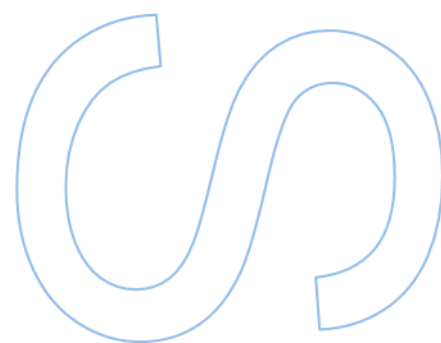
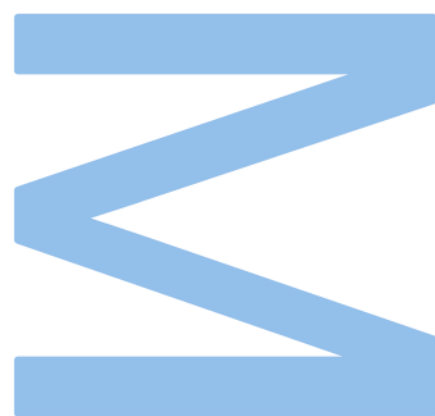




Análise de Modelos e Estratégias de E-commerce



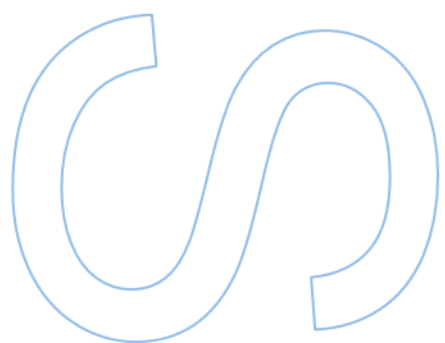
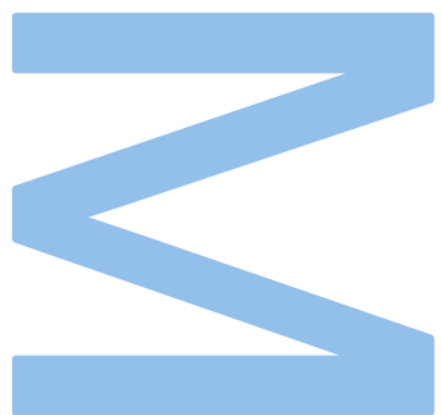
Paulo Daniel de Oliveira Ferreira

Mestrado em Estatística Computacional e Análise de Dados
Matemática
2022/2023

Orientador

Óscar António Louro Felgueiras, Professor Auxiliar, FCUP

idivia.



Para a Cláudia, que sempre me compreendeu e apoiou neste desafio.

Agradecimentos

Em primeiro lugar, quero agradecer à Searchmatters, Lda., em particular ao Vítor Ferreira, pela oportunidade de realizar este estágio, pelas sugestões e observações que me ajudaram ao longo destes meses de trabalho e também pelos essenciais dados de compras que serviram como base de trabalho.

Quero agradecer também ao professor Óscar Felgueiras, que tanto me apoiou e motivou a seguir em frente e me ajudou a ver uma oportunidade em cada dificuldade encontrada. A sua ajuda na parte teórica, na análise dos dados e na sua apresentação foi crucial.

Apesar de não terem cooperado diretamente comigo neste estágio, alguns dos meus colegas foram essenciais ao longo deste mestrado. Muito do que aprendi e que me ajudou neste estágio se deveu também ao trabalho desenvolvido com eles. Assim, aqui deixo uma sincera palavra de agradecimento aos meus colegas André, Cláudia, Daniel, Francisca, Henrique, Mafalda, Mariana, Tiago e Yulia.

Mas a maior palavra de agradecimento é para a minha família: a minha esposa Cláudia e os meus filhos Lucas e Simão Pedro. Sem o vosso apoio e compreensão para uma menor disponibilidade da minha parte, nada disto seria possível. Amo-vos menos do que merecem e mais do que consigo mostrar na maior parte das vezes.

Resumo

O *e-commerce* é um modelo de negócio que tem apresentado um significativo crescimento nos últimos anos, apresentando também grandes desafios relacionados com a concorrência e com a logística.

Neste relatório, pretende-se apresentar algumas conclusões resultantes da análise dos dados de vendas de uma jovem empresa de *e-commerce* na área de cosmética, no sentido de apresentar propostas de ação com vista ao seu crescimento. Nesta análise foram utilizadas ferramentas de *Data Mining* como Análise de Clusters e Regras de Associação.

Palavras-chave: *E-commerce*, *Data Mining*, Análise de Clusters, Regras de Associação

Abstract

E-commerce is a business model that has shown significant growth in recent years, also presenting great challenges related to competition and logistics.

In this report, it is intended to present some conclusions resulting from the analysis of the sales data of a young *e-commerce* company in the cosmetics area, in order to present proposals for action with a view to its growth. In this analysis Data Mining tools such as Cluster Analysis and Association Rules were used.

Keywords: E-commerce, Data Mining, Cluster Analysis, Association Rules

Índice

Lista de Tabelas	vi
Lista de Gráficos.....	vii
Lista de Figuras	viii
1. Introdução	1
1.1. Estrutura do relatório	1
2. O modelo de negócio de e-commerce	2
2.1. Origens do e-commerce	2
2.2. O presente do e-commerce.....	2
2.3. Que futuro para o e-commerce?	3
2.4. Objetivos de empresas de e-commerce.....	3
2.5. Vantagens do e-commerce	4
2.6. Desvantagens e desafios do e-commerce	5
2.7. O caso particular da empresa IDIVIA	7
3. Suporte teórico às ferramentas utilizadas	8
3.1. Análise de Clusters	8
3.1.1. Introdução	8
3.1.2. Clustering de K-Médias	14
3.1.3. Clustering Hierárquico	17
3.1.4. Clustering DBSCAN.....	19
3.1.5. Conclusão	26
3.2. Regras de Associação.....	27
3.2.1. Introdução e noções gerais	27
3.2.2. Medidas de associação	29
3.2.3. Processo de determinação de regras de associação	33
3.2.4. Algoritmo A-Priori.....	34
3.2.5. Algoritmo <i>FP-Growth</i>	37

4. Tratamento, exploração, apresentação e análise dos dados	41
4.1. Descrição dos dados.....	41
4.2. Limpeza e tratamento dos dados	42
4.3. Representação Visual dos Dados	45
4.4. Análise de Clusters	49
4.4.1. Clustering Hierárquico	49
4.4.2. Clustering DBSCAN.....	53
4.4.3. Conclusão	55
4.5. Regras de Associação.....	56
5. Conclusão	60
Referências Bibliográficas	62

Lista de Tabelas

Tabela 3.1. – Exemplo de transações e produtos em cada transação	28
Tabela 3.2. – Tabela de frequências para cada item.....	34
Tabela 3.3. – Tabela de frequências dos conjuntos com dois itens.....	35
Tabela 3.4. – Tabela de frequências dos conjuntos com três itens	35
Tabela 3.5. – Medidas de associação para as regras de associação.....	36
Tabela 3.6. – Tabela de frequências para cada item.....	37
Tabela 3.7. – Transações reprocessadas e ordenadas.....	38
Tabela 3.8. – Determinação dos conjuntos de itens frequentes	40
Tabela 4.1. – Variáveis de Product.csv e de Orders.xlsx e sua descrição.....	41
Tabela 4.2. – Variáveis de Orders with Brand and Category, sua descrição e origem	44
Tabela 4.3. – Características dos clusters obtidos por Clustering Hierárquico	53
Tabela 4.4. – Características dos clusters obtidos por Clustering DBSCAN	55

Lista de Gráficos

Gráfico 4.1. – Evolução diária (e médias a 7 e 30 dias) do volume de vendas.....	45
Gráfico 4.2. – Evolução mensal do volume de vendas.....	46
Gráfico 4.3. – Volume de negócios por país	46
Gráfico 4.4. – Volume de negócios por continente.....	47
Gráfico 4.5. – Volume de negócios por dia da semana.....	47
Gráfico 4.6. – Dendrograma (ligação simples)	49
Gráfico 4.7. – Dendrograma (ligação média)	50
Gráfico 4.8. – Dendrograma (ligação completa).....	50
Gráfico 4.9. – Dendrograma (ligação completa, 6 clusters).....	51
Gráfico 4.10. – Tamanho dos 6 clusters obtidos por Clustering Hierárquico	51
Gráfico 4.11. – Diagrama de dispersão (Regular Price - Quantity).....	52
Gráfico 4.12. – Clusters e Makeup (Y/N).....	52
Gráfico 4.13. – Distância ao k-vizinho mais próximo para minPts=15	53
Gráfico 4.14. – Tamanho dos 4 clusters obtidos por Clustering DBSCAN.....	54
Gráfico 4.15. – Distribuição dos continentes por cluster.....	54
Gráfico 4.16. – Distribuição de Health (Y/N) por cluster	54
Gráfico 4.17. – Distribuição de Men (Y/N) por cluster	54
Gráfico 4.18. – Distribuição de Oral (Y/N) por cluster.....	55

Lista de Figuras

Figura 3.1. – Jogo de tabuleiro Quarto.....	8
Figura 3.2 – Três clusters bem separados (pontos bidimensionais).....	9
Figura 3.3. – Quatro clusters baseados no centro (pontos bidimensionais)	9
Figura 3.4. – Oito clusters contíguos (pontos bidimensionais)	9
Figura 3.5 – Seis clusters densos (pontos bidimensionais)	10
Figura 3.6. – Interpretação geométrica das métricas de Minkowski para $n = 2$	12
Figura 3.7. – Sucessivas etapas do Clustering K-Médias	16
Figura 3.8. – Exemplo de um dendrograma	17
Figura 3.9. – Ponto central, ponto de fronteira e ponto de ruído	21
Figura 3.10. – Pontos centrais, ponto de fronteira e pontos de ruído	22
Figura 3.11. – Exemplo de clustering DBSCAN com minPts=3.	22
Figura 3.12. – Clustering DBSCAN (clusters com diferentes tamanhos e formatos) ...	23
Figura 3.13. – Exemplo de clustering DBSCAN com um cluster completamente rodeado por outro	23
Figura 3.14. – Exemplo de clustering DBSCAN com regiões de diferente densidade..	24
Figura 3.15. – <i>FP-Tree</i> resultante da aplicação do algoritmo <i>FP-Growth</i>	38
Figura 4.1. – Visão parcial do ficheiro Orders_Orange_Product.csv importado para o Orange	56
Figura 4.2. – Conjuntos de itens frequentes para suporte mínimo 0,5%	56
Figura 4.3. – Alguns dos conjuntos de itens frequentes com pelo menos 1 elemento para suporte mínimo 0,1%.....	57
Figura 4.4. – Alguns dos conjuntos de itens frequentes com pelo menos 2 elementos para suporte mínimo 0,1%.....	57
Figura 4.5. – Alguns dos conjuntos de itens frequentes com pelo menos 3 elementos para suporte mínimo 0,1%.....	57
Figura 4.6. – Alguns dos conjuntos de itens frequentes com pelo menos 4 elementos para suporte mínimo 0,1%.....	57
Figura 4.7. – Conjuntos de itens frequentes com pelo menos 5 elementos para suporte mínimo 0,1%.....	58
Figura 4.8. – Algumas das regras de associação para suporte mínimo 0,1% e confiança mínima de 80%.....	58
Figura 4.9. – Algumas das regras de associação (marcas) para suporte mínimo 0,5% e confiança mínima de 50%.....	59

1. Introdução

Neste relatório pretendo apresentar o trabalho desenvolvido no Estágio Curricular em Contexto Empresarial realizado no âmbito do Mestrado em Estatística Computacional e Análise de Dados. O estágio foi realizado remotamente na empresa Searchmatters, Lda., com a supervisão de Vítor Ferreira (Searchmatters, Lda.) e orientação de Óscar Felgueiras (FCUP).

Os objetivos do estágio passavam por conhecer o modelo de negócio de *e-commerce*, analisar dados de vendas de uma empresa de *e-commerce* na área de cosmética (IDIVIA) e propor estratégias de atuação com base nas análises realizadas.

As empresas estão cada vez mais interessadas em entender e tratar os seus clientes como indivíduos e não em massa, no sentido de lhes proporcionar uma melhor experiência enquanto clientes e aumentar as vendas. Com base nos dados das vendas, procurou-se encontrar padrões de compra e produtos que são usualmente comprados em conjunto.

Os dados analisados foram fornecidos pelo orientador e dizem respeito a compras online efetuadas entre outubro de 2021 e março de 2023. Foram utilizados diferentes programas na limpeza, preparação, apresentação e análise dos dados, nomeadamente Excel, Power BI, R Studio e Orange.

1.1. Estrutura do relatório

O relatório está organizado da seguinte forma:

O Capítulo 1 descreve o contexto em que o estágio foi realizado, bem como o objetivo e a abordagem utilizada. No Capítulo 2 é feita uma breve descrição do modelo de negócio de *e-commerce* em geral e da empresa IDIVIA em particular. O Capítulo 3 consiste no suporte teórico às ferramentas utilizadas na análise dos dados. No Capítulo 4 são expostos os passos dados no tratamento, exploração, apresentação e análise dos dados. Por fim, no Capítulo 5 são apresentadas as conclusões do trabalho, bem como algumas sugestões de trabalho futuro nesta área.

2. O modelo de negócio de *e-commerce*

2.1. Origens do *e-commerce*

Os primórdios do *e-commerce* remontam aos anos 90 do século passado, com a massificação do uso da internet. Uma das primeiras empresas de *e-commerce* foi a NetMarket, fundada em 1994, que comercializava livros, música e software. Em 1995 foi fundada a Amazon, que começou como uma livraria online, mas rapidamente se expandiu para um leque mais diversificado de produtos, sendo hoje um gigante tecnológico. Também por essa altura, surgiu o eBay, plataforma que intermedeia comércio entre utilizadores.

Com a crescente difusão e acessibilidade da internet, temos assistido ao crescimento e evolução do *e-commerce*. Na viragem do milénio, com o aparecimento de dispositivos móveis e a facilidade de acesso à internet móvel, tornou-se possível realizar compras em qualquer lugar e a qualquer hora, apenas com alguns toques no ecrã do telemóvel ou outro dispositivo móvel. Entrava-se numa nova era: o *m-commerce*. (Kumar, 2021)

2.2. O presente do *e-commerce*

Hoje, o *e-commerce* é um modelo de negócio em crescimento, que se vai transformando e inovando. É algo bem presente na vida de muitos consumidores e tornou-se num pilar fundamental da economia mundial, uma vez que as vendas online representam uma parcela considerável das vendas a retalho.

Com a concorrência das lojas online, muitas lojas físicas também se lançaram no *e-commerce* e adotaram novas tecnologias, no sentido de melhor atender os seus clientes e se manterem relevantes nesta realidade em rápida mudança.

Com a massificação do *m-commerce*, muitos consumidores fazem as suas compras usando dispositivos móveis, levando empresas de *e-commerce* a desenvolver versões dos sites adequados a dispositivos móveis, mas principalmente aplicações, no sentido de proporcionar a melhor experiência aos seus clientes.

Nos últimos anos, também as redes sociais têm registado um grande crescimento, pelo que empresas de *e-commerce* procuram tirar partido do seu alcance global para

impulsionar as suas vendas. Em particular, os *influencers* são parceiros que promovem marcas e empresas, atraindo potenciais clientes.

A inteligência artificial e *machine learning* têm desempenhado um papel fundamental no *e-commerce*, sendo usados de forma a personalizar a experiência de compra dos clientes, bem como na gestão e logística da cadeia de abastecimento.

Em resumo, o *e-commerce* caracteriza-se por uma transformação contínua, uma vez que as empresas procuram formas mais eficazes e automatizadas de alcançar e atender clientes nesta era digital, proporcionando-lhes uma boa experiência de compras e aumentando o seu lucro. (Kumar, 2021)

2.3. Que futuro para o *e-commerce*?

Tendo em conta a extrema relevância atual do *e-commerce* e a forma como outros negócios evoluíram ao longo dos tempos, é possível que nos próximos anos o *e-commerce* seja marcado por (Qayum, 2023):

- Maior difusão, tornando-se quase onnipresente, mesmo em áreas que hoje possam não nos ocorrer. Serão esses importantes nichos de mercado, que darão grandes oportunidades a quem neles apostar, enquanto outros operadores mais tradicionais no mercado poderão sofrer perdas por não inovar. (Apenas a título de exemplo, há apenas alguns anos não nos ocorreria que se pudesse encomendar comida ou pedir boleia através de uma aplicação. Mas alguém teve essas ideias, investiu nelas e lucrou milhões, enquanto outros que continuaram a fazer o mesmo de sempre tiveram perdas consideráveis). O *e-commerce* será, cada vez mais, um modelo de negócio no qual a inovação terá de ser uma constante, pois a estagnação poderá ditar perda de relevância, clientes e lucros.

- Crescimento através das redes sociais, de meios de pagamento inovadores, do uso da realidade virtual (de forma a proporcionar uma experiência mais realista), do recurso à inteligência artificial e a *machine learning*, bem como de modelos de subscrição (em que os clientes pagam uma assinatura, beneficiando de descontos e preços especiais).

2.4. Objetivos de empresas de *e-commerce*

Como qualquer outro negócio, o principal objetivo é **aumentar as vendas e os lucros**.

Aumentar o número de clientes. Para tal, são essenciais eficazes meios de publicidade, preços competitivos e uma plataforma de fácil utilização para o cliente.

Providenciar uma positiva e **agradável experiência de compra**, desde a facilidade de navegação na plataforma, ao seguimento da encomenda, até à receção do produto. (Por exemplo, é muito mais agradável para um cliente saber a data prevista da chegada da encomenda e poder recebê-la em mãos, em vez de ter de ir levá-la a um posto CTT. Empresas que proporcionam este tipo de serviços estarão mais perto de atrair e manter clientes.)

Manter-se competitivo. Para tal, é essencial inovar, não só em termos de oferta de produtos e serviços, mas também de meios tecnológicos e formas de alcançar clientes. Empresas que estagnam terão dificuldades em sobreviver num mercado tão competitivo.

Otimização de processos. Através da tecnologia e automatização de operações, será possível reduzir custos e tempos de espera, o que poderá aumentar o lucro e cativar mais clientes.

Potenciar o **reconhecimento da marca**, através da publicidade eficaz e contributo de *influencers*. Para tal, também será decisivo providenciar um produto ou serviço distinto (em termos de qualidade ou características) da concorrência. (Saaslyft, s.d.)

2.5. Vantagens do e-commerce

Conforto e acessibilidade. Atualmente, é possível realizar compras em qualquer lugar e a qualquer hora, sem necessidade de sair de casa. Mesmo pessoas com dificuldades em deslocar-se ou que não têm lojas físicas na sua área de residência conseguem, desta forma, adquirir os produtos ou serviços que pretendem. Em particular, devido à pandemia da CoViD-19, com muitos a evitar as lojas físicas (que estiveram fechadas por algum tempo), o e-commerce tornou-se a melhor alternativa, bem como a mais segura.

Preços competitivos. Com a diminuição de custos associados a transportes e intermediários, as empresas de *e-commerce* conseguem muitas vezes apresentar preços mais baixos do que as lojas físicas.

Comparação de preços. É possível comparar vários preços e características usando diferentes separadores de um computador ou mesmo um comparador de preços, sem necessidade de visitar várias lojas físicas.

Variedade. Negócios de *e-commerce* conseguem apresentar uma maior seleção de produtos (que podem existir em stock ou terem de ser encomendados a fornecedores), o que torna mais provável ao cliente encontrar o que pretende do que numa loja física.

Personalização. Usando inteligência artificial, cada cliente pode ter uma experiência personalizada de compra, com produtos e promoções recomendadas de acordo com o histórico de compras e de navegação.

Revisões de consumidores. Muitos vendedores permitem a avaliação e a revisão dos produtos para parte de clientes que adquiriram os produtos, permitindo aos potenciais compradores aferir a qualidade do produto e a satisfação

Alcance global. Atualmente, é possível vender produtos e serviços a clientes em qualquer lugar do mundo, trazendo vantagens para vendedores e clientes.

Eficiência. Com a automatização de processos na gestão do stock, atendimento e distribuição, a gestão do negócio torna-se mais eficiente e os custos podem ser reduzidos, aumentando o lucro. (Al-Lami & Alnoor, 2021)

2.6. Desvantagens e desafios do *e-commerce*

Impossibilidade de manuseamento. Em compras online, não é possível manusear, provar, cheirar ou experimentar os produtos, o que condiciona a experiência de compra e a determinação de ser o produto pretendido. (Quem nunca comprou um produto online e, mal o obteve, percebeu que não era bem o que pretendia?) Para minimizar esta situação, é importante os vendedores proporcionarem a maior informação possível dos produtos, como descrição detalhada das características do produto, imagens pormenorizadas (com vista 3D, se possível), excertos (por exemplo, no caso de livros), vídeos de apresentação e opiniões de compradores reais.

Falhas na segurança e privacidade. O uso de cartões de crédito ou outros meios de pagamento (PayPal, MBWay) podem trazer consigo a exposição a fraude. Também a disponibilização de dados confidenciais no momento da compra ou do registo no site pode resultar na divulgação indevida desses mesmos dados. Para ajudar os clientes a ultrapassar esse receio, é importante que as empresas invistam de forma a criar nos clientes uma sensação (e realidade) de segurança.

Dificuldades técnicas: Em caso de falhas no site ou na aplicação (devido a situações naturais ou a ataques informáticos), a experiência de compra pode ficar comprometida e resultar em perdas nas vendas. Uma forma de minimizar as consequências destes

imponderáveis é haver lojas físicas associadas às empresas de *e-commerce*, de modo que a faturação não seja zero no período em que o site ou aplicação se encontra indisponível. (É isso o que acontece com a empresa em estudo neste relatório e em muitos gigantes do retalho.)

Atrasos nas entregas. O envio e a entrega podem ser demorados, podendo ser necessário ao cliente esperar dias ou semanas para ter o produto encomendado e pago. Este pode ser um “quebra-negócios” (“dealbreaker”, ou seja, pode fazer a diferença entre a compra ser feita ou não), uma vez que afeta significativamente a experiência do cliente. Trata-se de um fator um pouco ingrato para a empresa, uma vez que está fora do seu controlo. Eventos regulares como greves ou condições meteorológicas, mas também situações raras como a recente invasão da Ucrânia ou o bloqueio no Canal do Suez, podem adiar em dias ou semanas os prazos previstos de entrega das encomendas.

Dificuldades com trocas e devoluções: Trocar ou devolver produtos adquiridos online pode ser bem mais complicado do que fazê-lo diretamente em loja física, pois pode implicar demoras e custos adicionais. Este condicionamento, aliado à impossibilidade de manusear o produto, pode levar o potencial cliente a repensar a sua compra. Uma forma de minimizar esta dificuldade é haver uma rede de espaços físicos associados à loja online que permitam ao cliente realizar trocas ou devoluções.

Dependência da tecnologia: O *e-commerce* depende muito da tecnologia e as empresas devem ter um site e um sistema de processamento de pagamentos de confiança para ter sucesso.

Falta de interação humana: Num mundo cada vez mais impessoal, o *e-commerce* dificulta (ou mesmo impossibilita) a interação humana, que normalmente permite o esclarecimento de dúvidas ou a apresentação de sugestões por parte de um funcionário de loja. A incorporação de inteligência artificial (chat robots, por exemplo) ou de sistemas de recomendação (o clássico “Quem comprou ... também comprou”) podem ajudar a mitigar a falta de contacto humano, mas nada substitui esta interação. Por essa razão, empresas que possam conciliar a vertente de *e-commerce* com a de loja física têm a possibilidade de alcançar um leque mais vasto de clientes.

Concorrência: O *e-commerce* tem-se tornado cada vez mais competitivo, obrigando os comerciantes a inovar de forma a manter-se relevantes e continuar a ter sucesso.

Apesar destes desafios, o *e-commerce* continua a crescer e evoluir, e as empresas que são capazes de superar obstáculos e dificuldades irão colher os benefícios da venda de produtos e serviços online. (Al-Lami & Alnoor, 2021)

2.7. O caso particular da empresa IDIVIA

A IDIVIA, uma marca da Searchmatters, Lda., é uma plataforma de *e-commerce* dedicada a produtos de cosmética e cuidados do corpo. Faz envios (com rastreamento) para todo o mundo, com portes grátis a partir de um determinado valor, e aceita meios de pagamento diversificados e na moeda do país do cliente (com raras exceções). Apresenta preços competitivos para mais de 6 mil produtos 100% originais e, apesar do pouco tempo de existência (outubro de 2021), tem registado um grande crescimento no volume de encomendas e de faturação.

A IDIVIA comercializa produtos não perecíveis e com grandes prazos de validade (ou mesmo sem prazo de validade), o que diminui os riscos associados ao tempo (de espera entre a encomenda e a entrega) e facilita a angariação de clientes em zonas remotas, como Ásia, América e Austrália.

A plataforma proporciona uma agradável experiência de navegabilidade, com diferentes línguas ao dispor, com os diferentes produtos divididos por categorias e facilidade de realização da compra. Dispõe também uma transparente política de reembolso e devoluções. Além disso, permite a subscrição de uma newsletter, de forma a manter a ligação aos potenciais clientes e informá-los de negócios vantajosos e promoções.

Quanto à divulgação da plataforma, para além da presença em algumas das principais redes sociais (Facebook, Twitter e Instagram), também tem publicidade paga através do Google Ads.

Quando é feita uma encomenda em *idivia.com*, verifica-se se os produtos têm stock em armazém. No caso de não haver disponibilidade do produto, é verificado se existe em stock numa das lojas físicas, nomeadamente farmácias, pertencentes ao grupo detentor da plataforma (que vendem alguns dos produtos que fazem parte do catálogo). Apenas em último caso é feita uma encomenda a um fornecedor. Assim que a encomenda tem todos os produtos disponíveis, é enviada para o destinatário.

3. Suporte teórico às ferramentas utilizadas

De forma a proceder à análise dos dados fornecidos, serão aplicados alguns algoritmos e procedimentos de *Data Mining*, nomeadamente Análise de Clusters e Regras de Associação, que a seguir se descrevem.

3.1. Análise de Clusters

3.1.1. Introdução

A análise de clusters é uma técnica fundamental em *Machine Learning* não supervisionado, que permite a descoberta de estruturas, padrões e relacionamentos ocultos nos dados. São várias as aplicações, como *Data Mining*, análise de imagens, segmentação de clientes ou deteção de erros.

O objetivo da análise de clusters é encontrar uma estrutura de clusters (ou agrupamentos) de forma que os objetos pertencentes a cada cluster partilhem alguma característica ou propriedade relevante para o domínio do problema em estudo, ou seja, são de alguma forma similares.

Por exemplo, as peças do jogo de tabuleiro Quarto podem ser agrupadas em dois clusters de quatro formas diferentes (altas/baixas, claras/escuras, circulares/quadradas, maciças/ocas).¹ Também poderiam ser consideradas combinações destas características (por exemplo, 4 clusters: altas e claras, altas e escuras, baixas e claras, baixas e escuras).



Figura 3.1 – Jogo de tabuleiro Quarto

¹ Apenas a título de curiosidade, o objetivo deste jogo é criar um cluster alinhado, ou seja, obter 4 peças em linha que partilhem uma mesma característica.

Intuitivamente, um cluster pode ser visto como um conjunto de objetos próximos ou que satisfazem alguma relação especial. No entanto, não existe uma definição única, formal e precisa deste conceito. Dependendo da área de estudo e dos objetivos do estudo, são vários as definições e os métodos para a determinação dos clusters.

Algumas definições de cluster (Barbara, 2000)

- **Cluster bem separado:** um cluster é um conjunto de pontos tal que qualquer ponto num cluster está mais próximo (ou mais semelhante) a todos os outros pontos do cluster do que a qualquer outro ponto fora do cluster.



Figura 3.2 – Três clusters bem separados (pontos bidimensionais).

- **Cluster baseado em centro:** um cluster é um conjunto de pontos tal que um ponto num cluster está mais próximo (ou mais semelhante) do “centro” de um cluster do que do centro de qualquer outro cluster. O centro de um cluster é frequentemente um centróide (a média de todos os pontos do cluster), ou um medióide, o ponto mais “representativo” de um cluster. Exemplo: Clustering de K-Médias.



Figura 3.3. – Quatro clusters baseados no centro (pontos bidimensionais).

- **Cluster contíguo ou encadeado (vizinho mais próximo ou clustering transitivo):** um cluster é um conjunto de pontos tal que um ponto num cluster está mais próximo (ou mais semelhante) a um ou mais outros pontos no cluster do que a qualquer ponto fora do cluster. Exemplo: Clustering Hierárquico.



Figura 3.4. – Oito clusters contíguos (pontos bidimensionais).

- **Definição baseada em densidade:** um cluster é uma região densa de pontos, que é separada de outras regiões de alta densidade por regiões de baixa densidade. Esta definição é mais utilizada quando os clusters são irregulares ou entrelaçados, e quando existem *outliers*. Exemplo: Clustering DBSCAN.

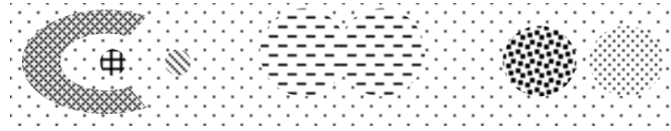


Figura 3.5. – Seis clusters densos (pontos bidimensionais).

- **Definição baseada em similaridade:** um cluster é um conjunto de pontos que são “semelhantes”, e objetos noutros clusters não são “semelhantes”.

Crítérios de clustering (J. Handl, 2005)

Cada definição de cluster resulta num critério de agrupamento, que é uma forma de selecionar uma estrutura para representar os clusters que melhor se ajuste a um determinado conjunto de dados. Os critérios de agrupamento podem ser agrupados em três categorias:

- **Compactação:** a compactação ou homogeneidade de um cluster é geralmente associada a uma pequena variação intracluster. Algoritmos que otimizam este tipo de critérios costumam ser bons a descobrir clusters esféricos e/ou bem separados, mas revelam-se pouco eficazes em estruturas mais complexas.

- **Encadeamento ou ligação:** trata-se de um conceito mais local, baseado na ideia de que objetos vizinhos devem pertencer ao mesmo cluster. É bastante apropriado para detetar clusters de formas arbitrárias, mas não funciona tão bem quando há pouca separação entre os clusters.

- **Separação espacial:** considera as distâncias entre os clusters, mas fornece pouca orientação durante o processo de agrupamento, podendo facilmente conduzir a soluções triviais.

Matriz de dados

Antes de prosseguir, convém introduzir a noção de matriz de dados. Para um conjunto de dados com n variáveis e m observações, é possível representar toda a informação contida nesse conjunto de dados através de uma matriz $m \times n$.

É usual proceder à transformação dos dados da matriz antes de os utilizar, uma vez que os dados estão frequentemente em diferentes escalas. Uma boa forma de o fazer é normalizar os dados, ou seja, para cada valor numérico).², subtrair a média da sua variável e dividi-lo pelo desvio padrão. Assim, para cada x_i^l (a l -ésima variável da i -ésima observação), o seu valor normalizado será

$$x_i'^l = \frac{x_i^l - \bar{x}^l}{s_{x^l}}$$

Proximidade

Um conceito fundamental na análise de clusters é a de proximidade entre objetos. Conceitos como “próximo” devem ser definidos de forma rigorosa para que se possam tirar conclusões adequadas. Para medir a proximidade, podem ser utilizadas medidas de similaridade (a mais comum é a correlação) ou medidas de dissimilaridade (a mais comum é a distância). No caso das medidas de similaridade (que variam entre 0 e 1), valores próximos de 1 indicam objetos próximos e valores próximos de 0 denotam objetos afastados. Já no caso das medidas de dissimilaridade, valores próximos de 0 denotam objetos próximos e valores elevados indicam objetos afastados. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2017)

Medidas de dissimilaridade

Qualquer medida de distância (que quantifica a dissimilaridade) deve satisfazer as seguintes propriedades para quaisquer três pontos $\mathbf{x}_i$, $\mathbf{x}_j$ e $\mathbf{x}_k$ de um conjunto de dados:³

- 1) $d(\mathbf{x}_i, \mathbf{x}_i) = 0$ (reflexividade)
- 2) $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$ (simetria)
- 3) $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$ (positividade)

Se também se verificaram as duas propriedades seguintes, a distância designa-se de métrica:

- 4) $d(\mathbf{x}_i, \mathbf{x}_j) = 0 \Leftrightarrow \mathbf{x}_i = \mathbf{x}_j$
- 5) $d(\mathbf{x}_i, \mathbf{x}_k) \leq d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_k)$ (desigualdade triangular)

² Para variáveis qualitativas, a abordagem será outra, como se verá à frente.

³ $\mathbf{x}_i = (x_i^1, x_i^2, \dots, x_i^n)$, onde n é a dimensão (ou número de variáveis) do conjunto de dados.

A métrica mais utilizada é a de Minkowski, que é definida (para diferentes valores de p , com $p \geq 1$) por:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[p]{\sum_{l=1}^n |x_i^l - x_j^l|^p}$$

As principais métricas de Minkowski são as seguintes:

- para $p = 1$, distância de Manhattan:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^n |x_i^l - x_j^l|$$

- para $p = 2$, distância de euclidiana (a mais utilizada em análise de clusters):

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^n |x_i^l - x_j^l|^2}$$

- para $p = \infty$, distância de Chebyshev (a diferença máxima entre atributos):

$$d(\mathbf{x}_i, \mathbf{x}_j) = \max_{1 \leq l \leq n} |x_i^l - x_j^l|$$

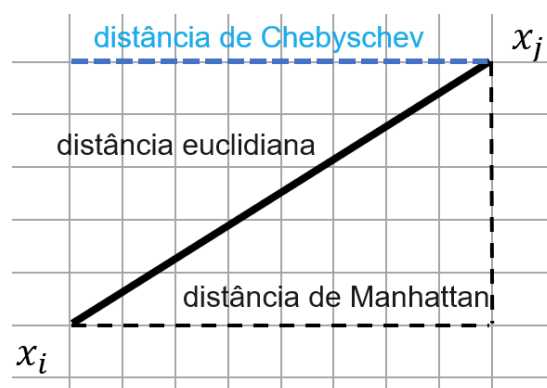


Figura 3.6. – Interpretação geométrica das métricas de Minkowski para $n = 2$.

Por vezes também é utilizada (em algoritmos hierárquicos) uma outra medida de distância, que consiste em ordenar todas as distâncias (euclidiana, por exemplo) entre pares de pontos por ordem crescente e atribuindo-lhes valores de 1 a $\frac{m(m-1)}{2}$. A distância entre um ponto e ele mesmo será 0.

Medidas de similaridade

Duas formas usadas para avaliar a similaridade entre pares de pontos são dadas pelo valor absoluto das medidas de separação angular e de correlação de Pearson.

A separação angular é dada pelo cosseno do ângulo entre dois vetores não nulos. Assim,

$$|\text{cosseno}(\mathbf{x}_i, \mathbf{x}_j)| = \left| \frac{\sum_{l=1}^n (x_i^l \times x_j^l)}{\sqrt{\sum_{l=1}^n (x_i^l)^2 \times \sum_{l=1}^n (x_j^l)^2}} \right|$$

Já a correlação de Pearson é dada pelo quociente entre a covariância e o produto dos desvios padrão. Assim, sendo $\bar{x}_i = \sum_{l=1}^n \frac{x_i^l}{n}$ e $\bar{x}_j = \sum_{l=1}^n \frac{x_j^l}{n}$,

$$\text{pearson}(\mathbf{x}_i, \mathbf{x}_j) = \frac{|\text{covariância}(\mathbf{x}_i, \mathbf{x}_j)|}{\sqrt{\text{variância}(\mathbf{x}_i) \times \text{variância}(\mathbf{x}_j)}} = \frac{|\sum_{l=1}^n (x_i^l - \bar{x}_i)(x_j^l - \bar{x}_j)|}{\sqrt{\sum_{l=1}^n (x_i^l - \bar{x}_i)^2 \times \sum_{l=1}^n (x_j^l - \bar{x}_j)^2}}$$

No caso de vetores binários (apenas zeros e uns), existem medidas de similaridade que avaliam a proximidade dos dois vetores. 0 significa que são totalmente diferentes e 1 significa que são exatamente iguais. Uma medida bastante usual é o coeficiente de similaridade simples, que consiste em dividir o número de casos coincidentes pelo número total de casos. Também se pode usar o Coeficiente de Jaccard, mas este não tem em conta os casos em que ambos os vetores assumem valor 0. Dependendo dos casos, pode ser mais conveniente uma ou outra medida.

Algo parecido pode ser feito em relação às variáveis qualitativas. Por exemplo, a distância de Hamming entre duas observações contabiliza o número de variáveis com valores diferentes. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2017)

Matriz de proximidade

A matriz de proximidade (que pode ser de similaridade ou de dissimilaridade) contém todas as similaridades ou dissimilaridades entre observações, sendo assim uma matriz $m \times m$, que será utilizada na etapa seguinte, central na análise de clusters.

Processo de clustering

São vários os algoritmos de clustering, que podem ser selecionados dependendo de vários critérios, como a natureza dos dados, o domínio específico, os objetivos, a

eficiência computacional, o número ou a forma dos clusters, facilidade de interpretação ou a forma como lidam com *outliers*.

Não existe um algoritmo de clustering que seja melhor do que todos os outros, uma vez que dependem de vários fatores. Por essa razão, é recomendável comparar e avaliar vários algoritmos para escolher o mais adequado para o que se pretende. É isso que será feito mais à frente.

Validação

O coeficiente de silhueta é uma métrica utilizada para avaliar a qualidade do clustering. Varia entre -1 e 1 , sendo que pontos com coeficiente de silhueta perto de 1 são considerados como bem atribuídos ao seu cluster, enquanto valores próximos de -1 sugerem uma má escolha de cluster ou que o ponto pode ser um *outlier*.

Para cada ponto de dados $\mathbf{x}_i$, é dado por $s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}$, onde $a(\mathbf{x}_i)$ é a distância média de $\mathbf{x}_i$ aos pontos do seu cluster e $b(\mathbf{x}_i)$ é a distância média de $\mathbf{x}_i$ aos pontos do cluster mais próximo e ao qual não pertence.

O coeficiente de silhueta médio (de todos os pontos no conjunto de dados) é frequentemente usado como uma medida global da qualidade do agrupamento. Este índice funciona bem com o clustering de k-médias e é usado para determinar o número ideal de clusters. (Rousseeuw, 1987)

Interpretação

A interpretação de clusters envolve entender e extrair ideias significativas dos grupos de pontos de dados que foram identificados como pertencentes ao mesmo cluster. Esta análise deve ter em conta o domínio específico e o contexto dos dados e permite identificar as características que contribuem de forma mais significativa para a separação ou formação de clusters.

3.1.2. Clustering de K-Médias

Introdução

K-Médias é um dos algoritmos de clustering mais utilizados e também dos mais simples. O seu objetivo é particionar um conjunto de dados em K clusters distintos, onde cada ponto de dados pertence ao cluster com a média mais próxima. O algoritmo minimiza iterativamente a variância dentro do cluster, procurando encontrar os centros de cluster ótimos. Encontrar estes centros é um problema computacionalmente difícil, mas através deste algoritmo é possível encontrar rapidamente ótimos locais.

Escolha do valor de K

Escolher o número adequado do valor de K é uma etapa crucial para o algoritmo. Algumas formas de o fazer são (Barbara, 2000):

- Método do cotovelo: representando, num gráfico, diferentes valores de K no eixo horizontal e a soma dos quadrados das distâncias no eixo vertical, o ponto de cotovelo, onde a soma dos quadrados das distâncias começa a estabilizar, costuma ser uma boa escolha para K.
- Calculando o coeficiente silhueta (que mede o quão bem os clusters estão separados), valores mais elevados correspondem a melhores valores de K.
- Usando técnicas de validação cruzada para validar o desempenho do algoritmo K-Médias para diferentes valores de K.

Posições iniciais dos K-centróides

São escolhidos K pontos no espaço do conjunto de dados, que se designam por K-centróides. As posições iniciais dos K-centróides podem impactar significativamente os resultados finais. Sendo escolhidos aleatoriamente, os resultados serão diferentes em cada execução do algoritmo. Existem métodos para escolher os K-centróides, como K-Médias++, que aumentam a probabilidade de encontrar uma boa solução. (Vassilvitskii, s.d.)

Processo Iterativo: Etapa inicial

Trata-se de um processo iterativo, onde os pontos de dados são atribuídos a clusters e os centróides são atualizados a cada iteração.

Cada ponto de dados é atribuído ao cluster com centroide mais próximo (usando uma métrica, como a distância euclidiana ou outra), formando K clusters. Assim, depois de percorridos todos os pontos de dados, obtém-se K clusters, com cada ponto de dados atribuído a um deles.

Processo Iterativo: Etapa de atualização

Assim que todos os pontos de dados tiverem sido atribuídos aos clusters, são determinados os novos centróides para cada cluster, que são calculados fazendo a média de todos os pontos de dados dentro de cada cluster.

Este processo é repetido iterativamente, podendo alguns pontos de dados mudar de cluster ao longo do processo. Este processo termina quando que determinadas

condições de convergência forem satisfeitas, ficando assim o conjunto de dados repartido pelos K clusters.

Verificação de convergência

Para verificar se o algoritmo convergiu, verifica-se se os houve alterações significativas nos centróides entre iterações ou se um número máximo predefinido de iterações foi alcançado. Por convergência do algoritmo entende-se que foi atingido um estado de estabilidade, no qual iterações adicionais não iriam alterar de forma significativa os centróides ou do cluster ou a que cada ponto de dados pertence.

Resultado final

Depois de verificada a convergência, os dados estão repartidos pelos K clusters como pretendido.

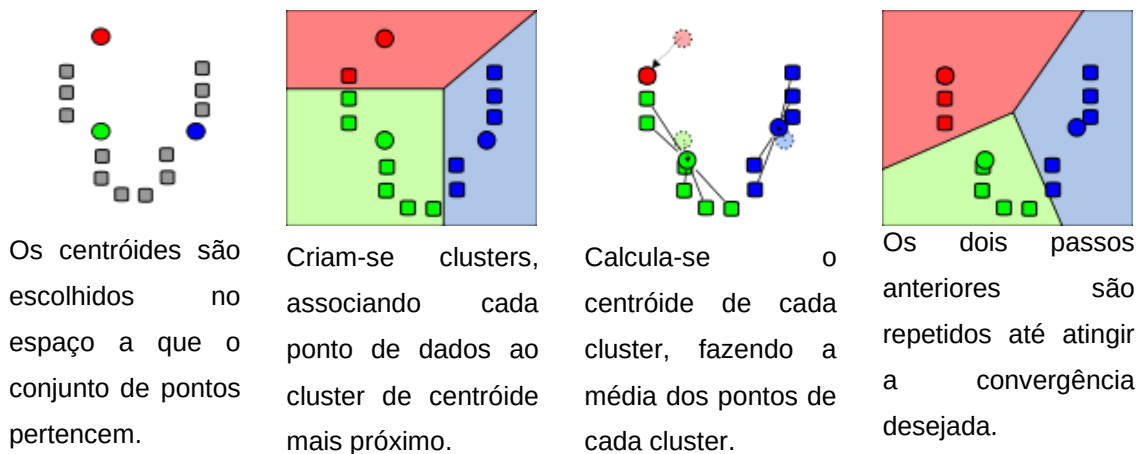


Figura 3.7. – Sucessivas etapas do Clustering K-Médias

Vantagens e limitações

Trata-se de um algoritmo popular, de simples compreensão e implementação. No entanto, a sua sensibilidade a *outliers* e a pressuposição de que os clusters são esféricos, torna desaconselhável o seu uso isoladamente. (Barbara, 2000)

Tendo em conta a natureza dos dados a analisar no Capítulo 4, decidiu-se não aplicar o clustering K-Médias ao conjunto de dados. O número limitado de variáveis numéricas e um grande número de variáveis dicotómicas e categóricas foram preponderantes para esta decisão. No entanto, uma vez que se trata de um dos mais populares algoritmos de clustering, optou-se pela inclusão de uma breve exposição teórica.

3.1.3. Clustering Hierárquico

Introdução

O Clustering Hierárquico é um método de análise de clusters que cria uma hierarquia de clusters, frequentemente representada por uma estrutura em forma de árvore à qual se dá o nome de dendrograma, que fornece uma representação visual das relações entre pontos de dados em diferentes níveis de detalhe. Ao contrário do K-Médias, não requer um número predefinido de clusters, pelo que é particularmente útil quando se pretende explorar a estrutura inerente aos dados sem especificar antecipadamente o número de clusters.

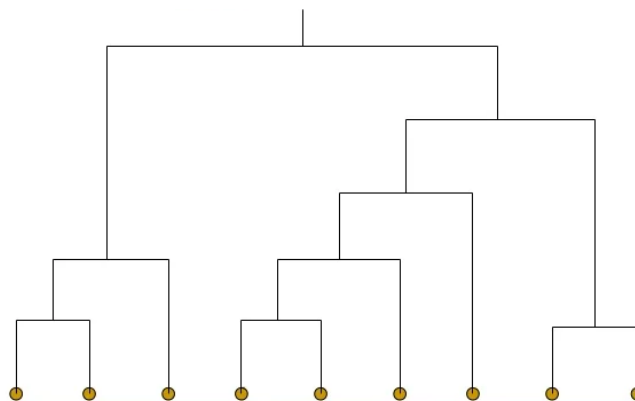


Figura 3.8. – Exemplo de um dendrograma

Existem duas abordagens principais de Clustering Hierárquico, que se descrevem a seguir: Aglomerativo (*bottom-up*) e Divisivo (*top-down*).

Clustering Hierárquico Aglomerativo (*bottom-up*)

Esta é a abordagem mais utilizada, na qual cada ponto de dados começa como um cluster separado e é sucessivamente agrupado com outros.

Início: O estado inicial do processo considera cada ponto de dados como o seu próprio cluster.

Similaridade entre pares: De seguida, calcula-se a semelhança ou dissimilaridade entre todos os pares de clusters. Para isso, usa-se uma métrica de distância escolhida, como distância euclidiana, distância de Manhattan ou coeficientes de correlação. Assim, mede-se o quão próximos ou distantes dois clusters estão.

Agrupar clusters mais próximos: identifique o par de clusters com a maior similaridade ou menor distância com base num dos métodos de ligação. Entre outros métodos de ligação, destacam-se os seguintes:

- Ligação mínima (ou simples): mede a distância entre os pontos de dados mais próximos em cada par de clusters.
- Ligação máxima (ou completa): mede a distância entre os pontos de dados mais distantes em cada par de clusters.
- Ligação média: mede a distância média entre todos os pontos de dados em cada par de clusters.
- Ligação centróide: mede a distância entre os centróides (pontos médios) de cada par de clusters.
- Ligação Ward: minimiza o aumento na variância intracluster.

Tendo determinado o par de clusters mais próximo, estes são agrupados num único cluster maior. Cada vez que isto é feito, o número de clusters diminui em um cluster.

Repetição do processo: Este processo é repetido, recalculando as semelhanças entre pares dos clusters restantes e agrupando o par de clusters mais próximo. Este processo iterativo continua até que haja apenas um cluster, que contém todos os pontos de dados.

Visualização do dendrograma: O resultado do agrupamento hierárquico pode ser representado como um dendrograma. No dendrograma, o eixo vertical representa a semelhança ou distância entre os clusters, e as linhas horizontais mostram onde os clusters são agrupados. A altura na qual dois clusters se fundem representa a sua semelhança ou distância.

Definição do número de clusters: Dependendo dos objetivos de análise e do dendrograma, o número de clusters pode ser definido cortando o dendrograma numa altura específica. Cortes mais abaixo produzem clusters mais refinados, enquanto cortes mais acima resultam em menos clusters maiores.

O clustering hierárquico oferece diversas vantagens, incluindo a capacidade de revelar relacionamentos hierárquicos dentro dos dados e a flexibilidade na seleção de clusters. Não requer a especificação de um número de clusters. No entanto, o custo computacional pode ser significativamente afetado pelo tamanho do conjunto de dados. Também não lida bem com *outliers* e ruído. Além disso, a ordem de entrada dos dados, a escolha da distância ou do método de ligação pode resultar em diferentes resultados de clustering. Portanto, é importante ter uma compreensão geral dos dados, bem como comparar os resultados obtidos por diferentes distâncias e métodos de ligação ou até mesmo com outros algoritmos de clustering. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2017)

Clustering Hierárquico Divisivo (*top-down*)

Esta é uma abordagem alternativa ao clustering hierárquico aglomerativo no qual se começa com todos os pontos de dados num único cluster e é sucessivamente dividido em subclusters menores.

Início: O estado inicial do processo considera um único cluster contendo todos os pontos do conjunto de dados.

Similaridade entre pares: Calcula-se a semelhança ou dissimilaridade entre todos os pontos de dados do cluster único.

Divisão de clusters: Identifica-se o par de pontos de dados dentro do cluster que são mais diferentes, ou seja, com a menor similaridade ou maior distância, com base num dos métodos de ligação. Este par será escolhido para dividir o cluster em dois subclusters menores.

Criação de subclusters: O cluster é dividido em dois subclusters, com cada ponto de dados a ser atribuído ao subcluster do ponto (do par determinado no passo anterior) mais próximo. Estes subclusters são clusters distintos dentro da hierarquia.

Repetição do processo: Este processo é repetido, recalculando as semelhanças entre pares de pontos dentro de cada cluster e dividindo-o. restantes e agrupando o par de clusters mais próximo. Este processo iterativo continua até que cada ponto de dados pertença ao seu próprio cluster individual.

Visualização do dendrograma e Definição do número de clusters:

Nestas duas etapas, o procedimento é o mesmo que é adotado no Clustering Hierárquico Aglomerativo.

Embora esta abordagem tenha algum interesse, é menos utilizada, uma vez que tende a ser mais dispendiosa do ponto de vista computacional, especialmente para conjuntos de dados maiores.

3.1.4. Clustering DBSCAN

DBSCAN⁴ é um algoritmo de clustering baseado em densidade proposto em 1996 por Martin Ester, Hans-Peter Kriegel, Jörg Sander e Xiaowei Xu. (Ester, Kriegel, Jörg, & Xu, 1996) Apesar de outros algoritmos baseados em densidade terem sido entretanto propostos (DENCLUE, por exemplo), trata-se de um algoritmo popular e bastante utilizado. Este algoritmo agrupa pontos de dados em grupos com base na sua

⁴ Density-Based Spatial Clustering of Applications with Noise. Em português, Clustering Espacial Baseado em Densidade de Aplicações com Ruído.

densidade espacial, definindo clusters como regiões densas separadas por regiões mais esparsas e também pode identificar *outliers* ou pontos de ruído que não pertencem a qualquer cluster. Não requer a especificação prévia do número de clusters e pode descobrir clusters com formas arbitrárias e não necessariamente convexas.

Algumas definições (Barbara, 2000)

Considere-se um conjunto de pontos de um espaço S no qual se pretende fazer clustering. Seja ε o raio de uma vizinhança em relação a algum ponto, d uma distância⁵ no espaço S e $minPts$ um número natural que indica o número mínimo de pontos que satisfazem uma determinada condição. ε e $minPts$ são os parâmetros que devem ser devidamente configurados para aplicar o algoritmo DBSCAN.

Para efeitos de clustering, os pontos de S são classificados como **pontos centrais**, **pontos de fronteira** ou **outliers**, da seguinte forma:

- A é um **ponto central** se pelo menos $minPts$ pontos (incluindo A) estiverem na vizinhança de A de raio ε . Estes pontos serão a base para a formação de clusters.
- Um ponto B é **diretamente acessível por densidade** a partir de A se o ponto B estiver na vizinhança de A de raio ε . Um ponto que seja diretamente acessível por densidade a partir de um ponto central diz-se **ponto diretamente acessível por densidade**. Um ponto diretamente acessível que não seja ponto central diz-se **ponto de fronteira**.
- B é um **ponto acessível por densidade** a partir de A se existir um caminho $A_1, A_2, \dots, A_n$ com $A_1 = A$ e $A_n = B$, onde cada A_{i+1} é diretamente acessível a partir de A_i (para $i = 1, 2, \dots, n - 1$). Assim, todos os pontos no caminho são pontos centrais, excetuando (eventualmente) B .
- Todos os pontos que não são diretamente acessíveis (ou seja, não são pontos centrais nem pontos de fronteira) dizem-se **outliers** ou **pontos de ruído**.

Na figura 3.9. estão representados pontos num espaço S bidimensional e três pontos A , B e C . Para $minPts = 7$, A é um ponto central, pois há 7 pontos (A incluído) na vizinhança de A de raio ε . B é um ponto fronteira pois há apenas 4 ($< minPts$) pontos na vizinhança de B de raio ε (e por isso não é ponto central), mas está na vizinhança de A de raio ε (sendo assim um ponto diretamente acessível). Já C é um ponto de

⁵ Pode ser definida como sendo a distância euclidiana, de Manhattan ou outra.

ruído, uma vez que não é ponto central (há apenas 3 pontos na vizinhança de B de raio ε) nem diretamente acessível (não está na vizinhança de raio ε de algum ponto central).

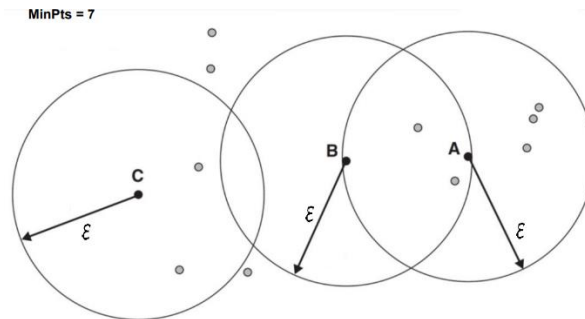


Figura 3.9. – Ponto central, ponto de fronteira e ponto de ruído.

Processo de clustering (Schubert, Sander, Martin, Kriegel, & Xu, 2017)

O algoritmo DBSCAN funciona da seguinte forma:

- Todos os pontos começam como não visitados, isto é, não foram considerados para o processo de clustering.
- Para cada ponto não visitado, é calculada a sua vizinhança, determinando se se trata ou não de um ponto central (se tem pelo menos $minPts$ na vizinhança de raio ε).
- Escolhendo um ponto central arbitrário, começa-se a criar um cluster, adicionando todos os pontos na vizinhança desse ponto central. Para cada ponto adicionado, é verificado se se trata de um ponto central. Se for um ponto central, a sua vizinhança também é adicionada ao cluster. (Caso contrário, trata-se de um ponto de fronteira.) Este processo continua até que não seja possível encontrar pontos acessíveis por densidade. Todos os pontos adicionados ao cluster são marcados como visitados, não sendo assim possível pertencerem a outro cluster.
- De seguida, o algoritmo seleciona outro ponto central não visitado e repete o processo para encontrar o cluster seguinte. Este processo continua até que todos os pontos centrais tenham sido visitados.
- Os pontos que não foram visitados, ou seja, aqueles que não são pontos centrais nem estão na vizinhança de pontos centrais, são considerados como pontos de ruído ou *outliers* e não pertencem a qualquer cluster.

A ordem pela qual os pontos centrais são escolhidos pode afetar a constituição dos clusters, mas apenas no que a pontos de fronteira diz respeito. Os pontos centrais

pertencerão sempre ao mesmo cluster, independentemente da ordem pela qual sejam selecionados.

Na figura 3.10., os pontos azuis são pontos centrais e o vermelho é ponto de fronteira. Todos estes pertencem a um mesmo cluster. Já os pontos verdes não pertencem a qualquer cluster, sendo classificados como pontos de ruído ou *outliers*.

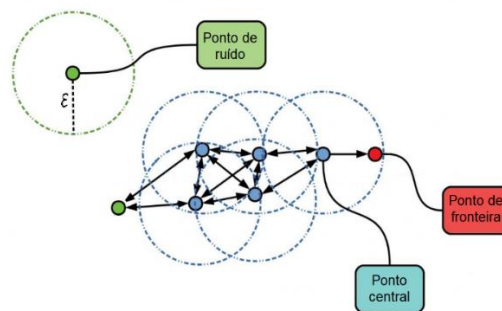


Figura 3.10. – Pontos centrais (azul), ponto de fronteira (vermelho) e pontos de ruído (verde), para $minPts = 2$.

A figura 3.11. representa o resultado de um processo de clustering DBSCAN, com $minPts = 3$, que resulta em dois clusters (roxo e verde) e dois pontos de ruído (preto).

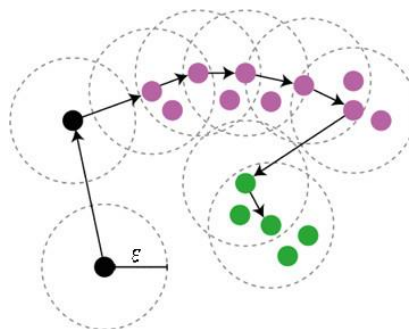


Figura 3.11. – Exemplo de clustering DBSCAN com $minPts = 3$. O processo resulta em 2 clusters e 2 pontos de ruído.

Complexidade do algoritmo

A complexidade do DBSCAN pode ser analisada em termos de tempo e de espaço.

A complexidade de tempo do DBSCAN é normalmente considerada como $O(n \log n)$, onde n é o número de pontos de dados. Essa complexidade resulta da necessidade de realizar consultas para determinar o número de pontos dentro da vizinhança de cada ponto de dados. Esta complexidade também depende da distância utilizada e da eficiência do cálculo de distâncias entre os pontos de dados.

A complexidade espacial é $O(n)$, onde n é o número de pontos de dados. Esta complexidade deve-se principalmente ao armazenamento de pontos de dados e

informações associadas, como estatuto de visitado, atribuição de cluster e informações de vizinhança.

No geral, as complexidades de tempo e espaço do DBSCAN são consideradas eficientes para muitos cenários práticos, tornando-o aplicável a grandes conjuntos de dados. (Ester, Kriegel, Jörg, & Xu, 1996)

Vantagens

- Não é necessário especificar antecipadamente o número de clusters (como acontece em outros algoritmos, como o k-médias). Isto permite a adaptação a conjuntos de dados com um número desconhecido ou variável de clusters.
- É um algoritmo bastante simples de entender concetualmente. Apesar de haver variantes que tratam de algumas das suas desvantagens, a sua simplicidade torna-o num algoritmo popular e bastante utilizado.
- Não pressupõe formatos nem tamanhos de clusters, consegue assim identificar clusters de formato arbitrário, incluindo formatos não lineares ou bastante irregulares. Pode até encontrar clusters completamente cercados por outros. (O parâmetro *minPts* reduz a possibilidade de ligação entre clusters por meio de uma fina linha de pontos.)

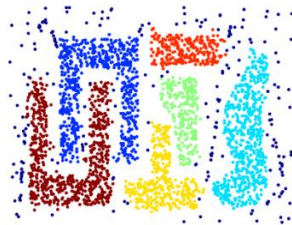


Figura 3.12. – Clustering DBSCAN com clusters com diferentes tamanhos e formatos. Os pontos de ruído estão assinalados a azul-escuro.

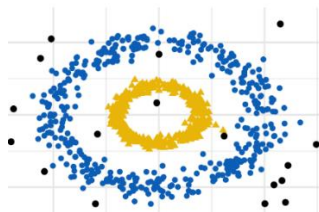


Figura 3.13. – Exemplo de clustering DBSCAN com um cluster completamente rodeado por outro. Os pontos de ruído estão assinalados a preto.

- Não é afetado por *outliers*, uma vez que os classifica como ruído. Consegue distinguir os pontos de ruído dos pontos centrais e mesmo dos pontos de fronteira.

- Apenas necessita da definição de dois parâmetros (ϵ e $minPts$), os quais podem ser ajustados tendo em conta a densidade, o tamanho e o número de clusters, entre outros fatores relevantes.
- É muito pouco sensível à ordem pela qual os pontos do conjunto de dados são considerados pelo algoritmo. Apenas os pontos de fronteira poderão variar entre clusters, dependendo da ordem pela qual os pontos centrais são selecionados.
- É computacionalmente eficiente na generalidade dos casos, podendo assim lidar com grandes conjuntos de dados de forma eficaz.

Desvantagens

- Este algoritmo não é totalmente determinístico, uma vez que os pontos de fronteira podem pertencer a diferentes clusters, dependendo da ordem pela qual os dados são considerados. No entanto, uma vez que tal não acontece com pontos centrais ou pontos de ruído, essa situação é pouco frequente e tem pouco impacto no resultado de clustering. Existem variantes deste algoritmo, como o DBSCAN*, que são determinísticos, tratando os pontos de fronteira como pontos de ruído.
- Para dimensões elevadas, o desempenho tende a diminuir (“maldição da dimensionalidade”), uma vez que a noção de distância (independentemente da sua escolha) se torna menos significativa, tornando mais difícil encontrar clusters significativos com precisão.
- Para conjuntos de dados que tenham regiões com grandes diferenças de densidades, o processo apresenta várias dificuldades, uma vez que pressupõe uma densidade semelhante em todo o conjunto. Como a escolha dos parâmetros ϵ e $minPts$ é a mesma para todo o conjunto de dados, em alguns casos o clustering resultante pode estar longe do desejável. Dependendo da escolha dos parâmetros, pode haver muitos clusters de pequena dimensão ou poucos clusters de maior dimensão, mas muitos pontos classificados como ruído.

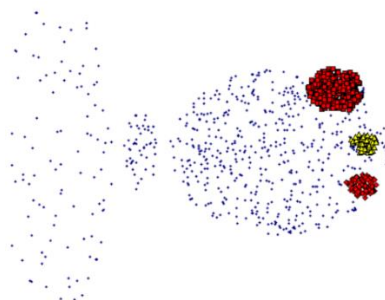


Figura 3.14. – Exemplo de clustering DBSCAN com regiões de diferente densidade. Os pontos são agrupados em 3 clusters, mas muitos pontos são (indevidamente) classificados como ruído.

Existem variantes deste algoritmo, como DBSCAN*, DENCLUE ou HDBSCAN, que permitem minimizar esta desvantagem ao introduzir a noção de ε variável. (Campello, Moulavi, & Sander, 2013)

- A escolha dos parâmetros tem impacto na formação dos clusters e não é fácil encontrar a combinação ideal, especialmente em conjuntos de dados com densidades variadas ou clusters de tamanhos e formas diferentes.

Escolha de parâmetros

Encontrar valores adequados para os parâmetros é crucial para obter resultados significativos de clustering. Não sendo determinados de forma automática, devem ser especificados por quem faz a análise. Para isso, há algumas sugestões que devem ser consideradas:

- Ter em conta as características dos dados, bem como algum conhecimento prévio sobre eles ou sobre o domínio. Também alguma informação sobre os tamanhos ou número esperado de clusters e densidade devem ser tidas em conta.

- Visualizar os dados de forma a obter informação sobre densidade, tamanhos de clusters e formas no conjunto de dados.

- Quanto ao valor de *minPts*, é conveniente ser maior ou igual a 3. (Valores inferiores darão o mesmo resultado que o clustering hierárquico de ligação mínima.) Em geral, é conveniente utilizar valores maiores, de forma a obter clusters mais significativos. Existem recomendações de utilizar $\text{minPts} \geq n + 1$ ou $\text{minPts} \geq 2n$, onde n é a dimensão do espaço do qual o conjunto de dados faz parte. No entanto, há que ter em atenção que estas são apenas sugestões e não regras a seguir de forma rígida. (Schubert, Sander, Martin, Kriegel, & Xu, 2017)

- Para escolher o valor de ε , pode ser utilizada a distância (euclidiana, por exemplo) de cada ponto ao k-vizinho mais próximo (este valor de k também terá de ser escolhido, mas pode ser $\text{minPts} - 1$). Ordenando estes valores por ordem crescente, e representando-os num gráfico, o ponto de cotovelo ajudará a perceber onde a variação nas distâncias diminui, indicando a transição de regiões densas para regiões esparsas.

- A escolha da distância a utilizar está intimamente ligada à escolha de ε e tem um grande impacto nos resultados. Em geral, será necessário primeiro identificar uma medida razoável de similaridade para o conjunto de dados, antes que o parâmetro ε possa ser escolhido.

- Devem ser avaliados os resultados do clustering, repetindo e refinando os valores dos parâmetros até obter resultados satisfatórios.
- É essencial encontrar um equilíbrio entre obter clusters significativos, evitar segmentação excessiva e lidar corretamente com *outliers*. A adequação dos valores dos parâmetros depende do conjunto de dados e dos objetivos da sua análise.

3.1.5. Conclusão

A utilização de algoritmos de clustering pode ser muito útil na identificação de padrões e estruturas nos dados, podendo ajudar na fundamentação da tomada de decisões. Havendo vários tipos de algoritmos, é importante testar diferentes opções, uma vez que cada um apresenta as suas vantagens e as suas limitações. Existe a questão de pressupostos associados a cada um deles, como o formato ou o número dos clusters, o tamanho da base de dados e a rapidez de aplicação do algoritmo. É de grande importância o recurso a conhecimento específico do domínio ao qual a base de dados diz respeito, tomando em conta as contribuições de especialistas.

3.2. Regras de Associação

3.2.1. Introdução e noções gerais

A análise de regras de associação surgiu como uma técnica popular para *Data Mining* de dados comerciais, que procura descobrir relações e padrões significativos entre elementos de conjuntos de dados de grande dimensão. Os pioneiros desta técnica foram Rakesh Agrawal, Tomasz Imieliński e Arun Swami que em 1993 apresentaram regras de associação para descobrir regularidades entre produtos vendidos (em grande escala) em supermercados. (Agrawal, Imieliński, & Swami, 1993) Com alguns desenvolvimentos nas últimas décadas, tem sido utilizada no comércio, análise de cestos de compras, análise de comportamentos de clientes, análise de navegação em sites, entre outros. São particularmente úteis na análise de cestos de compras, em que o objetivo é identificar itens que são frequentemente comprados juntos, como produtos num carrinho de compras ou vendidos em pacote. Recentemente, tem sido utilizada noutras áreas, como sistemas de recomendação, extração de conhecimento na internet, deteção de intrusos e bioinformática.

A aplicação desta técnica pode ajudar as empresas a tomar decisões de negócios, otimizar vendas e aprimorar estratégias de marketing baseadas em padrões de compra. Entre essas estratégias estão posicionamento de produtos (*product placement*), vendas cruzadas (*cross-selling*), gestão de prateleiras (*stocking shelves*), marketing cruzado (*cross-marketing*), elaboração de catálogos (*catalog design*) e segmentação de clientes.

Dois produtos e que sejam frequentemente comprados numa loja podem:

- ser colocados na mesma prateleira, de forma a que quem compra um dos produtos seja induzido a comprar o outro);
- ter descontos promocionais em apenas um dos produtos;
- serem combinados (quando possível), criando um novo produto, que se espera ser do agrado dos compradores desses dois produtos. (Ng & Soo, 2017)

Seja $A = (a_1, a_2, \dots, a_m)$ o universo de m itens, que podem ser produtos num supermercado, opções de serviços ou páginas na internet. Qualquer subconjunto I de A é denominado de **conjunto de itens** e diz respeito a qualquer grupo de produtos que podem ser comprados em conjunto.

Uma vez que nos próximos parágrafos serão introduzidos vários conceitos, é apresentado de seguida um exemplo, que servirá de base para exemplificar os diferentes conceitos.

Transação	Produtos
t_1	$\{maçã, cerveja, arroz, carne\}$
t_2	$\{maçã, cerveja, arroz\}$
t_3	$\{maçã, cerveja\}$
t_4	$\{maçã, pera\}$
t_5	$\{leite, cerveja, arroz, carne\}$
t_6	$\{leite, cerveja, arroz\}$
t_7	$\{leite, cerveja\}$
t_8	$\{leite, pera\}$

Tabela 3.1. – Exemplo de transações e produtos em cada transação

Neste exemplo, o conjunto de itens é $A = (maçã, cerveja, arroz, carne, pera, leite)$ e a dimensão é $m = 6$.

Seja $T = (t_1, t_2, \dots, t_n)$ um conjunto de n transações, que pode listar o conjunto de produtos adquiridos por um cliente num supermercado numa dada compra. Cada transação é um conjunto de itens, mas alguns conjuntos de itens podem não fazer parte de T (se for incluído o conjunto vazio, há 2^m conjuntos de itens possíveis, que usualmente é muito maior do que n).

No exemplo apresentado, o conjunto de transações é $T = (t_1, t_2, \dots, t_8)$.

Diz-se que uma transação $t \in T$ **suporta** o conjunto de itens I quando $I \subseteq t$, ou seja, a transação t contém todos os elementos de I . Por exemplo, as transações 5, 6 e 7 suportam o conjunto de itens $\{cerveja, leite\}$.

Uma **regra de associação** é definida como uma implicação do tipo $X \Rightarrow Y$, em que $X, Y \subseteq A$. Por outras palavras, uma regra de associação é da forma “se *antecedente* então *consequente*”, em que tanto o antecedente como o consequente são conjuntos de itens. A ideia de uma regra de associação é a de que, se os elementos de X pertencem a uma transação, então é de esperar que os elementos de Y também pertençam. As regras são calculadas a partir dos dados e não têm uma natureza determinística, mas probabilística, uma vez que não é de esperar que uma regra funcione em 100% dos casos. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2017)

Um exemplo de regra de associação pode ser $\{leite, arroz\} \Rightarrow \{cerveja\}$, o que significa que, se numa transação estiver leite e cerveja, também deverá estar arroz. Neste caso, isso é certo que aconteça, mas tal não é necessário, uma vez que não se trata de regras determinísticas, mas probabilísticas. Já a regra $\{leite, cerveja\} \Rightarrow \{arroz\}$ confirma-se em $\frac{2}{3}$ das situações, ou seja, em 2 das 3 transações onde estão leite e cerveja, também está arroz. Uma outra regra pode ser $\{leite, carne\} \Rightarrow \{cerveja, arroz\}$ que, embora funcione sempre, apenas se verifica numa das transações, o que é pouco expressivo.

Enquanto num exemplo com 6 produtos e 8 transações é relativamente simples encontrar algumas regras, numa base de dados com milhares de produtos e de transações, tal tarefa torna-se bastante complicada, até porque podem ser geradas regras sem qualquer interesse (através de algoritmos que serão apresentados mais à frente). Torna-se, portanto, necessário introduzir algumas medidas que ajudem a descobrir quais as regras relevantes de entre todas as regras possíveis. (Kirenz, 2020)

3.2.2. Medidas de associação

Suporte

O **suporte** de um conjunto de itens indica a frequência relativa com que esse conjunto surge nas transações, ou seja, é o quociente entre o número de transações em que todos os elementos desse conjunto surgem e o número total de transações. Também pode ser visto como a probabilidade de ocorrência de todos os elementos de um conjunto de itens. Este conceito também se aplica a regras, ou seja, o suporte de uma regra é o quociente entre o número de transações em que todos os elementos do antecedente e do conseqüente surgem e o número total de transações.

No exemplo dado, o suporte de $\{cerveja, arroz\}$ é 4 em 8, ou seja 50%, e o suporte de $\{leite, cerveja\} \Rightarrow \{arroz\}$ é 2 em 8, ou seja 25%.

$$support(\{cerveja, arroz\}) = \frac{4}{8} = 50\% \quad support(\{leite, cerveja\} \Rightarrow \{arroz\}) = \frac{2}{8} = 25\%$$

Em geral, $support(X) = P(X)$ e $support(X \Rightarrow Y) = P(X \cup Y) = P(E_X \cap E_Y)$.⁶

Confiança

A **confiança** de uma regra mede o grau de fiabilidade de uma regra, ou seja, o quão frequentemente a regra é verdadeira. É dada pelo quociente entre o número de

⁶ Convém clarificar o que se entende por $X \cup Y$. Aqui, consideram-se todos os elementos de X e também os de Y , daí $X \cup Y$. Em termos de transações, pretende-se as que incluem os elementos de X e também os elementos de Y . Assim, a reunião dos conjuntos X e Y conduz à interseção do conjunto de transações que incluem X (que se designa por E_X) e do conjunto de transações que incluem Y (que se designa por E_Y).

transações que incluem os elementos tanto do antecedente como do consequente e o número de transações que incluem apenas os elementos do antecedente. Também pode ser calculada como o quociente entre o suporte de uma regra e o suporte do antecedente. Em termos probabilísticos, a confiança fornece a probabilidade (condicionada) de o consequente ocorrer sabendo que o antecedente se verificou.

No exemplo dado, a confiança de $\{leite, cerveja\} \Rightarrow \{arroz\}$ é $\frac{2}{3}$. Também pode ser calculado como o quociente entre o suporte de $\{leite, cerveja\} \Rightarrow \{arroz\}$ ($\frac{2}{8}$) e o suporte de $\{leite, cerveja\}$ ($\frac{3}{8}$).

$$confidence(\{leite, cerveja\} \Rightarrow \{arroz\}) = \frac{2}{3}$$

$$confidence(\{leite, cerveja\} \Rightarrow \{arroz\}) = \frac{support(\{leite, cerveja\} \Rightarrow \{arroz\})}{support(\{leite, cerveja\})} = \frac{\frac{2}{8}}{\frac{3}{8}} = \frac{2}{3}$$

$$\text{Em geral, } confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} = \frac{P(X \cup Y)}{P(X)} = \frac{P(E_X \cap E_Y)}{P(E_X)} = P(E_Y | E_X).$$

Uma medida menos utilizada, mas que também convém mencionar é a **cobertura**, que avalia a frequência relativa do antecedente da regra no conjunto de transações. Assim, a cobertura de uma regra é o quociente entre o número de transações que incluem os elementos do antecedente e o número total de transações.

Por exemplo, a cobertura de $\{leite, cerveja\} \Rightarrow \{arroz\}$ é 3 em 8, ou seja 37,5%.

$$coverage(\{leite, cerveja\} \Rightarrow \{arroz\}) = support(\{leite, cerveja\}) = \frac{3}{8}.$$

$$\text{Em geral, } coverage(X \Rightarrow Y) = support(X).$$

Força

A **força**, também pouco utilizada⁷, mede quão mais frequente é o consequente do que o antecedente. Assim, a força de uma regra é o quociente entre o número de transações que incluem os elementos do consequente e o número de transações que incluem os elementos do antecedente.

Por exemplo, a força de $\{leite, cerveja\} \Rightarrow \{arroz\}$ é 4 em 2, ou seja 2.

$$strength(\{leite, cerveja\} \Rightarrow \{arroz\}) = \frac{support(\{arroz\})}{support(\{leite, cerveja\})} = \frac{\frac{4}{8}}{\frac{2}{8}} = 2.$$

$$\text{Em geral, } strength(X \Rightarrow Y) = \frac{support(Y)}{support(X)}.$$

⁷ São aqui referidas aqui a cobertura e a força, uma vez que são medidas utilizadas no software *Orange*.

Lift

O coeficiente de interesse, ou **lift**, reflete a noção estatística de independência entre o antecedente e o consequente. Indica quão mais provável é que o antecedente e o consequente sejam comprados em conjunto do que seria de esperar se fossem estatisticamente independentes. O *lift* de uma regra é o quociente entre a sua confiança e o suporte do consequente. Também pode ser visto como o quociente entre o suporte da regra e o produto dos suportes do antecedente e do consequente.

$$lift(X \Rightarrow Y) = \frac{confidence(X \Rightarrow Y)}{support(Y)} = \frac{support(X \cup Y)}{support(X) \times support(Y)} = \frac{P(X \cup Y)}{P(X) \times P(Y)} = \frac{P(E_X \cap E_Y)}{P(E_X) \times P(E_Y)}$$

- $lift > 1$ indica que a presença do antecedente na transação aumenta a probabilidade de o consequente também estar.

- $lift = 1$ indica que não há qualquer associação entre o antecedente e o consequente, uma vez que $P(E_X \cap E_Y) = P(E_X) \times P(E_Y)$, ou seja, X e Y são independentes.

- $0 \leq lift < 1$ indica que a presença do antecedente na transação reduz a probabilidade de o consequente também estar.

$$lift(\{leite, cerveja\} \Rightarrow \{arroz\}) = \frac{support(\{leite, cerveja, arroz\})}{support(\{leite, cerveja\}) \times support(\{arroz\})} = \frac{\frac{2}{8}}{\frac{3}{8} \times \frac{1}{2}} = \frac{4}{3} > 1,$$

pelo que a compra de leite e cerveja influencia positivamente a probabilidade de também comprar arroz. Também se pode ver da seguinte forma: uma vez que a probabilidade de comprar arroz é $\frac{1}{2}$ e, no caso de ter comprado leite e arroz, passa a $\frac{2}{3}$,

há um aumento na probabilidade, que é quantificado como $\frac{\frac{2}{3}}{\frac{1}{2}} = \frac{4}{3}$.

$$lift(\{maçã\} \Rightarrow \{cerveja\}) = \frac{support(\{maçã, cerveja\})}{support(\{maçã\}) \times support(\{cerveja\})} = \frac{\frac{3}{8}}{\frac{1}{2} \times \frac{3}{4}} = 1, \text{ pelo que a}$$

compra de maçã não influencia a probabilidade de também comprar arroz. De outra forma, uma vez que a probabilidade de comprar cerveja é $\frac{3}{4}$, independentemente de ter ou não comprado maçã, estes dois produtos não têm qualquer dependência estatística.

Leverage

Uma outra medida, que também procura detetar a dependência entre o antecedente e o consequente de uma regra é o **leverage**, que é dado pela diferença entre o suporte da regra e o produto dos suportes do antecedente e do consequente.

$$\begin{aligned} leverage(X \Rightarrow Y) &= support(X \cup Y) - support(X) \times support(Y) \\ &= P(E_X \cap E_Y) - P(E_X) \times P(E_Y) \end{aligned}$$

O *leverage* está relacionado com o *lift* e mede a dependência, embora com uma escala diferente. Enquanto o *lift* toma valores em $[0, +\infty[$, *leverage* toma valores em $] -1, 1[$.⁸

Algumas observações podem ser feitas em relação aos valores de *leverage*:

- $leverage > 0 \Leftrightarrow lift > 1$, ou seja, a presença do antecedente na transação aumenta a probabilidade de o consequente também estar;
- quanto mais próximo de 1 estiver o *leverage* (o que corresponde a valores elevados de *lift*), mais forte é a associação (positiva) entre o antecedente e do consequente;
- $leverage = 0 \Leftrightarrow lift = 1$, ou seja, a presença do antecedente na transação não interfere na probabilidade de o consequente também estar;
- quanto mais próximo de 0 estiver o *leverage* (o que corresponde a valores de *lift* próximos de 1), mais fraca é a associação entre o antecedente e do consequente;
- $leverage < 0 \Leftrightarrow lift < 1$, ou seja, a presença do antecedente na transação diminui a probabilidade de o consequente também estar;
- quanto mais próximo de -1 estiver o *leverage* (o que corresponde a valores de *lift* próximos de 0) mais forte é a associação (negativa) entre o antecedente e do consequente.

$$leverage(\{leite, cerveja\} \Rightarrow \{arroz\})$$

$$\begin{aligned} &= support(\{leite, cerveja, arroz\}) - support(\{leite, cerveja\}) \times support(\{arroz\}) = \\ &\frac{2}{8} - \frac{3}{8} \times \frac{1}{2} = \frac{4}{16} - \frac{3}{16} = \frac{1}{16} > 0, \text{ pelo que a compra de leite e cerveja influencia} \\ &\text{positivamente a probabilidade de também comprar arroz.} \end{aligned}$$

$$leverage(\{maçã\} \Rightarrow \{cerveja\})$$

⁸ *Leverage* não pode tomar valor 1, uma vez que para tal acontecer teria de ocorrer em simultâneo $P(E_X \cap E_Y) = 1$ e $P(E_X) \times P(E_Y) = 0$. Mas $P(E_X) \times P(E_Y) = 0 \Leftrightarrow P(E_X) = 0 \vee P(E_Y) = 0 \Leftrightarrow E_X = \emptyset \vee E_Y = \emptyset$ e, portanto, $E_X \cap E_Y = \emptyset$ e $P(E_X \cap E_Y) = 0$, contradizendo $P(E_X \cap E_Y) = 1$.

Também não pode tomar valor -1, uma vez que teria de ocorrer em simultâneo $P(E_X \cap E_Y) = 0$ e $P(E_X) \times P(E_Y) = 1$. Mas $P(E_X) \times P(E_Y) = 1 \Leftrightarrow P(E_X) = 1 \wedge P(E_Y) = 1 \Leftrightarrow E_X = A \wedge E_Y = A$. (A é o conjunto de todos os itens.) Portanto, $E_X \cap E_Y = A$ e $P(E_X \cap E_Y) = 1$, contradizendo $P(E_X \cap E_Y) = 0$.

$= \text{support}(\{maçã, cerveja\}) - \text{support}(\{maçã\}) \times \text{support}(\{cerveja\}) = \frac{3}{8} - \frac{1}{2} \times \frac{3}{4} = \frac{3}{8} - \frac{3}{8} = 0$ pelo que a compra de maçã não influencia a probabilidade de também comprar arroz.

Em termos práticos, o *leverage* é usado para filtrar regras de associação que representam ocorrências aleatórias. Regras com *leverage* positivo e próximo de 1 são consideradas mais interessantes e indicativas de associações reais entre itens.

Embora o *leverage* seja uma medida valiosa para encontrar de regras de associação, ela é frequentemente usada em conjunto com outras métricas, como confiança, suporte e *lift*, para avaliar de forma abrangente a qualidade e a importância das regras de associação.

Convicção

A **convicção** mede o quão convincente é a regra e é dada pelo quociente entre a frequência relativa de ausência do consequente nas transações e a probabilidade de erro da regra. Esta medida é útil para identificar regras com consequentes fortes.

Em geral, $\text{conviction}(X \Rightarrow Y) = \frac{1 - \text{support}(Y)}{1 - \text{confidence}(X \Rightarrow Y)}$.

$$\text{conviction}(\{leite, cerveja\} \Rightarrow \{arroz\}) = \frac{1 - \text{support}(\{arroz\})}{1 - \text{confidence}(\{leite, cerveja\} \Rightarrow \{arroz\})} = \frac{1 - \frac{1}{2}}{1 - \frac{2}{3}} = 1,5,$$

ou seja, seria de esperar que a regra estivesse errada 50% mais (1,5 vezes mais) se a associação entre $\{leite, cerveja\}$ e $\{arroz\}$ fosse meramente aleatória.

3.2.3. Processo de determinação de regras de associação

As regras de associação são determinadas procurando nos dados padrões frequentes, usando critérios como suporte, confiança ou *lift* para determinar quais os relacionamentos mais importantes.

Se as regras fossem construídas a partir da análise de todos os possíveis conjuntos de itens possíveis, para além do tempo despendido, haveria tantas regras que não teriam nenhum significado. É por isso importante escolher regras que sejam bem representadas pelos dados.

A utilização de regras de associação apresenta várias vantagens, como encontrar padrões que ajudam a compreender correlações e coocorrências entre conjuntos de dados. Um exemplo disso vem da medicina, onde pode ajudar a determinar a probabilidade condicional de uma doença comparando as relações de sintomas de casos anteriores. (Kirenz, 2020)

No entanto, também apresenta algumas desvantagens, como a dificuldade em definir adequadamente os patamares mínimos ou em encontrar o equilíbrio ideal entre a quantidade e a qualidade de regras de associação.

O processo de determinação de regras de associação passa por aplicar um algoritmo que começa por determinar conjuntos de itens frequentes, sendo depois criadas regras de associação a partir destes, tendo em conta critérios definidos, como confiança, suporte ou *lift*.

Serão apresentados dois dos mais populares algoritmos, A-Priori e *FP-Growth*. O primeiro pela sua relevância histórica, uma vez que se trata do algoritmo inicialmente proposto em 1994, e o segundo por ser o algoritmo usado pelo software Orange.

3.2.4. Algoritmo A-Priori

O Algoritmo A-Priori, apresentado por Agrawal e Srikant em 1994, foi o primeiro a ser usado para determinar conjuntos de itens frequentes e regras de associação. Ele realiza uma série de iterações, descobrindo progressivamente conjuntos de itens frequentes de tamanho crescente. O nome do algoritmo é A-Priori porque usa conhecimento prévio de propriedades de conjuntos de itens frequentes. (Agrawal, Imieliński, & Swami, 1993)

Contagem de itens e identificação de itens frequentes

O algoritmo começa por percorrer o conjunto de dados para contar a frequência de cada item e identificar itens frequentes (que satisfazem um patamar de suporte mínimo definido). No exemplo apresentado, a tabela de frequências seria a seguinte:

Item	Frequência
maçã	4
cerveja	6
arroz	4
carne	2
pera	2
leite	4

Tabela 3.2. – Tabela de frequências para cada item

Se o suporte mínimo for 0,3, significa que todos os itens com uma frequência inferior a $8 \times 0,3 = 2,4$ deverão ser descartados. Neste caso, são a carne e a pera. Os restantes 4 são considerados itens frequentes. Assim, numa primeira iteração do algoritmo, os

conjuntos de itens frequentes são $\{maçã\}$, $\{cerveja\}$, $\{arroz\}$ e $\{leite\}$ e escreve-se $C_1 = \{\{maçã\}, \{cerveja\}, \{arroz\}, \{leite\}\}$.

Segunda iteração do algoritmo

Na iteração seguinte, são considerados todos os conjuntos com dois itens que se podem formar a partir dos 4 elementos de C_1 e conta-se a frequência de cada um deles.

Conjuntos de itens	Frequência
$\{maçã, cerveja\}$	3
$\{maçã, arroz\}$	1
$\{maçã, leite\}$	0
$\{cerveja, arroz\}$	4
$\{cerveja, leite\}$	3
$\{arroz, leite\}$	2

Tabela 3.3. – Tabela de frequências dos conjuntos com dois itens

Se o suporte mínimo for 0,3, $C_2 = \{\{maçã, cerveja\}, \{cerveja, arroz\}, \{cerveja, leite\}\}$.

Terceira iteração do algoritmo

Procedendo como no passo anterior, obtém-se:

Conjuntos de itens	Frequência
$\{maçã, cerveja, arroz\}$	2
$\{maçã, cerveja, leite\}$	0
$\{maçã, arroz, leite\}$	0
$\{cerveja, arroz, leite\}$	2

Tabela 3.4. – Tabela de frequências dos conjuntos com três itens

Como nenhum dos conjuntos de itens satisfaz o suporte mínimo, $C_3 = \emptyset$ e o processo termina.

Em geral, o processo continua até que, numa iteração k , $C_k = \emptyset$ e o conjunto de itens frequentes é dado por $\bigcup_{i=1}^{k-1} C_i$.

No exemplo apresentado, os itens frequentes são $\{cerveja, arroz\}$, $\{cerveja, leite\}$, $\{cerveja, maçã\}$, $\{cerveja\}$, $\{arroz\}$, $\{leite\}$ e $\{maçã\}$.

Determinação de regras de associação

Assim que os conjuntos de itens frequentes tenham sido identificados, podem ser geradas regras de associação a partir deles. Este passo envolve calcular a confiança para cada regra possível.

As regras são geradas considerando diferentes combinações de antecedentes e consequentes dos conjuntos de itens frequentes. Estas regras podem ser podadas tendo em conta critérios definidos pelo utilizador como confiança mínima ou *lift* mínimo⁹, de forma a reter apenas as regras mais relevantes.

Por fim, o conjunto de final de regras de associação, bem como os valores de suporte e confiança, são apresentados aos utilizadores para interpretação e tomada de decisão.

No exemplo apresentado, as regras possíveis são descritas na tabela seguinte, bem como as medidas de associação.

Regra	Suporte	Confiança	Cobertura	Força	<i>Lift</i>	<i>Leverage</i>	Convicção
$\{cerveja\} \Rightarrow \{arroz\}$	0,5	0,(6)	0,75	0,(6)	1,(3)	0,125	1,5
$\{arroz\} \Rightarrow \{cerveja\}$	0,5	1	0,5	1,5	1,(3)	0,125	-
$\{cerveja\} \Rightarrow \{leite\}$	0,375	0,5	0,75	0,(6)	1	0	1
$\{leite\} \Rightarrow \{cerveja\}$	0,375	0,75	0,5	1,5	1	0	1
$\{cerveja\} \Rightarrow \{maçã\}$	0,375	0,5	0,75	0,(6)	1	0	1
$\{maçã\} \Rightarrow \{cerveja\}$	0,375	0,75	0,5	1,5	1	0	1

Tabela 3.5. – Medidas de associação para as regras de associação

Deve ser notado que, no exemplo apresentado, os conjuntos de itens frequentes não têm mais de 2 elementos, o que limita bastante as possibilidades de diferentes regras. Num exemplo de um conjunto de itens frequentes $\{maçã, cerveja, arroz\}$, as regras testadas seriam $\{maçã, cerveja\} \Rightarrow \{arroz\}$, $\{arroz\} \Rightarrow \{maçã, cerveja\}$, $\{maçã, arroz\} \Rightarrow \{cerveja\}$, $\{cerveja\} \Rightarrow \{maçã, arroz\}$, $\{cerveja, arroz\} \Rightarrow \{maçã\}$ e $\{maçã\} \Rightarrow \{cerveja, arroz\}$.

Em geral, um conjunto de itens frequentes com n elementos pode dar origem a $2^n - 2$ a diferentes regras.

⁹ Também podem ser consideradas outras métricas, como *leverage* ou *convicção*, para aferir a relevância das regras.

3.2.5. Algoritmo *FP-Growth*

O Algoritmo *FP-Growth* (*Frequent Pattern Growth*, Crescimento de Padrões Frequentes) é uma técnica amplamente usada em regras de associação para descobrir conjuntos de itens frequentes e gerar regras de associação a partir de bases de dados de transações de grandes dimensões, no qual se revela particularmente eficiente. (Jiawei, Jian, & Yiwen, 2000)

O Algoritmo *FP-Growth* constrói uma estrutura chamada *FP-Tree* (*Frequent Pattern Tree*, Árvore de Padrões Frequentes) a partir dos dados de transações. A *FP-Tree* é uma representação dos conjuntos de itens frequentes nos dados.

Contagem de itens frequentes

O algoritmo começa por percorrer o conjunto de dados para contar a frequência de cada item e identificar itens frequentes (que satisfazem um patamar de suporte mínimo definido).

Ordenação de frequências de contagem de itens

De seguida, estes itens frequentes são ordenados por ordem decrescente de frequência, sendo criada uma tabela de frequências. Esta ordenação facilita a construção da *FP-Tree* de forma eficiente.

No exemplo apresentado, a tabela de frequências seria a seguinte:

Item	Frequência
cerveja	6
maçã	4
leite	4
arroz	4
carne	2
pera	2

Tabela 3.6. – Tabela de frequências para cada item

Reprocessamento de transações

As transações são reprocessadas, considerando apenas os itens frequentes, ou seja, aqueles que satisfazem o suporte mínimo definido. Esses itens em cada transação são ordenados de acordo com a ordem na tabela de frequências. Os restantes itens são descartados.

Se o suporte mínimo for 0,3, significa que todos os itens com uma frequência inferior a $8 \times 0,3 = 2,4$ deverão ser descartados. Neste caso, são a carne e a pera. A tabela com as transações reprocessadas e ordenadas (isto é, com os itens ordenados de acordo com a ordem na tabela de frequências) é a seguinte:

t_1	{ <i>cerveja, maçã, arroz</i> }
t_2	{ <i>cerveja, maçã, arroz</i> }
t_3	{ <i>cerveja, maçã</i> }
t_4	{ <i>maçã</i> }
t_5	{ <i>cerveja, leite, arroz</i> }
t_6	{ <i>cerveja, leite, arroz</i> }
t_7	{ <i>cerveja, leite</i> }
t_8	{ <i>leite</i> }

Tabela 3.8. – Transações reprocessadas e ordenadas

Construção da *FP-Tree*

De seguida, cria-se a árvore de padrões frequentes, a *FP-Tree*. A estrutura da *FP-Tree* permite armazenar e representar os conjuntos de itens, bem como as suas relações. (Gama, Carvalho, Facelli, Lorena, & Oliveira, 2017)

Primeiro, cria-se a raiz da árvore, designada por *null*. Depois, todas as transações são processadas, criando a *FP-Tree* como no exemplo.

- $t_1 = \{cerveja, maçã, arroz\}$, gerando o percurso $null \rightarrow cerveja \rightarrow maçã \rightarrow arroz$. Cada um dos itens está associado a um nó da árvore e tem frequência 1.
- $t_2 = \{cerveja, maçã, arroz\}$, mas não se gera um novo percurso. Apenas se atualiza cada uma das frequências para 2.
- $t_3 = \{cerveja, maçã\}$. Novamente, não se gera um novo percurso. Apenas se atualiza as frequências de *cerveja* e *maçã* para 3.
- $t_4 = \{maçã\}$. Aqui, gera-se um novo percurso, $null \rightarrow maçã$. O nó de *maçã* tem frequência 1.
- $t_5 = \{cerveja, leite, arroz\}$, gerando o percurso $null \rightarrow cerveja \rightarrow leite \rightarrow arroz$. A frequência de *cerveja* é atualizada para 4, enquanto os nós de *leite* e *arroz*, agora criados, têm frequência 1.

- $t_6 = \{cerveja, leite, arroz\}$, mas não se gera um novo percurso. Apenas se atualizam as frequências de cerveja, leite e arroz para 5, 2 e 2, respetivamente.
- $t_7 = \{cerveja, leite\}$. Novamente, não se gera um novo percurso. Apenas se atualiza as frequências de cerveja para 6 e de leite para 3.
- $t_8 = \{leite\}$. Aqui, gera-se um novo percurso, $null \rightarrow leite$. O nó de leite tem frequência 1.

Assim, obtém-se a *FP-tree* correspondente.

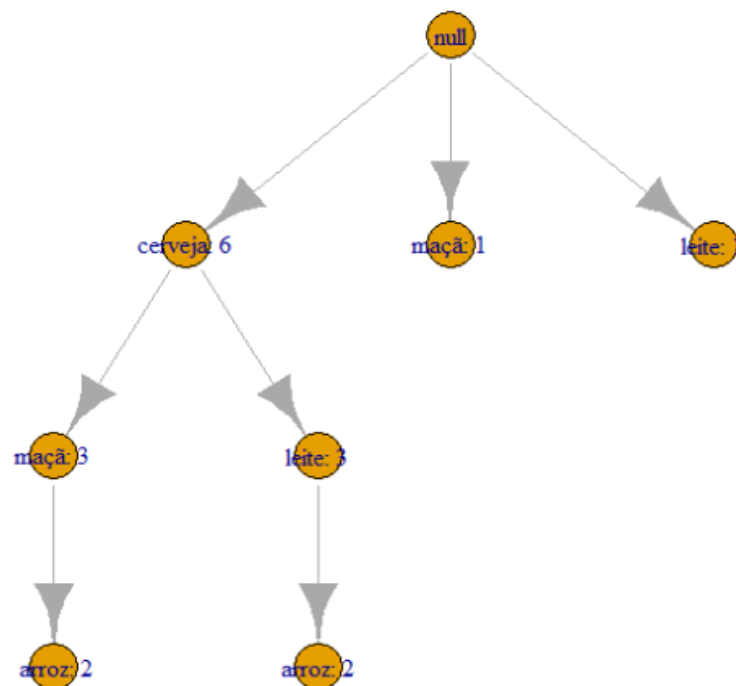


Figura 3.15. – *FP-Tree* resultante da aplicação do algoritmo *FP-Growth*

Determinação de conjuntos de itens frequentes

Depois de construir a *FP-Tree*, os conjuntos de itens frequentes podem ser extraídos de forma eficiente usando um método recursivo chamado “crescimento de padrão”. Este algoritmo leva ao crescimento dos conjuntos de itens frequentes, em vez de gerar candidatos e testá-los como no Algoritmo A-Priori.

Começando com o item menos frequente da tabela de frequências, os padrões vão crescendo explorando a *FP-Tree* de forma recursiva e são formados novos conjuntos de itens. Este processo gera os conjuntos de itens de forma eficiente.

Escolhendo o item menos frequente, procuram-se todas as transações que terminam nesse item. No exemplo apresentado, o item menos frequente é arroz e as transações são $\{cerveja, maçã, arroz\}$ e $\{cerveja, leite, arroz\}$. Cada uma delas tem contagem 2 (porque é a frequência do nó de arroz em ambas as transações).

Contam-se as frequências de cada um dos itens, com exceção de arroz, que são *cerveja*: 4, *maçã*: 2, *leite*: 2. Como maçã e leite não satisfazem o suporte mínimo (2,4), apenas será considerada *cerveja*, bem como arroz. Assim, dois conjuntos de itens frequentes são $\{cerveja, arroz\}$ e $\{arroz\}$, ambos com contagem 4.

Passa-se então ao item seguinte, leite. A transação que termina em leite é $\{cerveja, leite\}$, com contagem 3. Assim, o item *cerveja* tem contagem 3, satisfaz o suporte mínimo e portanto $\{cerveja, leite\}$ é um conjunto de itens frequentes com contagem 3. $\{leite\}$ também é um conjunto de itens frequentes com contagem 4.

O mesmo se faz para maçã e, por fim, para *cerveja*.

Item	Percursos com terminação no item	Contagem dos itens	FP-Tree condicional	Conjuntos de itens frequentes
<i>arroz</i>	<i>cerveja, maçã</i> : 2 <i>cerveja, leite</i> : 2	<i>cerveja</i> : 4 <i>leite</i> : 2 <i>maçã</i> : 2	<i>cerveja</i> : 4	$\{cerveja, arroz\}$: 4 $\{arroz\}$: 4
<i>leite</i>	<i>cerveja</i> : 3	<i>cerveja</i> : 3	<i>cerveja</i> : 3	$\{cerveja, leite\}$: 3 $\{leite\}$: 4
<i>maçã</i>	<i>cerveja</i> : 3	<i>cerveja</i> : 3	<i>cerveja</i> : 3	$\{cerveja, maçã\}$: 3 $\{maçã\}$: 4
<i>cerveja</i>	-	-	-	$\{cerveja\}$: 6

Tabela 3.8. – Determinação dos conjuntos de itens frequentes

Obtêm-se, assim, os conjuntos de itens frequentes e as suas frequências: $\{cerveja, arroz\}$: 4, $\{cerveja, leite\}$: 3, $\{cerveja, maçã\}$: 3, $\{cerveja\}$: 6, $\{arroz\}$: 4, $\{leite\}$: 4 e $\{maçã\}$: 4

Determinação de regras de associação

A partir daqui o procedimento é o mesmo que no Algoritmo A-Priori, descrito anteriormente. No exemplo apresentado, os conjuntos de itens frequentes são os mesmos que os que foram obtidos no Algoritmo A-Priori, pelo que não será aqui repetida tabela apresentada.

4. Tratamento, exploração, apresentação e análise dos dados

4.1. Descrição dos dados

Foram fornecidos pelo orientador Vítor Ferreira dois ficheiros (wc-product-export-22-2-2023-1677104637321.csv e orders-2023-03-06-17-42-32.xlsx, abreviados para Product.csv e Orders.xlsx, respetivamente) com dados sobre produtos e encomendas realizadas em idivia.com entre os dias 5 de outubro de 2021 e 6 de março de 2023.

As variáveis¹⁰ em cada ficheiro são a seguir apresentadas e sucintamente descritas.

	Variável	Descrição
Product.csv	SKU	Stock Keeping Unit, código utilizado para identificar e organizar produtos no inventário
	Name	Nome do produto
	Weight (g)	Peso (em gramas)
	Sale price	Preço de venda
	Regular price	Preço normal
	Categories	Categorias às quais o produto pertence (Makeup, Hair,...)
	Attribute 1 value(s)	Marca do produto
Orders.xlsx	Order Number	Número da encomenda
	Customer User ID	Identificação do Cliente (tem valor 0 no caso de o cliente não estar identificado)
	Order Date	Data da encomenda
	Country Code (Shipping)	Código ISO 3166-2 (de duas letras) que identifica o país de envio da encomenda
	SKU	Stock Keeping Unit, código utilizado para identificar e organizar produtos no inventário
	Item Name	Nome do produto
	Quantity	Número de unidades do produto na encomenda
	Item Cost	Custo do produto (sem impostos)
	Item Cost (inc. tax)	Custo do produto (com impostos)
	Brand	Marca do produto

Tabela 4.1. – Variáveis de Product.csv e de Orders.xlsx e sua descrição

¹⁰ “SKU” e “Order Number” são consideradas nesta tabela, apesar de se tratarem de identificadores e não de variáveis usadas nas análises realizadas.

O ficheiro Product.csv tem 10509 linhas, que correspondem a 10509 SKU (sendo alguns destes repetidos). O ficheiro Orders.xlsx tem 17494 linhas, que correspondem a 17494 produtos/encomenda.

4.2. Limpeza e tratamento dos dados

Antes de proceder à análise dos dados, foi necessário proceder ao seu tratamento e limpeza, de forma a lidar antecipadamente com valores omissos ou inconsistentes. Também foram criadas outras variáveis que facilitassem a interpretação e representação dos dados e posteriores análises. Importando ambos os ficheiros (Product.csv e Orders.xlsx) para o Power BI, foram feitas algumas transformações nos dados, usando o Power Query.

No ficheiro Product.csv, foram retirados os SKU duplicados, uma vez que havia o mesmo o mesmo SKU com nomes diferentes para mais do que uma língua. No ficheiro Orders.xlsx, foram retiradas as linhas com quantidades nulas, uma vez que correspondem a encomendas não efetuadas.

Uma vez que nem toda a informação pretendida constava de apenas um dos ficheiros, foi necessário intercalar os dois, de forma a complementar a informação. A ideia seria usar SKU como chave primária, mas alguns (1478, cerca de 8,4%) dos SKU no ficheiro Orders.xlsx não correspondiam a qualquer SKU de Product.csv. Em alternativa, foi usado o nome do produto (Item Name em Orders.xlsx e Name em Product.csv), mas mesmo assim houve 782 (cerca de 4,5%) produtos em Orders.xlsx para os quais não foi possível encontrar correspondência em Product.csv. Assim, para estes produtos, não foi possível encontrar informação relevante, como Regular Price ou Categories. Estes produtos foram incluídos nas análises posteriores, mas com valores omissos em algumas das variáveis.

Também havia alguns produtos (527, cerca de 3%) sem informação relativa à marca (Brand). Para encontrar essa informação, foi elaborada uma lista com todas as marcas e efetuada uma pesquisa usando a função "Text.Contains" do Power Query para verificar se o nome do produto continha, no seu texto, o nome de alguma das marcas. Desta maneira, foi possível reduzir o número de produtos sem marca para apenas 177 (cerca de 1%), sendo-lhes atribuída a marca "Other Brands". Assim, a variável Brand foi modificada, imputando valores onde a informação era inexistente.

Foram criadas outras variáveis que ajudariam na análise, como a seguir se descreve:

- Client Identification, variável dicotómica que assume o valor 0 se não houver identificação do cliente e 1 se houver.
- Weekday, variável que associa o respetivo dia da semana a cada data de encomenda. Foi utilizada a função “Date.DayOfWeekName” do Power Query.
- Continent, variável que associa o respetivo continente a cada código de país. Como há uma grande diversidade de países, foi tomada a decisão de agrupar os países por continente. Foi utilizada a informação contida no ficheiro all.txt, que se encontra nos Anexos.
- Skincare (Y/N), Suncare (Y/N), Makeup (Y/N), Baby (Y/N), Men (Y/N), Oral (Y/N), Hair (Y/N) e Health (Y/N), variáveis dicotómicas que, partindo da variável Categories de Product.csv, indica se o produto pertence (1) ou não (0) à categoria. Apesar de haver muitas outras categorias na variável Categories, foram selecionadas estas 8 pela sua relevância e destaque dado na plataforma online.

Além destas variáveis criadas para todos os pares produto/encomenda, também foram criadas outras, mas apenas para os casos em que tal era possível, devido à ausência ou qualidade da informação. Dessas, destaca-se a variável Recency, calculada contando o número de dias decorridos desde a encomenda anterior. Para tal, usou-se a função “Duration.Days” do Power Query. Esta variável toma o valor “null” no caso da primeira encomenda do cliente, uma vez que não existe encomenda anterior. Como algumas das encomendas foram feitas sem identificação do cliente, para esta variável apenas foram consideradas as encomendas feitas por clientes com identificação.

Uma variável cujo estudo seria muito interessante é o preço. Apesar de todos os produtos em Orders.xlsx terem um custo (Item Cost), este valor apresentava algumas discrepâncias, devido ao facto de esse valor não estar sempre em euros, mas por vezes na moeda do país de envio da encomenda. Para ultrapassar esta situação, procurou-se criar uma variável que fornecesse automaticamente a taxa de câmbio para euros em função da data da encomenda e do código do país. Tal foi possível na maioria dos casos, com exceção de alguns países cujo câmbio não foi possível encontrar e se pressupôs que a compra havia sido feita em dólares americanos. No entanto, foi possível perceber quase de imediato que mesmo esta solução apresentava debilidades, uma vez que, no mesmo país e para o mesmo produto, havia uma grande discrepância entre valores. A conclusão que foi possível retirar é a de que a moeda de compra não fica determinada pelo país de envio da encomenda, mas por outros critérios que não foi possível apurar. Não sendo possível na generalidade dos

casos determinar a valor apurado na venda do produto, a opção foi a de não considerar a variável Item Cost de Orders.xlsx, mas a variável Regular Price de Product.csv, apenas nos casos em que essa informação existe.

Na tabela seguinte estão descritas as variáveis que serão utilizadas na construção de visuais e nas análises. Algumas delas foram obtidas diretamente dos ficheiros Product.csv ou Orders.xlsx, enquanto outras foram criadas a partir das variáveis existentes. Em particular, a variável Brand, apesar de existir em Orders.xlsx, foi modificada através de obtenção de informação adicional a partir de outras variáveis.

Variável	Descrição	Origem
Order Number	Número da encomenda	Orders.xlsx
Order Date	Data da encomenda	Orders.xlsx
Weekday	Dia da semana da encomenda	Criada a partir de Order Date
Client Identification	Variável dicotómica (0 – cliente não identificado; 1 – cliente identificado)	Criada a partir de Customer User ID
Continent	Continente do país de envio	Criada a partir de Country Code (Shipping)
Item Name	Nome do produto	Orders.xlsx
Quantity	Número de unidades do produto	Orders.xlsx
Brand	Marca do produto	Modificada
Regular Price	Preço normal	Product.csv
Skincare (Y/N)	Variável dicotómica (0 – não pertence à categoria; 1 – pertence à categoria)	Criada a partir de Categories
Suncare (Y/N)		
Makeup (Y/N)		
Baby (Y/N)		
Men (Y/N)		
Oral (Y/N)		
Hair (Y/N)		
Health (Y/N)		

Tabela 4.2. – Variáveis de Orders with Brand and Category, sua descrição e origem

4.3. Representação Visual dos Dados

Depois de proceder à limpeza e ao tratamento dos dados, foram criados no Power BI alguns visuais, que nos permitem retirar algumas conclusões relevantes sobre o conjunto de dados das vendas.

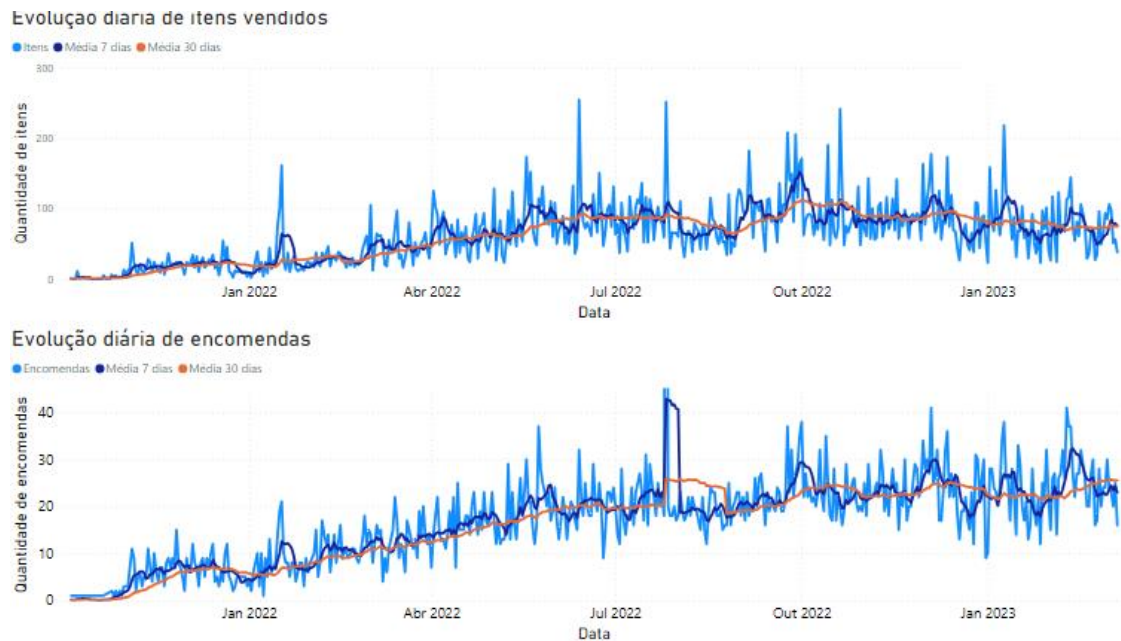
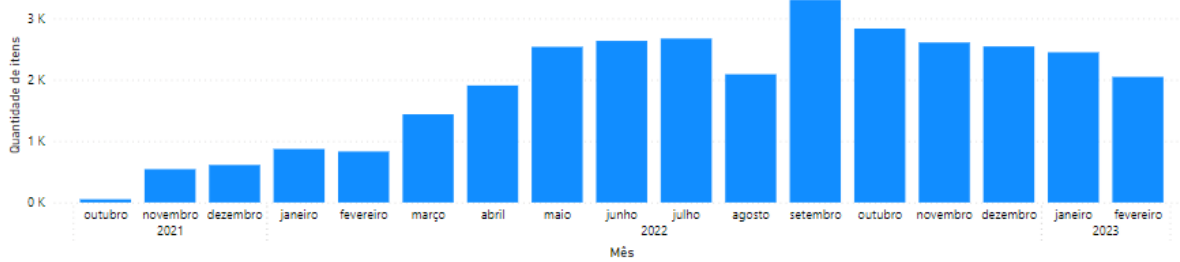


Gráfico 4.1. – Evolução diária (e médias a 7 e 30 dias) do volume de vendas

No Gráfico 4.1., é possível verificar uma ligeira tendência positiva no volume de produtos vendidos diariamente e também nas encomendas. Já no Gráfico 4.2, que ilustra a evolução mensal, é possível confirmar essa ligeira tendência. (Uma vez que as encomendas do mês de março de 2023 terminam no dia 6, esses dados foram excluídos dos visuais.) Verifica-se uma quebra nas vendas (quantidade e encomendas) no mês de agosto, o que pode ser explicado por ser um mês caracterizado pelas férias.

Evolução mensal de itens vendidos



Evolução mensal de encomendas

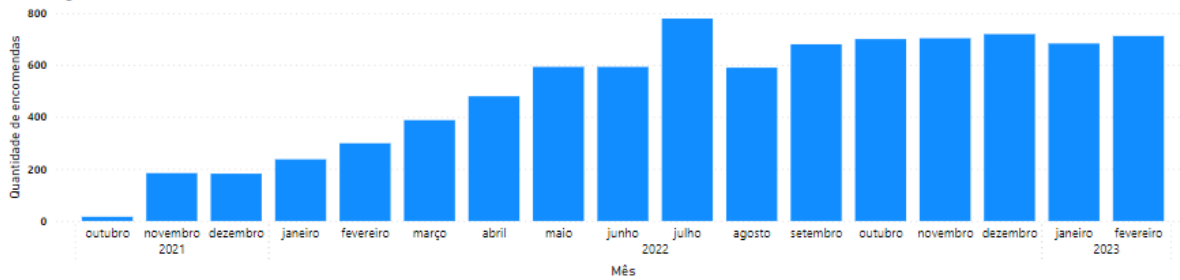
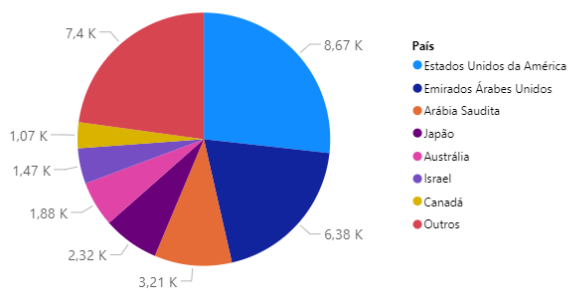


Gráfico 4.2. – Evolução mensal do volume de vendas

Quanto à localização dos compradores (mais precisamente, localização do envio), verifica-se que os países com maior quantidade e encomendas são Estados Unidos da América, Emirados Árabes Unidos e Arábia Saudita. É interessante que, apesar de se tratar de uma empresa que envia encomendas a partir de Portugal, apenas cerca de 3% das encomendas e 2% das quantidades são enviadas para Portugal, com mais de metade do volume de negócio a corresponder a países da América ou Médio Oriente.

Itens vendidos por país



Encomendas por país

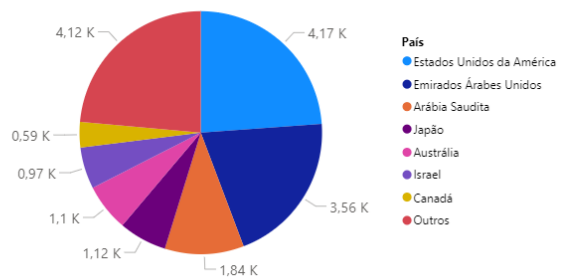


Gráfico 4.3. – Volume de negócios por país

Analisando por continentes, naturalmente os continentes com mais vendas são a Ásia e a América. As vendas para África são muito pouco relevantes e com alguma surpresa a Europa tem pequena expressão para esta empresa portuguesa.

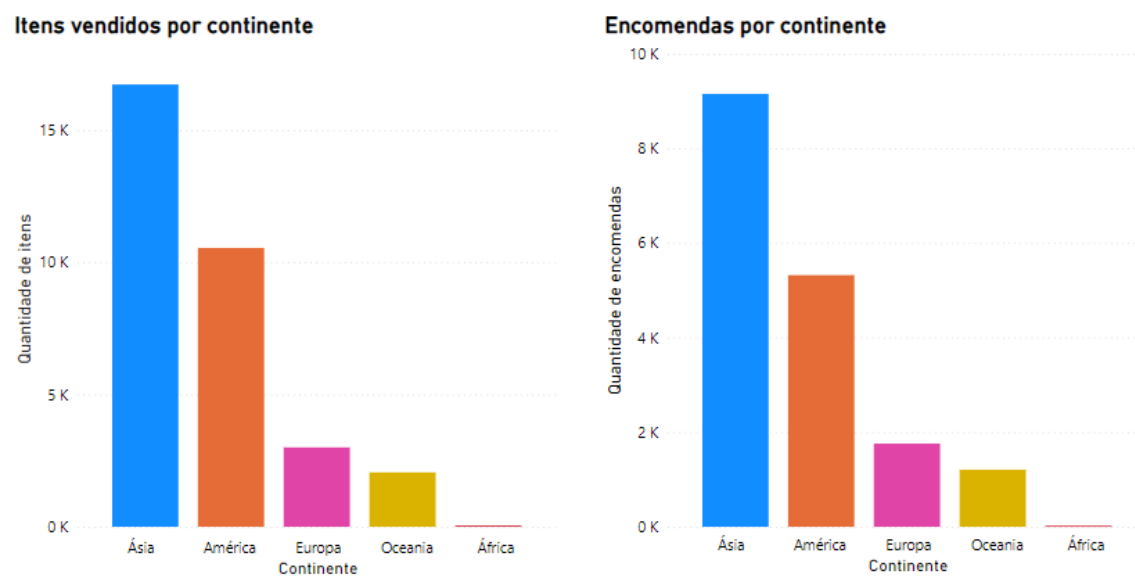


Gráfico 4.4. – Volume de negócios por continente

As compras não apresentam grande variação ao longo da semana, sendo o domingo o dia em que menos compras são feitas e a segunda-feira o dia com mais atividade por parte dos consumidores.

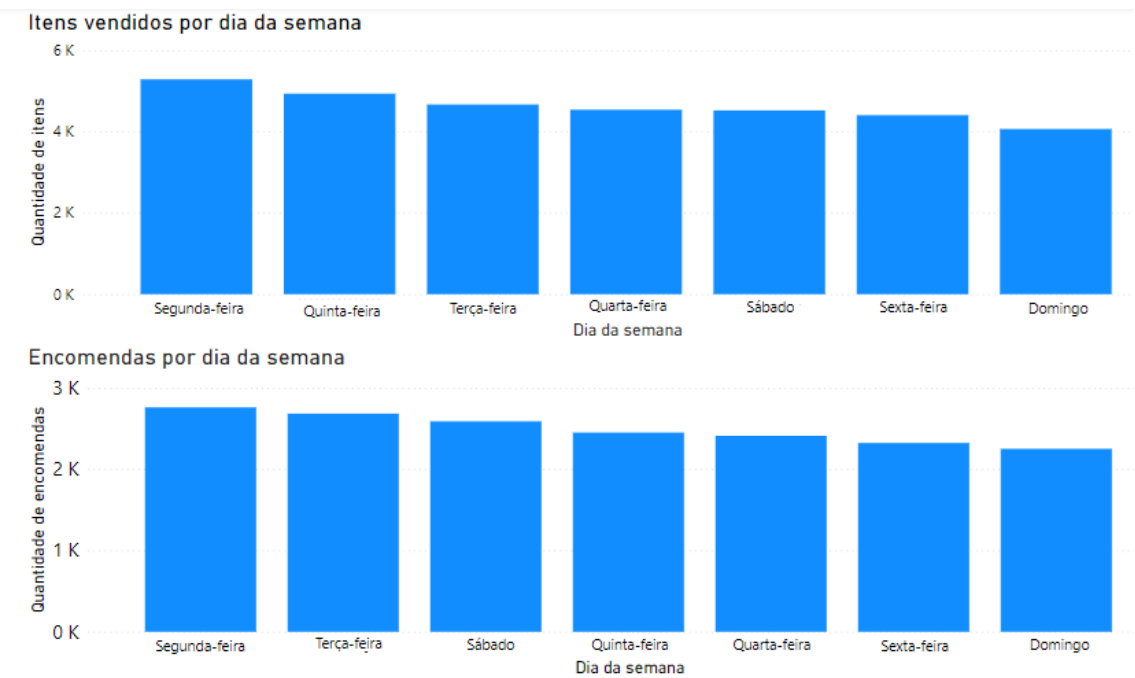


Gráfico 4.5. – Volume de negócios por dia da semana

Comparando as compras de clientes registados com as de clientes não registados, verifica-se que os primeiros tendem a realizar compras com maiores quantidades, 5,5 unidades por encomenda (17580 itens em 3210 encomendas) contra 2,8 (14824 itens

em 5290 encomendas), pelo que seria benéfico promover o registo de clientes, possivelmente oferecendo algum tipo de vantagem em contrapartida. Tal como referido anteriormente, é útil para o vendedor (e muitas vezes para o cliente) obter alguma informação do cliente, de forma a manter contacto com ele e apresentar oportunidades de negócio.

4.4. Análise de Clusters

Os dados transformados no Power BI foram exportados para um ficheiro de valores separados por vírgulas do Microsoft Excel, *Orders_BI.csv*, que foi de seguida importado para o programa de análise de dados *Orange*. Aí foram realizadas análises de *Data Mining*, nomeadamente Análise de Clusters e Regras de Associação.

No que à Análise de Clusters diz respeito, foram usados diferentes tipos de clustering (Hierárquico e DBSCAN) disponíveis no *Orange*, o que permitiu agrupar os dados em clusters, como a seguir se descreve.¹¹

4.4.1. Clustering Hierárquico

Importando o ficheiro *Orders_BI.csv* para o *Orange*, realizou-se o clustering hierárquico, com a distância euclidiana, obtendo-se os dendrogramas com diferentes métodos de ligação.

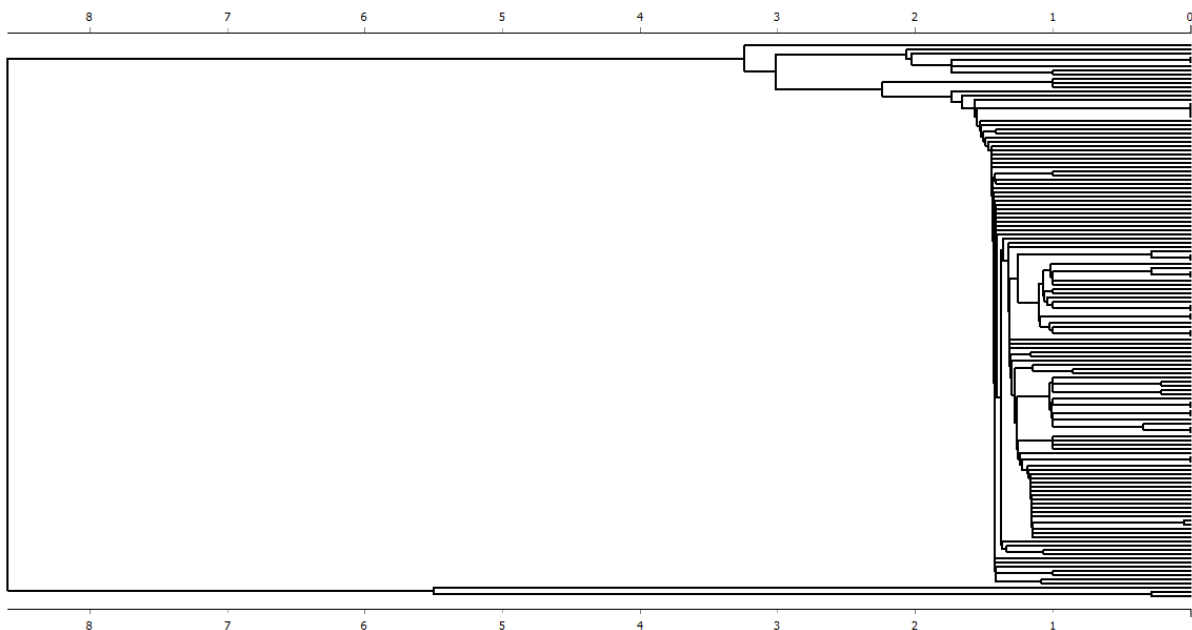


Gráfico 4.6. – Dendrograma (ligação simples)

¹¹ Como explicado anteriormente, optou-se por não realizar o clustering K-Médias. Apesar de ser um dos tipos de clustering mais populares e estar disponível no software *Orange*, não se adequava ao conjunto de dados, o que ditou a sua não utilização.

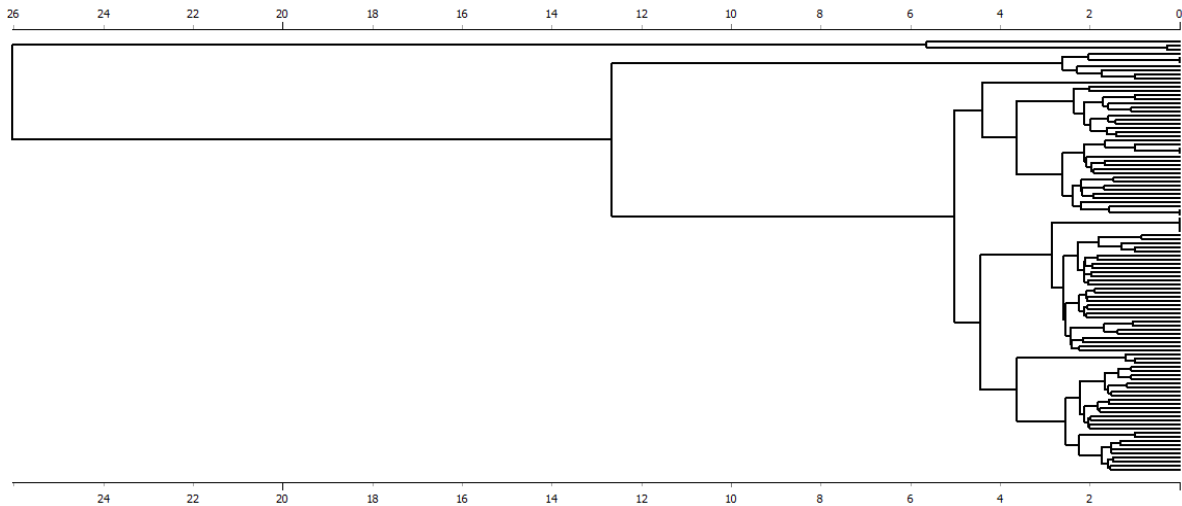


Gráfico 4.7. – Dendrograma (ligação média)

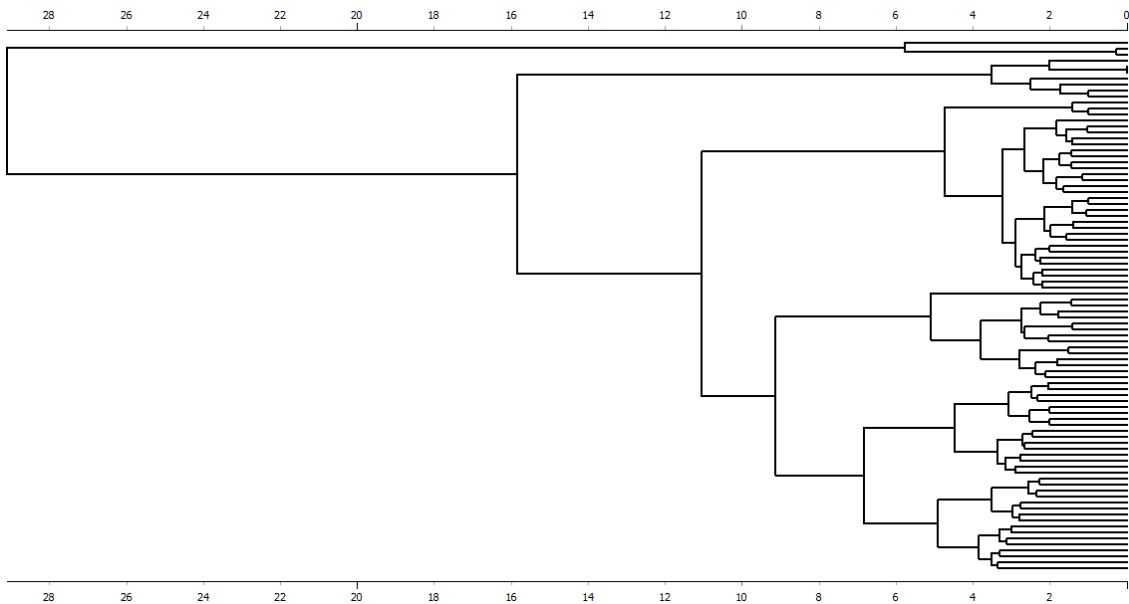


Gráfico 4.8. – Dendrograma (ligação completa)

Apesar do aspeto dos dendrogramas, que podem criar a ilusão (principalmente nos casos da ligação média e da ligação completa) de clusters de tamanhos relativamente equilibrados, na realidade os clusters obtidos através do método de ligação simples, média e completa têm tamanhos muito desequilibrados, com mais de 90% dos pontos num só cluster. Depois de analisadas várias possibilidades de ligação e de número de clusters, optou-se pelo método de ligação Ward, que resultou em clusters com tamanhos mais equilibrados. Assim, a partir daqui a análise será feita ao resultado obtido pelo método de ligação Ward, com 6 clusters. Foi escolhido este número de clusters por se tratar de um número próximo do obtido no Clustering K-Médias e

resultar em clusters com tamanhos mais equilibrados do que nos outros casos analisados.

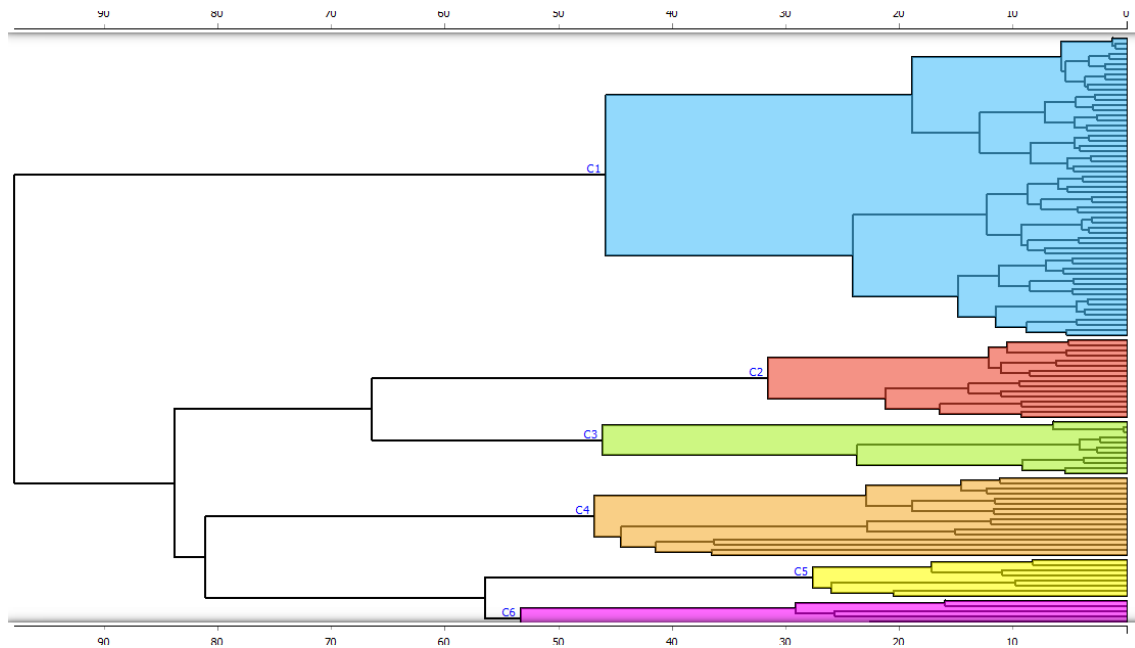


Gráfico 4.9. – Dendrograma (ligação Ward, 6 clusters)

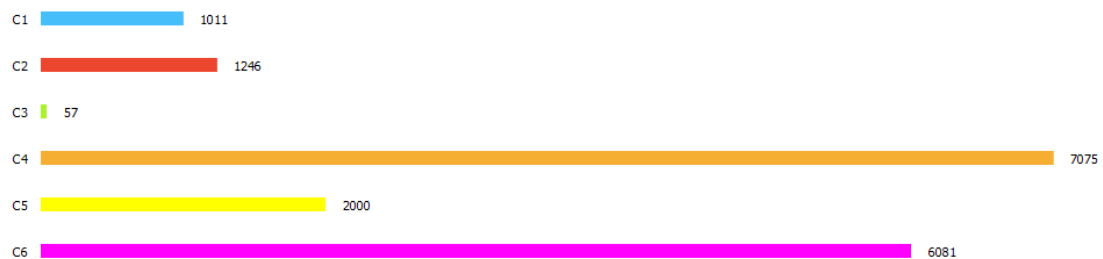


Gráfico 4.10. – Tamanho dos 6 clusters obtidos por Clustering Hierárquico

Os clusters 4 e 6 têm tamanhos relativamente próximos, os clusters 1, 2 e 5 têm um tamanho inferior, enquanto o cluster 3 diz respeito a um número muito reduzido de itens adquiridos.

De seguida, foi estudada, através de diagramas de dispersão e de barras, a forma como as variáveis mais relevantes influenciam a atribuição de cada ponto de dados ao respetivo cluster.

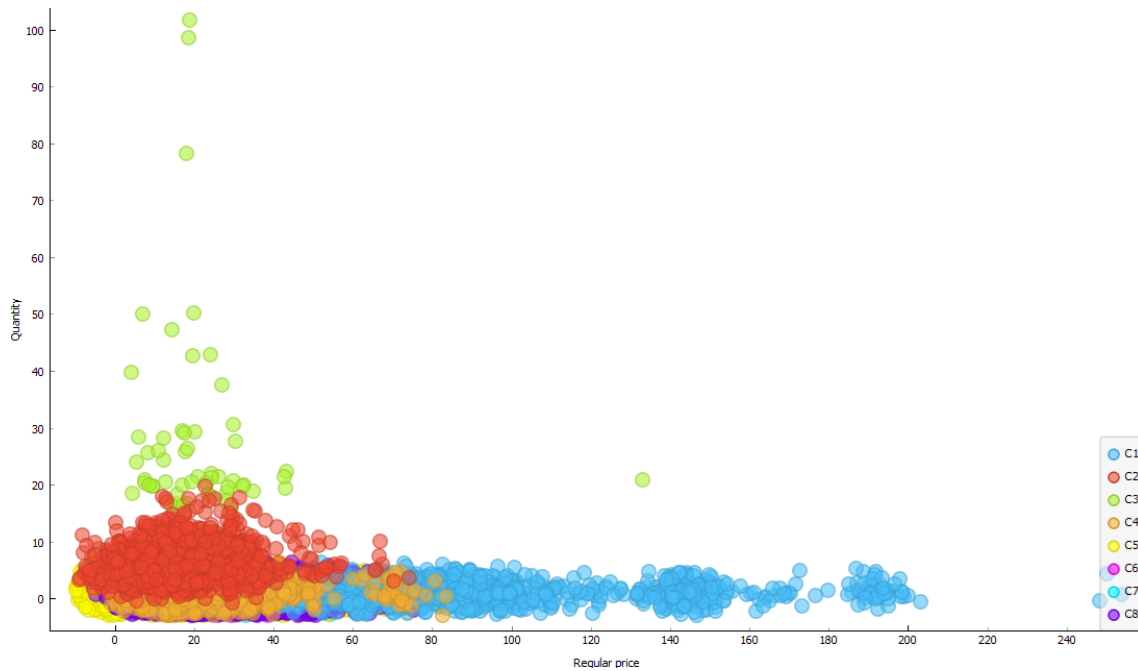


Gráfico 4.11. – Diagrama de dispersão (Regular Price - Quantity)

O cluster 3 caracteriza-se por quantidades elevadas, enquanto o cluster 2 por quantidades um pouco mais elevadas do que os restantes, com exceção do cluster 3. O cluster 1 corresponde a preços mais elevados, enquanto o cluster 6 a preços um pouco mais elevados do que os restantes, com exceção do cluster 1.

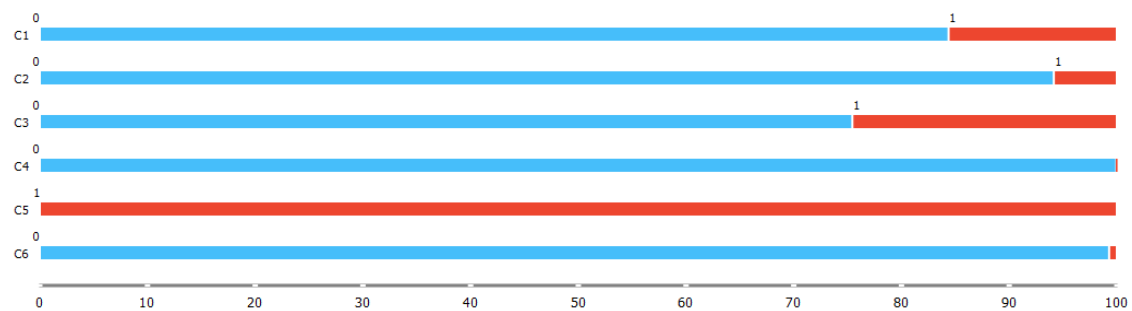


Gráfico 4.12. – Clusters e Makeup (Y/N)

O cluster 5 corresponde a itens adquiridos que estão na categoria Makeup. Os clusters 4 e 6 (quase na sua totalidade) dizem respeito a itens adquiridos que não estão na categoria Makeup. Nos restantes clusters não se verifica claro domínio no que à categoria Makeup diz respeito.

As restantes variáveis não aparentam ter particular impacto na definição dos clusters.

Em resumo, algumas das características dos clusters são as seguintes:

Cluster	Características
1	Preços elevados
2	Quantidades ligeiramente superiores à média
3	Quantidades elevadas
4	Produto não Makeup
5	Produto Makeup
6	Preços ligeiramente superiores à média, produto não Makeup

Tabela 4.3. – Características dos clusters obtidos por Clustering Hierárquico, ligação Ward

4.4.2. Clustering DBSCAN

Importando o ficheiro Orders_BI.csv para o *Orange*, é possível realizar o Clustering DBSCAN. No entanto, a escolha dos parâmetros (*minPts* e ϵ) não se revelou fácil, pois se obtinha um cluster com mais de 90% do conjunto de dados ou um elevado número de clusters.

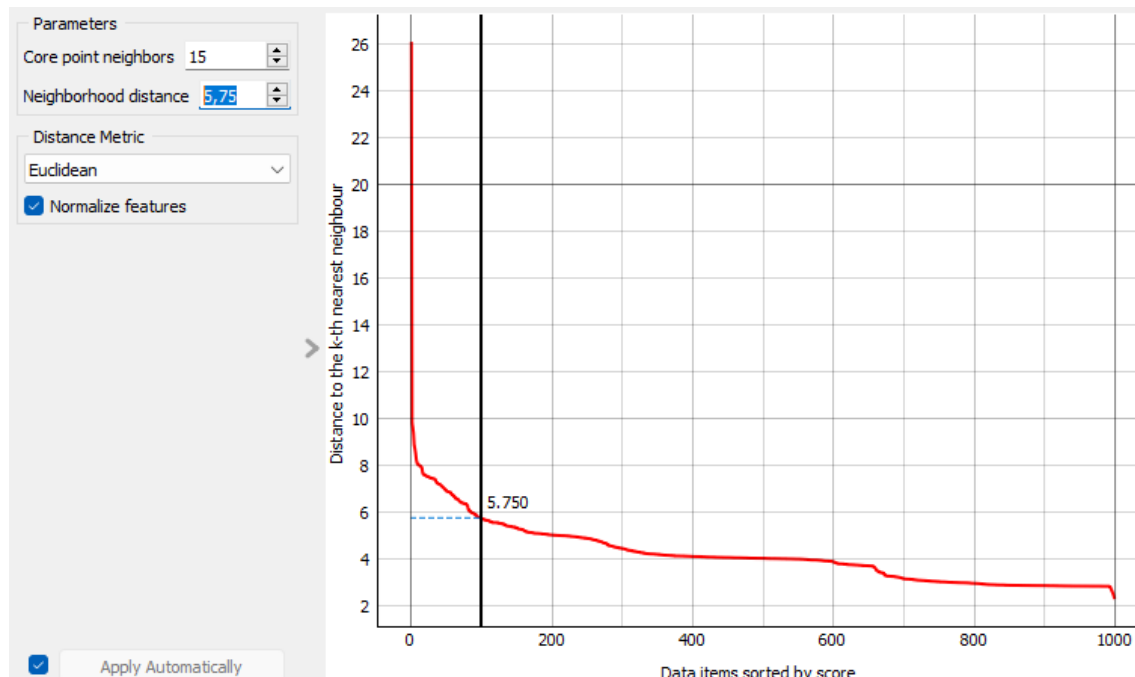


Gráfico 4.13. – Distância ao k-vizinho mais próximo para *minPts* = 15

No entanto, selecionando *minPts* = 15 (segundo uma das recomendações anteriormente apresentadas), o gráfico que apresenta a distância ao k-vizinho mais próximo sugere $\epsilon = 5,75$.

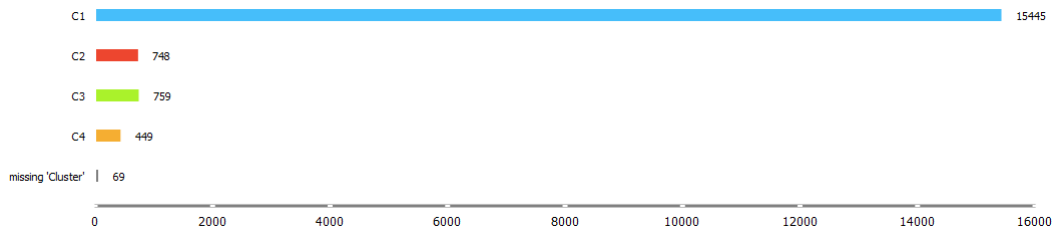


Gráfico 4.14. – Tamanho dos 4 clusters obtidos por Clustering DBSCAN

O clustering resulta num cluster dominante, 3 clusters mais pequenos e alguns pontos de ruído, sendo um pouco mais difícil tirar conclusões particularmente relevantes.

De seguida, foi estudada, através de diagramas de dispersão e de barras, a forma como as variáveis mais relevantes influenciam a atribuição de cada ponto de dados ao respetivo cluster.

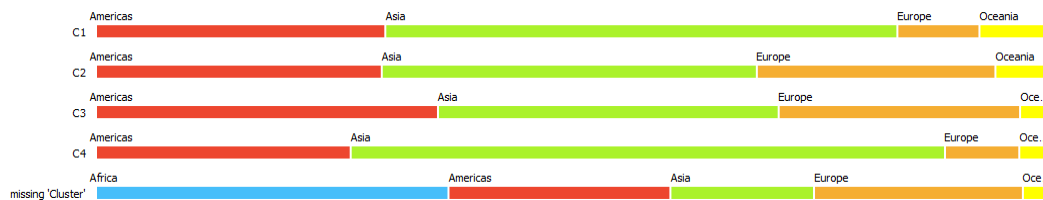


Gráfico 4.15. – Distribuição dos continentes por cluster

Todas as encomendas de África são consideradas ruído.



Gráfico 4.16. – Distribuição de Health (Y/N) por cluster

Os clusters 1, 3 e 4 corresponde a itens adquiridos que não estão na categoria Health.

O cluster 2 diz respeito a itens adquiridos que estão na categoria Health.



Gráfico 4.17. – Distribuição de Men (Y/N) por cluster

Os clusters 1, 2 e 3 corresponde a itens adquiridos que não estão na categoria Men. O cluster 4 diz respeito a itens adquiridos que estão na categoria Men.

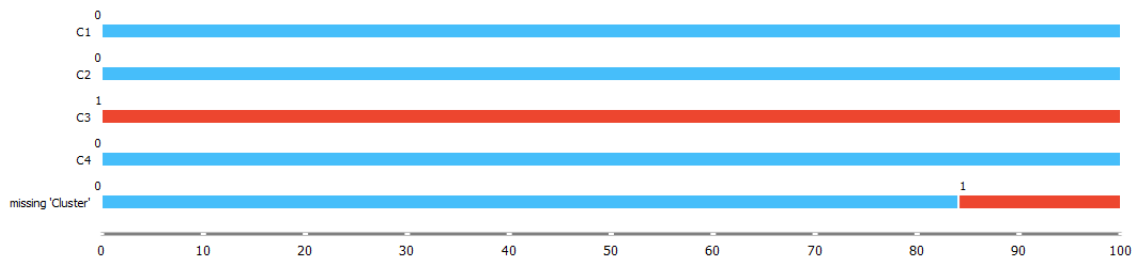


Gráfico 4.18. – Distribuição de Oral (Y/N) por cluster

Os clusters 1, 2 e 4 corresponde a itens adquiridos que não estão na categoria Oral. O cluster 3 diz respeito a itens adquiridos que estão na categoria Oral.

Algumas das características dos clusters são as seguintes:

Os clusters 3 e 4 corresponde a itens adquiridos que não estão na categoria Health.

Os clusters e 3 corresponde a itens adquiridos que não estão na categoria Men. O cluster 4 diz respeito a itens adquiridos que estão na categoria Men.

Os clusters 4 corresponde a itens adquiridos que não estão na categoria Oral. O cluster 3 diz respeito a itens adquiridos que estão na categoria Oral.

Cluster	Características
1	Nenhum produto de Health, Men ou Oral
2	Todos os produtos de Health, nenhum de Men ou Oral
3	Todos os produtos de Oral, nenhum de Health ou Men
4	Todos os produtos de Men, nenhum de Health ou Oral
ruído	Todas as encomendas de África são ruído

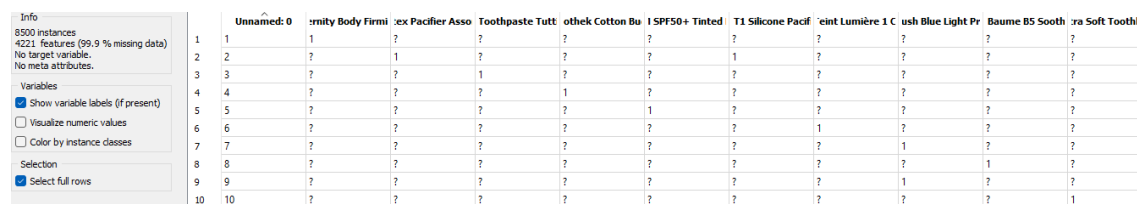
Tabela 4.4. – Características dos clusters obtidos por Clustering DBSCAN

4.4.3. Conclusão

Tal como referido anteriormente, é recomendável utilizar diferentes tipos de clustering para poder tirar conclusões mais abrangentes. Utilizando vários tipos de clustering, foi possível verificar que as variáveis Regular Price, Quantity e Skincare causaram alguma divisão nos clusters através do clustering Hierárquico (ligação Ward), enquanto no clustering DBSCAN, foram as variáveis Health, Men e Oral que se revelaram mais relevantes na definição dos clusters. Como referido anteriormente, a natureza dos dados levou a que não fosse utilizado o clustering K-Médias. Não parece óbvio, perante os resultados obtidos, que recomendações possam ser dadas à empresa de forma a melhorar o seu desempenho.

4.5. Regras de Associação

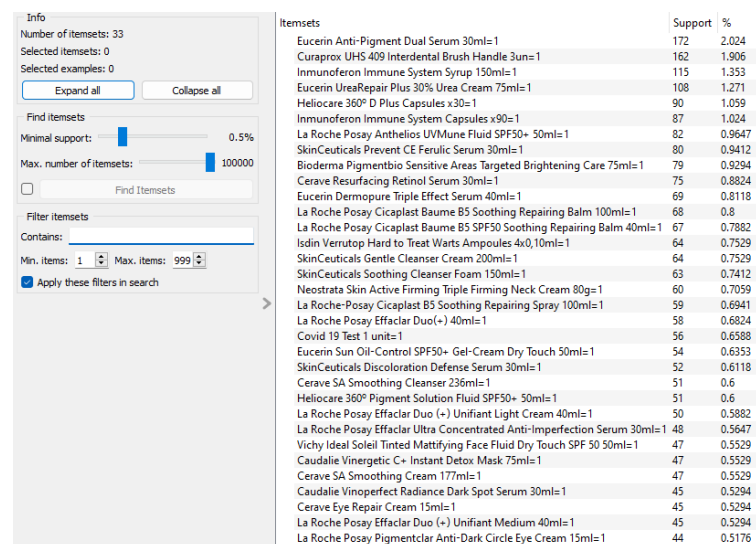
O ficheiro Orders_BI.csv foi importado para o R e aí transformado de forma que cada linha correspondesse a uma encomenda e cada coluna a um dos produtos vendidos. Uma célula igual a 1 significa que a encomenda inclui o produto e uma célula vazia significa que a encomenda não inclui o produto. O ficheiro resultante, Orders_Orange_Product.csv, foi importado para o *Orange*, onde foi realizada a análise de Regras de Associação. Registaram-se 8500 encomendas e 4221 produtos distintos.



	Unnamed: 0	Smiley Body Firmi	Sex Pacifier Asso	Toothpaste Tutt	othek Cotton Bu	SPF50+ Tinted	T1 Silicone Pacif	oint Lumière 1 C	ush Blue Light Pr	Baume B5 Sooth	ra Soft Toothb
1	1	1	?	?	?	?	?	?	?	?	?
2	2	?	1	?	?	?	1	?	?	?	?
3	3	?	?	1	?	?	?	?	?	?	?
4	4	?	?	?	1	?	?	?	?	?	?
5	5	?	?	?	?	1	?	?	?	?	?
6	6	?	?	?	?	?	1	?	?	?	?
7	7	?	?	?	?	?	?	1	?	?	?
8	8	?	?	?	?	?	?	?	1	?	?
9	9	?	?	?	?	?	?	1	?	?	?
10	10	?	?	?	?	?	?	?	?	1	?

Figura 4.1. – Visão parcial do ficheiro Orders_Orange_Product.csv importado para o *Orange*

Inicialmente, verificaram-se quais os conjuntos de itens frequentes. Escolhendo um suporte mínimo de 0,5%, apenas se obtiveram 33 conjuntos de itens frequentes, mas todos com um único elemento, que não têm interesse no estabelecimento de regras de associação.



Itemsets	Support	%
Eucerin Anti-Pigment Dual Serum 30ml=1	172	2.024
Curaprox UHS 409 Interdental Brush Handle 3un=1	162	1.906
Immunoforon Immune System Syrup 150ml=1	115	1.353
Eucerin UreaRepair Plus 30% Urea Cream 75ml=1	108	1.271
Heliocare 360° D Plus Capsules x30=1	90	1.059
Immunoforon Immune System Capsules x90=1	87	1.024
La Roche Posay Anthelios UVMune Fluid SPF50+ 50ml=1	82	0.9647
SkinCeuticals Prevent CE Ferulic Serum 30ml=1	80	0.9412
Bioderma Pigmentbio Sensitive Areas Targeted Brightening Care 75ml=1	79	0.9294
Cerave Resurfacing Retinol Serum 30ml=1	75	0.8824
Eucerin Dermopure Triple Effect Serum 40ml=1	69	0.8118
La Roche Posay Cicaplast Baume B5 Soothing Repairing Balm 100ml=1	68	0.8
La Roche Posay Cicaplast Baume B5 SPF50 Soothing Repairing Balm 40ml=1	67	0.7882
Isdin Verrutop Hard to Treat Warts Ampoules 4x0,10ml=1	64	0.7529
SkinCeuticals Gentle Cleanser Cream 200ml=1	64	0.7529
SkinCeuticals Soothing Cleanser Foam 150ml=1	63	0.7412
Neostrata Skin Active Firming Triple Firming Neck Cream 80g=1	60	0.7059
La Roche-Posay Cicaplast B5 Soothing Repairing Spray 100ml=1	59	0.6941
La Roche Posay Effaclar Duo(+) 40ml=1	58	0.6824
Covid 19 Test 1 unit=1	56	0.6588
Eucerin Sun Oil-Control SPF50+ Gel-Cream Dry Touch 50ml=1	54	0.6353
SkinCeuticals Discoloration Defense Serum 30ml=1	52	0.6118
Cerave SA Smoothing Cleanser 236ml=1	51	0.6
Heliocare 360° Pigment Solution Fluid SPF50+ 50ml=1	51	0.6
La Roche Posay Effaclar Duo (+) Unifiant Light Cream 40ml=1	50	0.5882
La Roche Posay Effaclar Ultra Concentrated Anti-Imperfection Serum 30ml=1	48	0.5647
Vichy Ideal Soleil Tinted Mattifying Face Fluid Dry Touch SPF 50 50ml=1	47	0.5529
Caudalie Vinergic C+ Instant Detox Mask 75ml=1	47	0.5529
Cerave SA Smoothing Cream 177ml=1	47	0.5529
Caudalie Vinopure Radiance Dark Spot Serum 30ml=1	45	0.5294
Cerave Eye Repair Cream 15ml=1	45	0.5294
La Roche Posay Effaclar Duo (+) Unifiant Medium 40ml=1	45	0.5294
La Roche Posay Pigmentclar Anti-Dark Circle Eye Cream 15ml=1	44	0.5176

Figura 4.2. – Conjuntos de itens frequentes para suporte mínimo 0,5%

Por essa razão, o foi testado o suporte mínimo de 0,1%, o que permitiu obter 464 conjuntos de itens frequentes com 1 elemento, 58 com 2 elementos, 24 com 3 elementos, 7 com 4 elementos e 1 com 5 elementos.

Info

Number of itemsets: 464
Selected itemsets: 0
Selected examples: 0

Expand all

Collapse all

Find itemsets

Minimal support: 0.1%

Max. number of itemsets: 100000

Find Itemsets

Filter itemsets

Contains:

Min. items: 1 Max. items: 999

☒ Apply these filters in search

Itemsets

	Support	%
✓ Eucerin Anti-Pigment Dual Serum 30ml=1	172	2.024
✓ Eucerin Anti-Pigment Day Cream SPF30 50ml=1	20	0.2353
Eucerin Anti-Pigment Night Cream 50ml=1	10	0.1176
Eucerin Anti-Pigment Night Cream 50ml=1	19	0.2235
Curaprox UHS 409 Interdental Brush Handle 3un=1	162	1.906
Immunoforon Immune System Syrup 150ml=1	115	1.353
✓ Eucerin UreaRepair Plus 30% Urea Cream 75ml=1	108	1.271
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	23	0.2706
✓ Eucerin DermatoCLEAN [HYALURON] Toner 200ml=1	14	0.1647
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	13	0.1529
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	13	0.1529
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	13	0.1529
✓ Intea Chamomile Natural Blonde Lotion 100ml=1	12	0.1412
✓ Eucerin DermatoCLEAN [HYALURON] Toner 200ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	11	0.1294
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	12	0.1412
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	11	0.1294
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	13	0.1529
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	13	0.1529

Figura 4.3. – Alguns dos conjuntos de itens frequentes com pelo menos 1 elemento para suporte mínimo 0,1%

Info

Number of itemsets: 58

Selected itemsets: 0

Selected examples: 0

Expand all

Collapse all

Find itemsets

Minimal support: 0.1%

Max. number of itemsets: 100000

☐

Find Itemsets

Filter itemsets

Contains:

Min. items: 2

Max. items: 999

☒ Apply these filters in search

Itemsets

	Support	%
✓ La Roche Posay Effaclar Duo (+) Unifiant Light Cream 40ml=1	17	0.2
La Roche Posay Effaclar Duo (+) Unifiant Medium 40ml=1	17	0.2
✓ Cerave SA Smoothing Cleanser 236ml=1	10	0.1176
Cerave Foaming Cleanser 236ml=1	10	0.1176
Cerave Moisturizing Lotion 473ml=1	9	0.1059
✓ Cerave Hydrating Facial Cleanser 8FL Oz/236ml=1	10	0.1176
Cerave Foaming Cleanser 236ml=1	9	0.1059
✓ Cerave Eye Repair Cream 15ml=1	9	0.1059
Uniaque Power 3 Anti-Perspirant Deodorant Duo=1	9	0.1059
✓ Eucerin Anti-Pigment Dual Serum 30ml=1	20	0.2353
✓ Eucerin Anti-Pigment Day Cream SPF30 50ml=1	20	0.2353
Eucerin Anti-Pigment Night Cream 50ml=1	10	0.1176
Eucerin Anti-Pigment Night Cream 50ml=1	19	0.2235
✓ Cerave SA Smoothing Cream 177ml=1	9	0.1059
Cerave SA Smoothing Cream 340gr=1	9	0.1059
✓ Bioderma Atoderm Shower Oil 1000ml=1	9	0.1059
Avene Tricera Nutrition Duo Balm Nutri Fluid=1	9	0.1059

Figura 4.4. – Alguns dos conjuntos de itens frequentes com pelo menos 2 elementos para suporte mínimo 0,1%

Info

Number of itemsets: 24

Selected itemsets: 0

Selected examples: 0

Expand allCollapse all

Find itemsets

Minimal support: 0.1%

Max. number of itemsets: 100000

Find Itemsets

Filter itemsets

Contains:

Min. items: 3Max. items: 999

Apply these filters in search

Itemsets

	Support	%
✓ Eucerin Anti-Pigment Dual Serum 30ml=1	10	0.1176
✓ Eucerin Anti-Pigment Day Cream SPF30 50ml=1	10	0.1176
Eucerin Anti-Pigment Night Cream 50ml=1	10	0.1176
✓ Eucerin UreaRepair Plus 30% Urea Cream 75ml=1	13	0.1529
✓ Eucerin DermatoCLEAN [HYALURON] Toner 200ml=1	13	0.1529
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	13	0.1529
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	13	0.1529
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	13	0.1529
Intea Chamomile Natural Blonde Lotion 100ml=1	11	0.1294
✓ Eucerin DermatoCLEAN [HYALURON] Toner 200ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	11	0.1294
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	12	0.1412
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	11	0.1294
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	11	0.1294
✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1	13	0.1529
Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1	13	0.1529

Figura 4.5. – Alguns dos conjuntos de itens frequentes com pelo menos 3 elementos para suporte mínimo 0,1%

Info		Itemsets		Support	%
Number of itemsets: 7		✓ Eucerin UreaRepair Plus 30% Urea Cream 75ml=1		11	0.1294
Selected itemsets: 0		Intea Chamomile Natural Blonde Lotion 100ml=1		11	0.1294
Selected examples: 0		✓ Eucerin DermatoCLEAN [HYALURON] Toner 200ml=1		11	0.1294
Expand all Collapse all		Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1		11	0.1294
Find itemsets		✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1		11	0.1294
Minimal support: 0.1%		Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1		11	0.1294
Max. number of itemsets: 100000		✓ Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1		11	0.1294
☐ Find Itemsets		Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1		11	0.1294
Filter itemsets		Intea Chamomile Natural Blonde Lotion 100ml=1		11	0.1294
Contains:		✓ Eucerin DermatoCLEAN [HYALURON] Toner 200ml=1		11	0.1294
Min. items: 4 Max. items: 999		Eucerin Hyaluron Filler + Elasticity SPF15 Eye Cream 15ml=1		11	0.1294
☒ Apply these filters in search		Eucerin UreaRepair PLUS 10% Urea Lotion 400ml=1		11	0.1294
		Bioderma Cicabio Soothing Repairing Cream 100ml=1		15	0.1765
		A-Derma Epitheliale AH Ultra Soothing Repairing Cream 40ml=1		15	0.1765
		Avene Cicafate Skin Repair Emulsion Post-Procedure 40ml=1		15	0.1765
		A-Derma Epitheliale AH Ultra Protective Repairing Cream SPF50+ 40ml=1		15	0.1765

Figura 4.6. – Alguns dos conjuntos de itens frequentes com pelo menos 4 elementos para suporte mínimo 0,1%

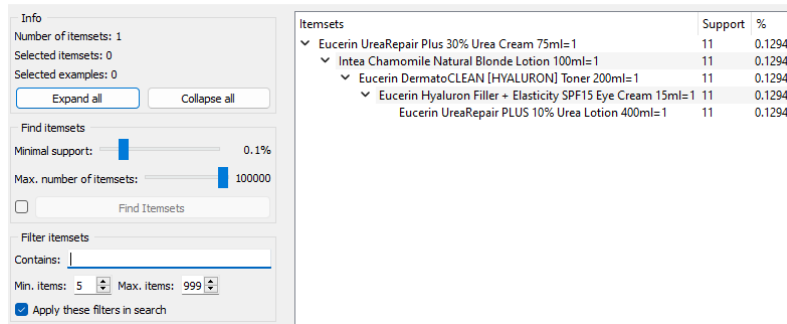


Figura 4.7. – Conjuntos de itens frequentes com pelo menos 5 elementos para suporte mínimo 0,1%

De seguida, foram determinadas regras de associação que satisfizessem o suporte mínimo de 0,1% e confiança mínima de 80%, obtendo assim 140 regras. Filtrando essas regras para aquelas que têm pelo menos 2 elementos no antecedente e 2 no consequente (que resultam dos conjuntos de itens com pelo menos 4 elementos), restam 46 regras.



Figura 4.8. – Algumas das regras de associação para suporte mínimo 0,1% e confiança mínima de 80%

Estas regras permitem identificar quais os produtos que são mais vezes vendidos em conjunto e podem ser tidas em conta na definição de campanhas promocionais ou outras estratégias de venda.

Apesar de estas regras terem uma confiança elevada, o suporte é bastante baixo, o que é uma consequência de haver muitos produtos diferentes em relação ao número de encomendas. Também o *leverage* é pouco maior do que 0, o que indica que, apesar de as regras funcionarem com grande fiabilidade, acabam por ser pouco expressivas no conjunto global de encomendas.

Por esta razão, e apenas para testar o estudo das regras de associação numa situação com menos colunas, foi realizado a mesma análise com marcas em vez de produtos. Aí foi possível encontrar regras com maior suporte, mas com confiança e *leverages* mais baixos.

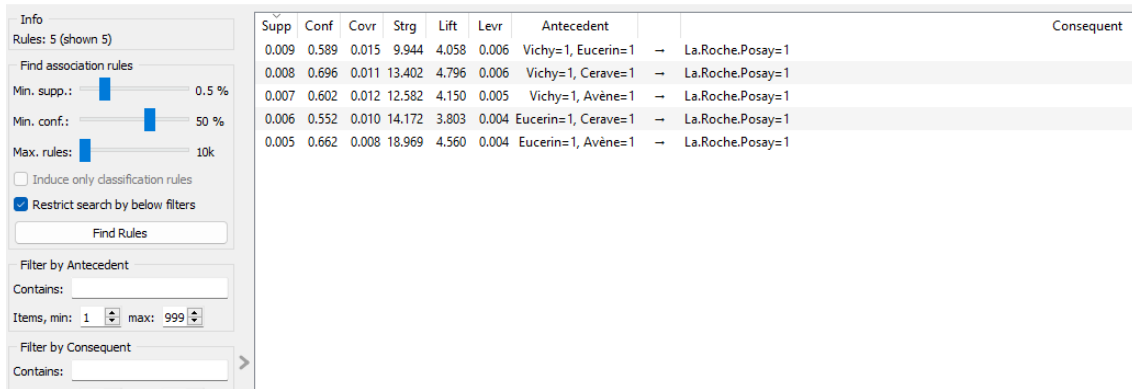


Figura 4.9. – Algumas das regras de associação (marcas) para suporte mínimo 0,5% e confiança mínima de 50%

Desta forma, foram determinadas 5 regras com suporte mínimo de 0,5% e confiança mínima de 50%, que permitem identificar marcas que são frequentemente encontradas numa mesma encomenda.

A lista de transações em Orders_BI.csv continha informação que, depois de devidamente tratada, pode ajudar a tirar algumas conclusões sobre os produtos (e também marcas) mais frequentemente adquiridos em conjunto. Os baixos valores de suporte, que se explicam pelo elevado número de produtos em comparação com as encomendas (o número de encomendas são apenas cerca do dobro do número de produtos) não permite obter regras que funcionem com grande frequência, o que torna difícil apresentar sugestões de ação com uma eficiência significativa.

5. Conclusão

Com este relatório, pretendeu-se abordar a temática do *e-commerce*, um modelo de negócio em franca expansão, que se tem vindo a adaptar a outras inovações sociais e tecnológicas, sendo hoje um meio quase incontornável para a realização de compras.

Foi apresentado um breve contexto histórico do *e-commerce*, bem como alguns desafios e vantagens deste modelo de negócio.

Tendo em conta os dados de vendas de uma empresa de *e-commerce* na área da cosmética, procedeu-se à limpeza, tratamento, apresentação e análise de dados. A limpeza e o tratamento dos dados foram realizados no Power Query do Power BI e permitiram seleccionar e transformar as variáveis relevantes nos passos seguintes do trabalho.

Ainda no Power BI, foram construídos diferentes visuais que permitiram obter uma melhor compreensão dos dados das vendas, ajudando a encontrar algumas sugestões de ação no mercado.

Importando os dados trabalhados no Power BI para o *Orange*, realizaram-se várias análises, nomeadamente Análise de Clusters e Regras de Associação. Estas ferramentas de *Data Mining* foram objeto de uma fundamentação teórica, em particular o Clustering DBSCAN e as Regras de Associação (uma vez que não fizeram parte dos conteúdos lecionados ao longo do Mestrado), que permitiram encontrar relações entre transações e entre produtos, o que pode ajudar a identificar segmentos de clientes e de produtos.

A realização do estágio de forma remota trouxe desafios, como a diminuição de interação humana, ausência de trabalho de equipa, e limitações na comunicação e obtenção de feedback quase imediato, mas tal foi sendo mitigado na medida do possível através de encontros regulares com o orientador Óscar Felgueiras e contactos telefónicos com o supervisor Vítor Ferreira.

Inicialmente pretendia-se estudar modelos de *pricing*, mas tal objetivo acabou por ser reformulado devido à escassez de alguma informação necessária e ao facto de nem sempre ser possível saber em que moeda a transação foi realizada. Por essa razão, mas também devido à complexidade do tema em questão, optou-se por Análise de Clusters e Regras de Associação.

Tendo consciência de que nem sempre os dados nos chegam com a qualidade e organização desejável (cabendo ao analista de dados tirar a maior informação

possível dos mesmos), algumas melhorias na qualidade dos dados trariam um valor acrescentado à análise, nomeadamente uma melhor organização, preenchendo informação em falta e uniformizando a moeda da compra, de forma a poder analisar melhor o volume de negócios e obter propostas de atuação mais fundamentadas.

Quanto a sugestões de trabalho futuro, teria interesse o estudo de algoritmos de sistemas de recomendação ou de modelos de *pricing*, de forma a permitir uma melhor experiência de compra para os clientes e otimizar vendas e lucros para os comerciantes. Para tal seria útil recorrer a mais informação, como a uniformização dos preços obtidos na venda de cada produto/transação e dados de clientes mais pormenorizados.

Os resultados estão longe de ter uma clara e imediata aplicação que conduza a uma melhoria no desempenho das vendas, mas o caminho acabou por ser tão ou mais compensador do que o destino. Permitiu perceber como funcionam certos programas, softwares e algoritmos desconhecidos até à realização do estágio, identificar lacunas próprias na compreensão de certos conceitos e ferramentas e crescer enquanto pessoa e analista de dados.

Este estágio permitiu desenvolver num contexto empresarial algumas das técnicas lecionadas durante o Mestrado, mas também algumas ferramentas não abordadas anteriormente. Assim, trouxe um valor acrescentado à formação académica enquanto analista de dados, tornando mais clara a forma como o trabalho nesta área é realizado fora de um contexto curricular. Resta dizer que o trabalho foi mais enriquecedor com a colaboração, incentivo, sugestões e supervisão do supervisor e do orientador de estágio.

Referências Bibliográficas

- Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*, 207.
- Al-Lami, G., & Alnoor, A. (2021). E-Commerce: Advantages and Limitations. *International Journal of Academic Research in Accounting Finance and Management Sciences* 11(1), 153-165.
- Barbara, D. (2000). *An Introduction to Cluster Analysis for Data Mining*. Obtido de https://www-users.cse.umn.edu/~hanxx023/dmclass/cluster_survey_10_02_00.pdf
- Campello, R., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. *Advances in Knowledge Discovery and Data Mining. Vol. 7819*, 160–172.
- Ester, M., Kriegel, H.-P., Jörg, S., & Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. Obtido de <https://cdn.aaai.org/KDD/1996/KDD96-037.pdf>
- Gama, J., Carvalho, A., Facelli, K., Lorena, A., & Oliveira, M. (2017). *Extração de Conhecimento de Dados*. Edições Sílabo.
- J. Handl, J. K. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15), 3201-3212.
- Jiawei, H., Jian, P., & Yiwen, Y. (2000). Mining frequent patterns without candidate generation. *ACM SIGMOD Record, Volume 29, Issue 2*, 1–12.
- Kirenz, J. (2020). *Introduction to Association Rule Mining in R*. Obtido de <https://www.kirenz.com/post/2020-05-14-r-association-rule-mining>
- Kumar, S. L. (2021). 20 years of Electronic Commerce Research. *Electron Commer Res* 21, 1–40.
- Ng, A., & Soo, K. (2017). *Numsense! Data Science for the Layman: No Math Added*. Leanpub.
- Qayum, A. (2023). *The future of ecommerce: How ecommerce will change in 2024*. Obtido de <https://www.oberlo.com/blog/future-of-ecommerce>

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*. 20, 53–65.

Saaslyft. (s.d.). 10 Objectives of Ecommerce Businesses. Obtido de <https://saaslyft.com/blog/objectives-of-ecommerce-businesses/>

Schubert, E., Sander, J., Martin, E., Kriegel, H.-P., & Xu, X. (2017). DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.* 42 (3), 19:1–19:21. Obtido de DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN.

Vassilvitskii, S. (s.d.). *K-means++: The Advantages of Careful Seeding*. Obtido de <http://theory.stanford.edu/~sergei/slides/BATS-Means.pdf>