 2.21% de todas as transações de supermercado. O valor de suporte é 1.02%, e indica que a combinação da regra é forte, ou seja, repete-se muitas vezes no total das transações.

O valor *lift* é 30.23. Isto indica que na compra do artigo *Compal Essencial de Manga*, a probabilidade de o artigo *Compal Essencial De Pêssego* ser também vendido sobe aproximadamente 30 vezes. A combinação dos sabores pêssego e manga é forte e isso poderá traduzir um gosto do consumidor do ECI.

A terceira regra mais importante é a combinação dos artigos *Royal Gelatina de Ananás Pack 4 embalagem 100 g ⇒ Royal Gelatina de Morango Pack 4 embalagem 100 g*. O valor da métrica confiança é 56.58%, ou seja, em 56.58% dos talões com o antecedente *Gelatina de Ananás* existe também o artigo *Gelatina de Morango*. O valor da confiança esperada significa que o artigo *Gelatina de Morango* aparece em 1.98% do total dos talões de supermercado. A métrica suporte é 0.76%, e indica que a força da combinação face ao número total de transações não é suficientemente forte. Isto explica também que a combinação de produtos surge com menor frequência. Por último, o valor do *lift* é de 28.62 isto indica que sempre que o artigo *Gelatina de Ananás* é vendido, a probabilidade do consequente *Gelatina de Morango* ser vendido sobe quase 29 vezes.

Em termos gerais, verifica-se que as regras com o valor do *lift* maior correspondem aos artigos das categorias: Bebidas, Iogurtes Especiais e Sobremesas para preparar. No caso da regra que associa os artigos Compal de diversos sabores, isto pode justificar-se pela razão da elevada similaridade dos produtos, e da preferência do consumidor em selecionar diversos tipos de sabores e de consumo

rápido. As gelatinas refletem o mesmo tipo de comportamento dos sumos néctar, no que toca à escolha de sabores.

As regras com valores do *lift* mais baixos são relativas aos produtos das categorias: Legumes e Frutas, Águas e Mercearia. Dada a extensa oferta de produtos dessas categorias, são poucas as regras que sobressaem entre elas, o que justifica o baixo valor do *lift* das regras relativas aos produtos destas categorias.

Relativamente à métrica suporte, confirma-se que as regras com valores mais altos são as combinações dos artigos das categorias Iogurtes Especiais, Legumes, Frutas Tropicais e Mercearia, nomeadamente da subcategoria farinha e açúcar. Isto quer dizer que a frequência da combinação antecedente e consequente é relativamente alta face às restantes combinações no total das transações. As combinações das regras com valores de suporte mais baixos são as dos artigos da categoria Sumos Néctar Compal e Gelatinas Royal. Isso poderá estar relacionado com o número de vezes que os produtos estão em promoção ao longo do ano, ou seja, os clientes tendem a comprá-los em conjunto só quando os mesmos estão em promoção.

Sobre a métrica confiança, em termos gerais, comprovou-se que, por um lado, as combinações de artigos com valores mais altos são os Iogurtes Especiais e as Gelatinas Royal. Este acontecimento poderá estar relacionado com as campanhas promocionais que as marcas realizaram, ou também pelo facto de os clientes gostarem de combinações entre produtos de diferentes sabores. Por outro lado, as combinações com valores mais baixos são as regras com os artigos Legumes e Frutas e Águas. A baixa relação entre os artigos Legumes e Frutas poderá estar associada à especificidade das suas caraterísticas. No caso das Águas Monchique, uma possível explicação será a preferência para comprar garrafões de 5 litros para consumir em casa e a compra de garrafas de plástico de 1.5 litros para transportar facilmente para qualquer lugar. Esta marca de água sem gás é umas das mais vendidas em toda a sua categoria.

No que diz respeito à métrica confiança esperada, apurou-se que as regras com valores mais altos são os artigos das categorias Legumes e Frutas e Mercearia. A forte associação na regra de frutas tropicais poderá traduzir, novamente, a tendência e preferência de consumo do cliente deste supermercado por produtos especializados, de qualidade e de origem mais rara. Em relação aos valores de confiança mais baixos, estes são os artigos das categorias Iogurtes Especiais e Gelatinas Royal de sabores. Os baixos valores podem confirmar a ideia de que os mesmos são vendidos em conjunto apenas em situações de campanha promocional.

Em suma, verifica-se que há um maior número de associações nos artigos pertencentes às categorias Bebidas, Iogurtes, Frescos e Mercearia.

4.2. ASSOCIAÇÃO DE DEPARTAMENTOS

Os principais resultados das regras de associação dos departamentos podem ser observados na Tabela 10.

Regra	*Lift*	Confiança esperada	Confiança	Suporte
Peixaria ⇒ Talho	1.57	45.26%	71.03%	9.24%
Peixaria ⇒ Pastelaria	1.24	27.89%	34.57%	4.50%
Talho ⇒ Pastelaria	1.21	45.26%	54.93%	15.32%
Peixaria ⇒ Frutaria	1.19	74.44%	88.32%	11.49%
Talho ⇒ Congelados	1.18	55.08%	64.95%	29.39%
Pastelaria ⇒ Frutaria	1.17	74.44%	87.43%	24.39%
Pastelaria ⇒ Charcutaria	1.17	67.15%	78.55%	21.91%
Perfumaria ⇒ Drogaria e Limpeza	1.16	74.13%	85.86%	42.32%
Peixaria ⇒ Congelados	1.15	55.08%	63.54%	8.27%
Animais ⇒ Perfumaria	1.15	49.28%	56.83%	6.51%
Talho ⇒ Frutaria	1.14	74.44%	84.90%	38.43%
Vinhos e Licores ⇒ Pastelaria	1.13	27.89%	31.61%	6.89%

Tabela 10 – Regras de associação entre departamentos

A primeira regra de associação explica que há uma relação entre o antecedente *Talho* com o consequente *Peixaria*. O valor da confiança da regra indica que 20.42% das transações com artigos do departamento *Talho* também contêm artigos do departamento *Peixaria*. O valor da confiança esperada indica que 13.01% dos artigos do departamento *Peixaria* aparecem no total das transações. O valor do suporte indica que a combinação do conjunto *Talho* ⇒ *Peixaria* está presente em 9.24% do número total de transações. O baixo valor da métrica de suporte é alarmante, pois representa um número baixo da combinação no total de transações. Por último, o valor do *lift* 1.57 indica que a possibilidade de vender artigos da *Peixaria* aumenta 57% sempre que artigos do *Talho* são adquiridos. O mesmo acontece quando a regra se apresenta de forma inversa (*Peixaria* ⇒ *Talho*).

A segunda regra de associação mais importante é *Peixaria* ⇒ *Pastelaria*. O valor do *lift* é 1.24, o que indica que a possibilidade de os artigos de *Pastelaria* serem vendidos aumenta 24% sempre que os artigos da *Peixaria* são também vendidos. O valor do suporte é de 4.50%, o que demonstra que a combinação *Peixaria* ⇒ *Pastelaria* não é muito elevada, e que a frequência da regra é baixa. Comparativamente aos valores de suporte das outras regras, esta apresenta o valor mais baixo de todas. O valor da métrica confiança é 34.57%, isto é, quer dizer que essa percentagem dos talões com o antecedente *Peixaria* também contém artigos do departamento *Pastelaria*. Por fim, o valor da métrica confiança esperada é 27.89% e indica a percentagem que o consequente *Pastelaria* aparece no total dos talões. Esta regra é curiosa, porque apresenta um valor do *lift* elevado, mas baixos valores nas outras métricas.

A terceira regra de associação que merece atenção é a *Talho* ⇒ *Frutaria*, especialmente por apresentar métricas com valores bastante equilibrados e fortes. O seu *lift* é 1.14. Há, portanto, uma relação de dependência entre o antecedente e consequente da regra. Adicionalmente, a probabilidade de venda dos artigos da Frutaria aumenta em 14% sempre que os artigos do *Talho* são comprados. O valor de suporte elevado é um bom indicador, pois demonstra que a combinação está presente em 38.43% no total das transações. Ainda, o valor da confiança é de 84%, o que traduz um valor muito positivo, pois indica que em 84% de transações com artigos do antecedente *Talho* também contêm artigos com o consequente *Frutaria*. A confiança esperada desta regra também apresenta um valor elevado. Isto

indica que o consequente *Frutaria* aparece 74.44% do total das transações, o que é um bom indicador, pois quererá dizer que os artigos da *Frutaria* são os preferidos do consumidor do supermercado.

Em termos gerais, e após analisar as principais métricas, comprovou-se que na métrica *lift*, os principais conjuntos de associações entre os departamentos *Talho*, *Peixaria*, *Pastelaria*, *Frutaria* e *Drogaria e Limpeza*. O sucesso dos valores dessas métricas poderá estar ligado ao facto dos produtos desses departamentos serem considerados exclusivos, especializados e de elevada qualidade. Outra causa poderá ser a proximidade entre departamentos, e também o facto de serem os primeiros a surgir no *frontend* da *app*. Por sua vez, os departamentos *Perfumaria* e *Animais* apresentam valores de *lift* mais baixos, o que poderá estar relacionado com o facto de pertencerem aos departamentos com menor volume de vendas, e também a categorias com fins de consumo muito particulares. Além disso, esses departamentos situam-se nas últimas posições da *app*, ou seja, são os últimos corredores a aparecer ao cliente. Para chegar até eles, o cliente terá de deslizar mais vezes o dedo no ecrã.

Em relação à métrica suporte, e de modo geral, as quatro regras com o suporte mais alto são: os conjuntos dos departamentos *Drogaria e Limpeza* ⇒ *Perfumaria*, *Talho* ⇒ *Frutaria*, *Congelados* ⇒ *Talho* e *Frutaria* ⇒ *Pastelaria*. Mais uma vez, podemos explicar que a percentagem de transações que contêm essas combinações é elevada devido a pertencerem às categorias com maior volume de vendas e também devido à proximidade entre departamentos na aplicação de supermercado. Os departamentos *Talho*, *Frutaria*, *Pastelaria* e *Congelados* estão posicionados nos primeiros lugares da *app*, pelo que facilita a sua visualização e combinação de artigos entre categorias mais próximas. Por outro lado, são também reconhecidas como sendo categorias especializadas, premium, exclusivas, de elevada qualidade, diversidade e garantia. Na combinação de *Drogaria e Limpeza* ⇒ *Perfumaria*, verifica-se que apesar de estarem próximas entre si, estão afastadas das outras categorias por estarem posicionadas no final da *app*. Além disso, há artigos de limpeza e drogaria que são pesados e volumosos, tais como embalagens de detergente em pó, ou rolos de papel higiénico, pelo que se torna mais cómodo adquiri-los no canal *online*. Além disso, estão com desconto promocional várias vezes ao longo do ano. As regras com valores de suporte mais baixos são os conjuntos: *Peixaria* ⇒ *Pastelaria*, *Animais* ⇒ *Perfumaria* e *Congelados* ⇒ *Peixaria*. O baixo valor de suporte da regra *Congelados* ⇒ *Peixaria* significa que quem procura produtos congelados tem menos interesse em produtos frescos. Os produtos frescos são tipicamente mais caros, quando comparados com os respetivos congelados, o que também poderá influenciar na decisão da compra. No que diz respeito à regra *Animais* ⇒ *Perfumaria* o valor de suporte é baixo, porque o departamento *Animais* é efetivamente aquele que apresenta menor volume de vendas. Além disso, a fraca combinação entre categorias poderá estar relacionada com a fraca complementaridade entre os mesmos.

Na métrica confiança, constatou-se que as regras com valores mais elevados são: *Perfumaria* ⇒ *Drogaria e Limpeza*, *Peixaria* ⇒ *Frutaria*, *Pastelaria* ⇒ *Frutaria* e *Pastelaria* ⇒ *Charcutaria*.

O conjunto *Perfumaria* ⇒ *Drogaria e Limpeza* poderá estar associado ao facto dos dois departamentos se encontrarem muito próximos um do outro e por serem ambos de limpeza. Também quer dizer que o cliente que compra um produto de limpeza pessoal também tem tendência a comprar um produto de limpeza para o seu lar.

A combinação *Pastelaria* ⇒ *Charcutaria* é curiosa. No entanto, pode estar relacionada com um hábito gastronómico do consumidor português em combinar pão com enchidos. A relação *Peixaria* ⇒ *Frutaria* acontece devido à proximidade entre os dois departamentos na aplicação.

Em relação, às combinações com os valores de confiança mais baixos estes são: Perfumaria ⇒ Animais, Frutaria ⇒ Peixaria; Pastelaria ⇒ Peixaria. Note-se, apesar disso, que quando a regra é inversa, os valores de confiança são elevados.

Relativamente à métrica confiança esperada comprovou-se que as combinações com valores mais altos são: *Frutaria* ⇒ *Talho*, *Charcutaria* ⇒ *Pastelaria*, *Frutaria* ⇒ *Peixaria*, *Drogaria e Limpeza* ⇒ *Perfumaria*, e *Congelados* ⇒ *Peixaria*. Isto quer dizer que seus consequentes no total das transações de supermercado são bastante significativos face às restantes associações.

As regras cuja confiança esperada com valores mais baixos são: *Animais* ⇒ *Perfumaria*, *Peixaria* ⇒ *Congelados*, *Congelados* ⇒ *Talho*, *Pastelaria* ⇒ *Talho* e *Talho* ⇒ *Peixaria*. Isto quer dizer que a percentagem do seu consequente no total das transações é baixo relativamente às outras regras de associação.

4.3. Análise de Clusters

Como abordado no ponto da metodologia, o método de *clustering* ou segmentação serve para classificar elementos em grupos, e no âmbito deste projeto, serviu para agrupar conjuntos de clientes consoante as suas variáveis de interesse. Tipicamente, o que se pretende na segmentação é obter *clusters* o mais homogéneos possível dentro de si, o mais heterogéneos possível entre si.

Neste estudo, o principal objetivo da análise de *clusters* é descobrir os diferentes tipos de clientes, de acordo com as segmentações realizadas para o valor e para o consumo. Relativamente à segmentação de clientes pelo valor, verificou-se que 3 é o número ideal de *clusters* para esta segmentação. Esta confirmação foi obtida através dos resultados da aplicação do método *Elbow Curve*, como se confirma no gráfico da Figura 8. O mesmo evidencia claramente que o decréscimo do R^2 se torna menos acentuado a partir dos 3 *clusters*. O gráfico também explica a relação entre o número de *clusters* e respetivo erro. A qualidade da divisão dos clientes pelos 3 *clusters* é explicada pelo valor de 40.2% do RSQ, ou R^2. Quanto maior for o valor do RSQ melhor será a qualidade dos *clusters*.

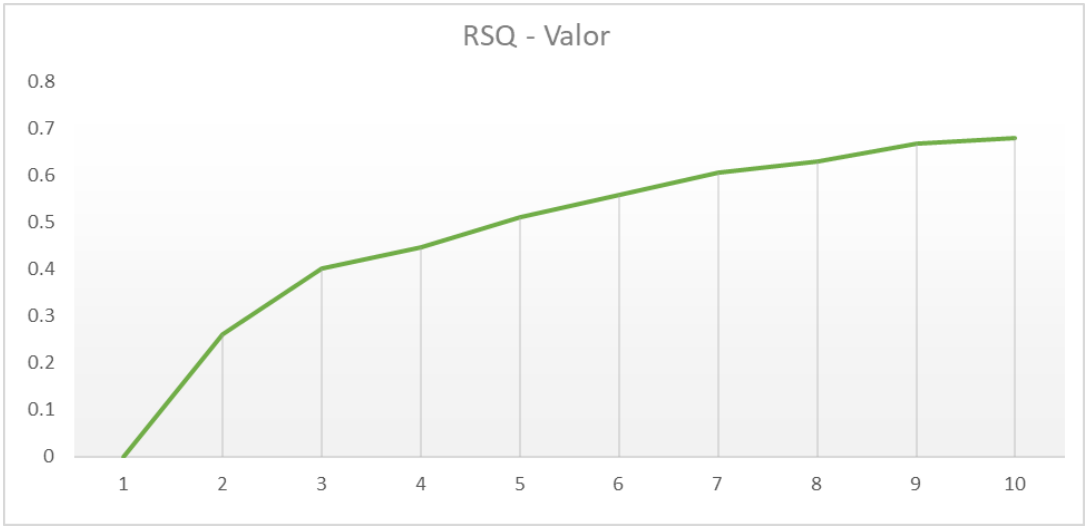

Figura 8 – R^2 em função do número de clusters na segmentação por valor

No que diz respeito à segmentação por consumo, a escolha do número de *clusters* foi também realizada através do mesmo método. De acordo com o gráfico da Figura 9, 4 *clusters* é também o

número ideal para esta segmentação. Como se confirma na figura, o valor do RSQ relativo à divisão dos clientes em 4 *clusters* de elementos distintos é de 26.6%.

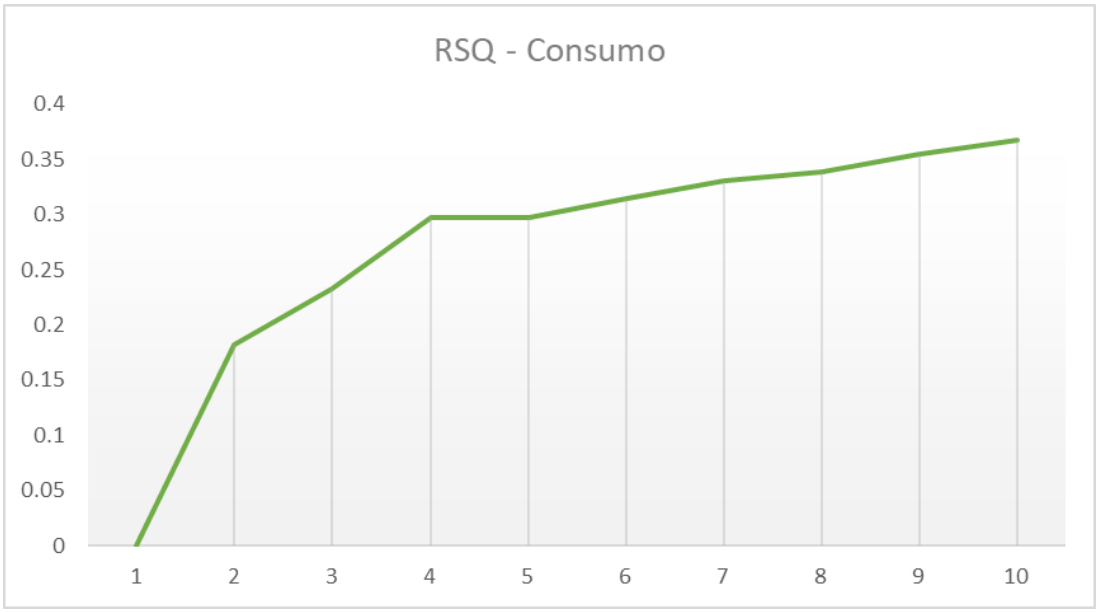

Figura 9 – R^2 em função do número de clusters na segmentação por consumo

4.3.1. Segmentação por valor

Para melhor compreender os *clusters*, procedeu-se à analise dos dados estatísticos das tabelas Tabela 12 e Tabela 13, presentes na secção 5, e que são caracterizados com a ajuda da Figura 10.

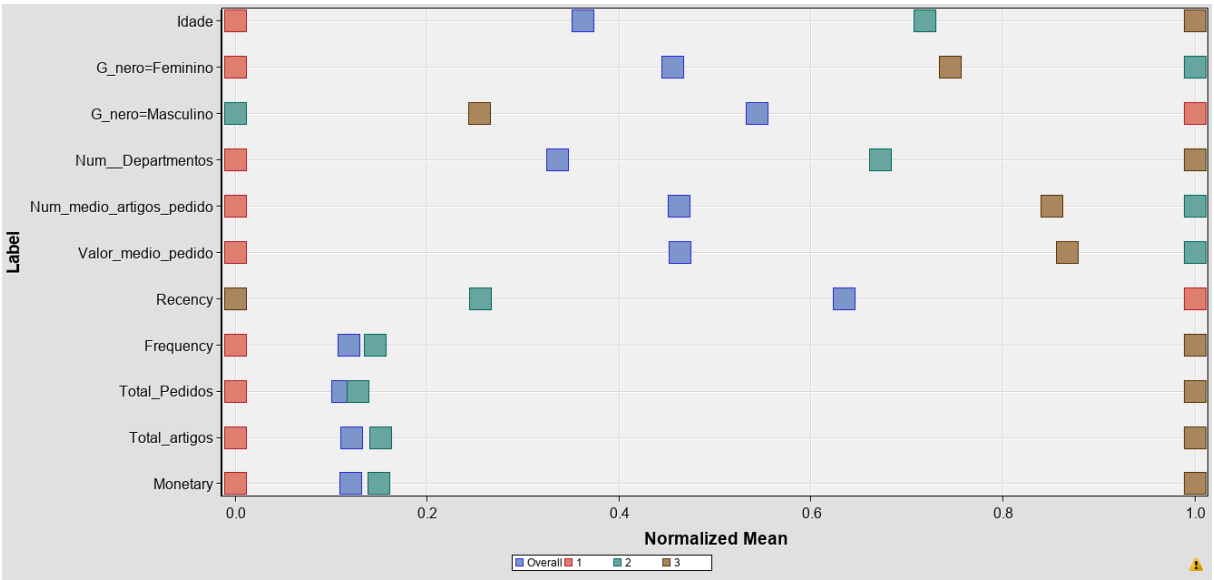

Figura 10 – Input means plot resultante da segmentação por valor ($k = 3$)

O segmento 1 – *Passageiros* – é o *cluster* com maior número de clientes (1046), dos quais 60.6% são do género feminino, e 31.4% do género masculino, perfazendo 52.7% dos clientes. Este *cluster* tem uma média de idade de 41.7 anos, que não difere muito da média de idades dos outros segmentos. O seu valor de recência é alto (49.9), e indica que a média de tempo desde a última compra é quase um ano, espelhando o grupo de clientes que fez compras menos recentemente. Ao nível dos valores das

variáveis de frequência (1.9) e valor monetário (182.39€) verifica-se que são clientes que gastam pouco e visitam com menor frequência. O número médio de artigos por cada pedido é de 35.2 itens. A média do total de pedidos é de 2. O número total de departamentos é de 8.8. O valor médio de pedido é de 94.65€. Este grupo é considerado importante, pois são clientes com um valor monetário baixo, mas a nível numérico trata-se do maior grupo, podendo representar futuros clientes passíveis de serem fidelizados. Conclui-se que para este segmento dever-se-á adotar estratégias de comunicação *online*, por serem mais económicas que as *offline*. Além disso, este grupo de clientes parece ser sensível a descontos, promoções diretas e indiretas, vales de desconto, campanhas de pague 2 e leve 3, pelo que poderão ser facilmente alvo de campanhas de *e-mail marketing*, mensagens *push* e publicidade nas redes sociais.

O segmento 2 – *Indecisos* – é o segundo maior grupo de clientes, e é constituído por 815 clientes, isto é, 41.06% do total de clientes. O seu público pertence maioritariamente ao género feminino, com 79.8% dos elementos desse *cluster*. A idade média é 44.4 anos. Estas clientes têm um valor de recência baixo (15.3), o que comprova que não compram continuamente. Os valores de frequência de compra são intermédios, ou seja, a média de número de pedidos é de 5.9. O valor monetário é alto (1035.48 €), o que indica que os elementos têm alto poder de aquisição. Tanto o número médio de artigos por pedido (75.1), o número total de pedidos (6.3), o número total de artigos comprados (373.6), como o valor médio por pedido (211.39 €) também se mostram superiores face ao primeiro segmento. Por último, a média de departamentos também é alta (12.7) comparativamente ao primeiro segmento. Estes clientes, apesar de terem os valores da frequência e recência competitivos, precisam de um reforço positivo, por forma a incentivar a frequência de visitas. Os elementos costumam comprar muitos artigos, de vários departamentos, muitas vezes, pelo que se devem aplicar políticas de recomendação e partilha da *app*, de modo a que estes elementos referenciem este serviço na sua comunidade de família e amigos. Através de um sistema de recomendação entre clientes da *app*, seria possível aumentar o número de clientes deste tipo. É crucial investir numa comunicação *online* e *offline*, por forma a garantir a captação da atenção do maior número de clientes. Criar eventos e experiências gratuitas e relevantes com base nas campanhas promocionais poderá ser uma forma de atraí-los para a experiência de compra *mobile*. Manter o interesse através da oferta de vales de desconto, cheques de oferta e descontos diretos ajudará não só a aumentar as interações na *app*, como a reter um maior número de clientes.

O segmento número 3 – *Clientes fiéis* - é um grupo restrito de clientes (117), 5.89% dos clientes. Os clientes são maioritariamente do género feminino, e a sua idade média é de 45.4 anos. O valor da recência é baixo (3.4), e isto reflete que os pedidos são muito recentes. O valor da frequência de compra é 29.6 semanas, e valor monetário é o mais alto (5876.61 €), o que significa que gastam muito. Também o número médio de artigos por pedido (69.1) e o número total de pedidos (35.2) são os valores mais altos em comparação com os outros grupos. O valor médio por pedido é de 195.89 €, e é o segundo valor mais alto do total de clientes. A média de departamentos é de 14.6, o que significa que é mais alta de todas, e que estes clientes têm a tendência de comprar muitos artigos de diversos departamentos.

Esses valores quererão traduzir que estes clientes poderão pertencer a um agregado familiar extenso e com elevado poder de compra, e que têm necessidade de comprar produtos específicos com características particulares, e de departamentos diversos, de modo a satisfazer as necessidades dos membros de toda a família. Estes clientes são fiéis à empresa e aos seus produtos, pois o seu

comportamento de compra é contínuo e consistente ao longo do período em análise. Enfim, com base nos valores dos resultados das variáveis, conclui-se que estes clientes são o melhor grupo, pelo que é necessário agir com muita atenção, pois são os menos sensíveis ao preço, os mais fiéis, e os mais valiosos para a empresa. De modo a assegurar a satisfação do grupo, devem-se assumir custos de fidelização, e assim adotar medidas de gratificação, como por exemplo a oferta do passe de entregas ao domicílio durante 6 meses, a oferta de cabazes de produtos na data de aniversário, ou numa data festiva específica, ou a oferta de *workshops* de alimentação saudável e de gastronomia em geral.

4.3.2. Segmentação por consumo

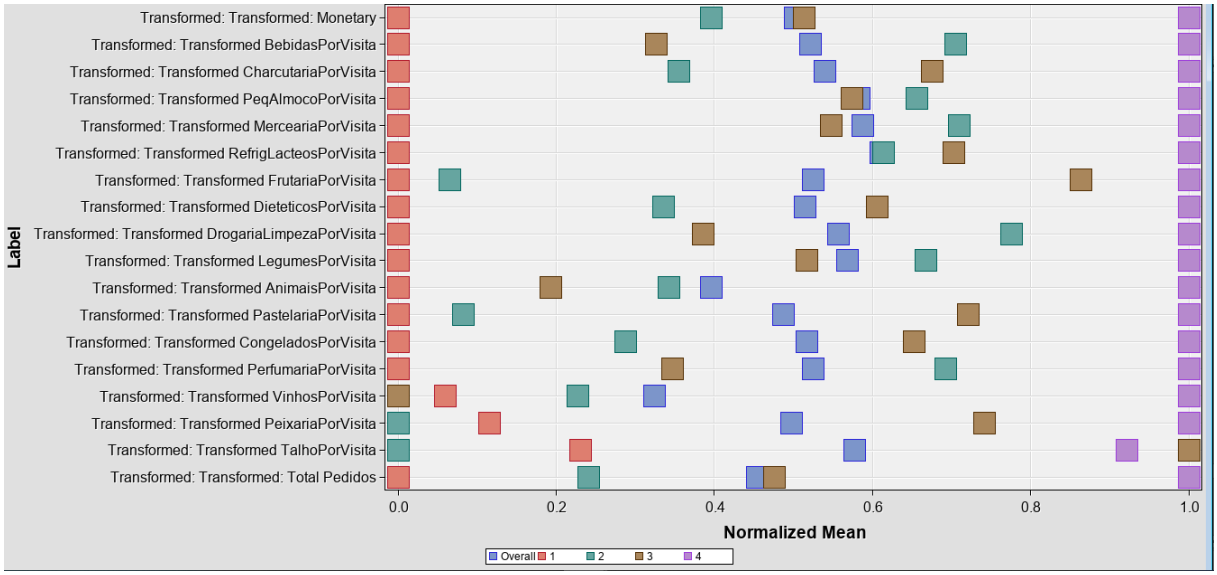

Figura 11 – Input means plot resultante da segmentação por consumo ($k = 4$)

O grupo relativo ao segmento 1 – *Baixo Consumo* – é constituído por 416 elementos, ou seja, corresponde a 21% de todos os clientes. O seu valor monetário gasto é de 73.89 €, e o número médio de pedidos realizados é de 1.4 pedidos, resultando em 52.77 € gastos por pedido. Após analisar o gráfico da Figura 11, verificou-se que este segmento, comparativamente aos segmentos 2 e 3, compra mais artigos dos departamentos Vinho e Peixaria do que os outros. No que diz respeito à análise feita ao segmento, e recorrendo aos dados estatísticos da Tabela 15 – Resultados estatísticos para a segmentação por consumo na secção 8.2, comprovou-se que o grupo é caracterizado por comprar muito pouco, em quase todos os departamentos. Os departamentos com maiores consumos são: Bebidas (2.18 €), Frutaria (2.32 €), Drogaria e Limpeza (1.99 €), Refrigerados e Lácteos (1.62 €), Mercearia (1.56 €), Pequeno-almoço (1.54 €), Charcutaria (1.53 €), e Talho (1.17 €). Os departamentos com menor consumo por pedido são: Animais (0.27 €), Pastelaria (0.28€), Peixaria (0.30 €), Vinhos (0.88 €), Congelados (0.94 €), Legumes (0.95€) e Dietéticos (0.98 €). Este grupo de clientes é o que menos consome, pelo que é importante aumentar a relação entre a empresa e os clientes através do aumento das interações, e promoção de produtos de compra habitual deste segmento.

O segmento número 2 – *Consumidores de não perecíveis* – é constituído por 478 elementos, isto é, 24.1% do total de clientes. A média do valor monetário total é 241.52 €, e a média do número total de pedidos é de 2.2 pedidos. O valor médio por pedido é 109.78 €. Este segmento em comparação com o segmento *Baixo Consumo*, gasta mais em quase todos os departamentos, menos nas categorias Peixaria e Talho. Face ao segmento número 3 (*Consumidores de perecíveis*), comprova-se que gasta

mais nas categorias Bebidas, Pequeno-almoço, Mercearia, Drogaria e Limpeza, Legumes, Animais, Perfumaria e Vinhos. Porém, este segmento número 2 tem o valor monetário e número de pedidos inferior ao segmento número 3. Quando comparado com o segmento número 4, este segundo segmento apresenta valores de consumo bem abaixo. Finalmente, este segmento gasta maioritariamente nos departamentos de Drogaria e Limpeza (14.96 €), Bebidas (8.36 €), Mercearia (7.84 €), Refrigerados e Lácteos (7.26 €) e Pequeno-almoço (6.99 €). Todos os outros restantes departamento têm valores de consumos pouco significativos, isto é, Animais (0.48 €), Peixaria (0.20 €), Pastelaria (0.33 €), Talho (0.36 €), e Vinhos (1.25 €).

Este grupo é caracterizado por ser sensível ao preço, devido ao seu valor por pedido elevado, e por comprarem poucas vezes. No entanto, este grupo é igualmente valioso para a empresa, pelo que é importante investir na sua satisfação e lealdade. Programas como a divulgação de promoções, e descontos em produtos que habitualmente compram são boas estratégias para combater a sensibilidade ao preço. Além disso, deve-se aumentar a relação e interação entre a empresa e esse grupo de clientes, através da partilha de experiências e de conteúdos relevantes para os mesmos.

O segmento 3 – *Consumidores de perecíveis* – tem o maior número de elementos (581) e representa 29.3% dos clientes. O valor monetário total é 342.33 €, a média total de pedidos é 3.3, e o valor médio gasto por pedido é de 103.73 €. Este segmento, em comparação com os segmentos 1 e 2, tem os valores das variáveis monetária e número de pedidos superiores. No entanto, face ao segmento 4, apresenta valores de consumo inferiores. Em relação aos departamentos foi possível verificar que este segmento gasta mais do que o segmento 1 em quase todos os departamentos, à exceção do departamento Vinhos. Em comparação com o segmento 2, este terceiro segmento gasta mais nos seguintes departamentos: Charcutaria, Refrigerados e Lácteos, Frutaria, Dietéticos, Pastelaria, Congelados e Peixaria. Comparativamente ao segmento número 4, este grupo consome muito menos em quase todos os departamentos, exceto na categoria Talho. O segmento 3, a nível individual, gasta mais nos departamentos Frutaria (11.40 €), Talho (9.27 €), Refrigerados e Lácteos (8.77 €), Congelados (6.17 €), Pequeno-almoço (5.93 €), Drogaria e Limpeza (5.89 €) e Charcutaria (5.76 €), e consome menos nas categorias de Animais (0.38 €), Pastelaria (0.82 €), Peixaria (1.04 €) , Dietéticos (3.32 €) e Vinhos (0.76 €). Este é, sem dúvida, o segundo grupo de clientes mais importante. Para este grupo é necessário motivar o aumento do consumo das categorias mais vendidas através da divulgação de campanhas promocionais de produtos, ofertas promocionais, descontos diretos e indiretos, campanhas temáticas, entre outras.

O quarto segmento – *Elevado Consumo* – é o segundo maior grupo de clientes e tem 510 elementos, o que corresponde a 25.7% dos clientes. Também tem o maior valor médio monetário (1452.70 €), o maior número de pedidos (7.1), e o maior valor médio gasto por pedido de 204.60 €. Este segmento é o mais importante de todos os grupos, consumindo muito em todos os departamentos, exceto na categoria Talho (7.77 €), onde é o segmento 3 que supera todos os gastos nesse departamento. Para este grupo que contém os clientes mais valiosos, mais leais e menos sensíveis ao preço, é necessário adotar uma estratégia de comunicação de incentivo à compra dos produtos habituais. Outra estratégia possível será através do *e-mail marketing*, mensagens *push,* enviar mensagens escritas com vales de oferta, e oferta de experiências em datas comemorativas, tais como no aniversário.

Em suma, é interessante comprovar que os departamentos com menor consumo em todos os segmentos são: Animais, Dietéticos, Pastelaria, Peixaria e Vinhos. No extremo oposto, os departamentos com maior consumo são: Drogaria e Limpeza, Frutaria Mercearia, Pequeno Almoço e

Refrigerados e Lácteos. De uma forma geral, a empresa deve adotar uma abordagem comercial diferente na venda de artigos dos departamentos menos consumidos.

4.3.3. Cruzamento de segmentações

		Consumo				
		Baixo Consumo	**Não Perecíveis**	**Perecíveis**	**Elevado Consumo**	**Todos**
Valor	**Passageiros**	13.2%	12.7%	16.1%	10.6%	**52.7%**
	Indecisos	6.8%	10.1%	11.9%	12.3%	**41.1%**
	Fiéis	0.8%	1.2%	1.2%	2.8%	**5.9%**
	Todos	**20.8%**	**24.0%**	**29.2%**	**25.7%**	**100.0%**

Tabela 11 – Cruzamento dos resultados das segmentações por valor e por consumo

Por último, e ainda na análise de *clusters*, foi importante juntar os resultados das duas segmentações (consumo e valor) de modo a descobrir a percentagem de tipo de clientes por combinação de *clusters*, ou seja, descobrir o comportamento de consumo dos *clusters* de valor e o tipo de valor dos *clusters* de consumo. Conforme a Tabela 11 – Cruzamento dos resultados das segmentações por valor e por consumo, foi possível compreender que o *cluster Passageiros* da segmentação por valor representa 55.1% do *cluster* Perecíveis (16.1% do total de clientes). Isto significa que a nível de comportamento de consumo, gastam mais nos departamentos Frutaria, Talho, Refrigerados e Lácteos, Congelados, Pequeno-almoço Drogaria e Limpeza e Charcutaria, e consomem menos nas categorias de Animais, Pastelaria, Peixaria, Dietéticos e Vinhos. Uma possível estratégia de captação destes clientes que consomem menos será o envio notificações sobre uma promoção de desconto direto nos produtos dos departamentos de Animais, Pastelaria, Peixaria, Dietéticos e Vinhos. Dado que são clientes que compram num vasto leque de departamentos também se poderá aplicar uma campanha de 50% desconto em talão para aumentar assim a probabilidade desses clientes regressarem. Aqui também a oferta de entregas ao domicílio gratuitas poderá aumentar o consumo.

O *cluster Indecisos* é um grupo muito importante, porque apresenta uma frequência e valor monetários intermédios, mas valor médio por pedido alto. Estes representam 47.9% do *cluster Elevado Consumo* (12.3% do total de clientes). Isto quer dizer que a nível de consumo, os *Indecisos* gastam mais em todos os departamentos, exceto em Talho. Para este grupo, é fundamental aumentar o seu número de visitas, e, para isso, uma sugestão seria a oferta de vales de desconto numa próxima visita, e numa categoria específica, enviados mensalmente sob a forma de mensagem *push*.

Relativamente ao *cluster Fiéis*, este é o grupo com maior valor para a empresa, e representa 10.9% do *cluster Elevado Consumo* (2.8% do total de clientes). Em relação, ao consumo gasta mais em todos os departamentos, menos em Talho. Para estes clientes é necessário manter as estratégias, uma vez que se revelam eficazes para este grupo.

4.3.4. Reordenação de corredores

Após obter e analisar os resultados da associação de produtos e da segmentação de clientes, optou-se por aplicar estratégias diferentes para cada tipo de resultados. Os resultados da associação de produtos de departamentos foram utilizados para reordenar os corredores de produtos da *app* de

supermercado, enquanto que os resultados da segmentação de clientes serviram para redefinir a comunicação das mensagens *in-app*.

Relativamente à reordenação dos corredores, a *app* do supermercado é composta atualmente por um menu que contém corredores (categorias), e estes, por sua vez, subdividem-se em subcorredores (subcategorias) com lineares de produtos. Estes estão ordenados por ordem numérica e categórica. Por exemplo, o corredor número 1 corresponde a uma categoria de produtos gerais, o corredor 2 à categoria das campanhas promocionais, o terceiro às ofertas, o quarto às bebidas, o quinto às marcas ECI, e assim por diante. Normalmente, as categorias mais importantes (com maior peso monetário) ocupam as primeiras posições numéricas e as categorias menos importantes ocupam as últimas posições numéricas. Os primeiros corredores estão sempre ordenados nas quatro primeiro posições e são reservados tanto para destaques de campanhas internas da empresa como para destaques das marcas. Os subcorredores estão ordenados por relevância monetária e também seguem a mesma estrutura de ordenação do supermercado *online*. Os lineares não apresentam um critério de organização bem definido, à exceção da proximidade de produtos da mesma família e marca.

De acordo com os resultados obtidos, proponho várias alterações na organização e disposição dos corredores, subcorredores e lineares da aplicação. Atualmente, a ordenação dos corredores é a seguinte:

1. Promoções
2. Campanhas (Folheto)
3. Ofertas
4. Destaques de marcas
5. Bebidas
6. As nossas marcas
7. Vinhos
8. Vinho do Porto. Licores e Destilados
9. Charcutaria e Queijos embalados
10. Charcutaria e Queijos ao corte
11. Padaria e Pastelaria
12. Pequeno-almoço
13. Talho
14. Peixaria
15. Frutas e Legumes
16. Lácteos e ovos
17. Iogurtes e sobremesas
18. Congelados
19. Refeições prontas
20. Comida do mundo
21. Produtos biológicos
22. Dietéticos
23. Produtos para celíacos
24. Conservas
25. Aperitivos e frutos secos
26. Mercearia

27. Sobremesas para preparar
28. Chocolates e doces
29. Bolachas
30. Doces, marmelada e mel
31. Azeites vinagres e especiarias
32. Bebé e criança
33. Perfumaria e cosmética
34. Homem
35. Higiene
36. Saúde e bem-estar
37. Limpeza
38. Drogaria
39. Cão
40. Gato
41. Outros animais

A nova ordenação proposta seguirá a seguinte estrutura:

1. Promoções
2. Campanhas (Folheto)
3. Ofertas
4. Destaques de marcas
5. As nossas marcas
6. Talho
7. Peixaria
8. Padaria e Pastelaria
9. Frutas e Legumes
10. Congelados e refeições prontas
11. Charcutaria
12. Drogaria e Limpeza
13. Higiene, saúde e bem-estar
14. Bebé e criança
15. Perfumaria e Cosmética
16. Vinhos, Licores e Destilados
17. Biológicos e Dietéticos
18. Doçaria
19. Mercearia
20. Comida do mundo
21. Bebidas
22. Lácteos, ovos e sobremesas
23. Animais

Figura 12 – Exemplo de ecrã com o menu dos novos corredores da app (1/3)

Figura 13 – Exemplo de ecrã com o menu dos novos corredores da app (2/3)

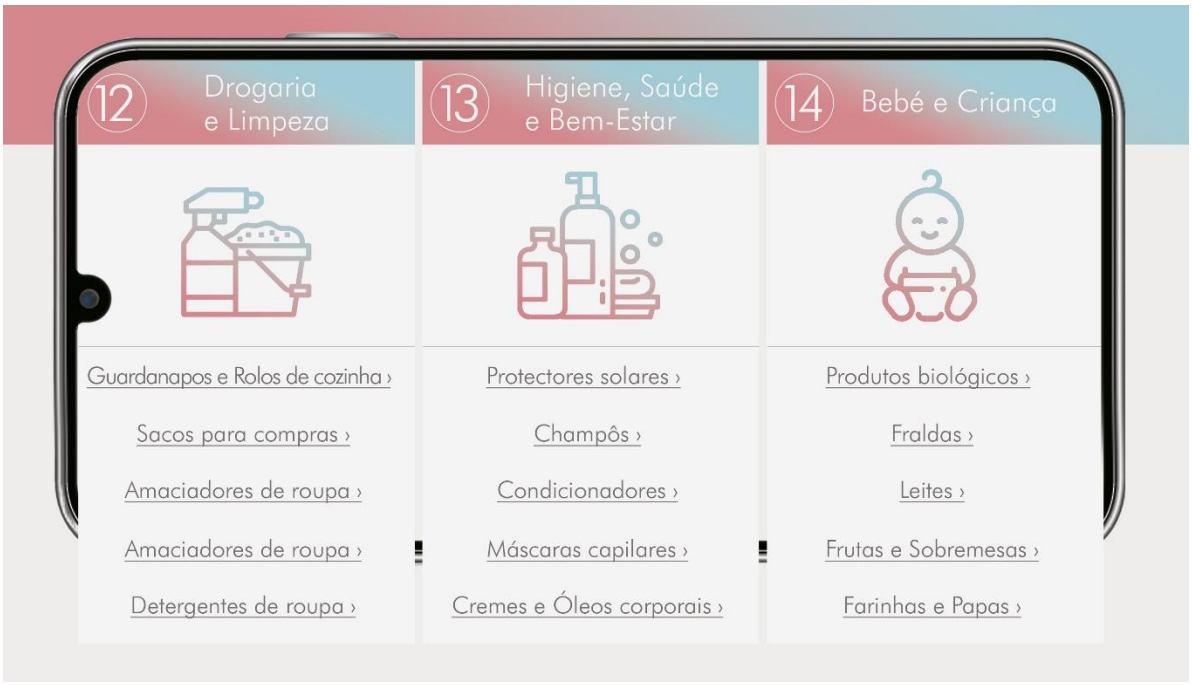

Figura 14 – Exemplo de ecrã com o menu dos novos corredores da app (3/3)

Nesta nova estrutura, aglomeraram-se vários corredores num só. Por um lado, isto simplifica a navegação pelos corredores, juntando produtos de categorias muito próximas. Por outro, esta união amplifica o *up-selling* e *cross-selling* desses mesmos produtos, que anteriormente estavam hierarquicamente distantes, mas categoricamente próximos. Prevê-se que esta proposta venha a aumentar as vendas destas categorias com menor volume de vendas e que têm menor visibilidade para o utilizador.

Os corredores que nesta proposta sofreram esta aglomeração são diversificados. Em primeiro lugar, o novo corredor da Mercearia comportará os antigos corredores de Mercearia, Azeites e Vinagres, Pequeno-almoço e também Aperitivos e Frutos secos. Em segundo, proponho a criação de um novo corredor Doçaria, onde ficarão contidos os anteriores corredores Sobremesas para preparar, Doces, Marmelada e Mel, Chocolates e Doces e Bolachas. Proponho ainda um único corredor Congelados, que inclui também as Refeições prontas, bem como um único corredor para Animais, que antes estava a dividido entre Cão, Gato e Outros animais. Por fim, sugiro a união do corredor Homem no de Higiene, e a ainda a sua aglomeração com o corredor de Saúde e bem-estar.

A nova sequência de corredores também poderia sofrer alterações de acordo com as regras de associação extraídas anteriormente. À exceção dos primeiros corredores (Promoções, Campanhas, etc.), os que os sucedem, e, portanto, surgem primeiramente, passam a ser o Talho, Peixaria, Pastelaria, Frutaria, Congelados, Charcutaria, Drogaria e Limpeza, Perfumaria, e Vinhos e Licores. Estes corredores movem-se para o topo da lista por apresentarem as regras de associação entre departamentos com valores de *lift* mais elevados. Já o corredor dos Animais, apesar de mostrar uma boa associação com o departamento de Perfumaria, este apresenta valores monetários muito baixos, pelo que se decidiu mantê-lo no fim da lista de corredores. No entanto, e por forma a potencializar as vendas deste departamento, sugiro que o mesmo seja, sempre que oportuno, reposicionado nas primeiras posições da *app*.

Relativamente à disposição de produtos nos lineares, os resultados das regras de associação por artigos sugerem uma alteração na organização dos mesmos. Estes resultados mostram que produtos do mesmo tipo e da mesma marca, tendem a ser comprados em conjuntos de diferentes sabores, tais como sumos da marca Compal, gelatinas da marca Royal, ou mesmo iogurtes da marca Lindahls. Esta observação pode ser estendida para muitos outros produtos, pelo que a nova disposição dos produtos nos lineares passa por agrupar estes produtos, não só por sabor, mas por marca e tamanho.

Uma estratégia adicional para a associação de produtos será a venda de espaços publicitários nos lineares aos fornecedores das marcas com produtos cujas regras de associação sejam mais predominantes.

O modelo de associação também é útil para recomendar os artigos com maior associação dentro da aplicação. Além disso, esse modelo é ótimo para o tipo de clientes que passam muito tempo a ver produtos na *app*, mas que compram pouco. A recomendação de produtos ajudará os clientes a encontrar intuitivamente e eficazmente produtos que possivelmente desejam comprar. Outra vantagem do sistema de recomendação é o aumento da venda do número de artigos por pedido através da estratégia *cross-sell*, ou seja, ao sugerir outros tipos de produtos da mesma categoria levará o cliente a comprar mais artigos. Por último, a recomendação de produtos a partir da sua associação fortalecerá a lealdade do cliente com a empresa. O facto de o cliente saber que a empresa onde compra já conhece os seus hábitos de consumo, as suas preferências e as suas necessidades fará com que ele opte pela empresa habitual e não pela concorrência que não o conhece.

4.3.5. Estratégias de conteúdos informativos comerciais

A segmentação de clientes permitirá enviar mensagens personalizadas de acordo com as preferências de consumo dos vários *clusters*. A personalização de conteúdos permitirá não só criar uma relação de proximidade com os clientes, mas também adequar os conteúdos informativos de acordo com o conhecimento obtido dos mesmos, ao contrário do que acontece com a publicidade massificada. Atrair os clientes através de uma forma objetiva e personalizada levará à fidelização do cliente à marca. Além do mais a personalização de conteúdos permitirá evitar desperdício de tempo e recursos em companhas para grupos de consumidores que nunca estarão interessados naqueles produtos.

Cada segmento de clientes tem um consumo diferente nos diversos departamentos, pelo que para cada grupo se recomenda o envio de mensagens *push*, aquando de campanhas promocionais (promoções, descontos diretos e indiretos, ofertas de produtos dessas categorias) de produtos dessas categorias. Através destas estratégias conseguir-se a atrair novos clientes e sobretudo vender mais aos clientes fidelizados.

Além disso, também é aconselhável estabelecer contactos com os fornecedores dessas categorias, por forma a incluir este serviço de comunicação no plano comercial anual de publicidade. O serviço de mensagens *push notification* é um excelente canal de comunicação entre fornecedores e os clientes da empresa que consomem essas categorias de produtos. Através dessas mensagens, as marcas poderão promover lançamento de novos produtos, divulgar promoções e passatempos, e comunicar ofertas exclusivas. Por outro lado, a empresa aumentará a fidelização dos seus clientes, pois oferecerá conteúdo referente às preferências de consumo dos seus clientes.

Outra oportunidade que o estudo de segmentação oferecerá é a aplicação de estratégias de *cross-selling* e *up-selling* de produtos de diferentes categorias, através do envio de mensagens, contendo promoções nos departamentos mais consumidos tornará a experiência de compra mais pessoal.

Com esta informação, a empresa poderá analisar detalhadamente o motivo de existirem departamentos com mais e com menos vendas, e posteriormente encontrar soluções para promover as vendas nos departamentos menos populares, e melhorar os departamentos mais populares. Para os departamentos menos populares, fica a sugestão de realizar um estudo mais detalhado, por forma a perceber melhor quem são esses clientes, e como podemos angariar novos clientes com características similares. Além disso, recomenda-se criar campanhas promocionais exclusivas para os departamentos com menor consumo. A angariação de clientes de faixas etárias mais baixas e mais altas também deve ser tomada em consideração, por forma a aumentar o nível de heterogeneidade dos clientes.

5. CONCLUSÕES

Este trabalho teve como principais objetivos procurar conhecer os hábitos de consumo, os padrões e perfis de clientes e sobretudo as correlações entre artigos. Ao atingir estes objetivos foi possível conhecer novas maneiras de organizar a disposição dos artigos e corredores da *app* de supermercado. Também foi possível conhecer os diversos perfis de clientes e assim redefinir estratégias de comunicação.

A partir da análise descritiva, realizaram-se duas análises distintas. A primeira foi referente aos resultados obtidos das regras de associação, tanto para os departamentos, como para os artigos. A segunda análise incidiu sobre os resultados da segmentação de clientes por valor e por consumo. Relativamente à primeira análise, concluiu-se que as regras de associação entre departamentos com valores de *lift* mais elevados são as regras que relacionam o Talho, Peixaria, Pastelaria e Frutaria, Drogaria e Limpeza. Os departamentos cuja associação tem menor *lift* são a Perfumaria e os Animais. Estes resultados serviram para a reordenação dos corredores da *app*. Já os valores das regras de associação dos artigos serviram para redefinir as estratégias de ordenação de produtos nos lineares. Além disso, as regras obtidas puderam ser estendidas a muitos outros produtos, pelo que a nova disposição dos produtos nos lineares passa por agrupar estes produtos, não só por marca e tamanho, mas também por sabor. Outra ação que advém da associação de produtos é a venda de espaços publicitários nos lineares aos fornecedores das marcas com produtos cujas regras de associação são mais predominantes. Uma última ação resultante do modelo de associação é a recomendação de produtos, dentro da *app*, de acordo com as associações entre artigos e departamentos.

Em relação à análise da segmentação de clientes, optou-se por fazê-lo em função do valor e do consumo. Ambas as segmentações foram utilizadas para redefinir as estratégias de comunicação, utilizando conteúdo de acordo com o tipo de segmentação.

Em suma, este projeto foi pioneiro para o departamento de supermercado, e serviu para obter dados que eram antes desconhecidos tanto relativamente a produtos e departamentos, como aos clientes. Os resultados obtidos neste estudo serão apresentados à direção e gestores departamentais para validação e implementação.

6. LIMITAÇÕES E RECOMENDAÇÕES PARA FUTURO TRABALHO

Um dos maiores desafios deste projeto foi a obtenção e tratamento dos dados. A empresa dispõe de várias plataformas de dados de clientes, pelo que a informação está dispersa e descentralizada. Conseguir reunir toda a informação de uma só vez e através de um só sistema de informação foi desafiante. O tratamento de dados foi igualmente desafiante, pois foi necessário proceder a muitas transformações até conseguir dados limpos e objetivos.

Uma recomendação importante será desenvolver um modelo preditivo, utilizando o modelo de árvores de decisão, para descobrir quais os *clusters* de clientes que compram os artigos do folheto promocional quinzenal do supermercado.

A segmentação de clientes servirá também para a empresa conhecer mais sobre os seus clientes. Através de pequenos questionários de satisfação, poderá, sempre que oportuno, aprofundar mais o conhecimento sobre os seus hábitos de consumo, comportamento de compra, e grau de satisfação.

Recomenda-se ainda o acrescento de mais dados sociodemográficos (profissão, habilitações, rendimentos, atividades, interesses pessoais, dietas alimentares, etc.) aos segmentos de clientes, por forma a criar e desenvolver *personas*, ou seja, personagens fictícias que representam os consumidores de cada tipo de *cluster*. Criar *personas* ajudará a empresa a compreender as necessidades, experiências, comportamentos e objetivos dos seus consumidores. Cada tipo de cliente tem diferentes tipos de necessidades, desejos e expectativas, pelo que deverá ser tratado de forma única e personalizada. Conhecê-lo ao pormenor permitirá desenvolver produtos, serviços, estratégias de *marketing* e conteúdos informativos exclusivos e pessoais. Aumentar e melhorar a relação através da humanização de serviços deverá ser um objetivo a atingir.

7. BIBLIOGRAFIA

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *20th Int. Conf. Very Large Data Bases (VLDB)*, (pp. 487–499). Santiago.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 1993 ACM SIGMOD Conference*, (pp. 1-9). Washington.

Azevedo, A., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A parallel overview. *IADIS European Conference on Data Mining* (pp. 182-185). Amesterdão: ResearchGate.

Azevedo, C. S., & Santos, M. F. (2005). *Data Mining: Descoberta de Conhecimento em Bases de Dados* . Lisboa: FCA- Editora Informática.

Barros, G. (2018). *El Corte Inglés: Organização do espaço de loja da parafarmácia.* Lisboa: Nova IMS Buisness School.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine Volume 17 Number 3*, 37-48.

Gama, J., Carvalho, A. P., Faceli, K., Lorena, A. C., & Oliveria, M. (2015). *Extração de Conhecimento de Dados: Data Mining.* Lisboa: Edições Sílabo.

Gil, A. C. (2008). Métodos e Técnicas de Pesquisa Social. Em *Métodos e Técnicas de Pesquisa Social* (Sexta edição ed., pp. 09-28). São Paulo: Editora Atlas.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining Concepts and Techniques, 3rd-Edition.* Waltham: Morgan-Kaufmann.

Henderson, G. R., & Rank-Christman, T. (2016). Diversity and consumer behavior. *Current Opinion in Psychology*, 148-153.

Kurniawan, F., Umayah, B., Hammad, J., Nugroho, S. M., & Hariadi, M. (2018). Market Basket Analysis to Identify Customer Behaviors by Way of Transaction Data. *Knowledge Engineering and Data Science, 1 (1)*, 20–25.

Lakatos, E., & Marconi, M. d. (1992). *Metodologia do trabalho científico.* São Paulo: Editora Atlas SA.

Larose, D. T. (2014). *Discovering Knowledge in data. An introducing to Data Mining* (Second Edition ed.). New Jersey: John Wiley & Sons, Inc.

Linoff, G. S., & Berry, M. J. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Support.* New York, USA: John Wiley & Sons.

Lucas, J. P., Luz, N., Moreno, M., Anacleto, R., Figueiredo, A. A., & Martins, C. (2013). A hybrid recommendation approach for a tourism system. *Expert Systems with Applications, 40(9)*, 3532–3550.

Maimon, O. Z., & Rokach, L. (2010). *Data Mining and Knowledge Discovery hand book* (2nd ed.). London: Springer.

Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques.* Berlim: Springer.

Pimenta, C., Ribeiro, R., Sá, V., & Belfo, F. P. (2018). Fatores que Influenciam o Sucesso Escolar das Licenciaturas numa Instituição de Ensino Superior Portuguesa. *18.ª Conferência da Associação Portuguesa de Sistemas de Informação.* Santarém: Instituto Superior de Contabilidade e Administração de Coimbra.

Puccinelli, N. M., Goodstein, R. C., Grewal, D., Price, R., Raghubir, P., & Stewart, D. (March de 2009). Customer Experience Management in Retailng: Understanding the Buying Process. *Journal of Retailing, 85*(1), 15-30.

Raeder, T., & Chawla, N. V. (28 de Agosto de 2011). Market basket analysis with networks. *Social Network Analysis and Mining*, 97–113.

Romero, C. L. (2006). *El comportamiento del consumidor ante el diseño de venta virtual: efectos e interacciones.* Cuenca: Ediciones de la Universidad de Castilla-La Mancha.

Sánchez, D., Vila, M., Cerda, L., & Serrano, J. (2009). Association rules applied to credit card fraud detection. *Expert Systems with Applications, 36(2 PART 2)*, 3630–3640.

Shankar, V., Inman, J., Mantrala, M., Kelley, E., & Rizley, R. (Julho de 2011). Inovations in Shopper Marketing: Current Insights and Future Research Issues. *Jiurnal of Retailing, 87*(Supplement 1), S29-S42.

Shearer, C. (2000). The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing, volume 5*, pp. 13-18.

Sorensen, H., Bogomolova, S., Anderson, K., Trinh, G., Sharp, A., Kennedy, R., . . . Wright, M. (2017). Fundamental patterns of in-store shopper behavior. *Journal of Retailing and Consumer Services, 37(C)*, 182-194.

Stilou, S., Bamidis, P. D., & Maglaveras, N. (2001). Mining association rules from clinical databases: An intelligent diagnostic process in healthcare. *MEDINFO Proceedings od the 10th World Conferences on Medical Informatics, Part 2* (pp. 1399–1403). Amesterdam : IOS Press.

Turban, E., Sharda, R., Delen, D., & King, D. (2007). *Business Intelligence: a Managerial Approach.* New Jersey: Prentice Hall.

Vijaylaxmi, Batra, G., & Alam, D. (2012). Preserving privacy in data mining using semma methodology. *In International Journal on Computer Science and Engineering, volume 4,*, 853-858.

8. ANEXOS

8.1. DADOS ESTATÍSTICOS DA SEGMENTAÇÃO POR VALOR

Variável	Cluster	Mínimo	Máximo	Média	Desvio Padrão
Idade	1	15	89	41.7	10.1
	2	23	93	44.4	10.1
	3	27	90	45.4	9.7
Recency	1	0	87	49.9	25.3
	2	0	85	15.3	18.4
	3	0	42	3.4	7.5
Frequency	1	1	12	1.9	1.6
	2	1	23	5.9	4.9
	3	9	69	29.6	13.2
Monetary	1	3.99 €	3446.93 €	182.39 €	202.54 €
	2	103.72 €	5094.80 €	1035.48 €	823.27 €
	3	1679.18 €	18790.32 €	5876.61 €	3042.60 €
Número médio de artigos por pedido	1	0	98	35.2	17.8
	2	11	387	75.1	42.1
	3	18	330	69.1	39.8
Número total de pedidos	1	1	21	2.0	2.1
	2	1	25	6.3	5.4
	3	9	123	35.2	20.1
Número total de artigos	1	0	502	67.9	67.6
	2	41	1490	373.6	288.8
	3	624	6729	2087.6	1036.5
Valor médio por pedido	1	3.99 €	513.70 €	94.65 €	52.10 €
	2	21.67 €	1255.08 €	211.39 €	133.62 €
	3	49.39 €	1035.35 €	195.89 €	122.53 €
Número total de departamentos	1	1	16	8.8	3.3
	2	4	16	12.7	2.1
	3	11	16	14.6	1.1

Tabela 12 – Resultados estatísticos para a segmentação por valor, referentes às variáveis numéricas

Variável	Cluster	Número de ocorrências	Percentagem
Género Feminino	1	718	68.6%
	2	650	79.8%
	3	90	76.9%
Género Masculino	1	328	31.4%
	2	165	20.2%
	3	27	23.1%

Tabela 13 – Resultados estatísticos para a segmentação por valor, referentes às variáveis categóricas

Cluster	Tamanho	Percentagem
1	1046	52.70%
2	815	41.06%
3	117	5.89%
Total	1985	100.0%

Tabela 14 – Dimensões dos clusters resultantes da segmentação por valor

8.2. DADOS ESTATÍSTICOS DA SEGMENTAÇÃO POR CONSUMO

Variável	Cluster	Mínimo	Máximo	Média	Desvio Padrão
Monetary	1	3.99 €	3446.93 €	73.89 €	9.88 €
	2	48.27 €	3430.45 €	241.52 €	10.76 €
	3	54.13 €	6081.83 €	342.33 €	11.82 €
	4	176.67 €	18790.32 €	1452.70 €	12.32 €
Total de pedidos	1	1	17	1.4	1.9
	2	1	33	2.2	2.8
	3	1	85	3.3	3.4
	4	1	123	7.1	4.2
Animais por pedido	1	0.00 €	130.40 €	0.27 €	1.19 €
	2	0.00 €	76.82 €	0.48 €	1.58 €
	3	0.00 €	91.25 €	0.38 €	1.26 €
	4	0.00 €	313.78 €	0.99 €	2.35 €
Bebidas por pedido	1	0.00 €	344.62 €	2.18 €	2.59 €
	2	0.00 €	212.56 €	8.36 €	2.30 €
	3	0.00 €	95.83 €	4.24 €	1.75 €
	4	0.00 €	219.88 €	13.72 €	1.63 €
Charcutaria por pedido	1	0.00 €	61.03 €	1.53 €	1.96 €
	2	0.00 €	66.29 €	3.24 €	2.10 €
	3	0.00 €	71.42 €	5.76 €	1.48 €
	4	0.00 €	136.72 €	9.83 €	1.28 €
Congelados por pedido	1	0.00 €	45.48 €	0.94 €	1.77 €
	2	0.00 €	108.90 €	2.46 €	2.38 €
	3	0.00 €	131.18 €	6.17 €	1.83 €
	4	0.00 €	158.79 €	13.38 €	1.55 €

Dietéticos por pedido	1	0.00 €	138.28 €	0.98 €	1.96 €
	2	0.00 €	178.59 €	2.05 €	2.44 €
	3	0.00 €	235.22 €	3.32 €	2.45 €
	4	0.00 €	102.86 €	6.18 €	1.98 €
Drogaria e Limpeza por pedido	1	0.00 €	238.11 €	1.99 €	2.86 €
	2	0.00 €	468.07 €	14.96 €	2.21 €
	3	0.00 €	111.86 €	5.89 €	2.10 €
	4	0.00 €	401.40 €	24.87 €	1.26 €
Frutaria por pedido	1	0.00 €	62.71 €	2.32 €	2.30 €
	2	0.00 €	51.59 €	2.67 €	2.15 €
	3	0.00 €	224.26 €	11.40 €	1.55 €
	4	0.00 €	252.16 €	14.26 €	1.56 €
Legumes por pedido	1	0.00 €	37.77 €	0.95 €	1.40 €
	2	0.00 €	131.92 €	5.47 €	1.62 €
	3	0.00 €	67.78 €	3.93 €	1.26 €
	4	0.00 €	90.51 €	10.76 €	1.14 €
Mercearia por pedido	1	0.00 €	49.24 €	1.56 €	1.72 €
	2	0.00 €	109.13 €	7.84 €	1.61 €
	3	0.00 €	63.70 €	5.66 €	1.14 €
	4	0.00 €	150.26 €	13.66 €	1.02 €
Pastelaria por pedido	1	0.00 €	82.70 €	0.28 €	0.85 €
	2	0.00 €	20.54 €	0.33 €	0.81 €
	3	0.00 €	81.18 €	0.82 €	1.08 €
	4	0.00 €	43.69 €	1.08 €	1.22 €
Peixaria por pedido	1	0.00 €	134.70 €	0.30 €	1.31 €
	2	0.00 €	75.82 €	0.20 €	0.95 €
	3	0.00 €	305.16 €	1.04 €	1.98 €
	4	0.00 €	171.64 €	1.45 €	2.15 €
Pequeno Almoço por pedido	1	0.00 €	63.02 €	1.54 €	1.80 €
	2	0.00 €	104.78 €	6.99 €	1.78 €
	3	0.00 €	70.54 €	5.93 €	1.35 €
	4	0.00 €	128.82 €	13.55 €	1.14 €
Perfumaria por pedido	1	0.00 €	109.48 €	0.65 €	1.54 €
	2	0.00 €	231.83 €	4.63 €	2.33 €
	3	0.00 €	64.68 €	2.06 €	1.66 €
	4	0.00 €	94.62 €	8.71 €	1.60 €
Refrigerados e Lácteos por pedido	1	0.00 €	40.40 €	1.62 €	1.77 €
	2	0.00 €	61.86 €	7.26 €	1.44 €
	3	0.00 €	53.62 €	8.77 €	0.92 €
	4	0.72 €	93.24 €	16.02 €	0.86 €
Talho por pedido	1	0.00 €	91.82 €	1.17 €	2.33 €
	2	0.00 €	40.93 €	0.36 €	0.99 €
	3	0.00 €	141.60 €	9.27 €	1.86 €
	4	0.00 €	224.88 €	7.77 €	2.51 €

Vinhos por pedido	1	0.00 €	476.25 €	0.88 €	3.08 €
	2	0.00 €	237.36 €	1.25 €	2.87 €
	3	0.00 €	93.26 €	0.76 €	1.75 €
	4	0.00 €	298.46 €	4.20 €	3.55 €

Tabela 15 – Resultados estatísticos para a segmentação por consumo

Cluster	Tamanho	Percentagem
1	416	21.0%
2	478	24.1%
3	581	29.3%
4	510	25.7%
Total	1985	100.0%

Tabela 16 – Dimensões dos clusters resultantes da segmentação por valor