INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Sistema de Recomendação e Análise de Dados no Retalho Alimentar: "Um caso real"

João Manuel Alferes Simões Vieira da Mota

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientador(a):
PhD, João Carlos Amaro Ferreira, Professor Auxiliar com Agregação,
ISCTE-IUL

Co-Orientador(a):
MSc, André Filipe Nova Marques, Partner da IMBS Consulting,
IMBS Consulting

Setembro, 2024

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Sistema de Recomendação e Análise de Dados no Retalho Alimentar: "Um caso real"

João Manuel Alferes Simões Vieira da Mota

Mestrado em Sistemas Integrado de Apoio à Decisão

Orientador(a):
PhD João Carlos Amaro Ferreira, Professor Auxiliar com Agregação
ISCTE-IUL

Co-Orientador(a):
MSc, André Filipe Nova Marques, Partner da IMBS Consulting,
IMBS Consulting

Setembro, 2024

Agradecimento

Alcançar este objetivo não teria sido possível sem o conhecimento adquirido ao longo dos últimos anos, nem sem as pessoas com as quais convivi neste mesmo período.

Aos meus orientador e coorientador, Professor João Carlos Ferreira e André Barbosa, agradeço toda a orientação e disponibilidade. Um agradecimento especial à IMBS pela disponibilização de dados, ferramentas para a realização desta dissertação.

À minha família, especialmente aos meus pais e ao meu irmão, por todo o apoio, paciência, motivação e amor. Sem eles, nada disto teria sido possível, tal como em tudo na minha vida. Considero-me realmente um rapaz com muita sorte pela família que tenho. Obrigado por estarem sempre ao meu lado e por acreditarem sempre nas minhas capacidades, principalmente quando eu próprio duvido.

Aos meus grupos de amigos, Palmelita (Coimbra) e Equipa Suprema (Lisboa), pelo apoio incondicional, pelas gargalhadas, pelos cafés para discutir ideias. Em especial, queria agradecer ao Miguel Conceição Valentim e à Catarina Pontes por terem feito este percurso de 2 anos comigo, e aos meus dois melhores amigos, Pedro Teixeira e Margarida Rosa.

Aos meus colegas de casa, António Leão Vicente, João Dias Figueira, Rodrigo Pinheiro pelo constante apoio e por me ouvirem diariamente.

Por fim, mas não menos importante, agradecer ao meu herói, Naruto Uzumaki, por me dar força e esperança quando mais precisei.

Um obrigado nunca vai ser suficiente.

Resumo

Devido ao acentuado crescimento do número de utilizadores de e-commerce, ocorrido nas últimas décadas, torna-se cada vez mais importante as empresas dominarem sistemas capazes de providenciar recomendações ao utilizador final com base nas suas preferências. Deste modo, os sistemas de recomendação não devem de forma alguma ser desvalorizados. Em casos como o do retalho alimentar, estes sistemas são essenciais para melhorar a experiência do consumidor e impulsionar vendas através da análise do histórico de compras para sugerir produtos personalizados, facilitando as decisões de compra e aumentando a satisfação do cliente e a faturação do negócio. Além de promover produtos relevantes, estes sistemas ajudam a fidelizar clientes, diferenciando as marcas da concorrência e contribuindo para o crescimento e sustentabilidade das empresas. O desenvolvimento deste projeto, pretende demonstrar um caso de implementação de um sistema de recomendação e posteriormente analisar o impacto da sua utilização na faturação da empresa. O projeto foi desenvolvido com dados de uma empresa multinacional na área do retalho alimentar, especificamente, dados de vendas dos anos de 2019 e 2020. A abordagem aos dados foi assente no modelo CRISP-DM, com a manipulação e análise dos dados realizada através de Python e PySpark. Após a integração e limpeza dos dados, foram aplicados métodos de visualização e o algoritmo de recomendação FP-growth. A análise resultante visa explorar o efeito da adoção de sistemas de recomendação sobre o desempenho financeiro da empresa e a satisfação do utilizador final.

Palavras-Chave: CRISP-DM; Sistemas de Recomendação; Análise de dados; FP-growth; Retalho alimentar; Vendas.

Abstract

Due to the sharp increase in the number of e-commerce users over the past decades, it has become increasingly important for companies to master systems capable of providing recommendations to the end user based on their preferences. In this sense, recommendation systems should not be undervalued. In sectors like food retail, these systems are essential for improving the consumer experience and boosting sales by analyzing purchase history to suggest personalized products. They facilitate purchasing decisions, increasing both customer satisfaction and business revenue. In addition to promoting relevant products, these systems help foster customer loyalty, differentiate brands from competitors, and contribute to business growth and sustainability. The aim of this project is to demonstrate a case of implementing a recommendation system and subsequently analyze its impact on the company's revenue. The project was developed using data from a multinational company in the food retail sector, specifically sales data from the years 2019 and 2020. The approach to the data was based on the CRISP-DM model, with data manipulation and analysis carried out using Python and PySpark. After data integration and cleaning, visualization methods were applied, along with the FP-Growth recommendation algorithm. The resulting analysis aims to explore the impact of adopting recommendation systems on the company's financial performance and end-user satisfaction.

Keywords: CRISP-DM; Recommendation Systems; Data Analysis; FP-growth; Food Sector; Sale

Índice

Índice de Figuras

Índice de Tabelas

CAPÍTULO 1

1 Introdução

1.1 Motivação

Nos últimos anos, tem-se assistido a um crescimento exponencial do número de utilizadores que utilizam *e-commerce* [1]. De acordo com o *"US Department of Commerce Retail Indicator Division"*, as vendas de comércio eletrónico atingiram os 870 biliões de dólares nos EUA em 2021, representando aumento de 14,2% relativamente a 2020 e 50,5% em relação a 2019 [2]. No ano de 2021, o *e-commerce* representou cerca de 13,2% do total das vendas do setor retalhista nos EUA [2]. Esta mudança de paradigma foi especialmente motivada pela conveniência em adquirir produtos a partir de casa, em ter acesso a um vasto leque de produtos não disponíveis localmente e ao tempo poupado [1]. Tendo em consideração este crescimento, existe cada vez mais a necessidade de implementar sistemas capazes de providenciar ao utilizador recomendações mais precisas, através das suas preferências.

Um sistema de recomendação tem como objetivo filtrar a informação em tempo real com o intuito de fornecer aos utilizadores recomendações personalizadas de produtos ou serviços. Uma das principais vantagens da implementação deste tipo de sistemas é o facto de lidarem com o crescente problema da sobrecarga de informação armazenada na internet devido ao e-commerce, pela sugestão de partes filtradas dessa informação e, dessa forma, permitirem melhorar a gestão da relação com o cliente. Outra vantagem apontada pela literatura refere-se ao impacto significativo no aumento das receitas de uma empresa [3].

1.2 Objetivos

A principal finalidade da presente dissertação é a demonstração de uma implementação de um sistema de recomendação e a análise do impacto da utilização destes sistemas na faturação de um negócio, assim como, a apresentação desses resultados às empresas. No presente trabalho é feita uma análise dos padrões de consumo dos clientes de uma empresa através de um modelo de *Market Basket Analysis* (MBA), que constitui o ponto de partida para a possível implementação futura de um sistema de recomendação totalmente funcional. Após a implementação do MBA, pretende-se ainda avaliar a viabilidade do sistema de recomendação obtido.

1.3 Empresa

Este projeto foi desenvolvido no âmbito de um estágio profissional na IMBS (Integrated Management Business Solutions). A IMBS é uma empresa nacional de consultoria de gestão fundada em 2014, que procura capacitar os seus clientes com melhores decisões de gestão através do desenvolvimento dos seus processos, apostando na Transformação Digital. Por razões de confidencialidade pré-estabelecidas, não é possível divulgar o nome do cliente a que os dados se referem.

1.4 Metodologia

Na condução desta dissertação foi utilizada a metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM). Esta metodologia foi utilizada devido à sua eficácia na organização e gestão de tarefas, e devido à sua natureza flexível.

As fases da metodologia CRISP-DM são:

1. *Business Understanding* – Nesta fase é necessário compreender o problema numa perspetiva de negócio;
2. *Data Understanding* – Fase de recolha, descrição e verificação da qualidade dos dados;
3. *Data Preparation* – Fase de preparação e limpeza dos dados "em bruto" (Adicionar e eliminar variáveis, integração de datasets, formatação, etc…);
4. *Modeling and Evaluation* – Aplicação de um modelo adequado aos objetivos do problema e respetiva avaliação da sua performance.
5. *Deployment* – Determinar uma estratégia de *Deployment*.

Especificamente nesta dissertação, opções por fazer uma adição na metodologia. Sendo assim criada a seguinte fase: *Exploratory Data Analysis* – Fase onde se visualizam e analisam descritivamente os dados finais.

O *Business Understanding,* está descrito na introdução deste documento; as demais fases da metodologia serão descritas nas seções seguintes.

Parte do projeto, mais especificamente nas fases de extração e modelação, recorreu-se à utilização da Arquitetura *Microsoft Azure,* apresentada na Figura 1.

De seguida, apresentam-se as ferramentas utilizadas e o seu devido propósito:

- Data Factory – extração de dados;
- Data Lake - repositório de armazenamento de dados;
- Databricks - transformação de dados e implementação de algoritmos;

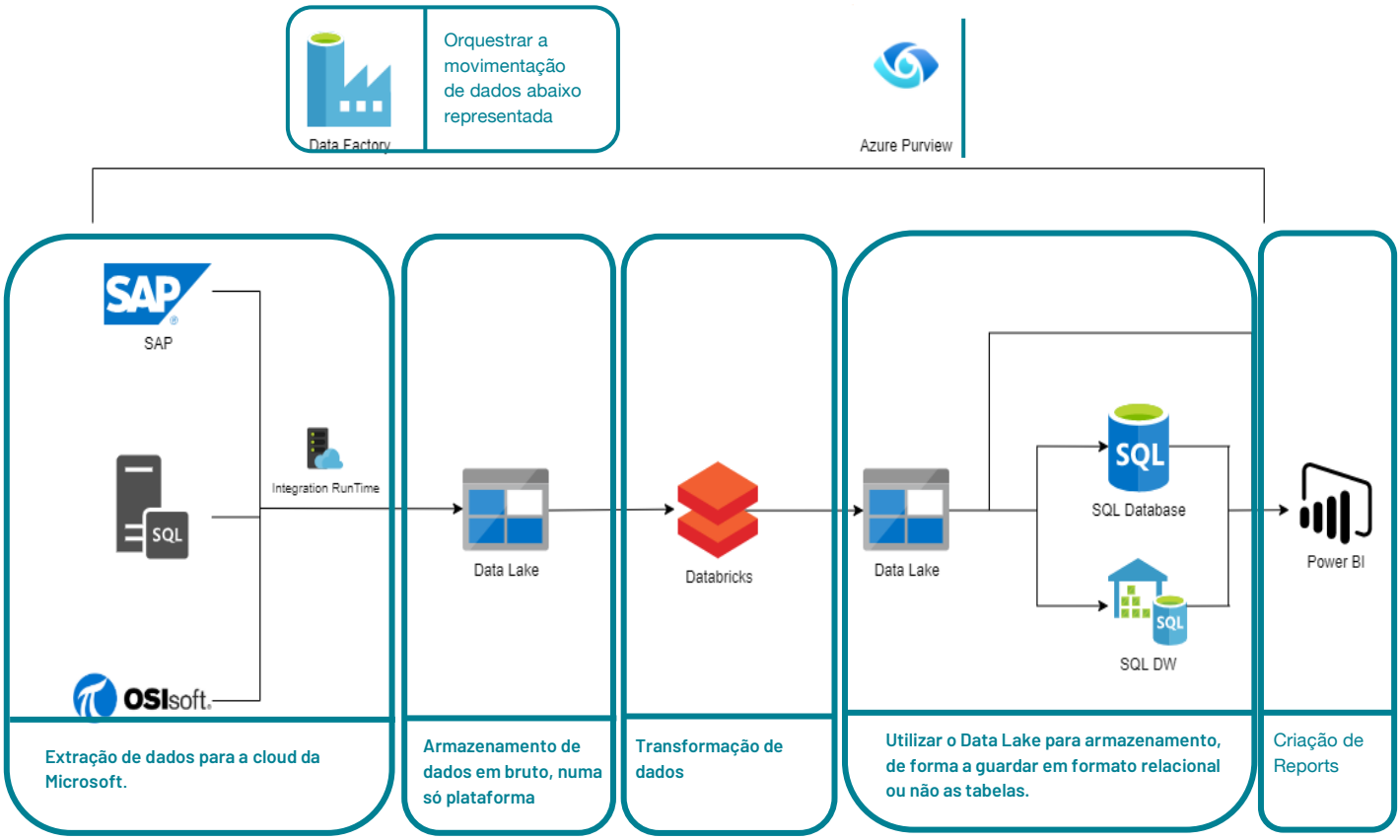


Figura 1: Arquitetura Azure

CAPÍTULO 2

2 Revisão da Literatura

2.1 Conceitos

2.1.1 Sistema de Recomendação

Os sistemas de recomendação podem ser definidos como algoritmos que sugerem itens ou produtos relevantes para os utilizadores. Várias técnicas de sistemas de informação têm sido propostas desde meados dos anos 90, e muitos tipos de *software* têm sido desenvolvidos recentemente para uma variedade de aplicações [4].

Em ambientes de informação complexo, onde existe uma grande quantidade de informação, como é o caso da Internet, um sistema deste tipo torna-se fundamental na melhoria da experiência do utilizador [1, 5]. No caso do comércio eletrónico, estes sistemas resolvem parte do problema de sobrecarga de informação através da filtragem da informação mais importante que é gerada de forma dinâmica, fornecendo aos utilizadores recomendações personalizadas e exclusivas de conteúdos e serviços [3]. Na ótica de uma empresa, um bom sistema de recomendação pode representar uma vantagem competitiva significativa, na medida em que melhora o processo de tomada de decisão [3, 4], potenciando assim, um aumento das receitas.

Na ótica do consumidor, torna-se bastante mais fácil escolher produtos de um cabaz altamente personalizado e ajustado ao seu historial de preferências, pela recomendação de produtos que sejam potencialmente mais relevantes [3].

Outro aspeto integral de um sistema de recomendação é o fornecimento de relatórios detalhados e atualizados, o que permitem à empresa tomar decisões informadas sobre a direção de uma campanha ou estrutura de uma página de produto [6].

Os dois principais tipos de algoritmos tradicionais de sistemas de recomendação são: ***Content-Based Filtering*** e ***Collaborative-Based Filtering,*** [7].

Um sistema de recomendação ***Content-Based*** sugere itens com um elevado grau de semelhança aos itens que um utilizador adquiriu ou mostrou interesse anteriormente, através de ações passadas ou *feedback* explícito. Este algoritmo dá especial ênfase aos atributos dos itens através da atribuição de uma categoria específica para cada item. Isto permite prever padrões de comportamento relacionados com algumas ligações existentes entre produtos semelhantes e tendências de consumo por parte dos utilizadores [8] Uma vantagem deste algoritmo é a capacidade de recomendar novos itens mesmo com a ausência de classificações fornecidas pelos utilizadores [3]. Além disto, é um sistema altamente adaptável na medida em que se ajusta rapidamente, caso as preferências dos utilizadores sofram algumas

alterações. Os utilizadores não são obrigados a partilhar o seu perfil para receber recomendações de produtos, assegurando assim a privacidade [3].

Por outro lado, este tipo de algoritmo apresenta algumas desvantagens. Uma delas é o facto de a precisão da recomendação estar inerentemente dependente da qualidade dos dados, ou seja, da descrição dos atributos dos itens e da organização do perfil do utilizador. Superespecialização é outro problema comum [10]. Muitas vezes, o utilizador sofre de algumas restrições quando lhe é recomendado o mesmo tipo de itens previamente definidos no seu perfil. Outra limitação deve-se à dificuldade em obter feedback do utilizador, na medida em que este sistema tipicamente não recomenda produtos com base na classificação atribuída por outros utilizadores. Assim sendo, não é possível verificar a qualidade da recomendação [10].

Collaborative filtering é um conjunto de técnicas que se baseia no comportamento passado dos utilizadores, através de uma análise muito extensa de uma vasta quantidade de dados relativa às preferências e interesses dos utilizadores [8].

Mais concretamente, é construída uma base de dados, muitas vezes representada por uma matriz de preferências de itens por utilizador. Tendo a base de dados formulada, o sistema faz a recomendação dos itens relevantes para um utilizador através da correspondência dos utilizadores com preferências semelhantes (*matching*), calculando o grau de semelhança entre os seus perfis. Os utilizadores com elevado grau de semelhança são agrupados num grupo denominado por vizinhança. A um utilizador pertencente a uma vizinhança é-lhe recomendado um certo item que este ainda não classificou, mas já teve uma classificação positiva pelos outros utilizadores da sua vizinhança [3].

A principal vantagem de um sistema de recomendação que se baseia nestas técnicas face a um sistema c*ontent-based,* é a possível vasta exposição a muitos produtos diferentes que são recomendados aos utilizadores. Para além disto, não é necessária informação extensiva acerca dos produtos e categorização dos itens.

Por outras palavras, o algoritmo mantém-se eficiente em situações onde não existe muita informação associada aos itens ou quando é difícil para um computador analisar opiniões de utilizadores acerca de produtos [3].

Quanto às desvantagens, uma delas está relacionada com o problema de *cold-start*, que ocorre quando existe pouca informação acerca de um utilizador, nomeadamente quando apresenta um perfil vazio (não classificou ou adquiriu nenhum produto) [10]. Isto reduz drasticamente a performance do algoritmo porque o sistema não conhece as preferências do utilizador. Outro problema derivado também da falta de informação é a esparsidade dos dados. Nem todos os utilizadores classificam todos os itens com os quais têm preferência, o que resulta na construção de uma matriz utilizador-item bastante esparsa e com fraca capacidade para definir vizinhanças [3, 11]. Consequentemente, as recomendações geradas pelo sistema são de fraca qualidade. A escalabilidade, aumento de clientes e produtos, também pode constituir um problema [3, 10]. À medida que o número de utilizadores e itens aumentam, torna-se

necessário ter à disposição um poder de computação crescente o que é bastante dispendioso, fazendo com que a implementação deste tipo de sistemas seja pouco viável para muitas empresas.

A Netflix tem um dos sistemas mais eficientes e conhecidos do mundo. De acordo com o *Netflix Research* [12], os algoritmos utilizados nos sistemas de recomendação são "*core*" do produto Netflix e, devido à sua importância, estão a ser feitas melhorias nas recomendações, fazendo avançar o estado da arte no terreno. De acordo com um artigo académico escrito por Gomez Uribe e pelo diretor de produtos da Netflix, Neil Hunt, o efeito combinado da personalização e recomendações é responsável por uma poupança anual de mais de um bilião de dólares [13]. Os sistemas de recomendação são aplicáveis a muitas outras indústrias. No caso da Amazon, 35% das compras dos clientes são provenientes de recomendações baseadas nos algoritmos mencionados [14]. Os resultados publicados num estudo demonstram que as recomendações no Youtube são responsáveis por cerca de 60% de todos os cliques em vídeos a partir da *home page* [15]. Com estes exemplos, é indiscutível o valor que estes sistemas de recomendação providenciam ao negócio de uma empresa, aumentando significativamente o valor das receitas.

Ainda hoje, têm sido feitos progressos na melhoria dos algoritmos e na introdução de novo *software* de sistemas de recomendação, alargando a sua utilidade para um vasto leque de aplicações e domínios tais como: *e-government*, *e-business*, *e-commerce/e-shopping*, *e-library*, *e-learning*, *e-tourism*, *e-resource services* e *e-group activities* [16]. Com a crescente utilização de *smartphones* com acesso à Internet já é possível oferecer recomendações personalizadas aos utilizadores móveis, sendo necessário desenvolver mais sistemas de recomendação móveis. No entanto, estão a ser conduzidas novas investigações em recomendações móveis sensíveis (dados móveis). Nos domínios de aplicação de *e-tourism* ou *e-shopping*, estão a ser feitos novos estudos devido à procura contínua em obter recomendações de lojas e produtos em tempo real com base na localização do utilizador [17]. Outro problema que à data não foi completamente resolvido, prende-se com a escassez de dados nos sistemas de recomendação (esparsidade). A transferência de técnicas de aprendizagem que combinam os dados relevantes de outros domínios para o domínio alvo, é um tópico muito debatido atualmente, pois representa uma boa oportunidade para resolver este problema.

Por fim, com o surgimento de *Big Data*, pode ser útil tirar partido desta grande quantidade de dados para modelar as preferências dos utilizadores de forma mais precisa e abrangente. No futuro, espera-se que sejam desenvolvidos mais sistemas de recomendação otimizados para processar grandes quantidades de dados [5].

2.1.2 Market Basket Analysis e Regras de Associação

Tal como referido na introdução, foi utilizado a *Market Basket Analysis* como ponto de partida para a possível implementação de um sistema de recomendação. A *Market Basket Analysis* (MBA) é uma técnica de *data mining* utilizada não só no retalho, mas também em muitas outras áreas. Através da análise do historial de compras de clientes, um sistema de MBA tem como principal objetivo analisar os produtos que os clientes tendem a comprar em conjunto [18]. O retalhista utiliza essa informação para desenvolver estratégias de venda como por exemplo, alteração do *layout* da loja física, *design* de catálogos de produtos, *marketing* cruzado em lojas online, *e-mails* personalizados com vendas, etc. O MBA permite aliviar o esforço de gestão através de um sistema que automatiza o processo para encontrar os produtos que são comprados em conjunto de acordo com um conjunto de regras de associação.

Exemplificando, um possível output seria: Um consumidor que compre um produto A, têm probabilidade P(x), de comprar um produto B.

As Regras de Associação são uma técnica de data mining utilizada em sistemas de MBA que tem como principal objetivo extrair correlações relevantes ou padrões frequentes entre itens em bases de dados de transações ou outros repositórios de dados com muitos registos. Estas relações, que são apresentadas como output num sistema de MBA, são essenciais para os retalhistas estudarem o comportamento dos seus clientes. Um dos métodos mais utilizados consiste em encontrar os conjuntos de itens (itemsets) mais frequentes de uma transação. [19, 20]

Uma transação refere-se à unidade de análise que engloba um conjunto de itens (cesto de compras, páginas de sessão de um utilizador num website, etc.) e um itemset é um conjunto de um ou mais itens. Um conjunto de itens de ordem K (K-itemset) é um conjunto de K itens. O comprimento (length) é o número de itens de um itemset.

Uma regra de associação é constituída pelo antecedente e consequente. Um antecedente é um item da base de dados e um consequente é um item encontrado em combinação com o antecedente. Matematicamente, uma regra de associação pode ser definida da seguinte maneira:

Sejam I=I1, I2, …, um conjunto de N atributos distintos e T uma transação que contenha um conjunto de itens tais que T⊆ I. Uma regra de associação é uma implicação sob a forma de X⇒Y, onde X, Y $\subset$ I são conjuntos de itens, e $\mathbf{X} \cap \mathbf{Y} = \emptyset$, X é o antecedente e o Y o consequente.

Existem outros dois parâmetros associados à medição da força das regras de associação: o Suporte e a Confiança. O suporte é a proporção de transações que contêm um conjunto de itens X e é dada pela fórmula:

$$Suporte(X) = \frac{frequência(X)}{N} = P(X)$$

Considerando, por exemplo, o itemset1={pão} e o itemset2={sabonete}, poderá haver mais transações que contêm pão do que transações que contêm sabonete, pelo que o itemset1 terá um valor de suporte mais elevado do que o itemset2.

Dada uma regra A => B, o Suporte diz respeito à proporção de transações que contêm tanto A como B do total de transações da base de dados. Por outras palavras, este parâmetro mede o quão frequente o conjunto de itens A e B aparecem juntos em todas as transações. O suporte de cada regra é então calculado por:

$$Suporte(A \Rightarrow B) = \frac{frequência(A,B)}{N} = P(AB)$$

Uma vez que a base de dados pode conter milhões de registos, é bastante comum definir um mínimo de suporte para selecionar apenas as regras mais relevantes através dos itens mais frequentes. Se, por exemplo, quisermos apenas considerar os conjuntos de itens que aparecem pelo menos 100 vezes num total de 10000 transações, teríamos de definir um suporte mínimo de 0,01.

A confiança é a proporção de transações que contêm A B do total de transações que contêm A, ou seja, é a probabilidade condicional da ocorrência do consequente dado o antecedente. Para uma regra de A => B, a Confiança resulta da proporção de transações onde B ocorre em A, sendo representada da seguinte forma:

$$Confiança(A \Rightarrow B) = \frac{P(AB)}{P(A)} = P(B/A) = \frac{frequência(A,B)}{frequência(A)}$$

Exemplificando, se tivermos a seguinte regra {Manteiga} => {Pão}, a confiança mede a percentagem de transações com o item manteiga que também contêm o item pão. Uma confiança de 70% significaria que 70% das transações que contêm manteiga também contêm pão. Tal como o suporte, também podemos definir um mínimo de confiança para descartar as regras de associação menos relevantes.

Se uma regra A => B satisfaz o mínimo suporte e a mínima confiança significa que a regra é forte. O objetivo do analista é encontrar todas as regras fortes. É importante realçar que os limites estabelecidos dependem muito do contexto de negócio do qual estas regras são aplicadas. Por exemplo, para um supermercado é expectável que exijam um suporte mínimo a rondar os 20% e um nível de confiança mínimo de 70% dada a dimensão da base de dados das compras dos clientes. Por outro lado, um analista de deteção de fraudes deverá reduzir os níveis de suporte e confiança mínimos porque existem poucas transações fraudulentas. [21]

O Lift é outro parâmetro que mede o grau de importância de uma regra. Considerando a regra A => B, o Lift fornece a correlação entre A e B, ou seja, mostra como um conjunto de itens A afeta o outro conjunto de itens B. Por outras palavras, podemos dizer que o Lift se refere à probabilidade de B estar

no carrinho com o conhecimento de A também estar presente sobre a probabilidade de B estar no carrinho sem qualquer conhecimento da presença de A. Matematicamente, temos que:

$$Lift\ (A \Rightarrow B)\ = \frac{Suporte(A \Rightarrow B)}{Suporte(A)Suporte(B)} = \frac{P(AB)}{P(A) * P(B)}$$

Se o *Lift* = 1, A e B são independentes e nenhuma regra pode ser extraída entre os dois conjuntos de itens. Se o *Lift* > 1, A e B dependem um do outro, ou seja, existe uma elevada associação entre os dois conjuntos de itens. Quanto mais elevado for o valor do *Lift*, maior é a probabilidade de preferência para adquirir B. Se o *Lift* < 1, a presença de A terá um efeito negativo em B (correlação negativa).

Exemplificando, consideremos a seguinte regra: pasta de dentes => leite. Sabemos que a probabilidade de o leite estar contido na transação/carrinho sabendo que a pasta de dentes também está é de 70%, isto é: Confiança (pasta de dentes => leite) = 0,7. Agora consideremos também que a probabilidade de haver leite na transação sem qualquer conhecimento da pasta de dentes é de 80%. Com estes resultados podemos concluir que a existência da pasta dos dentes reduz a probabilidade de o leite estar contido na transação (passou de 0,7 para 0,8). Assim, o Lift (pasta de dentes => leite) = 0,7/0,8 = 0,87, um valor inferior a 1.

2.1.3 Algoritmos de MBA

Existem alguns algoritmos baseados em regras de associação. O **Algoritmo Apriori** é um dos algoritmos mais utilizados para encontrar conjuntos de itens frequentes através das regras de associação. O funcionamento do algoritmo encontra-se resumido no Anexo 1.

É de realçar que o Algoritmo Apriori apresenta algumas desvantagens, uma vez que em bases de dados muito extensas, este algoritmo é lento, pouco eficiente e ocupa muita memória, pois tem de percorrer todos os registos mais do que uma vez, gerar um grande número de conjuntos de itens candidatos e verificar cada um deles.

Para colmatar estes inconvenientes, neste projeto foi utilizado o **algoritmo *FP-Growth*** na implementação do modelo de *Market Basket Analysis*. Este algoritmo é um dos métodos mais utilizados na mineração de conjuntos de itens frequentes e baseia-se na construção de uma *FP-Tree* em que cada nodo representa um item com um contador. Para isto, podemos dividir o processo em duas fases distintas: Pré-processamento e projeção da árvore.

A primeira fase consiste em:

1. Percorrer a base de dados que contém todas as transações e calcular o valor do suporte para cada conjunto de itens;
2. Eliminar os conjuntos de itens pouco frequentes, ou seja, descartar aqueles que não respeitam o suporte mínimo exigido;
3. Ordenar os conjuntos de itens de forma decrescente, de acordo com os valores de contagem do suporte.

O resultado desta fase deve ser uma tabela com os conjuntos de itens ordenados por ordem decrescente com a respetiva contagem.

De seguida, na **segunda fase**, o algoritmo constrói a *FP-Tree*, através dos seguintes passos:

1. Considera-se que o nodo raiz da árvore é um valor nulo;
2. Percorrer a base de dados e examinar as transações. Para a primeira transação, o conjunto de itens com maior frequência aparece no topo com ligação ao nodo nulo definido anteriormente. Os conjuntos de itens restantes dessa transação vão sendo adicionados à árvore por ordem decrescente (*top to bottom*) dos valores do suporte, que foram calculadas na fase de pré-processamento;
3. De seguida, o processo anterior é repetido para todas as transações (processo iterativo) e acaba quando não existirem mais transações a examinar. É de realçar que se algum conjunto de itens desta transação já estiver presente noutro ramo (em resultado de uma transação anterior), então basta incrementar em 1 o contador para esse conjunto de itens. Caso contrário, é necessário adicionar outro ramo;
4. Com a construção da *FP-Tree*, é necessário proceder à sua mineração. Para isso, é necessário examinar os nodos com valores de suporte mais baixos e examinar o caminho desde esse nodo até ao nodo raíz, listando todos os itens intermédios. A este caminho ou conjunto de caminhos dá-se o nome de *conditional pattern base*.
5. Contruir a *FP-Tree* condicional que resulta da contagem dos conjuntos de itens listados no *conditional pattern base*, sendo apenas contabilizados aqueles que respeitam o valor de suporte mínimo estabelecido;
6. A partir da *FP-Tree* condicional, são gerados os padrões frequentes que constituem as regras de associação.[22]

CAPÍTULO 3

3 Sales analytics process

3.1 Data Understanding

Numa primeira fase, foi necessário extrair os dados necessários em prol do objetivo. Posto isto, através da ferramenta fornecida pelo Azure, Data Factory, começou-se por criar diferentes pipelines.

Através desta ferramenta, foi possível fazer uma ligação com as Bases de Dados de SQL do cliente. Contudo, foi necessário abortar as seguintes etapas:

- Mapeamento das variáveis pretendidas extrair;
- Mapeamento do formato de cada coluna (exemplo: string, binário, integer, entre outras);
- Decisão do formato de ficheiro em que se iria guardar os dados (exemplo: csv ou parquet);

Após todas as extrações de dados efetuadas, é necessário perceber com que dados estamos a lidar e o seu significado, metadados.

O *dataset* final utilizado no projeto, denominado de "Fact_Vendas", possui dados relativos às vendas (transações) de um cliente da IMBS do setor do retalho alimentar nos anos de 2019 e 2020. O *dataset* é constituído por mais de 14,4 milhões de observações (transações) e 27 diferentes variáveis.

Como variáveis de maior foco e relevância realça-se as seguintes:

- Nº Documento: número da fatura;
- Moeda: *currency* em que a venda foi feita;
- Quantidade Bónus: produtos oferecidos em compensação de compras de grandes quantidades;
- Nº Material: código de produto;
- Quantidade de oferta: oferta de um produto (exemplo: cabaz de Natal);
- Nº Cliente: código de cliente;
- Data: data da venda do produto;
- Custos de Transporte: custo de transporte do produto;
- Custos de Bónus: custa da venda dos produtos bónus;
- Quantidade faturada: quantidade de produto vendido;
- Custo: custos de produção do produto vendido;
- Faturação EUR: valor faturado já com o câmbio de moedas realizado;
- Custo EUR: valor dos custos já com o câmbio de moedas realizado.

Posto isto, ficámos com cerca de 12 milhões de observações. De notar que, por questões de confidencialidade, não se teve acesso ao nome dos produtos e clientes, pelo que a análise foi realizada apenas com a sua codificação.

3.2 Data Preparation

A seguinte fase do CRISP-DM trata-se do processo de limpeza e tratamento de dados foi uma das primeiras tarefas realizadas nos dados, da qual se realça os seguintes aspetos:

a) Alteração à natureza da variável "Data". Esta variável era considerada como uma variável "object", pelo que se procedeu à mudança de tipo para "datetime". Desta forma, a variável foi transformada numa variável de tempo que devolve corretamente o dia, mês e ano de cada venda;
b) Análise de valores duplicados. No *dataset* existiam 568 casos duplicados que foram removidos;
c) Análise da existência de dados não atribuídos (NA). No *dataset* existiam 191 casos de valores não atribuídos que foram removidos;
d) Outras análises de verificação. A análise de valores nulos ou omissos no *dataframe*, tal como a verificação de outliers, revelou a inexistência destes casos;
e) Criação da variável margem bruta absoluta. Esta variável resultou da diferença entre a faturação e o custo;
f) Criação da variável quantidade. Esta variável é a quantidade total de um produto por fatura e resultou da soma das três variáveis de quantidade (Quantidade Bónus, Quantidade de oferta e Quantidade faturada). Previamente, estas variáveis foram reclassificadas, i.e., o tipo foi retificado (de *String* para *Integer*);
g) Criação de um novo index para o produto. Esta variável foi necessária para melhorar a performance do algoritmo de *FP Growth*. É apenas utilizado no algoritmo. A criação do index está explicada na preparação dos dados para a modelação.

3.3 Exploratory Data Analysis

Para a condução da Análise Exploratória de Dados foram utilizadas algumas bibliotecas da linguagem Python tais como: *pandas*, *numpy*, *matplotlib* e *seaborn*. Nesta etapa, foram realizadas análises exploratórias aos dados com dois objetivos: um de caracterização da empresa cliente da IMBS em termos de vendas, nomeadamente faturação, produtos vendidos, custos e margem bruta, e outro de caracterização de produtos e de clientes da empresa com maiores margens brutas absolutas.

Tal como já mencionado, a variável relativa à margem bruta absoluta foi adicionada à base de dados e resulta da diferença entre a faturação e o custo. Apesar desta margem não representar o lucro final da empresa, fornece uma perspetiva realista do lucro operacional.

3.3.1 Caracterização do cliente em termos de vendas

Para a condução da Análise Exploratória de Dados foram utilizadas algumas bibliotecas da linguagem Python tais como: *pandas*, *numpy*, *matplotlib* e *seaborn*. Nesta etapa, foram realizadas análises exploratórias aos dados com dois objetivos: um de caracterização da empresa cliente da IMBS em termos de vendas, nomeadamente faturação, produtos vendidos, custos e margem bruta, e outro de caracterização de produtos e de clientes da empresa com maiores margens brutas absolutas.

Tal como já mencionado, a variável relativa à margem bruta absoluta foi adicionada à base de dados e resulta da diferença entre a faturação e o custo. Apesar desta margem não representar o lucro final da empresa, fornece uma perspetiva realista do lucro operacional. Desta maneira tornou-se possível calcular a Margem Bruta e Quantidade Faturada, nos anos de 2019 e 2020, como a Tabela 1 demonstra.

	Faturação	Custo	Margem Bruta	Quantidade Faturada
2019	521.129.586,44€	307.459.737,50€	213.669.848,95€	210.202.118
2020	437.222.845,64€	262.182.874,91€	175.039.970,73€	105.187.344
Δ	-16,1 %	-14,1 %	-18,1%	-99,8%

Tabela 1: Vendas do Cliente da IMBS em 2019 e 2020

Para melhor compreender as vendas do cliente da IMBS, foi também conduzido um estudo com base mensal para as mesmas variáveis. Este estudo está apresentado no anexo 2. Esta análise confirma que o ano de 2019 apresentou resultados mais favoráveis para a empresa, com a exceção de alguns meses. Observando o comportamento da margem bruta absoluta mensalmente (Figura 3 do Anexo 2), verifica-se que a maioria dos meses apresenta maiores margens em 2019, nomeadamente em abril e maio, onde as taxas de variação são de -82,3% e -54,1%, respetivamente. Em fevereiro, julho e setembro verificam-se maiores margens em 2020.

3.3.2 Produtos e Clientes de maiores margens

Para além da análise anterior, realizou-se a caracterização de produtos e de clientes da empresa com maiores margens brutas absolutas, pela apresentação dos top 10 produtos com maior margem e top 10 clientes com maior margem nos dois anos analisados.

Relativamente aos produtos com maior margem absoluta, o primeiro passo consistiu em criar um *subset*, denominado de *fact_vendas_subset*, a partir da base de dados principal (*fact_vendas*) e selecionar as colunas representativas do código do produto (*Nº Material*) e margem absoluta (*Margem Absoluta EUR*). No segundo passo, realizou-se uma agregação da margem bruta absoluta por produto através do comando *group_by.sum()*. Neste caso concreto procedeu-se à soma das margens, pois pretendeu-se ter para cada produto o total da margem bruta absoluta. Por fim, com o intuito de obter uma perspetiva mais clara do peso da margem de cada produto em relação à margem bruta absoluta total, foi adicionada uma terceira coluna de percentagem, resultando na Tabela 2 apresentada no Anexo 3. A informação desta tabela está graficamente apresentada na Figura 2. De relembrar que, por questões de confidencialidade, não se teve acesso ao nome dos produtos, pelo que na figura estes foram substituídos por P (P de produto) e um número por ordem de maior margem; em que P1 é o produto de maior margem, P2 é o produto com a segunda maior margem, etc. Na Tabela 2 do anexo 3 estão apresentados os códigos dos produtos.

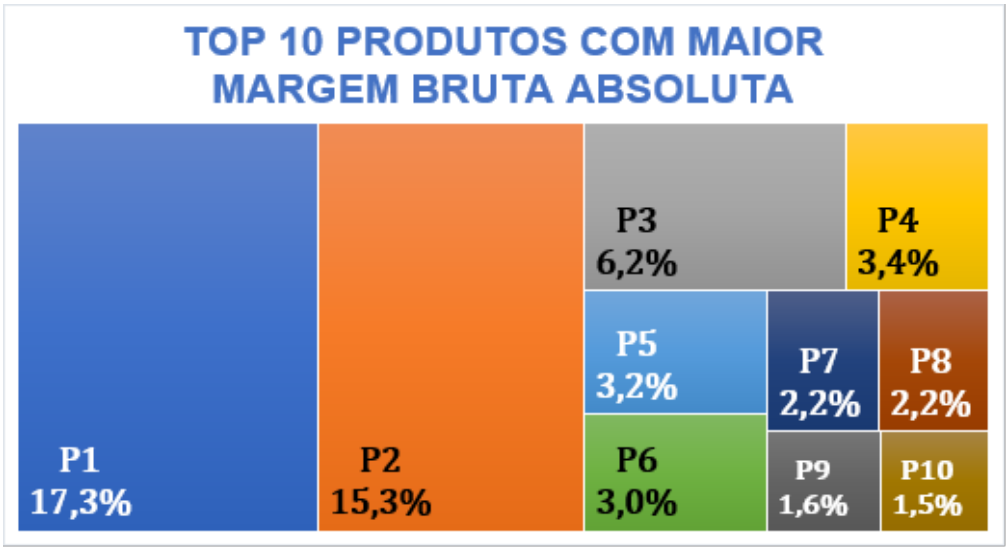


Figura 2: Top 10 produtos com maior margem bruta

Tal como podemos observar na Figura 2, conclui-se que os dois primeiros produtos com maior margem, P1 e P2 (de código *5014000* e *5015000 respetivamente*) são responsáveis por cerca de um terço da margem bruta absoluta total, mais precisamente 32,6%. Para além disto, é de realçar que o conjunto destes 10 produtos com maiores margens perfazem 55,9% da margem bruta absoluta total.

Relativamente ao top 10 clientes com maiores margens, seguiu-se o mesmo processo. Neste caso, o comando *group_by.sum()* operou sobre as variáveis *"Nº Cliente"* e *"Margem Absoluta EUR"*. O top 10 clientes com maior margem bruta estão apresentados na Figura 3. No Anexo 3 apresenta-se a mesma informação na Tabela 3 com o código dos clientes.

Figura 3: Top 10 clientes com maior margem bruta

Analisando a Figura 3, pode verificar-se que apenas três clientes totalizam cerca de 12% da margem bruta total e que os 10 clientes que possuem margens mais elevadas perfazem 21,6% da margem bruta absoluta total.

CAPÍTULO 4

4 Modelling and Evaluation

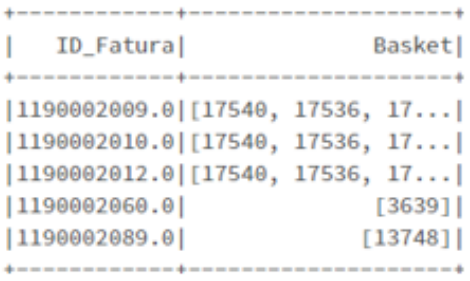
4.1 Preparação dos dados para modelação

Antes de dar início à fase de modelação, realizou-se a preparação dos dados com o foco naquilo que se pretende obter, isto é, a obtenção das regras.
Foi desenvolvida uma nomenclatura para todas as variáveis da base de dados de input à modelação. Deste modo, os códigos de fatura e de produto passaram a ser denominados de "ID_Fatura" e "ID_Artigo", respetivamente.

Para a fase de modelação, definiu-se que o algoritmo *FP Growth* iria ser aplicado tendo em conta apenas o código das faturas, códigos de produtos e quantidade adquiridas por produto, mais especificamente as variáveis: "ID_Fatura" (antes nomeada de "Nº Documento"), "ID_Artigo" (cujo nome original era "Nº Material"), "Quantidade Bónus", "Quantidade de oferta" e "Quantidade faturada". De seguida, foi efetuada uma fusão, através da soma, das três variáveis de quantidade numa única variável, "Quantidade", que se refere à quantidade total de um produto por fatura.

Os códigos dos produtos (variável ID_Artigo) eram compostos por valores muito extensos, o que iria ter um impacto negativo na performance do algoritmo. Por isso, optou-se pela criação de um novo *index*, através de um dicionário, para cada produto diferente, começando no número 1 e acabando no número total de diferentes produtos disponíveis na base de dados. Apresenta-se uma ilustração desta atribuição no anexo 3. De realçar que este novo index de produtos é usado apenas pelo algoritmo; o index não tem significado para o cliente, mas sim o código de produto disponível em "ID_Artigo". Por isso, após a execução do algoritmo, o processo reverte-se para que o *output* final contenha os códigos originais de cada produto e, desta forma, faça sentido para o cliente.

Por fim, os produtos vendidos por fatura foram agrupados (através de um *groupby*), ou seja, a cada fatura foi associado um *basket* de produtos. A Figura 4 apresenta o resultado deste comando para cinco faturas. De referir que o output gerado possui 2.650.207 milhões de faturas diferentes, cada uma com um conjunto de produtos (*basket*) associado.



ID_Fatura	Basket
1190002009.0	[17540, 17536, 17...]
1190002010.0	[17540, 17536, 17...]
1190002012.0	[17540, 17536, 17...]
1190002060.0	[3639]
1190002089.0	[13748]

Figura 4: Associação de um basket de produtos às faturas

4.2 Aplicação do Algoritmo

Para a execução do algoritmo *FP Growth*, é necessário definir o suporte mínimo e a confiança mínima. O suporte é dado por um número x, estabelecido pelo investigador, a dividir pelo número total de faturas diferentes, de acordo com a seguinte formula:

$$Suporte = \frac{x}{Nº\ de\ diferentes\ faturas} = \frac{x}{2.650.207}$$

A definição do valor para o suporte mínimo vai definir a força das regras de associação de acordo com a frequência dos produtos nas transações. A Tabela 2 apresenta três colunas, em que a primeira coluna refere-se ao x que atribuímos para o suporte, a segunda apresenta o tempo que o algoritmo demorou a ser executado e, por fim, a terceira coluna refere-se ao número de regras de associação criadas após a execução do algoritmo.

Suporte(x)	Tempo Total	NºRegras de Asso.
1500	5m24s	1.500.000
1600	4m45s	1.074.200
1750	4m31s	660.000

Tabela 2: Tempos de execução do algoritmo

Após análise de resultados, optou-se pelo valor de x de 1600, ou seja, por um suporte mínimo de:

$$Suporte = \frac{1600}{2.650.207} = 0{,}0006$$

Com o suporte mínimo definido e uma confiança mínima de 0,10 (10%), o algoritmo foi executado tendo-se obtido mais de um milhão de regras de associação. A Tabela 3 apresenta um excerto do *output* obtido.

antecedent	consequent	confidence	lift	support
[18158, 17622, 17...]	[18154]	0.9793296089385475	227.86884854400364	6.614577653745537E-4
[18158, 17622, 17...]	[17572]	0.9776536312849162	160.67124502087958	6.603257783259949E-4
[18158, 17622, 17...]	[17521]	0.9994413407821229	177.79073952410843	6.7504160995726E-4
[18158, 17622, 17...]	[17570]	0.9849162011173185	95.76021023606006	6.652310555364166E-4
[13857, 13661, 13...]	[1005]	0.940415964024733	133.82897335392235	6.312714440796512E-4
[13852, 13661, 13...]	[1005]	0.9803240740740741	139.50822764214303	6.391953534195631E-4
[13858, 13857, 13...]	[1005]	0.9767718880285885	139.00272217454662	6.188195865455038E-4
[18435, 18430, 18...]	[18463]	0.9993993993993994	1336.3346539273884	6.278754829339746E-4
[18435, 18430, 18...]	[18436]	0.990990990990991	1327.771112872225	6.225928767073666E-4
[18435, 18430, 18...]	[18456]	0.9933933933933934	1335.039617101889	6.241021927721117E-4
[18435, 18430, 18...]	[18438]	0.9705705705705706	1343.1921253890966	6.09763690157033E-4
[18435, 18430, 18...]	[18428]	0.9861861861861861	1354.8976329362022	6.195742445778764E-4
[13857, 13661, 13...]	[13653]	0.9449275362318841	147.5781466800927	6.15046296383641E-4
[13857, 13661, 13...]	[1005]	0.9860869565217392	140.32833350065022	6.41836656532867E-4
[13860, 13699, 13...]	[13857]	0.9587203302373581	236.35417025463795	7.010773120741134E-4
[13860, 13699, 13...]	[13653]	0.9339525283797729	145.86407733984163	6.829655192971719E-4
[13860, 13699, 13...]	[1005]	0.8957688338493293	127.47531728772644	6.550431720993869E-4

Tabela 3: Excerto do output obtido pela execução do MBA

Após o Market Basket Analaysis realizado, foram efetuadas algumas alterações ao output final. Como referido anteriormente, foi reposto o código de produto original presente nas colunas de "antecent" e "consequent". Foram também criadas duas variáveis: "N_Basket" que funciona como um index para os baskets e "N_Antecentes" que possui o número total de produtos que cada cesto contém. Por fim, o nome das variáveis foi alterado de inglês para português. A Tabela 4 apresenta um excerto do output final das regras de associação.

N_Basket	Antecedente	Consequente	N_Antecedentes	Confiança	Interesse	Suporte
1	012681	012682	1	0.3433231396534149	425.5740823065751	6.354220632577003E-4
2	012681	6254029	1	0.4401630988786952	14.478562794499252	8.146533459461846E-4
3	012682	012681	1	0.7876520112254444	425.5740823065752	6.354220632577003E-4
4	012870	5028323	1	0.4347138337647397	16.473841697030764	0.001140665615931...
5	016501	500000	1	0.5838283828382839	130.9356069219514	6.674950296335343E-4
6	016501	5957007	1	0.7795379537953795	125.24625291992672	8.91251136232E-4
7	016501	5957007	2	0.933295647258338	149.95008532486074	6.229702057235529E-4
7	500000	5957007	2	0.933295647258338	149.95008532486074	6.229702057235529E-4
8	016501	500000	2	0.6989839119390348	156.7616193880184	6.229702057235529E-4
8	5957007	500000	2	0.6989839119390348	156.7616193880184	6.229702057235529E-4
9	016502	640106	1	0.19277446904146534	27.855201312631518	0.001294615854535...
10	016502	640350	1	0.12686818743679065	16.159319383971482	8.520089185486266E-4
11	016502	1053025	1	0.11349589841555231	2.6998745552580212	7.622046126962913E-4
12	016502	600311	1	0.16597370491066413	14.164509389134297	0.001114629913814...
13	016502	600205	1	0.34891560849533654	19.222504688568765	0.002343213190516816
14	016502	1058046	1	0.12119339251601303	2.70973481367478	8.13898687913812E-4
15	016502	649001	1	0.3280143836386111	38.0074333516847	0.002202846796495519

Tabela 4: Excerto do output final das regras de associação

CAPÍTULO 5

5 Discussão e Resultados

5.1 Resultados

Após a fase de modelação, propôs-se fazer uma avaliação do impacto financeiro potencial para a empresa cliente com a implementação das regras de associação, ou seja, estimar para cada *basket* e para cada produto a margem absoluta acrescida à margem já existente pela compra do produto sugerido. A análise deste objetivo está efetuada no *notebook* denominado por *"Resultados"*.

O primeiro desafio consistiu em determinar para cada produto a sua margem absoluta unitária, onde bastou efetuar o quociente entre a margem absoluta total e a quantidade faturada no notebook da "Análise Exploratória" esta operação já tinha sido realizada, portanto apenas bastou exportar a tabela para o novo notebook "Resultados". A Tabela 5 apresenta o resultado pretendido para 5 produtos.

	Margem Absoluta Unitária EUR
Nº Material	
5014000	12.93
5015000	12.12
5010000	14.91
5028323	4.77
6254045	10.52

Tabela 5: Margem Absoluta Unitária

A Tabela 5 e o output final da modelação foram importados para o *notebook* "Resultados" para determinação da margem acrescida para cada *basket* e cada produto. Dado que o consequente representa o produto com maior probabilidade de ser comprado com base num antecedente ou conjunto de antecedentes, pretende-se estimar, para cada *basket*, a margem absoluta acrescida. Para isto, efetuando um *merge* das duas tabelas mencionadas anteriormente com base na coluna em comum (*"Nº Material"* e *"Consequente"* representam ambos o código de produto). De seguida, foi calculada a margem absoluta do *basket* pela soma das margens absolutas de todos os antecedentes e calculada a margem absoluta acrescida" que resulta no produto da confiança (probabilidade de adquirir o consequente dado o antecedente) com a margem absoluta unitária do consequente. A Tabela 6 apresenta o output gerado para apenas 5 *baskets*.

	N_Basket	N_Antecedentes	Antecedente	Consequente	Confiança	Margem Absoluta Unitária EUR	Margem Absoluta Acrescida
0	1	1	12681	012682	0.34	0.22	0.07
94333	12338	1	1058035	1020222	0.14	0.28	0.04
224274	15888	1	1058050	1058000	0.12	2.21	0.27
224275	16709	1	1058054	1058000	0.26	2.21	0.57
224276	16732	1	1058055	1058000	0.33	2.21	0.73

Tabela 6: Margem Absoluta Acrescida por Basket

Na Tabela 6, verifica-se, por exemplo, que o *basket* 1 obteve uma margem absoluta de 0,22€. Com as regras de associação postas em prática, um cliente que compre o produto 12681 (antecedente) tem uma probabilidade de 34% (confiança) em adquirir o produto 012682 (consequente), traduzindo-se num potencial ganho de 0,07€ (margem absoluta acrescida).

Para estimar a margem absoluta potencial para cada produto, foi criado um *subset* do *output* anterior (Tabela 6) com as variáveis *"Consequente"*, *"Margem Absoluta Unitária EUR"* e *"Margem Absoluta Acrescida"*. Através de um *group_by* do consequente pela média, determinou-se a margem absoluta acrescida média por produto. Por outras palavras, conseguiu-se determinar, em média, qual o valor acrescentado em termos de margem absoluta acrescida de todos os produtos que aparecem pelo menos uma vez como consequente nas regras de associação. Por fim, foram adicionadas, uma coluna da taxa de variação da margem pelo quociente entre a margem absoluta acrescida e uma da margem absoluta unitária. A Tabela 7 mostra o resultado para 5 produtos.

Consequente	Margem Absoluta Unitária EUR	Margem Absoluta Acrescida	TxVarMargem
1053060	38.99	31.19	80.0
1030026	29.28	24.89	85.0
1028955	23.87	20.77	87.0
5121000	17.69	14.67	82.9
5011010	13.53	13.19	97.5

Tabela 7: Margem bruta acrescida por produto e respetiva taxa de variação

Através da Tabela 7, observa-se na primeira linha que a margem absoluta do produto *1053060* é de 38,99€. Cada vez que este produto aparece como consequente nas regras de associação (ou seja, nos diferentes *baskets*), estima-se que, em média, a sua margem absoluta acrescida seja de 31,19€. Isto significa que, para este produto, se as regras de associação forem colocadas em prática a margem absoluta acrescida pode atingir um ganho potencial máximo de 80%.

Para avaliar o potencial máximo de rendimento que a empresa pode alcançar com estas regras na sua globalidade, optou-se por elaborar um *boxplot* (Figura 7) em relação à taxa da margem absoluta acrescida:

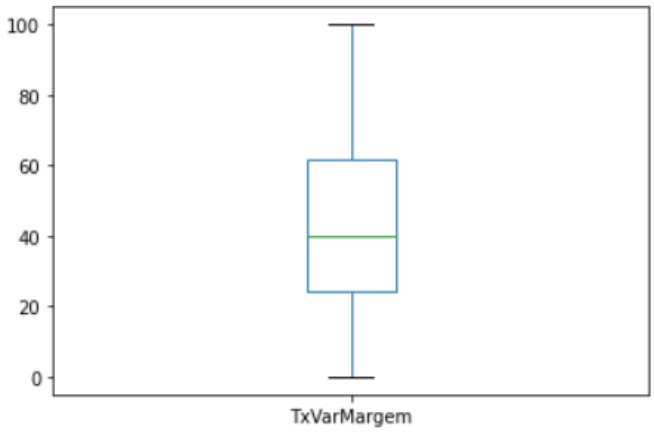


Figura 5: BoxPlot da Margem Absoluta Acrescida

Com um suporte mínimo de 0,06% e uma confiança mínima de 10%, verificou-se que apenas três produtos não foram recomendados e a recomendação em 12 produtos aumentou a margem para o dobro. Verificou-se também que o aumento médio da margem com a implementação de um sistema de recomendação com este suporte mínimo e esta confiança mínima foi de cerca de 45,1%, com desvio padrão de 25,4% o que indicia muita dispersão. Analisando o gráfico da Figura 7, verifica-se que, em metade das regras, a margem aumenta em mais de 40% e em um quarto aumenta em mais de 60%.

□

5.2 Conclusões

Devido ao aumento acentuado do número de utilizadores de *e-commerce* nas últimas décadas, tornou-se ainda mais crucial o domínio de sistemas capazes de fornecer recomendações personalizados aos utilizadores finais, com base nas suas preferências e comportamentos de compra.

Isto deve-se ao grande impulsionamento na área da transformação digital e pelo consequente aumento da concorrência no mercado, obrigando as empresas a procurar soluções inovadoras para melhorar a experiência do cliente e, simultaneamente, aumentar as suas receitas. Neste contexto, os sistemas de recomendação desempenham um papel fundamental e não devem ser subestimados.

No sector do retalho alimentar, em particular, estes sistemas são essenciais para otimizar a experiência do consumidor e impulsionar as vendas. Ao analisar o histórico de compras e outros dados relevantes do cliente, os sistemas de recomendação podem sugerir produtos personalizados, adaptados às necessidades e preferências individuais. Esta personalização facilita significativamente as decisões de compra, fornecendo aos clientes sugestões mais precisas, o que, por sua vez, aumenta a sua satisfação e lealdade para com a marca. Ao mesmo tempo, a capacidade de recomendar produtos com precisão tem um impacto direto nas receitas da empresa, uma vez que incentiva os consumidores a explorar novas ofertas e a fazer compras adicionais.

Para além de promoverem produtos relevantes para cada cliente, os sistemas de recomendação desempenham um papel estratégico na promoção da fidelização dos clientes. Ao oferecer uma experiência de compra mais intuitiva e personalizada, as marcas conseguem diferenciar-se da concorrência e construir uma relação de confiança com os consumidores, que se sentem valorizados. Este fator é fundamental não só para o crescimento do negócio, mas também para a sua sustentabilidade a longo prazo.

Neste projeto analisou-se a implementação de um sistema de recomendação em contexto real, utilizando dados de uma empresa multinacional do sector do retalho alimentar. O conjunto de dados é composto por registos de vendas dos anos de 2019 e 2020, abrangendo um período em que o comércio eletrónico teve uma expansão significativa. Este cenário constitui uma excelente oportunidade para avaliar o impacto concreto da utilização de sistemas de recomendação no aumento das vendas e na melhoria da satisfação dos clientes.

A abordagem adotada para a análise de dados seguiu o modelo CRISP-DM (Cross-Industry Standard Process for Data Mining), um dos métodos mais reconhecidos e eficazes para a exploração de dados e o desenvolvimento de modelos preditivos. A manipulação e a análise dos dados foram efectuadas utilizando Python e PySpark, que se revelaram essenciais para lidar com grandes volumes de dados e garantir uma análise robusta e escalável. Após a integração e limpeza dos dados, foram aplicados

vários métodos de visualização de dados para identificar padrões e tendências relevantes, seguidos da implementação do algoritmo de recomendação FP-Growth, um dos algoritmos mais eficientes para descobrir padrões frequentes em grandes conjuntos de dados.

A análise das vendas da empresa permitiu verificar que atualmente um terço da margem bruta total é obtida através da venda de apenas dois produtos e os 10 produtos com maiores margens perfazem mais de metade (56%) da margem bruta total. No que se refere aos clientes, apenas três clientes totalizam cerca de 12% da margem bruta total e os 10 clientes que possuem margens mais elevadas perfazem 21,6% da margem bruta total.

Os resultados permitiram verificar que, usando um suporte mínimo muito reduzido (0,06%), as regras de associação apontam para um aumento médio da margem bruta unitária de cerca de 45% (desvio padrão de 25%). Para além disso, verificou-se que 75% dos produtos aumentou, no máximo, a sua margem em 60% e metade aumentou em 40%. Tais resultados permitem concluir que a variação na margem bruta estimada de cada produto foi positiva e elevada, podendo afirmar-se que as regras de associações permitem aumentar as margens das empresas que implementam sistemas de recomendação.

Neste estudo, o suporte mínimo considerado foi muito pequeno e, no entanto, resultou num ganho médio de 45% na margem bruta. Outros valores de suporte mínimo e de confiança mínima deverão ser equacionados e analisados o seu impacto económico.

A análise resultante visa explorar o impacto da adoção de sistemas de recomendação não só no desempenho financeiro da empresa, mas também na satisfação e retenção dos utilizadores. A hipótese é que a implementação deste tipo de sistema trará melhorias significativas na experiência de compra do cliente e simultaneamente aumentará o volume de vendas da empresa, confirmando a relevância estratégica dos sistemas de recomendação no atual panorama do retalho alimentar.

5.3 Trabalho futuro

Apesar dos resultados, derivados da implementação do sistema de recomendação ao retalho alimentar, outras áreas podem ser exploradas e melhoradas.

Uma das hipóteses para estudo futuro envolve a incorporação de dados externos no sistema de recomendação, como informações de redes sociais, dados meteorológicos ou mesmo eventos sazonais. Esses tipos de dados podem fornecer informações adicionais sobre o comportamento do consumidor, permitindo que o sistema ajuste suas sugestões com base em fatores externos, como tendências de mercado ou condições em tempo real.

Em resumo, embora os sistemas de recomendação sejam já alargadamente adotados, existem ainda vastas oportunidades de inovação e melhoria, o trabalho futuro deve centrar-se no reforço da sua eficácia, precisão e impacto a longo prazo no sector retalhista ou qualquer outro setor onde o mesmo se aplique.

Referências

[1] Alexander, R. (2020). How recommendation systems are driving change in the retail sector. Disponível online em https://www.prescouter.com/2020/05/how-recommendation-systems-are-driving-change-in-the-retail-sector/.

[2] Goldberg, J. (2022). E-Commerce Sales Grew 50% to $870 Billion During The Pandemic. Forbes. Disponível em https://www.forbes.com/sites/jasongoldberg/2022/02/18/e-commerce-sales-grew-50-to-870-billion-during-the-pandemic/?sh=322375d84e83

[3] Isinkaye, F.O., Folajimi, Y.O., Ojokoh, B.A. (2015). Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16(3), 261-273.

[4] Lu, J., Wu, D., Mao, M., Wang, W., Zhang, G. (2015). Recommender system application developments: A survey. Decision Support Systems, 74, 12-32.

[5] Low, J., Tan, I., Ting, C. (2019). Recent Developments in Recommender Systems. Retirado de: https://www.researchgate.net/publication/337051634_Recent_Developments_in_Recommender_Systems

[6] The 10 Benefits of a Recommendation Engine: Increasing AOV and More. (n.d.). Retrieved June 18, 2022, from https://kibocommerce.com/blog/recommendation-engine-benefits-aov/

[7] Recommendation System -Understanding The Basic Concepts, n.d. https://www.analyticsvidhya.com/blog/2021/07/recommendation-system-understanding-the-basic-concepts/

[8] The Importance of Recommender Systems | by Commons | Medium, n.d. https://medium.com/@Commons/the-importance-of-recommender-systems-36f86f92181

[9] Sahoo, L., Das, D., & Datta, S. (2017). A Survey on Recommendation System. Article in International Journal of Computer Applications, 160(7), 975–8887. https://doi.org/10.5120/ijca2017913081

[10] Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System. In International Journal of Computer Applications (Vol. 110, Issue 4).

[11] What Are Recommendation Systems in Machine Learning? | Analytics Steps, n.d. https://www.analyticssteps.com/blogs/what-are-recommendation-systems-machine-learning

[12] Netflix Research, n.d. https://research.netflix.com/research-area/recommendations

[13] Gomez-Uribe, C., Hunt, N. (2015). The Netflix Recommender System: Algorithms, Business Value, and Innovation. https://dl.acm.org/doi/pdf/10.1145/2843948

[14] How Retailers Can Keep up with Consumers | McKinsey, n.d.
https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

[15] Davidson, J., Liebald, B., Liu, J., Nandy, P., & van Vleet, T. (2010). The YouTube Video Recommendation System. www.youtube.com/videos.
https://www.inf.unibz.it/~ricci/ISR/papers/p293-davidson.pdf

[16] Lu, J., Wu, D., Mao, M., Wang, W., & Zhang, G. (2015). Recommender system application developments: A survey. Decision Support Systems, 74, 12–32.
https://doi.org/10.1016/J.DSS.2015.03.008

[17] Gavalas, D., Konstantopoulos, C., Mastakas, K., & Pantziou, G. (2014). Mobile recommender systems in tourism. Journal of Network and Computer Applications, 39(1), 319–333.
https://doi.org/10.1016/J.JNCA.2013.04.006

[18] Market Basket Analysis | Guide on Market Basket Analysis, n.d.
https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/

[19] Kutuzova, T., & Melnik, M. (2018). Market basket analysis of heterogeneous data sources for recommendation system improvement. Procedia Computer Science, 136, 246–254.
https://doi.org/10.1016/J.PROCS.2018.08.263

[20] Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the Changing Trends of Market Data Using Association Rule Mining. Procedia Computer Science, 85, 78–85.
https://doi.org/10.1016/J.PROCS.2016.05.180

[21] https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.6295&rep=rep1&type=pdf

[22] Frequent Pattern (FP) Growth Algorithm In Data Mining, n.d.
https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/

[23] Início Da Epidemia de COVID-19 Em Portugal Caracterizado Por Disseminação Massiva de Variante Do SARS-CoV-2 Com Mutação Específica - INSA, n.d. https://www.insa.min-saude.pt/inicio-da-epidemia-de-covid-19-em-portugal-caracterizado-por-disseminacao-massiva-de-variante-do-sars-cov-2-com-mutacao-especifica/

[24] Quais São Os Setores Mais Afetados Pelo Coronavírus? – ECO, n.d.
https://eco.sapo.pt/2020/03/27/quais-sao-os-setores-mais-afetados-pela-coronavirus/

Anexos

Anexo 1: Funcionamento do Algoritmo Apriori

Imaginemos que temos uma tabela em que a primeira coluna representa o número da transação e a segunda o conjunto de itens de ordem K que foram adquiridos. O primeiro passo consiste em calcular o suporte para cada item individual. Com base nessa informação, o analista deve definir qual é o suporte mínimo para que o algoritmo considere apenas os itens mais frequentes, sendo que os itens que têm um suporte abaixo do mínimo estipulado não são considerados nos passos seguintes. Seguidamente repete-se a passagem anterior, mas desta vez adicionando um item de cada vez ao conjunto de itens e descartando os conjuntos de itens que não respeitam o suporte mínimo requerido. Este ciclo de filtragem de itens termina quando não existem mais conjuntos de itens frequentes (suporte do K-itemset < suporte mínimo). A esta etapa dá-se o nome de processo de geração de candidatos e já se está em condições para listar todas as regras de associação dos conjuntos de itens frequentes, calculando o suporte e confiança para cada uma e descartando aquelas que não verificam os limites.

Anexo 2: Caracterização da empresa cliente

Para melhor compreender as vendas do cliente, foi também conduzido um estudo com base mensal para as mesmas variáveis nos dois anos em questão. Para visualizar os resultados, foi necessário proceder a algumas transformações nos datasets de 2019 e 2020:

1. Agrupar os dados por mês através de uma operação de groupby da data com frequência mensal, e somar os valores de todas as transações;
2. Passar para uma lista, os valores de faturação, custo, margem bruta absoluta e quantidade faturada, separadamente;
3. Criar uma lista com o nome de todos os meses;
4. Criar um dicionário Mês/Faturação, através das listas;
5. Com o dicionário, criar o dataframe final;
6. Proceder à visualização dos resultados

A Figura 6 apresenta um gráfico comparativo da faturação mensal nos anos de 2019 e 2020.

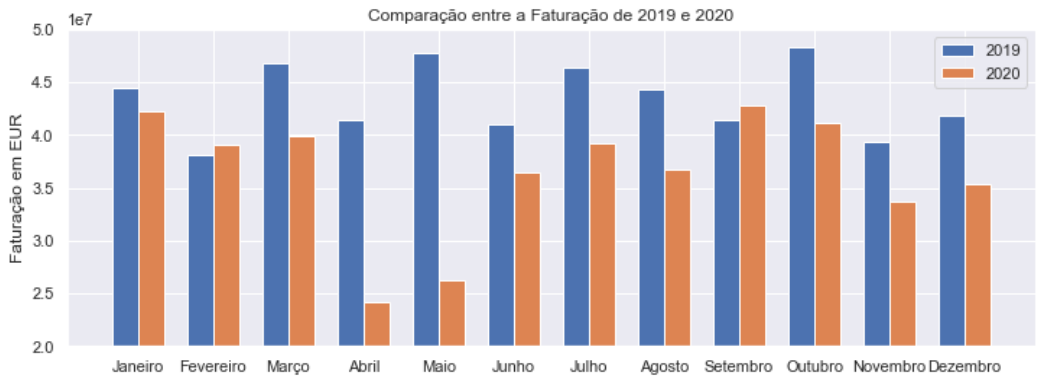


Figura 6: Faturação mensal em 2019 e 2020

Como se pode observar no gráfico, em 2020 apenas os meses de fevereiro e setembro (mês com maior faturação) apresentaram valores de faturação ligeiramente superiores a 2019(39.091.278,18€ em 2019 e 42.821.622,76€ em 2020). Para além disto, é de realçar o fraco desempenho nos meses de abril (faturação de 24.172.701,64€) e maio (faturação de 26.306.738,25€) de 2020. Estes valores são expectáveis pois o período entre 14 de março e 9 de abril caraterizou-se pelo aumento exponencial da COVID-19 em Portugal [23], tendo sido o comércio um dos setores da economia mais afetados pela pandemia [24]. Quanto a 2019, destacam-se pela positiva os meses de março, maio, julho e outubro que atingiram valores de faturação acima dos 45 milhões de euros.

Com o intuito de realizar uma análise do custo mensal, foi feito um gráfico comparativo dos dois anos, apresentado na Figura 7.

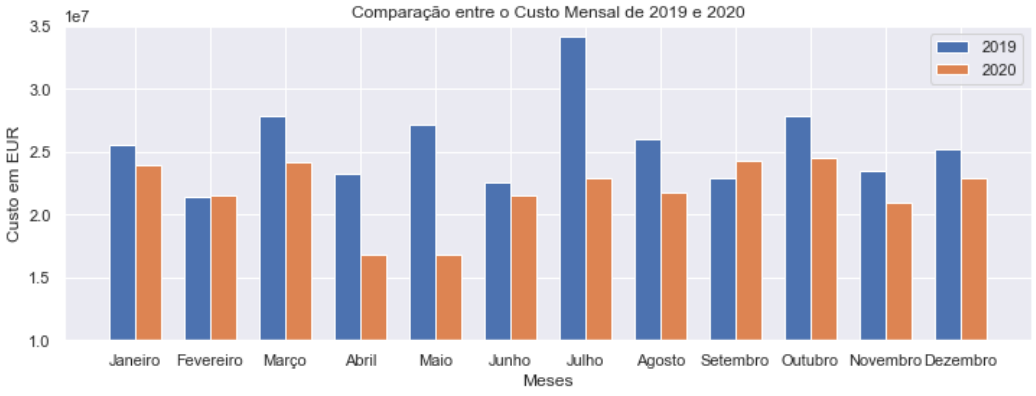


Figura 7: Custo mensal em 2019 e 2020

Relativamente ao custo total dos produtos, também se observa uma diferença mensal entre os dois anos, no entanto não é tão acentuada como da faturação. Em 2019, é de realçar o mês de julho que apresentou custos totais dos produtos bastante superiores a todos os outros meses (34.127.680,26€) e o mês de fevereiro com custos mais baixo (21.456.290,88€). O único mês que evidenciou custos mais reduzido em 2019 comparado com 2020 foi fevereiro, com uma ligeira diferença de 126.698,84€. Em 2020, tal como foi observado na análise da faturação, pode-se constatar que os meses de abril e maio foram caracterizados por custos significativamente mais reduzidos quando comparados com os restantes meses do ano. A Figura 3 apresenta o gráfico comparativo da margem bruta absoluta mensal.

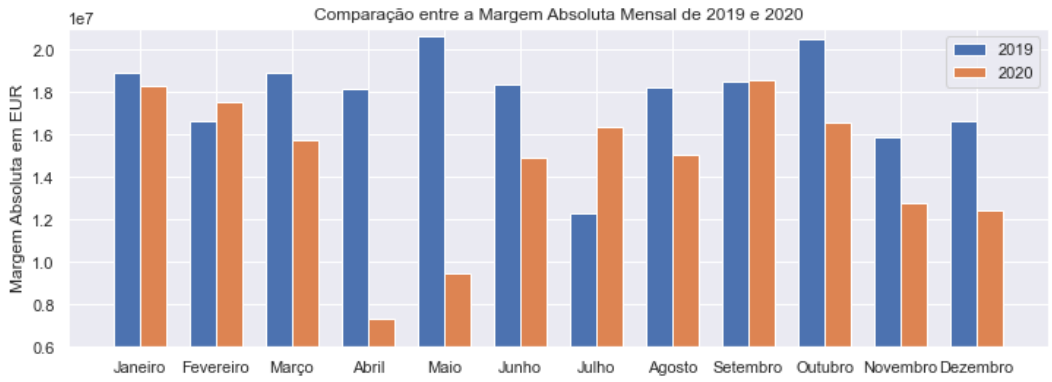


Figura 8: Margem bruta absoluta mensal em 2019 e 2020

Observando o gráfico, é relevante mencionar a ampla diferença na maior parte dos meses comparando os dois anos, nomeadamente os meses de abril e maio que apresentaram quedas de 82,3% e 54,1%, respetivamente. Com a exceção de fevereiro, julho e setembro, todos os restantes meses manifestaram valores de margens inferiores em 2020, comparativamente com 2019.

Concluindo, claramente o ano de 2019 apresentou resultados mais favoráveis para a empresa, com a exceção de alguns meses. Podemos constatar que a pandemia teve um impacto significativo na faturação e custos dos produtos.

Anexo 3: Top 10 Produtos e Top 10 Clientes com maior margem bruta

Nº Material	Margem Absoluta EUR	Margem Absoluta %
5014000	67323654.61	17.3
5015000	59325931.49	15.3
5010000	24264575.23	6.2
5028323	13228629.02	3.4
6254045	12586948.14	3.2
5011010	11503387.42	3.0
600910	8640943.68	2.2
5018000	8593431.96	2.2
5034001	6083261.96	1.6
5028324	5688975.87	1.5

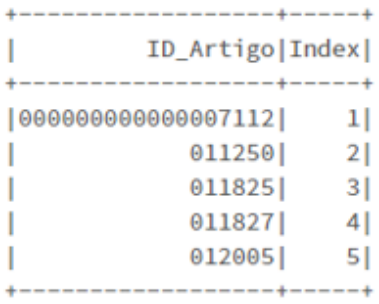
Tabela 8: Top 10 Produtos com maior margem bruta

Nº Cliente	Margem Absoluta EUR	Margem Absoluta %
103872	18251105.84	4.7
9000002	15190897.74	3.9
91723	14315288.63	3.7
708176	7221144.52	1.9
28311	6971350.95	1.8
325	5641243.82	1.5
9000010	4541786.69	1.2
38371	4302547.66	1.1
9000011	3946764.76	1.0
9000013	3045507.76	0.8

Tabela 9: Top 10 Clientes com maior margem bruta

Anexo 4: Afetação do novo index ao código de produto

A figura seguinte pretende ilustrar, para apenas 5 produtos, como foi realizada a afetação do código de produto (variável ID_Artigo) ao novo index de produto (variável Index).



ID_Artigo	Index
0000000000007112	1
011250	2
011825	3
011827	4
012005	5

Figura 9: Afetação do novo index ao código de produto