INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Segmentação de Clientes de Farmácias Comunitárias com aplicação em KNIME *Analytics Platform*

Mariana Marques Carapinha

Mestrado em Métodos Analíticos para a Gestão (Business Analytics)

Orientadora:
Doutora Patrícia Andreia da Silva Filipe, Professora Associada,
ISCTE – Instituto Universitário de Lisboa

Outubro, 2024

BUSINESS
SCHOOL

Departamento de Métodos Quantitativos para Gestão e Economia

Segmentação de Clientes de Farmácias Comunitárias com aplicação em KNIME *Analytics Platform*

Mariana Marques Carapinha

Mestrado em Métodos Analíticos para a Gestão (Business Analytics)

Orientadora:
Doutora Patrícia Andreia da Silva Filipe, Professora Associada,
ISCTE – Instituto Universitário de Lisboa

Outubro, 2024

Agradecimento

Gostaria de expressar a minha profunda gratidão à empresa, cuja identidade permanece anonimizada por razões de confidencialidade, pela disponibilização dos dados essenciais e pelo valioso apoio, que permitiram o início desta investigação da forma mais eficiente e adequada.

À Professora Patrícia Filipe, deixo um especial agradecimento pela dedicação, pelo tempo disponibilizado, e pela constante motivação e incentivo que foram determinantes para a conclusão deste trabalho com sucesso.

Aos meus pais, Dina e José, sou imensamente grata pelo carinho, apoio incondicional e paciência ao longo das fases mais desafiantes deste percurso. As vossas palavras de sabedoria e conselhos foram fundamentais, e a concretização desta etapa só foi possível graças a vocês.

Ao meu namorado, Diogo, agradeço por todo o amor, compreensão, e apoio na concretização dos meus objetivos, e pela paciência incansável mesmo nos dias menos agradáveis.

Aos meus amigos, o meu sincero agradecimento pelo carinho, por estarem ao meu lado nos momentos de descontração, essenciais para manter o equilíbrio entre trabalho e vida pessoal, e pelo apoio constante.

Por fim, agradeço a todos os docentes que acompanharam o meu percurso ao longo deste mestrado, pelos conhecimentos partilhados e pela orientação prestada durante todo o processo de investigação.

Obrigada!

Mariana Carapinha

Resumo

A segmentação de clientes em farmácias comunitárias é um tema ainda pouco explorado, com a maioria dos estudos focados em *E-Pharmacy*, porém bastante critico, tendo em conta a competitividade deste setor. Este estudo busca fornecer uma perspetiva de como a segmentação de clientes pode aprimorar a personalização das estratégias de *marketing* em farmácias comunitárias. Para tal, são desenvolvidos dois modelos de segmentação, baseados numa amostra de dados de uma farmácia comunitária, com um total de 6 805 clientes e abrangendo o período de 2022 a 2024. Foi criado um modelo de análise RFM, com recurso ao algoritmo de K-means, e para a segmentação demográfico-produto é aplicado o algoritmo de agrupamento hierárquico. Como resultado, foram identificados seis segmentos de clientes em cada uma das análises, permitindo caracterizar cada segmento e definir estratégias específicas. Além disso, foi realizada uma análise conjunta das segmentações demográfico-produto e RFM, obtendo-se cruzamentos que possibilitam desenvolver estratégias de *marketing* ainda mais direcionadas. O contributo com maior relevância desta investigação é fornecer conhecimento sobre a segmentação de clientes direcionado para farmácias comunitárias e, assim, permitir que estas definam com maior precisão estratégias de *marketing* mais eficazes.

Palavras-chave: Segmentação de clientes, Análise RFM, K-means, Agrupamento Hierárquico, Farmácia Comunitária

***JEL Classification System*:** C38, L81

Abstract

Customer segmentation in community pharmacies is an underexplored topic, with most studies focusing on E-Pharmacy, however, it is critical due to the competitiveness of this sector. This study provides a perspective on how customer segmentation can improve the personalization of marketing strategies in community pharmacies. Therefore, two types of segmentation models are developed based on a sample of data from a community pharmacy, with a total of 6,805 customers and covering the period from 2022 to 2024. An RFM analysis model was created using the K-means algorithm, and the hierarchical clustering algorithm was applied for demographic-product segmentation. As a result, six customer segments were identified in each analysis, making it possible to characterize each segment and define specific strategies. In addition, a joint analysis of the demographic-product and RFM segmentation was carried out, obtaining cross-checks that allow even more targeted marketing strategies to be developed. The main contribution of this research is to provide knowledge on customer segmentation for community pharmacies, enabling them to define customer profiles more accurately and thereby establishing more effective marketing strategies.

Keywords: Customer Segmentation, RFM Analysis, K-means, Hierarchical Clustering, Community Pharmacy

***JEL Classification System*:** C38, L81

Índice Geral

Agradecimento i

Resumo iii

Abstract v

1. Introdução 1

1.1 Contextualização 1

1.2 Motivação e contribuição 2

1.3 Questão de investigação e objetivos 3

1.4 Estrutura da dissertação 4

2. Revisão da literatura 5

2.1 Segmentação de clientes 5

2.2 Segmentação de clientes de farmácias comunitárias 11

2.2.1 Farmácias Comunitárias e Produtos OTC 11

2.2.2. Importância da segmentação de clientes de farmácia 11

2.2.3 Aplicações de segmentação de clientes de farmácias comunitárias 12

3. Metodologia 13

3.1 Compreensão do Negócio 14

3.2 Compreensão dos dados 15

3.3 Preparação dos dados 20

3.4 Modelação 26

3.5 Avaliação 29

4. Resultados e Discussão 31

4.1 Caraterização dos clientes da Farmácia S 31

4.2 Avaliação dos modelos e caraterização dos clusters 33

4.2.1 Análise Segmentação Demográfica-Produto 33

4.2.2 Análise RFM 35

5. Conclusões e Recomendações 41

5.1. Sumário da investigação 41

5.2. Limitações 42

5.3 Contribuições 42
5.4 Recomendações e Pistas Futuras 43
Referências Bibliográficas 45
Anexos 51
Anexo A - Workflow preparação de dados em KNIME Analytics Platform 51
Anexo B - Elbow Method em KNIME 51
Anexo C – Microsoft Power BI & KNIME Analytics Platform 52
Anexo D – Aplicação dos algoritmos e resultados segmentação “Demográfico-produto” 53
Anexo E – Resultados análise RFM 54

Índice de Figuras

Figura 1: Fases da metodologia CRISP-DM ..... **Erro! Marcador não definido.**
Figura 2: Distribuição das farmácias pelo território português 16
Figura 3: Distribuição dos clientes pelo território português 16
Figura 4: Distribuição do género dos clientes 18
Figura 5: Top 5 segmentos mais vendidos 19
Figura 6: Código criação Faixas Etárias 21
Figura 7: Workflow com o processo de criação das faixas etárias 21
Figura 8: Preparação da variável Valor de Venda 22
Figura 9: Processo de filtrar os dados pela farmácia com maior volume de vendas 22
Figura 10: Processo de Ranking da segmentação relacionada com o produto 23
Figura 11: Transformação das variáveis categóricas em numéricas 24
Figura 12: Preparação de dados para a análise RFM 24
Figura 13: Algoritmos utilizados na fase da Modelação 26
Figura 14: Fluxograma do processo do algoritmo K-means 27
Figura 15: Gráfico Elbow Method: variação da soma dos erros quadrados em função do número de clusters 29
Figura 16: Nó coeficiente de silhueta 30
Figura 17: Distribuição dos clientes por género e faixa etária 31
Figura 18: Distribuição da faixa etária por segmento favorito (Top 5) 32
Figura 19: Distribuição do género por responsável comercial favorito (Top 5) 32
Figura 20: Características dos segmentos criados através do Hierarchical Clustering 34
Figura 21: Distribuição dos clientes por faixa etária e cluster 34
Figura 22: Distribuição dos clusters por género 34
Figura 23: Gráficos de dispersão dos clusters finais pelas variáveis RFM 38
Figura 24: Cross table entre os clusters RFM e "Demográfica-Produto" 40

Índice de Tabelas

Tabela 1: Dicionário de dados, tabela "Histórico de Compras" 17
Tabela 2: Algoritmos da segmentação Demográfica-Produto e respetivos parâmetros utilizados 28
Tabela 3: Resultados da segmentações Demográfica-Produto..... 30
Tabela 4: Resultados dos clusters da segmentação Demográfica-Produto (Hiearchical Clustering) .. 30
Tabela 5: Médias dos clusters de cada variável RFM 36
Tabela 6: Resultados dos clusters da variável RFM Score (K-means) 37
Tabela 7: Caraterísticas do clusters a nível das variáveis RFM (Média e Valor mais Comum) 37

1. Introdução

1.1 Contextualização

As farmácias comunitárias têm um papel relevante na disponibilização de cuidados de saúde essenciais à sociedade, na gestão da utilização de medicamentos e na oferta de serviços farmacêuticos relacionados. No entanto, o setor farmacêutico está cada vez mais competitivo devido a fatores como a globalização, a expansão das farmácias *online* e as mudanças nas preferências dos consumidores. Consequentemente, as farmácias comunitárias precisam de desenvolver táticas de marketing baseadas em dados atuais sobre o comportamento dos consumidores, para atrair novos clientes e construir uma base de clientes fiéis (Kevrekidis *et al*., 2018).

A segmentação é considerada uma das estratégias de marketing com maior eficácia para conseguir atender os clientes com necessidades específicas, permitindo criar e sustentar uma vantagem competitiva sobre a concorrência (Wijayanti *et al*., 2024). Este método permite desenvolver campanhas de marketing específicas, através da obtenção de *insights* valiosos sobre os clientes, permitindo melhorar a satisfação dos mesmos e, assim, criar programas de fidelização mais eficientes.

Esta investigação procura explorar como a segmentação de clientes pode ser aplicada de forma eficaz nas farmácias comunitárias, contribuindo para um atendimento mais eficiente e direcionado, preenchendo lacunas existentes na literatura e nas práticas atuais. Para o desenvolvimento da investigação são utilizadas informações sobre clientes de 9 farmácias nacionais, como dados demográficos, dados relacionados com o produto e dados sobre o comportamento de compra, relativos ao período de 2022 a 2024.

O Regime jurídico dos medicamentos de uso humano refere que está proibido por lei a publicidade de medicamentos ou tratamentos médicos em que a sua obtenção está sujeita a receita médica (Decreto-Lei nº 176/2006, 2006), como resultado, a investigação incide apenas sobre dados históricos de compras de OTC (*Over-The-Counter*) (medicamentos que podem ser adquiridos sem necessidade de prescrição médica), uma vez que as estratégias de marketing poderão ser relacionadas para a promoção de produtos.

A venda de medicamentos com receita médica é a principal razão pela qual os clientes recorrem às farmácias e é considerada uma das maiores fontes de receita das mesmas. Contudo, o suporte às vendas de OTC é essencial para prosperidade financeira das farmácias, uma vez que o consumo destes está a aumentar continuamente e as suas margens

de lucro são bastante benéficas. Focar na segmentação de OTC's é essencial para uma gestão eficiente direcionada para o cliente das farmácias (Kevrekidis *et al*., 2018).

O objetivo é fornecer um modelo prático e acessível que possa ser aplicado por farmácias comunitárias, ajudando a melhorar a sustentabilidade e o impacto positivo na saúde das comunidades que servem.

1.2 Motivação e contribuição

Como referido anteriormente, a segmentação de clientes é fundamental para compreender melhor os comportamentos e necessidades dos consumidores, especialmente em setores competitivos como é o caso do setor farmacêutico, em que existe uma extensa variedade de produtos e perfis de clientes, tornando a identificação de segmentos crucial (Kevrekidis *et al*., 2018).

O desenvolvimento de uma segmentação eficiente ajudará na otimização de recursos, uma vez que permite a alocação direcionada dos recursos de marketing e operacionais, potencializando o seu crescimento e a rentabilidade. Outra motivação, inclui a melhoria do atendimento personalizado, contribuindo para a inovação na gestão do marketing das farmácias, permitindo-lhes entender melhor os seus clientes, melhorando a qualidade do seu atendimento de acordo com as especificidades de cada um e consequentemente aumentar a satisfação dos mesmos (Cooil *et al*., 2008; Kevrekidis *et al*., 2018). Além disso, será possível a adaptação a mudanças no comportamento dos clientes, tendo em conta que estes são recorrentes, e impactantes em situações de crise de saúde, como, por exemplo, na pandemia COVID-19, em que é necessário conseguir uma adaptação rápida das estratégias de marketing.

Existe um uma extensa base de dados com investigações sobre segmentação de clientes, porém, existe um *gap* na segmentação direcionada para farmácias. Possivelmente, por ser considerada uma área em que os dados são bastante sensíveis, tornando a investigação importante para a comunidade farmacêutica, e ainda aplicável a outros tipos de negócios, uma vez que os dados recolhidos serão genéricos a outros setores.

O desenvolvimento da parte prática da investigação é efetuado através da plataforma KNIME *Analytics* versão 5.2.3 (Berthold *et al*., 2009), que poderá ser útil para a Literatura Académica, uma vez que a comunidade KNIME promove a partilha de conteúdos, conhecimento e a colaboração entre investigadores. Para uma análise visual detalhada das características dos clientes, utiliza-se também o Microsoft Power BI. A integração direta com o KNIME *Analytics Platform* permite a interligação eficiente das duas ferramentas,

possibilitando uma exploração mais profunda dos dados e uma visualização gráfica aprimorada. Essa combinação oferece uma melhor solução para a análise de dados, aproveitando as vantagens de ambas as ferramentas.

1.3 Questão de investigação e objetivos

Esta dissertação tem como objetivo desenvolver um modelo de segmentação de clientes para farmácias utilizando algoritmos de *clustering*, incluindo a técnica de RFM (*Recency*, *Frequency*, *Monetary*), utilizando dados de fatores demográficos e comportamentais dos consumidores, e dados relacionados com o produto.

A questão de investigação a que a dissertação pretende responder é a seguinte: Como a segmentação de clientes pode ajudar a melhorar a personalização de estratégias de marketing das farmácias comunitárias?

Para que esta pergunta seja corretamente respondida é essencial identificar segmentos e caraterizar os perfis de clientes, que será efetuado através de técnicas de *clustering*, e análises descritivas. Para isso, será necessário alcançar determinados objetivos:

O.1. Caraterização dos clientes das farmácias

- Este objetivo encontra-se cumprido aquando da elaboração da análise descritiva dos dados após terminada a fase da preparação dos dados, por forma a verificar a qualidade antes de iniciar a modelação.

O.2. Identificação dos segmentos (segmentação)

- Este objetivo é alcançado aquando da aplicação e validação de algoritmos de *clustering*, evidenciada pelo coeficiente de silhueta, que deve apresentar um valor idealmente próximo de 0,7 ou superior. Além disso, os segmentos identificados devem demonstrar relevância e utilidade para a tomada de decisões estratégicas no negócio.

O.3 Caraterização dos segmentos criados

- Este objetivo será alcançado quando os segmentos identificados forem caracterizados por meio de análises descritivas e perfis claros forem definidos para cada grupo.

1.4 Estrutura da dissertação

A estrutura da dissertação é baseada na metodologia adotada CRISP-DM (*Cross-Industry Standard Process for Data Mining*), sendo esta a metodologia mais adequada para investigações que têm por base segmentação de clientes, uma vez que a sua abordagem é iterativa, estruturada e facilita a compreensão dos dados e construção dos modelos de *clustering*.

Após a introdução, no Capítulo 2º é abordada a revisão da literatura, onde são descritos os principais estudos e teorias relacionadas ao tema, assim como as metodologias e abordagens frequentemente utilizadas em investigações anteriores na área de estudo.

No 3º Capítulo é detalhada toda a metodologia utilizada, descrevendo as fases que esta compreende, exceto a fase da implementação, acompanhando a seguinte ordem: compreensão do negócio, compreensão de dados, preparação de dados, modelação e avaliação. É importante realçar que ao longo do processo a alternação entre a fase de preparação de dados e a fase de modelação é relevante para assegurar a melhor qualidade dos dados para posteriormente aplicar os modelos.

No Capítulo 4 são apresentados os resultados da investigação, e posteriormente são avaliados e discutidos, por forma a validar os objetivos inicialmente propostos para responder à questão de investigação.

Por fim, no Capítulo 5 são descritas as conclusões, limitações identificadas e recomendações, visando a orientar futuras pesquisas, com o objetivo de maximizar benefícios provenientes desta investigação.

2. Revisão da literatura

A revisão da literatura é essencial para reunir os resultados de pesquisas efetuadas, apresentá-los numa perspetiva mais abrangente e identificar áreas que requerem mais investigação, o que se torna relevante para aumentar as bases de dados de referências das temáticas e construir modelos conceptuais (Snyder, 2019).

A base de dados científica principal utilizada para encontrar artigos relacionados com a temática foi a *Scopus*, visto que esta plataforma permite avaliar os resultados de pesquisa, tem uma cobertura ampla e inclusiva, e ainda é considerada uma das melhores para cobertura de temas específicos como farmácias (Burghardt *et al*., 2020; Pranckutė, 2021).

2.1 Segmentação de clientes

A introdução do tema segmentação de clientes é apresentada por W. R. Smith (1956), em que refere que é possível ganhar competitividade ao reconhecer segmentos específicos de mercado.

A segmentação é fundamental para o DSS (*Decision Support System*) das empresas. Consiste numa técnica de análise de dados com aplicação a Marketing essencial que ajuda na definição de perfis de clientes. Embora as necessidades e expectativas dos consumidores sejam únicas, muitos partilham características bastante similares (Hicham *et al*. 2022).

A segmentação tem várias vantagens, como a maximização do valor do cliente e a utilização eficiente dos recursos, contribuindo diretamente para a eficiência dos métodos de gestão e relacionamento com o cliente. O CRM (*Customer Relationship Management*) tem um papel crítico na gestão do relacionamento com o cliente, sendo a segmentação uma das funções mais importantes. Através da segmentação, as empresas podem identificar e agrupar clientes com perfis semelhantes, o que facilita a criação de campanhas de marketing mais direcionadas para esses. Essa personalização permite maximizar o valor de cada cliente e adaptar as ofertas e comunicações às suas necessidades específicas. Além disso, como os recursos das empresas são limitados, a segmentação contribui para uma utilização mais eficiente desses recursos, pois ajuda a direcionar os esforços para segmentos que oferecem maior retorno. Ao compreender e atender melhor os diferentes perfis de clientes é possível alocar os recursos de forma mais eficiente (Cooil *et al*., 2008; Chen *et al*., 2006; Christy *et al*., 2021; Tabianan *et al*., 2022).

2.1.1 Categorias de segmentação de clientes

Existem várias categorias que podem ser utilizadas para efetuar segmentação de clientes, as cinco principais são (Camilleri, 2018):

Segmentação demográfica – consiste no agrupamento dos clientes com base em características observáveis e baseadas em factos, como, por exemplo, idade, género, nacionalidade. Este tipo de segmentação é bastante utilizado devido à menor dificuldade na obtenção de dados.

Segmentação geográfica – divide os clientes com base na localização, considerando variáveis como o clima e densidade populacional, permitindo identificar diferenças entre os clientes de várias regiões. Esta categoria poderá ser crucial na definição de campanhas de marketing direcionadas para as necessidades e preferências regionais.

Segmentação psicográfica – segmenta o mercado baseado na personalidade, estilos de vida, valores dos clientes. Esta abordagem permite obter uma visão mais profunda dos clientes, ajudando as empresas a personalizar as mensagens de marketing para os diferentes segmentos.

Segmentação comportamental – reparte os clientes com fundamento nos seus comportamentos de compra, como os seus hábitos, frequência e volume de compras. Analisando padrões de compra, as empresas conseguem facilmente identificar os clientes mais valiosos e criar programas de fidelização e promoções direcionadas para estes. Através desta segmentação é possível aplicar a técnica RFM (*Recency, Frequency, Monetary*) que é baseada no histórico de consumo do cliente, permitindo categorizar os clientes com base em vários fatores comportamentais.

Segmentação relacionada com o produto – organiza os clientes com base em características específicas do produto ou dos serviços, identificando fatores como benefícios e lealdade à marca. Esta análise permite alinhar ofertas de produtos com as preferências dos clientes, e consequentemente melhorar a satisfação destes a longo prazo.

2.1.2 Métodos de Segmentação

O método RFM é considerado eficaz para a segmentação de mercado, que analisa informação sobre o comportamento de consumo dos clientes.(Gustriansyah *et al*., 2019). Os segmentos provenientes desta técnica são ordenados de ordem objetiva, do mais valioso para o menos valioso, e assim utilizado para pontuar os clientes (Hopf *et al*., 2023).

Esta ferramenta é robusta e importante na área de marketing, é bastante utilizada para classificar os consumidores de acordo com o seu histórico de compras em todo o tipo de setores, como, por exemplo, setor de compras *online*, o setor retalho, entre outros. Este método tem por base três dimensões para agrupar os clientes: *Recency, Frequency e Monetary,* (Christy *et al*., 2021).

A variável *Recency* consiste na diferença de dias entre o dia da última compra do cliente e o fim da recolha de dados. Neste caso quanto maior o seu valor, menor a probabilidade de o cliente visitar em breve a empresa. *Frequency* é o número de vezes que o cliente efetua compras durante um determinado período, indicando que quanto maior este valor maior a propensão a realizar compras futuras. A variável *Monetary* indica o valor total ou médio gasto pelo cliente durante um determinado período, o que poderá significar que quanto maior este valor mais fiel é o cliente à empresa (Wu *et al*., 2020).

Não obstante, desta técnica apresentar diversas vantagens, como mencionado anteriormente, esta possui algumas desvantagens, como, por exemplo, o reduzido número de variáveis, uma vez que são apenas consideradas três variáveis, não captando toda a complexidade do comportamento do cliente. Além disso, a falta de contexto sobre o cliente em outras dimensões, também é considerada uma limitação, uma vez que ao basear-se apenas em fatores comportamentais os restantes fatores que podem igualmente fornecer *insights* importantes são excluídos, como, por exemplo, incluir dados demográficos (Ernawati *et al*., 2021).

A análise de *clustering* envolve o agrupamento de uma série de padrões com base nas suas semelhanças, onde os elementos dentro de um grupo são mais parecidos entre si do que com os de outros grupos (Jain *et al*., 1999). Esta técnica é considerada bastante útil em diversas áreas e facilita a compreensão de grandes conjuntos de dados, sendo crucial a sua utilização à medida que os conjuntos de dados continuam a expandir em tamanho e complexidade, é uma das principais técnicas utilizadas em Data Mining (Chaudhry *et al*., 2023).

Além de ser uma ferramenta essencial em *Data Mining*, a técnica de *clustering* pode ser utilizada em diferentes algoritmos para agrupar os dados, como, por exemplo, o algoritmo *K-means, Hierarchical Clustering*, DBSCAN (*Density Based Spatial Clustering of Applications with Noise*), que são alguns dos métodos mais aplicados de *clustering*, que posteriormente são apresentados detalhadamente pela sua relevância na análise de agrupamentos, destacando-se o funcionamento e as principais características de cada um. Porém, existem outros métodos que também são aplicados na literatura como, por exemplo, o *Fuzzy C – Means* (Christy *et al*., 2021) e o *Two-Step Clustering* (Kevrekidis *et al*., 2018).

O algoritmo *K-means* é um dos métodos mais conhecidos no contexto de *clustering*, este foi introduzido por Macqueen em 1967 (Wu *et al*., 2020). A simplicidade de implementação e a rápida abordagem são vantagens do algoritmo, porém o seu comportamento complexo torna-se um desafio para o uso e interpretação precisa dos resultados. Portanto, a compreensão deste é vital para o uso eficaz (Davidson, 2002).

Através da identificação de uma divisão de n observações em k *clusters* o algoritmo *K-means* consegue que cada observação esteja associada ao *cluster* mais semelhante à média. Este por sua vez, necessita de um parâmetro inicial k, que define o número de partições a serem criadas, e atribui os objetos aos grupos visando a minimização do erro quadrático (Sinaga *et al*. 2020). Este algoritmo não impõe uma estrutura hierárquica e calcula todos os *clusters* viáveis ao mesmo tempo (Jain, 2010). Contudo, este algoritmo apesar de ser o mais aplicado, apresenta algumas limitações como, por exemplo, o facto dos *outliers* poderem influenciar negativamente a qualidade da análise, que pode ser ultrapassado através de um tratamento de dados adequado (Hayasaka, 2024). Outra limitação, está relacionada com o facto de o algoritmo não conseguir estimar um número de *clusters* presente nos dados, sendo necessário definir um número de *clusters* previamente, o que pode originar um enviesamento dos resultados. Este problema pode ser solucionado através da aplicação do *"Elbow Method*" que ajuda a determinar o número de *clusters* apropriado para aplicar ao algoritmo (Umargono *et al*., 2020).

O algoritmo *hierarchical clustering* estrutura elementos de entrada numa árvore binária que agrupa elementos que partilham caraterísticas similares. Na árvore binária resultante, a distância entre os elementos de entrada reflete a sua semelhança (Bar-Joseph *et al*., 2001). O objetivo deste algoritmo é construir uma árvore hierárquica que reúna todos os elementos numa única árvore. Para um conjunto de n amostras calcula-se uma matriz triangular superior de similaridade que contém pontuações para todos os pares de amostras, sendo que o par de amostras mais parecido é unido, criando-se um nó com um perfil de expressão médio. Esse processo é repetido até restar apenas um elemento (Eisen *et al*., 1998).

Existem três tipos de ligação de *hierarchical clustering, single linkage* que consiste em medir a distância entre os pontos mais próximos de dois *clusters*, originando *clusters* distintos, a *Complete Linkage* calcula a distância entre os pontos mais distantes de dois *clusters*, resultando em *clusters* consistentes, a *Average linkage* calcula a distância entre todos os pares de pontos de dois *clusters*, sendo uma combinação entre os métodos *single* e *complete* (Hayasaka, 2024).

Este algoritmo tem algumas limitações, sendo este possivelmente demorado para determinados conjuntos de dados, devido à complexidade acrescida com o aumento do

número de observações. Além disso, demonstra sensibilidade a ruído e *outliers*, potencialmente prejudicando a formação de *clusters*, contudo esta limitação é ultrapassada ao executar uma limpeza assertiva dos dados (Papakyriakou *et al*. 2022; Tomar, 2013).

O algoritmo DBSCAN produz *clusters* de forma aleatória, estes são formados através da observação da densidade dos pontos. Áreas de alta densidade de pontos revelam a presença de *clusters*, enquanto áreas com reduzida densidade de pontos revela a presença de possível ruído ou *outliers*.(Al-Mhairat *et al*., 2019). Este algoritmo não necessita que seja pré-definido o número de clusters e consegue lidar com clusters, seja qual for o seu tamanho ou forma (Jeena *et al*., 2023), contudo também apresenta desvantagens relativamente à sensibilidade dos parâmetros de entrada, a definição pode ser complexa e depende do conjunto de dados.

2.1.2.1 Métricas de avaliação de Clustering

Avaliar um algoritmo de *clustering* é complexo, uma vez que depende da qualidade de separação dos dados que este efetua de acordo com uma referência, ao invés de considerar os valores absolutos dos rótulos dos *clusters*. A métrica de avaliação deve refletir a similaridade dentro dos *clusters* e a desigualdade entre diferentes *clusters* (Al-Mhairat *et al*., 2019).

Existem várias métricas para efetuar a avaliação dos resultados de *clustering*, como, por exemplo, o e coeficiente de silhueta (Christy *et al*., 2021) e o *Adjusted Rand Index* (Coombes *et al*., 2021), sendo estas as técnicas mais populares, existem outras que também são utilizadas na literatura como, por exemplo, o *Fowlkes-Mallows scores* (Al-Mhairat *et al*., 2019).

O coeficiente de silhueta é utilizado para medir a similaridade entre um objeto e o *cluster* ao qual pertence, em relação a outros *clusters*, numa escala que varia de -1 a +1 (Ma *et al*., 2023), este indica que a correspondência é agrupada com qualidade quando este se aproxima de 1 ou o contrário quando se aproxima de -1, neste caso se o objeto apresentar um coeficiente de silhueta perto de 0 significa que não está claramente diferenciado entre os *clusters* (Sinaga *et al*. 2020).

O coeficiente de silhueta tem a vantagem de depender exclusivamente da distribuição real dos objetos em *clusters*, e não do algoritmo de *cluster* utilizado para obtê-la. Outro benefício recai sobre o facto deste coeficiente ser uma ferramenta eficaz para aprimorar os resultados da análise de *clusters* (Rousseeuw, 1987).

O ARI (*Adjusted Rand Index*) é bastante utilizado na validação de agrupamento, visto que é considerada uma medida de concordância entre duas partições: uma que é definida por critérios externos e outras que é criada pelo algoritmo de *clustering* (Al-Mhairat *et al*., 2019).

O ARI ajusta a avaliação levando em consideração a concordância que poderia ocorrer aleatoriamente entre duas partições comparadas, aplicando uma pontuação variando de -1 a +1. Neste caso, um valor de 1 indica uma perfeita concordância entre duas partições (Coombes *et al*., 2021). Esta métrica é amplamente utilizada e essencial na avaliação de agrupamento de dados, porém apresenta desvantagens como, por exemplo, a sua aplicação necessita de um conhecimento prévio das classes dos dados, o que geralmente não está disponível na prática (Buitinck *et al*., 2013).

2.1.4 Aplicações de Segmentação de clientes

Patankar *et al*. (2021) desenvolveram uma metodologia para a segmentação de clientes utilizando técnicas de *machine learning*, para efetuar a segmentação utilizaram o algoritmo *K-means Clustering,* baseando-se em dados relacionados com o comportamento dos clientes. As conclusões obtidas desta investigação indicam que a melhor solução para o problema existente na segmentação de clientes é dividir os clientes com base nas características comportamentais, e que a escolha do algoritmo *K-means* foi considerada acertada para a abordagem.

Christy *et al*. (2021) efetuaram uma investigação de segmentação de cliente baseada na renda anual e nível de gastos dos clientes para criar os *clusters* com objetivo de desenvolver estratégias de marketing. Foram utilizados alguns algoritmos de *machine learning,* como o *K-means e Fuzzy C-Means* para construção dos modelos. Além disto, foram utilizadas técnicas como análise RFM para classificar os clientes em diferentes segmentos com base nos dados comportamentais destes, podendo assim personalizar as estratégias de marketing. Concluiu-se que o *K-means* é o algoritmo que consome menos tempo e reduz o número de iterações.

Afzal *et al*. (2024) realizaram um estudo que explora aplicação do algoritmo *Hierarchical clustering* na segmentação de clientes de *shoppings*, com objetivo de detetar *insights* sobre o comportamento e preferências dos clientes. Para permitir uma melhor compreensão dos resultados foi aplicada a métrica coeficiente de silhueta e foram elaboradas diferentes visualizações como histogramas e *scatterplots* para análise dos resultados. Foi concluído através desta pesquisa que a utilização do algoritmo hierárquico, abrangendo análises univariadas, bivariadas e multivariadas, demonstrou ser eficaz na captura de estruturas hierárquicas e na identificação de relações entre segmentos de clientes diferentes.

2.2 Segmentação de clientes de farmácias comunitárias

2.2.1 Farmácias Comunitárias e Produtos OTC

Farmácias comunitárias são uns dos principais prestadores de cuidados de saúde primários à comunidade, proporcionando suporte na devida utilização de produtos medicinais e uma nos serviços farmacêuticos suportados. Estes estabelecimentos ajudam na promoção da saúde pública, educando os pacientes sobre prevenção de doenças, hábitos saudáveis e como referido anteriormente na utilização responsável de medicamentos (Kevrekidis *et al*., 2018).

Os produtos OTC (*Over the Counter*) são medicamentos que podem ser adquiridos sem necessidade de prescrição médica, regularmente utilizados para tratar condições de saúde ligeiras, como dor, alergias, entre outros. A fácil acessibilidade e necessidade deste tipo de medicamentos torna-os uma escolha popular para os consumidores aquando do alívio rápido e eficaz de sintomas comuns (FasterCapital, 2024).

2.2.2. Importância da segmentação de clientes de farmácia

A transição de farmácias comunitárias para farmácias online tem se intensificando nos últimos anos, impulsionada pela acessibilidade e baixo custo das compras online (Gray, 2011). Por este motivo é fundamental que as farmácias comunitárias adotem estratégias de marketing para permanecerem competitivas e responderem às necessidades dos consumidores.

O desempenho dos farmacêuticos de farmácias comunitárias aumentou ao longo dos anos na maior parte dos países, consequentemente leva a uma maior satisfação e reconhecimento do papel deste na área da saúde. Contudo, estes necessitam de ser capazes de avaliar as necessidades dos clientes e ter uma função fundamental e ativa no cuidado de saúde dos seus clientes (Al-Arifi, 2012).

O setor farmacêutico é considerado bastante competitivo e compete às farmácias desenvolverem estratégias de marketing eficientes como, por exemplo, o atendimento personalizado, para alcançar a fidelidade dos clientes, e ainda conseguir atrair novos clientes, através de campanhas de marketing direcionadas (Castaldo *et al*., 2016).

Perante a crescente competitividade no setor farmacêutico, a segmentação de clientes é essencial para as farmácias comunitárias. Identificar perfis de clientes e as suas necessidades, permite desenhar ações de marketing mais direcionadas, melhorando a satisfação e fidelidade. Além da habitual retenção de clientes, as farmácias podem atrair

novos com campanhas direcionadas. Este processo garante a sua competitividade perante as farmácias online e a sua relevância no mercado.

2.2.3 Aplicações de segmentação de clientes de farmácias comunitárias

A literatura na área de segmentação de clientes de farmácias comunitárias ainda é pouco explorada, possivelmente devido ao nicho específico do tema, sendo que existem muitas pesquisas focadas no *E-Pharmacy* (farmácia eletrónica) como exemplificado pelo estudo de Patak *et al*. (2014). Além disso, a maioria dos estudos em farmácias comunitárias concentra-se na saúde e não em estratégias de marketing. A sensibilidade dos dados de saúde e as questões de privacidade também limitam a disponibilidade de dados para investigações (Fantonelli *et al*., 2023). Contudo, foi encontrado um estudo que se foca na segmentação de clientes de farmácias comunitárias e nos produtos OTC.

Kevrekidis *et al*. (2018) investigaram as preferências dos consumidores na escolha de farmácias e de medicamentos de venda livre. Para desenvolver a investigação utilizaram um questionário estruturado para a recolha dos dados e aplicaram o algoritmo *Two-Step Clustering* para identificar os segmentos de clientes.

Como conclusões identificaram três segmentos distintos de clientes com base nas suas preferências e características, e ainda, que o mercado de farmácias comunitárias é dividido em segmentos de clientes com preferências variadas quanto à escolha de farmácias e produtos OTC, bem como na avaliação de serviços e produtos farmacêuticos, e nas suas características demográficas.

3. Metodologia

A metodologia a aplicar na dissertação é a CRISP-DM (*Cross-Industry Standard Process for Data Mining*), que é constituída por seis fases (Chapman *et al*., 2000), representado através da figura 1, compreensão de negócio, compreensão dos dados, preparação de dados, modelação, avaliação e implementação, no entanto, apenas será realizada até à fase de avaliação dos modelos (foco nos resultados e conclusões), sem incluir a fase de implementação. A implementação não será realizada nesta investigação, uma vez que o foco reside na avaliação dos resultados obtidos e na seleção das melhores abordagens, permitindo identificar as soluções mais eficazes sem avançar para a sua aplicação prática.

Esta metodologia é frequentemente adotada em data mining para a compreensão do comportamento dos consumidores foi utilizada previamente em vários casos práticos de segmentação de clientes, como é exemplificado, entre muitos outros, por: Silva (2022) e Alexandre *et al*. (2017).

Esta investigação insere-se na área de Data Science, com foco em Data Mining, uma subdisciplina dedicada à extração de padrões e *insights* provenientes de grandes volumes de dados. O uso de modelos matemáticos e analíticos caracteriza o processo de Data Mining, que é amplamente utilizado para a segmentação de clientes, previsão de comportamentos e outras aplicações empresariais. Embora seguir estritamente uma metodologia de projeto possa ser um desafio para as equipas dessa área, a metodologia CRISP-DM pode ser útil e eficaz (Saltz, 2021).

Figura 1: Fases da metodologia CRISP-DM

Fonte: Adaptado de(Nascimento et al., 2018)

3.1 Compreensão do Negócio

A primeira fase da metodologia CRISP-DM consiste na compreensão do negócio, uma vez que é essencial obter uma visão clara, detalhada e abrangente do mesmo. Esta fase compreende quatro etapas, determinar os objetivos do negócio, avaliar a situação, definir os objetivos específicos e criar um plano de projeto detalhado (Saltz, 2021).

A compreensão do negócio, especificamente no que se refere à compreensão das temáticas relativas à segmentação e ao setor farmacêutico, foi inicialmente tratada no Capítulo 2, que apresenta a revisão da literatura, onde foram abordados os principais temas relevantes para esta investigação. Relativamente aos objetivos do negócio, estes consistem em caraterizar os clientes, identificar os segmentos de clientes e definir os perfis de clientes através dos segmentos para desenvolver estratégias de marketing direcionadas. Compreender os diferentes comportamentos e preferências dos clientes permite às farmácias personalizarem as suas ofertas e melhorarem a experiência destes.

A etapa de avaliação da situação, envolve uma análise a vários fatores. Identificar se as bases de dados provenientes da empresa de consultoria possuem qualidade para efetuar as análises necessárias ao desenvolvimento da segmentação (fase da compreensão dos dados). Análise dos recursos necessários, neste caso são utilizadas durante o desenvolvimento da investigação as ferramentas Microsoft Excel, Microsoft Power BI e principalmente KNIME *Analytics Platform* (posteriormente designada por apenas KNIME).

KNIME é uma plataforma que permite analisar dados, sendo este caraterizado por ser acessível e iterativo, é uma ferramenta que utiliza nós para processar dados, explorar resultados de análises e/ou de modelos criados, ajuda na criação de relatórios e integração de dados de diversas fontes, e apresenta escalabilidade suficiente para trabalhar dados em ambiente de Big Data. Por último, porém igualmente relevante, esta plataforma oferece uma versão gratuita (Berthold *et al.*, 2009; Fyson, 2024).

Por fim, respeitante aos riscos, constrangimentos e limitações, é possível encontrar na literatura exemplos de estudos que ajudam a antecipar esses desafios, como é descrito no estudo Kevrekidis *et al*. (2018) é necessário prestar devida atenção à escolha de uma amostra representativa, a seleção do método de análise exploratória adequados e à influência das variáveis escolhidas para a análise.

Com base nos objetivos estipulados para o negócio são desenhados os objetivos específicos do projeto. Para este estudo, os objetivos são a caracterização dos clientes através de análises descritivas iniciais, por forma a conhecer melhor os clientes antes de iniciar a segmentação, e posteriormente utilizar técnicas de *clustering* para identificar

segmentos de clientes e caraterizar perfis de clientes com base nos padrões de compra de produtos OTC, tendo em conta caraterísticas demográficas, relacionadas com o produto e comportamentais (análise RFM). Através da análise dos dados do histórico de compras, pretende-se identificar grupos de clientes com comportamentos de compra semelhantes permitindo uma compreensão mais detalhada dos diferentes segmentos presentes na base de clientes da farmácia.

A criação de um plano de projeto detalhado é importante para garantir que a sua execução seja cumprida com sucesso. Este plano segue a metodologia CRISP-DM, que fornece um *framework* estruturado para conduzir o tipo de projeto que se pretende desenvolver. O plano inclui as fases do processo CRISP-DM até à avaliação, e se necessário é possível avançar e retroceder entre as fases conforme o desenvolvimento do projeto.

3.2 Compreensão dos dados

Na fase da compreensão dos dados é crucial recolher os dados das diversas fontes disponíveis, explorar e descrever as suas características, e ainda verificar a qualidade dos mesmos e definir ações caso seja necessário melhorá-los (Schröer *et al*., 2021), sendo assim, repartida em quatro etapas, recolha, descrição, exploração e verificação da qualidade dos dados (Saltz, 2021).

A fase prática inicia-se pela etapa da recolha dos dados necessários para elaboração da investigação. A base de dados principal, partilhada pela empresa de consultoria, contém dados sobre o histórico de compras dos clientes das farmácias comunitárias, e as duas bases de dados secundárias são provenientes da plataforma Github (Lafuente, 2018) que contêm os códigos postais, e o nome dos distritos e ilhas correspondentes. As 3 bases de dados acima referidas encontram-se inicialmente em forma de tabela em ficheiros CSV *(comma-separated values*).

Relativamente à descrição, exploração e verificação da qualidade dos dados, foram efetuados dicionários de dados (exemplo apresentado na tabela 1) e realizado um conjunto de análises descritivas, de modo a melhor caraterizar os dados.

Figura 3: Distribuição dos clientes pelo território português
Fonte: Elaboração própria, em Power BI

Figura 2: Distribuição das farmácias pelo território português
Fonte: Elaboração própria, em Power BI

A base de dados principal contém informações, como anteriormente referido, sobre o histórico de compras de 35 382 clientes e 9 farmácias comunitárias, com um horizonte temporal de 2022 a 2024. Através da figura 3, observa-se que a distribuição da origem dos clientes abrange todo o território nacional, incluindo ilhas. No entanto, há uma maior concentração nos distritos de Lisboa, Setúbal e Santarém. Isso deve-se à localização das farmácias, conforme demonstrado na figura 2, que se encontram situadas nesses mesmos distritos.

Os clientes das farmácias comunitárias são caracterizados como indivíduos que realizam a compra de produtos ou serviços, podendo ou não ser os consumidores finais. Em alguns casos, o comprador e o consumidor são a mesma pessoa, contudo, também é comum que o comprador adquira produtos para terceiros, como familiares ou colegas.

Nesta base de dados estão presentes dados demográficos dos clientes, como data de nascimento, código-postal e género, dados relacionados com o produto, como o nome do produto, segmentos, marca, e dados comportamentais de compra, como o valor da compra, as quantidades e a data da compra.

A etapa da análise da qualidade dos dados foi efetivamente iniciada após eliminação dos *missing values* da variável ID Cliente (detetados 612 872 *missing*s), visto que para a elaboração da segmentação requer apenas dados relacionados com os clientes, a ausência do código identificador inviabiliza a inclusão desses registos na análise. Esta tarefa permitiu que subsequentemente fossem realizadas análises mais precisas e confiáveis dos padrões e características dos clientes. A tabela 1 apresenta o dicionário de dados após eliminar os *missing values* da variável ID Cliente totalizando 35 382 clientes e 397 926 linhas de vendas:

Tabela 1: Dicionário de dados, tabela "Histórico de Compras"
Fonte: Elaboração própria

Nome da variável	Descrição	Tipo de variável	*Missing Values*
ID_Cliente	Código identificador único do cliente	Texto	0
ID_Farmacia	Código identificador único da farmácia	Texto	0
Produto_ID	Código identificador único do produto	Texto	0
Data_Venda	Data em que foi realizada a venda	Data (formato: YYYYMMDD)	0
Valor_Venda	Valor total da venda	Numérico	0
Quantidade_ Vendida	Quantidades vendidas	Numérico	0
Responsavel_ Comercial	Responsável comercial do produto vendido	Texto	10 191
Marca_Produto	Marca do produto vendido	Texto	1 731
Segmento_Produto	Segmento do produto vendido	Texto	0
Codigo_Postal	Código postal do Cliente	Texto	0
Tipo_Pessoa	Se o tipo de cliente é Singular ou Coletivo	Texto	0
Genero	O género do cliente	Texto	110 722
Data_Aniversario	Data de aniversário do cliente	Data (formato: YYYYMMDD)	122 135

Após inserção das bases de dados em KNIME através do nó *Excel Reader* e utilizado o nó *Joiner* para junção das diversas tabelas, para deteção de *missing values* é aplicado o nó *Value Count* e para análise estatística foi utilizado o nó *Statistic View*.

A análise da qualidade dos dados é efetuada através do KNIME e foram detetadas duas variáveis Data de Venda e Data de Aniversário consideradas no KNIME por tipo texto incorretamente, são efetuadas ações de preparação de dados para alterar as mesmas para o tipo data (do formato YYYYMMDD para DD/MM/YYYY) para permitir uma análise temporal mais precisa e facilitar a manipulação destes dados para efetuar a preparação da análise RFM.

Relativamente às variáveis das quantidades vendidas e valor da venda, estas apresentam valores negativos ou igual a zero, por exemplo, “-1”, possivelmente relacionados com erros de introdução de dados. Para solucionar este problema, na fase da preparação de dados estes valores são eliminados, pois estão presentes em apenas 3 976 linhas de compras para a variável quantidades vendidas e 1 339 linhas de compras para a variável valor de venda (corresponde a menos de 1% do total de 397 926 linhas de compras).

No que diz respeito, às variáveis género e data Aniversário são as que apresentam maior número de *missing values* o que poderá estar relacionado com a sensibilidade de disponibilização destes dados por parte dos clientes. Contudo, relativamente ao número total de clientes estas variáveis não apresentam um número elevado de *missing*s (em Data Aniversário totaliza um número de 2 897 clientes (8% do total), no género, 1 231 clientes (3% do total) como é possível constatar na figura 4, são consequentemente eliminados na fase da preparação de dados.

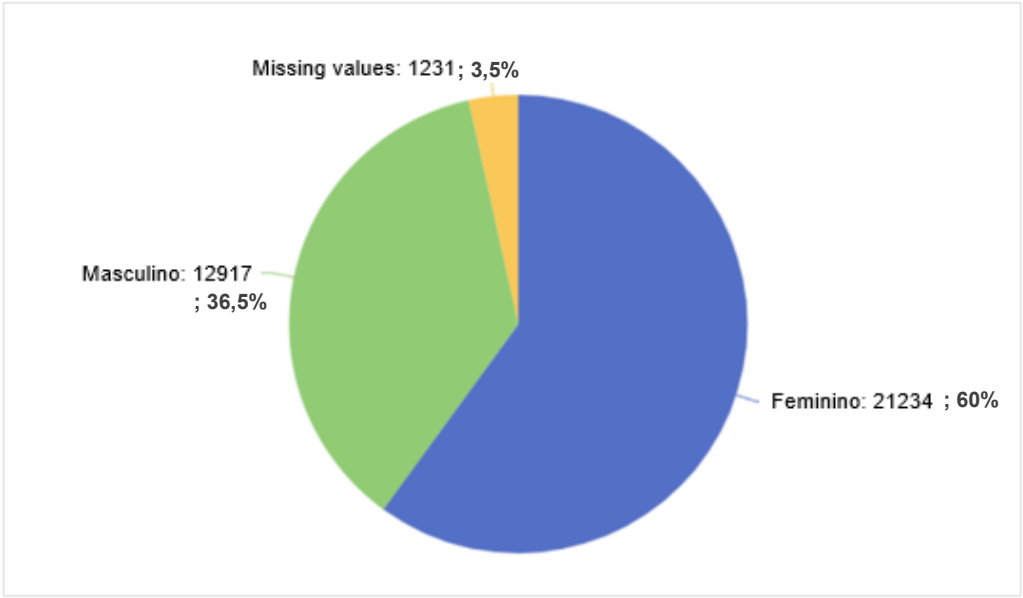

Figura 4: Distribuição do género dos clientes
Fonte: Elaboração própria (em KNIME)

Relativamente aos dados relacionados com o produto, existem quatro variáveis Produto ID, marca do produto, responsável comercial e segmento do produto, através da análise destas variáveis é possível identificar 93 754 produtos, 1 889 marcas, 521 responsáveis comerciais e 18 segmentos do produto.

Responsável comercial de produtos de farmácias são entidades responsáveis pela gestão e execução das operações comerciais, incluindo vendas, distribuição e promoção de produtos, sendo estes encarregues das negociações com fornecedores e distribuidores. Na variável responsável comercial existem 10 191 *missing values* e na variável marca do produto

existem 1 731 *missing values*, na fase de preparação estes missing values serão eliminados, por forma a não prejudicar as análises futuras.

Na base de dados são apresentados 18 segmentos do produto, e é possível verificar através do gráfico da figura 5, os top cinco segmentos favoritos dos clientes, que os dois segmentos com mais quantidades compradas ao longo dos 3 anos são e "Dermofarmacia, cosmética e acessórios" e “Sistema Respiratório”. Os segmentos com menos vendas são a “Saúde Animal” e “Solares” (376 e 1 926 quantidades vendidas, respetivamente).

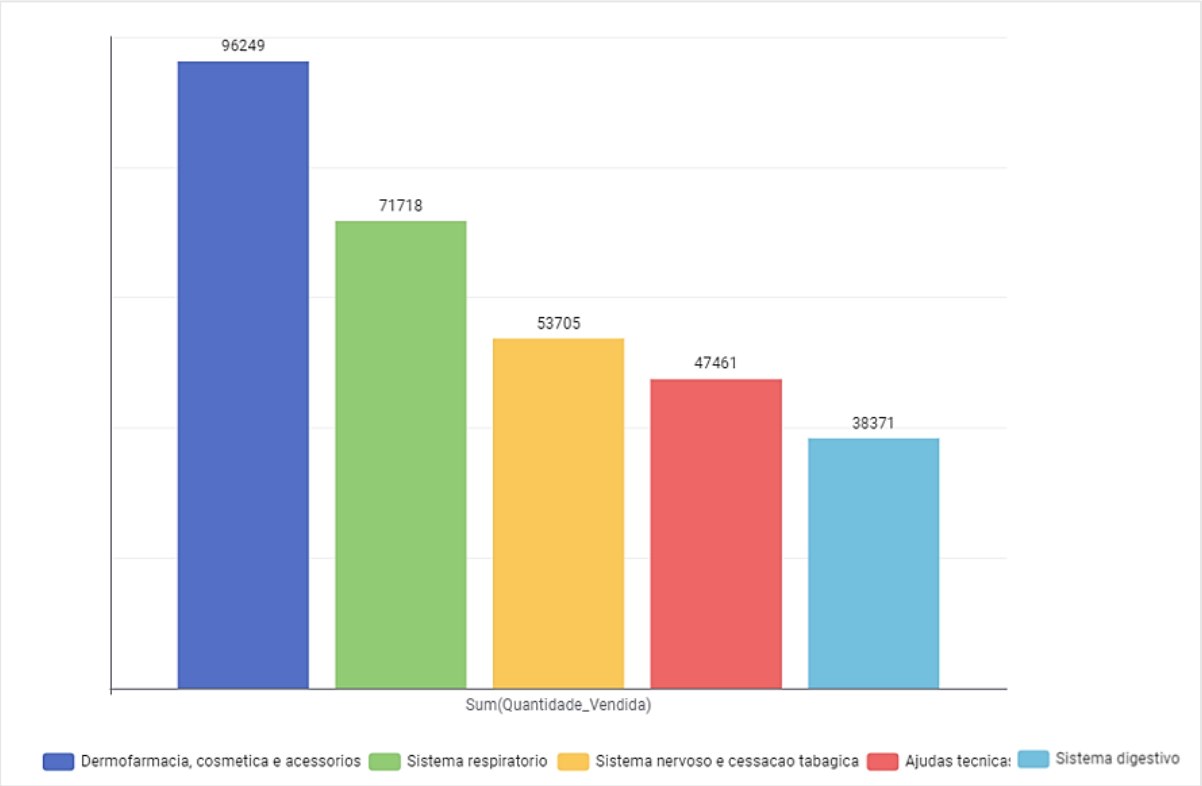

Figura 5: Top 5 segmentos mais vendidos
Fonte: Elaboração própria (em KNIME)

Relativamente aos dados provenientes das tabelas da plataforma Github, é apenas necessário recorrer a alterações na tabela códigos postais efetuadas diretamente na plataforma Excel, como a seleção das variáveis úteis à análise, que neste caso são o código distrito, o número código postal e a extensão código postal, eliminando variáveis relacionadas com os concelhos, uma vez que apenas são utilizados os distritos para análise. Posteriormente na preparação de dados é necessário criar a variável código postal através da junção das variáveis número código postal e extensão código postal para obtenção do distrito correspondente aos códigos postais dos clientes (por exemplo, o número “2480” e

extensão “220” juntamente corresponde a “2840-220” que indica um concelho pertencente ao distrito de Setúbal).

A fase da compreensão de dados será realizada também durante a preparação de dados, uma vez que ao transformar diversas variáveis será necessário entender o estado em que se encontram, e se é necessário efetuar outras alterações para garantir a qualidade das mesmas.

3.3 Preparação dos dados

A preparação de dados é considerada a fase de maior aplicação de esforços durante o desenvolvimento do projeto, em que é despendido cerca de 50% a 80% do tempo (Barapatre *et al*. 2017), posto isto é feita a explicação detalhada de algumas etapas, com o objetivo de aprofundar o conhecimento sobre esta fase na aplicação KNIME Analytics Platform. Esta fase é constituída por cinco etapas (Saltz, 2021): a seleção, a limpeza, a construção, a integração e a formatação das variáveis, para garantir a qualidade das variáveis para a fase da modelação.

Relativamente à seleção dos dados foram selecionadas apenas as variáveis que seriam necessárias à análise, eliminando as restantes variáveis ID Produto e Utente Código Postal. As variáveis selecionadas para análise foram renomeadas para melhorar a clareza e compreensão dos dados, através da seleção de nomes descritivos e intuitivos.

Antes de iniciar as etapas de limpeza, através da utilização do nó *Rule Engine*, é criada a variável Localidade Farmácia inserindo as localidades destas através de códigos desenvolvidos, sendo que as mesmas não foram disponibilizadas na base de dados inicial.

Como referido anteriormente no dicionário de dados, a variável Data Aniversário é a variável que contém o maior número de *missing value*s (valores que apresentam "00:00"), e para eliminar os mesmos é efetuada uma remoção destes através do nó *Row Filter*. Após tratamento dos *missing value*s é necessário alterar o tipo da variável para data, visto que, apresenta o tipo texto, é utilizado o nó *String to Date&Time* indicando o tipo de data que apresenta “DDMMYYYY”.

A data de aniversário é importante para análise, dado que, permite a criação da variável Idade através da diferença desta e da data atual. Essa operação é realizada pelo nó *Date&Time Difference*, resultando no número exato de anos do cliente até à data corrente.

Com recurso novamente ao nó *Row Filter* é estabelecido um intervalo da variável Idade de 18 a 100 anos focando nos clientes que representam o típico público-alvo, eliminando outliers e possíveis erros de registo, garantindo que análise seja mais precisa e relevante.

Para facilitar a análise é criada a variável Faixa Etária recorrendo ao nó *Rule Engine*, agrupando os clientes em categorias de idade, permitindo uma segmentação mais eficaz. Como essa categorização são capturadas as informações essenciais sobre a idade dos clientes, as variáveis detalhadas são, portanto, descartadas. O código utilizado no nó *Rule Engine* é apresentado na figura 6:

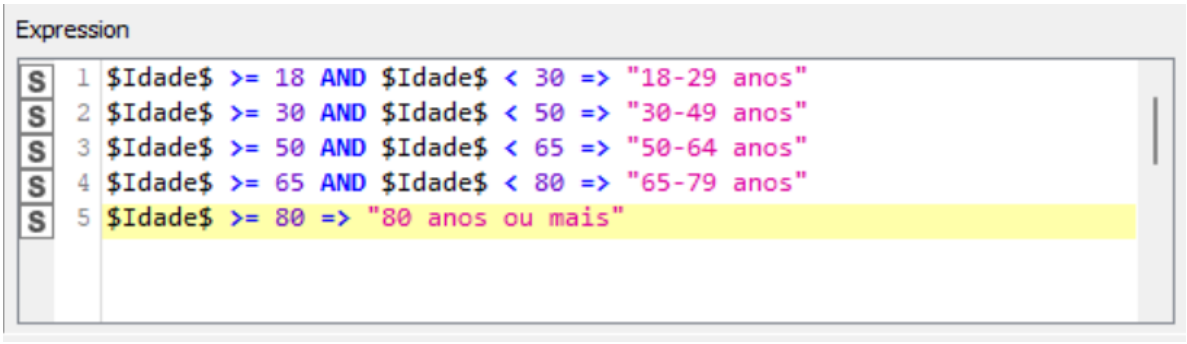

Figura 6: Código criação Faixas Etárias
Fonte: Elaboração própria (em Knime)

Por fim, são eliminadas as variáveis Data Aniversário e Idade, pois a variável Faixa Etária é considerada mais relevante para a fase da segmentação. O fluxo de trabalho do processo descrito anteriormente pode ser visualizado na Figura 7.

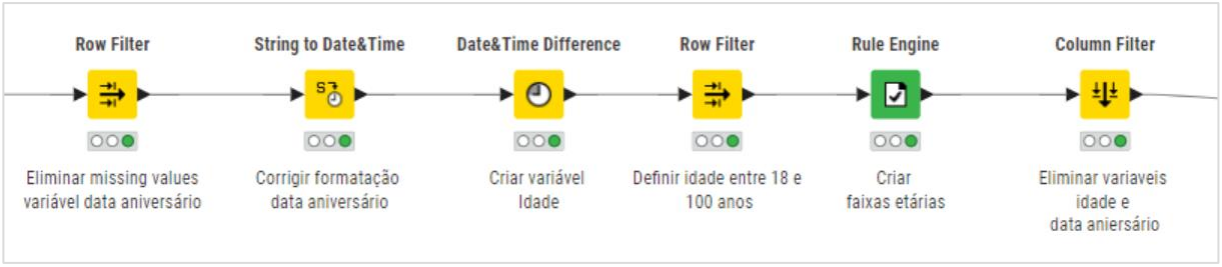

Figura 7: Workflow com o processo de criação das faixas etárias
Fonte: Elaboração própria (em Knime)

A variável Quantidade Vendida é uma das variáveis que apresenta valores incorretos (1 007 valores), que neste caso apresentam valores de quantidades negativas ou nulas, as quantidades negativas podem indicar incorreta inserção de dados, são eliminados através do nó *Row Filter*, uma vez que podem ser casos de erros ou ofertas e brindes, e estes são irrelevantes para a segmentação, sendo que não refletem o comportamento de compra

regular dos clientes. Relativamente à variável Valor de Venda, esta também sofreu alterações, para conseguir alterar o seu tipo inicial de texto para numérica, com recurso ao nó *String Manipulation* são substituídos os "." por "," resultando em valores como "12,38", o que permitiu através do nó *String to Number* passar a variável para numérica. É possível verificar o processo de preparação de dados da variável Valor da Venda através da figura 8:

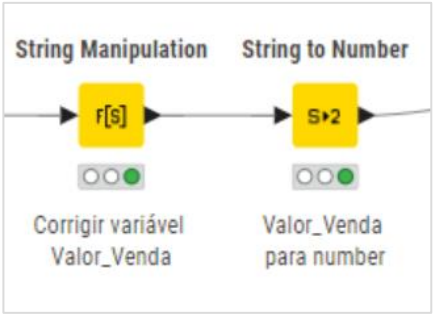

Figura 8: Preparação da variável Valor de Venda
Fonte: Elaboração própria (em Knime)

Com o intuito de facilitar o processo é utilizado o nó *Missing Value* para eliminar os valores omissos das restantes variáveis: Género, Responsável Comercial e Marca.

Em relação à variável Data da Venda, que estava originalmente no formato "DDMMYYYY", é convertida para formato de data "DD/MM/YYYY" utilizando o nó *String to Data&Time*, para conseguir obter o horizonte temporal das vendas.

No que respeita ao passo final da primeira fase do tratamento de dados, workflow em detalhe no Anexo A (figura A1 e A2), é utilizado o nó *Row Filter*, figura 9, para filtrar a base de dados pela farmácia com maior número de vendas, que neste caso é a farmácia S (totalizou um número de vendas de 49 575 em 238 792), segmentando assim apenas os clientes dessa farmácia em específico. Caso necessário, é possível ajustar o filtro para outras farmácias ou removê-lo completamente para obter uma visão global de todas.

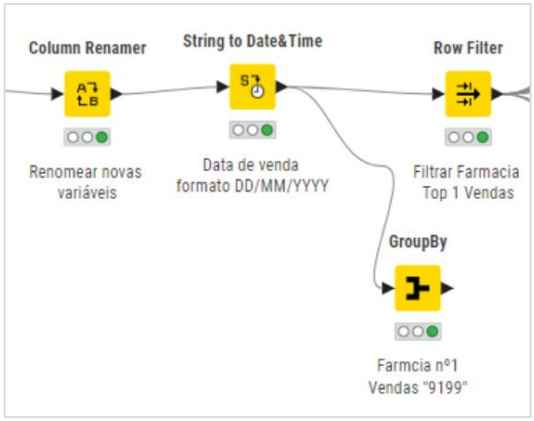

Figura 9: Processo de filtrar os dados pela farmácia com maior volume de vendas
Fonte: Elaboração Própria

A segunda parte da preparação de dados é específica para cada uma das segmentações serão elaborados 3 tipos de segmentação diferentes, demográfica, comportamental e relacionada com o produto. O processo de preparação de dados para os dados relacionados com o produto (variáveis Marca, Responsável Comercial e Segmento) envolvem o processo de identificação das preferências dos clientes. Como os clientes não escolhem sempre a mesma marca, é essencial analisar as suas compras para determinar qual a sua marca favorita, assim como o segmento e o responsável comercial da sua preferência.

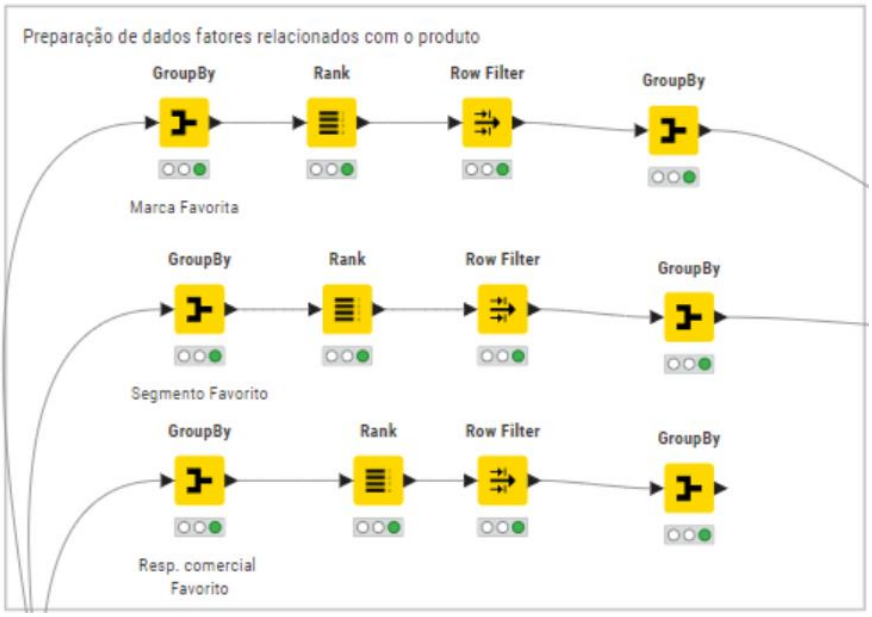

Figura 10: Processo de Ranking da segmentação relacionada com o produto
Fonte: Elaboração própria (em KNIME)

Para que seja definido o "favorito" é efetuado o processo de *ranking*, observado na figura 10, que no caso da Marca Favorita é inicialmente utilizado o nó *GroupBy* agrupando os clientes e as marcas pela soma das quantidades vendidas, de seguida é utilizado o nó *Rank* para ordenar as quantidades vendidas por ordem decrescente e atribuído um valor de ordem decrescente, posteriormente é utilizado o nó *Row Filter* para filtrar as linhas com o *Rank* igual a 1, significa que o cliente comprou um maior número desses produtos, obtendo assim a marca favorita por cliente. Para completar o processo de preparação das variáveis, Responsável Comercial e Marca é feita a anonimização destes com recurso a nós específicos (*Anonymization* e *String Manipulation*), com o propósito de poder mostrar as informações destes na visualização dos resultados. Relativamente aos segmentos foi ainda criada uma variável que englobando os segmentos em segmentos macro da variável, originando apenas 9 segmentos (em vez de 18), por forma a simplificar o processo de segmentação.

Antes de iniciar a fase da modelação, neste caso a aplicação dos algoritmos *K-means, DBSCAN e Hierarchical Clustering* é necessário transformar as variáveis categóricas em

dummy, sendo que esses algoritmos utilizam distâncias matemáticas e suas funções objetivo são baseadas em valores numéricos (Cheung *et al*. 2013). Para este processo, como comtemplado na figura 11, é utlizado o nó *One to Many*, que tem como objetivo transformar as variáveis categóricas em dummy. No caso específico da variável **Faixas Etárias**, classificada como uma variável ordinal, foi aplicada uma escala numérica por meio do nó **Rule Engine** para simplificar o processo. Por fim, as variáveis são normalizadas para se encontrarem todas na mesma escala, e o funcionamento dos algoritmos seja feito nas condições apropriadas.

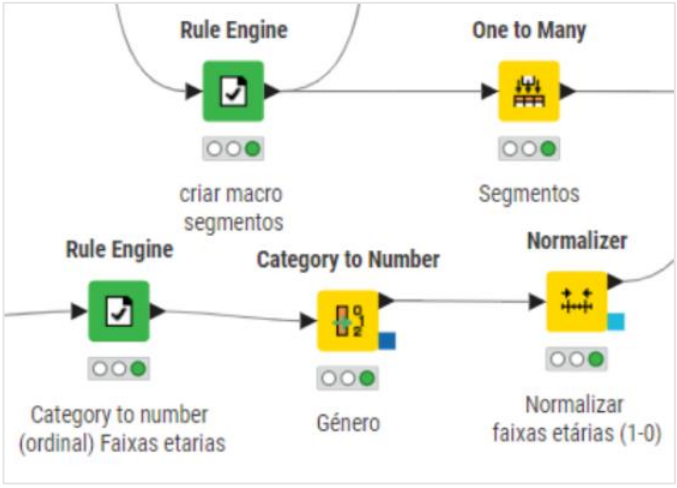

Figura 11: Transformação das variáveis categóricas em numéricas
Fonte: Elaboração Própria (Em KNIME)

A terceira fase da preparação de dados recai sobre a análise RFM, uma vez que nesta etapa é essencial adaptar as variáveis às três dimensões essenciais: Recência (Recency), Frequência (Frequency) e Valor Monetário (Monetary). Esta etapa é essencialmente baseada em um exemplo disponibilizado pelo KNIME (Hopf *et al*. 2023).

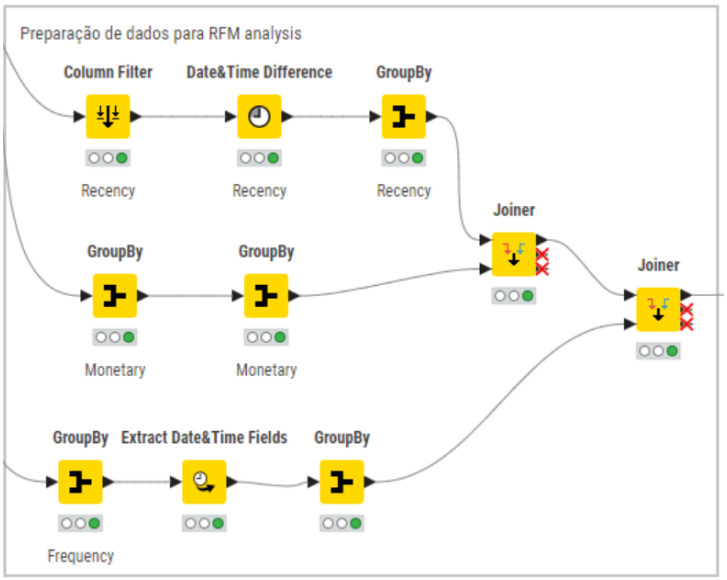

Figura 12: Preparação de dados para a análise RFM
Fonte: Elaboração própria (em KNIME)

A figura 12 demostra o *workflow* da preparação inicial das variáveis Recência, Monetário e Frequência. A variável Recência é criada com base na variável Data de Venda para obter a diferença de dias entre o dia da última compra do cliente e o fim de recolha de dados, esta fórmula é efetuada através do nó *Date&Time Difference*, posteriormente é agrupada a variável ID Cliente de acordo com o mínimo de dias obtido do nó *Date&Time Difference*.

A variável Monetário é criada com base na variável Valor de Venda para obter o ticket médio de compra, de seguida é agrupada a variável ID Cliente e Data da Venda para obter o Valor de Venda médio gasto por data, e de seguida é agrupada variável ID Cliente efetuando a média do valor de venda por cliente.

Para criar a variável Frequência é necessária a variável Data da Venda, sendo que é preciso obter a quantidade de vezes que o cliente foi à farmácia durante os 3 anos. Para iniciar o processo são agrupados os clientes por datas de venda, posteriormente é utilizado o nó *Extract Date&Time Fields* para obter o ano da venda, depois é agrupado a variável ID Cliente pela contagem da variável proveniente do nó anterior (contagem da quantidade de vezes que o cliente frequentou a farmácia).

Finalizando este processo é necessário verificar se a qualidade dos dados se mantém, neste caso são analisados os outliers de cada uma das 3 variáveis RFM, através dos gráficos *Box Plot*. Por fim, é necessário remover os outliers encontrados para não influenciarem negativamente a utilização dos algoritmos, resultando na diminuição do número de clientes de 6508 para 5886.

Concluída a fase de preparação dos dados, procede-se à elaboração de uma última análise descritiva dos clientes da farmácia S utilizando a ferramenta Power BI (workflow presente no Anexo C, figura C2), uma vez que é possivel através da utilização de diferentes nós ligar automaticamente o KNIME ao Power BI. Esta análise visa proporcionar uma compreensão aprofundada das características dos clientes, permitindo à empresa conhecê-los melhor e, assim, apoiar na tomada de decisões estratégicas antes de avançar para a segmentação, que será realizada com o auxílio de algoritmos na fase da modelação. Os resultados desta análise são abordados posteriormente na seção dos resultados.

3.4 Modelação

A fase de modelação envolve a seleção dos algoritmos e técnicas, construção do modelo, considerando o problema de negócio e os dados disponíveis, sendo necessário a definição de parâmetros específicos (Schröer *et al*., 2021). O software de *low-code* KNIME, torna-se apropriado para este tipo de análise, uma vez que este contém os seguintes fatores essenciais: tem escalabilidade suficiente para suportar dados em ambiente *Big Data*, é uma ferramenta gratuita, e permite a análise de dados, criação de relatórios e integração de dados de diversas fontes (Fyson, 2024).

Na fase de modelação do processo de segmentação de clientes, são aplicados três algoritmos diferentes *K-means,* DBSCAN e *Hierarchical clustering,* utilizando os nós KNIME apresentados na figura 13, o workflow da aplicação destes algoritmos encontra-se em detalhe no Anexo D, figura D1.

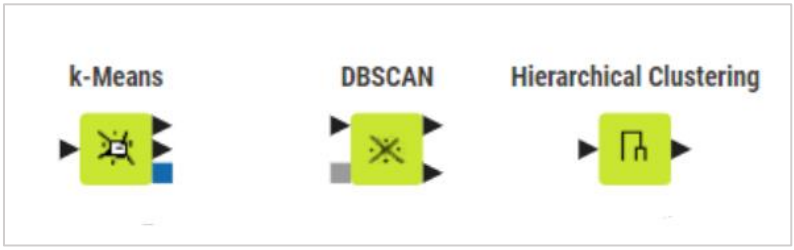

Figura 13: Algoritmos utilizados na fase da Modelação
Fonte: Elaboração própria (em KNIME)

O algoritmo *K-means*, é uma escolha justificada por ser um dos métodos mais utilizados para segmentação em diversos estudos, incluindo exemplos anteriormente referidos na revisão da literatura. A utilização do algoritmo *K-means* deve-se à sua eficácia em agrupar clientes com base em características semelhantes, o que é essencial no contexto de farmácias, onde a personalização das estratégias de marketing pode ter um impacto significativo. O algoritmo permite a criação de *clusters* distintos, facilitando a segmentação demográfica, comportamental e relacionada ao produto.

O algoritmo *K-means* é um método de *clustering* não hierárquico que agrupa dados em K *clusters*, maximizando a similaridade interna dos grupos formados e minimizando a similaridade entre os grupos. O processo do algoritmo é ilustrado na figura 14, primeiro envolve a escolha de K centroides iniciais, atribuindo dados ao centróide mais próximo, recalculando os centroides, e repetindo até que os centroides estabilizem, utilizando a distância euclidiana para determinar a proximidade entre os pontos. Quando não há mais mudanças nos centroides, o processo de *clustering* encontra-se completo (Gustriansyah *et al*., 2019).

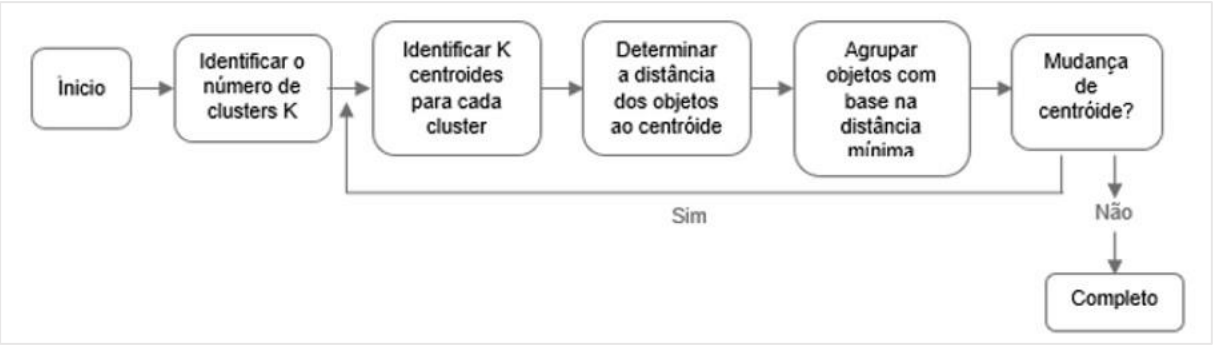

Figura 14: Fluxograma do processo do algoritmo K-means
Fonte: Adaptado de (W. Smith, 2015)

É fundamental que sejam testados mais do que um algoritmo para garantir que os resultados obtidos sejam os mais adequados aos dados analisados. Cada algoritmo atua de diferentes formas sobre a natureza dos dados e pode revelar padrões distintos. Neste caso foram selecionadas o *Hierarchical Clustering*, apesar de não ser uma escolha tão popular como o *K-means*, também é frequentemente utilizado para segmentação de clientes. E o algoritmo DBSCAN, sendo este menos popular que os anteriores, mas com diversas vantagens como, por exemplo, o facto de não ser necessário especificar o número de *clusters* e conseguir lidar bem com ruído existente nos dados (Tomar, 2013).

Com o objetivo de simplificar o processo de análise RFM, que isoladamente é considerado complexo, aplicou-se exclusivamente o algoritmo *K-means*, amplamente utilizado para a segmentação do comportamento de compra de clientes. Conforme destacado no estudo realizado por D. Chen *et al*. (2012), que utilizou o modelo RFM e o algoritmo *K-means* para a gestão de relacionamento com o cliente, os resultados experimentais demonstram que o modelo proposto é um método eficaz para a análise de valor do cliente. Na seção de resultados, o processo de segmentação RFM é explicado de forma mais detalhada para facilitar a compreensão da mesma.

Foram excluídas da segmentação as variáveis Marca Favorita e Responsável Comercial devido à alta dimensionalidade dos dados, uma vez que os algoritmos de *clustering* não lidam bem com bases de dados de grandes dimensões. A variável Marca apresentava aproximadamente 700 categorias distintas, e a variável Responsável Comercial cerca de 200, o que pode comprometer o desempenho e a qualidade dos *clusters* gerados. O aumento no número de dimensões impacta a medição de distância entre os pontos, e como os algoritmos de *clustering* dependem dessas medidas, é crucial que os objetos de um *cluster* estejam mais próximos entre si do que de outros *clusters*, caso contrário, os resultados podem perder significado (Steinbach *et al*., 2003).

A variável localização não foi incluída na análise de segmentação, pois a sua a relevância para a caracterização dos clientes é considerada limitada. Além de não agregar valor significativo para a identificação dos perfis dos clientes, a sua inclusão influenciava negativamente a aplicação dos algoritmos, sobretudo ao ser combinada com a variável "segmento", que possui maior importância na definição e distinção dos grupos e de estratégias de marketing direcionadas.

Foram aplicados os três algoritmos anteriormente referidos às segmentações, presentes na tabela 2, relacionada com o produto e demográfica (denominada posteriormente como “Demográfica-Produto”), os parâmetros apresentados na tabela são os que foram definidos para alcance dos melhores resultados.

Tabela 2: Algoritmos da segmentação Demográfica-Produto e respetivos parâmetros utilizados
Fonte: Elaboração própria

Tipo de segmentação	Variáveis Incluídas	Algoritmo	N.º *Clusters* definidos	Outros Parâmetros
Demográfica-Produto	Faixa etária, género e segmento do produto	*K-means*	6	N/A
		Hierarchical Clustering	6	*Euclidian distance, Linkage Average*
		DBSCAN	N/A	*0,5 Epsilon, 4 Minimum Points*

O *Elbow Method* é utilizado para determinar o número ideal de clusters. Este método permite identificar o melhor número de clusters através de uma visualização gráfica, como mostrado na Figura 15 que avalia a soma dos erros quadrados (SSE) entre os dados e os centroides dos clusters para diferentes valores de k. O ponto de inflexão observado no gráfico ocorreu em k=5, ponto que se encontra mais distante da reta traçada entre as duas extremidades da curva (Delgado *et al.*, 2015). No entanto, ao experimentar k=6, foram detetadas melhorias adicionais, proporcionando uma segmentação mais eficiente e minimizando o erro. O aumento posterior do número de clusters não traria melhorias substanciais, justificando, assim, a escolha de k=6 como a configuração ideal para o modelo (Berthold *et al.*, 2009; Delgado *et al.*, 2015; Umargono *et al.*, 2020). O *workflow* associado à aplicação deste algoritmo encontra-se no Anexo B.

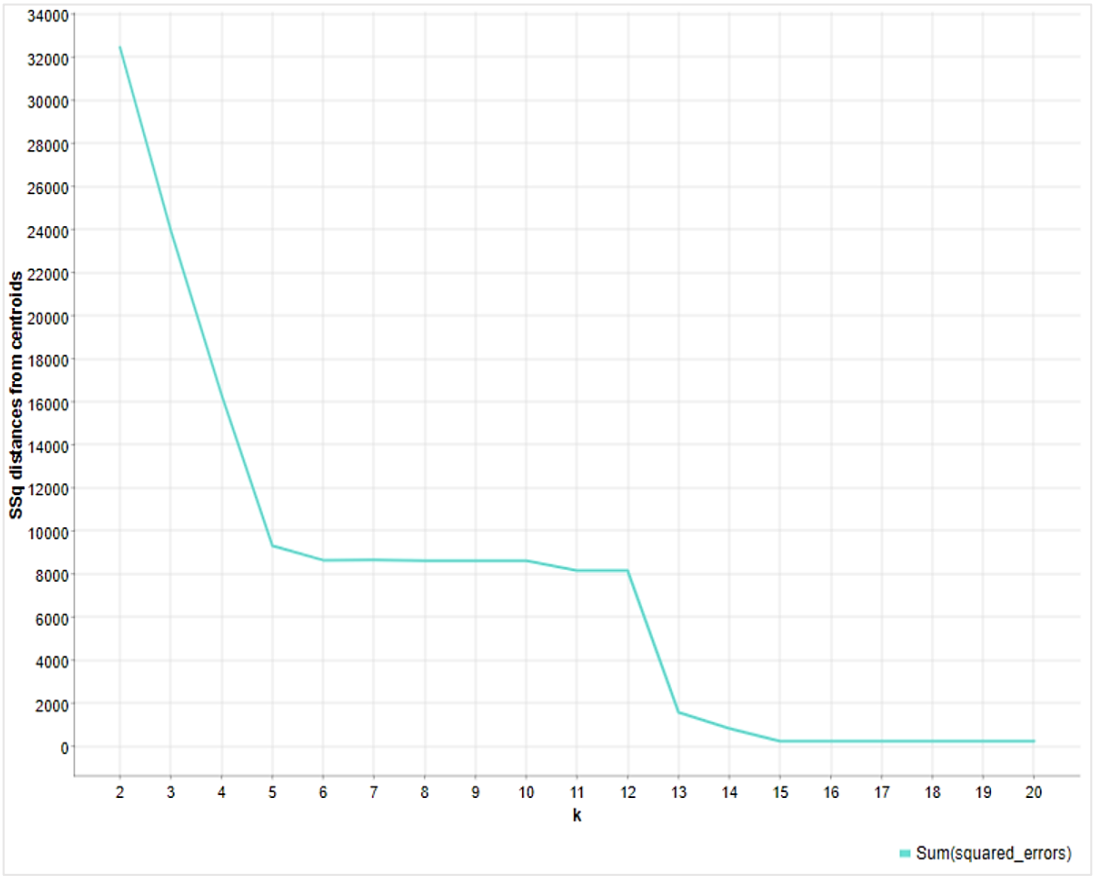

Figura 15: Gráfico Elbow Method: variação da soma dos erros quadrados em função do número de clusters
Fonte: Elaboração própria (em KNIME)

3.5 Avaliação

Nesta fase realiza-se a análise dos resultados obtidos em relação aos objetivos de negócio estabelecidos e, com base nos resultados, se necessário, são estipuladas ações adicionais (Schröer *et al.*, 2021).

Inicialmente é avaliada a coerência teórica dos modelos, posteriormente é avaliada a qualidade dos *clusters* gerados pelos algoritmos. A avaliação dos resultados dos algoritmos é efetuada através da métrica coeficiente de silhueta, que mede a coesão e a separação entre os grupos formados, assegurando que as segmentações realizadas sejam coerentes e relevantes para os objetivos do estudo. A fórmula do coeficiente de silhueta é apresentada através do KNIME da seguinte forma:

$$SC = \frac{b - a}{\max(a, b)}$$

Em que “a” representa a distância média entre os pontos dentro do mesmo *cluster*, e “b” corresponde à distância média até o *cluster* mais próximo. O algoritmo calcula tanto a média dos coeficientes de silhueta individuais quanto o valor geral, com resultados que variam de -1 e 1, representando a pior e melhor qualidade dos *clusters*, respetivamente (Massaro, 2022). Na figura 16 é possível visualizar o nó *Silhouette Coefficient*.

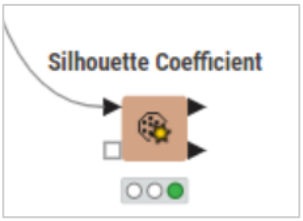

Figura 16: Nó coeficiente de silhueta
Fonte: Elaboração própria (em KNIME)

Após a elaboração a fase da modelação, em que foram aplicados os modelos, estes são expostos a uma fase de avaliação dos resultados obtidos, com o propósito de determinar qual o modelo que mais contribui efetivamente para os objetivos estabelecidos. A análise destes resultados (tabela 3 e 4) é feita na seção 4 (Resultados e Discussão).

Tabela 3: Resultados da segmentações Demográfica-Produto
Fonte: Elaboração Própria

Tipo de segmentação	Variáveis Incluídas	Algoritmo	N.º *Clusters* definidos	Outros Parâmetros	Coeficiente Silhueta
Demográfica-Produto	Faixa etária, género e segmento do produto	*K-means*	6	N/A	0,634
		Hierarchical Clustering	6	*Euclidean distance, Linkage Average*	*0,65*
		DBSCAN	N/A	*0,5 Epsilon, 4 Minimum Points*	*0,717*

Nota: para o algoritmo DBSCAN foi aplicado o Numeric Distance com: distância Euclidean, sendo que este é apropriado para lidar com variáveis mistas (neste caso binárias e ordinais)

Tabela 4: Resultados dos clusters da segmentação Demográfica-Produto (Hiearchical Clustering)
Fonte: Elaboração Própria

Segmentação Demográfica-Produto	Nº clientes	Coeficiente de silhueta
***Cluster* 0**	1569	0,72
***Cluster* 1**	2176	0,72
***Cluster* 2**	48	0,53
***Cluster* 3**	172	0,53
***Cluster* 4**	763	0,55
***Cluster* 5**	1780	0,57
Overall	**6508**	**0.65**

Para a segmentação RFM também é utilizado o coeficiente silhueta para avaliação do algoritmo *K-means*, contudo como anteriormente referido na fase da modelação, esta será detalhada na seção dos resultados.

4. Resultados e Discussão

4.1 Caraterização dos clientes da Farmácia S

Neste capítulo é apresentada a caracterização dos clientes da Farmácia S com base numa análise descritiva iterativa realizada em Power BI, cujos detalhes completos podem ser consultados no Anexo C, figura C2. Face ao objetivo analítico inicialmente estabelecido, estas análises foram conduzidas no final da fase de preparação dos dados, com o propósito de obter uma compreensão mais detalhada das características dos clientes antes de iniciar a fase de modelação.

As análises descritivas apresentaram resultados sobre os 6508 clientes da farmácia S, sobre caraterísticas demográficas, relacionadas com o produto e comportamentais (valores das variáveis RFM). Através das análises aos valores das variáveis RFM, verificou-se que a média da recência é de 164,55 dias, o ticket médio dos clientes é de 11,69€, e a frequência média de visitas ao longo dos últimos três anos é de 5,03 vezes. Esses resultados oferecem uma visão preliminar sobre o comportamento dos clientes, permitindo uma abordagem mais informada para as etapas subsequentes.

Analisando a figura 17, é constatado que o género feminino é predominante (cerca de 62%), e a maior parte deste género enquadra-se na faixa etária dos “30-49 anos”, o mesmo acontece com o género masculino (corresponde a 38% dos clientes). Por sua vez, a faixa etária com menos clientes é a “80 anos ou mais”, o que poderá indicar a menor mobilidade dessas pessoas para visitar farmácias fisicamente, ou o fato de muitos dependerem de cuidadores ou familiares para adquirir medicamentos e outros produtos farmacêuticos.

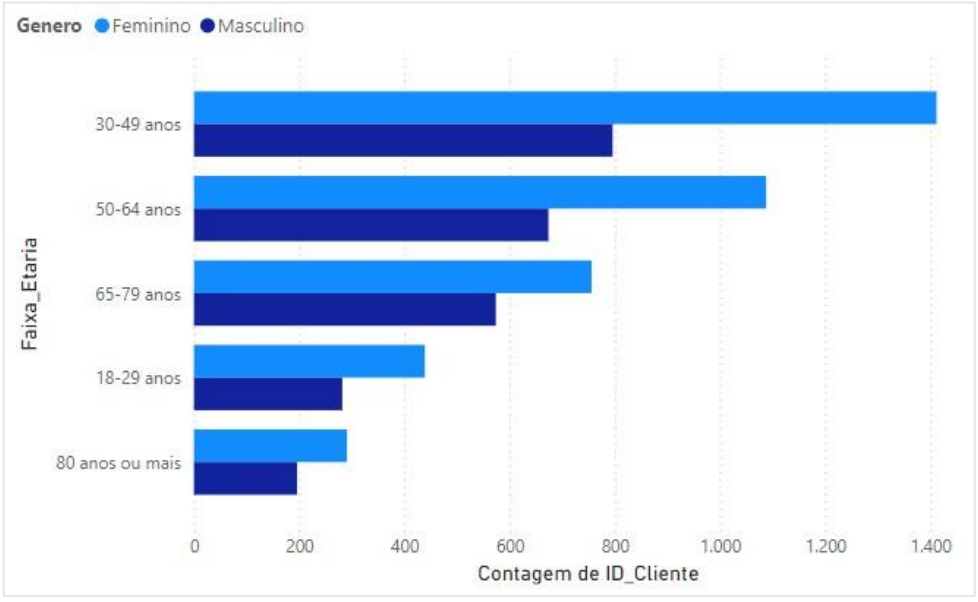

Figura 17: Distribuição dos clientes por género e faixa etária
Fonte: Elaboração própria (em Power BI)

Pode verificar-se através do gráfico presente na figura 18, os grupos de “30-49 anos” e “50-64 anos” juntos representam mais de 50% dos clientes em todos os segmentos. A faixa etária dos “30-49 anos” (34,73%) é predominante no segmento “Higiene, saúde e beleza”, enquanto nos “Testes e Monitorização de saúde” predomina o grupo “65-79 anos” (26,74%). A faixa etária dos “80 anos ou mais” representa a menor percentagem em quase todos os segmentos (exceto no segmento “cuidados com saúde geral”). Na figura 19 é possível observar o Top responsáveis comerciais favoritos, o género feminino predomina nos 5 responsáveis comerciais. (referindo novamente que os responsáveis comerciais são anonimizados em códigos por motivos de confidencialidade).

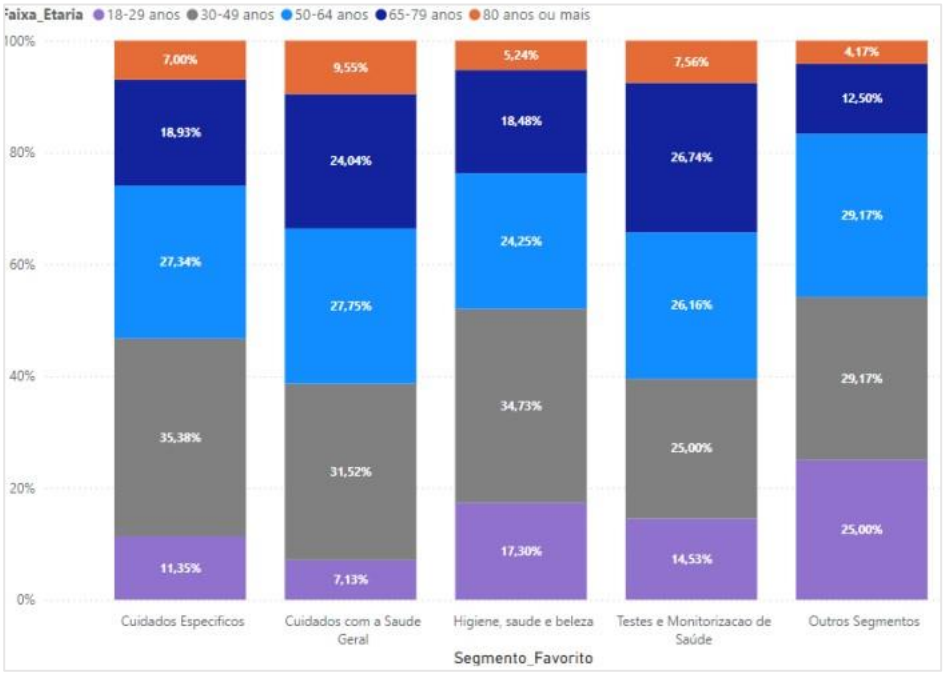

Figura 18: Distribuição da faixa etária por segmento favorito (Top 5)
Fonte: Elaboração própria (em Power BI)

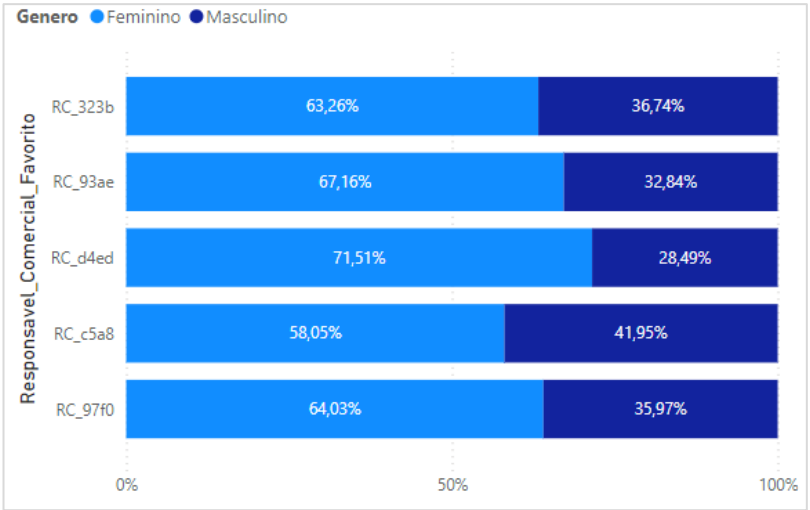

Figura 19: Distribuição do género por responsável comercial favorito (Top 5)
Fonte: Elaboração própria (em Power BI)

4.2 Avaliação dos modelos e caraterização dos clusters

Neste capítulo, são apresentados os resultados obtidos a partir dos modelos desenvolvidos, com o propósito de validar os objetivos de investigação inicialmente propostos.

4.2.1 Análise Segmentação Demográfica-Produto

Face aos resultados apresentados na seção 3.5, tabela 3, o algoritmo DBSCAN apresentou o melhor coeficiente silhueta, de 0,717, em comparação com os algoritmos *K-means* e *Hierarchical Clustering* (coeficiente silhueta de 0,634 e 0,65, respetivamente), indica uma segmentação de boa qualidade, dado que valores acima de 0,7 geralmente sugerem uma boa separação entre os *clusters* e uma coesão interna adequada, cumprindo com parte da validação do segundo objetivo estipulado sobre a caraterização dos segmentos (Seção 1.3). Contudo, a segunda parte da validação, relacionada com o valor dos segmentos não é cumprida, uma vez que este ao formar 10 *clusters*, e tendo em conta que este repartiu entre segmento do produto e género (5 segmentos do produto x 2 géneros = 10 *clusters*), originou uma fragmentação excessiva criando muitos *clusters* pequenos, o que dificulta a execução de campanhas de marketing eficazes.

Os 6 *clusters* formados pelo *Hierarchical Clustering* fornecem mais valor em termos de caraterização dos segmentos do que o algoritmo DBSCAN, apesar de apresentarem um coeficiente de silhueta inferior (0,65). No entanto, esse valor ainda supera o desempenho do algoritmo *K-means* (coeficiente silhueta de 0,63), destacando a relevância dos clusters no contexto da análise. Os resultados dos coeficientes silhueta provenientes da aplicação dos algoritmos DBSCAN e *K-means* encontram-se em detalhe no Anexo D, figura D2 e D3, respetivamente.

Os *clusters* originados pelo *Hierarchical Clustering* representam grupos de grande dimensão com interesses claros, como "Cuidados Específicos" e "Cuidados com a Saúde Geral", o que pode facilitar a criação de campanhas e ofertas direcionadas. Ao analisar individualmente os *clusters*, através da tabela 4 (seção 3.5) observa-se que o *Cluster* 1 e 0 apresentam o coeficiente mais elevado (0,72), o que sugere uma excelente coesão interna e clara distinção em relação aos outros grupos. Por outro lado, os *clusters* 2, 3, 4 e 5 têm coeficientes menores, variando entre 0,57 e 0,53, o que, embora ainda indique uma segmentação aceitável, pode sugerir alguma sobreposição ou menos clareza na diferenciação entre esses grupos. Para analisar os segmentos provenientes dos *clusters* formados através da aplicação do algoritmo *Hierarchical Clustering* foram realizadas análises descritivas, presentes nas figuras 20, 21 e 22. Estas análises servem para ajudar a melhor caraterizar cada segmento, como estipulado na seção 1.3., os segmentos identificados devem demonstrar relevância e utilidade para a tomada de decisões estratégicas no negócio.

Cluster String	Faixa_Etaria String	Count... Number ...	Genero String	Count... Number ...	Segmento_Favorito String	Count Number
cluster_0	30-49 anos	525	Masculino	1569	Cuidados Especificos	1569
cluster_1	30-49 anos	800	Feminino	2176	Cuidados Especificos	2176
cluster_2	30-49 anos	14	Feminino	27	Outros Segmentos	48
cluster_2	50-64 anos	14	Feminino	27	Outros Segmentos	48
cluster_3	65-79 anos	46	Feminino	93	Testes e Monitorizacao de Saúde	172
cluster_4	30-49 anos	265	Feminino	486	Higiene, saude e beleza	763
cluster_5	30-49 anos	561	Feminino	1204	Cuidados com a Saude Geral	1780

Figura 20: Características dos segmentos criados através do Hierarchical Clustering
Fonte: Elaboração própria (em KNIME)

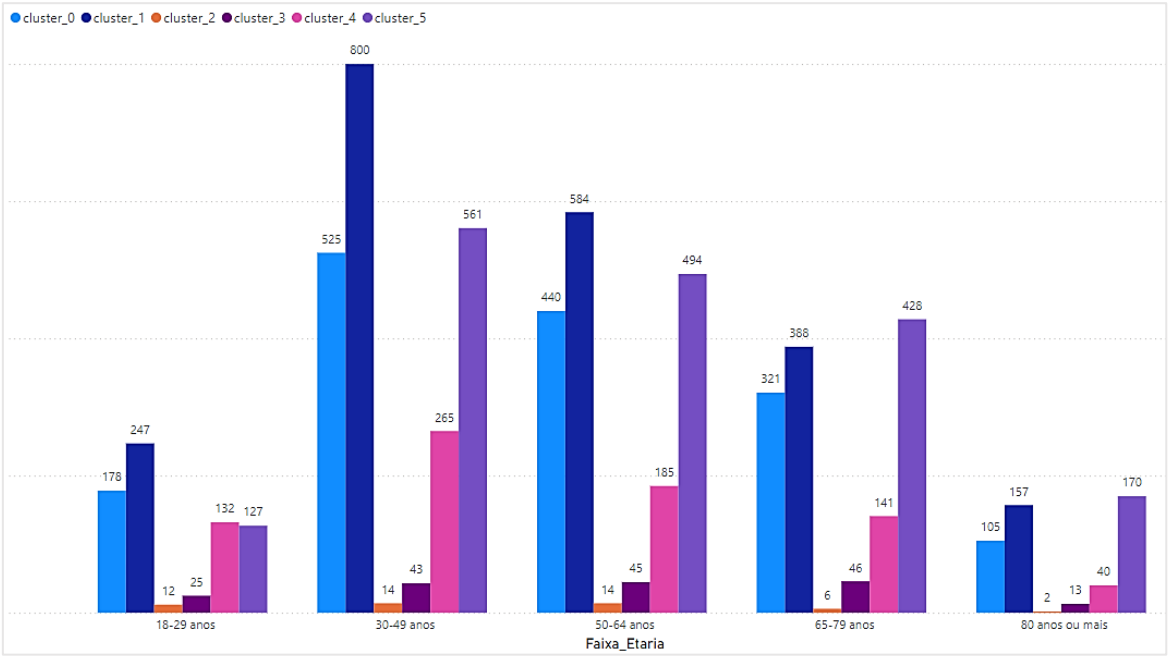

Figura 21: Distribuição dos clientes por faixa etária e cluster
Fonte: Elaboração própria (em Power BI)

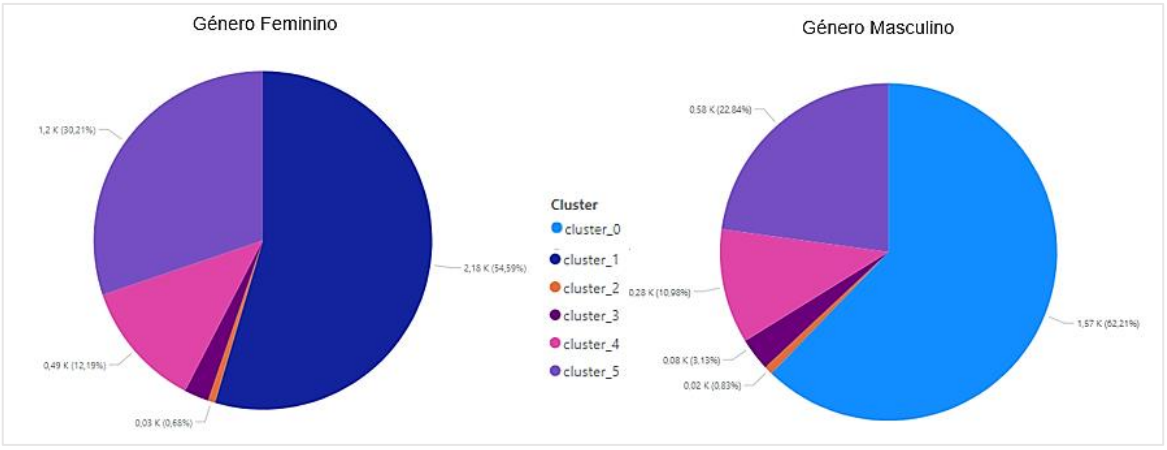

Figura 22: Distribuição dos clusters por género
Fonte: Elaboração própria (em Power BI)

O *cluster* 0 representa cerca de 24% dos clientes totais, denominados “Homens Cuidadosos”, é representado totalmente pelo género masculino, analisando a figura 22, e maioritariamente é constituído pela faixa etária dos “30-49 anos”, estes são focados principalmente em produtos para cuidados específicos, como os relacionados aos sistemas digestivo, respiratório e nervoso. O mesmo acontece para o segmento “Mulher Cuidadosas”, pertencentes ao *cluster* 1, é constituido por 33% do total de clientes, representados pelo mesma faixa etária (figura 20, identificam-se os 800 clientes pertencentes ao *cluster* 1), porém representados totalmente pelo género feminino, o que é possível concluir ao analisar a figura 22.

O *cluster* 2 “Mulheres Alternativas” totaliza 48 clientes, apresenta duas linhas na figura 20, este *cluster* identifica maioritariamente dois grupos etários, em concreto, entre os “30-49 anos” e entre os “50-64 anos”, em ambos os casos são referidos 14 clientes, sendo que em todo o espectro existem 27 clientes do sexo feminino, este segmento interessa-se pelo segmento de produto “Outros Segmentos”, tais como, homeopatia e saúde animal. Observando o gráfico da figura 21 é o grupo que apresenta o menor número de clientes em todas as faixas etárias.

O *cluster* 3 representado por 2% dos clientes, “Mulheres de Controlo”, encontram-se maioritariamente entre a faixa etária dos “65-79 anos” e são do género feminino, este segmento tem um forte foco na prevenção, monitorização e testes de saúde.

No *cluster* 4 denominado de “Mulheres de Glamour”, também predomina o género feminino (com 486 clientes, num total de 762), sendo que a principal aglomeração ocorre na faixa etária dos “30-49 anos”, sendo notório o interesse em produtos do segmento “Higiene, Saúde e Beleza”, como cosméticos, cuidados orais e solares.

Por fim, o *cluster* 5 designado “Mulheres Prudentes”, é interessado em cuidados gerais de saúde, como nutrição, suplementos alimentares e ajudas técnicas, sendo um grupo composto por 27% dos clientes totais, em que 67% destes são do género feminino (1204 clientes), e maioritariamente da faixa etária dos “30-49 anos”.

4.2.2 Análise RFM

Inicialmente, o algoritmo *K-means* foi aplicado separadamente a cada uma das variáveis da análise RFM, o *workflow* desta análise encontra-se em detalhe no Anexo E, figura E1. O parâmetro k=5 foi definido, resultando na formação de cinco *clusters* para cada uma dessas variáveis, estas variáveis apresentaram coeficientes silhuetas entre 0,53 e 0,68, são considerados aceitáveis, pois, embora não sejam muito altos, contribuem para a segmentação

final eficaz, isso é comprovado pelo *k-means* aplicado à variável *RFM score* (resultados a descrever mais adiante). Cada *cluster* originado agrupa os clientes com características semelhantes em relação a cada variável RFM, permitindo uma segmentação inicial dos dados de acordo com os padrões observados de consumo.

Após a definição dos *clusters*, os grupos foram classificados numa escala de 0 a 4, de acordo com o desempenho médio observado para cada *cluster*. Esse procedimento foi realizado para todas as variáveis do modelo RFM, possibilitando a criação de uma classificação ordenada, onde 0 representa o menor desempenho e 4 o maior desempenho.

A Tabela 5 apresenta as médias de cada uma das variáveis para cada um dos *clusters*, permitindo uma análise quantitativa e comparativa das classificações atribuídas. Neste caso podemos verificar que quanto maior a recência (o número de dias desde a última compra do cliente até fim de recolha de dados) menor a classificação, sendo considerados clientes menos propensos a realizar compras novamente. Por outro lado, o valor monetário quanto maior, melhor a classificação, demonstrando que está a investir mais dinheiro nos produtos ou serviços. Através do Anexo E, figura E2, é possível analisar uma amostra dos resultados dos clientes da farmácia S, com as classificações atribuídas por variável RFM.

Tabela 5: Médias dos clusters de cada variável RFM
Fonte: Elaboração própria

Escala de classificação	Recência Média	Monetário Média	Frequência Média
0	419,9	3,9	1,2
1	289,6	7,5	4,7
2	174,1	10,1	13,2
3	91,1	13,3	28,0
4	24,0	17,6	63,8

Com as classificações atribuídas a cada uma das variáveis, procedeu-se à soma das destas correspondentes a Recência, Frequência e Valor Monetário. Esta soma resultou na variável *RFM Score*, uma métrica que integra as três variáveis e reflete o comportamento agregado de cada cliente. A variável *RFM Score* foi, então, utilizada como base para uma nova aplicação do algoritmo *K-means*, visando identificar padrões de consumo mais abrangentes e refinados, com k=6, a fim de identificar segmentos de consumidores com padrões semelhantes de consumo. O resultado foi a formação de seis *clusters* distintos, que representam diferentes grupos de consumidores, segundo a classificação RFM agregada. As tabelas 6 e 7 resumem as características dos seis *clusters* formados, apresentando a média da variável *RFM Score* (resultante da aplicação do algoritmo *K-means*), o número de clientes em cada *cluster* e o coeficiente de silhueta, uma medida que reflete a qualidade e coesão

interna dos *clusters*, no Anexo E, figura E3, é apresentada uma amostra dos resultados das variáveis por cliente durante o processo, para facilitar a compreensão destes.

Tabela 6: Resultados dos clusters da variável RFM Score (K-means)
Fonte: Elaboração própria

Segmentos	RFM Score	N.º clientes	Coeficiente de silhueta	RFM *K-means*
Clientes Premium	8.6	998	0.42	*Cluster* 0
Clientes Passivos	2.2	1 443	0.42	*Cluster* 1
Clientes Recentes	4.0	918	1.0	*Cluster* 2
Clientes Intermitentes	5.0	933	1.0	*Cluster* 3
Clientes Fieis	7.0	695	1.0	*Cluster* 4
Clientes Moderados	6.0	899	1.0	*Cluster* 5
		5 886	**0.76**	

Nota: a variável RFM score varia entre 0 e 12

Os *clusters* obtidos (tabela 6) foram avaliados quanto à sua homogeneidade interna e distinção entre si, segmentando um total de 5 886 clientes e atingindo um coeficiente de silhueta de 0,76, dado que valores acima de 0,7 geralmente sugerem uma boa separação entre os *clusters* e uma coesão interna adequada, cumprindo a 1ª validação do segundo objetivo estipulado, sobre a identificação dos segmentos (Seção 1.3). Tanto no *Cluster* 0 quanto no *cluster* 1, o coeficiente de silhueta de 0,42 sugere uma coesão relativamente baixa, indicando possíveis sobreposições com outros *clusters*. Os *clusters* 2, 3, 4 e 5 apresentam um coeficiente de silhueta de 1,0 sugere uma excelente segmentação, com clara distinção de outros grupos.

Tabela 7: Caraterísticas do clusters a nível das variáveis RFM (Média e Valor mais Comum)
Fonte: Elaboração Própria

		Recência (0-4)	Monetério (0-4)	Frequência (0-4)	RFM Score (0-12)
***Cluster* 0**	Média	3.8	2.9	1.8	8.6
	Valor Comum	4	3	2	
***Cluster* 4**	Média	3.5	2.4	0.9	7.0
	Valor Comum	4	2	1	
***Cluster* 5**	Média	3.2	2.2	0.6	6.0
	Valor Comum	4	2	0	
***Cluster* 3**	Média	2.6	1.9	0.4	5.0
	Valor Comum	3	2	0	
***Cluster* 2**	Média	2.1	1.6	0.2	4.0
	Valor Comum	2	1	0	
***Cluster* 1**	Média	1.1	1.0	0.1	2.2
	Valor Comum	1	0	0	
Médias Totais		**2.57**	**1.95**	**0.65**	**5.2**

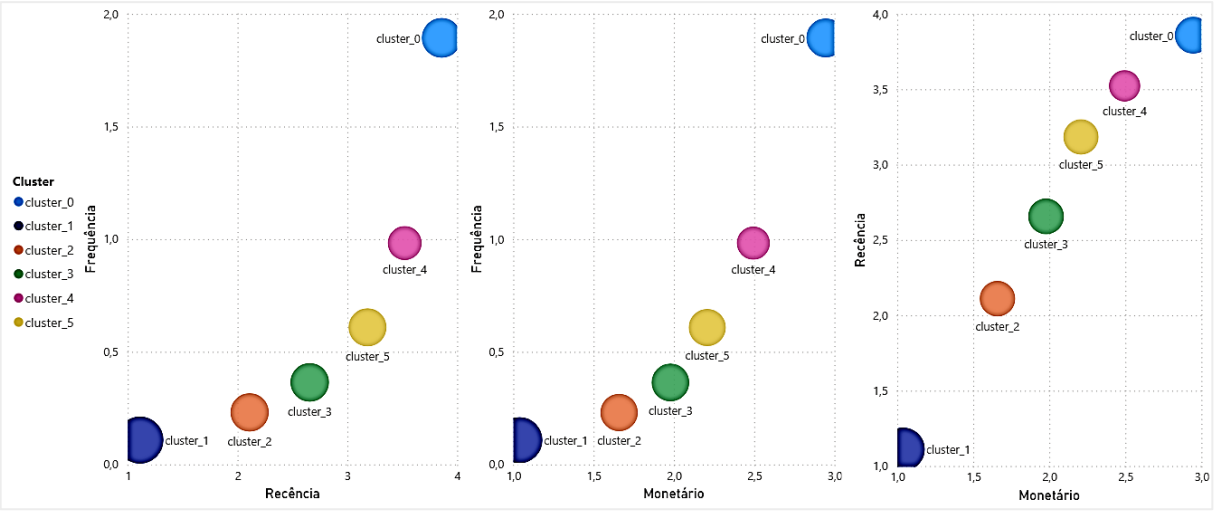

Figura 23: Gráficos de dispersão dos clusters finais pelas variáveis RFM
Fonte: Elaboração Própria

Caraterização dos segmentos de clientes de acordo com o resultados apurados presentes nas tabelas 6 e 7, e figura 23:

Os clientes do *cluster* 0 (16,7% do total), representado pela cor azul clara (figura 23) denominados “Clientes Premium” (Score 8.6), são considerados o segmento mais importante para a farmácia, representam grande valor para as mesmas, uma vez que compram muitas vezes, gastam muito e compraram recentemente, caracterizados por apresentarem os resultados mais altos nas três variáveis: frequência (1.8), recência (3.8) e monetário (2.9). O *cluster* 0 destaca-se dos demais, principalmente pela análise de frequência/recência apresentada no gráfico de dispersão da Figura 23, evidenciando uma frequência significativamente superior.

Em contraste, os clientes do *cluster* 1 (24,5% do total, segmento de maior dimensão), representado pela cor azul escura (figura 23), denominados “Clientes Passivos” (Score 2.2), realizam poucas compras, gastam pouco e não compram com frequência, todos as médias das variáveis encontram-se todas bastante abaixo da média (frequência 0.1, recência 1.0 e monetário 1.1), portam um valor insignificativo para a farmácia. No entanto, por representarem uma porção significativa de clientes, são aconselháveis ações para aumentar a sua reativação. Embora o *cluster* 1 apresente uma frequência semelhante à do *cluster* 2 (figura 23), destaca-se negativamente por possuir médias inferiores nas variáveis monetário e recência.

O *cluster* 2 corresponde a 15,5% do total dos clientes, representado pela cor laranja (figura 23), denominados “Clientes Recentes” (Score 4.0) é caraterizado por gastar pouco em comparação com a maior parte dos segmentos e ter uma frequência de compras baixa, indicando que têm potencial para serem desenvolvidos através de técnicas de retenção. O

cluster 2, embora apresente melhores resultados nas variáveis Monetário e Frequência em comparação ao *cluster* 1, diferencia-se por exibir uma recência significativamente superior, como evidenciado na Figura 23 (análise Recência/Monetário).

Os clientes do *cluster* 3 “Clientes Intermitentes” correspondem a 15,8% do total, representado pela cor verde (figura 23), têm uma atividade moderada, encontram-se a baixo da média do score (com 5,0) fazem compras de valor médio, a recência é ligeiramente mais alta que a média, mas a frequência de compras precisa de mais subsistência, posto isto campanhas para incentivar ao retorno são valiosas para este tipo de cliente. Clientes que também necessitam deste tipo de estratégias são os clientes do *cluster* 5 “Clientes Moderados” (Score 6.0 com 15,3% dos clientes totais), representado pela cor amarela (figura 23), uma vez que apesar da recência (igual a 3,2) e do valor gasto se situarem acima da média (igual a 2,2), a frequência encontra-se ligeiramente abaixo, com média de 0,6.

Por fim, o *cluster* 4 “Clientes Fieis” engloba 11,8% dos clientes com um score de 7,0, representado pela cor rosa (figura 23), caraterizado por apresentar todas as variáveis RFM acima da média, contudo não o suficiente para ser considerado “Cliente Premium”, principalmente devido à frequência relativamente mais baixa (0.9), estratégias para inventivar ao retorno também podem ser aplicadas para melhorar a média da frequência, e possivelmente converter estes clientes em “Clientes Premium”. O cluster 4, apresenta maior valor para a farmácias, do que os segmentos “Cliente Moderados” e “Clientes Intermitentes” principalmente devido à sua média de frequência, evidenciando o facto de ser denonimado “Clientes fieis”, uma vez que frequentam mais usualmente as farmácias do que os dois segmentos anteriormente descritos.

Finalizando esta análise, verifica-se que o objetivo O.3 (secção 1.3) também se encontra validado, uma vez que os segmentos criados foram caraterizados e foram definidos perfis claros para cada grupo, por forma a ajudar as farmácias a conhecer melhor os seus tipos de clientes.

RFM_Cluster *String*	Cluster *String*	Frequency *Number (double)*	Row Percent *Number (double)*	Total Row Count *Number (double)*
Clientes Moderados	Mulheres Prudentes	617	61.824	998
Clientes Premium	Mulheres Cuidadosas	579	40.125	1,443
Clientes Passivos	Mulheres Cuidadosas	363	39.542	918
Clientes Recentes	Mulheres Cuidadosas	339	36.334	933
Clientes Intermitentes	Mulheres Cuidadosas	317	35.261	899
Clientes Fieis	Mulheres Prudentes	290	41.727	695

Figura 24: Cross table entre os clusters RFM e "Demográfica-Produto"
Fonte: Elaboração própria (em KNIME)

Face aos resultados apresentados através da segmentação Demográfica-Produto e RFM é elaborado o cruzamento entre os *clusters* formados de cada um, presente na figura 24, e são revelados *insights* importantes para estratégias de marketing. Analisando os resultados é possível verificar que o único *cluster* RFM que tem uma forte compatibilidade com um c*luster* da segmentação Demográfica-Produto, é o segmento “Clientes Moderados” que 62% da sua constituição pertence ao segmento “Mulheres Prudentes”, através desta combinação podem ser efetuadas campanhas ainda mais direcionadas, sendo que combina as caraterísticas demográficas, relacionadas com o produto e comportamentais. Por outro lado, não será conveniente tomar a mesma consideração sobre as restantes combinações, uma vez que as percentagens são relativamente baixas, demonstrando uma fraca compatibilidade (35% a 42%).

5. Conclusões e Recomendações

5.1. Sumário da investigação

A segmentação de clientes de farmácias comunitárias ainda é um tema pouco desenvolvido, sendo a maior parte dos estudos direcionado para as *E-Pharmacy.* Este estudo permite obter uma perspetiva de como a segmentação de clientes ajuda a melhorar a personalização de estratégias de marketing das farmácias comunitárias.

A revisão da literatura permitiu alargar o conhecimento tanto a nível da área farmacêutica, como sobre a segmentação de clientes, e verificar lacunas existentes na literatura que conjuga as duas vertentes, foram utilizadas informações de cerca de 50 fontes durante a elaboração da revisão da literatura, incluindo artigos, jornais e capítulos de livros.

A investigação é conduzida através da metodologia CRISP-DM percorrendo todas as fases exceto a implementação, e ao longo destas fases são cumpridos os objetivos da investigação, na fase da preparação dos dados foi possível validar o primeiro objetivo, com recursos a análises descritivas é feita a caraterização dos clientes das farmácias, após garantir a qualidade dos dados para a modelação dos dados. Posteriormente, para validar o segundo objetivo, que consiste na identificação dos segmentos, foram utilizados vários algoritmos para a segmentação Demográfico-Produto para obtenção dos melhores resultados, porém o *Hierarchical Clustering* permitiu identificar seis segmentos com melhor precisão para o negócio, visto que o obteve um coeficiente silhueta de 0,65, e os segmentos criados portam valor para que sejam elaboradas estratégias de marketing direcionadas, o mesmo acontece para a análise RFM, que identificou seis segmentos, através da utilização do algoritmo *K-means,* e o resultado obtido no final do processo para o coeficiente silhueta é de 0,76 o que indica uma formação de *clusters* de alta qualidade.

Para concluir a validação, o terceiro objetivo analítico definido é validado através da análise das caraterísticas dos segmentos, com recurso a análises descritivas, por forma a traçar um perfil que permita facilitar o processo de criação de estratégias direcionadas para os clientes. Durante estas análises é elaborado um cruzamento entre os segmentos provenientes da análise Demográfico-Produto e análise RFM, o que permitiu identificar pelo menos um grupo com forte compatibilidade, cerca de 62% dos “Clientes Moderados” são “Mulheres Prudentes”, o que permite definir estratégias ainda mais personalizadas.

Ao realizar este estudo foi possível responder à questão de investigação inicialmente proposta: “Como a segmentação de clientes pode ajudar a melhorar a personalização de estratégias de marketing das farmácias tradicionais?”, a segmentação de clientes elaborada

no estudo permitiu segmentar os clientes de acordo com as suas características demográficas e relacionadas com o produto (originados 6 *clusters*) e ainda através do comportamento de compra (originados 6 *clusters*). Através de análises descritivas, foi possível traçar perfis de clientes que, por sua vez, permite às farmácias definir estratégias de marketing direcionadas para cada segmento.

5.2. Limitações

Durante o estudo foram detetadas algumas limitações, como é o caso de os dados serem apenas relativos apenas à farmácia com o maior volume de vendas, os resultados não são aplicáveis a todas as farmácias, seria necessário filtrar por cada farmácia e ajustar os modelos para obter os melhores resultados.

Outra limitação recai sobre a dificuldade dos algoritmos de *clustering* utilizados conseguirem lidar com um grande número de variáveis, o problema da dimensionalidade, o que reduziu bastante a quantidade de informação para analisar que, porém, seria relevante para ajudar à definição de campanhas e ofertas mais direcionadas.

Existem ainda outras limitações, que nesta investigação não foram abordadas, como, por exemplo, as promoções e ofertas que as farmácias aplicam que influência diretamente na decisão de compra do consumidor. Outro fator que impacta essa escolha, é a sazonalidade da procura, visto que alguns produtos têm maior procura em determinados períodos, como medicamentos para gripes e constipações de inverno, ou protetores solares no verão.

Por fim, uma das limitações encontradas foi o elevado tempo necessário para a execução do *Hierarchical Clustering*, esse tempo prolongado restringe a capacidade de realizar múltiplas iterações ou ajustes nos parâmetros do algoritmo de forma eficiente. Além disso, em contextos práticos, essa demora não é viável para empresas que precisem de atualizações frequentes nos perfis dos clientes, especialmente em cenários em que os dados são dinâmicos e as decisões precisam ser tomadas de forma ágil.

5.3 Contribuições

A investigação oferece uma contribuição significativa para ajudar as farmácias a definir estratégias de marketing mais eficazes. Através da aplicação de modelos avançados de segmentação de clientes, as farmácias conseguem identificar com maior precisão os vários perfis de clientes, o que facilita a adoção de ações personalizadas e orientadas para cada grupo. Este nível de detalhe na análise permite otimizar os recursos aplicados, ajudando

também a aumentar a competitividade e a satisfação dos clientes, enquanto melhora a eficiência operacional e a fidelização dos clientes às farmácias.

Além disso, a metodologia CRISP-DM é amplamente aplicável a outros setores além das farmácias. Os modelos desenvolvidos, que utilizam dados facilmente recolhidos em diversos ambientes empresariais, podem ser adaptados a qualquer negócio que tenha a necessidade de segmentar clientes. Isso torna o estudo relevante não apenas para o setor farmacêutico, mas também para outras áreas que possam beneficiar de uma melhor compreensão dos comportamentos e preferências dos consumidores.

Por fim, este trabalho também oferece um contributo relevante para a comunidade científica e a prática empresarial, será disponibilizado o workflow no repositório de exemplos práticos da comunidade KNIME, com um caso de aplicação detalhado da segmentação de clientes, desde a preparação dos dados até à avaliação dos resultados. A junção das áreas de farmácia e segmentação de clientes é ainda pouco explorada na literatura, o que reforça a importância desta investigação tanto a nível científico quanto para os profissionais dessas áreas.

5.4 Recomendações e Pistas Futuras

Como próximos passos, é essencial expandir a aplicação dos modelos de segmentação de clientes desenvolvidos. Inicialmente, a segmentação foi realizada em apenas uma farmácia para simplificar o processo de construção dos modelos. No entanto, para garantir uma análise mais abrangente e resultados mais robustos, será necessário aplicar esses modelos em todas as farmácias com dados disponibilizados, possibilitando uma visão mais completa e personalizada do comportamento dos clientes.

Além disso, é importante explorar outros tipos de segmentação, incluindo diferentes algoritmos, como a utilização de algoritmos de Redes Neuronais. A aplicação desses métodos permitirá avaliar se os resultados obtidos são mais vantajosos, comparados aos modelos atuais. A implementação de abordagens baseadas em *deep learning*, também pode representar um avanço significativo na segmentação de clientes, providenciando *insights* mais detalhados.

Outro passo relevante é incluir características dos produtos na segmentação, que ajudará na possível elaboração de um sistema de recomendação. Essa integração ajuda a oferecer produtos mais apropriados de acordo com cada grupo de clientes, aprimorando a experiência de compra e aumentando a satisfação.

O desenvolvimento de uma segmentação baseada no *Customer Lifetime Value* também será uma opção viável, uma vez que os dados necessários estão incluídos na base de dados disponibilizada. Essa abordagem permite que os clientes sejam agrupados de acordo com o valor que portam para a empresa, ajudando a facilitar a identificação de estratégias de marketing de diferentes níveis (Groeger *et al*. 2015).

Por fim, uma vez que a metodologia seguiu o ciclo até a fase de "Avaliação", onde foram verificados a qualidade dos modelos e os *insights* obtidos, será crucial avançar para a fase de "Implementação". Este passo envolve a aplicação direta dos modelos em um ambiente de negócio, ou o desenvolvimento de uma solução prática baseada nos resultados, de forma a transformar os *insights* em ações concretas e mensuráveis.

Referências Bibliográficas

Afzal, A., Khan, L., Hussain, M. Z., Hasan, M. Z., Mustafa, M., Khalid, A., Awan, R., Ashraf, F., Khan, Z., & Javaid, A. (2024). Customer Segmentation Using Hierarchical Clustering. *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, 1–6. https://doi.org/10.1109/I2CT61223.2024.10543349

Al-Arifi, M. N. (2012). Patients' perception, views and satisfaction with pharmacists' role as health care provider in community pharmacy setting at Riyadh, Saudi Arabia. *Saudi Pharmaceutical Journal*, *20*(4), 323–330. https://doi.org/https://doi.org/10.1016/j.jsps.2012.05.007

Alexandre, C., Bocsanean, P., Mangana, J., Santos, J., Monteiro, D., & Santos, P. (2017). Marketing behaviors analysis in a mobile wallet solution using data mining. *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*, 88–92. https://doi.org/10.1109/CICN.2017.8319362

Al-Mhairat, M., Alabbadi, R., Shaban, R., & Alqudah, A. (2019). *Performance Evaluation of clustering Algorthims* [University of Jordan]. https://www.researchgate.net/publication/334971445_Performance_Evaluation_of_clustering_Algorthims

Barapatre, D., & A, V. (2017). Data preparation on large datasets for data science. *Asian Journal of Pharmaceutical and Clinical Research*, *10*, 485. https://doi.org/10.22159/ajpcr.2017.v10s1.20526

Bar-Joseph, Z., Gifford, D., & Jaakkola, T. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, *17*(suppl_1), S22–S29. https://doi.org/10.1093/bioinformatics/17.suppl_1.S22

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., & Wiswedel, B. (2009). KNIME: The Konstanz Information Miner. Em C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 319–326). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_38

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., & Varoquaux, G. (2013). *API design for machine learning software: experiences from the scikit-learn project*. https://arxiv.org/abs/1309.0238

Burghardt, K. J., Howlett, B. H., Khoury, A. S., Fern, S. M., & Burghardt, P. R. (2020). Three Commonly Utilized Scholarly Databases and a Social Network Site Provide Different, But Related, Metrics of Pharmacy Faculty Publication. *Publications*, *8*(2). https://doi.org/10.3390/publications8020018

Camilleri, M. A. (2018). Market Segmentation, Targeting and Positioning. Em M. A. Camilleri (Ed.), *Travel Marketing, Tourism Economics and the Airline Product: An Introduction to Theory and Practice* (pp. 69–83). Springer International Publishing. https://doi.org/10.1007/978-3-319-49849-2_4

Castaldo, S., Grosso, M., Mallarini, E., & Rindone, M. (2016). The missing path to gain customers loyalty in pharmacy retail: The role of the store in developing satisfaction and trust. *Research in Social and Administrative Pharmacy*, *12*(5), 699–712. https://doi.org/10.1016/j.sapharm.2015.10.001

Chapman, P., Clinton, J., Kerben, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0: Step-by-step data mining guide*. https://api.semanticscholar.org/CorpusID:59777418

Chaudhry, M., Shafi, I., Mahnoor, M., Vargas, D. L. R., Thompson, E. B., & Ashraf, I. (2023). A Systematic Literature Review on Identifying Patterns Using Unsupervised Clustering Algorithms: A Data Mining Perspective. *Symmetry*, *15*(9). https://doi.org/10.3390/sym15091679

Chen, D., Sain, S., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, *19*, 197–208. https://doi.org/10.1057/dbm.2012.17

Chen, Y., Zhang, G., Hu, D., & Wang, S. (2006). Customer Segmentation in Customer Relationship Management Based on Data Mining. Em K. Wang, G. L. Kovacs, M. Wozny, & M. Fang (Eds.), *Knowledge Enterprise: Intelligent Strategies in Product Design, Manufacturing, and Management* (pp. 288–293). Springer US. https://doi.org/10.1007/0-387-34403-9_40

Cheung, Y., & Jia, H. (2013). Categorical-and-numerical-attribute data clustering based on a unified similarity metric without knowing cluster number. *Pattern Recognition*, *46*(8), 2228–2238. https://doi.org/https://doi.org/10.1016/j.patcog.2013.01.027

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2021). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University -*

Computer and Information Sciences, *33*(10), 1251–1257. https://doi.org/https://doi.org/10.1016/j.jksuci.2018.09.004

Cooil, B., Aksoy, L., & Keiningham, T. L. (2008). Approaches to Customer Segmentation. *Journal of Relationship Marketing*, *6*(3–4), 9–39. https://doi.org/10.1300/J366v06n03_02

Coombes, C. E., Liu, X., Abrams, Z. B., Coombes, K. R., & Brock, G. (2021). Simulation-derived best practices for clustering clinical data. *Journal of Biomedical Informatics*, *118*, 103788. https://doi.org/https://doi.org/10.1016/j.jbi.2021.103788

Davidson, I. (2002). *Understanding K-means non-hierarchical clustering*. https://www.researchgate.net/publication/228574607_Understanding_K-means_non-hierarchical_clustering

Decreto-Lei n.º 176/2006, D.R. I Série. 176/2006 (30-06-06) 6297-6383.

Delgado, H., Anguera, X., Fredouille, C., & Serrano, J. (2015, Outubro). Novel Clustering Selection Criterion for Fast Binary Key Speaker Diarization. *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*. https://doi.org/10.13140/RG.2.1.3073.2886

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Ernawati, E., Baharin, S. S. K., & Kasmin, F. (2021). A review of data mining methods in RFM-based customer segmentation. *Journal of Physics: Conference Series*, *1869*(1). https://doi.org/10.1088/1742-6596/1869/1/012085

Fantonelli, M. dos S., Zanotto, W. L., de Melo, F. M. F., Celuppi, I. C., Lacerda, T. C., de Oliveira, F. M., Hammes, J. F., Cunha, C. L., Felisberto, M., Santos, R. A. dos, Scandolara, D., da Rosa, J. S., de Oliveira, J. M. D., Demarchi, I. G., Wazlawick, R. S., & Dalmarco, E. M. (2023). Organization and management of sensitive personal health data in electronic systems in countries with implemented data protection laws, lessons to Brazil: A brief systematic review. *Computer Law & Security Review*, *51*. https://doi.org/10.1016/j.clsr.2023.105872

FasterCapital. (2024, Junho 16). *OTC Drugs: Understanding the Basics and Benefits*. https://fastercapital.com/content/OTC-Drugs--Understanding-the-Basics-and-Benefits.html

Fyson, H. (2024, Maio 27). *Why choose an open source tool for data science*. KNIME. https://www.knime.com/blog/why-choose-open-source-tool-data-science

Gray Nicola. (2011). THE EVOLUTION OF ONLINE PHARMACIES. *The journal of consumer-led health (SelfCare)*, 76–86. https://selfcarejournal.com/wp-content/uploads/2015/09/Gray-2.376-86.pdf

Groeger, L., & Buttle, F. (2015). Customer Lifetime Value. Em *Wiley Encyclopedia of Management* (pp. 1–3). https://doi.org/https://doi.org/10.1002/9781118785317.weom090070

Gustriansyah, R., Suhandi, N., & Antony, F. (2019). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*(1), 470 – 477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Hayasaka, S. (2024, Maio 6). *Cluster analysis: What it is, types & how to apply the technique without code*. https://www.knime.com/blog/what-is-clustering-how-does-it-work

Hicham, N., & Karim, S. (2022). Analysis of Unsupervised Machine Learning Techniques for an Efficient Customer Segmentation using Clustering Ensemble and Spectral Clustering. *International Journal of Advanced Computer Science and Applications*, *13*, 122–130. https://doi.org/10.14569/IJACSA.2022.0131016

Hopf, T., & Nikolaidou, A. (2023). Customer Valuation. Em R. Cadili & F. Ordenes (Eds.), *Meet Your Customer* (pp. 147–155). KNIME Press. https://www.knime.com/

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv.*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jeena, S., Chaudhary, A., & Thakur, A. (2023). *Implementation & Analysis of Online Retail Dataset Using Clustering Algorithms*. 1–6. https://doi.org/10.1109/ICIEM59379.2023.10166552

Kevrekidis, D. P., Minarikova, D., Markos, A., Malovecka, I., & Minarik, P. (2018). Community pharmacy customer segmentation based on factors influencing their selection of pharmacy and over-the-counter medicines. *Saudi Pharmaceutical Journal*, *26*(1), 33–43. https://doi.org/https://doi.org/10.1016/j.jsps.2017.11.002

Lafuente, R. (2018). *centraldedados / codigos_postais*. GitHub. https://github.com/centraldedados/codigos_postais

Ma, M., Liang, M., & Ji, Y. (2023). Comparison and Evaluation of Clustering Algorithms. *2023 International Conference on Computers, Information Processing and Advanced Education (CIPAE)*, 213–219. https://doi.org/10.1109/CIPAE60493.2023.00047

Massaro, A. (2022). Advanced Control Systems in Industry 5.0 Enabling Process Mining. *Sensors*, *22*(22), 8677. https://doi.org/10.3390/s22228677

Nascimento, R., Fagundes, R., & Júnior, G. (2018). Mineração de Dados Educacionais: Um Estudo Sobre Indicadores da Educação em Bases de Dados do INEP. *RENOTE*, *16*. https://doi.org/10.22456/1679-1916.85989

Papakyriakou, D., & Barbounakis, I. (2022). Data Mining Methods: A Review. *International Journal of Computer Applications*, *183*, 5–19. https://doi.org/10.5120/ijca2022921884

Patak, M., Lostakova, H., Curdova, M., & Vlckova, V. (2014). The E-Pharmacy Customer Segmentation Based on the Perceived Importance of the Retention Support Tools. *Procedia - Social and Behavioral Sciences*, *150*, 552–562. https://doi.org/10.1016/j.sbspro.2014.09.075

Patankar, N., Dixit, S., Bhamare, A., Darpel, A., & Raina, R. (2021). Customer Segmentation Using Machine Learning. Em *Recent Trends in Intensive Computing* (Vol. 39, pp. 239–244). https://doi.org/10.3233/APC210200

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications*, *9*(1). https://doi.org/10.3390/publications9010012

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/https://doi.org/10.1016/0377-0427(87)90125-7

Saltz, J. S. (2021). CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps. *2021 IEEE International Conference on Big Data (Big Data)*, 2337–2344. https://doi.org/10.1109/BigData52589.2021.9671634

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Silva, J. (2022). *Segmentação de clientes B2B e previsão estratégica de oportunidades futuras com Inteligência Artificial* [Dissertação de Mestrado, Universidade do Minho, Escola de Engenharia]. https://hdl.handle.net/1822/82125

Sinaga, K. P., & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Smith, W. (2015). *Parque Natural Municipal Corredores de Biodiversidade: pesquisas e perspectivas futuras*. https://www.researchgate.net/publication/313852594_Parque_Natural_Municipal_Corredores_de_Biodiversidade_pesquisas_e_perspectivas_futuras

Smith, W. R. (1956). Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Marketing*, *21*, 3–8. https://api.semanticscholar.org/CorpusID:49060196

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/https://doi.org/10.1016/j.jbusres.2019.07.039

Steinbach, M., Ertöz, L., & Kumar, V. (2003). The Challenges of Clustering High Dimensional Data. *Univ. Minnesota Supercomp. Inst. Res. Rep.*, *213*. https://doi.org/10.1007/978-3-662-08968-2_16

Tabianan, K., Velu, S., & Ravi, V. (2022). K-Means Clustering Approach for Intelligent Customer Segmentation Using Customer Purchase Behavior Data. *Sustainability*, *14*(12). https://doi.org/10.3390/su14127243

Tomar, D. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio - Science and Bio - Technology*, *5*, 241–266. https://doi.org/10.14257/ijbsbt.2013.5.5.25

Umargono, E., Suseno, J., & Gunawan, S. K. (2020). K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula. *The 2nd International Seminar on Science and Technology (ISSTEC 2019)*, 121–129. https://doi.org/10.2991/assehr.k.201010.019

Wijayanti, S., Ihalauw, J. J. O. I., Abdi, A. S., & Kusuma, L. (2024). A Mini-Theory of Hidden Marketing Strategy as Competitive Advantage. *Quality - Access to Success*, *25*(200), 59 – 68. https://doi.org/10.47750/QAS/25.200.07

Wu, J., Shi, L., Lin, W.-P., Tsai, S.-B., Li, Y., Yang, L., & Xu, G. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering*, *2020*, 1–7. https://doi.org/10.1155/2020/8884227

Anexos

Anexo A - Workflow preparação de dados em KNIME Analytics Platform

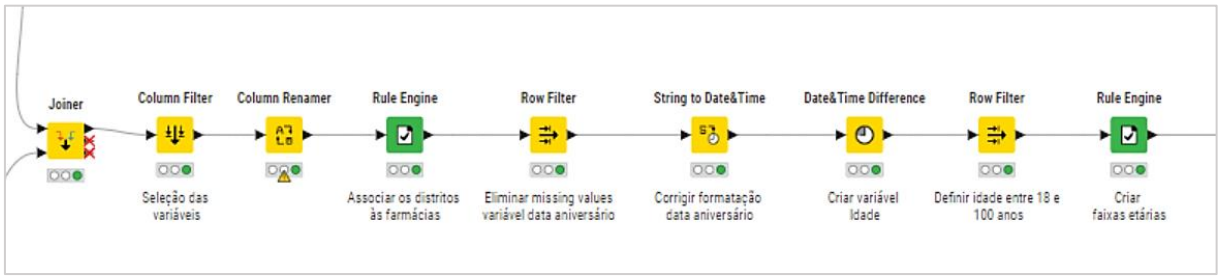

Figura A1: 1º Fase da preparação de dados (1/2) (workflow)
Fonte: Elaboração própria (em KNIME)

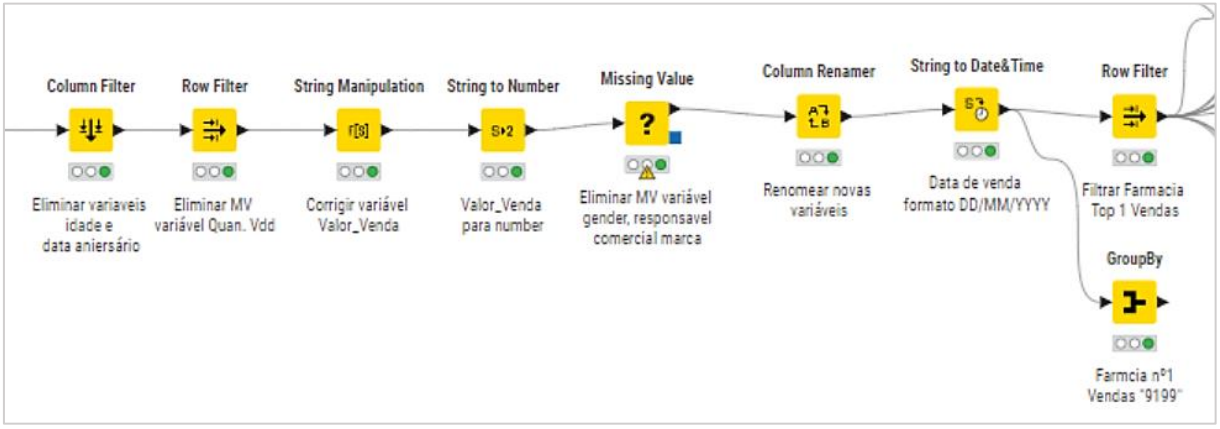

Figura A2: 1º Fase da preparação de dados (2/2) (workflow)
Fonte: Elaboração própria (em KNIME)

Anexo B - Elbow Method em KNIME

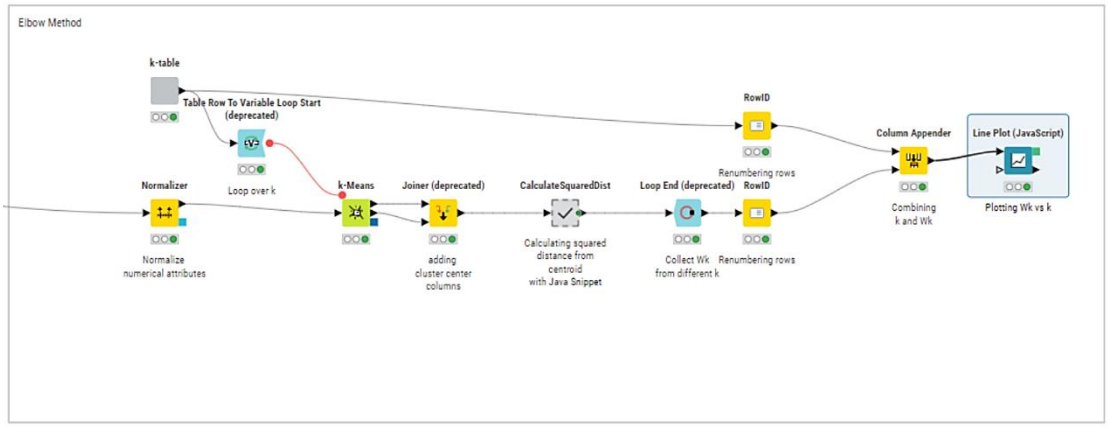

Figura B1: Workflow em KNIME Elbow Method

Fonte: (Berthold *et al.*, 2009)

Anexo C – Microsoft Power BI & KNIME Analytics Platform

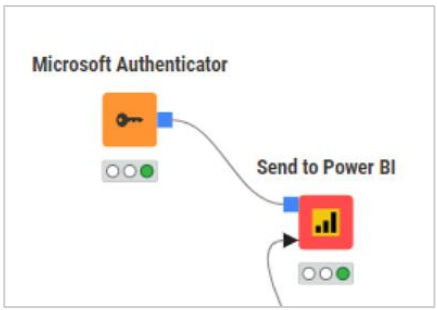

Figura C1: Workflow em KNIME Elbow Method

Fonte: Elaboração própria (em KNIME)

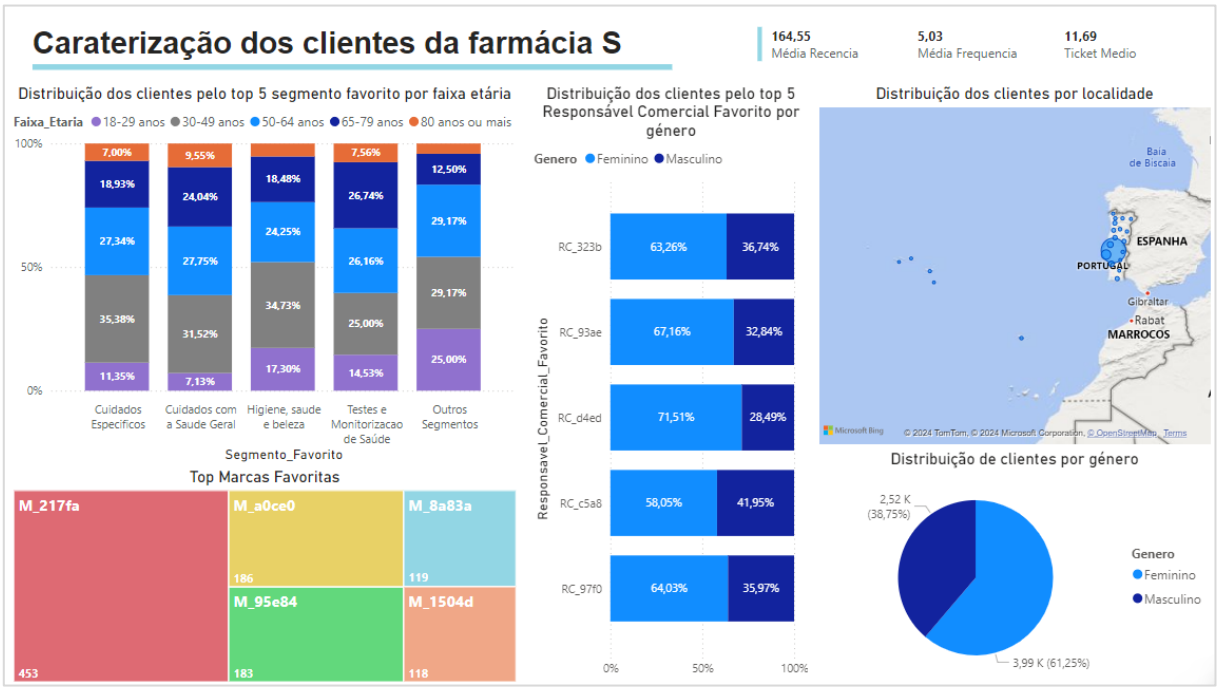

Figura C2: Workflow em KNIME Elbow Method

Fonte: Elaboração própria (em Power BI)

Anexo D – Aplicação dos algoritmos e resultados segmentação “Demográfico-produto”

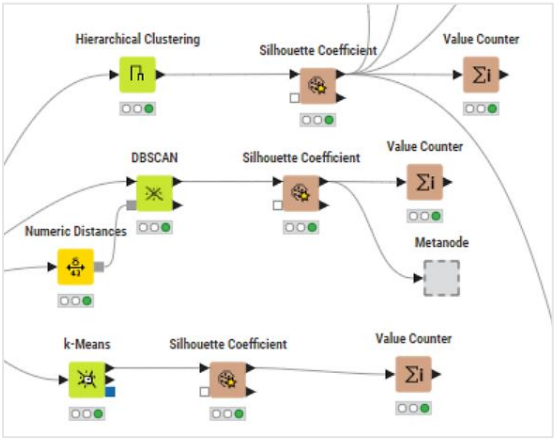

Figura D1: Workflow aplicação dos algoritmos segmentação “Demográfico-Produto”

Fonte: Elaboração própria (em KNIME)

#	RowID	Mean Silhouette Coefficient *Number (double)*
1	Cluster_7	0.717
2	Cluster_5	0.726
3	Cluster_6	0.72
4	Cluster_9	0.718
5	Cluster_3	0.687
6	Cluster_8	0.701
7	Cluster_2	0.704
8	Cluster_4	0.705
9	Cluster_0	0.661
10	Cluster_1	0.776
11	Overall	0.717

Figura D2: Resultados do coeficiente silhueta sobre o algoritmo DBSCAN

Fonte: Elaboração própria (em KNIME)

RowID	Mean Silhouette Coefficient *Number (double)*
cluster_3	0.397
cluster_1	0.75
cluster_0	0.769
cluster_2	0.627
cluster_4	0.704
cluster_5	0.473
Overall	0.634

Figura D3: Resultados do coeficiente silhueta sobre o algoritmo K-means

Fonte: Elaboração própria (em KNIME)

Anexo E – Resultados análise RFM

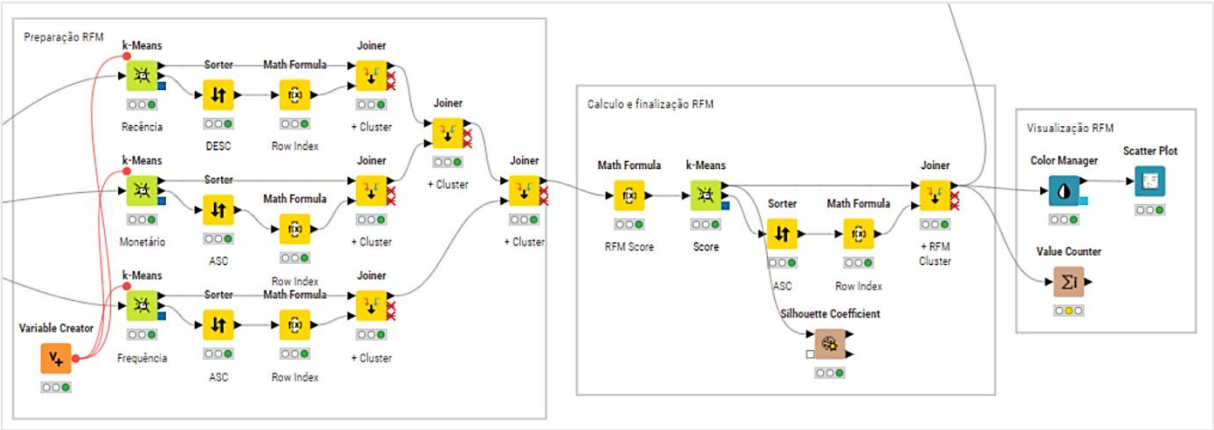

Figura E1: Workflow aplicação dos algoritmos análise RFM

Fonte: Elaboração própria (em KNIME)

ID_Cliente String	Recencia Number (lo...	Monetario Number (do...	Frequencia Number (inte...	Recencia_Cluster Number (integer)	Monetario_Cluster Number (integer)	Frequencia_Cluster Number (integer)
91990000000001...	66	8.5	1	3	1	0
91990000000008...	32	11.502	19	4	2	2
91990000000013...	146	6.56	1	2	1	0
91990000000023...	110	8.384	4	3	1	1
91990000000026...	19	15.802	26	4	4	3
91990000000030...	66	12.02	17	3	3	2
91990000000039...	25	9.998	13	4	2	2
91990000000040...	27	11.3	1	4	2	0
91990000000042...	73	12.689	62	3	3	4
91990000000052...	5	7.994	26	4	1	3

Figura E2: Amostra dos resultados da aplicação do algoritmo K-means a cada variável RFM

Fonte: Elaboração própria (em KNIME)

ID_Cliente String	Recencia Number (lo...	Monetario Number (do...	Frequencia Number (inte...	Recencia_Cluster Number (integer)	Monetario_Cluster Number (integer)	Frequencia_Cluster Number (integer)	RFM Score Number (dou...	RFM_Cluster Number (integer)
91990000000001...	66	8.5	1	3	1	0	4	1
91990000000008...	32	11.502	19	4	2	2	8	5
91990000000013...	146	6.56	1	2	1	0	3	0
91990000000023...	110	8.384	4	3	1	1	5	2
91990000000026...	19	15.802	26	4	4	3	11	5
91990000000030...	66	12.02	17	3	3	2	8	5
91990000000039...	25	9.998	13	4	2	2	8	5
91990000000040...	27	11.3	1	4	2	0	6	3
91990000000042...	73	12.689	62	3	3	4	10	5

Figura E3: Amostra dos resultados da aplicação do algoritmo K-means à variável RFM_Score e classificação atribuída (variável RFM_Cluster)

Fonte: Elaboração própria (em KNIME)