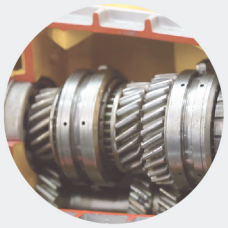
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA E GESTÃO INDUSTRIAL






ANÁLISE DE PADRÕES DE CONSUMO E DESENVOLVIMENTO DE UM SISTEMA DE RECOMENDAÇÃO DE PRODUTOS

JOÃO MIGUEL LEITE MARTINS
setembro de 2023

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

ANÁLISE DE PADRÕES DE CONSUMO E DESENVOLVIMENTO DE UM SISTEMA DE RECOMENDAÇÃO DE PRODUTOS

João Miguel Leite Martins

2023

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

P.PORTO


P.PORTO

Instituto Superior de Engenharia do Porto

ANÁLISE DE PADRÕES DE CONSUMO E DESENVOLVIMENTO DE UM SISTEMA DE RECOMENDAÇÃO DE PRODUTOS

João Miguel Leite Martins

Estudante n.º 1210190

Dissertação apresentada ao Instituto Superior de Engenharia do Porto para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia e Gestão Industrial, realizada sob a orientação do Doutor Carlos Manuel Abreu Gomes Ferreira.

2023

Instituto Superior de Engenharia do Porto

Departamento de Engenharia Mecânica

P.PORTO

AGRADECIMENTOS

Em primeiro lugar, gostaria de conferir os meus agradecimentos à instituição que me acolheu nestes dois anos de mestrado, o Instituto Superior de Engenharia do Porto. Com especial atenção para todos os professores que fizeram parte do meu percurso académico, contribuindo assim para o meu crescimento a nível profissional e pessoal.

Ao Doutor Carlos Manuel Abreu Gomes Ferreira, meu orientador académico, pelas suas recomendações e pela sua flexibilidade na disponibilidade para me ajudar nos momentos críticos.

A todos os meus amigos que me apoiaram e apoiam diariamente em tudo a que me proponho, sem eles a vida não teria o mesmo significado.

Para finalizar, mas acima de tudo, quero deixar a minha imensa gratidão aos meus pais, Rui Martins e Carla Leite. Respeito todo o trabalho que tiveram para me tornar no homem que sou hoje e por me terem proporcionado as ferramentas necessárias para que eu pudesse atingir os meus objetivos.

página propositadamente em branco

RESUMO

Atualmente, no setor do retalho, as estratégias inovadoras são fundamentais para apoiar os retalhistas a destacaram-se da concorrência e a atender às necessidades dos clientes. A integração eficaz de sistemas de recomendação baseados em regras de associação no retalho, não só permite os retalhistas conhecerem melhor os seus clientes, permitindo criar recomendações personalizadas, mas também fazer uma gestão mais otimizada dos seus inventários, garantindo que os produtos estejam disponíveis no tempo e lugar certo, quando os clientes os desejam.

No mercado atual, existe uma enorme diversidade de opções disponíveis para atender às necessidades dos clientes, tornando imperativa a necessidade de os retalhistas estabelecerem estratégias precisas e ponderadas que os permitem destacar da concorrência. Deste modo, nesta presente dissertação, conduziu-se uma investigação com o intuito de identificar padrões de consumo baseados em regras de associação e sequências frequentes, a partir de um conjunto de dados disponibilizado pela plataforma *kaggle*, denominado de *Instacart Market Basket Analysis.* Além disso, foi desenvolvido um sistema de recomendação baseado nas regras de associação identificadas e diversos modelos de *Machine Learning* foram aplicados de forma a selecionar o mais adequado para prever a recompra de produtos por parte dos clientes.

Numa primeira fase, o processo para encontrar as regras de associação presente no conjunto de dados começou com uma definição clara dos objetivos de investigar os padrões de consumo e relacionamento entre os produtos. De seguida, a ferramenta selecionada para realizar a análise foi o *Knime*, que possui um nó designado de "Association Rule Learner", baseado no algoritmo *Apriori*, para descobrir a relação entre os produtos. Diferentes limites mínimos de suporte foram utilizados para identificar associações significativas. O foco recaiu sobre as regras de associação que apresentaram elevados valores das métricas de suporte, confiança e *lift*, uma vez que tais associações se demonstraram mais significativas do que as que ocorriam ao acaso. Para o desenvolvimento do sistema de recomendação baseado em regras de associação foram realizados dois estudos de precisão, um com base em altos níveis de confiança e suporte e outro com base em altos níveis de *lift* e suporte. Para ambos os estudos foram removidos produtos de forma aleatória para testar a eficácia do sistema. Por último, diversos modelos de *Machine Learning* foram aplicados com o objetivo de identificar o que oferecia melhor precisão, a fim de ser selecionado para prever a recompra de produtos.

Com base nas regras de associação encontradas, foram sugeridas diversas estratégias de negócio, como promoções e posicionamento estratégico dos produtos com o objetivo de incentivar os clientes a comprar determinado produto e, por consequência, a impulsionar o volume de vendas. O sistema de recomendação baseado em métricas de confiança e suporte de valores altos demonstrou ser o mais eficaz na previsão correta dos produtos, pelo que seria o selecionado para entrar em produção. O modelo *Stacking* revelou ser o modelo mais eficaz, com uma precisão de 75%, tornando-o a escolha preferencial para prever se um cliente irá adquirir novamente um determinado produto.

PALAVRAS-CHAVE

Retalho, Associação, Recomendação, Modelos, *Machine Learning,* Precisão

página propositadamente em branco

ABSTRACT

In today's retail sector, innovative strategies are key to helping retailers stand out from the competition and meet their customers' needs. The effective integration of recommendation systems based on association rules in retail not only allows retailers to get to know their customers better, enabling them to create personalized recommendations, but also to manage their inventories more optimally, ensuring that products are available at the right time and place, precisely when customers require them.

In today's market, there is a huge diversity of options available to meet customer needs, making it imperative for retailers to establish precise and thoughtful strategies that allow them to stand out from the competition. In this way, this dissertation conducted an investigation with the aim of identifying various association rules and frequent sequences, based on a set of data made available by the kaggle platform, called Instacart Market Basket Analysis. In addition, a recommendation system was developed based on the association rules identified, and several Machine Learning models were applied in order to select the most appropriate one to predict product repurchase by customers.

Initially, the process of finding the association rules present in the data set began with a clear definition of the objectives of investigating consumption patterns and relationships between products. Next, the tool selected to carry out the analysis was Knime, which has a node called "Association Rule Learner", based on the Apriori algorithm, to discover the relationship between products. Different minimum support thresholds were used to identify significant associations. The focus was on association rules that had high values of the support, confidence and lift metrics, since these associations proved to be more significant than those that occurred randomly. To develop the recommendation system based on association rules, two accuracy studies were carried out, one based on high levels of confidence and support and the other based on high levels of lift and support. For both studies, products were randomly removed to test the effectiveness of the system. Finally, several Machine Learning models were applied with the aim of identifying the one that offered the best accuracy to be selected to predict product repurchase.

Based on the association rules found, various business strategies were suggested, such as promotions and strategic positioning of products with the aim of encouraging customers to buy a particular product and, consequently, boosting sales volume. The recommendation system based on confidence metrics and high-value support proved to be the most effective in correctly predicting products. Therefore, was the one selected to go into production. The Stacking model proved to be the most effective model, with an accuracy of 75%, making it the preferred choice for predicting whether a customer will purchase a particular product again.

KEYWORDS

Retail, Association, Recommendation, Models, Machine Learning, Accuracy

página propositadamente em branco

ÍNDICE

página propositadamente em branco

ÍNDICE DE FIGURAS

página propositadamente em branco

ÍNDICE DE TABELAS

página propositadamente em branco

LISTAS DE SIGLAS E SÍMBOLOS

Lista de Siglas

ISEP	Instituto Superior de Engenharia do Porto
P.Porto	Instituto Politécnico do Porto
CRISP-DM	*Cross Industry Standard Process for Data Mining*
RFM	Recência, Frequência e Monetário
KNN	*K-nearest neighbors*
FP-Growth	*Frequent Pattern Growth*
KDD	*Knowledge Discovery in Database*
OCVR	*Overall Variability of Association Rule*
CSV	*Comma-separated values*
MBA	*Market Basket Analysis*
MLP	*Multilayer Perceptrons*

página propositadamente em branco

1. INTRODUÇÃO

Neste capítulo será fornecida uma breve descrição do contexto em que o projeto será realizado bem como a relevância do tema. Seguidamente, é enunciado a questão que dá origem à necessidade desta investigação, os objetivos da mesma e as metodologias utilizadas no desenvolvimento deste projeto.

1.1. Problema de investigação, enquadramento e pertinência

O retalho refere-se à venda direta de bens e serviços aos consumidores para uso pessoal. As vendas podem ser realizadas a partir de lojas físicas, plataformas online ou até mesmo vendedores ambulantes.

O retalho tem tido um papel preponderante na economia. Todavia, nos últimos anos, tem vindo a sofrer alterações significativas devido à evolução acelerada das tecnologias.

O surgimento da implementação de sistemas de recomendação e regras de associação tem vindo a permitir aos retalhistas conhecerem melhor os seus clientes, os seus hábitos de consumo e, por sua vez realizar recomendações personalizadas para os mesmos (Konstan & Riedl, 2012).

Os sistemas de recomendação utilizam algoritmos de *Machine Learning* para analisar dados relacionados com o histórico de compras dos clientes e outras informações relevantes de forma a realizar recomendações de produtos ou serviços que eles possam estar interessados. Os dados utilizados pelo sistema de recomendação podem surgir de diversas fontes, nomeadamente através de websites, informação demográfica ou até mesmo de feedback explícitos fornecido pelos próprios utilizadores sobre determinado produto (Kane, 2018).

Regras de associação são utilizadas de forma a identificar padrões e relações dentro de um conjunto de dados. Por exemplo, se um retalhista perceber que os clientes que compram cerveja também tendem a comprar amendoins eles podem usar essa informação para fazer recomendações para clientes que só compraram cerveja.

Com o surgimento do comércio online, existe uma enorme variedade de informação disponível acerca do comportamento e das preferências dos consumidores. Esses dados podem ser utilizados para criar recomendações personalizadas bem como posicionar promoções de forma estratégica, impulsionando as vendas e melhorando a experiência de compra dos consumidores.

Com o crescente aumento da concorrência existe uma enorme variedade de opções de compra disponíveis para os clientes, por esse motivo os retalhistas precisam de definir estratégias para se destacarem da concorrência. Os sistemas de recomendação e as regras de associação são ferramentas cruciais para o sucesso dos retalhistas dado que o uso das mesmas permitirá melhorar a experiência de consumo, reter clientes e impulsionar a rentabilidade do negócio.

1.2. Questão e objetivos de investigação

Graças à crescente necessidade de inovação e de implementação de novas soluções no mundo do retalho, é necessário conhecer os padrões de consumo dos consumidores para ir de encontro com as necessidades dos mesmos. Daqui surge a necessidade da criação de um modelo que permita

efetuar uma sugestão de compra, encontrando sequências e tipos de produtos comprados conjuntamente pelos clientes.

Deste modo, espera-se como resultado desta dissertação o desenvolvimento de um modelo que permita efetuar sugestões de produtos com base nas compras dos consumidores e nos padrões de consumo encontrados, com os seguintes objetivos específicos: implementar e seguir as etapas da metodologia CRISP-DM, utilizando a ferramenta *Knime* para elaborar resultados, encontrar regras de associação, encontrar sequências frequentes, desenvolver o modelo de recomendação de produtos para os diferentes tipos de clientes e comparar diferentes modelos de *Machine Learning* e selecionar o mais adequado para entrar em produção.

1.3. Opções metodológicas

De acordo com o autor Kothari (2004, p.1), Clifford Woody mencionou que "a investigação compreende a definição e redefinição de problemas, a formulação de hipóteses ou soluções, a recolha, organização e avaliação de dados, a realização de deduções e de soluções".

Para uma investigação ser reconhecida por ter boa qualidade deve ser realizada de uma forma sistemática, isto é, deve seguir um conjunto específico de regras e passos numa ordem específica. A pesquisa deve ser orientada pelos princípios do raciocínio lógico. Empírica, ou seja, é baseada em situações do mundo real e trata de dados práticos e concretos. Replicável, isto é, pode ser verificada por meio de repetição do estudo, o que auxilia a fornecer uma base sólida para a tomada de decisão (Com, 2019).

Segundo Com (2019), existem duas abordagens possíveis numa investigação: qualitativa e quantitativa. A abordagem qualitativa concentra-se na avaliação subjetiva de atitudes, opiniões e comportamentos. Este tipo de pesquisa é realizado com base nas perceções e impressões do pesquisador e gera resultados que podem não ser quantificáveis ou adequados para uma análise quantitativa rigorosa. A abordagem quantitativa trata-se da recolha de dados que podem ser analisados, utilizando métodos formais e rigorosos.

Esta dissertação será realizada numa perspetiva quantitativa, na medida que se trata de um estudo objetivo, baseado em análise de dados e implementação de métodos, capazes de gerar resultados mensuráveis e de possível análise estatística.

De acordo com (Fernandes et al., 2020) existem diversas estratégias que compõem o plano de investigação, como investigação experimental, inquérito, estudo de caso e investigação-ação. Nesta dissertação o método de investigação utilizado é a investigação-ação. A investigação-ação visa solucionar problemas práticos e reais na organização, com a envolvência ativa do investigador na resolução (Fernandes et al., 2020). É uma abordagem metodológica que combina ação e investigação, utilizando um processo cíclico que alterna entre ação e reflexão crítica (Coutinho, 2011). Neste projeto, diferentes modelos de recomendação serão analisados, passando por um processo crítico de revisão, sendo selecionado o mais adequado para entrar em produção.

É importante referir que nesta dissertação a metodologia CRISP-DM será utilizada de forma a converter os dados em informações úteis, capazes de auxiliar na tomada de decisão, sendo um dos processos mais utilizados para efetuar projetos de análise preditiva de dados (Kelleher et al., 2015).

A ferramenta *Knime* será a ferramenta utilizada para a elaboração dos resultados. *Knime* é uma plataforma de análise de dados gratuita que permite que os utilizadores criem visualmente fluxos de dados, integrem várias fontes de dados e realizem tarefas de manipulação, transformação e análise de dados (Berthold et al., 2007).

2. REVISÃO BIBLIOGRÁFICA

De forma a familiarizar o leitor com a indústria do retalho, o presente capítulo começa por expor e destacar as principais caraterísticas desse setor. De seguida, com o intuito de esclarecer a metodologia adotada nesta dissertação, são delineados os aspetos fundamentais do método CRISP-DM, seguindo-se da apresentação de conceitos e metodologias *Machine Learning*, regras de associação, sistemas de recomendação e análise RFM. No que diz respeito ao trabalho relacionado são expostos diferentes artigos para destacar o que já foi feito de importante na área em questão bem como os resultados obtidos.

2.1. Retalho

Berman, Evans e Chatterjee (2018) referem que "A venda a retalho engloba as atividades comerciais envolvidas na venda de bens e serviços aos seus consumidores para uso pessoal, familiar ou doméstico".

O retalho não inclui apenas a venda de bens tangíveis, mas também a venda de serviços e bens digitais. O retalho é o processo de venda de bens e serviços a consumidores, diretamente ou através de intermediários. Ele desempenha um papel significativo na economia pois emprega um grande número de pessoas e, em termos monetários, em 2015 as vendas representaram 4,785 triliões de dólares nos Estados Unidos, o que representou um terço da economia total.

O retalho é a última etapa na distribuição de bens e serviços (figura 1).

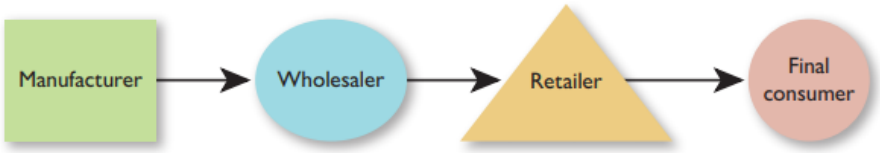


Figura 1- Cadeia de distribuição de bens e serviços (Berman et al., 2018)

Os retalhistas muita das vezes atuam como intermediários entre fabricantes e consumidores, possuindo uma grande quantidade de produtos de diversas fontes com o propósito de os vender em menores quantidades (figura 2). Servem também de meio de comunicação entre os fabricantes e o cliente, fornecendo informações sobre a disponibilidade e caraterísticas dos bens e serviços, permitindo assim que os fabricantes alcancem mais clientes, reduzam custos, melhorem o seu fluxo de caixa e aumentam as suas vendas (Berman et al., 2018).

Atualmente, os consumidores abordam as suas compras de maneira diferente, dado ao crescimento do comércio eletrónico e à crescente utilização da tecnologia no processo de compra. Por esse motivo, os retalhistas precisam de planear e se adaptar às mudanças. O domínio do canal online é evidente, o que levou ao crescimento de retalhistas online e estratégias omnicanal por parte dos retalhistas tradicionais de lojas físicas. Os clientes não compram apenas online, eles também vão ao online para iniciar o processo de compra e verificar o que os outros compradores

pensam sobre determinado produto o que demonstra que é impensável descartar o meio online como um ponto de venda (Berman et al., 2018).

Omnicanal refere-se ao desenvolvimento de "uma abordagem integrada em toda a operação do retalho que oferece uma resposta sem falhas à experiência do consumidor através de todos os canais de compra disponíveis, quer seja em dispositivos móveis de Internet, computadores, lojas, televisão e catálogos" (Saghiri et al., 2017).

De forma a serem competitivos os retalhistas devem constantemente se adaptar às mudanças nas necessidades e preferências dos consumidores. O investimento em novas tecnologias, como Inteligência Artificial e Machine Learning, bem como a análise e uso de dados é determinante para compreender o comportamento do consumidor e identificar oportunidades de crescimento.

A segmentação de base de clientes fundamentada em *Big data* é uma estratégia utilizada para dividir os clientes em grupos distintos com base em dados e informações coletadas sobre eles. A segmentação de clientes pode ser uma estratégia valiosa para as empresas, pois permite um melhor entendimento dos seus clientes e uma melhor personalização no atendimento dos mesmos (Berman et al., 2018).

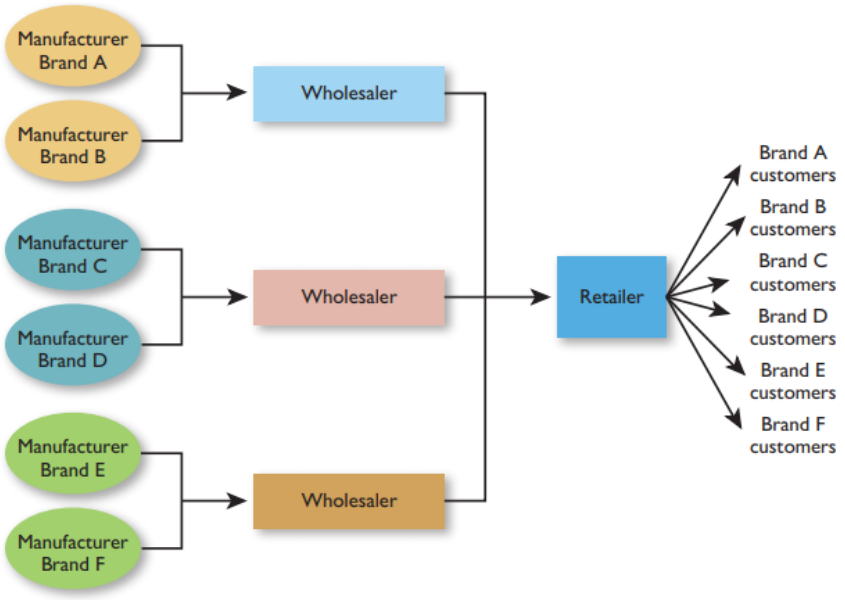


Figura 2- O papel do retalhista na venda aos consumidores (Berman et al., 2018)

2.2. CRISP-DM

A metodologia *Cross Industry Standard Process for Data Mining* (CRISP-DM), foi criada em 1996 pelas empresas DaimlerChrysler, atualmente conhecida por Mercedez-Benz Group, SPSS e NCR.

Estas empresas sentiram a necessidade de criar um modelo de processo-padrão, disponível de forma gratuita e que não pertencesse a nenhuma empresa em particular. Esta metodologia seria

capaz de dar resposta aos problemas de aprendizagem por tentativa e erro e, consequentemente, demonstrar um certo grau de maturidade diante de potenciais clientes (Chapman et al., 2000).

Com financiamento da Comissão Europeia, estas empresas formaram um consórcio e criaram o "CRISP-DM Special Interest Group", de forma a recolher diferentes opiniões de uma ampla rede de profissionais (Chapman et al., 2000).

CRISP-DM é uma sequência de seis etapas que começa pelo entendimento do problema empresarial (ou organizacional) que está a ser tratado, de forma a conceber uma solução analítica para a mesma e termina com a implantação da solução que satisfaz a necessidade específica do problema.

A primeira etapa, designada por Compreensão do negócio (*Business Understanding*), como o próprio nome indica, trata-se da parte de entender o problema empresarial em questão (Kelleher et al., 2015). Nesta etapa deve ser desenvolvido um plano para recolher e analisar os dados e um orçamento deve ser estabelecido para dar suporte ao estudo desenvolvido (Turban et al., 2013).

A segunda etapa, designada por Compreensão dos dados (*Data Understanding*), é a etapa onde devem ser identificados e compreendidos os dados disponíveis nos diferentes conjuntos de dados em questão, bem como o tipo de dados que estão contidos nessas fontes (Kelleher et al., 2015).

A terceira etapa, designada por Preparação dos dados (*Data Preparation*) tem como objetivo converter os dados num formato adequado para a fase de modelagem (Kelleher et al., 2015). A preparação dos dados é uma etapa importante no processo de mineração dos dados, pois pode impactar significativamente a qualidade e precisão dos modelos

A quarta etapa, conhecida por Modelação (*Modeling*) é a etapa onde diferentes modelos são selecionados e aplicados ao conjunto de dados em estudo (Kelleher et al., 2015).

A quinta etapa, designada por Avaliação (*Evaluation*) é a fase onde os diferentes modelos aplicados são avaliados de forma a garantir que são adequados para o efeito. "Esta fase do CRISP-DM abrange todas as tarefas de avaliação necessárias para demonstrar que um modelo de previsão será capaz de realizar previsões precisas e que não sofrerá de *overfitting* ou *underfitting*" (Kelleher et al., 2015, p.53).

Por último, a sexta etapa, designada por Implantação (*Deployment*) envolve todo o trabalho necessário para integrar com sucesso um modelo de Machine Learning nos processos de uma organização.

2.3. Machine Learning

Segundo Géron (2017) *Machine Learning* é uma subárea da Inteligência Artificial que envolve o uso de algoritmos e modelos estatísticos para permitir que os computadores aprendam e façam previsões ou decisões sem precisar de serem explicitamente programados para isso. Para que tal aconteça, é necessário efetuar um treino de um modelo num conjunto de dados para que o modelo possa usar esses dados para aprender como realizar uma tarefa específica.

Um exemplo da aplicação do *Machine Learning* é na previsão do comportamento dos clientes: Onde um algoritmo pode ser treinado com dados históricos sobre as compras dos clientes para efetuar previsões sobre o que um determinado cliente é mais propenso a comprar no futuro.

Os modelos em *Machine Learning* são utilizados para efetuar previsões ou obter conhecimento a partir de dados. Os modelos podem ser preditivos ou descritivos, ou seja, são usados para efetuar previsões sobre eventos futuros ou são utilizados para entender e explicar padrões e relacionamentos existentes nos dados, respetivamente.

A eficiência do modelo, incluindo a sua complexidade, é importante em determinadas aplicações, pois determina a rapidez e eficácia com que o modelo pode ser usado para prever ou extrair conhecimento (Alpaydin, 2014).

Existem vários tipos de algoritmos de *Machine Learning* devido ao seu amplo uso. A tabela 1 descreve, de forma resumida, os principais tipos de aprendizagem relacionados a esta tecnologia.

Tabela 1- Descrição dos diferentes tipos de aprendizagem

Tipo de aprendizagem	**Descrição**
Aprendizagem supervisionada	O algoritmo é treinado com um conjunto de dados já conhecidos, ou seja, os dados incluem tanto as caraterísticas de entradas como as caraterísticas correspondentes de saída. O objetivo da aprendizagem supervisionada é, a partir do modelo construído, gerar um conjunto de previsões mais precisas quando são introduzidos novos dados de entrada (Mohri et al., 2018).
Aprendizagem não supervisionada	Os dados são agrupados em conjuntos semelhantes, de forma a se correlacionarem entre si. Este tipo de aprendizagem tenta encontrar um padrão de forma a classificar corretamente a informação recebida. Neste tipo de aprendizagem pode ser difícil avaliar o desempenho do algoritmo, pois não existem exemplos iniciais para comparar com as previsões realizadas (Mohri et al., 2018).
Aprendizagem semi-supervisionada	O algoritmo é treinado com um conjunto de dados que inclui informação conhecida e desconhecida. O modelo deverá utilizar os dois tipos de informação para efetuar previsões para correspondências de saída ainda desconhecidas (Mohri et al., 2018).
Aprendizagem por reforço	O algoritmo recolhe informações através das suas interações com o ambiente. Neste tipo de aprendizagem existe um agente associado capaz de observar o ambiente que, posteriormente, seleciona e realiza ações de forma a obter recompensas (ou penalidades na forma de recompensas negativas). Com base nestas ações é gerado um modelo com a melhor estratégia definida, de forma a obter a maior recompensa ao longo do tempo (Géron, 2017).

Os tipos de aprendizagem são frequentemente divididos em análises descritivas (aprendizagem não supervisionada) e análises preditivas (aprendizagem supervisionada). Dado que esta dissertação está relacionada com análises preditivas e análises descritivas, na secção 2.2.1 e 2.2.3 será

explicado em maior detalhe a aprendizagem supervisionada e aprendizagem não supervisionada, respetivamente.

2.3.1. Aprendizagem supervisionada

Aprendizagem supervisionada é um tipo de aprendizagem de *Machine Learning* onde o algoritmo é treinado com um conjunto de dados já conhecidos, ou seja, os dados incluem tanto as caraterísticas de entradas como as caraterísticas correspondentes de saída. O objetivo é aprender uma função capaz de mapear as caraterísticas de entrada para as saídas corretas (Mohri et al., 2018). Esses modelos preditivos são instruídos de forma clara sobre o que devem aprender e como o devem fazer. O conjunto de treino contém um conjunto completo de dados, incluindo os valores alvo de saída, que não estão presentes no conjunto de teste (Mohri et al., 2018).

Segundo Liu e Wu (2012), sendo o conjunto (x_i,y_i) uma amostra de treino, onde **x** representa a entrada do sistema, **y** representa a saída do sistema e *i* representa o índice da amostra: Uma entrada de treino $\mathbf{x}_i$ é fornecida ao sistema de aprendizagem e o sistema gera um output $\tilde{\mathbf{y}}_i$. O $\tilde{\mathbf{y}}_i$ é então comparado com o valor real armazenado em $\mathbf{y}_i$ em que a diferença entre eles é calculada. A diferença, designada de *Error Signal* é então enviada para o Sistema de aprendizagem para ajustar os parâmetros do elemento que está a ser treinado. O objetivo deste processo de aprendizagem é obter um conjunto de parâmetros ótimos que possam minimizar as diferenças entre $\tilde{\mathbf{y}}_i$ e $\mathbf{y}_i$, para todos os *i*, ou seja, minimizar o erro total ao longo de todo o conjunto de dados. É de notar que um erro de treino mínimo não indica necessariamente um bom desempenho nos testes, isto porque os dados de treino são utilizados somente para treinar o modelo. O que dita se o modelo é capaz são os dados de teste que não foram incluídos na etapa anterior.

A figura 3 ilustra a forma de utilização da aprendizagem supervisionada.

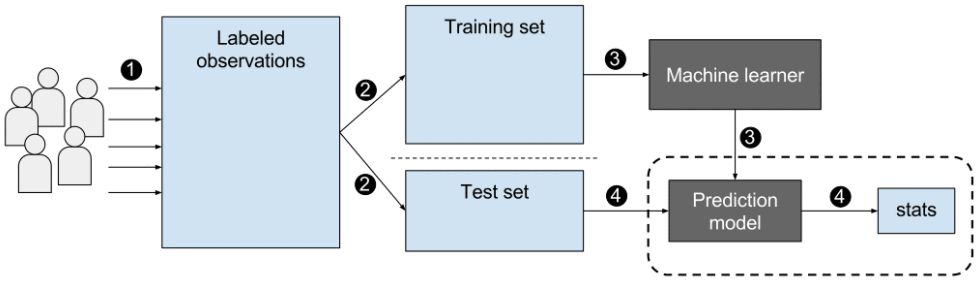


Figura 3- Aprendizagem supervisionada (Salian, 2018)

Existem dois principais tipos de aprendizagem supervisionada: Classificação e Regressão.

2.3.1.1. Regressão

Nos tipos de regressão o objetivo é prever o valor de uma variável numérica contínua, como por exemplo o preço de um carro. O algoritmo é treinado com um conjunto de dados já conhecidos, onde as caraterísticas de entrada são usadas para prever a correspondência numérica de saída (Géron, 2017).

2.3.1.2. Classificação

Nos tipos de classificação o objetivo é prever uma classe ou categoria. Esses modelos são utilizados para resolver problemas de classificação, nos quais o objetivo é prever a classe ou categoria à qual o novo conjunto de dados pertence, com base em exemplos previamente conhecidos. O exemplo comumente conhecido para descrever os modelos de classificação é o caso de prever se um e-mail é "spam" ou "não spam" (Géron, 2017). Dentro do tipo de aprendizagem supervisionada, de modelos de classificação, existem diversos algoritmos de *Machine Learning* que podem ser utilizados, nomeadamente árvores de decisão (*Decision trees*), *K-Nearest Neighbours* e *Multilayer perceptrons.*

K-Nearest Neighbours

O K-nearest neighbours (KNN) é um método simples e eficaz, que apesar da sua simplicidade, é um algoritmo capaz de gerar resultados com um bom desempenho (Sun & Huang, 2010).

O KNN utiliza um número "k", definido pelo utilizador, que representa os "k" vizinhos mais próximos de um dado desconhecido que se pretende classificar. Para isso, o algoritmo calcula a distância entre o dado desconhecido e os "k" pontos de dados mais próximos recorrendo ao uso da métrica "Distância Euclidiana" (Hastie et al., 2009). Se os "k" pontos de dados mais próximos pertencerem a classes diferentes, o algoritmo atribuirá o dado desconhecido à classe mais comum entre esses pontos (Witten et al., 2011).

O KNN consiste em duas fases: Determinação dos vizinhos mais próximos e determinação da classe a que pertence (Cunningham & Delany, 2020).

A figura 4 representa de uma forma simplificada o algoritmo KNN.

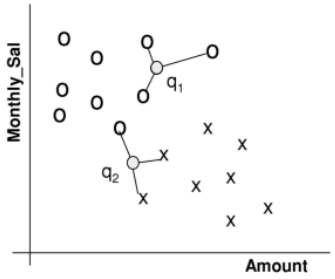


Figura 4- Representação simplificada do algoritmo KNN (Cunningham & Delany, 2020)

Árvore de decisão

O algoritmo de árvore de decisão consiste num modelo de aprendizagem supervisionada, aplicável tanto para tarefas de classificação como para regressão, que divide um conjunto de dados em subgrupos cada vez menores com base em uma série de divisões binárias, com o objetivo de criar um modelo capaz de realizar previsões sobre novos dados.

A árvore de decisão é composta por nós de decisão internos que indicam uma decisão a ser tomada sobre um atributo, ramos que representam uma regra de decisão e folhas terminais que representam o resultado. Em cada nó é aplicado um teste aos dados para determinar qual será o próximo ramo a seguir. O processo continua até que se alcance uma folha, momento este em que é produzido um output (um código de classe no caso de modelos de classificação ou um valor numérico no caso de regressão) (Alpaydin, 2014).

Os subconjuntos criados pelas divisões são designados de nós. Os subconjuntos que não são divididos são designados de nós terminais (Leo Breiman et al., 1987).

As árvores de decisão são não paramétricas, o que significa que elas não assumem nenhuma forma em particular. Cresce e adapta-se à complexidade dos dados conforme estes são aprendidos (figura 5).

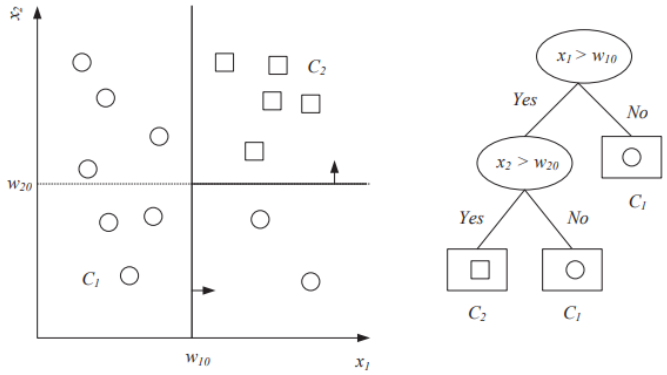


Figura 5- Conjunto de dados e sua árvore de decisão correspondente (Alpaydin, 2014)

Perceptrons e Multilayer perceptrons

De acordo com Nielsen (2015) uma rede neuronal é um sistema de algoritmos que foi projetado com o intuito de reconhecer padrões. É composto por neurónios artificiais que são inspirados na estrutura e função dos neurónios no cérebro humano.

Um tipo de neurónio artificial é chamado de *perceptron*, que foi desenvolvido na década de 1950 e 1960 por Frank Rosenblatt. Consiste num único neurónio com múltiplas entradas (figura 6), em que cada uma tem um peso associado. A saída do neurónio é determinada pela soma ponderada dos pesos das entradas, que depois é comparada com o valor estipulado (*threshold value*). Se o resultado for maior que o valor estipulado, o valor da saída do neurónio é 1, caso contrário o valor é 0.

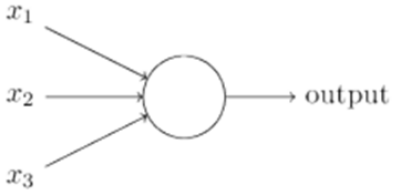


Figura 6- Rede neuronal simples (Nielsen, 2015)

Os *perceptrons* possuem várias entradas binárias x_1, x_2, …, x_n e produzem uma única saída binária, y.

$$y = \begin{cases} 0, & \Sigma_j w_j \cdot x_j \leq threshold \\ 1, & \Sigma_j w_j \cdot x_j > threshold \end{cases} \qquad (2.1)$$

Segundo Gardner & Dorling (1998) o *multilayer perceptron* é um tipo de rede neuronal artificial que é composta por nós/neurónios simples, interconectados. Estes neurónios estão conectados por pesos e transmitem sinais entre si conforme a informação que recebem.

O *Perceptron* de múltiplas camadas é organizado em camadas, em que a camada mais à esquerda observada na figura 7 é a camada de entrada e os neurónios presentes nessa camada são designados de neurónios de entrada. A camada mais à direita é a camada de saída, que contém o único neurónio de saída. A camada do meio é designada de camada oculta, dado que nesta camada não existem entradas nem saídas (Nielsen, 2015).

O número de camadas ocultas e nós pode variar consoante o problema em questão. Uma maior quantidade de nós oferece uma maior sensibilidade ao problema, porém aumenta o risco de *overfitting* [1](Murtagh, 1991).

A camada de entrada atua como um condutor para os dados de entrada e a camada de saída produz a saída final da rede.

As camadas ocultas processam os dados de entrada e os transmitem para as camadas de saída. O *perceptron* de múltiplas camadas é conhecido por ser uma rede neuronal *feed-forward*, porque a informação flui em uma direção específica através da rede, da camada de entrada para a camada de saída (Gardner & Dorling, 1998).

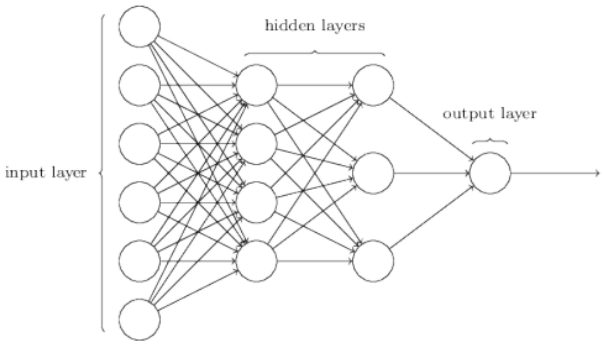


Figura 7- Rede neuronal de múltiplas camadas (Nielsen, 2015)

Métodos de E*nsemble*

Métodos de *ensemble* são técnicas de aprendizagem de *machine learning* que combinam diversos modelos para produzir um único modelo preditivo ótimo (Lutins, 2017). “São algoritmos de

[1] Cenário que ocorre quando um modelo estatístico se ajusta muito bem aos dados de treino, porém demonstra ser ineficaz na previsão de novos resultados.

aprendizagem que constroem um conjunto de modelos e, em seguida, classificam novos pontos de dados através de uma votação ponderada das suas previsões." (Dietterich, 2000, p. 1).

Estas técnicas são muito estudadas na área de aprendizagem supervisionada, uma vez que, em muitos casos, as técnicas de *ensemble* são capazes de produzir resultados mais precisos do que os modelos de previsão individuais que os compõem. Nas técnicas de *ensemble* é importante que os modelos sejam precisos e diversificados, o que significa que devem cometer erros diferentes na previsão dos novos dados (Dietterich, 2000). Idealmente, as saídas dos modelos devem ser independentes ou preferencialmente negativamente correlacionadas (Zhang & Ma, 2012).

Os métodos de *ensemble* são geralmente utilizados em duas configurações: seleção de modelos (*classifier selection)* e fusão de modelos (*classifier fusion).* Na seleção de classificadores, cada modelo é treinado como sendo o "especialista local" em alguma vizinhança específica de todo o espaço de caraterísticas. Dada uma nova instância, o modelo treinado com os dados mais próximos da vizinhança desta instância, é então escolhido para determinar a decisão final, ou recebe o maior peso na contribuição para a decisão. Na fusão de modelos (*classifier fusion)*, todos os modelos são treinados em todo o espaço de características, e depois combinados para obter um classificador composto com menor variância e, consequentemente, menor erro. *Bagging, Random Forest, Boosting/AdaBoost, Stacking* são exemplos desta abordagem.

Bagging*,* também conhecido como *bootstrap aggregation*, é um método utilizado para reduzir a variância num conjunto de dados "ruidosos" (IBM, 2023a) e é mais adequado para problemas com conjuntos de dados de treino relativamente pequenos (Zhang & Ma, 2012). No *bagging* é selecionada uma amostragem aleatória de um conjunto de dados de treino, com reposição, o que permite que certos pontos de dados sejam escolhidos mais do que uma vez. Após a geração de várias amostras de dados, esses modelos fracos são treinados de forma independente e, no final, agregam as suas previsões individuais para formar uma previsão final, partindo da premissa de que a previsão combinada apresenta um desempenho superior e mais preciso (IBM, 2023a). Um exemplo prático é o uso de múltiplas árvores de decisão: Após cada árvore de decisão calcular a sua previsão, todos os resultados são agregados, formando a previsão teoricamente mais precisa. Em vez de se confiar somente na previsão de uma árvore de decisão, os métodos de *ensemble* permitem ter em conta uma amostra de árvores de decisão, calcular e decidir quais os atributos a considerar para cada uma, e criar uma previsão final com base nos resultados agregados (figura 8).

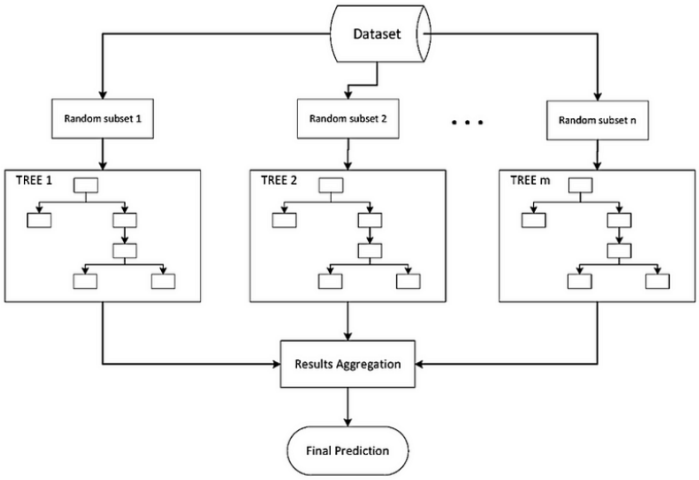


Figura 8- Método de *Bagging* com árvores de decisão (Lutins, 2017)

O algoritmo ***random forest*** pode ser visto como uma abordagem semelhante ao *bagging*, mas com uma pequena diferença crucial. Enquanto no *bagging* tradicional todas as árvores de decisão são construídas com o conjunto completo de características para cada amostra e, embora as amostras possam ser ligeiramente diferentes, os dados tendem a se dividir principalmente nas mesmas características em todos os modelos. Em contraste, os modelos de *random forest* decidem onde fazer as divisões com base em uma seleção aleatória de características (Lutins, 2017).

Em vez de dividir os dados usando características semelhantes em cada nó durante o processo de construção da árvore, os modelos de *random forest* implementam um nível de diferenciação, pois cada árvore realiza divisões com base em características distintas (figura 9). Esse nível de diferenciação proporciona um conjunto diversificado de modelos para serem combinados, resultando num modelo preditivo mais preciso.

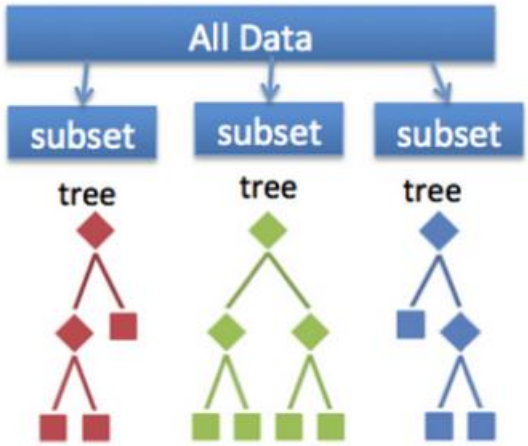


Figura 9- *Bagging* com *Random Forest* (Lutins, 2017)

Boosting é uma técnica de *Machine Learning* que visa criar um modelo forte ao combinar vários aprendizes fracos (weak *learners*), tendo como objetivo principal reduzir os erros de treino (IBM, 2023b). No processo de *boosting* (figura 10), é selecionada uma amostra aleatória de dados, e um modelo é ajustado a essa amostra. Em seguida, o próximo modelo é treinado de forma sequencial, onde cada modelo tenta corrigir as fraquezas do modelo anterior (Zhang & Ma, 2012). Esse ajuste sequencial permite redistribuir os pesos dos dados, o que permite ao algoritmo identificar os parâmetros mais relevantes para melhorar o desempenho do modelo (IBM, 2023b). Ao contrário do *bagging*, em que os modelos fracos são treinados em paralelo, o *boosting* realiza o treino de forma sequencial, aprimorando cada modelo para obter um desempenho final mais robusto e preciso (IBM, 2023b). Além disso, há uma distinção nos cenários de aplicação entre *bagging* e *boosting*. O *bagging* é utilizado em situações onde os *learners* fracos apresentam alta variância e baixo *bias*. Por outro lado, o *boosting* é mais adequado para casos com baixa variância e alto *bias*, pelo que é mais indicado para casos em que haja necessidade de diminuir o *bias* (IBM, 2023b).

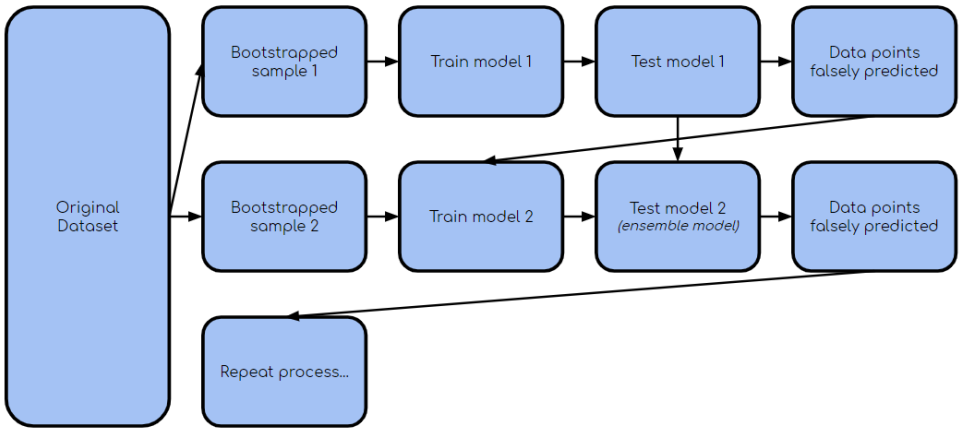


Figura 10- Método de *boosting* (Shin, 2020)

Adaboost *(Adpative Boosting)* segue o procedimento mencionado acima. Inicialmente, atribui-se o mesmo peso para cada conjunto de dados. Durante a fase de treino, o *Adaboost* seleciona de forma iterativa instâncias nos conjuntos seguintes, com base em uma distribuição atualizada das amostras de treino. Em seguida, os modelos são combinados por meio de uma votação ponderada, em que esses pesos são determinados pelos erros de treino dos modelos. Essa abordagem garante que as instâncias classificadas incorretamente pelo modelo anterior recebam maior peso, aumentando sua probabilidade de serem incluídas nos dados de treino do próximo modelo, possibilitando sua correção na próxima iteração (Zhang & Ma, 2012). ***Decision stumps*** são comumente selecionadas como *weak learners* no algoritmo de *Adaboost* quando há um conhecimento limitado sobre o domínio do problema de aprendizagem. Trata-se de árvores de decisão simples com apenas uma divisão, ou seja, apresentam somente duas folhas. Apesar da sua simplicidade, as *decision stumps* apresentam um bom desempenho quando utilizados em algoritmos de *boosting.* São considerados *weak learners* porque têm um desempenho ligeiramente superior a um palpite aleatório e a sua simplicidade torna-os menos propensos a *overfitting*. As *decision stumps* são uma escolha popular devido à sua simplicidade, eficácia e capacidade de servir como um ponto de partida para modelos mais complexos (Kégl, 2014).

A tabela 2 apresenta, de forma resumida, as diferenças entre os métodos de *bagging* e *boosting.*

Tabela 2- Comparação entre os métodos de *bagging* e *boosting* (Adaptado de Aljamaan & Elish, 2009; IBM, 2023b)

	Bagging	***Boosting***
Similaridades	- Utiliza votação - Combina modelos do mesmo tipo - Computacionalmente dispendioso	
Diferenças	Os modelos individuais são construídos separadamente	Cada novo modelo é influenciado pelo desempenho do modelo anterior
	É atribuido um peso igual a todos os modelos	Pondera a contribuição de um modelo pelo seu desempenho

	Bagging	***Boosting***
	Lida com problemas de sobreajuste (*overfitting)*	Não lida com problemas de sobreajuste
	Reduz a variância	Reduz o *Bias*
	-	Eficaz computacionalmente[2]

Ao contrário dos métodos de *bagging* e *boosting*, que utilizam combinadores não treináveis (*nontrainable combiners*), onde os pesos das combinações são definidos durante a fase de treino dos modelos, impossibilitando a identificação do que cada modelo aprendeu sobre uma determinada divisão do espaço de características, o método de ***stacking*** surge como uma alternativa. O ***stacking*** utiliza combinadores treináveis (*treinable combiners)*, o que permite determinar quais modelos são mais eficazes em diferentes espaços de características, e combiná-los com base nessa probabilidade (Zhang & Ma, 2012).

O método de ***stacking*** é uma técnica que combina diferentes modelos em quatro etapas (figura 11):

- Criação e treino dos modelos de nível 1: São selecionados T modelos de nível 1, em que todos eles são treinados usando validação cruzada (*k-fold cross-validation)*;

- Criação do conjunto do meta-modelo de nível 2: As saídas dos modelos de nível 1 são combinadas e utilizadas como o novo conjunto de treino para o meta-modelo;

- Treino do meta-modelo: O meta-modelo, normalmente um modelo mais robusto, é treinado utilizando as previsões dos modelos de nível 1. O objetivo é aprender como combinar as previsões dos modelos de nível 1 de forma a otimizar o desempenho geral do conjunto;

- Previsão de novos pontos de dados: Após o treino do meta-modelo, o modelo de *stacking* está preparado para efetuar novas previsões em novos pontos de dados desconhecidos. Quando surge um novo ponto de dado, os modelos de nível 1 efetuam as suas previsões e essas previsões são passadas para o meta-modelo de nível 2, que as utiliza como entrada e determina a previsão final do conjunto.

[2] Dado que os algoritmos de *boosting* têm a capacidade de selecionar somente os parâmetros que aumentam a capacidade de efetuar as previsões na fase de treino, podem ajudar a diminuir a dimensionalidade e a aumentar a eficiência computacional (IBM, 2023b).

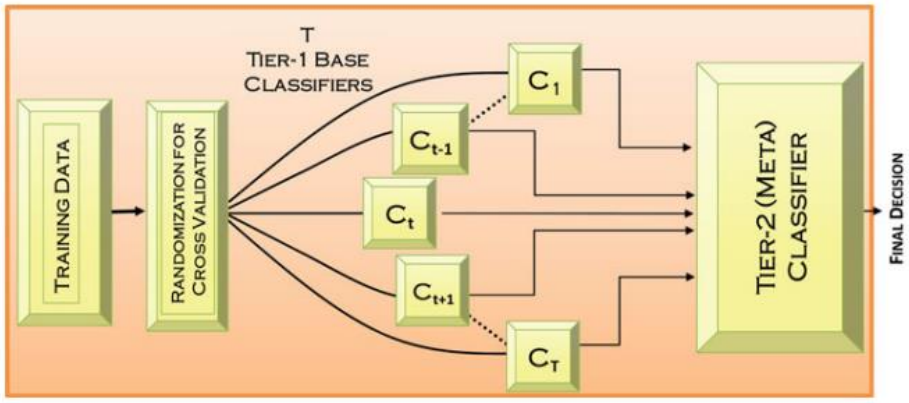


Figura 11- Método de *stacking* (Zhang & Ma, 2012)

O método de *stacking* tem a vantagem de proporcionar maior precisão de previsão em comparação com *bagging* e *boosting*. No entanto, devido à sua capacidade de combinar modelos heterogêneos, incluindo modelos de *bagging* e *boosting*, ele requer mais tempo computacional e recursos (Kalirane, 2023).

Na tabela 3 estão representadas as principais caraterísticas dos métodos de *bagging, boosting* e *stacking.*

Tabela 3- Principais caraterísticas dos métodos de *bagging*, *boosting* e *stacking* (adaptado de (Kalirane, 2023))

	Bagging	***Boosting***	***Stacking***
Objetivo	Reduzir variância	Reduzir *bias*	Melhorar a precisão
Tipos de modelos	Modelos homogêneos	Modelos homogêneos	Modelos heterogêneos[3]
Método de agregação	Paralelo	Sequencial	Hierárquico
Ponderação	É atribuído um peso igual a todos os modelos	Pondera a contribuição de um modelo pelo seu desempenho	Utiliza um meta-modelo para aprender a combinação mais adequada a partir previsões dos modelos fracos

2.3.2. Aprendizagem não supervisionada

Aprendizagem não supervisionada é um tipo de aprendizagem de *Machine Learning* que utiliza algoritmos para analisar e agrupar dados com caraterísticas desconhecidas sem intervenção humana (IBM Corp., 2023). Para além disso, tem como objetivo construir representações dos dados de entrada que podem ser utilizados no auxílio da tomada de decisão (Ghahramani, 2004).

[3] Modelos heterogêneos são modelos de *machine learning* diferentes e diversificados, provenientes de diferentes famílias de algoritmos, enquanto modelos homogêneos são modelos do mesmo tipo ou família e que compartilham a mesma estrutura básica.

De acordo com Géron (2017) os métodos de aprendizagem não supervisionada mais importantes são: agrupamento, visualização e redução da dimensionalidade e associação.

Agrupamento

Agrupamento é um método utilizado para organizar dados não rotulados em grupos, com base em similaridades ou diferenças. Este método utiliza algoritmos para processar dados não classificados em grupos, representados por padrões (IBM Corp., 2023). Por exemplo, este método poderia ser utilizado para agrupar os visitantes de um *blog* com base nos dados dos mesmos. O algoritmo utilizado será capaz de encontrar conexões entre os diferentes visitantes sem que o utilizador refira a qual grupo o visitante irá pertencer (Géron, 2017).

Na figura 12, encontra-se ilustrado um exemplo de agrupamento, com os grupos formados.

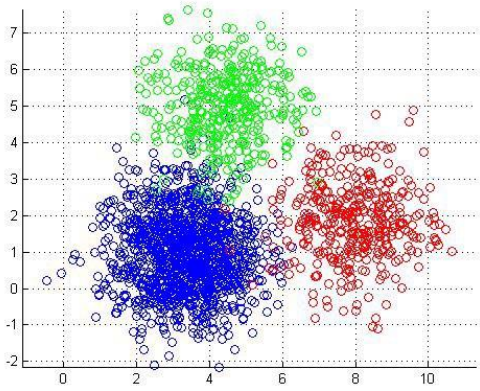


Figura 12- Agrupamento (N. S. Chauhan, 2019)

Visualização e redução da dimensionalidade

Visualização e redução da dimensionalidade é um método utilizado para diminuir o número de caraterísticas de um conjunto de dados, tornando-o mais fácil de trabalhar sem comprometer a integridade dos mesmos (IBM Corp., 2023). De acordo com Géron (2017) a redução da dimensionalidade pode ser alcançada combinando múltiplas caraterísticas correlacionadas em uma única caraterística. Por exemplo, a quilometragem e a idade de um automóvel estão intimamente relacionadas, e, portanto, estas duas caraterísticas podem ser combinadas numa só em que represente o desgaste geral do veículo.

Associação

Regras de associação é uma técnica utilizada para identificar relações entre variáveis num conjunto de dados. É utilizada para análise de cesto de compras, auxiliando as empresas a conhecer os padrões de consumo dos seus clientes e, desse modo, a melhorarem as suas estratégias de venda bem como a desenvolver bons sistemas de recomendação (IBM Corp., 2023).

A análise de cestos de compras (figura 13) visa identificar padrões nos hábitos de compras dos clientes (Cios et al., 2007). Um exemplo fornecido pelo autor Géron (2017) ilustra bem este caso: “Suponha que você é o dono de um supermercado. Ao executar uma regra de associação no seu registo de vendas poderá ser revelado que as pessoas que compram molho para churrasco e batatas fritas também tendem a comprar carne. Assim, estes artigos poderão ser posicionados de forma estratégica, próximos entre si.” (Géron, 2017, p.34)

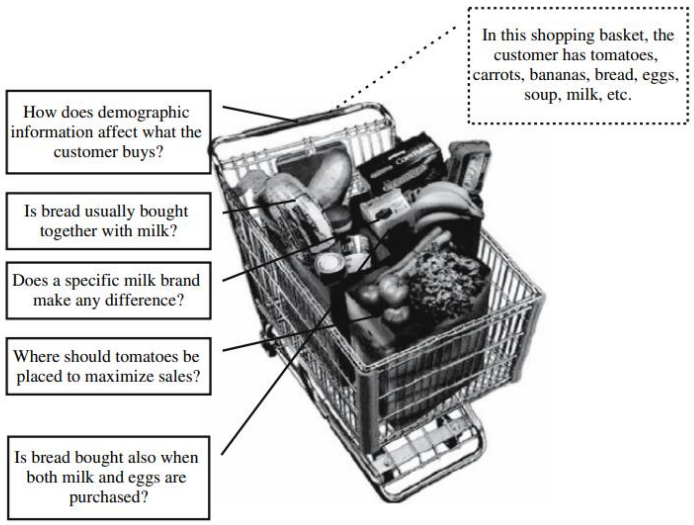


Figura 13- Ilustração de um cesto de compras (Cios et al., 2007)

Existem vários algoritmos diferentes, utilizados para gerar regras de associação, tais como: Apriori, FP-Growth, AIS, entre outros. Todavia, o algoritmo Apriori é o mais amplamente utilizado (IBM Corp., 2023).

Algoritmo Apriori

O algoritmo Apriori (figura 14) inicia com o conjunto de transações de artigos de um cliente (T), o conjunto de artigos frequentes (L) e o conjunto de artigos candidatos (C). O nível mínimo de suporte é estabelecido antecipadamente. O algoritmo é dividido em duas fases: a primeira consiste em criar o conjunto de artigos candidatos, no qual o conjunto de artigos frequentes é gerado combinando $\mathbf{L}_{k-1}$ * $\mathbf{L}_{k-1}$. A segunda fase consiste na criação do conjunto de artigos frequentes, onde os artigos acima do suporte mínimo são retirados do conjunto e um conjunto de artigos frequentes é formado. O algoritmo termina quando o conjunto de artigos candidatos $\mathbf{C}_k$ se torna vazio. O conjunto de artigos frequentes é utilizado para identificar artigos relacionados com o comportamento do consumidor, de forma a efetuar recomendações (Jooa et al., 2016).

$$
\begin{aligned}
&\text{Apriori}(T, \epsilon) \\
&\quad L_1 \leftarrow \{\text{large } 1-\text{itemsets}\} \\
&\quad k \leftarrow 2 \\
&\quad\quad \textbf{while } L_{k-1} \neq \emptyset \\
&\quad\quad\quad C_k \leftarrow \{c | c = a \cup \{b\} \wedge a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\} \\
&\quad\quad\quad \textbf{for transactions } t \in T \\
&\quad\quad\quad\quad C_t \leftarrow \{c | c \in C_k \wedge c \subseteq t\} \\
&\quad\quad\quad\quad \textbf{for candidates } c \in C_t \\
&\quad\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
&\quad\quad\quad L_k \leftarrow \{c | c \in C_k \wedge\ count[c] \geq \epsilon\} \\
&\quad\quad\quad k \leftarrow k + 1 \\
&\quad\quad \textbf{return } \bigcup_k L_k
\end{aligned}
$$

Figura 14- Algoritmo Apriori (Jooa et al., 2016)

2.3.3. Avaliação de desempenho dos algoritmos

A avaliação de desempenho dos algoritmos é fundamental para avaliar a sua eficácia e confiabilidade. Se um modelo não é avaliado de forma correta, poderá resultar em previsões fracas no futuro (N. Chauhan, 2020). A medição do desempenho permite comparar diferentes modelos e abordagens, identificando quais são os mais eficazes para dar resposta a um determinado problema. Para além disso, estes tipos de análises são fundamentais para detetar a presença de fenómenos de *overfitting*, em que o modelo tem um desempenho excelente, porém quando os dados de teste são aplicados o resultado não é igualmente satisfatório, ou *underfitting*, em que o modelo não tem a capacidade de encontrar relações entre as variáveis e detetar a complexidade dos dados (Géron, 2017).

Uma das métricas mais simples de avaliação de desempenho dos algoritmos é a matriz de confusão. A matriz de confusão é uma representação matricial dos resultados de previsão, onde as linhas representam os valores reais e as colunas os valores previstos (figura 15).

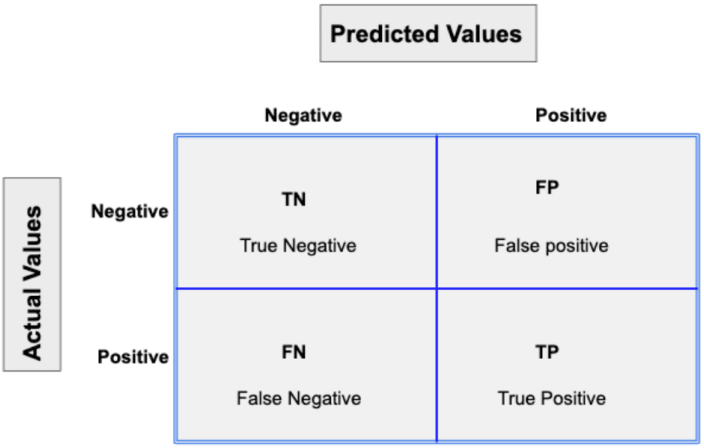


Figura 15- Matriz de confusão

Cada previsão pode ter como resultado um dos quatro quadrantes acima representados:

- Verdadeiro positivo (TP): Previsto verdadeiro e ser verdadeiro na realidade;
- Verdadeiro negativo (TN): Previsto falso e ser negativo na realidade;
- Falso positivo (FP): Previsto positivo e ser negativo na realidade;
- Falso negativo (FN): Previsto negativo e ser positivo na realidade.

A métrica *accuracy* mede "com que frequência a classificação está correta". A precisão é uma métrica de avaliação para problemas de classificação e é representada pela seguinte fórmula:

$$Accuracy = \frac{(TP + TN)}{total} \qquad (2.2)$$

A métrica *precision* mede com que frequência a classificação está correta, quando a previsão é positiva e é representada pela seguinte fórmula:

$$Precision = \frac{TP}{FP + TP} \qquad (2.3)$$

A métrica *recall ou sensivity* mede com que frequência se prevê resultados positivos, quando o resultado é efetivamente positivo, sendo representada por:

$$Recall = \frac{TP}{FN + TP} \qquad (2.4)$$

A métrica *specificity* resulta na medição da frequência com que se prevê o não, quando é não, sendo representada por:

$$Specificity = \frac{TN}{TN + FP} \qquad (2.5)$$

Por último, a métrica F1-*score* também designada por *F-Measure* relaciona as métricas de *precision* e *recall* e é representada por:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \qquad (2.6)$$

2.3.3.1. Validação cruzada

A validação cruzada é uma técnica utilizada para avaliar a capacidade de generalização de modelos de previsão, de forma a evitar o *overfitting* sendo o método mais simples e mais amplamente utilizado para estimar o erro de previsão (Hastie et al., 2009a).

Como os dados muitas vezes são escassos, utiliza-se o método *K-fold Cross-validation*, em que parte dos dados são utilizados para treinar o modelo e outra parte para o testar.

No caso representado na figura 16, os dados foram divididos em 5 partes iguais (K=5), sendo a terceira parte utilizada para validação (k^{th} *part*). Para as restantes k-1 partes o modelo é treinado e é calculado o erro de previsão do modelo ajustado quando se prevê a k^{th} *part.* Este procedimento é executado para k = 1,2…,K. Em cada iteração, os subconjuntos de treino e de teste são alternados, e são combinadas as K estimativas do erro de previsão (Hastie et al., 2009a).

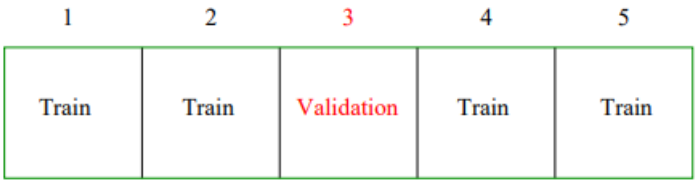


Figura 16- Validação cruzada (K=5)

2.4. Regras de associação

Regras de associação são modelos descritivos, utilizados de forma a identificar padrões e relações dentro de um conjunto de dados. Uma regra de associação expressa na forma X-> Y é interpretado como tendo a capacidade de satisfazer X e, portanto, é provável que satisfaçam Y.

Segundo (Berry et al., 2004) as relações são apresentadas sobre a forma "if-then", como "se um cliente compra o produto A, então é provável que compre B".

Essas regras podem ser úteis para identificar padrões frequentes em dados (Zhang & Zhang, 2002), para tomar decisões sobre como vender ou promover produtos (Berry et al., 2004) e têm aplicações diretas em diferentes áreas.

Na indústria do retalho, os modelos de *Machine Learning* podem ser utilizados para efetuar uma análise dos carrinhos de compra dos clientes, que envolve encontrar associações entre produtos que são frequentemente comprados pelos clientes. Por exemplo, se um cliente compra o produto X, este também está propenso a comprar o produto Y. Ao ser possível determinar estas associações, os retalhistas poderão direcionar os clientes para produtos específicos promovendo o *cross-selling*.

Esta informação pode ser valiosa para uma empresa como a *Walmart* em termos de tomar decisões sobre como promover e vender os seus produtos. Por exemplo, eles podem escolher exibir barras de chocolate de maneira prominente perto das bonecas Barbie ou oferecer promoções que incluam os dois produtos (Berry et al., 2004).

Para encontrar estas associações os retalhistas podem estimar a probabilidade condicional de P(Y|X), onde Y é o produto que está a ser condicionado por X, que é o produto ou o conjunto de produtos que o cliente já comprou (Alpaydin, 2014).

Existem três tipos de regras comuns geradas pelas regras de associação: o acionável (*the actionable*), o trivial (*the trivial)* e o inexplicável (*the inexplicable*).

O acionável é um padrão que foi identificado, possivelmente através da análise de dados. Exemplo disso é o caso do *Walmart* que foi apresentado um parágrafo acima.

Resultados triviais são aqueles que já são conhecidos ou esperados por qualquer pessoa familiarizada com o negócio. Por exemplo, pode-se esperar que os clientes que adquirem contratos de manutenção de eletrodomésticos também adquiram os próprios eletrodomésticos ou vice-versa, já que eles geralmente são anunciados e vendidos juntos. Embora esses resultados sejam válidos e bem apoiados por dados, eles não são particularmente úteis ou práticos (Berry et al., 2004).

O inexplicável é algo que resulta de por exemplo: um padrão foi observado que nos dados de uma empresa de hardware os piaçabas eram um artigo comumente vendido quando uma nova loja abria. No entanto, este padrão não oferece nenhuma compreensão sobre o comportamento do consumidor ou sugere alguma ação para a empresa (Berry et al., 2004).

A importância das regras de associação é quantificada através do cálculo das métricas *lift, support* e *confidence.*

Lift é uma métrica que mede a correlação entre dois conjuntos de artigos, A e B. Trata-se da razão entre a probabilidade de A e B ocorrerem juntos e a probabilidade de ocorrerem de forma independente. Se o valor do *lift* for maior que um, os conjuntos de artigos estão positivamente

correlacionados, enquanto se o valor for menor que um, indica que estão negativamente correlacionados (Han et al., 2011).

O *lift* entre A e B pode ser representado por:

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)} \qquad (2.7)$$

Support representa a "percentagem de transações existente numa base de dados de transações que determinada regra satisfaz" (Han et al., 2011, p.21). É descrito como sendo a probabilidade $\mathrm{P}(\mathrm{X} \cup \mathrm{Y})$, onde $\mathrm{X} \cup \mathrm{Y}$ indica que determinada transação contém tanto X como Y, ou seja, a união do conjunto de artigos X e Y (Han et al., 2011).

Por último, a métrica *confidence* avalia o grau de certeza da associação detetada. É representada pela probabilidade condicional $\mathrm{P}(\mathrm{X}|\mathrm{Y})$, que é a probabilidade de uma transação que contém X também conter Y.

As métricas *support* e *confidence* podem ser definidas por:

$$support(X => Y) = P(X \cup Y) \qquad (2.8)$$

$$confidence(X => Y) = P(X|Y) \qquad (2.9)$$

Com a identificação destas associações entre diferentes artigos, os negócios poderão compreender melhor os seus clientes, as suas preferências e efetuar recomendações mais personalizadas com base nas caraterísticas de compras específicas de determinado cliente.

2.5. Sistemas de recomendação

Sistemas de recomendação são modelos utilizados para efetuar recomendações de artigos aos consumidores com base em ações ou preferências passadas. Estes sistemas recolhem dados sobre os utilizadores, como os artigos que eles visualizaram ou compraram e utilizam essa informação para conhecer melhor os interesses dos mesmos. O sistema é capaz de comparar esses dados com o comportamento similar de outros utilizadores e, através dessa informação recomendar produtos que terão uma maior probabilidade de serem do interesse do cliente. Os dados utilizados pelo sistema de recomendação podem surgir de diversas fontes, nomeadamente através de websites, informação demográfica, ou até mesmo feedback explícito fornecido pelos próprios utilizadores sobre determinado produto (Kane, 2018).

O comportamento explícito refere-se a ações que podem ser interpretadas como sendo do interesse do utilizador. Por exemplo, se um utilizador clica num link ou numa publicidade, pode-se assumir que esse utilizador tem interesse em determinado artigo. Todavia, o utilizador poderá ter selecionado esse link/publicidade por acidente e, por isso, é importante considerar diversas fontes de dados, quando se arquiteta um modelo de recomendação, de forma a fornecer uma maior precisão e confiança nas recomendações efetuadas. Fontes estas que poderão ser os artigos que

realmente foram comprados pelos utilizadores, isto porque certamente houve um interesse por parte do consumidor naquele artigo em específico.

Nos sistemas de recomendação a utilidade de um artigo é normalmente representada por uma classificação que indica se um determinado utilizador gostou de um certo produto, porém pode ser calculada pelo aplicativo através de uma função baseada em lucro. Um desafio presente nos sistemas de recomendação é que a função de utilidade só está definida para um subconjunto do conjunto geral que constitui os utilizadores e artigos, exigindo que a função seja extrapolada para o restante conjunto. Essa extrapolação pode ser realizada recorrendo a heurísticas ou estimando a função de utilidade que otimiza um certo critério de desempenho. Após se realizar a extrapolação de forma a estimar as classificações desconhecidas, as recomendações podem ser realizadas selecionando a classificação mais alta estimada para esse utilizador. Em alternativa, pode-se recomendar um conjunto de utilizadores a um artigo ou os N melhores artigos a um utilizador (Adomavicius & Tuzhilin, 2005).

Quando se produz um sistema de recomendação é importante conhecer as fontes de dados utilizadas e determinar quais dados são mais relevantes e confiáveis para o sistema. É também importante ter uma quantidade de dados suficiente de forma a produzir sistemas de recomendação precisos (Kane, 2018).

Os sistemas de recomendação são normalmente classificados nas seguintes categorias: sistemas de filtragem colaborativa, sistemas de filtragem baseada em conteúdo e sistemas de recomendação híbridos.

Sistemas de filtragem colaborativa são sistemas de recomendação de produtos baseado nas preferências de outros utilizadores com interesses similares. Por exemplo, se um utilizador A gosta de laranjas, kiwis e abacaxi, enquanto o utilizador B gosta de laranjas, kiwis e uvas, é provável que A goste de uvas e B de abacaxi, dado que apresentam interesses semelhantes (Techlabs, 2021).

Estes sistemas usam técnicas de factorização matricial para identificar padrões e efetuar recomendações (Isinkaye et al., 2015).

Segundo Jooa et al. (2016), os sistemas de filtragem colaborativa possuem dois métodos principais: método baseado no utilizador e método baseado no artigo. O método baseado no utilizador gera recomendações através das semelhanças existentes entre os clientes e o método baseado no artigo gera recomendações com base nas semelhanças entre produtos. O método baseado no utilizador é o método mais fácil e rápido de implementar, uma vez que o método baseado no artigo pode não ser preciso pelo facto de os utilizadores possuírem preferências variadas.

De acordo com Adomavicius & Tuzhilin (2005), a utilidade $u(c,s)$ de um artigo s para o utilizador c é estimada com base nas utilidades $u(c_j,s)$ atribuídas ao artigo s por utilizadores semelhantes, $cj \in C$, em que C representa o conjunto de utilizadores e cj representa um utilizador semelhante ao utilizador c.

Sistemas de filtragem baseada em conteúdo, efetuam recomendações de artigos com base nas caraterísticas dos próprios artigos. Por exemplo, sistemas de *streaming* sugerem filmes aos utilizadores com base nos géneros e nos atores de filmes previamente visualizados pelo mesmo (Isinkaye et al., 2015).

Em sistemas de filtragem baseada em conteúdo, a utilidade *u(c,s)* de um artigo *s* para o utilizador *c* é estimada com base nas utilidades $u(c,s_i)$ atribuídas pelo utilizador *c* ao artigo $s_i \in S$, que são semelhantes ao artigo *s* (Adomavicius & Tuzhilin, 2005).

Sistemas de recomendação híbridos são sistemas de recomendação de produtos que utilizam uma abordagem híbrida, combinando métodos de filtragem colaborativa com filtragem baseada em conteúdo. A combinação destes dois métodos possibilita a eliminação de certas limitações existentes quando se utiliza uma destas abordagens de forma isolada (Adomavicius & Tuzhilin, 2005).

De acordo com Adomavicius & Tuzhilin (2005), existem diferentes maneiras possíveis de combinar estes dois métodos: 1. implementar os dois métodos de forma isolada e combinar as suas previsões; 2. incorporar algumas caraterísticas de filtragem baseada em conteúdo em filtragem colaborativa ou vice-versa; 3. construir um modelo abrangente que combina elementos de filtragem baseada com elementos de filtragem colaborativa.

Na primeira abordagem acima indicada (1) os sistemas de recomendação híbridos podem ser construídos através da combinação das saídas dos sistemas colaborativos com as saídas dos sistemas baseados em conteúdo utilizando uma combinação linear ou através de um esquema de votação. Uma outra abordagem seria através do uso dos sistemas de forma individual em que o sistema selecionado seria o que apresentasse um melhor desempenho.

Na segunda abordagem (2), os sistemas de recomendação híbridos combinam as técnicas colaborativas com algumas das caraterísticas de filtragem baseada em conteúdo. Esta abordagem permite recomendar artigos aos utilizadores não só com base nas avaliações geradas pelos utilizadores com perfis similares como também com base no conteúdo.

Na terceira abordagem (3), os autores Adomavicius & Tuzhilin (2005) referem que os sistemas de recomendação híbridos podem ser melhorados ao incorporar técnicas baseadas em conhecimento para aumentar a precisão e superar as limitações dos sistemas tradicionais. Todavia, a principal desvantagem desses sistemas é a necessidade de aquisição de conhecimento, que pode ser um *bottleneck* para aplicações de inteligência artificial.

Os sistemas de recomendação são uma ferramenta essencial para as empresas. Estes sistemas não só melhoram a experiência do consumidor como também impulsionam as receitas e ajudam as empresas a manterem-se competitivas (Techlabs, 2021). Por esse motivo, o desenvolvimento de um bom sistema de recomendação é um investimento estratégico que pode trazer benefícios significativos para as empresas.

2.6. Análise RFM

A análise RFM é um método utilizado pelas empresas para segmentar clientes através da análise dos seus hábitos de consumo (Blattberg et al., 2008).

Segundo os autores Bult & Wansbeek (1995) a técnica de seleção mais utilizada é a análise de Recência, Frequência e Valor monetário (análise RFM).

A recência (R) representa a última vez que o cliente realizou uma compra. Frequência (F) é uma medida que indica a frequência com que o cliente comprou num determinado período (Bult & Wansbeek, 1995). É frequentemente medido pelo número de ocasiões de compras efetuadas desde

a primeira compra realizada (Blattberg et al., 2008). O valor monetário (M) é uma medida de quanto o cliente gastou em compras (Bult & Wansbeek, 1995).

De acordo com Wu & Lin (2005), os investigadores referem que quanto maior forem os valores de recência e frequência, mais provável será que os clientes façam novas compras com a empresa. Para além disso, quanto maior for o valor monetário, mais provável será que os clientes comprem novamente os produtos e serviços da empresa.

O método RFM começa por ordenar os clientes pela sua data de compra, com os clientes mais recentes no topo, permitindo que os clientes sejam agrupados em várias categorias. Em seguida, a frequência e o valor monetário são padronizados e ordenados da mesma maneira. Assim, ao multiplicar R*F*M, o valor RFM para cada cliente pode ser determinado. Adicionalmente, os clientes podem ser agrupados com base numa proporção (Wu & Lin, 2005). De acordo com Birant (2011), para cada variável (R, F e M) os clientes podem ser divididos em cinco grupos iguais. Os melhores 20% recebem uma pontuação de 5 e os restantes recebem uma pontuação de 4, 3, 2 e 1. Assim, todos os clientes são classificados através da combinação das pontuações das três variáveis.

Os autores Bult & Wansbeek (1995) referem que as duas principais desvantagens deste modelo são o número limitado de variáveis em uso, o que limita o número de caraterísticas envolvidas nesta análise e "a menos que as variáveis RFM sejam todas mutuamente independentes (o que não é expectável), o modelo não tem em consideração a redundância (duplas contagens)" (Bult & Wansbeek, 1995, p.380).

2.7. Trabalho relacionado

Gurudath (2020), com base na análise do comportamento de compra dos clientes de um retalhista da *Instacart*, propôs uma solução para prever o comportamento futuro dos consumidores com a implementação de um sistema de recomendação baseado em regras de associação. Para a realização deste projeto foram importados cinco conjuntos de dados que contêm informação pessoal de clientes não identificados, com mais de três milhões de registos de compras em supermercados.

O sistema proposto, recorre ao uso dos algoritmos *Apriori* (Agrawal & Srikant, 1994) e *Frequent Pattern Growth* (Hirate et al., 2004) com o objetivo de identificar padrões nos produtos que são frequentemente comprados juntos. *FP- Growth* é uma versão melhorada do algoritmo *Apriori* utilizado para analisar padrões ou encontrar associações de conjuntos de dados em casos frequentes. A principal vantagem do algoritmo *FP-Growth* é que ele é mais rápido que o algoritmo *Apriori*, especialmente para conjuntos de dados de grandes dimensões.

Para além disso, Gurudath (2020) fez uso do *framework* Flask (*Flask*, 2023) para criar uma aplicação web do sistema de recomendação de compras. A aplicação web contém três páginas. A primeira página contém a informação relativa ao número de produtos e o número de regras geradas pelas regras de associação. A segunda página (página de recomendação) permite ao utilizador selecionar produtos para o carrinho de compras e exibe esses mesmos produtos juntamente com uma lista de produtos recomendados que dizem respeito aos produtos que o utilizador acrescentou previamente. A terceira página, página de análise exploratória dos dados, são exibidos gráficos dos produtos populares e os resultados do algoritmo utilizado para minerar as regras de associação.

De forma a avaliar a capacidade de previsão das regras de associação foram utilizadas as métricas *Support, Confidence* e *Lift*. Para além disso, foi utilizado o NetworkX (*NetworkX*, 2023) que se trata de um programa para desenvolvimento e análise de redes complexas capaz de testar a relação entre antecedentes e consequentes obtidas após a regra de relação. Através da análise da figura 17 é possível verificar que, por exemplo: "banana é um artigo popular que combina com morangos, maçãs, pepino, maça, abacate e limão. Portanto, se uma pessoa compra um destes seis artigos a probabilidade de comprar banana é alta (o contrário também se aplica)" (Gurudath, 2020, p.59).

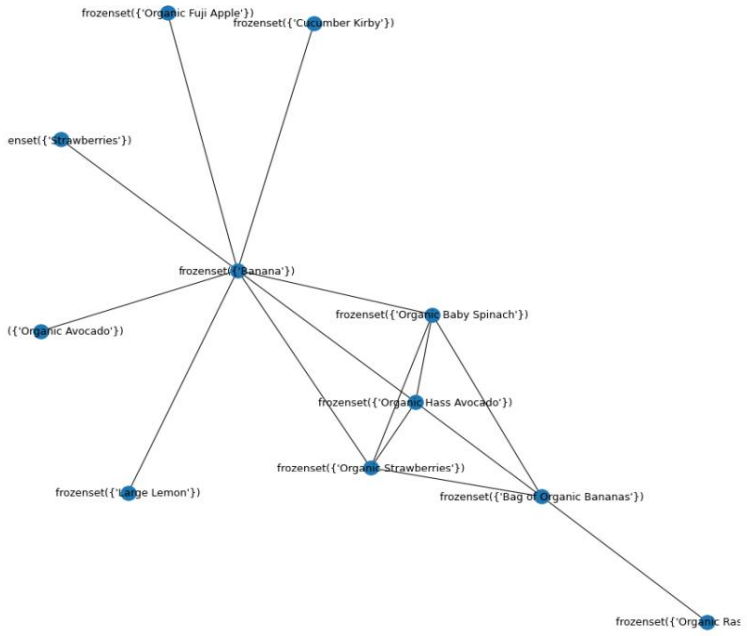


Figura 17- Gráfico de rede das regras de associação (Gurudath, 2020)

Kutuzova & Melnik (2018) tiveram como objetivo melhorar o desempenho de um sistema de recomendação para supermercados. O primeiro conjunto de dados utilizado consiste numa coleção de pedidos de clientes registados ao longo do tempo, com o objetivo de prever quais os produtos estarão presentes no próximo pedido do cliente. O conjunto de dados contém mais de 3 milhões de pedidos de mais de 200000 utilizadores do *InstaCart*. Para além disso, inclui a sequência de produtos comprados em cada pedido, a semana e hora do dia em que o pedido foi realizado e uma medida relativa de tempo entre pedidos (*Kaggle*, 2023)

O primeiro conjunto de dados (k_{prior}) foi utilizado para construir um modelo prévio (P) do comportamento dos clientes que seria uma representação de um sistema ideal de recomendação. O segundo conjunto de dados (k_{train}) foi utilizado para construir um sistema de recomendação original (R). Um terceiro conjunto de dados (G_{orig}) surge como sendo uma fonte de dados externa que deve ser adaptável para construir um sistema de recomendação. O conjunto k_{train} foi agrupado para definir que ordens do G_{orig} (fonte de dados externa) são mais adequadas para serem usadas na melhoria do sistema de recomendação.

Um sistema de integração foi proposto para incorporar fontes de dados externas ao sistema de recomendação já existente, usando técnicas como regras de associação, filtragem colaborativa e análise de clusters.

Foram construídos dois sistemas de recomendação: O primeiro baseado no conjunto de dados G_{orig} e o segundo baseado num conjunto de dados adaptado G_{adapat}. Através do uso de duas métricas, *confidence distance CD* e *recommendation conformity RC*, foi possível medir a eficácia do sistema de recomendação e comparar com as versões original e modificada. A métrica *confidence distance CD* é uma medida da diferença nos valores de confiança para a mesma regra de associação XX=>YY obtida a partir da avaliação do sistema de recomendação R e do modelo anterior P. A métrica *recommendation conformity RC* calcula o grau de consistência entre os conjuntos de produtos recomendados aos clientes com base nas suas transações, comparando o sistema de recomendação R e o modelo anterior P.

Os resultados indicam que é possível melhorar o sistema de recomendação que usa fontes de dados externa. No entanto, os conjuntos de dados que foram utilizados eram significativamente diferentes, pelo que o autor deixa evidente a necessidade de realizar novamente este estudo com dados mais adequados e com novas alterações realizadas nos modelos utilizados.

Alfiqra & Khasanah (2020) realizaram um estudo com o objetivo de descobrir quais os tipos de produtos são comprados pelos consumidores numa única transação. Foi utilizado um conjunto de dados que representa um mês de vendas de um supermercado localizado em Yogyakarta composto por 57784 transações envolvendo 41248 artigos. Os dados foram divididos em quatro períodos e o algoritmo *Apriori* foi aplicado a cada período para implementar regras de associação. Para além disso, foi utilizado o software R para correr o algoritmo. *Support, Confidence* e *Lift* foram utilizados para medir a força das regras de associação e determinar a sua utilidade em tomar decisões ou a efetuar previsões. A pesquisa seguiu o processo KDD (*Knowledge Discovery in Database*), que inclui etapas para selecionar, preparar, transformar e interpretar os dados.

Os resultados demonstraram que as regras obtidas para cada período foram diferentes, o que fez com que surgisse a necessidade de uma análise posterior. Para isso, uma análise OCVR (*Overall Variability of Association Rule*) foi efetuada. Trata-se de um indicador para analisar o grau de variação nas regras de associação de um período para o outro, com o intuito de determinar quais as regras são mais importantes na MBA (*Market Basket Analysis*).

Com base na análise OCVR duas estratégias de marketing foram implementadas para promover as vendas do supermercado. Foram definidos grupos de produtos para venda a preços especiais e a disposição das prateleiras foi alterada, de forma a ajudar os clientes a fazer associações entre produtos e, com isso, aumentar potencialmente as vendas.

Uma aplicação para uso no mundo retalhista foi desenvolvida por Azri (2020) que utiliza "o módulo *TensorRec* em *Python*, um algoritmo de *Machine Learning* que utiliza *Tensorflow* e *Keras* como *backend* com uma interface bastante simplificada".

O autor afirma que um modelo de filtragem colaborativa poderá ser suficiente para um pequeno supermercado, porém se o caso for uma cadeia de supermercados com milhões de transações diárias é necessário um algoritmo mais complexo capaz de produzir resultados relevantes. Por isso, Azri (2020) sugere a combinação das livrarias de *Python*, *Pandas*, *Numpy* e *SKLearn* com o módulo *TensorRec* para produzir um modelo robusto e sólido capaz de gerar bons resultados.

O objetivo deste trabalho é "perceber como aplicar um sistema de recomendação a retalhistas que contêm muitos clientes ou sistemas que consistem em múltiplos utilizadores" (Azri, 2020).

Para a realização deste projeto, o autor utilizou um conjunto de dados que contém registos de janeiro de 2011 a 2014. O conjunto de dados é composto por três tabelas referente a *customers* (informação sobre os clientes), *transaction* (transações realizadas pelos clientes) e *product hierarchy* (informação dos produtos). Dado que o conjunto de dados fornece informação relevante acerca da interação entre clientes e produtos bem como algumas caraterísticas dos próprios clientes, o autor aplicou um sistema de recomendação híbrido que contém tanto um sistema de filtragem colaborativa como um sistema de filtragem baseada em conteúdo.

Azri (2020), de forma a retirar informação adicional acerca dos clientes utilizou o método RFMV e um algoritmo de aprendizagem não supervisionada, *k-means clustering,* com o objetivo de produzir padrões, grupos ou conhecimento que não são explicitamente visíveis.

O autor refere que poderá não ser correto aceitar modelos somente com uma precisão igual a 100%, dado que existe uma perda na precisão para os artigos que nunca foram adquiridos anteriormente, mas se pretendem recomendar. Por esse motivo, o autor aceita modelos com uma precisão por volta dos 90%.

Por fim, o autor acrescenta que para melhorar os sistemas de recomendação é importante considerar o negócio em questão, o pedido e manipular os dados de entrada. Por vezes, para melhor a precisão dos modelos, estes podem necessitar de uma mudança nas caraterísticas de entrada em vez de um ajuste nos parâmetros do algoritmo.

Ozkan & Kocakoc (2021) propuseram um modelo RFM "modificado" para ser utilizado na segmentação de clientes para retalhistas. A diversidade de produtos comprados pelos clientes é acrescentada ao modelo como o parâmetro V. Desta forma, as empresas têm a possibilidade de categorizar tanto os seus clientes como os produtos.

Neste estudo, os dados utilizados continham informação sobre 60284 clientes e 6406485 transações ocorridas durante um período de 6 meses de todas as lojas da região do Egeu.

Com base na análise RFM-V efetuada os autores também propuseram uma matriz designada por *CustomerProduct Depth Matrix*. Esta matriz permite que os clientes sejam divididos em quatro quadrantes com base no seu envolvimento com o produto.

Os autores referem que "Os resultados da análise também podem ser associados aos dados de *Basket Analysis*, a fim de se desenvolverem estratégias de marketing e realizar sugestões promocionais eficazes" (Ozkan & Kocakoc, 2021, p.1).

Os autores indicam que o método RFM-V e a matriz *CustomerProduct Depth Matrix* permitem aos retalhistas analisar a relação entre os clientes e os produtos. Desta forma, é possível tomar medidas para mover os clientes para os quadrantes desejados. Por esse mesmo motivo, Ozkan & Kocakoc (2021) referem que os retalhistas devem implementar estratégias para melhorar as relações com os clientes de forma a movê-los para o quadrante superior direito da matriz, pois nenhum dos *clusters* está dentro do grupo *best* ou *hi-spender group* (figura 18).

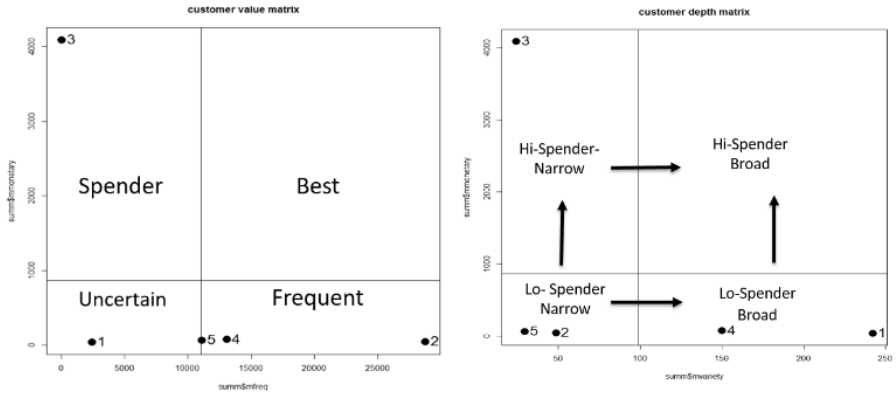


Figura 18- Representação dos quadrantes definidos (Ozkan & Kocakoc, 2021)

3. CASO DE ESTUDO

No presente capítulo, inicialmente, será apresentado um enquadramento do modelo de negócio em questão, de forma a realçar as suas principais caraterísticas e particularidades que o definem. Posteriormente, será apresentada uma análise minuciosa dos conjuntos de dados utilizados neste estudo, com o intuito de identificar e compreender as informações disponíveis, assim como o seu formato e a sua natureza.

Além disso, será conduzida uma etapa de preparação dos dados, visando a adequação destes para o formato necessário à investigação dos padrões de consumo e do desenvolvimento do sistema de recomendação de produtos. Posto isto, será conduzido o estudo de investigação dos padrões de consumo por meio da aplicação de regras de associação e do desenvolvimento do sistema de recomendação de produtos.

Por fim, um subcapítulo será dedicado à aplicação de modelos de previsão, destinados a determinar a probabilidade de recompra de determinado produto por clientes específicos. Para cada modelo em análise, serão delineadas as manipulações necessárias nos dados, bem como a apresentação dos resultados obtidos. Adicionalmente, será fornecida uma breve conclusão referente aos resultados de cada abordagem.

3.1. Compreensão do negócio

De acordo com a metodologia CRISP-DM, o processo de compreensão do negócio, como o próprio nome indica, trata-se da parte de entender o problema empresarial em questão (Kelleher et al., 2015).

No caso desta dissertação, o objetivo principal é determinar quais os produtos consumidos com uma maior frequência por parte dos clientes, utilizando ferramentas de regras de associação para identificar sequências de consumo frequentes. Com essas informações, será possível desenvolver um sistema de recomendação capaz de sugerir produtos relevantes aos clientes, mantendo a fidelidade dos mesmos e demonstrando um tratamento especializado por parte da empresa.

No contexto de um retalhista específico que detenha um sistema de recomendação de produtos baseado em regras de associação, surgem vantagens significativas. Esses sistemas têm a capacidade de sugerir produtos complementares ou relacionados ao que o cliente está à procura, resultando em uma experiência de compra mais personalizada e conveniente. Isso, por sua vez, aprimora a satisfação do cliente e favorece a sua fidelização. Para além disso, as sugestões de produtos têm o potencial de impulsionar o aumento do volume das vendas, o que para uma empresa é bastante atrativo, uma vez que contribuirá para o crescimento do negócio. A adoção de um sistema de recomendação de produtos tem a capacidade de posicionar uma empresa como líder num determinado setor, pelo que a capacidade de inovação e atenção às necessidades reais dos clientes cria uma vantagem significativa em relação às demais (Techlabs, 2017).

3.2. Compreensão dos dados

Neste presente capítulo, serão identificados e compreendidos os dados disponíveis nos diferentes conjuntos de dados em questão, bem como o tipo de dados que estão contidos nessas fontes.

3.2.1. Recolha de dados

O conjunto de dados utilizado no desenvolvimento desta dissertação denomina-se por *Instacart Market Basket Analysis* (*Kaggle*, 2023) e fornece informações sobre encomendas de clientes da Instacart, uma plataforma de compras online. Este conjunto de dados foi disponibilizado em 2017 para um concurso *Kaggle* e contém mais de 3 milhões de encomendas de produtos alimentares de mais de 200000 utilizadores *Instacart*.

Estes dados encontram-se disponíveis em sete ficheiros CSV *(Comma-separated Values)* que contêm diferentes caraterísticas descritas em maior detalhe na fase seguinte, compreensão dos dados. Um dos ficheiros denominado de "order_products_prior.csv" contém, originalmente, 32434489 exemplos e 4 atributos, porém dada à limitação do *hardware* disponível para processar os dados serão apenas considerados 1500000 exemplos deste ficheiro. Já o ficheiro "order_products_train.csv" será utilizado na sua totalidade. Com estas limitações associadas ao projeto, o conjunto de dados contém um total de 2884617 encomendas, das quais 279754 são distintas, 162823 utilizadores, 43800 produtos, 21 secções e 134 corredores diferentes.

3.2.2. Descrição dos dados

Nas tabelas 4, 5, 6, 7 e 8 são apresentados e compreendidos os dados disponíveis nos diferentes conjuntos de dados em questão, bem como o tipo de dados que estão contidos nessas fontes.

Tabela 4- Informação dos atributos da tabela orders.csv

Variável	Tipo	Descrição
Order_id	Inteiro	Chave única para cada ordem.
User_id	Inteiro	Chave única para cada cliente.
Eval_set	Texto	Informa a que conjunto (prior, train, test) cada ordem pertence.
Order_number	Inteiro	Sequência das ordens executadas por cliente.
Order_dow	Inteiro	Dia da semana.
Order_hour_of_day	Texto	Hora do dia (em horas).
Days_since_prior_order	Decimal	Diferença de dias entre 2 ordens.

A tabela 4 contém informação acerca de cada ordem efetuada.

Tabela 5- Informação dos atributos das tabelas order_products_prior.csv e order_products_train.csv

Variável	Tipo	Descrição
Order_id	Inteiro	Chave única para cada ordem.
Product_id	Inteiro	Chave única para cada produto.

Variável	Tipo	Descrição
Add_to_cart_order	Inteiro	Sequência na qual os produtos são adicionados ao cesto.
Reordered	Boleano	Indica se o produto foi comprado novamente

A tabela 5 contém informação sobre os produtos que foram encomendados em cada encomenda.

Tabela 6- Informação dos atributos da tabela products.csv

Variável	Tipo	Descrição
Product_id	Inteiro	Chave única para cada produto.
Product_name	Texto	Nome do produto.
Aisle_id	Inteiro	Chave única para o corredor onde está inserido o produto.
Department_id	Inteiro	Chave única para a secção em que está inserida o produto.

A tabela 6 contém informação detalhada sobre os produtos.

Tabela 7- Informação dos atributos da tabela aisles.csv

Variável	Tipo	Descrição
Aisle_id	Inteiro	Chave única para o corredor onde está inserido o produto.
Aisle_name	Texto	Nome do corredor.

A tabela 7 contém informação sobre os corredores.

Tabela 8- Informação dos atributos da tabela departments.csv

Variável	Tipo	Descrição
Department_id	Inteiro	Chave única para o corredor onde está inserido o produto.
Department_name	Texto	Nome da secção.

A tabela 8 contém informação sobre as secções.

O supermercado possui 134 corredores distintos, cada um com uma média de 328 produtos diferentes. Para além disso, em média, cada secção abrange 6 corredores distintos.

3.2.3. Exploração dos dados

Nesta presente secção é realizada uma análise individual das variáveis e também uma análise à relação existente entre as mesmas.

Após uma análise ao conteúdo dos dados é possível aferir que existem 96268 valores omissos (tabela 9) para o atributo “days_since_prior_order” (diferença de dias entre 2 ordens). Através do atributo “order_number” (sequência das ordens executadas por cliente) é possível verificar que quando o valor é igual a 1 (primeira ordem) o atributo “days_since_prior_order” é sempre omisso. Portanto, os valores que se encontram omissos serão substituídos por 0.

Tabela 9- Número de valores omissos

	Número de valores omissos
add_to_cart_order	0
reordered	0
order_number	0
order_dow	0
order_hour_of_day	0
days_since_prior_order	96268

Para uma rápida avaliação geral dos dados foi realizado um estudo estatístico. Os resultados indicam que o atributo “add_to_cart_order“ tem um valor máximo de 127, o que sugere que uma ordem (ordem 61355) continha 127 produtos. O valor médio desse atributo é de aproximadamente 8,545, o que significa que, em média, um carrinho de compras contém esse número de produtos. O atributo "order_number", que representa a sequência das ordens realizadas por um cliente, mostra que o número máximo de ordens executadas por um cliente é 100. O horário médio preferido pelos clientes (“order_hour_of_day”) para realizar suas ordens é às 13,49 horas (Apêndice A). O atributo "days_since_prior_order" indica a diferença de dias entre duas ordens, sendo que o valor máximo é 30 e a média é de 14,039 dias (Apêndice A). Por fim, tendo em consideração a existência de enviesamento nos dados, é importante notar que os atributos “add_to_cart_order”, "order_number" e "days_since_prior_order" possuem valores superiores a 0. Como resultado, a maioria dos outliers nesses atributos tende a estar concentrada na cauda direita da distribuição. O atributo “order_hour_of_day” possui um enviesamento de -0,077 o que indica que há uma assimetria na distribuição dos dados, inclinando-se para a esquerda. Neste caso é provável que maior parte dos outliers estejam concentrados na cauda esquerda da distribuição (Apêndice A).

3.2.3.1. Análise Univariada

Order_hour_of_day

De forma a analisar o período de maior afluência foi construído o gráfico da figura 19 em que se constata que a maior parte das ordens são posicionadas entre as 8h e as 18h.

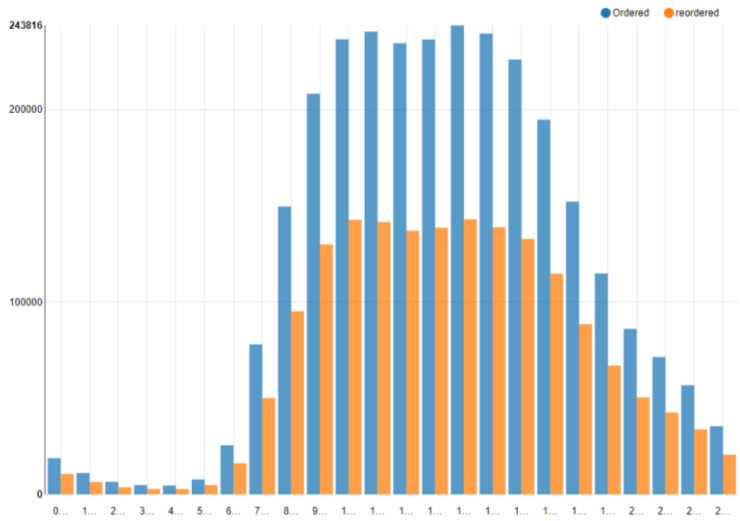


Figura 19- Período de maior afluência

Order_dow

Através da análise da figura 20 é possível auferir que grande parte das ordens são posicionadas no dia 0 e 1, sendo que no dia 3 e 4 são os dias em que existem menos ordens. Em relação ao reabastecimento de produtos, as pessoas tendem a executá-lo no dia 0. É importante referir que na descrição dos conjuntos de dados não se encontra qualquer informação acerca que valores representam que dia. Todavia, assumindo que o dia 0 seja o domingo, a razão pela qual as ordens são frequentemente posicionadas nesse dia pode ser atribuída ao facto de as pessoas não estarem a trabalhar, o que lhes permite aproveitar para realizar as suas compras semanais.

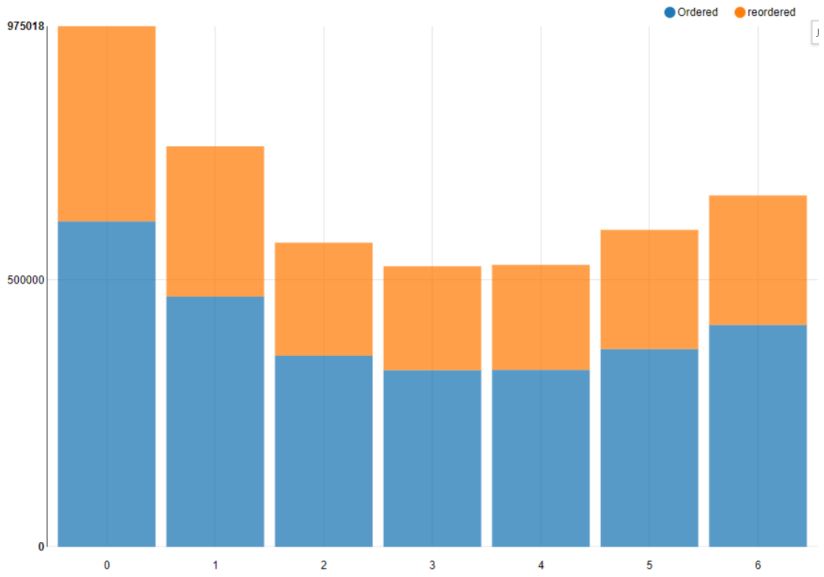


Figura 20- Dias de maior consumo

Days_since_prior_order

Com base na figura 21, observa-se que a maioria dos clientes efetua novas compras após uma semana ou um mês, o que sugere que a preferência da maioria é por adquirir produtos com durabilidade semanal ou mensal. Além disso, constata-se uma tendência crescente nos dias 1 a 6,

o que indica que alguns compradores são frequentes e optam por fazer reposições de forma mais frequente em um intervalo de tempo menor.

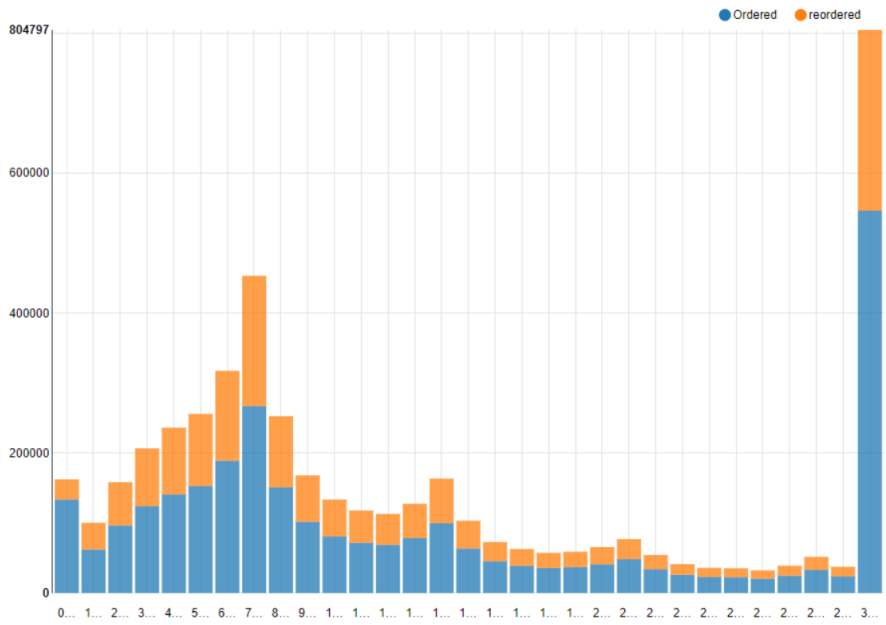


Figura 21- Frequência de compras realizadas

3.2.3.2. Análise multivariada

Uma matriz de correlação foi criada para estudar a relação entre as variáveis, como mostrado na figura 22. A análise da matriz destaca uma correlação média entre o “reordered” e o atributo "order_number". Além disso, os restantes pares de variáveis apresentam correlações fracas, o que sugere que as variáveis em questão não possuem uma relação linear forte entre si, ou seja, uma mudança em uma variável não está fortemente associada a mudanças na outra variável.

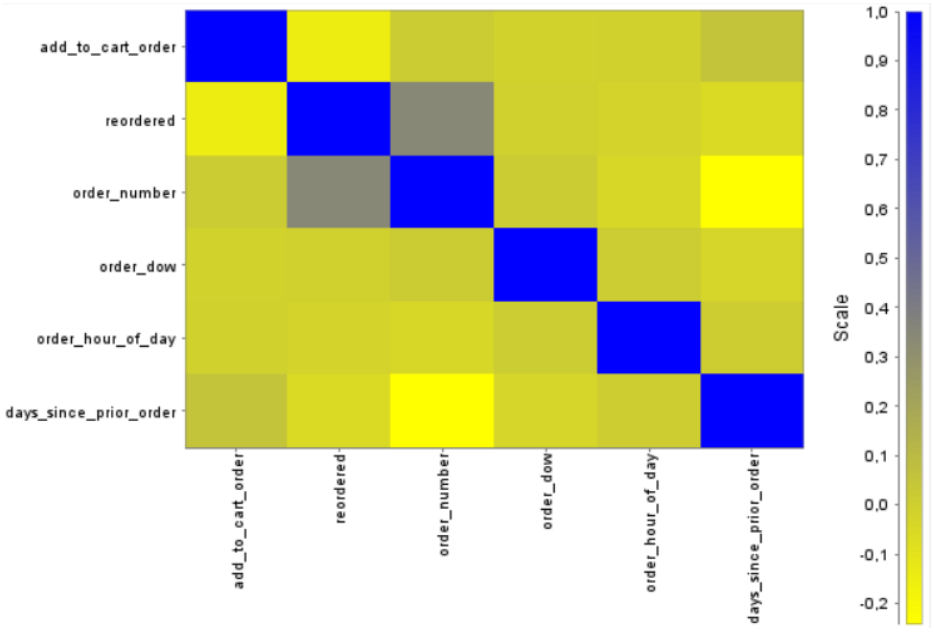


Figura 22- Matriz de correlação

Através da análise da figura 23 verifica-se que os utilizadores, maior parte das vezes compram em torno de 5 artigos, sendo as bananas o artigo mais popular (figura 24). Para além disso, 43800 produtos existentes, 34647 foram recomprados, destacando-se os produtos ilustrados na figura 24.

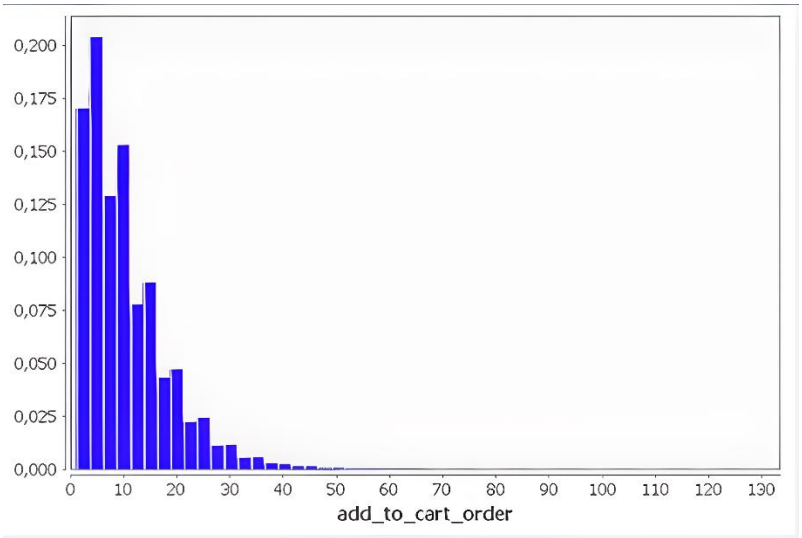


Figura 23- Número de artigos comprados por cliente

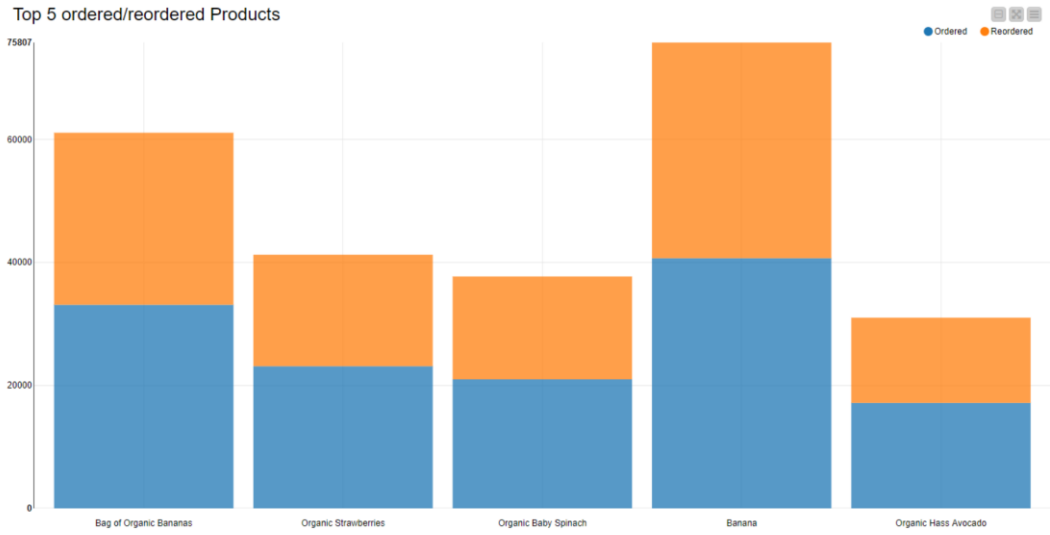


Figura 24- Top 5 produtos mais vezes comprados/recomprados

As bananas são frequentemente selecionadas como o primeiro artigo a ser adicionado ao carrinho (conforme indicado na figura 25).

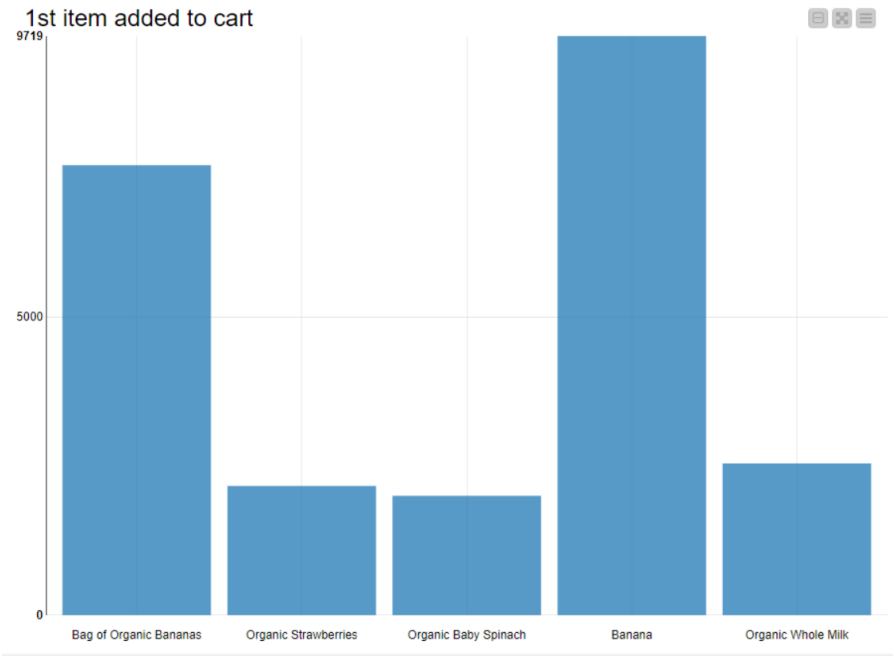


Figura 25- Primeiros artigos a serem colocados nos cestos de compras

Os corredores que possuem um maior número de vendas são *os fresh fruit* (frutas frescas), *fresh vegetables* (vegetais frescos), *packaged vegetables fruits* (vegetais e frutas embaladas), *yogurt* (iogurte) e *packaged cheese* (queijo embalado). Já os produtos frequentemente recomprados são os *fresh fruit, fresh vegetables, packaged vegetables fruits, yogurt* e *milk* (figura 26). Esse padrão poderá ser explicado pelo facto de que esses produtos são consumidos diariamente e têm um prazo de validade mais reduzido.

Em contrapartida, os corredores *de beauty* (produtos de beleza), *kitchen supplies* (utensílios de cozinha) e *first aid* (primeiros socorros) contêm produtos que possuem um longo período de duração, pelo que são menos vezes recomprados.

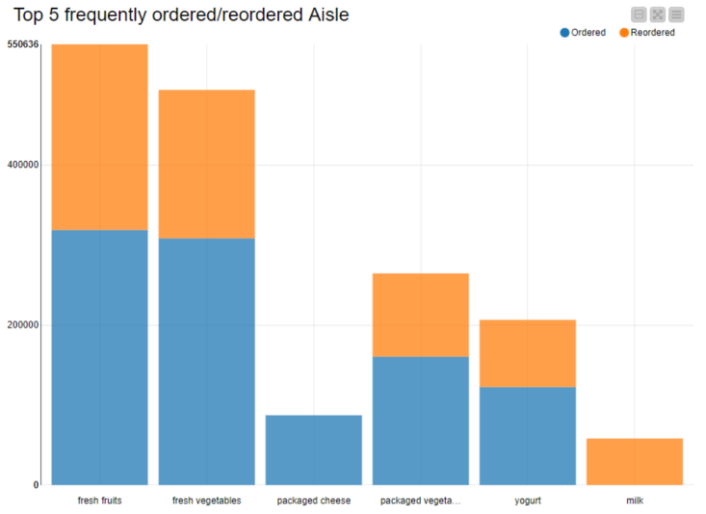


Figura 26- Top 5 corredores com maior número de vendas

As cinco secções com um maior número de vendas e, portanto, os que apresentam produtos mais frequentemente recomprados são os *produce* (produtos hortícolas), *dairy eggs* (lacticínios e ovos), *beverages* (bebidas), *snacks* (aperitivos) e *frozen* (congelados) (figura 27).

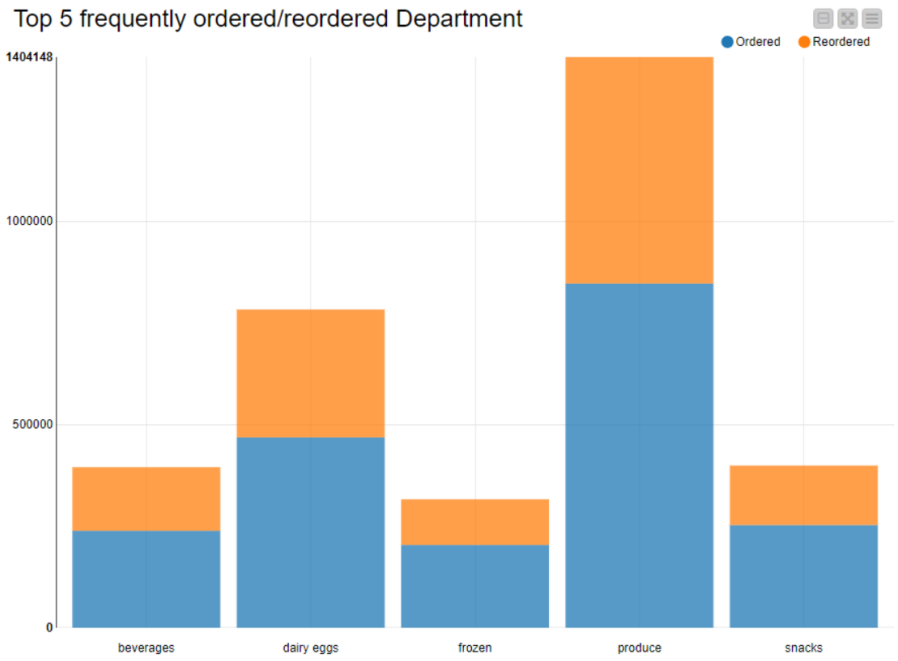


Figura 27- Top 5 secções com maior número de vendas

Como foi constado anteriormente, os três corredores com um maior número de vendas são os *fresh fruit* (frutas frescas), *fresh vegetables* (vegetais frescos) e *packaged vegetables fruits* (vegetais e frutas embaladas), todos pertencentes à secção *produce* (produtos hortícolas) (figura 28).

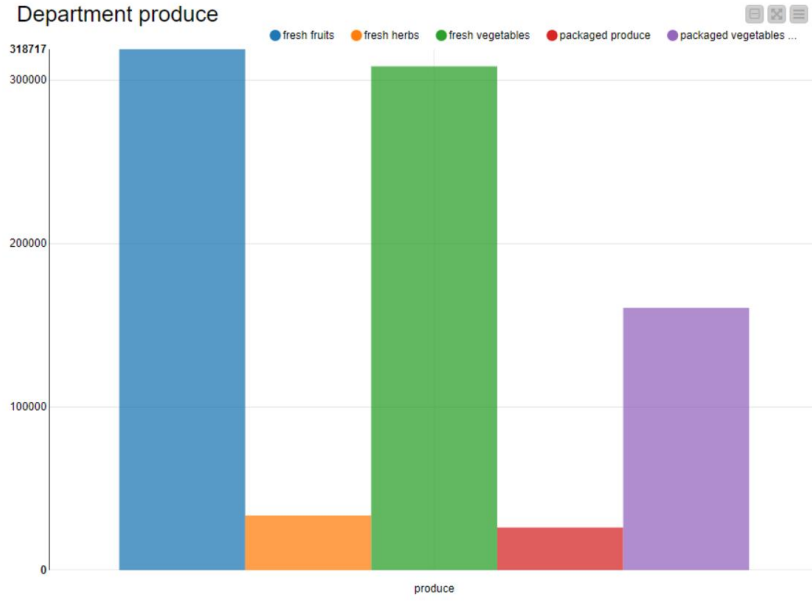


Figura 28- Secção *produce*

3.2.3.3. Hábitos de consumo

Clientes que realizaram sempre recompras de produtos

Aproximadamente 18026 clientes efetuaram exclusivamente recompras de produtos, o que sugere que os mesmos compraram os mesmos produtos repetidamente, demonstrando que possuem sempre os mesmos hábitos de consumo. Entre esses 18026 clientes, destacam-se os clientes com o user_id 97865 e 91826 com o maior número produtos recomprados.

Ao observar a lista de produtos adquiridos pelos clientes 97865 e 91826, não se observa um produto em particular que se destaque como o mais adquirido. Em contrapartida, um conjunto de produtos foi comprado em quantidades semelhantes.

Clientes que realizaram mais ordens/compras

O cliente identificado pelo *user_id* 34340 efetuou um total de 16 ordens diferentes. Apesar disso, o cliente 100330 foi o que adquiriu uma maior quantidade de produtos, porém numa menor quantidade de ordens, cerca de 6.

Destaca-se que o produto preferido do cliente 34340 é claramente a soda, pois em 11 das 16 ordens que realizou consta a inclusão desse produto (figura 29).

É possível perceber que o cliente 100330 tem preferência pelo horário compreendido entre as 10h e as 14h para realizar as suas compras.

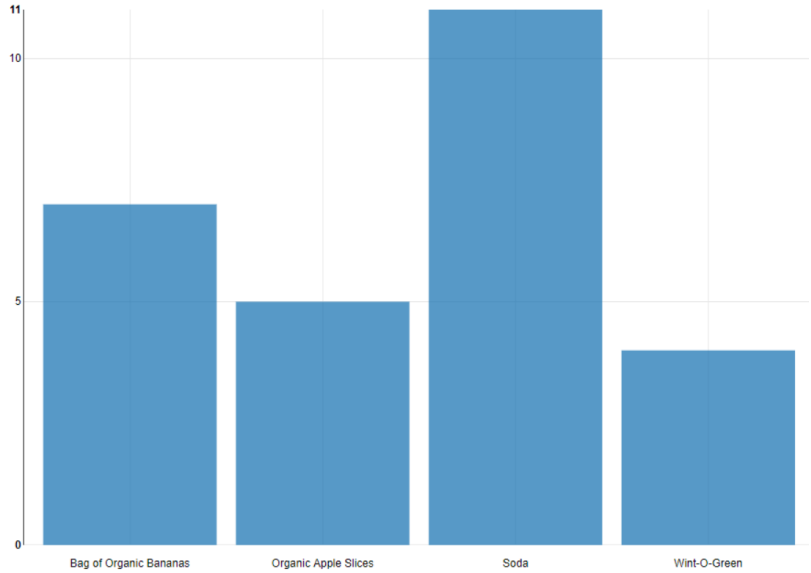


Figura 29- Preferências do cliente 34340

Produtos que nunca foram recomprados

Um total de 9017 produtos foram adquiridos apenas uma vez pelos clientes, sendo que entre eles se destacam os produtos *Seasoning* (condimentos)*, Ground Turmeric* (açafrão moído)*, Ground Allspice* (Pimenta da Jamaica moída) e *Organic Allspice* (Pimenta da Jamaica orgânica) como os produtos mais vezes adquiridos somente uma vez pelos clientes. Grande parte desses produtos pertence à secção *pantry* (despensa), mais especificamente ao corredor *spice seasonings* (especiarias). Essa observação pode indicar que houve uma grande expectativa em torno desses

produtos, mas eles não conseguiram atender às necessidades e expectativas dos clientes. Portanto, é possível que a secção/corredor precise ser reformulado para melhor atender às necessidades dos consumidores. Outra ilação que pode ser retirada é que, uma vez que se tratam de produtos destinados a temperar alimentos, estes têm longa durabilidade, o que resulta em compras menos frequentes por parte dos clientes.

Produtos que sempre foram recomprados

Foram identificados 3053 produtos que apresentam apenas registos de recompras, destacando-se entre eles o *Organic Grade A Raw Whole Milk* (leite integral orgânico grau A)*, Baked Whole Grain Wheat Parmesan Garlic Triscuit* (triscuit integral de trigo assado com parmesão e alho*, Crystals Cat Litter* (areia para gatos) e *Artisanal Nonfat Coffee Greek Yogurt* (iogurte grego sem gordura artesanal com café). Este fenómeno pode indicar que estes produtos apresentam uma alta qualidade e agradam a maioria dos consumidores, sendo considerados produtos confiáveis e capazes de garantir a fidelidade dos clientes.

3.3. Preparação dos dados

Neste capítulo, serão descritas as manipulações efetuadas no conjunto de dados para encontrar regras de associação e desenvolver o sistema de recomendação de produtos.

3.3.1. Preparação dos dados para as regras de associação

De forma a utilizar os dados transacionais, é necessário efetuar um pré-processamento, convertendo-os num formato que possa ser utilizado para a extração de regras de associação. Para isso, foi efetuado um agrupamento dos dados por transação, em que em cada transação contém a lista de produtos que foram comprados (figura 30).

Row ID	order_id	Sorted list(product_name)
Row0	1	[Bag of Organic Bananas,Bulgarian Yogurt,Cucumber Kirby,...
Row1	2	[All Natural No Stir Creamy Almond Butter,Carrots,Classic Bl...
Row2	3	[Air Chilled Organic Boneless Skinless Chicken Breasts,Lemo...
Row3	4	[Chewy 25% Low Sugar Chocolate Chip Granola,Energy Dri...
Row4	5	[2% Reduced Fat Milk,American Slices Cheese,Apricot Pres...
Row5	6	[Clean Day Lavender Scent Room Freshener Spray,Cleanse...
Row6	7	[Orange Juice,Pineapple Chunks]
Row7	8	[Original Hawaiian Sweet Rolls]
Row8	9	[100% Apple Juice Original,Baby Spinach,Distilled Water,...]
Row9	10	[Baby Portabella Mushrooms,Banana,Boneless Beef Sirloin S...
Row10	11	[Extra Virgin Olive Oil,Mango Pineapple Salsa,Teriyaki & Pine...
Row11	12	[100% Cranberry Juice,100% Juice No Added Sugar Orang...

Figura 30- Amostra do conjunto de dados utilizado para as regras de associação

3.3.2. Preparação dos dados para o sistema de recomendação de produtos

Para o primeiro estudo descrito na seção 3.4.2 (Sistemas de Recomendação com métricas de Confiança e Suporte de valores altos), foi utilizado o nó *row sampling* de forma que 10 carrinhos de compras fossem selecionados de forma aleatória. Posto isto, uma vez mais, o nó *row sampling* volta a ser utilizado para remover 1 a 2 produtos aleatoriamente desses mesmos carrinhos. O objetivo

dessa abordagem é avaliar se o sistema de recomendação desenvolvido na secção 3.4.2 é capaz de recomendar os produtos que foram previamente removidos dos carrinhos.

Os 10 carrinhos de compras selecionados para serem utilizados como cestos de compras em estudo foram removidos do conjunto de dados antes da aplicação do nó *association rule learner*. Essa remoção foi feita para garantir que o algoritmo não utilize esses carrinhos na procura por conjuntos frequentes de produtos.

Com o uso do nó *subset matcher* é possível comparar todos os subconjuntos da primeira tabela de entrada com todos os conjuntos da segunda tabela de entrada, para encontrar correspondência entre eles, ajudando a identificar transações que se encaixam em conjuntos de itens específicos.

No apêndice B é possível consultar a estrutura de nós que foi desenvolvida para realizar o primeiro estudo referente ao desenvolvimento do sistema de recomendação avaliado com base nas métricas de confiança e suporte de valores altos.

No segundo estudo, mencionado na seção 3.4.2 (Sistemas de Recomendação com métricas de *Lift* e Suporte de valores altos), foram aplicados dois nós *Rule-based Row Filter.* O primeiro nó foi utilizado para filtrar pelos 10 carrinhos de compras que foram selecionados, de forma aleatória, para o primeiro estudo. O segundo nó foi empregue para filtrar, em cada pedido, pelos "artigos a serem sugeridos" que foram selecionados, de forma aleatória, no primeiro estudo. Essa abordagem permite comparar a eficácia dos dois estudos e determinar qual deles apresenta um desempenho superior. Além disso, antes de fornecer o input para o nó *subset Matcher*, foi aplicado um nó *Joiner* para remover os artigos a serem sugeridos do carrinho de compras original (apêndice C).

3.4. Regras de associação e desenvolvimento do sistema de recomendação de produtos

Neste capítulo, serão realizados dois estudos. No primeiro subcapítulo será realizada uma investigação dos padrões de consumo por meio da aplicação de regras de associação. Em seguida, no segundo subcapítulo, será desenvolvido um sistema de recomendação com base nas regras de associação identificadas anteriormente.

3.4.1. Regras de associação

As regras de associação consistem em uma técnica de mineração de dados que tem por objetivo identificar relações entre os produtos presentes em um conjunto de dados específico. Esse tipo de análise é valioso para identificar padrões de consumo, compreender o comportamento do cliente e ajudar as empresas a tomar decisões mais fundamentadas sobre seus produtos. Por exemplo, um supermercado, após a análise, descobre que o cliente que compra cerveja também compra tremoços. Com base nessa informação a loja poderá aplicar descontos estratégicos num destes produtos ou até mesmo posicionar esses artigos de forma estratégica, com o objetivo de incentivar os clientes a comprarem ambos os produtos. Essa ação poderá resultar num aumento de vendas e por consequência numa maior satisfação por parte dos clientes.

3.4.1.1. Cálculo e avaliação das regras de associação

Para identificar padrões nas transações do conjunto de dados em estudo, foi utilizado o nó *Association Rule Learner* na ferramenta KNIME. Esse nó baseia-se no algoritmo *Apriori* e permite descobrir relações entre produtos que os clientes geralmente compram em conjunto.

Para além disso, esse nó tem como output os valores das três métricas utilizadas (*support*, *confidence* e *lift*) para avaliar a qualidade das regras de associação geradas (figura 31).

D Support	D Confide...	D Lift	S Conseq...	S implies	[...] Items
0.005	0.205	1.409	Banana	<---	[Sparkling Water Grapefr...
0.005	0.034	1.409	Sparkling W...	<---	[Banana]
0.005	0.17	1.167	Banana	<---	[Organic Cucumber]
0.005	0.034	1.167	Organic Cuc...	<---	[Banana]
0.005	0.179	1.515	Bag of Orga...	<---	[Seedless Red Grapes]
0.005	0.043	1.515	Seedless Re...	<---	[Bag of Organic Bananas]
0.005	0.15	1.817	Organic Stra...	<---	[Organic Yellow Onion]
0.005	0.061	1.817	Organic Yell...	<---	[Organic Strawberries]
0.005	0.127	1.695	Organic Bab...	<---	[Organic Whole Milk]
0.005	0.068	1.695	Organic Wh...	<---	[Organic Baby Spinach]
0.005	0.229	1.577	Banana	<---	[Organic Gala Apples]
0.005	0.035	1.577	Organic Gal...	<---	[Banana]
0.005	0.152	1.044	Banana	<---	[Organic Yellow Onion]

Figura 31- Exemplo de um output gerado pelo nó *Association Rule Learner*

Definiu-se um limite inicial de suporte de 0,001, visando identificar combinações de produtos com um alto nível de confiança (tabela 10).

Tabela 10- 10 primeiras regras com um alto nível de confiança e um limite mínimo de suporte de 0,001

A	B	*Support*	*Confidence*	*Lift*
Organic Strawberries, Organic Raspberries, Organic Hass Avocado	Bag of Organic Bananas	0,001	0,539	4,554
Organic Navel Orange, Organic Hass Avocado	Bag of Organic Bananas	0,001	0,505	4,263
Strawberries, Honeycrisp Apple	Banana	0,001	0,502	3,45
Lime Sparkling Water, Sparkling Lemon Water	Sparkling Water Grapefruit	0,001	0,489	20,012
Organic Raspberries, Organic Hass Avocado	Bag of Organic Bananas	0,004	0,4863	4,106
Organic Hass Avocado, Organic Kiwi	Bag of Organic Bananas	0,001	0,483	4,08
Strawberries, Organic Fuji Apple	Banana	0,001	0,477	3,28
Organic Strawberries, Organic Navel Orange	Bag of Organic Bananas	0,001	0,468	3,954
Organic Raspberries, Organic Kiwi	Bag of Organic Bananas	0,001	0,466	3,933
Sparkling Water Grapefruit, Sparkling Lemon Water	Lime Sparkling Water	0,001	0,46	31,252

Com base nos resultados apresentados da tabela 10, podemos concluir que os produtos orgânicos, em particular as frutas orgânicas como *organic* strawberries (morangos orgânicos), *organic raspberries* (framboesas orgânicas), *organic hass avocado* (abacate orgânico), *organic navel Orange*

(laranja navel orgânica), *organic kiwi* (kiwi orgânico) e *organic fuji* apple (maçã fuji orgânica), tendem a ser comprados em conjunto com *Bag of Organic Bananas* (saco de bananas orgânicas). Essa associação forte entre esses produtos orgânicos e bananas orgânicas sugere que os consumidores que compram um desses produtos orgânicos provavelmente também comprarão bananas orgânicas.

Do ponto de vista de negócio, esta informação poderá ser útil para criar estratégias para impulsionar as vendas das bananas orgânicas, tais como:

- Criar um pack de produtos orgânicos para incentivar a compra das bananas orgânicas. Por exemplo, um pack poderá ser constituído por morangos orgânicos, framboesas orgânicas e abacate orgânico a um preço com desconto para incentivar os clientes a comprar mais produtos orgânicos e também aumentar as vendas das bananas orgânicas.
- Criar uma campanha publicitária destacando os benefícios dos produtos orgânicos. Esta estratégia poderá ajudar a conscientizar os clientes sobre as vantagens dos produtos orgânicos em relação aos convencionais, incentivando à compra desses produtos, incluindo as bananas orgânicas.

A tabela 10 apresenta dois casos relevantes para análise: a compra conjunta dos produtos *Lime Sparkling Water* (água com gás de lima) e *Sparkling Lemon Water* (água com gás de limão), e dos produtos *Sparkling Water Grapefruit* (água com gás de toranja) e *Sparkling Lemon Water* (água com gás de limão). Ambos os casos apresentam valores altos de confiança e *lift*. Partindo da análise do primeiro caso, a alta confiança de 0,489 indica que os clientes que compram *Lime Sparkling Water* (água com gás de lima) e *Sparkling Lemon Water* juntos têm uma alta probabilidade de também comprarem *Sparkling Water Grapefruit.* O *lift* muito forte (20,012) sugere que essa associação é muito mais forte do que o esperado ao acaso. A probabilidade de *Sparkling Water Grapefruit* aumenta em 20,012 vezes quando os produtos *Lime Sparkling Water* e *Sparkling Lemon Water* são comprados. Isso indica que há um segmento de clientes que preferem comprar água com gás com sabor cítrico e que optam por experimentar diferentes sabores.

Com base nesses resultados, pode-se inferir que a criação de um pacote que inclua os produtos *Lime Sparkling Water* e *Sparkling Lemon Water* com um desconto promocional pode impulsionar a venda do produto *Sparkling Water Grapefruit*. Além disso, o produto *Sparkling Water Grapefruit* pode ser mantido com seu preço original, já que a compra do mesmo é quase garantida quando os clientes compram os outros dois produtos juntos. Essa estratégia pode ser eficaz para impulsionar as vendas desses produtos e atender às preferências dos clientes que buscam sabores cítricos em suas bebidas.

Ao manter o limite inicial de suporte em 0,001, foi observado que as regras de associação com os maiores valores de *lift* são formadas por produtos do corredor de iogurtes e da secção de ovos lácteos (tabela 11).

Tabela 11- 10 primeiras regras com um alto nível de lift e um limite mínimo de suporte de 0,001

A	B	*Support*	*Confidence*	*Lift*
Blueberry Yoghurt	Strawberry Rhubarb Yoghurt	0,001	0,302	87,373
Strawberry Rhubarb Yoghurt	Blueberry Yoghurt	0,001	0,295	87,373

A	B	*Support*	*Confidence*	*Lift*
Nonfat Icelandic Style Strawberry Yogurt	Non Fat Raspberry Yogurt	0,001	0,372	80,323
Non Fat Raspberry Yogurt	Nonfat Icelandic Style Strawberry Yogurt	0,001	0,238	80,323
Non Fat Raspberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0,002	0,426	76,606
Icelandic Style Skyr Blueberry Non-fat Yogurt	Non Fat Raspberry Yogurt	0,002	0,354	76,606
Nonfat Icelandic Style Strawberry Yogurt	Icelandic Style Skyr Blueberry Non-fat Yogurt	0,001	0,425	76,39
Icelandic Style Skyr Blueberry Non-fat Yogurt	Nonfat Icelandic Style Strawberry Yogurt	0,001	0,226	76,39
Non Fat Raspberry Yogurt	Non Fat Acai & Mixed Berries Yogurt	0,001	0,22	74,057
Non Fat Acai & Mixed Berries Yogurt	Non Fat Raspberry Yogurt	0,001	0,343	74,057

Analisando a regra de associação com o maior valor de *lift* (87,37) e uma confiança de 0,302, em que o item A é *Blueberry Yoghurt* (iogurte de mirtilo) e o consequente B é *Strawberry Rhubarb Yoghurt* (iogurte de morango e ruibarbo), conclui-se que a compra desses dois itens em conjunto é muito mais comum do que o esperado pelo acaso. Embora a confiança não seja a mais elevada (indicando que em 30,2% das vezes em que *Blueberry Yoghurt* é comprado, *Strawberry Rhubarb Yoghurt* também é comprado), o *lift* é bastante significativo, sugerindo que a probabilidade de comprar *Strawberry Rhubarb Yoghurt* aumenta em 87,37 vezes quando *Blueberry Yoghurt* é comprado. A partir da análise dos resultados da tabela 11, torna-se evidente a correlação existente entre a disposição física dos produtos no mesmo corredor e na mesma seção, e o consequente aumento nas vendas dos produtos similares.

Com o objetivo de estudar exclusivamente as regras de associação mais relevantes, que apresentassem um suporte mínimo de 0,005, foi necessário alterar o valor mínimo de suporte para 0,005 (tabela 12).

Tabela 12- 10 primeiras regras com um alto nível de confiança e um limite mínimo de suporte de 0,005

A	B	*Support*	*Confidence*	*Lift*
Organic Strawberries	Bag of Organic Bananas	0,021	0,256	2,162
[Bag of Organic Bananas	Organic Strawberries	0,021	0,179	2,162
Organic Hass Avocado	Bag of Organic Bananas	0,019	0,308	2,604
Bag of Organic Bananas	Organic Hass Avocado	0,019	0,16	2,604
Organic Strawberries	Banana	0,017	0,209	1,439
[Banana]	Organic Strawberries	0,017	0,119	1,439
Organic Avocado	Banana	0,017	0,301	2,069
Banana	Organic Avocado	0,017	0,114	2,069
Organic Baby Spinach	Bag of Organic Bananas	0,016	0,218	1,842
Bag of Organic Bananas	Organic Baby Spinach	0,016	0,138	1,842

Embora a tabela 12 apresente valores de suporte mais elevados do que aqueles exibidos na tabela 7, o conjunto de produtos é o mesmo, sendo compostos por produtos orgânicos e mostrando que há uma alta probabilidade de adquirir bananas. Isso evidencia um padrão de consumo de produtos saudáveis e naturais.

Além das regras mencionadas anteriormente, há quatro outras regras de associação na tabela 13 que são consideradas interessantes para análise, pois não envolvem produtos orgânicos ou bananas.

Tabela 13- Regras de associação relevantes para produtos não orgânicos e que não incluem bananas

A	B	*Support*	*Confidence*	*Lift*
Zero Calorie Cola	Soda	0,001	0,417	36,804
Soda	Zero Calorie Cola	0,001	0,099	36,804
Apples	Clementines	0,001	0,332	33,648
Clementines	Apples	0,001	0,108	33,648

A análise da tabela 13 mostra que existem associações relevantes entre alguns pares de produtos. Observa-se uma alta probabilidade (0,417) de quem compra *Soda* (refrigerante) também adquirir *Zero Calorie Cola* (coca-cola zero), enquanto quem compra *Zero Calorie Cola* não tem uma alta probabilidade de adquirir *Soda* (0,099). Isso sugere que, do ponto de vista promocional, seria interessante fazer uma promoção no produto *Soda*, já que a probabilidade de o cliente adquirir *Zero Calorie Cola* é elevada. Por outro lado, colocar a promoção no produto *Zero Calorie Cola* não seria tão interessante, pois há um maior risco de o cliente adquirir apenas esse produto, sem levar também o *Soda*. O mesmo padrão pode ser observado no caso das *Clementines* (clementinas). Seria interessante colocar uma promoção nas clementinas, já que a probabilidade de compra do produto *Apples* (maçãs) é alta (0,332). Além disso, a probabilidade de compra do produto *Apples* aumenta 33,648 vezes quando o produto *Clementines* é adquirido, o que sugere que há uma grande chance de o cliente comprar ambos os produtos.

Uma vez que este conjunto de dados contém uma enorme variedade de produtos, os valores de suporte obtidos são relativamente baixos, visto que a proporção de produtos por transação é menor. Com o objetivo de estudar combinações de produtos de uma forma menos granular, optou-se por aumentar um nível na hierarquia e realizar o mesmo estudo, mas desta vez por corredor (tabela 14). No entanto, ao definir um limite mínimo de suporte de 0,001 no nível hierárquico por secção, foram encontradas inúmeras regras de associação, tornando difícil o processo de processamento e análise. Portanto, decidiu-se estabelecer um limite mínimo inicial de suporte de 0,10.

Tabela 14- Regras de associação com os valores mais altos de suporte (suporte mínimo=0,10)

A	B	*Support*	*Confidence*	*Lift*
fresh vegetables	fresh fruits	0,322	0,72	1,301
fresh fruits	fresh vegetables	0,322	0,582	1,301
packaged vegetables fruits	fresh fruits	0,279	0,743	1,342
fresh fruits	packaged vegetables fruits	0,279	0,504	1,342
packaged vegetables fruits	fresh vegetables	0,243	0,648	1,448

A	B	*Support*	*Confidence*	*Lift*
fresh vegetables	packaged vegetables fruits	0,243	0,544	1,448
fresh vegetables, packaged vegetables fruits	fresh fruits	0,195	0,801	1,448
fresh fruits, packaged vegetables fruits	fresh vegetables	0,195	0,699	1,563
fresh fruits, fresh vegetables	packaged vegetables fruits	0,195	0,605	1,611
yogurt	fresh fruits	0,187	0,721	1,303

Através da análise da tabela 14, verifica-se que o conjunto de artigos pertencentes ao corredor *fresh vegetables*, *fresh fruits* e *packaged vegetables fruits* aparecem com muita frequência nas transações deste conjunto de dados. Por exemplo, a combinação de produtos pertencentes ao corredor *fresh vegetables* e *fresh fruits* aparecem juntos em 32,2% de todas as transações registadas neste conjunto de dados, o que sugere que produtos pertencentes a estes corredores são adquiridos com muita frequência. Para além disso, há uma alta probabilidade de que os clientes que compram produtos dos corredores *fresh vegetables* e *packaged vegetables fruits* também comprem produtos do corredor *fresh fruits*. Já o valor do *lift* é superior a 1, o que sugere que essas associações ocorrem com mais frequência do que o esperado aleatoriamente. Portanto, estratégias de posicionamento de produtos ou até mesmo de promoções podem ser adotadas, tais como:

- Aumentar a disponibilidade de produtos de *fresh fruits* próximo aos produtos de *fresh vegetables* e *packaged vegetables fruits* para incentivar os clientes a comprarem esses produtos em conjunto;

- Criar promoções especiais que incluam produtos de 3 diferentes corredores;

- Aprimorar o layout da loja, colocando os produtos dos corredores *fresh fruits* próximos aos produtos dos corredores *fresh vegetables* e *packaged vegetables fruits*, ou mesmo criando uma única seção ou área aberta que englobe esses três corredores.

Na tabela 15, encontram-se representadas as regras de associação com um maior grau de confiança com um limite mínimo de suporte de 0,10.

Tabela 15- Regras de associação com os valores mais altos de confiança (suporte mínimo=0,10)

A	B	*Support*	*Confidence*	*Lift*
yogurt, packaged vegetables fruits	fresh fruits	0,108	0,828	1,496
yogurt, fresh vegetables	fresh fruits	0,120	0,825	1,49
fresh vegetables, packaged vegetables fruits	fresh fruits	0,195	0,801	1,448
packaged cheese, fresh vegetables	fresh fruits	0,109	0,778	1,406
packaged vegetables fruits	fresh fruits	0,279	0,743	1,342
yogurt	fresh fruits	0,187	0,721	1,303
fresh vegetables	fresh fruits	0,322	0,72	1,301
fresh fruits, packaged vegetables fruits	fresh vegetables	0,195	0,699	1,563
soy lactosefree	fresh fruits	0,119	0,698	1,26
bread	fresh fruits	0,119	0,698	1,26

Partindo da análise da tabela 15, observa-se que independentemente do antecedente em questão, as regras de associação apresentam uma maior probabilidade de compra dos produtos do corredor *fresh fruits*. Isso confirma que a colocação dos produtos dos corredores *yogurt*, *packaged vegetables fruits*, *fresh vegetables* e até mesmo *soy lactosefree* (soja sem lactose) e *bread* (pão) em uma área próxima ou comum pode potencializar a compra em conjunto, aumentando a probabilidade de compra. É importante destacar que as regras de associação são baseadas em padrões históricos de compra e não garantem que um cliente comprará determinado produto na próxima visita. No entanto, essas regras podem ser úteis para o retalhista ajustar a disposição dos produtos e otimizar as vendas em conjunto.

3.4.2. Sistemas de recomendação baseado em regras de associação

O sistema de recomendação de produtos baseado em regras de associação são modelos utilizados para efetuar recomendações de artigos aos consumidores com base em histórico de transações passadas. As regras de associação são geradas, revelando as preferências e hábitos de consumo dos clientes, através das associações frequentes encontradas entre os diferentes artigos. Estes sistemas exploram a relação existente entre artigos, num determinado conjunto de dados, de forma a identificar padrões de consumo e, consequentemente recomendar produtos relevantes aos clientes. Os sistemas de recomendação são uma ferramenta essencial para as empresas, na medida em que permitem impulsionar as suas vendas e fornecem a capacidade de efetuar recomendações personalizadas que aumentam a fidelidade dos clientes.

Serão realizados dois estudos de precisão para sistemas de recomendação: um com base em altos níveis de confiança e suporte, e outro com base em altos níveis de *lift* e suporte. Cada estudo envolverá 20 carrinhos de compras. Desses, 10 carrinhos terão os produtos agrupados por nome de produto, enquanto os outros 10 serão agrupados por corredor. Durante cada estudo, um a dois produtos serão selecionados aleatoriamente e removidos dos carrinhos, a fim de testar a precisão dos sistemas de recomendação. Os detalhes sobre como os dados são transformados para essa seleção de carrinhos podem ser encontrados na seção acima, na seção 3.3.

Os carrinhos de compra usados no estudo, agrupados por nome de produto, correspondem aos carrinhos de número 3, 10, 21, 26, 48, 53, 54, 66, 77 e 87. Por outro lado, os carrinhos agrupados por corredor são identificados pelos números de ordem 3, 5, 11,20, 21, 51, 53, 64, 74 e 76.

3.4.2.1. Sistemas de Recomendação com métricas de Confiança e Suporte de valores altos

Os estudos subsequentes seguem a mesma sequência de estudo. Será apresentada em detalhe a lógica implementada para o primeiro carrinho de compras, e para os demais carrinhos, será fornecida uma tabela resumo indicando se o sistema de recomendação "acertou" ou "não acertou", em pelo menos um dos produtos, tendo por base as métricas de confiança e suporte de valores altos.

Carrinhos agrupados por nome do produto

No primeiro carrinho de compras analisado, inicialmente havia 8 produtos. Aleatoriamente, os produtos *Air Chilled Organic Boneless Skinless Chicken Breasts* (peitos de frango orgânicos sem osso

e sem pele) e *Organic Ezekiel 49 Bread Cinnamon Raisin* (pão orgânico ezekiel 49 com canela e passas) foram removidos com o propósito de testar se o sistema de recomendação será capaz de recomendar um desses produtos.

Conforme evidenciado na Figura 32, os produtos recomendados pelo sistema de recomendação foram Banana e *Total 2% Lowfat Greek Strained Yogurt With Blueberry* (iogurte grego magro total 2% com mirtilo), pelo que o sistema de recomendação não conseguiu acertar em nenhum dos dois produtos que foram inicialmente removidos aleatoriamente do carrinho de compras.

Figura 32- Sistema de recomendação de produtos baseado em regras de associação (nome de produtos)

De seguida, será apresentada uma tabela resumo (tabela 16) mencionando se o sistema de recomendação "acertou" ou "não acertou" para os demais carrinhos de compras, juntamente com a taxa de acerto do sistema de recomendação de produtos para carrinhos agrupados por nome dos produtos.

Tabela 16- Taxa de acerto do sistema de recomendação de produtos (nome de produtos) | Confiança e Suporte

Carrinho de compras (order_id)	**Artigos a sugerir**	**Artigo(s) sugerido(s)**	**Acertou/Não acertou**
3	Air Chilled Organic Boneless Skinless Chicken Breasts; Organic Ezekiel 49 Bread Cinnamon Raisin"	Banana; Total 2% Lowfat Greek Strained Yogurt With Blueberry	Não acertou
10	Spinach Peas & Pear Stage 2 Baby Food; Yellow Onions	Limes	Não acertou
21	Frozen Organic Blueberries; Organic Firm Tofu	Não foram encontrados subconjuntos para este carrinho de compras	-
26	Banana; Boneless Skinless Chicken Breasts	Banana	**Acertou**
48	Low Sodium Beef Broth; Organic Unrefined Toasted Sesame Oil	Banana; Organic Strawberries	Não acertou

Carrinho de compras (order_id)	Artigos a sugerir	Artigo(s) sugerido(s)	Acertou/Não acertou
53	Organic Blueberries; Organic Hass Avocado	Organic Baby Spinach; Bag of Organic Bananas	Não acertou
54	Instant French Vanilla Pudding & Pie Filling; Sliced Peaches	Organic Strawberries; Organic Avocado	Não acertou
66	Roman Raspberry Sorbetto; Soft Baked Chocolate Chip Cookies Gluten Free	Banana; Organic Strawberries	Não acertou
77	Orange Bell Pepper; Organic Tomato Basil Pasta Sauce	Orange Bell Pepper; Banana	**Acertou**
87	Chopped Green Chiles; Clear Tone Pink Rosa Antiperspirant Deodorant	Banana; Bag of Organic Bananas	Não acertou
		Taxa de acerto	20%

Ao analisar a Tabela 16, observa-se que o sistema de recomendação de produtos, baseado em regras de associação com métricas de confiança e suporte de valores altos, apresenta uma taxa de acerto de 20%.

Carrinhos agrupados por corredor

Uma vez que o presente conjunto de dados engloba uma vasta gama de produtos, a fim de realizar uma análise semelhante, porém com uma menor granularidade, os carrinhos de compras foram agrupados por corredor.

No primeiro carrinho de compras agrupados por corredores existem 6 corredores, nomeadamente *bread*, *fresh vegetables*, *packaged vegetables fruits*, *poultry counter*, *soy lactosefree* e *yogurt*. O corredor *yogurt* foi retirado de forma aleatória com o intuito de testar se o sistema de recomendação é capaz de recomendar este produto.

De acordo com a Figura 33, os corredores recomendados pelo sistema de recomendação, com base em valores mais altos de confiança e suporte, foram *fresh fruits* e *packaged cheese*, sendo que o sistema de recomendação não conseguiu acertar no corredor *yogurt*.

Figura 33- Sistema de recomendação de produtos baseado em regras de associação (corredores de produtos)

A seguir, apresenta-se a Tabela 17, que resume se o sistema de recomendação acertou ou não nos produtos para os demais carrinhos de compras. Além disso, no final da tabela, é calculada a taxa de acerto do sistema de recomendação de produtos para carrinhos de compras agrupados por corredores.

Tabela 17- Taxa de acerto do sistema de recomendação de produtos (corredores) | Confiança e Suporte

Carrinho de compras (order_id)	**Artigos a sugerir**	**Artigo(s) sugerido(s)**	**Acertou/Não acertou**
3	yogurt	fresh fruits; packaged cheese	Não acertou
5	dry pasta; milk	yogurt; milk	**Acertou**
11	fresh dips tapenades; frozen meals	fresh vegetables; fresh fruits	Não acertou
20	energy sports drinks; fresh vegetables	fresh vegetables; packaged vegetables fruits	**Acertou**
21	coffee; packaged cheese	fresh vegetables; packaged vegetables fruits	Não acertou
51	fresh herbs; fresh vegetables	fresh fruits; fresh vegetables	**Acertou**
53	fruit vegetable snacks; packaged cheese	fresh vegetables; packaged cheese	**Acertou**
64	milk; paper goods	fresh fruits; fresh vegetables	Não acertou
74	condiments; hot dogs bacon sausage	fresh fruits; packaged vegetables fruits	Não acertou
76	fresh vegetables; packaged vegetables fruits	fresh vegetables; packaged vegetables fruits	**Acertou**

Carrinho de compras (order_id)	Artigos a sugerir	Artigo(s) sugerido(s)	Acertou/Não acertou
		Taxa de acerto	50%

Ao analisar a Tabela 17, é possível observar que o sistema de recomendação de produtos, ao recomendar produtos com base nos corredores, apresenta uma taxa de acerto de 50%.

3.4.2.2. Sistemas de Recomendação com métricas de *Lift* e Suporte de valores altos

Os estudos posteriores adotam a mesma ordem de investigação. Irá ser fornecida uma explicação detalhada sobre a lógica aplicada ao primeiro carrinho de compras, enquanto para os carrinhos subsequentes, irá ser apresentada uma tabela resumindo se o sistema de recomendação teve sucesso ou falhou em pelo menos um dos produtos, com base em métricas de confiança e suporte de valores alto.

Carrinhos agrupados por nome do produto

No primeiro carrinho de compras analisado, que possui o ID do pedido igual a 3, foram selecionados e removidos os produtos *Air Chilled Organic Boneless Skinless Chicken Breasts* e *Organic Ezekiel 49 Bread Cinnamon Raisin* para testar a capacidade do sistema de recomendação em sugerir esses mesmos produtos.

De acordo com a primeira linha da tabela 18, os produtos recomendados pelo sistema de recomendação foram *Total 2% Lowfat Greek Strained Yogurt With Blueberry* e *Total 2% Lowfat Greek Strained Yogurt with Peach* (iogurte grego magro total 2% com pêssego), indicando que o sistema não conseguiu acertar em nenhum dos dois produtos que foram inicialmente removidos do carrinho de compras.

Tabela 18- Taxa de acerto do sistema de recomendação de produtos (nome de produtos) | Lift e Suporte

Carrinho de compras (order_id)	Artigos a sugerir	Artigo(s) sugerido(s)	Acertou/Não acertou
3	Air Chilled Organic Boneless Skinless Chicken Breasts; Organic Ezekiel 49 Bread Cinnamon Raisin	Total 2% Lowfat Greek Strained Yogurt With Blueberry; Total 2% Lowfat Greek Strained Yogurt with Peach	Não acertou
10	Spinach Peas & Pear Stage 2 Baby Food; Yellow Onions	Organic Jalapeno Pepper; Organic Garbanzo Beans	Não acertou
21	Frozen Organic Blueberries; Organic Firm Tofu	Não foram encontrados subconjuntos para este carrinho de compras	-
26	Banana; Boneless Skinless Chicken Breasts	Organic Grape Tomatoes; Large Lemon	Não acertou

Carrinho de compras (order_id)	Artigos a sugerir	Artigo(s) sugerido(s)	Acertou/Não acertou
48	Low Sodium Beef Broth; Organic Unrefined Toasted Sesame Oil	Organic Strawberries; Bag of Organic Bananas	Não acertou
53	Organic Blueberries; Organic Hass Avocado	Organic D'Anjou Pears; Organic Ginger Root	Não acertou
54	Instant French Vanilla Pudding & Pie Filling; Sliced Peaches	Organic Fuji Apple; Bartlett Pears	Não acertou
66	Roman Raspberry Sorbetto; Soft Baked Chocolate Chip Cookies Gluten Free	Organic Strawberries; Organic Avocado	Não acertou
77	Orange Bell Pepper; Organic Tomato Basil Pasta Sauce	Red Peppers Yellow Bell Pepper	Não acertou
87	Chopped Green Chiles; Clear Tone Pink Rosa Antiperspirant Deodorant	Blackberries Packaged Grape Tomatoes	Não acertou
		Taxa de acerto	0%

Ao analisar a Tabela 18, fica evidente que o sistema de recomendação de produtos, baseado em regras de associação com métricas de *lift* e suporte de valores altos, não obteve nenhum acerto, resultando em uma taxa de acerto de 0%. Ao fazer uma comparação com o estudo anterior, verifica-se que o sistema de recomendação de produtos, que utiliza valores altos de confiança e suporte, alcançou uma taxa de acerto superior, demonstrando ser mais eficaz na previsão dos produtos corretos.

Carrinhos agrupados por corredor

No primeiro carrinho de compras agrupados por corredores, com o ID de pedido igual a 3, originalmente havia 6 corredores: *bread*, *fresh vegetables*, *packaged vegetables fruits*, *poultry counter*, *soy lactose-free* e *yogurt*. No estudo anterior, o corredor *yogurt* foi removido aleatoriamente, portanto, neste estudo também será o item que será removido para testar a eficácia do sistema de recomendação.

De acordo com a primeira linha da tabela 19, os produtos recomendados pelo sistema de recomendação foram *tofu meat alternatives* (alternativas de carne de tofu) e *meat counter* (balcão de carnes), indicando que o sistema não conseguiu acertar no produto que inicialmente foi removido do carrinho de compras.

Tabela 19- Taxa de acerto do sistema de recomendação de produtos (corredores) | Lift e Suporte

Carrinho de compras (order_id)	**Artigos a sugerir**	**Artigo(s) sugerido(s)**	**Acertou/Não acertou**
3	yogurt	tofu meat alternatives; meat counter	Não acertou
5	dry pasta; milk	hair care; cleaning products	Não acertou
11	fresh dips tapenades; frozen meals	latino foods; canned jarred vegetables	Não acertou
20	energy sports drinks; fresh vegetables	fruit vegetable snacks; canned fruit applesauce	Não acertou
21	coffee; packaged cheese	frozen vegan vegetarian; asian foods	Não acertou
51	fresh herbs; fresh vegetables	Butter; dry pasta	Não acertou
53	fruit vegetable snacks; packaged cheese	frozen vegan vegetarian; canned meals beans	Não acertou
64	milk; paper goods	paper goods; dog food care	**Acertou**
74	condiments; hot dogs bacon sausage	ice cream toppings; canned jarred vegetables	Não acertou
76	fresh vegetables; packaged vegetables fruits	preserved dips spreads; fruit vegetable snacks	Não acertou
		Taxa de acerto	10%

Ao analisar a tabela 19, verifica-se que o sistema de recomendação de produtos (corredores), que se baseia em regras de associação de métricas de *lift* e suporte de valores altos, apresenta uma taxa de acerto de 10%, tendo somente acertado em um caso em estudo. Comparando a eficácia do sistema de recomendação atual com o anterior (conforme evidenciado na tabela 14), torna-se evidente que o sistema anterior, com uma taxa de acerto de 50%, é superior, o que indica a sua capacidade de fornecer melhores resultados.

Tendo por base as informações fornecidas ao longo do capítulo 3.4.2, é possível retirar algumas conclusões:

- Para os carrinhos de compras agrupados por nome de produto: O sistema de recomendação que utiliza métricas de confiança e suporte de valores altos apresentou uma taxa de acerto de 20%. Já o sistema de recomendação baseado em regras de associação com métricas de *lift* e suporte de

valores altos não obteve nenhum acerto, pelo que indica que o sistema de recomendação baseado em métricas de confiança e suporte é mais eficaz na previsão correta dos produtos;

- Para os carrinhos de compras agrupados por corredores: O sistema de recomendação de métricas de confiança e suporte de valores altos alcançou uma taxa de acerto de 50%. Por outro lado, o sistema baseado em regras de associação de *lift* e suporte de valores altos obteve uma taxa de acerto de 10%. Novamente, o sistema de recomendação de métricas de confiança e valores altos demonstra-se superior, e, portanto, mais eficaz na correta previsão dos produtos.

Portanto, com base nos resultados obtidos (por nome de produto e por corredor), o sistema de recomendação baseado em regras de confiança e suporte de valores altos demonstrou resultados superiores, com taxas de acerto de 20% e 50%, pelo que é recomendado que este sistema de recomendação seja selecionado para entrar em produção.

3.4.2.3. Comprovação da eficácia do sistema de recomendação para produtos frequentes

O objetivo do seguinte estudo é comprovar a eficácia do sistema de recomendação ao sugerir produtos frequentemente comprados. Com base nos estudos anteriores, observou-se que a maioria dos casos com acertos ocorreu em produtos com alta frequência de compra. Portanto, este estudo utilizará carrinhos agrupados por nome de produto com os números 5, 10, 27, 53 e 88 e será realizado com base no sistema de recomendação de métricas de confiança e suporte de valores altos (tabela 20), dado que demonstrou ser o sistema mais eficaz no estudo anterior. A partir desses carrinhos, os produtos mais frequentes serão removidos para testar a eficácia do sistema de recomendação nessas condições.

Tabela 20- Comprovação da eficácia do sistema de recomendação para produtos frequentes

Carrinho de compras (order_id)	**Artigos a sugerir**	**Artigo(s) sugerido(s)**	**Acertou/Não acertou**
5	Bag of Organic Bananas; Organic Hass Avocado	Bag of Organic Bananas; Organic Strawberries	**Acertou**
10	Banana; Organic Strawberries	Limes; Banana	**Acertou**
27	Bag of Organic Bananas; Organic Avocado	Bag of Organic Bananas; Banana	**Acertou**
53	Banana; Organic Hass Avocado	Banana; Organic Baby Spinach	**Acertou**
88	Banana; Organic Avocado	Banana	**Acertou**
		Taxa de acerto	100%

Ao analisar a tabela 20, verifica-se que a taxa de acerto é de 100% no caso de produtos frequentemente comprados, o que comprova a eficácia do sistema de recomendação de produtos para produtos frequentes.

3.5. Modelos de previsão para prever a recompra de um determinado produto

No âmbito deste capítulo serão empregues modelos de previsão para determinar se um cliente específico irá recomprar determinado produto. Esta fase, conhecida por modelação, segundo a metodologia CRISP-DM, é a etapa onde diferentes modelos são selecionados e aplicados ao conjunto de dados em estudo (Kelleher et al., 2015).

Para além disso, neste capítulo, serão abordadas as manipulações necessárias para a aplicação dos diversos modelos, assim como a discussão dos resultados obtidos. Será realizada uma avaliação da eficácia dos modelos utilizando métricas como *accuracy, precision, recall e* F1-*score*. O foco principal será a métrica de *accuracy*, uma vez que foi selecionada para avaliar o desempenho dos modelos.

Neste presente estudo, os modelos serão utilizados de modo a prever uma variável categórica, designada por "reordered". Esta variável possui dois valores possíveis, 0 ou 1, sendo que 0 representa que o cliente não recomprou o produto em questão e 1, indica que o produto foi comprado novamente.

3.5.1. Configuração das experiências

Neste subcapítulo serão introduzidos os diversos modelos utilizados, juntamente com a apresentação dos parâmetros e respetiva gama de valores utilizada para conduzir a análise de sensibilidade dos parâmetros. Para a configuração e realização das diferentes experiências, a ferramenta *knime* foi utilizada.

Em todos os modelos, a métrica *accuracy* foi selecionada para ser maximizada e, portanto, será a métrica utilizada para avaliar o desempenho dos diferentes modelos.

3.5.1.1. *K-Nearest Neighbors*

O primeiro modelo utilizado para prever a recompra de um determinado produto foi o modelo K-Nearest Neighbors (KNN). O modelo KNN é um modelo amplamente utilizado para modelos de classificação e regressão (Witten et al., 2011).

Dado que o modelo KNN, para prever um novo ponto de dado, procura pela distância dos "k" vizinhos mais próximos, neste caso em estudo, a partir da distância euclidiana, é necessário efetuar a normalização dos dados, dado que a distância entre os diversos pontos de dados pode ser distorcida por diferenças nas escalas dimensionais. Além disso, o desempenho pode ser afetado pela dimensionalidade elevada, pois o cálculo das distâncias pode ser computacionalmente exigente.

De forma a realizar uma análise de sensibilidade ao parâmetro "k", neste presente estudo, foram testados os valores de 3 a 10, com um incremento de um a um.

Para além disso, em todos os estudos realizados nesta dissertação, a estratégia de pesquisa utilizada neste nó foi *brute force* em que todas as combinações de parâmetros possíveis são verificadas e a melhor é devolvida (Berthold et al., 2007).

Para todos os modelos em estudo, foi realizado validação cruzada com um total de 10 iterações de validação cruzada. A divisão das amostras foi feita de forma estratificada, garantindo que as partições fossem selecionadas aleatoriamente, mas mantendo a distribuição das classes da coluna "reordered".

O fluxo que permitiu a construção deste modelo encontra-se representado no apêndice D.

3.5.1.2. Árvore de decisão

O segundo modelo utilizado para prever a recompra de um determinado produto foi o modelo árvore de decisão.

O modelo de árvore de decisão não requer a normalização dos dados, dado que este modelo realiza divisões nos dados com base em testes de atributos individuais, e, portanto, não depende da escala absoluta dos valores. Para além disso, modelos baseados em árvores de decisão não são dependentes da distância onde as caraterísticas têm efeito uma sobre a outra (Arya, 2022).

De forma a medir a qualidade de divisão da árvore de decisão, recorreu-se ao uso do método *Gini Index* que calcula a probabilidade de um atributo específico ser classificado incorretamente quando é feita uma seleção aleatória. A escala de Gini Index varia entre 0-0,5, em que o valor mínimo de 0 é o melhor valor e 0,5 o pior valor possível (Arya, 2022).

O método utilizado para realizar a poda foi o “Minimal Description Length” (MDL), de forma a reduzir o tamanho da árvore, evitando o *overfitting* e, por conseguinte, aumentando a qualidade da previsão (Berthold et al., 2007). O método MDL procura encontrar o equilíbrio entre a complexidade do modelo e a capacidade de descrever os dados. Além do mais, ressalta que a melhor descrição dos dados é fornecida pelo modelo que os comprime da melhor forma possível (Barron & Cover, 1991).

Na etapa de pós-processamento um método simples de poda é efetuado para cortar a árvore, começando pelas folhas, cada nó é substituído pela sua classe mais popular, porém somente se a precisão da previsão não diminuir (Berthold et al., 2007).

O valor de divisão para atributos numéricos é determinado com o valor médio dos dois valores do atributo que separam as duas partições (Berthold et al., 2007). Os atributos nominais são divididos de forma binária, em dois subconjuntos (um para cada filho), sendo que as divisões binárias são mais difíceis de calcular, porém resultam em árvores mais precisas (Berthold et al., 2007). Dada a dificuldade de calcular as divisões binárias, é definido um número máximo de valores nominais para qual todos os subconjuntos possíveis são calculados. Acima desse limite, é aplicada uma heurística que calcula primeiro o melhor valor nominal para a segunda partição, depois o segundo melhor valor, por aí adiante, até que nenhuma melhoria possa ser alcançada, assim esse cálculo não será tão dispendioso de ser efetuado (Berthold et al., 2007).

Se a pontuação alcançar um nó, no qual o valor do atributo é desconhecido, é aplicada uma estratégia de retornar a classe mais frequente do último nó. Para além disso, no caso de presença de valores em falta, é utilizada a última classe conhecida.

O parâmetro *minimum number of records*, que representa o número mínimo de registos necessários em cada nó, foi analisado e os valores de 2 a 15, com um incremento de um a um, foram testados. Se o número de registos for menor ou igual ao número selecionado, a árvore não cresce mais, correspondendo a um critério de paragem (pré-poda) (Berthold et al., 2007).

O fluxo que permitiu a construção deste modelo encontra-se representado no apêndice E.

3.5.1.3. *Random Forest*

O terceiro modelo utilizado para prever a recompra de um determinado produto foi o modelo *random forest.*

O modelo *random forest* como utiliza um conjunto de árvores de decisão independentes e, cada árvore é construída com uma amostra aleatória de dados que se baseiam em diferentes caraterísticas individuais (Breiman, 2001), trata-se de um modelo que não necessita de normalizar os dados.

O parâmetro que diz respeito à profundidade máxima das árvores foi testado com valores de um a quinze (incremento um a um). De forma a investigar o valor ideal para a profundidade máxima das árvores (*Maximum Depth*) foram utilizadas, inicialmente, 10 árvores de decisão para conduzir o estudo.

Com o propósito de analisar como o número de árvores de decisão afeta a precisão do modelo, realizou-se uma investigação adicional que envolveu a variação na quantidade de árvores de decisão utilizadas. Nesse estudo, a profundidade máxima das árvores de decisão foi fixada em 18, um valor previamente identificado como ideal para essa análise. No início, os valores 10, 50, 100 e 200 e 300 foram estabelecidos. Dado que o presente estudo é computacionalmente exigente, se uma melhoria na eficácia do modelo fosse observada, o número de árvores de decisão em análise seria ampliado para incluir novos valores, nomeadamente 500 e 600.O objetivo era determinar se o aumento no número de árvores de decisão, nesse contexto específico, impacta ou não a melhoria da eficácia do modelo.

O fluxo que permitiu a construção deste modelo encontra-se representado no apêndice F.

3.5.1.4. *AdaBoost*

O quarto modelo utilizado para prever a recompra de um determinado produto foi o modelo *AdaBoost* em que o modelo base selecionado (*weak learner*) foi o *DecisionStump.*

Por se tratar de uma árvore de decisão simples a normalização de dados não foi aplicada, dado que as árvores de decisão não são dependentes da distância onde as caraterísticas têm efeito uma sobre a outra (Arya, 2022).

De forma a realizar uma análise de sensibilidade ao parâmetro *number of iterations* os valores de três a doze foram testados com um incremento de três em três, sendo que o número de iterações especifica quantas vezes o algoritmo irá iterar para melhorar o modelo.

O algoritmo *AdaBoost* é utilizado para melhorar o desempenho dos modelos de *Machine Learning* mais fracos (geralmente conhecidos por “weak learners”) e transformá-los num modelo forte (Zhang & Ma, 2012).

O fluxo que permitiu a construção deste modelo encontra-se representado no apêndice G.

3.5.1.5. *Multilayer perceptrons*

O quinto modelo utilizado para prever a recompra de um determinado produto foi o modelo *Multilayer perceptrons.*

A normalização dos dados foi aplicada de forma a garantir que as caraterísticas têm escalas comparáveis e para garantir que algumas caraterísticas não tenham dominância em relação às outras devido às diferenças nas unidades de medida. A normalização dos dados permite obter uma média próxima de zero e acelerar o processo de aprendizagem, levando a uma convergência mais rápida. Para além disso, a normalização dos dados garante que haja tantos valores positivos como valores negativos utilizados como entradas para a próxima camada, tornando assim a aprendizagem mais flexível. Para as redes neuronais, a normalização dos dados, é importante para garantir que a magnitude dos valores de entrada assumidos pelas caraterísticas de entrada são similares, dado que o processo de aprendizagem, para as redes neuronais, depende da magnitude dos valores de entrada (Stöttner, 2019).

O parâmetro *number of hidden layers* (número de camadas ocultas) e o parâmetro *number of hidden neurons per layer* (número de neurónios ocultos por camada) foram testados, a fim de descobrir qual par de valores se encaixa melhor no objetivo de otimizar a função objetivo. Os valores em estudo foram de um a cinco e de um a dez, com um incremento de um a um, respetivamente. A estratégia de pesquisa utilizada foi *brute force* em que todas as combinações de parâmetros possíveis são verificadas e a melhor é devolvida (Berthold et al., 2007). Para além disso, foi definido um número máximo de cinquenta iterações.

A rede neural utilizada aplica uma estratégia de *feedforward,* onde os dados de entrada só podem fluir numa direção e não podem ser reciclados de volta para a entrada (Berthold et al., 2007).

O fluxo que permitiu a construção deste modelo encontra-se representado no apêndice H.

3.5.1.6. Stacking

Por último, o modelo *Stacking* foi utilizado de forma a combinar os diferentes modelos previamente estudados. O modelo de árvore de decisão, KNN, *random forest* e *multilayer perceptron* foram utilizados como modelos de nível 1 e uma regressão logística binária foi utilizada como meta modelo de nível 2.

As configurações dos variados parâmetros nos algoritmos de primeiro nível foram definidas com base na análise prévia desses parâmetros. Assim, foram selecionados os parâmetros que demonstraram proporcionar a melhor precisão nos estudos anteriores.

O modelo de regressão logística binária, utilizada como meta modelo de nível 2, é uma técnica estatística utilizada para classificação e análise preditiva. É utilizada quando a variável dependente é dicotómica ou binária (Landwehr, 2005). Para a classificação binária, uma probabilidade menor que 0,5 prevê 0, caso contrário se prevê 1 (IBM, 2023c). Este modelo tem como objetivo modelar as probabilidades de as instâncias pertencerem a classes diferentes, permitindo a classificação de novas instâncias, selecionando a classe com a estimativa de probabilidade mais elevada (Landwehr, 2005). A estimativa dos parâmetros do modelo é feita utilizando o critério de máxima

verossimilhança, ajustando os valores para obter a melhor correspondência aos dados. Após isso, as probabilidades previstas são calculadas e usadas para prever resultados binários (IBM, 2023c). O algoritmo *LogitBoost* é utilizado para melhorar o ajuste e selecionar os atributos mais relevantes para o modelo (Landwehr, 2005).

No que diz respeito à estrutura do fluxo desenvolvido, inicialmente foi efetuada uma partição que segmenta o conjunto de dados em um conjunto de treino e um conjunto de teste. Nesse processo, 80% dos dados são alocados para o conjunto de treino, enquanto os restantes 20% compõem o conjunto de teste. É importante salientar que tal divisão foi efetuada por meio de uma abordagem estratificada, com a variável "reordered" sendo selecionada para assegurar a sua representação equilibrada na amostra

O fluxo que permitiu a construção deste modelo encontra-se representado no apêndice I.

3.5.2. Apresentação dos resultados

Os resultados alcançados para os diversos parâmetros analisados em cada modelo, encontram-se detalhados no apêndice J. Além disso, a tabela 21 a seguir resume as melhores precisões obtidas para os diferentes modelos utilizados.

Tabela 21- Apresentação dos melhores resultados da métrica "precisão" por modelo (%)

Data set	**KNN**	**Árvore de Decisão**	***Random Forest***	***AdaBoost***	**MLP**	***Stacking***
InstaCart Market Basket Analysis	66,8%	70%	71,5%	66,2%	68,7%	75%

Através da análise da tabela 21, é possível auferir que o modelo de *stacking* apresenta um desempenho superior em relação aos demais, atingindo uma precisão de 75%. O sucesso desse modelo pode ser atribuído à sua capacidade de combinar diversos tipos de modelos, cada um com suas próprias virtudes e limitações. Essa abordagem permite que o modelo aprenda de diversas maneiras, o que resulta, geralmente numa melhoria de precisão. Além disso, o modelo de *stacking* incorpora um segundo nível de aprendizagem, no qual um meta-modelo é treinado com base nas previsões geradas pelos modelos base. Essa estratégia aprimora a capacidade do meta-modelo em prever novos pontos de dados com eficácia, dado que o mesmo aprende com as contribuições individuais dos modelos base treinados no nível 1.

4. CONCLUSÃO

Neste presente capítulo serão apresentadas as principais conclusões derivadas da pesquisa realizada nesta dissertação. Para além disso, serão abordadas as limitações identificadas e delineadas perspetivas para desenvolvimentos futuros, com o propósito de investigar vias alternativas que não foram exploradas neste estudo.

4.1. Conclusões finais

Esta dissertação envolveu a realização de uma pesquisa que se concentrou na aplicação de técnicas *Machine Learning* no setor do retalho. O objetivo foi analisar padrões de consumo, criar sistema de recomendação de produtos e desenvolver modelos de previsão. Uma vez que não existiam empresas dispostas a partilhar os seus dados, optou-se por utilizar o conjunto de dados público do *InstaCart,* que regista transações de produtos alimentares.

O conjunto de dados *InstaCart Market Basket Analysis* foi analisado e trabalhado através da ferramenta *Knime* com o intuito de desenvolver três estudos: a investigação de padrões de consumo por meio da aplicação de regras de associação, o desenvolvimento de um sistema de recomendação de produtos com base nas regras de associação identificadas anteriormente e a aplicação de diferentes modelos de previsão para determinar se um cliente específico irá recomprar determinado produto.

A abordagem aplicada respeita as fases da metodologia CRISP-DM, o que permitiu conduzir uma análise mais estruturada e objetiva ao definir um processo sistemático para o desenvolvimento do presente estudo.

A compreensão do negócio em questão foi fundamental para a estruturação e realização dos três estudos. Nesta fase, foi possível conhecer o problema empresarial em questão e qual a motivação que poderá levar a um retalhista a aplicar este desenvolvimento em contexto real. No contexto de um retalhista que implemente um sistema de recomendação de produtos, poderá destacar-se em relação aos demais, pelo facto de demonstrar uma maior atenção aos detalhes e ao oferecer um serviço mais personalizado aos clientes. Isso, por sua vez, resulta em uma vantagem competitiva.

O primeiro estudo, referente à investigação de padrões de consumo por meio da aplicação de regras de associação, foi realizado com dois tipos de granularidade distintas. Numa primeira abordagem foram estudados os produtos na sua forma mais granular possível, porém dado que o conjunto de dados contém uma enorme variedade de produtos, os valores de suporte relevaram-se ser relativamente baixos, visto que a proporção de produtos por transação é pequena. Posto isto, numa segunda abordagem foi conduzido o mesmo raciocínio lógico, porém num nível de hierarquia mais elevado, por corredor.

No decorrer do primeiro estudo, diversas abordagens foram empregues, incluindo uma variedade de níveis de suporte, confiança e *lift* distintos. As manipulações, com diferentes ajustes realizadas nos limites mínimos de suporte proporcionou informações valiosas sobre as potenciais estratégias adotadas pelos retalhistas. De forma resumida, foi possível observar uma forte associação entre produtos orgânicos, como morangos orgânicos tendem a ser comprados juntamentos com bananas orgânicas. Isso sugere que a implementação de estratégias promocionais que envolvem a inclusão

de produtos orgânicos impulsionará a compra de bananas orgânicas. Para além disso, produtos de água com gás, como *Lime Sparkling Water* (água com gás de lima) e *Sparkling Lemon Water* (água com gás de limão), estão fortemente associados a *Sparkling Water Grapefruit* (água com gás de toranja), o que sugere a oportunidade de implementar estratégias de promoção cruzada. Também foram identificadas fortes associações entre os produtos *Zero Calorie Cola* (coca-cola zero) e *Soda* (refrigerante). Foi observada uma alta probabilidade (0,417) de quem compra *Soda* também adquirir *Zero Calorie Cola*, porém, quem adquire *Zero Calorie Cola* não tem uma alta probabilidade (0,099) de adquirir *Soda*. Isso sugere que, do ponto de vista promocional, seria interessante efetuar uma promoção no produto "Soda" uma vez que iria atrair os clientes a comprar o produto *Zero Calorie Cola*. Por outro lado, aplicar a promoção no produto *Zero Calorie Cola* não seria tão vantajoso, pois há um maior risco do cliente adquirir apenas esse produto.

Outra descoberta relevante foi a associação frequente encontrada entre os produtos dos corredores *fresh vegetables* (vegetais frescos), *fresh fruits* (frutas frescas*)* e *packaged vegetables fruits* (vegetais e frutas embaladas). Essa informação pode ser utilizada para otimizar o *layout* da loja e criar promoções que incentivem compras conjuntas nesses corredores.

Este tipo de análise permite identificar que as regras de associação proporcionam informações relevantes para melhorar as estratégias de vendas, promoções e posicionamento dos produtos.

No segundo estudo, foi efetuada uma investigação sobre a eficácia da implementação de um sistema de recomendação de produtos, baseado em regras de associação, considerando as métricas de suporte, confiança e *lift*. O estudo foi dividido em dois conjuntos: o primeiro conjunto utilizou regras de associação com altos níveis de suporte e confiança e o outro conjunto era constituído por regras de associação com altos níveis de suporte e *lift*. Uma vez mais, foram analisados carrinhos de compras agrupados por produtos e por corredor.

Os resultados revelaram que o sistema de recomendação de produtos baseado em métricas de suporte e confiança elevados é mais eficaz na correta previsão dos produtos. O sistema de recomendação de produtos baseado em métricas de suporte e confiança elevados apresentou uma taxa de acerto de 20% para carrinhos de compras agrupados por produto e 50% para carrinhos agrupados por corredor. Em contrapartida, o sistema baseado em métricas de suporte e *lift* elevados obteve uma taxa de acerto de 0% para carrinhos agrupados por produtos e 10% para carrinhos agrupados por corredor. Portanto, é recomendado que o sistema baseado em métricas de suporte e confiança elevados seja selecionado para implementação.

Para além disso, o sistema de recomendação de produtos revelou ser 100% eficaz para produtos frequentemente comprados, o que destacou a importância das métricas de confiança e suporte altos para melhorar a precisão dos sistemas de recomendação de produtos, especialmente para produtos frequentemente adquiridos. Estas conclusões poderão ser benéficas para empresas que pretendem implementar sistemas de recomendação de produtos eficazes, capazes de impulsionar as suas vendas e aumentar a satisfação dos clientes.

Por fim, no último estudo realizado, foram utilizados diferentes modelos de previsão, de forma a prever a recompra de um determinado produto. Neste presente estudo, o objetivo era prever uma variável categórica designada de "reordered" que se trata de uma variável booleana, que representa se o cliente recompra ou não o produto.

As métricas de *accuracy, precision, recall* e F1-*score* foram consideradas para avaliar o desempenho dos modelos, porém a métrica *accuracy* foi a principal métrica utilizada para avaliar o respetivo desempenho, uma vez que se trata da métrica selecionada para maximizar a função objetivo em todos os modelos.

Os modelos utilizados neste estudo foram o KNN, árvore de decisão, *random forest*, MLP, *adaboost* e *stacking.* O melhor desempenho foi alcançado pelo modelo *stacking*, com uma precisão de 75%, o que representa um aumento de 3,5 pontos percentuais em relação ao segundo classificado, que foi o modelo de *random forest* com uma precisão de 71,5%. O pior desempenho foi obtido pelo modelo *Adaboost*, com uma precisão de 66,2%.

4.2. Limitações e investigação futura

Dado à limitação computacional do dispositivo utilizado para processar os dados, foi necessário reduzir um dos conjuntos de dados que compõem o conjunto *InstaCart Market Basket Analysis* para 1500000 exemplos, sendo que originalmente contém 32434489 exemplos.

Do ponto de vista de investigação futura, para aprimorar a descoberta de regras de associação, seria interessante segmentar os clientes em *clusters* com gostos semelhantes. Esse nível de granularidade permitiria identificar padrões potencialmente mais específicos dentro de cada grupo. As regras de associação encontradas poderiam ser mais precisas e direcionadas, dado que os clusters resultariam em grupos de clientes com interesses comuns, o que por sua vez iria permitir efetuar recomendações mais personalizadas, atendendo melhor às necessidades reais de cada segmento de clientes. Para além disso, a análise do comportamento de compra em grupos distintos proporcionaria uma melhor compreensão a esse nível, fornecendo conhecimentos valiosos que poderiam orientar as decisões estratégicas de uma forma mais informada. Oportunidades de nicho poderiam ser mais facilmente identificadas por uma agrupação em clusters, o que resultaria em recomendações mais personalizadas e direcionadas para este tipo de clientes.

Para o último estudo realizado, com o intuito de prever a recompra de determinado produto, a aplicação de mais modelos de previsão poderia potencializar os resultados, conduzindo à descoberta de modelos alternativos com desempenho superior que, posteriormente poderiam ser incluídos no modelo *stacking* resultando em um aumento da precisão final do modelo.

Para além disso, a aplicação do processo de seleção de variáveis, conhecido como “seleção de caraterísticas” (*feature selection*), poderia ser utilizado para escolher as variáveis ou caraterísticas mais relevantes do conjunto de dados, resultando num melhor desempenho dos modelos de previsão.

REFERÊNCIAS BIBLIOGRÁFICAS

Adomavicius, G., & Tuzhilin, A. (2005). *Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions*.

Agrawal, R., & Srikant, R. (1994). *Fast Algorithms for Mining Association Rules*.

Alfiqra, & Khasanah, A. U. (2020). Implementation of Market Basket Analysis based on Overall Variability of Association Rule (OCVR) on Product Marketing Strategy. *IOP Conference Series: Materials Science and Engineering*, *722*(1). https://doi.org/10.1088/1757-899X/722/1/012068

Aljamaan, H., & Elish, M. (2009). An Empirical Study of Bagging and Boosting Ensembles for Identifying Faulty Classes in Object-Oriented Softwar. *IEEE Xplore*.

Alpaydin, E. (2014). *Introduction to Machine Learning* (Third Edition). The MIT Press.

Arya, N. (2022, July 25). *Does the Random Forest Algorithm Need Normalization?* https://www.kdnuggets.com/2022/07/random-forest-algorithm-need-normalization.html

Azri, T. (2020). *Recommendation System for Retail Customer (A Hands-On Example Using Python TensorRec Module)*. https://taufik-azri.medium.com/recommendation-system-for-retail-customer-3f0f80b84221

Barron, A., & Cover, T. (1991). Minimum Complexity Density Estimation. *IEEE TRANSACTIONS ON INFORMATION THEORY*, *37*, 1034–1053.

Berman, B., Evans, J. R., & Chatterjee, P. (2018). *Retail management: A strategic approach*.

Berry, M. J., Linoff, G. S., & Marketing, F. (2004). *Customer Relationship Management Second Edition Data Mining Techniques* (Second Edition).

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Meinl, T., Ohl, P., Thiel, K., & Wiswedel, B. (2007). *KNIME:The Konstanz Information Miner, in: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer. http://labs.knime.org

Birant, D. (2011). *Data Mining Using RFM Analysis*. https://doi.org/10.5772/13683

Blattberg, R. C., Kim, B.-D., & Neslin, S. A. (2008). *Database Marketing: Analyzing and Managing Customers*. Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-0-387-72579-6

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bult, J. R., & Wansbeek, T. (1995). *Optimal Selection for Direct Mail*. Marketing Science 14(4):378-394.

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

Chauhan, N. (2020, May 28). *Model Evaluation Metrics in Machine Learning*. https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html

Chauhan, N. S. (2019, September 27). *What is Hierarchical Clustering?* KDnuggets. https://www.kdnuggets.com/2019/09/hierarchical-clustering.html

Cios, K. J., Swiniarski, R. W., Pedrycz, W., & Kurgan, L. A. (2007). *Unsupervised Learning: Association Rules*. Springer, Boston, MA.

Com, M. (2019). *Research Methodology (study material)*.

Coutinho, C. P. (2011). *Metodologia de Investigação em Ciências Sociais e Humanas: Teoria e Prática* (2ª Edição). Almedina.

Cunningham, P., & Delany, S. J. (2020). *k-Nearest Neighbour Classifiers: 2nd Edition (with Python examples)*. https://doi.org/10.1145/3459665

Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning*. http://www.cs.orst.edu/~tgd

Fernandes, J., Machado, C. F., & Amaral, L. (2020). *Methodology Used for Determination of Critical Success Factors in Adopting the New General Data Protection Regulation in Higher Education Institutions*. Springer, Cham. http://www.springer.com/series/11690

Flask. (2023). https://flask.palletsprojects.com/en/2.2.x/

Gardner, M. W., & Dorling, S. R. (1998). Artificial Neural Networks (The Multilayer Perceptron)- A Review of Applications in the Atmospheric Sciences. In *Atmospheric Environment* (Vol. 32, Issue 14).

Géron, A. (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems*. http://oreilly.com/safari

Ghahramani, Z. (2004). *Unsupervised Learning*. Springer. http://www.gatsby.ucl.ac.uk/~zoubin

Gurudath, S. (2020). *Market Basket Analysis & Recommendation System Using Association Rules*.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques* (Third Edition). Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. (2009a). *Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction*.

Hastie, T., Tibshirani, R., & Friedman, J. (2009b). *The Elements of Statistical Learning (Data Mining, Inference, and Prediction)* (Second Edition). Springer.

Hirate, Y., Iwahashi, E., & Yamana, H. (2004). *TF2P-growth: An Efficient Algorithm for Mining Frequent Patterns without any Thresholds*.

IBM. (2023a). *What is Bagging?* IBM.Com. https://www.ibm.com/topics/bagging

IBM. (2023b). *What is boosting?* IBM.Com. https://www.ibm.com/topics/boosting

IBM. (2023c). *What is logistic regression?* IBM.Com. https://www.ibm.com/topics/logistic-regression

IBM Corp. (2023). *What is unsupervised learning?* IBM. https://www.ibm.com/topics/unsupervised-learning

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. In *Egyptian Informatics Journal* (Vol. 16, Issue 3, pp. 261–273). Elsevier B.V. https://doi.org/10.1016/j.eij.2015.06.005

Jooa, J., Bangb, S., & Parka, G. (2016). Implementation of a Recommendation System Using Association Rules and Collaborative Filtering. *Procedia Computer Science*, *91*, 944–952. https://doi.org/10.1016/j.procs.2016.07.115

Kaggle. (2023). https://www.kaggle.com/c/instacart-market-basket-analysis

Kalirane, M. (2023, January 20). *Ensemble Learning Methods: Bagging, Boosting and Stacking*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/

Kane, F. (2018). *Building Recommender Systems with Machine Learning and AI*.

Kégl, B. (2014). *Introduction to AdaBoost*.

Kelleher, J. D., mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. MIT Press.

Kothari, C. R. (2004). *Research methodology: Methods & Techniques*. New Age International (P) Ltd.

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

Landwehr, N. (2005). *Logistic Model Trees ** (Vol. 59). Springer Science + Business Media, Inc. Manufactured in The Netherlands.

Leo Breiman, by, Friedman, J. H., & Olshen, R. A. (1987). Classification and Regression Trees. In *Cytometry* (Vol. 8).

Liu, Q., & Wu, Y. (2012). Supervised Learning. In *Encyclopedia of the Sciences of Learning* (pp. 3243–3245). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_451

Lutins, E. (2017, August 2). *Ensemble Methods in Machine Learning: What are They and Why Use Them?* Towards Data Science. https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning second edition* (Second Edition). MIT press.

Murtagh, F. (1991). *Neurocomputing 2 (1990/1991) 183/197*.

NetworkX. (2023). https://networkx.org/

Nielsen, M. (2015). *Neural Networks and Deep Learning*. http://neuralnetworksanddeeplearning.com

Ozkan, P., & Kocakoc, I. D. (2021). *A customer segmentation model proposal for retailers: RFM-V*. https://digitalcommons.usf.edu/m3publishing/vol5/iss2021/104

Saghiri, S., Wilding, R., Mena, C., & Bourlakis, M. (2017). Toward a three-dimensional framework for omni-channel. *Journal of Business Research*, *77*, 53–67. https://doi.org/10.1016/j.jbusres.2017.03.025

Salian, I. (2018). *SuperVize Me: What's the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?* Nvidia. https://blogs.nvidia.com/blog/2018/08/02/supervised-unsupervised-learning/

Shin, T. (2020, July 24). *Ensemble Learning, Bagging, and Boosting Explained in 3 Minutes*. Towards Data Science. https://towardsdatascience.com/ensemble-learning-bagging-and-boosting-explained-in-3-minutes-2e6d2240ae21

Stöttner, T. (2019, May 16). *Why Data should be Normalized before Training a Neural Network*. Towards Data Science. https://towardsdatascience.com/why-data-should-be-normalized-before-training-a-neural-network-c626b7f66c7d

Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*.

Techlabs, M. (2017, October 5). *5 Advantages Recommendation Engines can Offer to Businesses*. Towards Data Science. https://towardsdatascience.com/5-advantages-recommendation-engines-can-offer-to-businesses-10b663977673

Techlabs, M. (2021). *Types of Recommendation Systems & Their Use Cases*. Medium. https://medium.com/mlearning-ai/what-are-the-types-of-recommendation-systems-3487cbafa7c9

Turban, E., Delen, D., & Sharda, R. (2013). *Business intelligence and Analytics : Systems for Decision Support* (Tenth Edition).

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edition).

Wu, J., & Lin, Z. (2005). *Research on Customer Segmentation Model by Clustering*. https://doi.org/https://doi.org/10.1145/1089551.1089610

Zhang, C., & Ma, Y. (2012). Ensemble Machine Learning. In *Ensemble Machine Learning*. Springer New York. https://doi.org/10.1007/978-1-4419-9326-7

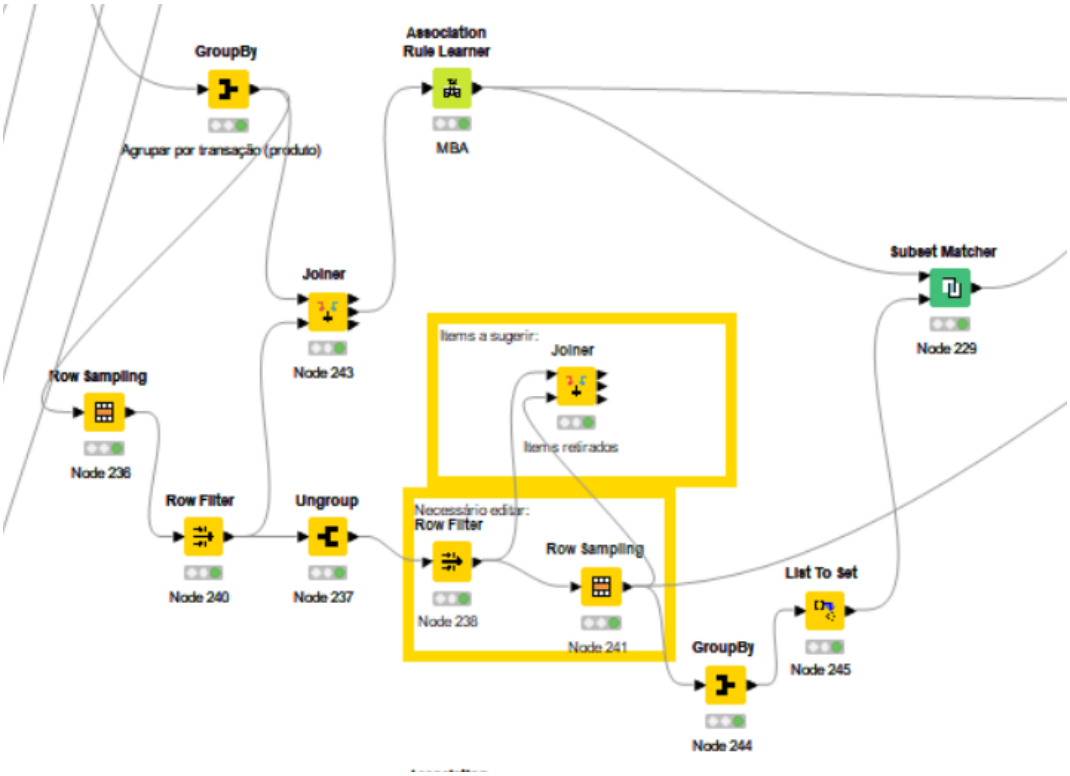
Zhang, C., & Zhang, S. (2002). *Association Rule Mining: Models and Algorithms*.

APÊNDICE A- ESTUDO ESTATÍSTICO PARA OS DIFERENTES ATRIBUTOS

S Column	D Min	D Max	D Mean	D Std. de...	D Variance	D Skewness	D Kurtosis	D Overall ...	I No. mis...	I No. NaNs	I No. +oos	I No. -oos	D Median	I Row co...	Histogram
add_to_cart_order	1	127	8.545	7.271	52.86	1.77	5.249	24,649,709	0	0	0	0	?	2884617	
order_number	1	100	17.207	17.18	295.142	1.967	4.354	49,634,288	0	0	0	0	?	2884617	
order_hour_of_day	0	23	13.49	4.243	18.001	-0.077	0.005	38,914,096	0	0	0	0	?	2884617	
days_since_prior...	0	30	14.039	10.072	101.452	0.535	-1.209	39,145,159	96268	0	0	0	?	2884617	

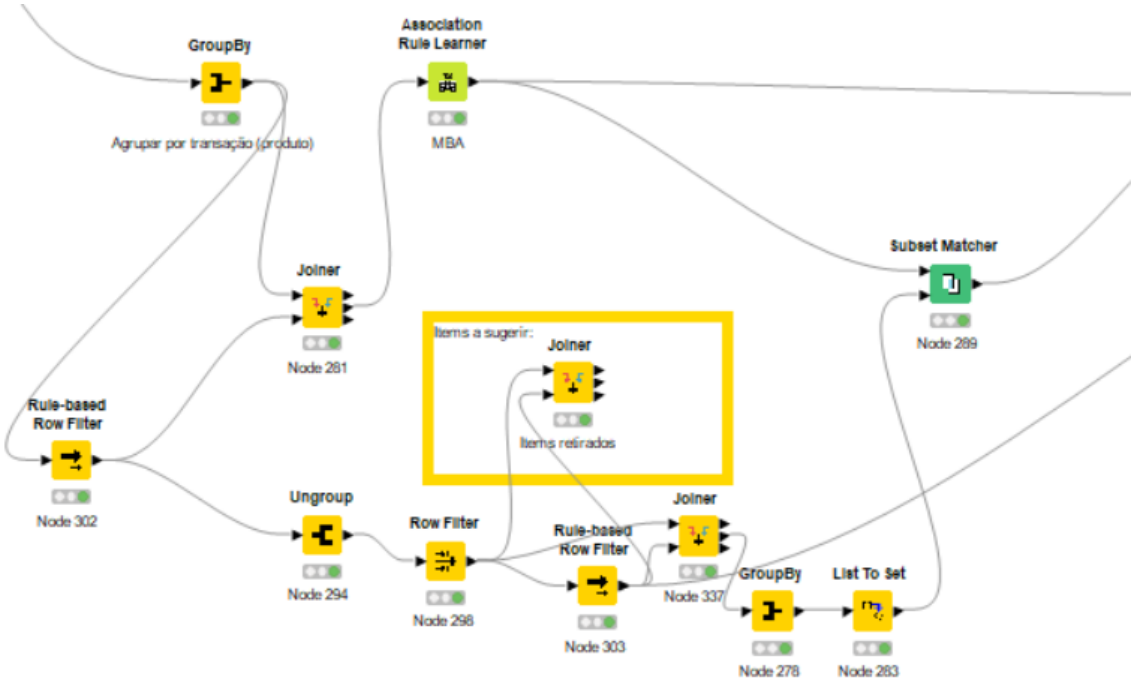
Apêndice A- Estudo estatístico dos diferentes atributos

APÊNDICE B- ESTRUTURA DO SISTEMA DE RECOMENDAÇÃO DE PRODUTOS (CONFIANÇA E SUPORTE)



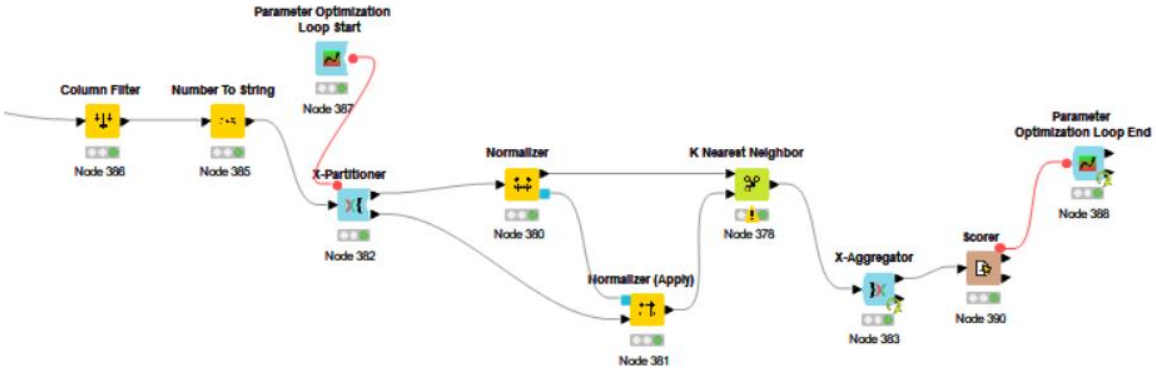
Apêndice B- Estrutura de nós para realizar o estudo referente ao desenvolvimento do sistema de recomendação avaliado com base nas métricas de confiança e suporte de valores altos

APÊNDICE C- ESTRUTURA DO SISTEMA DE RECOMENDAÇÃO DE PRODUTOS (*LIFT* E SUPORTE)



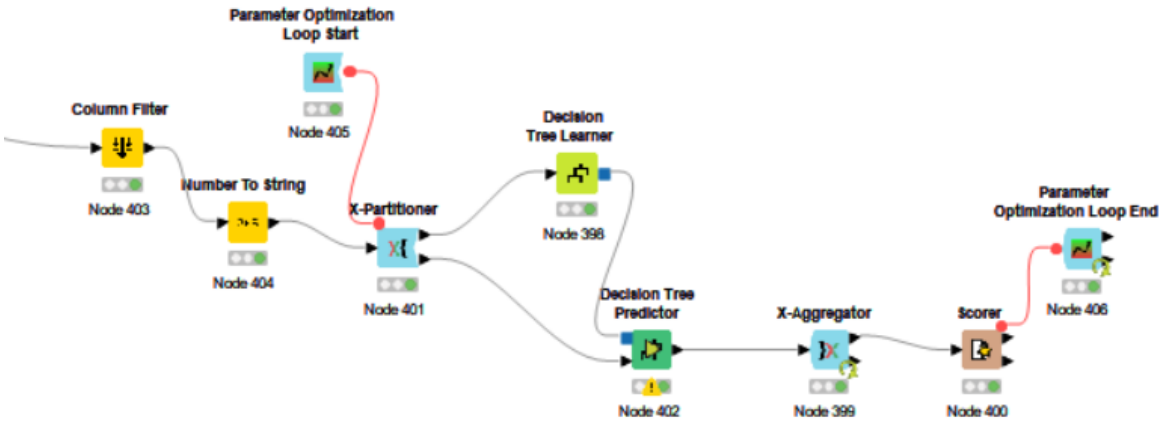
Apêndice C- Estrutura de nós para realizar o estudo referente ao desenvolvimento do sistema de recomendação avaliado com base nas métricas de *lift* e suporte de valores altos

APÊNDICE D- FLUXO DO MODELO K-NEAREST NEIGHBORS



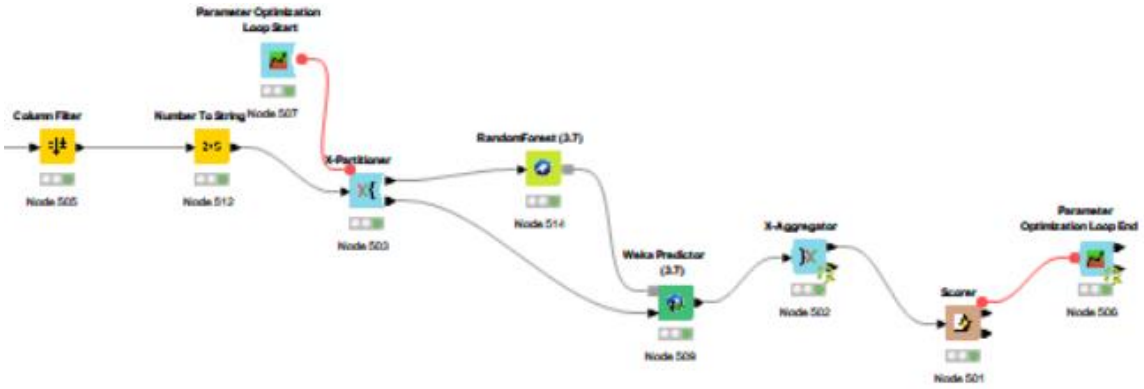
Apêndice D- Fluxo de trabalho para o modelo KNN

APÊNDICE E- FLUXO DO MODELO DE ÁRVORE DE DECISÃO



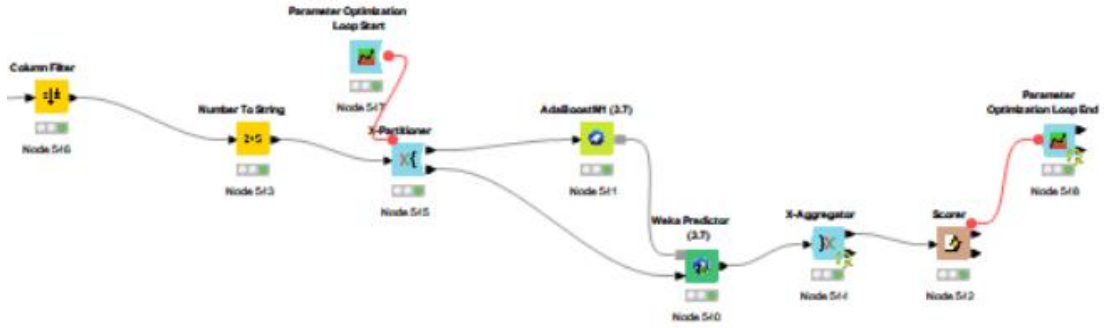
Apêndice E- Fluxo de trabalho para o modelo de árvore de decisão

APÊNDICE F- FLUXO DO MODELO DE *RANDOM FOREST*



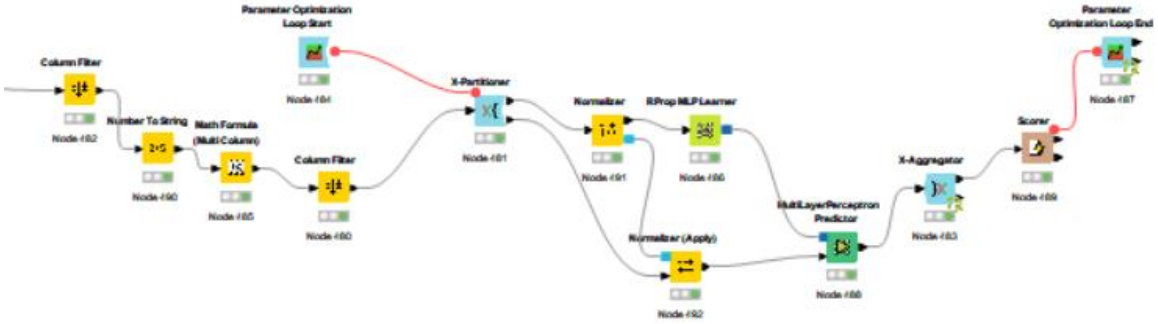
Apêndice F- Fluxo de trabalho para o modelo de *Random Forest*

APÊNDICE G- FLUXO DO MODELO DE *ADABOOST*



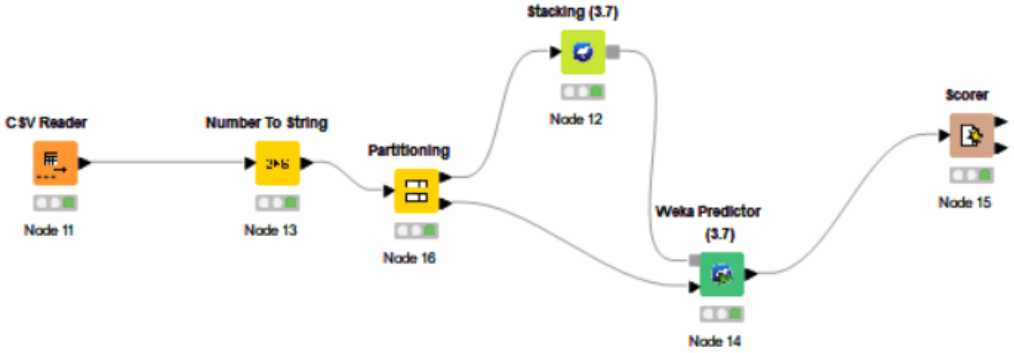
Apêndice G- Fluxo de trabalho para o modelo Adaboost

APÊNDICE H- FLUXO DO MODELO DE MLP



Apêndice H- Fluxo de trabalho para o modelo Adaboost

APÊNDICE I- FLUXO DO MODELO DE *STACKING*



Apêndice I- Fluxo de trabalho para o modelo stacking

APÊNDICE J- MODELOS DE PREVISÃO

K-Nearest Neighbors

Tabela J- 1- Diferentes valores de "k" em estudo e o seu respetivo desempenho

Modelo KNN	
k	***Accuracy***
3	0,648
4	0,65
5	0,657
6	0,659
7	0,663
8	0,664
9	0,667
10	0,668

Na tabela J-1 é possível visualizar os diferentes valores de *accuracy* obtidos para os diferentes valores de "k" vizinhos mais próximos, sendo que o valor de "k" que melhor desempenho oferece ao modelo é o valor de 10 com uma *accuracy* de 66,8%.

Árvore de decisão

Tabela J- 2- Diferentes valores de "MinNumOfRecords" em estudo e o seu respetivo desempenho

Árvore de decisão	
MinNumOfRecords	***Accuracy***
2	0,699
3	0,7
4	0,699
5	0,699
6	0,699
7	0,699
8	0,699
9	0,699
10	0,699
11	0,699
12	0,699
13	0,699
14	0,699
15	0,699

Na tabela J-2 é possível visualizar os diferentes valores de *accuracy* obtidos para os diferentes valores de "MinNumOfRecords", sendo que a variação do valor de "MinNumOfRecords" em nada contribuiu para uma melhor *accuracy* do modelo.

Random Forest

Tabela J- 3- Diferentes valores de *MaxDepth* em estudo e o seu respetivo desempenho

Random Forest	
MaxDepth	***Accuracy***
1	0,625
2	0,662
3	0,678
4	0,682
5	0,687
6	0,692
7	0,694
8	0,697
9	0,699
10	0,701
11	0,704
12	0,706
13	0,707
14	0,708
15	0,709
16	0,710
17	0,710
18	0,711
19	0,710
20	0,710
21	0,710
22	0,709
23	0,709
24	0,708
25	0,707

Na tabela J-3 é possível observar os diferentes valores de *accuracy* obtidos para os diferentes valores de máxima profundidade das árvores em estudo, sendo que a profundidade máxima de 18 é a que oferece uma melhor precisão ao modelo.

Tabela J- 4- Variação do número de árvores de decisão no modelo *Random Forest*

Random Forest	
Número de árvores de decisão	***Accuracy***
10	0,711
50	0,715
100	0,715
200	0,715
300	0,715

Conforme evidenciado na tabela J-4 constata-se que a precisão do modelo não varia significativamente à medida que o número de árvores de decisão aumenta, pelo que não justifica efetuar o mesmo estudo para os restantes valores. Sendo assim, o valor de 50 unidades para o número de árvores de decisão pode ser considerado como o ponto de referência para esse parâmetro.

Adaboost

Tabela J- 5- Variação do número de iterações no algoritmo de *AdaBoost*

AdaBoost	
NumIterations	***Accuracy***
3	0,662
6	0,662
9	0,662
12	0,662

Na Tabela J-5, denota-se a variação dos valores de precisão obtidos para as diversas iterações consideradas. Nota-se que o aumento no número de iterações não resulta em um aprimoramento da precisão do modelo. Portanto, o valor de referência pode ser estabelecido em 3, uma vez que um valor maior somente aumentaria o tempo de processamento do algoritmo.

Multilayer Perceptron

Tabela J-6- Variação do número de *hidden layers* e *hidden neurons* no algoritmo MLP

Multilayer Perceptron		
Camadas ocultas	**Neurônios ocultos**	***Accuracy***
1	1	0,684
2	1	0,684
3	1	0,646
4	1	0,611
5	1	0,613
1	2	0,684
2	2	0,684
3	2	0,684
4	2	0,639
5	2	0,647
1	3	0,684
2	3	0,685
3	3	0,685
4	3	0,685
5	3	0,675
1	4	0,685
2	4	0,685
3	4	0,685
4	4	0,685

Multilayer Perceptron		
Camadas ocultas	**Neurônios ocultos**	***Accuracy***
5	4	0,676
1	5	0,686
2	5	0,686
3	5	0,686
4	5	0,685
5	5	0,685
1	6	0,686
2	6	0,686
3	6	0,685
4	6	0,685
5	6	0,685
1	7	0,687
2	7	0,686
3	7	0,686
4	7	0,687
5	7	0,686
1	8	0,687
2	8	0,686
3	8	0,686
4	8	0,686
5	8	0,686
1	9	0,687
2	9	0,687
3	9	0,687
4	9	0,686
5	9	0,686
1	10	0,687
2	10	0,687
3	10	0,686
4	10	0,687
5	10	0,686

Na tabela J-6 é possível visualizar os diferentes valores de precisão obtidos para os diferentes valores de “Nr of hidden layers” e “Nr of hidden neurons”. Verifica-se que diversas configurações distintas de camadas ocultas e neurônios ocultos levam à mesma precisão. Por exemplo, a escolha da configuração com 1 camada oculta e 7 neurônios ocultos resulta em um modelo com precisão de 68,7%.