IPS Instituto Politécnico de Setúbal
Escola Superior de Ciências Empresariais

Data Mining na procura de nova informação: *Market Basket Analysis* aplicado a um *dataset* público

Joana Raquel Carias de Oliveira

Relatório de Dissertação do Mestrado em Gestão de Sistemas de Informação

ORIENTADOR

Professor Victor Barbosa

Dezembro de 2019

(esta página foi intencionalmente deixada em branco)

“Success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do.” - Pelé

Agradecimentos

O meu sincero agradecimento ao Professor Victor Barbosa, orientador da minha dissertação de mestrado, por toda a disponibilidade, dedicação e auxílio na elaboração desta tese de mestrado. Graças aos seu apoio, críticas construtivas e revisão do trabalho foi possível tornar esta dissertação de mestrado uma realidade.

Agradeço à minha família, em especial à minha mãe, pelo constante apoio, paciência, por nunca me deixar desmotivar e por me manter determinada a conseguir concluir este projeto.

Aos meus amigos, em especial à Raquel Santos e à Ines Cklamovska, pela sua amizade e palavras de incentivo.

Muito Obrigado!

Índice Geral

1 Introdução 1
1.1 Enquadramento 1
1.2 Âmbito do trabalho e Objetivos 2
1.3 Estrutura da Relatório 3
2 Metodologia 4
3 Enquadramento Teórico 9
3.1 Data Mining 9
3.1.1 Aplicações e Benefícios do Data Mining 10
3.1.2 Falácias associadas ao Data Mining 12
3.1.3 Tecnologias utilizadas no Data Mining 13
3.1.4 Linguagens de Programação mais utilizadas em *Data Mining* 14
3.1.5 Técnicas de Data Mining 18
3.2 Conhecer os Dados 19
3.2.1 Aquisição 19
3.2.2 Pré-Processamento 20
3.2.2.1 Sumarização Descritiva 20
3.2.2.2 Limpeza dos Dados 21
3.2.2.3 Integração 23
3.2.2.4 Transformação 24
3.2.2.5 Redução 25
3.3 Regras de Associação 26
3.3.1 Market Basket Analysis 27
3.3.1.1 Algoritmo Apriori 30
3.3.1.2 Exemplo de Geração de Regras de Associação 34
3.3.1.3 Interpretação das Regras de Associação 35
3.3.1.4 Algoritmos Eclat e FP-Growth 36

4 Análise Exploratória dos Dados 40
4.1 Descrição do Dataset e Modelo Relacional 40
4.1.1 Descrição das Variáveis 42
4.2 Principais Métricas e Indicadores 43
5 Análise aos carrinhos de compras 54
5.1 Tecnologias utilizadas 54
5.2 Preparação dos dados 55
5.3 Modelação 56
5.4 Comparação de Resultados obtidos com os dados de Treino e com os dados completos 64
5.5 Avaliação 66
6 Conclusões 73
7 Limitações e Recomendações para trabalhos futuros 75
Referências Bibliográficas 77
8 Anexos 81
8.1 Gráfico: Número de Compras consoante a Hora do Dia 81
8.2 Gráfico: Número de Dias até à próxima Compra 82
8.3 Gráfico: Contagem de Produtos por Departamentos e Corredores 83
8.4 Gráfico: Produtos mais adquiridos 84
8.5 Gráfico: Produtos mais requisitados por Departamentos e Corredores (Top 10) 85
8.6 Gráfico: Produtos menos requisitados por Departamentos e Corredores (Top 5) 86
8.7 Gráfico: Número de Produtos por Compra 87
8.8 Código Tratamento do Ficheiro 88
8.9 Script de Aplicação do Algoritmo Apriori 89
8.10 Aplicação do Algoritmo Apriori para compras com 11 ou mais produtos 90
8.11 Scripts para guardar regras de associação e outputs 91

Lista de Tabelas

Tabela 1 - Visão geral das tarefas CRISP e dos seus outputs 6
Tabela 2- Transações registadas ao longo do dia pelo agricultor..... 29
Tabela 3 - Dados em formato tabular referentes às transações registadas ao longo do dia pelo agricultor 32
Tabela 4 - Candidatos a 2-itemset 33
Tabela 5 - Regras de Associação para dois Antecedentes, exemplo da venda de vegetais 34
Tabela 6 - Regras de Associação para um Antecedente, exemplo da venda de vegetais35
Tabela 7- Descrição dos tipos de variáveis do Dataset analisado..... 42
Tabela 8 - Regras de Associação para o conjunto de dados Train..... 58
Tabela 9 - Regras de Associação A, B --> C para o conjunto de dados Train considerando 11 ou mais produtos por compra..... 59
Tabela 10 - Regras de Associação com Mínimo Suporte 0,006; Mínimo Confiança 0,4 e mínimo Lift 1,1 para o conjunto Train com 11 ou mais produtos por compra..... 60
Tabela 11- Regras de Associação para o conjunto de dados Histórico (Prior) 61
Tabela 12 - Regras de Associação Mínimo Suporte 0,001; Mínimo Confiança 0,5 e mínimo Lift 1,1 conjunto de dados histórico 62
Tabela 13 - Regras de Associação Mínimo Suporte 0,0025; Mínimo Confiança 0,4 e mínimo Lift 1,1 conjunto de dados histórico 62
Tabela 14 - Regra de Associação A, B --> C para conjunto de dados histórico considerando compras com 11 ou mais produtos 63
Tabela 15- Regras de Associação A, B--> C Mínimo Suporte 0,008; Mínimo Confiança 0,2 e mínimo Lift 1,1 para conjunto de dados histórico..... 63
Tabela 16 - Regras de Associação Mínimo Suporte 0,002; Mínimo Confiança 0,5 e mínimo Lift 1,1 para conjunto de dados histórico com mais de 10 produtos por compra..... 64
Tabela 17 - Comparação Regras de Associação obtidas Conjunto Train e Conjunto Histórico..... 65
Tabela 18 - Número de compras dos conjuntos de dados..... 66
Tabela 19- Número de Compras a que corresponde o Suporte Mínimo em cada subconjunto..... 67
Tabela 20- Contagem de Regras geradas com confiança mínima de 20% 67
Tabela 21- Contagem de Regras geradas com confiança mínima de 25% 68
Tabela 22 - Contagem de Regras geradas com confiança mínima de 50% 70

Lista de Figuras

Figura 1 - Metodologia CRISP 4
Figura 2 - Metodologia SEMMA 7
Figura 3 - Ferramentas mais utilizadas em Data Science entre 2017 e 2019 15
Figura 4- Comparação entre utilização da linguagem de programação R e Python (Kaggle, 2019) 15
Figura 5- Evolução da utilização da linguagem de programação Python e R (Piatetsky, 2019) 16
Figura 6 - Exemplo de árvore do Eclat 38
Figura 7 - Exemplo de união de dois itemsets no Eclat 38
Figura 8 - Modelo Relacional do Conjunto de Dados a analisar 41
Figura 9 – Número Total de Compras 43
Figura 10 - Número de Compras consoante a Hora do Dia. 43
Figura 11 - Número de Compras consoante período do dia 44
Figura 12- Compras por dia da semana 44
Figura 13 - Número de Utilizadores 45
Figura 14 - Compras por Utilizador 45
Figura 15 - Total de Utilizadores por número de Compras 45
Figura 16- Utilizadores por Total de Compras 46
Figura 17 - Média de Compras por Utilizador 46
Figura 18 - Mediana de Compras por Utilizador 46
Figura 19 - Número de Dias até próxima Compra 47
Figura 20 - Produtos mais adquiridos (Top 10) 48
Figura 21 - Produtos mais vezes recomprados 49
Figura 22 - Percentagem de compras com produtos recomprados 49
Figura 23 - Top 10 Número de Produtos por Compra 50
Figura 24 - Média de Produtos Adquiridos por Compra 50
Figura 25 - Mediana de Produtos adquiridos por Compra 50
Figura 26 - Ordem de adição de itens ao carrinho de compras 51
Figura 27 - Produtos com elevada percentagem em serem selecionados como primeiro produto a colocar na cesta de compras 52
Figura 28 - Departamentos com mais compras e percentagem de seleção de primeiro item no carrinho de compras 52
Figura 29 - Departamentos com maior percentagem de produtos selecionados em primeiro lugar no carrinho de compras. 53
Figura 30 – Formato original dos dados das compras 55

Figura 31 – Representação das compras numa só linha 55
Figura 32 – Carregamento dos dados para algoritmo pesquisa de Regras de Associação 56
Figura 33 - Gráfico Regras de Associação Geradas - Dataset Treino (confiança=25%).. 69
Figura 34 - Gráfico Regras de Associação Geradas - Dataset Histórico (confiança=25%) 69
Figura 35 - Gráfico Regras de Associação Geradas – Compras com mais de 20 produtos (confiança=50%) 70
Figura 36 – Comparação do número de regras obtidas com o valor real de suporte mínimo 71
Figura 37 - Anexo 1 Número de Compras consoante a Hora do Dia 81
Figura 38- Anexo 2 Gráfico Número de Dias até próxima Compra 82
Figura 39 - Anexo 3 Gráfico Contagem de Produtos por Departamentos e Corredores .. 83
Figura 40 - Anexo 4 Produtos mais adquiridos 84
Figura 41- Anexo 5 Produtos mais requisitados por Departamentos e Corredores 85
Figura 42 - Anexo 6 Produtos menos requisitados por Departamentos e Corredores 86
Figura 43 - Anexo 7 Número de Produtos por Compra 87
Figura 44 – Script para guardar regras em ficheiro 91
Figura 45 – Script para converter regras do ficheiro para formato legível 91

Lista de Siglas e Abreviaturas

APED: Associação Portuguesa das Empresas de Distribuição;

CRISP: *Cross Industry Standard Process for Data Mining;*

CSV: *Comma Separated Values;*

ECLAT*: Equivalence Class Clustering and bottom-up Lattice Traversal;*

JSON: *JavaScript Object Notation;*

MBA: *Market Basket Analysis*;

PDF: *Portable Document Format;*

SEMMA: *Sample, Explore, Modify, Model e Assess***;**

SQL: *Structured Query Language;*

Resumo

Hoje em dia, a população encontra-se sobrecarregada com dados, quando todas as atividades realizadas pelas organizações e pessoas, no seu dia-a-dia, geram dados. Contudo, o facto de termos acesso a um enorme volume de dados não significa que tenhamos acesso a muita informação ou conhecimento. É, portanto, importante trabalhar os dados por forma a gerar informação relevante para a tomada de decisão, pois num mundo globalizado e extremamente competitivo, um minuto pode ser fulcral para fechar um negócio e, para tal, é necessário ter acesso à informação atual, correta e sumarizada.

Face ao volume de dados existente e a necessidade de criar vantagens competitivas para as empresas sobreviverem nos seus mercados importa analisar os dados por forma a identificar informação que poderia estar oculta ou padrões nos comportamentos dos consumidores. É aqui que entra o *data mining*, cujo principal objetivo é analisar os dados e encontrar anomalias, padrões ou novas informações que auxiliem na tomada de decisão.

O setor do retalho é um dos setores que mais valor monetário gera mundialmente e um dos setores onde a concorrência é mais feroz, pelo que quanto mais conhecimento e informações as empresas tiverem ao seu dispor maior será a probabilidade de conseguirem adquirir vantagens competitivas. Nesta procura de informação temos como exemplo as regras de associação, uma técnica de *data mining* cujo objetivo é encontrar itens que ocorrem frequentemente e em conjunto nos cestos de compras dos clientes. Um dos algoritmos concebidos para a geração de regras de associação é o algoritmo Apriori em que a sua génese foi baseada na análise de compras efetuadas num supermercado. Ao aplicar algoritmos para obter regras de associação ao setor do retalho é comum indicar-se que se usou uma técnica de *market basket analysis*.

Este trabalho tem como principais objetivos a análise exploratória de um *dataset* público com um grande conjunto de compras (Instacart) e a geração de regras de associação recorrendo à utilização do algoritmo Apriori. Consoante os resultados obtidos serão sugeridas ideias para implementar novas estratégia de *marketing*.

Este trabalho iniciou-se com a revisão da literatura, investigando os conceitos de *data mining*, regras de associação e *market basket analysis*. Como bússola orientadora para a aplicação de técnicas de *data mining* seguiu-se a metodologia CRISP. Para a análise exploratória dos dados foi utilizado o *software Power BI e* para a transformação dos dados e aplicação do algoritmo Apriori e consequentemente a geração das regras de associação recorreu-se à linguagem *Python*.

Palavras – chave: *Data Mining*; Regras de Associação; Algoritmo Apriori

Abstract

Nowadays, the population is overloaded with data, where practically all activities performed generate it. However, the fact that we have access to a huge amount of data does not mean that we have access to a lot of information. Therefore, it is important to process the data in order to generate relevant information for decision making as well as to separate the relevant and non-relevant data in order to speed up processes that generate important information in decision making, because in a globalized and extremely competitive world, a minute can decide the fate of closing a deal.

Given the volume of data that exists and the need to create competitive advantages for companies to survive in their markets, it is important to analyze the data to identify information that could be hidden or patterns in consumer behavior. This is where data mining comes in, its main goal is to analyze data and find anomalies, patterns or new information to assist in decision making.

The retail area is one of the most monetarily generated areas in the world and one of the fiercest competition areas, so companies that have better knowledge and information are likely to gain competitive advantage. In this information quest, we have as an example the association rules, a data mining technique where the objective is to find items that occur frequently and together. One of the algorithms designed to generate association rules is the Apriori algorithm, its genesis was based on the analysis of purchases made in a supermarket. Thus, when applying the rules of association to the retail sector, it is usual to indicate that the market basket analysis technique was used.

This work has as main objectives the exploratory data analysis of a public *dataset* with a huge number of market baskets records (Instacart) and the generation of association rules using the Apriori algorithm. Depending on the results obtained, ideas for implementing a new marketing strategy will be suggested.

This work started with the literature review, investigating the concepts of data mining, association rules and market basket analysis. An example of the application of the Apriori algorithm for the generation of association rules was also verified. As a guiding compass for the application of data mining techniques, the CRISP methodology was followed. For exploratory data analysis is was used Power BI software and for data transformation and application of the Apriori algorithm and consequently the generation of association rules was used Python.

Keywords: Data Mining; Association Rules; Apriori Algorithm

1 Introdução

1.1 Enquadramento

De acordo com o relatório *Food and Grocery Retail Market Analysis* (2018), a indústria global do retalho de alimentos tem crescido constantemente nos últimos anos - uma tendência que deve continuar ao longo do período de previsão. Nesse mesmo relatório prevê-se que a dimensão global do mercado do retalho de alimentos e supermercados atinja 12,24 triliões de dólares até 2020.

Em Portugal também se verifica o crescimento do volume de vendas no retalho, com o presidente da Associação Portuguesa das Empresas de Distribuição (APED), Jorge Jordão, a indicar que o setor do retalho registou em 2017 um desempenho positivo, com um crescimento global de 3,8% face a 2016 e superando o patamar de 20 mil milhões de euros de volume de vendas, sendo que estes dados são referentes ao retalho alimentar, na ordem de 3,9% e no retalho não alimentar o crescimento registado foi de 3,8% (Agência Lusa, 2018).

O relatório *Global Powers of Retailing* 2018 (Deloitte, 2018), refere que as regras do retalho estão a mudar. Conceitos como inovação, colaboração, consolidação, integração, e automação são necessários para aplicar e revigorar o comércio, tendo um profundo impacto na forma como os retalhistas fazem negócios agora e no futuro.

De acordo com Finlay (2014), nos últimos anos as organizações tornaram-se cada vez mais interessadas nos intervalos entre as transações dos clientes e nos caminhos que os levaram às decisões que tomam. À medida que se fazem mais coisas eletronicamente, ficam disponíveis informações que fornecem *insights* sobre os processos de pensamento e as influências que levam a participar numa atividade em vez da outra. Toda esta informação sobre as pessoas é incrivelmente útil por vários motivos, mas o principal motivo é que permite prever o comportamento futuro. Ao usar informações sobre estilos de vida, movimentos e comportamentos passados das pessoas, as organizações podem prever o que provavelmente farão, quando farão e onde essa atividade irá ocorrer. Essas previsões são utilizadas para adaptar a forma como as organizações interagem com as pessoas, por forma a influenciar o comportamento das pessoas, com o objetivo de maximizar o valor das relações que as pessoas têm com as organizações.

Desta forma, depreende-se que num mundo globalizado, onde as alterações são constantes e ocorrem a um ritmo elevado, o sucesso das organizações depende da sua capacidade de antecipar essas mudanças ou de conseguirem reagir rapidamente quando

surgem. Assim sendo, o sector do retalho não é uma exceção e perante este vasto mercado, onde existe muita competição e esta é implacável, as empresas procuram obter vantagens competitivas face aos seus concorrentes por forma a conseguirem ter um lugar neste sector. Estas vantagens competitivas são muitas vezes obtidas através da análise de dados que permitem às empresas adquirir conhecimento acerca do negócio, clientes e concorrência.

São várias as técnicas utilizadas pelas organizações para aquisição de conhecimento que permitem gerar vantagens competitivas e fazer com que estas se destaquem da concorrência. Entre elas, encontram-se o *Data Mining*, ou mineração de dados, este conceito, segundo Azevedo & Santos (2005), recorre a algoritmos específicos ou a mecanismos de pergunta, tentando descobrir padrões discerníveis e tendências nos dados e inferir regras para os mesmos. A análise de dados pode fornecer um conhecimento adicional acerca do negócio, ao permitir ir além dos dados armazenados. É a partir dessa possibilidade que a utilização de *data mining* evidencia visíveis benefícios.

Para o sector do retalho são várias as técnicas de *data mining* que podem ser utilizadas por forma a adquirir conhecimento e identificar padrões e tendências através dos dados. A *Market Basket Analysis* (MBA) é uma das técnicas frequentemente utilizada pelos retalhistas. Os autores Giudici & Figini (2009) mencionam que a MBA tem o objetivo de identificar produtos ou grupos de produtos que tendem a estar juntos (são associados) nas transações de compra. Posteriormente, o conhecimento obtido será valioso para a organização, possibilitando por exemplo, determinado supermercado reorganizar o seu *layout*, colocando produtos frequentemente vendidos juntos ou localizando-os perto.

Esta afirmação é corroborada pelo autor Gayathri (2017) que indica que a MBA é uma técnica poderosa, especialmente no retalho, pois lida com milhares de itens e analisa os hábitos de compra dos clientes identificando os produtos que tendem a ser comprados em conjunto. A descoberta dessas associações ajuda os retalhistas a desenvolver estratégias de *marketing* e a tomar decisões de negócio.

1.2 Âmbito do trabalho e Objetivos

Face à importância do setor de retalho para a economia mundial e sendo o *data mining* um conceito atual e relevante para a obtenção de vantagens competitivas, considera-se pertinente a análise exploratória de dados e a aplicação da metodologia MBA ao conjunto de dados selecionado para este estudo. O tema da dissertação foi definido em conjunto com o orientador, pela sua atualidade, potencial interesse e aplicabilidade em contexto real realizando os devidos ajustes.

Assim sendo, o principal objetivo deste trabalho é demonstrar a aplicação da técnica de *Market Basket Analysis,* através da identificação dos produtos que os clientes compraram frequentemente em conjunto, identificando padrões nas compras que ajudem a elaborar ações de *marketing*.

Para atingir o principal objetivo do estudo, foram definidos os seguintes objetivos específicos:

- Conhecer a literatura aplicável referente à utilização de técnicas de *Data Mining*, mais concretamente a aplicação da metodologia de *Market Basket Analysis* através da identificação/criação de Regras de Associação;
- Caraterizar o conjunto de dados a estudar e fazer a sua análise, obtendo indicadores das transações analisadas;
- Criar um modelo e avaliar a sua qualidade e comportamento com os dados;
- Explorar as configurações do modelo e o seu impacto nos resultados alcançados.

Para a elaboração deste trabalho foi utilizado um conjunto de dados público (Instacart, 2017). Este apresenta dados referentes ao ano de 2017 e é composto por uma amostra de mais de 3 milhões de compras de supermercado com mais de 200.000 utilizadores da Instacart, empresa norte americana de entrega de produtos de mercearia *online*.

A escolha deste conjunto de dados em específico foi devido ao facto de se tratar de um conjunto de dados público, uma vez que existiram grandes dificuldades em obter conjuntos de dados reais referentes a empresas portuguesas, pois as mesmas apresentam bastante relutância em partilhar os seus dados devido a questões de segurança, proteção dos dados dos consumidores e receio acerca da utilização dos seus dados.

1.3 Estrutura da Relatório

Este relatório encontra-se dividido em sete capítulos. No primeiro capítulo é feito um breve enquadramento, é mencionado o âmbito e objetivos do projeto e indicada a estrutura do relatório. No segundo capítulo aborda-se a metodologia utilizada para a realização deste projeto. No capítulo seguinte é realizada a revisão bibliográfica, onde são abordados conceitos relacionados com *data mining, market basket analysis* e regras de associação. No quarto capítulo carateriza-se o conjunto de dados que é objeto de estudo e identificam-se indicadores relevantes para o negócio obtidos durante a análise exploratória dos dados. O próximo capítulo versa sobre a aplicação do algoritmo de Apriori e a obtenção das regras de associação. No capítulo sexto são realizadas as conclusões e confere-se se objetivos propostos foram alcançados. No último capítulo mencionam-se as dificuldades encontradas e aborda-se trabalho a desenvolver no futuro.

2 Metodologia

Segundo Prodanov & Freitas (2013), uma metodologia permite a identificação e aplicação de procedimentos e técnicas para obtenção de conhecimento, com o objetivo de evidenciar e comprovar a sua validade e utilidade em diversos domínios da sociedade. Por sua vez Campomar (1991) indica que a metodologia também chamada de método científico é de extrema importância na investigação académica pois se a mesma não existisse os resultados das investigações não teriam grande aceitação. É através do método científico que a sociedade valida o conhecimento adquirido empiricamente. Verifica-se então que uma metodologia serve para identificar e aplicar procedimentos e técnicas que nos ajudam a percorrer o caminho para chegar a determinado conhecimento e que sem a aplicação de uma metodologia em investigação os resultados obtidos não terão credibilidade no meio académico.

Para a aplicação de técnicas de *data mining* as duas principais metodologias utilizadas são a CRISP (*Cross Industry Standard Process for Data Mining*) e a SEMMA (*Sample, Explore, Modify, Model e Assess*) de acordo com Azevedo & Santos (2008). Larose & Larose (2014) indicam que a CRISP foi desenvolvida por analistas que representavam a Daimler-Chrysler, SPSS e NCR e fornece um processo padrão para adaptar a tarefa de mineração de dados na estratégia geral de solução de problemas.

Segundo Wirth & Hipp (2000), a metodologia CRISP fornece uma *framework* para a realização de projetos de *data mining*. Este modelo é independente do setor de atividade e da tecnologia utilizada e pretende fazer com que grandes projetos de mineração de dados sejam mais confiáveis, mais fáceis de gerir, proporcionem reduções de custos, sejam repetitivos e mais rápidos de implementar.

O modelo é iterativo e composto por seis fases. Na Figura 1 ilustram-se as fases do modelo que a seguir se descrevem, segundo Provost & Fawcett (2015) e Wirth & Hipp (2000).

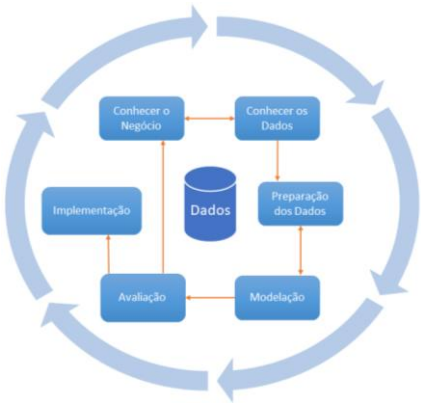

Figura 1 - Metodologia CRISP
Fonte: Adaptado de Provost & Fawcett, (2015) Figure 2-2.

1. **Conhecer o Negócio:** foco no entendimento do problema a ser resolvido. Nos projetos, raramente são pré-definidos os objetivos da mineração de dados de forma clara e inequívoca. Muitas vezes, reformular o problema e projetar uma solução é um processo iterativo. A formulação inicial pode não estar completa ou otimizada, pelo que podem ser necessárias múltiplas iterações.
2. **Conhecer os Dados:** esta fase inicia-se com a recolha inicial de dados. Procura-se identificar problemas relacionados com a qualidade dos dados, detetar subconjuntos interessantes para formular hipóteses e descobrir *insights* nos dados. Existe uma estreita ligação entre a fase de conhecer o negócio e a fase de conhecer os dados, pois apenas conhecendo o negócio se pode concluir que os dados recolhidos são adequados para o problema que se pretende estudar e apenas após a análise dos dados se pode verificar que os mesmos proporcionam *insights* úteis para o negócio.
3. **Preparação dos Dados:** as tecnologias analíticas que podemos usar normalmente impõem certos requisitos aos dados que utilizam. Geralmente exigem que os dados estejam em determinados formatos sendo por isso necessário efetuar conversões, limpeza e transformação nos dados recolhidos.
4. **Modelação:** esta etapa engloba a seleção e aplicação das técnicas apropriadas de *data mining* para o projeto em questão. Deste modo, é importante ter algum conhecimento das ideias fundamentais da mineração de dados, incluindo os tipos de técnicas e algoritmos existentes, porque é nesta etapa que a maior parte da ciência e da tecnologia será utilizada.
5. **Avaliação:** o objetivo desta etapa é avaliar os resultados da mineração de dados com rigor e ter a certeza de que estes são válidos e confiáveis antes de seguir em frente. Se procurarmos bastante em qualquer conjunto de dados, encontraremos padrões, mas temos de ter confiança de que estes são verdadeiras regularidades e não anomalias.
6. **Implementação:** na implementação, as técnicas aplicadas e o modelo gerado são disponibilizados para utilização em contexto real. Independentemente do sucesso da implementação, o processo geralmente retorna à fase de conhecer o negócio, uma vez que foi produzida uma grande quantidade de informações sobre o negócio e uma segunda iteração pode produzir uma melhor solução. Só a experiência de pensar sobre o problema de negócio, os dados e os objetivos que se pretendem alcançar originam novas ideias para melhorar o desempenho dos negócios. De notar que não é necessário falhar na implementação para iniciar o ciclo novamente, na fase de Avaliação os resultados podem não ser suficientemente bons, sendo

preciso ajustar a definição do problema, obter dados diferentes ou até mesmo recorrer a outras técnicas.

Os autores Wirth & Hipp (2000) elaboraram uma tabela que refere, além das fases do modelo CRISP, as tarefas genéricas associadas a cada fase e o *output* esperado. A Tabela 1 apresenta essa tarefas e outputs.

Tabela 1 - Visão geral das tarefas CRISP e dos seus outputs

Fonte: Adaptado de Wirth & Hipp (2000) figura 3

Conhecer o Negócio	Conhecer os Dados	Preparação dos Dados	Modelação	Avaliação	Implementação
Determinar os objetivos do Negócio *(Background*, objetivos de negócio e critérios de sucesso do negócio)	**Recolha Inicial dos dados** (Relatório da recolha inicial de dados)	(Conjunto de dados) (Descrição do Conjunto de dados)	**Seleção da técnica de Modelação** (Técnica de Modelação e Suposições do Modelo)	**Avaliação dos Resultados** (Avaliação dos resultados do *Data Mining* face aos critérios de sucesso do negócio, aprovação dos modelos)	**Plano de Implementação** (Plano de Implementação)
Aferir a situação (Inventário de recursos e requisitos, suposições e restrições, riscos e contingências, terminologia e custos e benefícios)	**Descrição dos dados** (Relatório com a descrição dos dados)	**Seleção dos dados** (Justificativa para Inclusão/Exclusão)	**Geração de teste de Design** (Teste de Design)	**Revisão do Processo** (Revisão do Processo)	**Plano de Monitorização e Manutenção** (Plano de Monitorização e Manutenção)
Determinar os objetivos do Data Mining (Data Mining objetivos e critérios de sucesso)	**Explorar os dados** (Relatório da exploração dos dados)	**Limpeza dos dados** (Relatório da limpeza dos dados)	**Construção do Modelo** (Definições de parâmetros, modelos e descrição do modelo)	**Determinação dos próximos passos** (Lista de possíveis ações e decisões)	**Relatório Final do Produto** (Relatório Final e apresentação final)
Elaboração do Plano do Projeto (Plano do Projeto e aferição inicial das ferramentas e técnicas a utilizar)	**Verificação da Qualidade dos Dados** (Relatório referente à qualidade dos dados)	**Construção dos dados** (Atributos derivados e registos gerados)	**Aferição do modelo** (Aferição do modelo e revisão das definições dos parâmetros)		**Revisão do Projeto** (Experiência e Documentação)
		Integração dos dados (Unir dados)			
		Formatar dados (Dados formatados)			

A metodologia SEMMA foi desenvolvida pelo SAS Institute, um dos maiores produtores de software de *business intelligence* e estatística. Segundo Olson & Delen (2008), a metodologia SEMMA pretende facilitar a aplicação de técnicas exploratórias estatísticas e de visualização, selecionando e transformando as variáveis preditivas mais significativas, modelando as variáveis para prever resultados e confirmar a precisão do modelo. Na

Figura 2 ilustra-se o modelo para aplicar a referida metodologia e posteriormente explicam-se as suas diferentes fases.

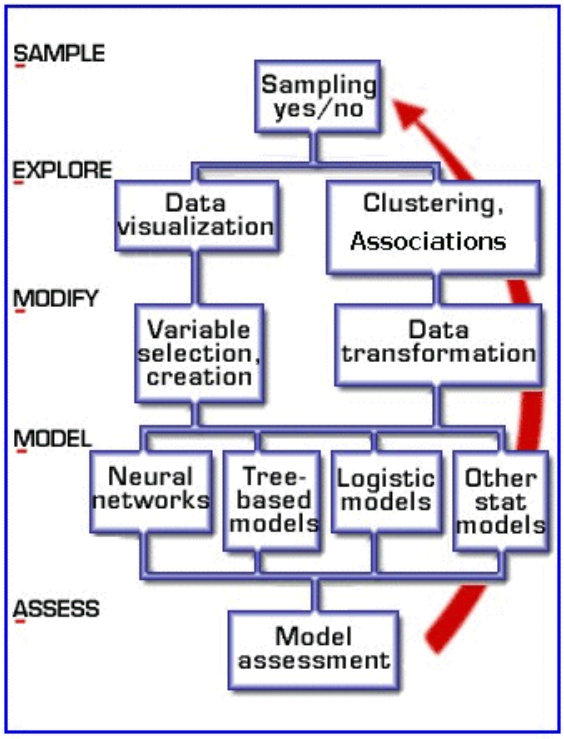

Figura 2 - Metodologia SEMMA

Fonte: Retirado de (SAS, 2017)

De acordo com Olson & Delen (2008) estas são as fases que constituem o SEMMA:

1. **Sample:** onde uma parte de um grande conjunto de dados é extraída (grande o suficiente para conter a informação significativa e ainda pequena o suficiente para se manipular rapidamente). Para um ótimo custo e desempenho computacional, é defendida uma estratégia de amostragem, que usa uma amostra confiável e estatisticamente representativa dos dados completos.
2. **Explore:** depois da fase da amostragem, o próximo passo é explorar os dados, visual ou numericamente, para identificar tendências ou agrupamentos inerentes. A exploração ajuda a refinar e redirecionar o processo de descoberta.
3. **Modify:** é nesta fase que o investigador seleciona e transforma as variáveis sobre as quais vai centrar o processo de construção do modelo. Com base nas descobertas na fase de exploração, pode ser necessário manipular os dados para incluir informações ou para introduzir novas variáveis/atributos. Também pode ser necessário procurar *outliers* e reduzir o número de variáveis, para reduzi-las às mais significativas.
4. **Model:** esta é a fase de escolha do modelo a adotar para o problema que está a ser estudado. Cada tipo de modelo tem pontos fortes e é apropriado em situações

específicas de mineração de dados, dependendo dos dados, cabe ao investigador/cientista saber escolher o modelo apropriado.

5. **Assess:** esta é a etapa final e é onde o investigador avalia a utilidade e a confiabilidade das descobertas do processo de *data mining*. Uma forma de avaliar o modelo é aplicá-lo a uma parte do conjunto de dados que não tenha sido utilizado durante a construção do modelo durante a fase de amostragem. Se o modelo for válido, deve funcionar para esta amostra reservada, bem como para a amostra utilizada na sua construção

Tanto a metodologia CRISP como a SEMMA são metodologias iterativas e possíveis de ajustar consoante o projeto que estamos a estudar. Para este projeto, optou-se por seguir a metodologia CRISP, uma vez que é o *standard* mais utilizado na indústria (Huber, Wiemer, Schneider, & Ihlenfeldt, 2018), na literatura, como em (Moro, Laureano, & Cortez, 2011; Sharma, Stranieri, Ugon, Vamplew, & Martin, 2017), e na comunidade (Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, 2014), e a qual permite facilmente fazer adaptações consoante o problema a estudar.

3 Enquadramento Teórico

3.1 Data Mining

Segundo Torgo (2017) o *data mining* é uma área de investigação relativamente recente e o seu principal objetivo é a análise de dados para a obtenção de conhecimento. Trata-se de uma área que ultimamente tem sido alvo de muita atenção pois hoje em dia recolhemos dados na maior parte das atividades que desempenhamos. Os dados recolhidos podem conter informação oculta acerca das atividades desempenhadas que podem ser úteis e dar vantagem competitiva às empresas.

De acordo com Kaur & Kang (2016), atualmente existem elevadas quantidade de dados nas bases de dados de vários setores, como o retalho, instituições bancárias, serviços relacionados com a área da saúde, entre outros. Contudo, nem toda a informação disponível é útil para o utilizador, sendo por isso muito importante extrair as informações úteis dessa grande quantidade de dados. Ao processo de extrair informações úteis do conjunto de dados em análise é dado o nome de *data mining*.

Os autores Tripathi *et al.* (2018) referem que *data mining* significa minerar elevadas quantidades de dados e apresentar previsões com base nesses dados. Além disso, referem também que ao realizar-se a mineração dos dados consegue-se detetar padrões e comportamentos nos dados.

A atual sobrecarga de dados, sem fim à vista, é também mencionada por Chakrabarti *et al.* (2009). Os autores indicam que esta se deve ao facto de atualmente termos ao dispor unidades de armazenamento a custos acessíveis, possibilitando salvar dados que anteriormente teriam sido destruídos. A tecnologia regista todas as nossas decisões como por exemplo: as nossas escolhas no supermercado, os nossos hábitos financeiros, as viagens que fazemos.

Os mesmos autores referem que as pessoas procuram padrões nos dados desde que a vida humana começou. Como exemplo temos: caçadores que procuram padrões no comportamento de migração animal, os agricultores procuram padrões no crescimento das culturas, os políticos na opinião do eleitor, entre outros. Face ao volume de dados que temos para analisar, o *data mining* torna-se imprescindível para nos elucidar acerca de padrões subjacentes nos dados. Dados analisados de forma inteligente são um recurso valioso, podendo levar a novos *insights* e, nos negócios, a vantagens competitivas. O *data mining* é definido como o processo de descoberta de padrões nos dados. Estes padrões descobertos devem ser significativos, permitindo a geração de vantagens.

Deste modo, constata-se que hoje em dia, temos ao nosso dispor uma quantidade imensurável de dados, sendo que esta abundância não nos permite dispor de tempo suficiente para os analisar a todos e tomar uma decisão atempada. Vivemos num mundo globalizado, onde a cada segundo ocorrem mudanças e as nossas decisões têm de ser tomadas de forma célere, mas informada. Para auxiliar na tomada de decisão temos a possibilidade de aplicar técnicas de *data mining*.

3.1.1 Aplicações e Benefícios do Data Mining

São várias as áreas de atuação, onde a utilização de técnicas de *data mining* pode ser útil e trazer valor para as organizações. De seguida apresentam-se alguns exemplos de aplicações de *data mining* segundo Guohua & Francis (2017):

- **Marketing** – O *data mining* é utilizado para identificar as preferências dos clientes e os padrões de compra. Esses resultados serão posteriormente utilizados para determinar quem são tipicamente os clientes de um determinado conjunto de produtos, permitindo segmentar os mesmos e saber a quem direcionar cada campanha para que esta tenha o seu retorno maximizado.
- **Ciência** – O *data mining* ajuda os humanos na investigação científica e novas descobertas. Ao examinar enormes conjuntos de dados, consegue encontrar padrões em estruturas moleculares, dados genéticos, mudanças climáticas globais e muito mais.
- **Finanças** –Algumas técnicas de *data mining* são usadas para detetar padrões em possíveis transações fraudulentas em cartões de crédito. Além disso, o *data mining* no setor financeiro é também utilizado para prever as flutuações da taxa de juros e de câmbio, bem como, para avaliações de risco de crédito e previsão de falência em empréstimos e classificação de títulos.
- **Indústria** –A tecnologia de mineração de dados tem sido usada com sucesso em áreas de produção, como programação e planeamento de produção, controlo de qualidade da produção e otimização da alocação de recursos.

Além das inúmeras aplicações de *data mining* nas diversas áreas de atuação, importa salientar também os seus principais benefícios. De seguida enumeram-se os principais benefícios de acordo com Finlay (2014):

- **Velocidade -** quando são usados modelos preditivos como parte de um sistema de tomada de decisão automatizado, milhões de clientes podem ser avaliados e tratados em apenas alguns segundos. Se um banco quiser produzir uma lista de clientes de cartão de crédito que também possa ser boa para um crédito automóvel, um modelo preditivo permite que isso seja feito rapidamente e com custo quase nulo. Avaliar todos os clientes de cartão de crédito dos bancos manualmente para encontrar potenciais clientes interessados seria completamente impraticável;
- **Melhores previsões -** o uso de modelos de análise preditiva geralmente cria melhores previsões. As melhores previsões dependem do problema em questão e são normalmente difíceis de quantificar;
- **Consistência -** um determinado modelo preditivo gerará sempre a mesma previsão quando utilizado com os mesmos dados. Este não é o caso das decisões efetuadas por humanos, pois até mesmo o especialista mais competente chegará a conclusões diferentes e tomará uma decisão diferente sobre algo dependendo de seu humor, da hora do dia, se eles estão com fome ou não entre outras séries de fatores.

Podemos constatar que são vários os benefícios associados à utilização de técnicas de *data mining*, observando exemplos de grandes marcas que utilizam com bastante sucesso o *data mining* para obter vantagens competitivas e proliferar nos seus negócios. Petersen (2016), identifica várias empresas famosas na sua área de atuação que recorrem ao *data mining* para obter benefícios para os seus negócios, seguem alguns exemplos providenciados pelo autor:

- **Delta:** Grandes companhias aéreas como a Delta monitorizam *tweets* para descobrir como os seus clientes se sentem com relação a atrasos, atualizações e entretenimento a bordo. Quando um cliente escreve um *tweet* negativo acerca de bagagem perdida, a companhia aérea envia um representante ao destino do passageiro, apresentando-lhe um bilhete gratuito de *upgrade* de primeira classe em seu retorno, juntamente com as informações sobre a bagagem perdida, prometendo entregá-la assim que a pessoa sair do avião.
- **Kohl's:** os clientes estão mais propensos a aceitar uma oferta quando estão na loja a comprar. É por isso que a Kohl's oferece ofertas personalizadas em tempo real. Os compradores podem optar por ofertas através de seus *smartphones*. Assim, se um comprador está no departamento de sapatos, por exemplo, pode receber um desconto nos sapatos que visualizou *online* mas que não chegou a comprar.
- **T-MOBILE:** A mineração de dados ajuda a reduzir a taxa de rotatividade de clientes. Ao analisar os dados, a T-Mobile consegue determinar as principais causas de

rotatividade, permitindo-lhes implementar soluções eficazes que manterão mais clientes na empresa.

- **WALMART:** O mecanismo de pesquisa mais recente do retalhista Walmart inclui dados semânticos. O Polaris, *software* desenvolvido internamente, depende de análise de texto e *machine learning* para produzir resultados de pesquisa relevantes. O WalMart refere que com a adição da pesquisa semântica registou-se um aumento na ordem dos 10% e 15% dos compradores *online* que concluíram uma compra. Para o WalMart, esse aumento equivale a ganhos de biliões de dólares.

Verifica-se então que o *data mining* apresenta potencial para ser utilizado em qualquer setor e são claros os benefícios da utilização do mesmo, salientando-se a velocidade de tratamento dos dados e a ausência de ideias pré-concebidas e julgamentos aquando a formulação das informações para a tomada de decisão.

3.1.2 Falácias associadas ao Data Mining

Apesar de serem várias as aplicações que fazem uso de técnicas de *data mining* e existirem inúmeros benefícios na sua utilização, existem também falácias, ideias erradas, que estão constantemente a ser associadas ao *data mining*. Considera-se, portanto, fundamental identificar e esclarecer as falácias que são associadas à mineração de dados. Segundo Larose & Larose (2014) estas são as principais falácias referentes ao *data mining*:

1. Existem ferramentas de *data mining* que podemos executar diretamente nos nossos dados e encontrar respostas para nossos problemas. **Realidade:** Não existem ferramentas automáticas de *data mining*, que automaticamente resolvem os problemas. A mineração de dados é um processo, que utiliza métodos para ajustar esse mesmo processo no plano de ação de negócios ou investigação.

2. O processo de mineração de dados é autónomo, exigindo pouca ou nenhuma supervisão humana. **Realidade**: O *data mining* não é mágico, sem supervisão humana especializada, o uso de *software* de mineração de dados fornecerá a resposta errada à pergunta errada aplicada ao tipo errado de dados. Importa salientar que análises erradas são piores que nenhuma análise, pois levam a recomendações erradas que podem acarretar elevados custos para a organização.

3. O *data mining* tem um retorno de investimento rápido. **Realidade**: As taxas de retorno variam, dependendo dos custos iniciais, custos de pessoal, custos de preparação do *data warehouse*, entre outros.

4. O *software* de mineração de dados é intuitivo e fácil de usar. **Realidade:** Mais uma vez, a facilidade de uso varia. No entanto, independentemente do que alguns *softwares* possam reivindicar, não se pode simplesmente comprar um *software* de *data mining* instalá-lo, sentar-se e vê-lo resolver todos os problemas. Por exemplo, os algoritmos exigem formatos de dados específicos, que podem exigir pré-processamento substancial. Os analistas de dados devem combinar o conhecimento do assunto com uma mente analítica e conhecimento do negócio.
5. A mineração de dados identificará as causas dos problemas nos negócios ou investigação. **Realidade**: O processo de descoberta de conhecimento irá ajudar a descobrir padrões de comportamento. Mais uma vez, cabe aos humanos identificar as causas.
6. O *data mining* limpará automaticamente a confusão que está na nossa base de dados. **Realidade**: Bem, não automaticamente. Numa fase preliminar do processo, a preparação dos dados geralmente lida com dados que não foram examinados ou usados durante muito tempo. Portanto, as organizações que começam uma nova operação de mineração de dados, muitas vezes, são confrontadas com o problema dos dados que têm sido mantidos por anos, estão obsoletos e precisam de ser atualizados.
7. A mineração de dados fornece sempre resultados positivos. **Realidade**: Não há garantia de resultados positivos. A mineração de dados não é uma panaceia para resolver problemas dos negócios. Mas, usado corretamente, por pessoas que entendem os modelos envolvidos, com os dados, requisitos e os objetivos gerais do projeto definidos, a mineração de dados pode fornecer resultados acionáveis e altamente lucrativos.

3.1.3 Tecnologias utilizadas no Data Mining

O conceito de *data mining* está relacionado com a descoberta de padrões nos dados. Para os descobrir, recorre frequentemente a técnicas que pertencem a outros ramos. Segundo Han, Kamber & Pei (2012) o *data mining* incorporou muitas técnicas de outros domínios, como estatística, *machine learning*, reconhecimento de padrões, bases de dados e sistemas de armazenamento de dados, recuperação de informação, visualização, algoritmos, alto desempenho computacional e muitos outros domínios de aplicação. A interdisciplinaridade da natureza da investigação e desenvolvimento da mineração de dados contribui significativamente para o sucesso da mesma nas suas extensas aplicações. Os mesmos autores descrevem as principais disciplinas utilizadas:

- **Estatística:** estuda a recolha, análise, interpretação ou explicação e apresentação de dados. O *data mining* tem uma ligação inerente com a estatística. Um modelo estatístico é um conjunto de funções matemáticas que descrevem o comportamento de objetos numa classe em termos de variáveis aleatórias e suas probabilidades.
- ***Machine Learning*:** investiga como os computadores podem aprender (ou melhorar seu desempenho) baseando-se em dados. Um dos principais objetivos nestas áreas de investigação é fazer com que os programas de computador aprendam automaticamente a reconhecer padrões complexos e tomar decisões inteligentes com base em dados.
- **Bases de Dados e *Data-warehouses*:** a pesquisa em bases de dados é focada na criação, manutenção e uso de bases de dados para organizações e utilizadores. Especialistas em bases de dados desenvolveram um conjunto de princípios reconhecidos em modelos de dados, linguagens de consulta, processamento de consultas e métodos de otimização, armazenamento de dados e métodos de indexação e acesso. Um *data warehouse* integra dados provenientes de múltiplas fontes e vários períodos temporais, consolidando-os no espaço multidimensional para formar cubos de dados.

3.1.4 Linguagens de Programação mais utilizadas em *Data Mining*

Duas das linguagens de programação mais populares no universo da ciência dos dados e consequentemente no *data mining* são *Python* e *R.* Esta afirmação é corroborada pela KDNuggets e pela Kaggle através de questionários realizados à comunidade (Kaggle, 2019; Piatetsky, 2019) onde questionam acerca das ferramentas, linguagens e plataformas de dados que utilizam frequentemente.

Na Figura 3 coloca-se o gráfico que mostra as ferramentas mais utilizadas entre 2017 e 2019, de acordo com os resultados do questionário da KDNuggets. O questionário da Kaggle obteve como resposta à questão "*What programming languages do you use on a regular basis?*" 83,4% das respostas para Python e 35,5% para a linguagem R, num total de 18828 respostas, como é relatado por Hayes (2019).

A Figura 4 mostra que, de acordo com as respostas, existem muitos mais utilizadores a usar apenas Python do que apenas R. Os dados dos dois questionários mostram que o uso do Python prevalece, mas mostram também que é comum utilizar-se mais do que uma linguagem e/ou ferramenta, como se pode validar pela observação de taxas de utilização elevadas de cada item da Figura 3. Os dados de Piatetsky (2019) mostram que nos últimos

anos tem havido algum crescimento na utilização de Python e o inverso na linguagem R, como representado na Figura 5.

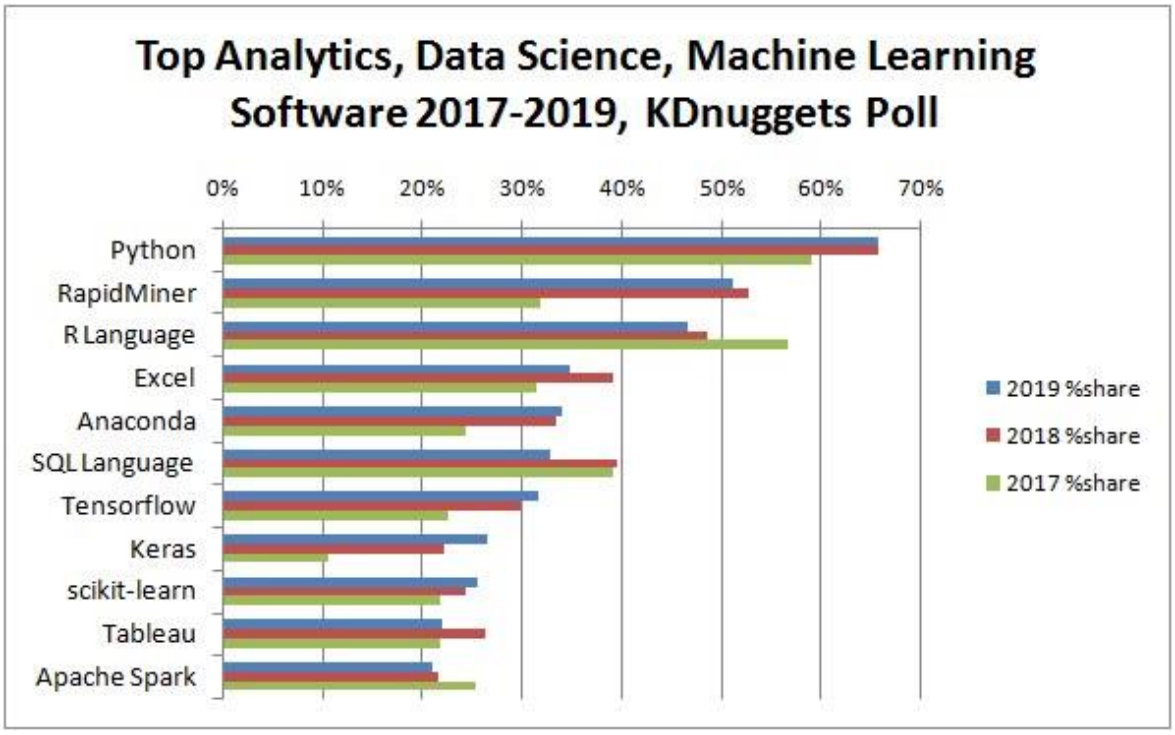

Figura 3 - Ferramentas mais utilizadas em Data Science entre 2017 e 2019

Fonte: Retirado de Piatetsky (2019)

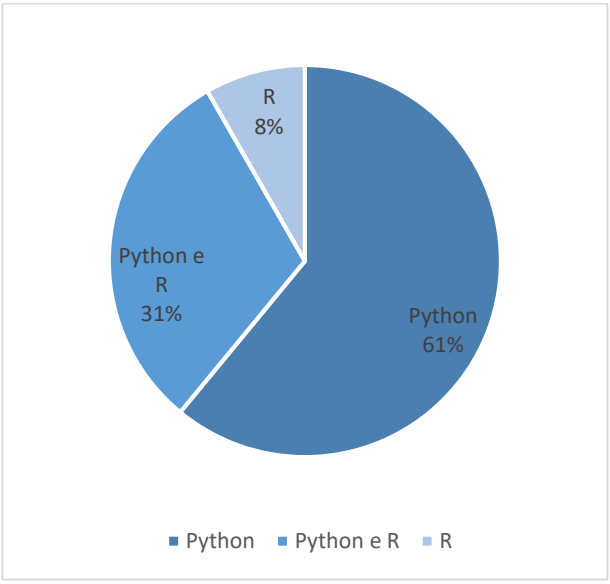

Figura 4- Comparação entre utilização da linguagem de programação R e Python (Kaggle, 2019)

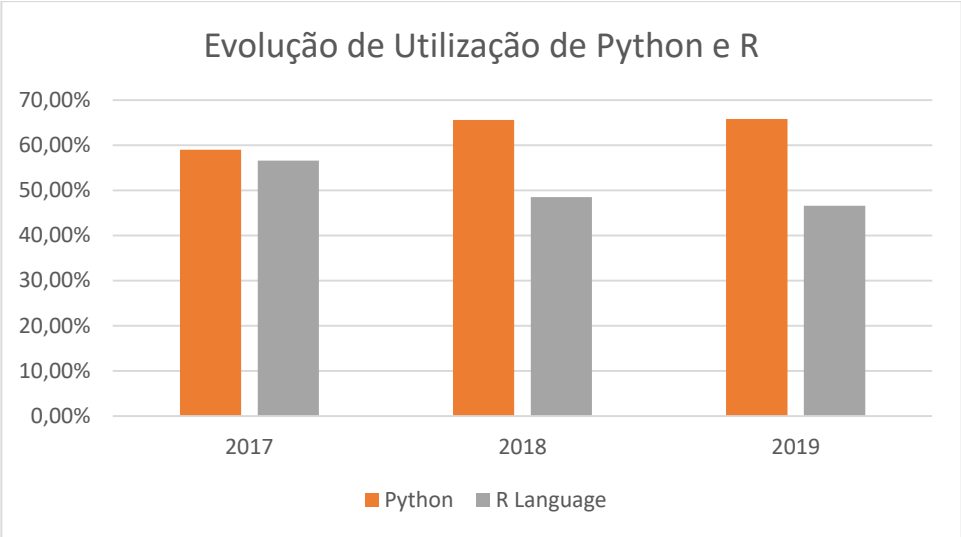

Figura 5- Evolução da utilização da linguagem de programação Python e R (Piatetsky, 2019)

De acordo com o artigo publicado pela Data-Driven Science (2018), a linguagem *Python* surgiu em 1989 e tem como filosofia a legibilidade e eficiência do código. É uma linguagem de programação orientada a objetos, o que significa que agrupa dados e códigos em objetos que podem interagir e modificar uns aos outros. Essa abordagem permite que se realizem tarefas com melhor estabilidade, modularidade e legibilidade de código. Por outro lado, a linguagem *R* foi desenvolvida em 1992, é uma linguagem procedimental que funciona decompondo uma tarefa de programação em uma série de etapas, procedimentos e sub-rotinas, o que é uma vantagem quando se trata de construir modelos de dados, porque torna relativamente fácil entender como operações complexas são realizadas; no entanto, muitas vezes essa vantagem é obtida à custa do pior desempenho e pouca legibilidade do código.

Esse mesmo artigo da Data-Driven Science (2018) efetua uma análise acerca das duas linguagens de programação tendo em consideração etapas frequentemente utilizadas quando se desempenham atividades ligadas à ciência dos dados, tais como a recolha de dados, exploração de dados, modelação de dados e visualização de dados. Infra coloca-se as principais funcionalidades de cada linguagem de programação consoante a tarefa a desempenhar:

- **Recolha de Dados:**
  - *Python* – Suporta diversos formatos de dados, sendo por exemplo possível importar ficheiros CSV (*comma separated values*), JSON (*JavaScript Object Notation),* importar diretamente tabelas SQL para o código, criar conjuntos de dados através de dados recolhidos em diferentes websites utilizando a biblioteca *Python requests library.*

- *R* – Também suporta diversos formatos de dados, sendo por exemplo possível importar ficheiros *Excel*, CSV e de texto e apesar de não ser tão versátil como o *Python* consegue também recolher dados de vários websites utilizando os pacotes *Rvest* e *magrittr*.

- **Exploração de Dados**
  - *Python* – Possibilita a descoberta de *insights* nos dados através da utilização do *Pandas*, a biblioteca de análise de dados do *Python*. Esta consegue conter e tratar grandes quantidades de dados sendo mais rápido o tratamento em comparação com o *Excel*, possibilitando a realização de operações como filtrar, classificar e exibir dados em questão de segundos.
  - *R* – Este *software* foi criado para fazer análises estatísticas e numéricas de grandes conjuntos de dados, por isso apresenta inúmeras opções de exploração de dados tais como: construir distribuições de probabilidade; aplicar uma variedade de testes estatísticos aos dados e usar técnicas padrão de *machine learning* e *data mining*.
- **Modelação de Dados**
  - *Python* – Possibilita a execução de uma análise de modelagem numérica com o *Numpy* e permite fazer cálculos científicos com o *SciPy*. Além disso, disponibiliza uma grande variedade de algoritmos de *machine learning* através da biblioteca *scikit-learn* que apresenta uma interface intuitiva.
  - *R* - Para fazer análises de modelações específicas, por vezes será necessário recorrer a pacotes fora da funcionalidade principal do *R*. Há muitos pacotes disponíveis para análises específicas, como a distribuição de *Poisson* e misturas de leis de probabilidade.
- **Visualização de Dados**
  - *Python* - O *IPython Notebook* que acompanha o *Anaconda* tem muitas opções poderosas para visualizar dados tais como a biblioteca *Matplotlib* para gerar gráficos básicos a partir dos dados incorporados no *Python*. Se o pretendido for um gráfico mais avançados ou melhor design, existe também a biblioteca *Plot.ly*.
  - *R* – Este *software* foi construído com o propósito de realizar análises estatísticas e demonstrar os resultados. Deste modo, fornece um ambiente poderoso adequado para visualização científica com muitos pacotes especializados em exibição gráfica de resultados. O módulo gráfico básico permite criar todos os gráficos básico. Com o *R* é também

possível salvar ficheiros em formatos de imagem, como *jpg*, ou salvá-los como PDFs *(portable document format)* separados. Caso seja necessário utilizar gráficos mais avançados existe a possibilidade de usar o *ggplot2* que cria por exemplo gráficos de dispersão complexos com linhas de regressão.

3.1.5 Técnicas de Data Mining

Na mineração de dados são várias as técnicas que podem ser utilizadas consoante o problema a resolver. Considera-se importante descrever as principais técnicas utilizadas, mas de uma forma resumida. Os autores Azevedo & Santos (2005) identificaram as seguintes técnicas:

- **Classificação:** a classificação tem como objetivo encontrar uma função que associa um caso a uma classe dentro de diversas opções de classificação, de forma a classificar um novo objeto de acordo com o modelo definido. São utilizados conjuntos de treino com exemplos pré classificados com a finalidade de construir modelos adequados à descrição das classes, que posteriormente são aplicados a dados não classificados. Por exemplo, a classe de mamíferos possui atributos que descrevem o mamífero; se um animal satisfizer as propriedades de classificação do mamífero, então esse animal pode ser classificado como um mamífero. A classificação pode ser usada em deteção de fraudes, aplicações de risco, tendências nos mercados financeiros e identificação de objetos em grandes bases de dados.
- **Previsão:** a previsão conjetura valores futuros ou desconhecidos de outras variáveis de interesse com base em algumas variáveis e na descoberta de padrões (variáveis contínuas) a partir de exemplos. Por exemplo, com base nas habilitações académicas, emprego atual e padrões da indústria, pode-se prever o salário a receber no futuro.
- **Regressão linear:** trata de encontrar uma função para uma previsão de uma variável, ou seja, procura uma função que representa de uma forma aproximada comportamentos de variáveis. Os métodos de regressão linear permitem a discriminação dos dados através da combinação dos atributos de entrada, o que equivale a determinar retas de separação dos dados.
- **Segmentação (*Clustering*):** identifica um conjunto finito de categorias ou segmentos para descrever os dados, isto é, identifica grupos homogéneos de objetos em que cada grupo é uma classe, dentro da mesma classe os objetos são semelhantes e entre classes são diferentes. Como exemplos de aplicações que recorrem à técnica de segmentação temos: detetar defeitos no fabrico de produtos,

encontrar grupos de afinidades para cartões de crédito e subgrupos homogéneos de consumidores em bases de dados de *Marketing*.

- **Associação ou Dependência:** esta técnica serve para identificar grupos de dados tipicamente associados e identificar factos que possam ser direta ou indiretamente associados. A associação é normalmente utilizada em aplicações onde se pretende identificar dados que possam ser colocados juntos num mesmo pacote, por exemplo, "Que produtos são normalmente comprados em conjunto?".

3.2 Conhecer os Dados

Independentemente da técnica de *data mining* escolhida, antes de se avançar para a construção do modelo importa conhecer detalhadamente os dados que temos ao dispor. Importa verificar que tipo de atributos têm os dados, como é que estes estão distribuídos e se à vista aparecem alguns valores que se destacam. Além disso, quando os dados chegam à nossa posse tipicamente não apresentam o formato específico para avançar com a aplicação do modelo e temos também frequentemente dados incompletos e/ou com valores em falta. Todo este "ruído" ao redor dos dados tem implicações na construção e nos resultados do modelo, importa, portanto, conhecer muito bem os dados por forma a saber interpretá-los e aplicar as devidas alterações nos mesmos para a construção de modelos eficazes.

Nos próximos pontos são abordados alguns processos para identificação e tratamento dos dados. Estas tarefas estão incluídas na metodologia CRISP, mais propriamente nas tarefas de Conhecer os Dados e Preparação dos Dados.

3.2.1 Aquisição

Os dados utilizados nas aplicações de *data mining* podem ser provenientes de diversas fontes de dados heterogéneas e não apresentarem o formato necessário para a aplicação dos algoritmos de *data mining*. Desta forma, importa saber quais os tipos de variáveis que são mais utilizados pelos algoritmos na mineração de dados para rapidamente identificar os tipos de dados que recolhemos e como os vamos utilizar na aplicação dos algoritmos.

Segundo Chakrabarti *et al* (2009), devemos ter em consideração três tipos de variáveis: Nominais, Ordinais e Contínuas.

- As variáveis nominais, também chamadas de categóricas, descrevem valores que não possuem as propriedades de ordem. Por exemplo, para descrever uma unidade habitacional podemos utilizar as categorias casa, apartamento ou alojamento partilhado. Ao olharmos para estas categorias as mesmas não nos indicam nenhuma ordem ou escala. Do ponto de vista dos algoritmos de *data mining*

é importante que as variáveis nominais não sejam representadas como inteiros pois a aplicação dos algoritmos pode inadvertidamente adicionar números e apresentar resultados erróneos. Ao representar as variáveis nominais em *softwares* as mesmas devem ser definidas como *strings* (sequência de carateres geralmente utilizada para representar palavras, frases ou textos) por forma a evitar estes erros.

- Variáveis ordinais ou de escala ordenada são variáveis categóricas, mas com a noção de ordem associada, ou seja, conseguimos identificar a ordem, mas não a escala. Como exemplo temos os níveis de risco associados a padrões de pagamentos com cartão de crédito definidos como alto, médio e baixo. Facilmente compreende-se a ordem, alto é maior que médio e médio é maior que baixo. Contudo, não se consegue afirmar que a diferença entre alto e médio é a mesma entre médio e baixo em termos de escala. Aplicando o mesmo raciocínio utilizado nas variáveis nominais, percebe-se que também que estas variáveis não devem de ser representadas como inteiros.
- As variáveis contínuas ou reais são as mais fáceis de utilizar e interpretar pois têm todas as propriedades desejáveis das variáveis: ordem, escala e distância. Além disso, possuem também os significados de zero e valores negativos definidos. Medidas reais são representadas por números reais com uma razoável precisão.

Tendo em mente os três tipos de variáveis descritos acima, uma das primeiras tarefas a realizar quando adquirimos os dados passará então por decidir sobre o tipo de dados a serem utilizados para cada variável.

3.2.2 Pré-Processamento

A fase do pré-processamento dos dados é fundamental no processo de mineração de dados pois se os dados tiverem pouca qualidade levarão a resultados também de pouca qualidade. Hoje em dia, com as enormes bases de dados e as inúmeras fontes de dados heterogéneas, deparamo-nos frequentemente com dados em falta, incompletos e até mesmo inconsistentes. É por estas razões que importa antes de avançarmos para a aplicação dos algoritmos com os dados que temos, analisar e pré-processar os dados para garantir a qualidade dos mesmos e deste modo conseguirmos obter resultamos com mais qualidade e credíveis.

Nos pontos abaixo abordam-se as técnicas que podem ser utilizadas na fase de pré-processamento dos dados.

3.2.2.1 Sumarização Descritiva

Os autores Chakrabarti *et al* (2009), referem que de modo a que o pré-processamento dos dados seja bem-sucedido, é importante ter uma visão geral dos mesmos. As técnicas de

sumarização descritiva ajudam a identificar as propriedades típicas dos dados e permitem através da sumarização dos dados, identificar rapidamente ruídos e valores que se destaquem (*outliers*). Podemos saber mais acerca dos dados se aplicarmos medidas de tendência central e dispersão dos dados. Medidas de tendência central incluem média, mediana e moda, enquanto medidas de dispersão de dados incluem quartis e variância.

De acordo com Larose & Larose (2014), as medidas do centro são um caso especial de medidas de localização, resumos numéricos que indicam onde, numa sequência de números, uma certa característica da variável reside. Para determinar a média aritmética, basta somar todos os valores do campo e dividir pelo tamanho da amostra. Para variáveis que não são extremamente distorcidas, a média geralmente não está muito distante do centro da variável. No entanto, para conjuntos de dados extremamente distorcidos, a média torna-se menos representativa do centro da variável. Além disso, a média é sensível à presença de *outliers* (valor que apresenta um grande afastamento em relação aos demais da amostra). Por esse motivo, os analistas utilizam também outras medidas alternativas de centro, como a mediana, definida como o valor do campo no meio quando os valores são classificados em ordem crescente. A mediana é resistente à presença de *outliers*. Outros analistas podem preferir usar a moda, que representa o valor que ocorre com mais frequência. A moda pode ser usada com um valor numérico ou dados categóricos, mas nem sempre está associado ao centro da variável.

Importa referir que por si só as medidas de tendência central não são suficientes para resumir uma variável, podemos ter casos em que temos as mesmas médias e os valores estarem mais dispersos na situação **A** e menos dispersos na situação **B**. Desta forma, é também importante verificar a dispersão dos dados.

Os autores Han, Kamber & Pei (2012), referem que os percentis mais utilizados além da mediana são quartis. O primeiro quartil, denotado por **Q1**, é o percentil 25; e o terceiro quartil, denotado por **Q3**, é o percentil 75. Os quartis, juntamente com a mediana, dão alguma indicação do centro, da distribuição e da forma de uma distribuição. Outra medida de dispersão frequentemente utilizada é o desvio padrão, este é a raiz quadrada da variância e mede a dispersão sobre a média e apenas deve ser utilizado quando a média é escolhida como medida de tendência central.

3.2.2.2 Limpeza dos Dados

A probabilidade de termos dados incompletos, inconsistentes ou em falta, é bastante elevada, pelo que é necessário proceder à limpeza dos mesmos por forma a garantir resultados de qualidade aquando da aplicação das técnicas de mineração de dados.

Segundo Aggarwal, (2015) as principais tarefas associadas a limpeza dos dados são o tratamento de dados em falta e o tratamento de dados incorretos.

Relativamente ao tratamento dos dados em falta, Aggarwal (2015) refere que existem três tipos de técnicas que podem ser utilizadas, estas são:

- Qualquer registo de dados que contenha um valor em falta pode ser totalmente eliminado. Salienta-se que esta abordagem pode não ser prática quando a maioria dos registos contém valores em falta.
- Os valores em falta podem ser estimados ou imputados. No entanto, os erros criados pelo processo de imputação podem afetar os resultados do algoritmo de mineração de dados.
- A fase analítica é projetada de tal forma que está preparada para funcionar com valores ausentes. Esta abordagem é geralmente a mais desejável, porque evita alguns dos erros associados ao processo de imputação.

Para o tratamento de dados incorretos ou inconsistências, Aggarwal (2015) indica que se pode recorrer a três métodos:

- **Deteção de Inconsistências:** acontece normalmente quando os dados estão disponíveis em diferentes fontes em diferentes formatos. Por exemplo, o nome de uma pessoa pode ser escrito na íntegra numa fonte, enquanto a outra fonte pode conter apenas as iniciais e um sobrenome.
- **Domínio de Conhecimento:** importa ter-se um bom conhecimento acerca do negócio e dos seus atributos e regras. Por exemplo, se o campo do país for "Estados Unidos", o campo cidade não pode ser “Shanghai”.
- **Métodos Centrados em dados**: o comportamento estatístico dos dados é usado para detetar *outliers*. Estes podem ter surgido devido a erros no processo de recolha, mas podem também significar comportamentos interessantes do sistema. Portanto, qualquer valor atípico detetado pode precisar ser examinado manualmente antes de ser descartado. Desta forma, o uso de métodos centralizados em dados para limpeza dos dados pode, às vezes, ser perigoso porque pode resultar na remoção de conhecimento útil.

Verifica-se que na fase do pré processamento dos dados a tarefa de limpeza dos dados terá um papel fundamental na garantia de resultados com qualidade. Face à proveniência de dados de diversas fontes com diversos formatos, esta tarefa será realizada na maior parte dos casos pois existirá uma elevada probabilidade de existirem valores em falta ou incompletos.

3.2.2.3 Integração

Os autores Monem, El-Bastawissy & Elwakil (2016) indicam que a etapa de integração dos dados é o processo de recolha de dados de diferentes fontes com o objetivo de apresentar ao analista uma visão unificada dos dados que apresenta as respostas aos requisitos inicialmente definidos. A qualidade das respostas estará intrinsecamente ligada com a qualidade dos dados nas fontes e poderá ser aumentada se durante o processo de integração forem aplicadas medidas que garantam a qualidade dos dados e sejam integrados dados que sejam significativos para o problema em questão.

Segundo Monem, El-Bastawissy & Elwakil (2016) os indicadores da qualidade dos dados que podem afetar o processo de integração dos dados e que os analistas devem de ter em consideração os seguintes:

- **Preenchimento dos dados:** necessário ter em atenção se existem dados em falta ou com valores a *null* para os dados que se estão a analisar e garantir que todos os dados que são necessários para a análise do problema se encontram disponíveis e podem ser integrados.
- **Validade dos dados:** ao analisar-se os dados tem de se garantir que estes são válidos tendo em consideração as especificidades do negócio. Se para o problema em questão os dados não forem válidos não se devem integrar pois iriam diminuir a qualidade do modelo.
- **Precisão dos dados:** garantir que os dados recolhidos das diferentes fontes apresentam valores aproximados dos valores esperados. Caso existam *outliers*, ou seja, os dados recolhidos não são aproximados aos valores esperados, será necessário realizar uma análise mais aprofundada para garantir que não se trata de anomalias geradas na altura da recolha dos dados ou mesmo na inserção dos mesmos nas respetivas fontes.
- **Período temporal dos dados:** é importante saber a que período temporal se referem os dados, importa saber não só a data de recolha dos dados como também a data em que foram criados. Exemplo, os dados existentes no conjunto de dados que se está a analisar podem datar de 2012 a 2019 e serem integrados apenas os dados referentes ao período de 2015 a 2018.

Por sua vez, Chakrabarti *et al.* (2009) explicam através de um exemplo uma integração que pode correr mal pois os analistas podem não ter forma de se certificarem que o ID do cliente na Base de Dados **A** é igual ao número do cliente na Base de dados **B** e estão a ser recolhidos dados de ambas as bases de dados. A solução passa pela utilização de metadados para cada atributo, onde são incluídos nome, significado, tipo de dados,

intervalo de valores permitido para o atributo e regras nulas para manipulação de valores em branco, zero ou nulos. Esses metadados devem ser usados para ajudar e evitar erros na integração do esquema. Outra situação a ter em consideração quando se faz integração dos dados é a redundância. Um atributo pode ser redundante se puder ser "derivado" de outro atributo ou conjunto de atributos. Inconsistências no nome de atributos ou dimensões também podem causar redundâncias.

Os mesmos autores referem também que outra questão importante na integração de dados é a deteção e resolução de conflitos de valores de dados. Por exemplo, para a mesma entidade do mundo real, valores de atributo de diferentes fontes podem diferir. Isto pode ser devido a diferenças na representação, dimensionamento ou codificação. Por exemplo, um atributo de peso pode ser armazenado em unidades métricas em um sistema e unidades imperiais britânicas em outro. Assim sendo, ao combinar atributos de uma base de dados para outra durante a integração, deve ser dada especial atenção à estrutura dos dados. Isto para garantir que quaisquer dependências funcionais de atributos e restrições referenciais na origem correspondem àqueles no sistema de destino.

Constata-se então que uma boa integração dos dados irá ajudar a obter resultados com mais qualidade e melhorará também a velocidade de processamento dos dados pois já foram previamente detetadas e eliminadas possíveis redundâncias.

3.2.2.4 Transformação

A transformação dos dados é a etapa responsável por transformar os dados que temos ao nosso dispor em dados que possam ser utilizados pelos algoritmos de *data mining*. Segundo García, Luengo & Herrera (2015), a transformação dos dados combina os atributos originalmente recolhidos com fórmulas matemáticas aplicáveis ao negócio.

Os mesmos autores mencionam que na aplicação de algoritmos de *data mining*, por vezes os dados recolhidos não são úteis no seu estado atual pois foram desenhados a pensar nos sistemas em que atualmente operam e não em aplicações de algoritmos de *data mining*, sendo por isso necessário manipulá-los para os transformar ou criar novos atributos baseados nos originais.

Existem inúmeras transformações que podem ser aplicadas aos dados, vão-se mencionar apenas algumas que podem ser executadas. Uma das principais técnicas utilizadas na transformação dos dados é a normalização. Segundo García, Luengo & Herrera (2015) esta técnica não gera novos atributos, mas transforma os originais em novos conjuntos de valores com as propriedades necessárias para a aplicação dos algoritmos de *data mining*. Estes autores indicam que a normalização tem como objetivo atribuir uma nova escala, ou

seja, um novo intervalo composto por um valor mínimo e um valor máximo, a todos os valores numéricos de determinado atributo.

Outra transformação referenciada pelos mesmos autores é a agregação de informação contida em vários atributos, como exemplos de agregação temos as vendas de um determinado negócio, estas podem ser agregadas mostrando valores diários, mensais, anuais, etc. A atribuição de *ranking* nos atributos mencionada por García, Luengo & Herrera (2015) permitirá também estabelecer uma ordem e hierarquias nos atributos.

São várias as transformações que podem ser realizadas aos dados e estas dependem sempre do problema que se está a estudar e do algoritmo de *data mining* que se pretende utilizar. Cabe ao analista analisar os dados que tem ao dispor e escolher quais os tipos de transformações que pretende executar.

3.2.2.5 Redução

São várias as vezes em que os conjuntos de dados que temos de analisar são de enormes dimensões e se aplicássemos os algoritmos de *data mining* ao conjunto total de dados a velocidade de processamento dos dados e os tempos de respostas obtidos seriam impraticáveis não trazendo qualquer vantagem às organizações. Desta forma, quando dispomos de conjuntos de dados com grandes dimensões por vezes temos de aplicar a técnica da redução de modo a serem aplicados os respetivos algoritmos de *data mining*.

De acordo com Chakrabarti *et al* (2009), as técnicas de redução de dados podem ser aplicadas para obter uma representação reduzida do conjunto de dados que é muito menor em volume, mas mantém de perto a integridade dos dados originais. A mineração no conjunto de dados reduzido deve ser mais eficiente e, mesmo assim, produzir os mesmos (ou quase os mesmos) resultados analíticos. As seguintes estratégias para a aplicação da redução de dados são as seguintes:

- Agregação de cubo de dados, em que as operações de agregação são aplicadas aos dados na construção de um cubo de dados.
- Seleção de subconjunto de dados, onde atributos ou dimensões irrelevantes ou redundantes podem ser detetados e removidos.
- Redução de dimensionalidade, onde os mecanismos de codificação são usados para reduzir o tamanho do conjunto de dados.
- Redução da numerosidade, em que os dados são substituídos ou estimados por alternativa, com representações de dados menores, como modelos paramétricos (que precisam ser armazenados apenas os parâmetros do modelo em vez dos dados reais) ou métodos não paramétricos como *clustering*, amostragem e uso de histogramas.

- Discretização e geração de hierarquia de conceitos, em que valores de dados brutos para os atributos são substituídos por intervalos ou níveis conceituais mais altos.
- A amostragem pode também ser usada como uma técnica de redução de dados porque permite conjunto de dados serem representado por uma amostra de dados aleatórios muito menor (ou subconjunto).

3.3 **Regras de Associação**

De acordo com Yuan (2017), a primeira aplicação da regra de associação data de 1994 quando Agrawal e Srikan analisaram a performance de compras num supermercado desenvolvendo o algoritmo Apriori (Agrawal & Srikant, 1994). A autora refere também que a partir desse momento a utilização da regra de Associação em *data mining* passou a desempenhar um papel importante não só na análise de dados comerciais como também noutros setores.

Segundo Yuan (2017), a regra de Associação faz parte de um dos ramos mais importantes na mineração de dados, identificando as associações e padrões entre itens num determinado conjunto de dados. Tem como objetivo descobrir itens que são registados frequentemente em conjunto tendo em consideração um determinado patamar e gerar regras de associação que cumpram as restrições definidas previamente.

Os autores Lai & Lu (2018) referem que a regra de associação descobre a probabilidade da coocorrência de itens, ativos ou objetivos num determinado conjunto de dados. Os resultados exibem relações entre itens, ativos e objetivos coocorrentes. A mineração de regras de associação é um dos métodos mais utilizados para detetar e extrair informações úteis de dados em larga escala, podendo revelar várias relações de associação.

De acordo com Nidhi et al. (2018), a mineração de regras de associação destina-se a encontrar conjuntos de itens, correlações e associações de vários tipos de bases de dados, como bases de dados relacionais, bases de dados transacionais, bases de dados sequenciais, entre outras. A principal aplicação das regras de associação é a *market basket analysis, ou* análise do cesto/carrinho de compras. A regra de Associação pode ser definida como X → Y, onde X, Y são os *itemsets* antecedente e consequente, respetivamente, significando que a presença de X está associada à presença de Y, em simultâneo, no cesto de compras, com uma determinada probabilidade.

Os autores Kabir, Ludwig & Abdullah, (2018) indicam que matematicamente uma regra de associação é definida como A → B onde A (antecedente) e B (consequente) são predicados

lógicos construídos por predicados booleanos. Um predicado lógico numa regra de associação consiste numa ou mais condições Booleanas e estas estão interligadas pelo operador lógico *AND* (∧). Num conjunto de dados transacionais (por exemplo, conjunto de talões de compra de um supermercado), podemos ter uma regra de associação como (item = leite) ∧ (item = pão) ⇒ (item = manteiga), o que significa que quando um cliente compra leite e pão é mais provável que ele também compre manteiga.

Constata-se então que a regra da Associação é uma técnica bastante utilizada em *data mining*, sendo o seu principal foco a análise de dados comerciais, podendo ser também aplicada noutros setores. O seu principal objetivo passa por identificar nos dados disponíveis os itens que frequentemente são adquiridos em conjunto por forma a posteriormente conseguir calcular a probabilidade de determinado item C ser adquirido se no nosso cesto de compras já tivermos adquirido o item A e B.

Nos pontos seguintes será explicado em detalhe como são geradas as regras de associação na *market basket analysis*.

3.3.1 Market Basket Analysis

Segundo Gayathri (2017), o termo *market basket analysis* (MBA) aplicado ao setor do retalho refere-se às informações que se conseguem proporcionar aos retalhistas acerca do comportamento dos seus clientes. Estas informações ajudarão os retalhistas a perceber melhor as necessidades dos seus clientes, planeando ações de *marketing* por forma a manter e atrair novos clientes. Os *layouts* das lojas poderão ser planeados de acordo com os comportamentos dos clientes e as campanhas publicitárias poderão ser personalizadas de acordo com as necessidades/hábitos dos clientes. A título de exemplo o autor refere que clientes que normalmente compram pão compram também leite. Assim sendo, ao colocar-se o pão perto da área do leite as probabilidades de venda de ambos os itens aumentam drasticamente.

Para o autor Kantardzic (2011) um cesto de compras é um conjunto de itens adquiridos por um cliente numa única transação. Os retalhistas armazenam o histórico de todas as transações ocorridas, pelo que é comum analisar-se essas transações por forma a se tentar descobrir novas informações. Uma das análises mais frequente quando se está a olhar para os registos das transações efetuadas é a identificação do conjunto de itens que aparecem frequentemente em simultâneo nas transações, a esta análise dá-se o nome de *market basket analysis.*

De acordo com Kantardzic (2011), a descoberta de itens que são adquiridos frequentemente em conjunto não é um problema simples de resolver, porque normalmente o número de clientes e transações registados na base de dados é bastante elevado e não

é possível de ser processado através de uma memória central de um computador. O outro obstáculo é que o potencial número de itens adquiridos em conjunto é exponencial ao número de itens diferentes que existem, apesar do número de itens adquiridos em conjunto poder ser inferior.

Os autores Solnet, Boztug, & Dolnicar (2016) indicam que a ideia base por detrás da MBA está assente na premissa que os consumidores raramente tomam decisões de compra isoladas, sendo que raramente adquirem um produto por compra, preferindo comprar uma cesta completa de produtos, geralmente produtos de diferentes categorias. O uso de informações sobre as cestas de mercado permite que se saiba não só apenas quais produtos e categorias de produtos tendem a ser comprados juntos, mas também determinam quais produtos ou categorias de produtos que são fatores determinantes para a compra de certos produtos. Esse conhecimento permite que os gestores desenvolvam estratégias com vista a influenciar o comportamento de compra, incluindo promover a procura de determinados produtos, promover categorias específicas de produtos ou oferecer promoções para produtos que tenham influência na compra de outros produtos e que provavelmente aumentarão os gastos gerais por compra.

A MBA é uma técnica utilizada principalmente para descobrir itens adquiridos em conjunto no setor do retalho que recorre à aplicação de regras de associação.

Por forma a exemplificar o funcionamento da *market basket analysis* coloca-se infra o exemplo descrito pelos autores Larose & Larose (2014).

Supondo que um agricultor tem para venda os seguintes itens: espargos, feijões, brócolos, milho, pimentos, abóbora e tomates. Doravante, este conjunto de itens passará a ser descrito como *I.* Ao longo do dia o agricultor registou várias vendas, subconjuntos do conjunto *I.* Na Tabela 2 apresenta-se uma lista com as transações efetuadas, descrevendo os itens adquiridos e o número de cada transação.

Sendo *D* o conjunto de transações representadas na Tabela 2 onde cada transação *T* em *D* representa um conjunto de itens presentes em *I* e supondo que temos um conjunto de itens *A* {feijões e abóbora} e um conjunto de itens *B* {espargos} então pode-se originar uma regra de associação, "*Se A Então B" (A →B)* onde o antecedente *A* e o consequente *B* são subconjuntos de *I,* sendo *A* e *B* exclusivos entre si.

Duas medidas fundamentais na aplicação da regra de associação são o suporte e a confiança.

Tabela 2- Transações registadas ao longo do dia pelo agricultor

Fonte: Adaptado de (Larose & Larose, 2014) tabela 12.1 pág 249

Número da Transação	Itens Adquiridos
1	Brócolos, pimentos, milho
2	Espargos, abóbora, milho
3	Milho, tomates, feijões, abóbora
4	Pimentos, milho, tomates, feijões
5	Feijões, espargos, brócolos
6	Abóbora, espargos, feijões, tomates
7	Tomates, Milho
8	Brócolos, tomates, pimentos
9	Abóbora, espargos, feijões
10	Feijões, milho
11	Pimentos, brócolos, feijões, abóbora
12	Espargos, feijões, abóbora
13	Abóbora, milho, espargos, feijões
14	Milho, pimentos, tomates, feijões, brócolos

Segundo Larose & Larose (2014) o suporte *s* de uma regra de associação *A→B* é calculado pela proporção de transações em *D* que contêm tanto *A* como *B.* Ou seja,

$$suporte = P(A \cap B) = \frac{número\ de\ transações\ que\ contêm\ A\ e\ B}{número\ total\ de\ transações}$$

O autor Aggarwal (2015) refere também que os itens correlacionados irão frequentemente aparecer em conjunto nas transações e apresentarão elevados valores de suporte. As regras de associação serão geradas tendo em consideração um valor mínimo de suporte, assim a definição desse valor terá um impacto significativo nos resultados pois se for utilizado um valor baixo para o suporte mínimo irá ser gerado um maior numero de itens e se for colocada uma fasquia muito alta para o suporte mínimo corre-se o risco de não se encontrar padrões frequentes.

Os mesmos autores referem que por sua vez a confiança *c* de uma regra de associação *A→B* serve para medir a precisão da regra e é calculada através da percentagem de transações em *D* que contêm *A* mas que também contêm *B.* Matematicamente,

$$confiança = P(B \setminus A) = \frac{P(A \cap B)}{P(A)} = \frac{número\ de\ transações\ que\ contêm\ A\ e\ B}{número\ de\ transações\ que\ contêm\ A}$$

De acordo com Larose & Larose (2014), a geração de regras de associação fortes estará intrinsecamente ligada aos valores definidos de suporte e confiança. A definição destes valores estará a cargo do analista que consoante o problema que pretende estudar irá determinar qual a percentagem de suporte e confiança que pretende. Se o analista estiver interessado em descobrir quais os itens que são comprados em conjunto num supermercado pode definir como suporte mínimo 20% e como confiança mínima 70%. Contudo, se o problema que se pretende analisar estiver relacionado com deteção de fraudes ou de ataques terroristas o nível de suporte terá que ser drasticamente reduzido pois níveis de 1% ou menos pois serão poucas as transações fraudulentas ou relacionadas com o terrorismo.

Para se perceber como funciona a aplicação da regra da Associação importa definir mais alguns conceitos.

Larose & Larose (2014) referem que um *itemset* é um conjunto de itens presentes em *I* e que um *k-itemset* é um *itemset* que contêm k itens. Ou seja, o conjunto {feijões, abóboras} é um *2-itemset* e o conjunto {Brócolos, pimentos, milho} é um *3-itemset,* cada conjunto pertencente ao conjunto *I* do exemplo do agricultor. A frequência do *itemset* é calculada pela soma de transações que contêm esse *itemset* específico*.* Um *itemset* frequente é um *itemset* que ocorre pelo menos um certo número mínimo de vezes, tendo como frequência de *itemset* $\geq \emptyset$. Supondo que é definido que $\emptyset = 4$, então os *itemsets* que ocorrerem mais de quatro vezes são considerados frequentes. Os *itemsets* que são frequentes são identificados como F_k.

Para a aplicação da técnica de *data mining* referente à regra de associação, Larose & Larose (2014) indicam que é um processo realizado em dois passos:

1. Descobrir todos os *itemsets* frequentes, ou seja, todos os *itemsets* com frequência $\geq \emptyset$.
2. Dos *itemsets* frequentes gerar regras de associação que satisfaçam o suporte e a confiança mínima previamente definida.

Para descobrir todos os *itemsets* frequentes pode-se recorrer a vários algoritmos, um dos algoritmos mais utilizado e popular é o algoritmo Apriori que se descreve no próximo ponto.

3.3.1.1 Algoritmo Apriori

Segundo Singh, Garg & Mishra (2018), o algoritmo Apriori foi proposto por Agrawal e Srikant (em (Agrawal & Srikant, 1994)) e é um dos mais conhecidos e utilizados em *data mining,* principalmente quando se pretende descobrir conjuntos de itens que são adquiridos frequentemente em conjunto. O algoritmo Apriori é o algoritmo base da Regra de

Associação e a sua criação foi responsável por impulsionar a investigação em *data mining*. Em 2006 a IEEE *International Conference on Data Mining (ICDM)* colocou este algoritmo no top 10 dos algoritmos de *data mining* com influência na comunidade científica.

O autor Aggarwal (2015) refere que o algoritmo Apriori utiliza a propriedade "*Downward Closure*" para eliminar espaço de pesquisa de candidatos a itens frequentes. Assim, se um conjunto de itens for identificado como pouco frequente, não existem vantagens em continuar a analisar esse conjunto para geração de candidatos, elimina-se esse conjunto de itens, evitando-se contagens de níveis de suporte desnecessárias pois tratam-se de itens não frequentes. O algoritmo Apriori gera primeiro os candidatos com menor comprimento k e contabiliza os seus suportes antes de gerar os candidatos de comprimento (k+1). Os k-*itemsets* frequentes resultantes são utilizados para restringir o número de (k+1) – candidatos com a propriedade "*Downward Closure*". Uma vez que a parte da contagem de candidatos que tenham determinado suporte é a que requer maior capacidade computacional na geração de padrões frequentes, é importante termos um baixo número de candidatos, pois quanto maior for o número maior será a capacidade computacional exigida.

Os autores Larose & Larose (2014) fazem também referência à propriedade do algoritmo Apriori, indicando que se um conjunto de itens Z não for frequente adicionar outro item A ao conjunto de itens Z não tornará Z mais frequente, ou seja, se Z não é frequente, $Z \cup A$ também não será frequente, nem nenhum superconjunto de Z (conjunto de itens contendo Z). Esta propriedade é bastante útil pois permite reduzir significativamente o espaço de pesquisa.

Por forma a demonstrar como o algoritmo Apriori funciona, vai-se retomar o exemplo do ponto anterior indicado por Larose & Larose (2014) e utilizar as transações registadas na Tabela 2.

Segundo Larose & Larose (2014), considerando a lista de transações *D* presentes na Tabela 2 e considerando que um *itemset* é considerado frequente em *D* se ocorrer 4 ou mais vezes, ou seja, $\emptyset \geq 4$, o primeiro passo é encontrar F_1, os 1-*itemsets* frequentes (conjuntos com um produto), que representam, neste caso, o próprio vegetal. Para auxiliar nos cálculos, criou-se a Tabela 3 que identifica os vegetais e o número das transações registando para cada transação se determinado vegetal foi ou não adquirido.

Tabela 3 - Dados em formato tabular referentes às transações registadas ao longo do dia pelo agricultor

Fonte: Adaptado de Larose & Larose (2014) tabela 12.3 pág 250

Número da Transação	Espargos	Feijões	Brócolos	Milho	Pimentos	Abóbora	Tomate
1	0	0	1	1	1	0	0
2	1	0	0	1	0	1	0
3	0	1	0	1	0	1	1
4	0	1	0	1	1	0	1
5	1	1	1	0	0	0	0
6	1	1	0	0	0	1	1
7	0	0	0	1	0	0	1
8	0	0	1	0	1	0	1
9	1	1	0	0	0	1	0
10	0	1	0	1	0	0	0
11	0	1	1	0	1	1	0
12	1	1	0	0	0	1	0
13	1	1	0	1	0	1	0
14	0	1	1	1	1	0	1
SOMA	6	10	5	8	5	7	6

Verifica-se então que todos os 1-*itemsets* são frequentes pois ocorrem mais de 4 vezes para as transações observadas. De acordo com Larose & Larose, (2014) F_1 = {espargos, feijões, brócolos, milho, pimentos, abóbora , tomates}. O próximo passo será identificar os 2-*itemsets*, isto é feito criando um conjunto C_k de candidatos k-*itemsets* unindo F_{k-1} consigo próprio. O conjunto C_k posteriormente é podado utilizando a propriedade do algoritmo Apriori. O conjunto de items de C_k que não forem eliminados passam a formar $F_{k..}$ Na Tabela 4 são identificadas todas as possíveis combinações de dois itens e contabilizadas as vezes que ocorrem nas transações.

Tendo em consideração $\emptyset = 4$, F_2 = {{espargos, feijões}, {espargos, abóbora}, {feijões, milho}, {feijões, abóbora}, {feijões, tomates}, {brócolos, pimentos}, {milho, tomates}}. Continuando com a explicação dos autores Larose & Larose (2014), o próximo passo passará por utilizar o conjunto de itens frequentes F_2 para gerar $C_{3,}$ o candidato 3-*itemset*. Tal como foi feito para a geração do C_2, para a geração do C_3 unimos F_2 consigo próprio onde os *itemsets* são unidos se tiverem o primeiro k-1 items em comum. Por exemplo, {espargos, feijões} e {espargos, abóbora} têm o primeiro k-1 = 1 em comum – espargos.

Deste modo são unidos gerando um novo candidato composto por {espargos, feijões, abóbora}. o mesmo ocorre com {feijões, milho} e {feijões, abóbora} que geram o candidato {feijões, milho, abóbora}. Os candidatos {feijões, milho, tomate} e {feijões, abóbora, tomate} são também gerados de igual modo. Assim sendo temos C_3 = {{espargos, feijões, abóbora}, {feijões, milho, abóbora}, {feijões, milho, tomate}, {feijões, abóbora, tomate}}.

Tabela 4 - Candidatos a 2-itemset

Fonte: Adaptado de Larose & Larose, (2014) tabela 12.4 pág 252

Combinação	Número de vezes que ocorre	Combinação	Número de vezes que ocorre
Espargos, feijões	5	**Brócolos, milho**	2
Espargos, brócolos	1	**Brócolos, pimentos**	4
Espargos, milho	2	**Brócolos, abóbora**	1
Espargos, pimentos	0	**Brócolos, tomate**	2
Espargos, abóbora	5	**Milho, pimentos**	3
Espargos, tomates	1	**Milho, abóbora**	3
Feijões, brócolos	3	**Milho, tomate**	4
Feijões, milho	5	**Pimentos, abóbora**	1
Feijões, pimentos	3	**Pimentos, tomates**	3
Feijões, abóbora	6	**Abóbora, tomates**	2
Feijões, tomate	4		

Mais uma vez, seguindo as explicações de Larose & Larose (2014), vai-se efetuar a poda do novo conjunto gerado, C_3. para cada item *s* em C_3 são gerados e examinados conjuntos k-1. Se algum destes conjuntos não forem frequentes, *s* não pode ser frequente, logo será eliminado. Como exemplo temos *s*= {espargos, feijões, abóbora}, o subconjunto de tamanho *k-1*= 2 é gerado da seguinte forma: {espargos, feijões}, {espargos, abóbora}, {feijões, abóbora}. Analisando a Tabela 4 verifica-se que todos os subconjuntos são frequentes e desta forma C_3 não será podado. O mesmo acontecerá para {feijões, milho, tomate}.

Continuando a análise de Larose & Larose (2014), desta feita para o conjunto {feijões, milho, abóbora}, verifica-se que o subconjunto {milho, abóbora} tem uma frequência de 3 que é inferior a 4 que foi a medida definida para estabelecer quando um determinado item seria considerado frequente. Tendo em conta a propriedade Apriori o conjunto {feijões, milho, abóbora} não pode ser considerado frequente, logo não constará em F_3. O mesmo acontece com o conjunto {feijões, abóbora, tomate}.

Por último, contabilizando o *itemset* {espargos, feijões, abóbora} constatamos que ocorre quatro vezes na lista de transações e o *itemset* {feijões, milho, tomate} ocorre apenas três vezes. Assim sendo, o *itemset* {feijões, milho, tomate} será também eliminado, sendo o *itemset* frequente F_3= {espargos, feijões, abóbora}.

A tarefa de se encontrar os itens que ocorrem em conjunto frequentemente para o exemplo da venda de vegetais do agricultor encontra-se terminada. Contudo, o processo ainda não terminou, estando em falta a geração das Regras de Associação.

3.3.1.2 Exemplo de Geração de Regras de Associação

Após encontrar o *itemset* frequente o próximo passo é gerar as regras de associação. Segundo Larose & Larose (2014) os passos a tomar para a geração de regras de associação são os seguintes:

- Geração de todos os subconjuntos de *s*, onde *s* é o *itemset* frequente;
- Representar um subconjunto não vazio de *s* através de *ss.* Considerar a regra de Associação R: *ss* → (*s-ss*), onde (*s-ss*) indica o conjunto s sem ss. Gerar R se R preenche o requisito mínimo do parâmetro de confiança e fazer esta ação para cada subconjunto de ss de s.

Retomando o exemplo dado no ponto anterior, verifica-se que o *itemset* frequente identificado foi F_3= {espargos, feijões, abóbora}. Segundo Larose & Larose (2014), os subconjuntos de *s* são: {espargos}, {feijões}, {abóbora}, {espargos, feijões}, {espargos, abóbora}, {feijões, abóbora}. Ao considerar a regra R onde *ss* = {espargos, feijões} então (*s –ss)* = {abóbora},temos {espargos, feijões} → {abóbora}. Para esta regra, o suporte apresenta o valor de 28,6%, pois das 14 transações registadas a combinação {espargos, feijões, abóbora} ocorreu em conjunto 4 vezes. O valor da confiança é de 80% pois a combinação {espargos, feijões} ocorreu 5 vezes das 14 transações e das 5 vezes que ocorreu verificou-se que 4 transações também continham o item {abóbora}, ou seja 4/5 = 80%. A Tabela 5 identifica as regras de associação nas situações de dois antecedentes. assumindo que a confiança mínima pretendida é de 60% e que queremos apenas um consequente.

Tabela 5 - Regras de Associação para dois Antecedentes, exemplo da venda de vegetais
Fonte: Adaptado de Larose & Larose (2014) tabela 12.5 pág 253

Se Antecedente, então Consequente	Suporte	Confiança
{espargos, feijões} → {abóbora}	4/14 = 28.6%	4/5 = 80%
{espargos, abóbora} → {feijões}	4/14 = 28.6%	4/5 = 80%
{abóbora, feijões} → {espargos}	4/14 = 28.6%	4/6= 66,7%

Posteriormente, verifica-se as regras de associação com apenas um único antecedente e consequente. Utilizando os *itemsets* frequentes em F_2 obtém-se a Tabela 6.

Tabela 6 - Regras de Associação para um Antecedente, exemplo da venda de vegetais
Fonte: Adaptado de Larose & Larose (2014) tabela 12.6 pág 254

Se Antecedente, então Consequente	Suporte	Confiança
{espargos} → {feijões}	5/14 = 37,50%	5/6= 83,3%
{feijões} → {espargos}	5/14 = 37,50%	5/10= 50%
{espargos} → {abóboras}	5/14 = 37,50%	5/6= 83,3%
{abóboras} → {espargos}	5/14 = 37,50%	5/7 = 71,4%
{feijões} → {milho}	5/14 = 37,50%	5/10= 50%
{milho} → {feijões}	5/14 = 37,50%	5/8= 62,5%
{feijões} → {abóbora}	6/14= 42.9%	6/10= 60%
{abóbora} → {feijões}	6/14= 42.9%	6/7=85,7%
{feijões} → {tomate}	4/14 = 28,6%	4/10= 40%
{tomate} → {feijões}	4/14 = 28,6%	4/6= 66,7%
{brócolos} → {pimentos}	4/14 = 28,6%	4/5=80%
{pimentos} → {brócolos}	4/14 = 28,6%	4/5 = 80%
{milho} → {tomate}	4/14 = 28,6%	4/8=50%
{tomate} → {milho}	4/14 = 28,6%	4/6 = 66,7%

Identificando os itens frequentes e descobrindo este tipo de padrões os retalhistas passam a conhecer melhor os seus clientes e a poder planear estratégias de *marketing* adequadas ao comportamento dos sues clientes.

3.3.1.3 Interpretação das Regras de Associação

Os valores gerados nas Regras de Associação dependem dos valores predefinidos de suporte e confiança. A definição destes valores é extremamente importante e tem um grande impacto na geração das regras. Assim sendo, a definição destes valores deve ficar a cargo de analistas experientes.

De acordo com Han, Kamber & Pei, (2012), considera-se uma regra de associação forte quando esta apresenta valores acima dos valores mínimos de suporte e confiança. Por vezes, mesmo regras fortes podem não ser interessantes ou levaram a conclusões erradas. Por exemplo, analisando a compra de jogos de computador e vídeos, das 10.000 transações analisadas, 6000 incluíram jogos de computador, 7500 incluíram vídeos e 4000 incluíram ambos os artigos. Definindo como suporte mínimo o valor de 30% e confiança

60% a seguinte regra de associação foi descoberta: compra (X, “Jogos de computador”) → compra (X, “Videos”), com suporte de 40% e confiança de 66%.

Os autores Han, Kamber & Pei, (2012) referem que como se trata de uma regra forte a mesma será reportada pois cumpre com os critérios mínimos previamente estabelecidos. Contudo, a esta regra pode levar a conclusões erradas, pois a probabilidade de comprar vídeos é de 75% que é superior aos 66% de confiança. Neste caso, a compra de um destes itens diminui a probabilidade de compra do outro, pelo que importa perceber esta situação antes de se tomar uma decisão de negócio.

Assim sendo, o suporte e a confiança podem não chegar para filtrar regras de associação que não sejam interessantes. De acordo com Han, Kamber & Pei, (2012), por forma a obter melhores resultados deve-se recorrer a medidas de correlação. Existem várias medidas de correlação que podem ser utilizadas, sendo que uma das mais utilizadas e simples é o *lift.* Esta indica que a ocorrência do *itemset* A é independente da ocorrência do *itemset* B se P(A $\cup B$) = P(A) P(B). De outra forma os *itemsets* A e B são dependentes e correlacionados.

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

Han, Kamber & Pei (2012) indicam que se o valor for inferior a 1, então a ocorrência de A está correlacionada negativamente com a ocorrência de B, o que significa que a ocorrência de um pode significar a ausência de outro. Se o valor for superior a 1, então A e B estão positivamente correlacionados, o que significa que a ocorrência de um implica a ocorrência também do outro. Se o valor for igual a 1, significa que A e B são independentes e não existe correlação entre eles. No exemplo dos jogos de computador e dos vídeos, verifica-se que a probabilidade de comprar um jogo de computador é de 60%, a probabilidade de comprar um vídeo é de 75% e a probabilidade de comprar ambos é de 40%. Aplicando a formula do *lift* temos 0.40/(0.60*0.75) que dá o valor de 0,89. Como o valor é inferior a 1 temos uma correlação negativa.

3.3.1.4 Algoritmos Eclat e FP-Growth

Importa referir que o algoritmo Apriori não é o único utilizado para a identificação dos itens frequentes e consequente geração de Regras de Associação. Desde a descoberta da geração de itens frequentes por Agrawal em 1993, esta tem sido uma temática que tem recebido muita atenção por parte da comunidade científica, tendo sido criados ou melhorados algoritmos relacionados com a geração de itens frequentes de acordo com Goethals (2004).

Por sua vez, Heaton (2016) menciona que os algoritmos Apriori, Eclat e FP-Growth são os algoritmos mais comuns para mineração de conjuntos de itens frequentes.

O algoritmo Apriori foi o pioneiro na geração de itens frequentes e já foi referido no capítulo 3.3.1.1. Relativamente ao algoritmo Eclat, Heaton (2016) menciona que este foi desenvolvido por Zaki, Parthasarathy, Ogihara e Li (1997) e que o seu nome é um acrónimo *para Equivalence Class Clustering e bottom up Lattice Traversa*. A principal diferença entre este algoritmo e o Apriori é que neste a primeira pesquisa é feita de forma recursiva. Os parâmetros de entrada para o Eclat são ligeiramente diferentes dos utilizados no Apriori. É utilizado um prefixo I que especifica o padrão que deve estar presente em qualquer conjunto de itens encontrado através da chamada ao Eclat.

Essa alteração permite usar a recursividade na construção dos conjuntos de itens. A iteração inicial do Eclat usa um valor I de {} (vazio), o que significa que nenhum prefixo específico é utilizado. Esta chamada inicial encontrará todos os conjuntos de itens frequentes de item único. Por sua vez, o modo de funcionamento do algoritmo Eclat é chamar-se recursivamente a si mesmo, aumentando I em cada iteração, adicionando conjuntos de itens que contêm o valor I na função com que foi chamado, mas com um item adicional. Este processo continuará até o valor de I ter comprimento suficiente, assegurando que o algoritmo tenha processado cestos de todos os comprimentos existentes.

De acordo com (Heaton J. , 2016), existem vários métodos para armazenar o suporte dos *itemsets* descobertos pelo algoritmo Eclat, sendo o mais comum a utilização de uma estrutura chamada em formato de árvore. Esta estrutura, em formato de árvore, contém sempre um nó vazio na raiz e à medida que os conjuntos de itens vão sendo encontrados vão sendo adicionados inserindo um nó para cada item que compõe o conjunto de itens. O item mais à esquerda corresponde a um filho do nó raiz. O segundo item corresponde a um filho do primeiro item deste conjunto frequente. Nenhum pai jamais teria mais de um filho do mesmo item; no entanto, um item pode aparecer em várias localizações. A árvore é gerada para que o algoritmo possa encontrar rapidamente o suporte de um conjunto de itens, percorrendo rapidamente a árvore uma vez que os itens no conjunto são lidos da esquerda para a direita. O nó que contém o item mais à direita contém o suporte para esse conjunto de itens. Enquanto o algoritmo processa o conjunto de dados a árvore é percorrida procurando cada conjunto de itens descoberto. São criados nós, se necessário, para preencher a árvore com todos os conjuntos de itens. Se os nós já existirem, o nó do item mais à direita no conjunto de itens tem o seu suporte aumentado.

A Figura 6 mostra um exemplo de uma árvore se suporte ao algortimo Eclat. A cada nó da árvore está associada uma lista de identificadores de transações, resultando que na prática, depois de ter essas listas para os nodos de primeiro nível, apenas é necessário fazer a interceção das listas de cada par de *itemsets* para obter a lista de identificadores e respetiva contagem, como exemplificado na Figura 7.

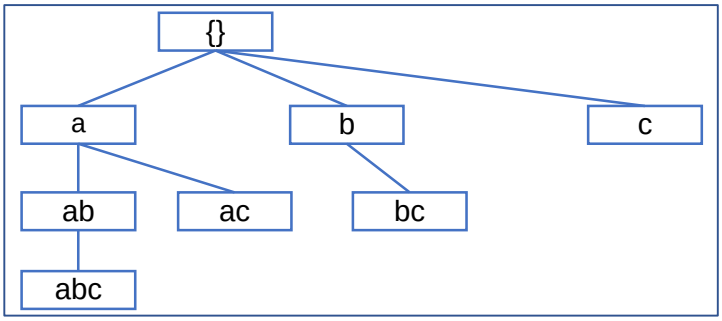

Figura 6 - Exemplo de árvore do Eclat
(Adaptado de (Maciag, Hepting, Ślęzak, & Hilderman, 2007))

A
1
2
3
5
7
8

∧

B
1
3
4
5
6
7

=

AB
1
3
5
7

Figura 7 - Exemplo de união de dois itemsets no Eclat

Relativamente ao algoritmo FP-Growth, segundo Borgelt (2005), trata-se de um dos algoritmos atualmente mais rápidos e populares para mineração de conjuntos de itens frequentes. Baseia-se na representação numa árvore de prefixos das transações da base de dados chamada FP Tree. Esta representação permite economizar quantidades consideráveis de memória para armazenar as transações. Este algoritmo pode ser descrito como um processo de eliminação recursiva, pois numa etapa de pré-processamento exclui todos os itens das transações que não são frequentes individualmente. De seguida, seleciona todas as transações que contêm o item menos frequente (menos frequente entre os frequentes) e exclui também esse item. Repete-se o processo por forma a obter uma base de dados reduzida, tendo em consideração que os conjuntos de itens que foram encontrados na recursão partilham o item eliminado através do prefixo. No retorno, remove-se o item processado também do conjunto de dados que contém todas as transações e começa-se de novo, ou seja, processa-se o segundo item frequente. Nas etapas de processamento, a árvore de prefixos é melhorada por *links* entre as filiais, é explorada para encontrar rapidamente as transações que contêm um determinado item e também para remover esse item das transações após o processamento.

Segundo Heaton (2016), o algoritmo FP-Growth foi introduzido por Han, Pei e Yin (2000) para evitar a geração de candidatos. Para isso, passou-se a utilizar uma árvore que armazena os valores reais dos cestos, em vez de armazenar candidatos como o Apriori e Eclat. As diferenças deste algoritmo para o Apriori e para o Eclat são que o Apriori é um algoritmo horizontal, o Eclat é um algoritmo vertical, de profundidade e, por sua vez, a estrutura do FP-Growth fornece uma visão vertical dos dados. No entanto, o FP-Growth também adiciona uma tabela de cabeçalho para cada item individual que possui suporte acima do suporte mínimo. Esta tabela de cabeçalho contém uma lista vinculada que todos os nós do mesmo tipo. A tabela de cabeçalho fornece ao FP-Growth uma visão horizontal dos dados, além da visão vertical fornecida pela árvore.

Segundo Heaton (2016), o algoritmo Apriori é de fácil entendimento e é por isso, regra geral, o algoritmo utilizado quando começamos a estudar a temática dos itens frequentes. Contudo, este algoritmo apresenta problemas de escalabilidade e esgota a memória disponível muito mais rapidamente que o Eclat ou o FP-Growth.

4 Análise Exploratória dos Dados

Tendo como bússola orientadora a metodologia CRISP, começou-se por conhecer o negócio, conhecer os dados e preparar os dados. O negócio em questão é a entrega de produtos de mercearia nos Estados Unidos da América. Os clientes fazem as suas compras *online* escolhendo os retalhistas locais que têm acordo com a Instacart através da aplicação móvel Instacart ou através do *site* Instacart.com e uma pessoa (*personal shopper)* da Instacart vai pessoalmente levantar as compras realizadas e entregá-las no local e hora indicado pelo cliente. Os preços apresentados no *site* da Instacart podem divergir dos preços dos retalhistas locais, podendo ser mais baixos, mais altos ou os mesmos. No *site* da Instacart existem também artigos em promoções e cupões promocionais que oferecem descontos em determinados produtos. Cada entrega tem um custo associado, existindo também a possibilidade de subscrever um plano mensal ou anual com a Instacart que contemplará entregas gratuitas.

Ao passar para a etapa de conhecer os dados e ao analisar os mesmos através de uma análise exploratória adquiriu-se conhecimento acerca do negócio e validou-se que os mesmos eram úteis para o estudo do problema em questão. Conhecendo bem o negócio e os dados é também mais fácil preparar os mesmos, pois mais facilmente identificamos a necessidade de efetuar transformações e limpezas nos dados que auxiliem na resolução do problema.

4.1 Descrição do Dataset e Modelo Relacional

As primeiras ações referentes à análise do conjunto de dados obtido em Instacart (2017) consistiram na leitura das informações relacionadas com os ficheiros disponibilizados. A Instacart ao disponibilizar este conjunto de dados forneceu também a indicação que estes fazem parte do *"The Instacart Online Grocery Shopping Dataset 2017"* (Stanley, 2017) e que para cada utilizador disponibilizaram entre 4 a 100 compras.

Este conjunto de dados é composto por vários ficheiros que descrevem as compras dos clientes ao longo do tempo. Além disso, os ficheiros ligam-se entre si através da aplicação de um modelo relacional.

Os ficheiros foram disponibilizados em formato CSV e são os seguintes:

- Corredores;
- Departamentos;
- Compras;

- Produtos:
- Compras_Produtos_Histórico.
- Compras_Produtos_Treino.

O ficheiro correspondente aos corredores tem apenas duas colunas, o id do corredor e o nome do mesmo. O mesmo acontece para o ficheiro correspondente aos Departamentos, temos id do departamento e nome do departamento. O ficheiro referente às compras é composto pelo id da compra, pelo id do cliente, pela indicação de qual o conjunto que uma compra pertence (anterior, treino, teste), o número da compra, a coluna order_dow que se refere ao dia da semana, a coluna que indica a que hora do dia foi realizada a compra e os dias que passaram desde a última compra. O ficheiro que identifica os produtos apresenta como colunas o id do corredor, o id do departamento, o id do produto e o nome do produto. Os ficheiros Compras_Produtos (Histórico e Treino) são ficheiros auxiliares que servem para estabelecer a relação entre o ficheiro Produtos e o ficheiro Compras, nele constam as colunas id da compra, id do produto, número de ordem em que foi adicionado à compra e se é a repetição da compra do mesmo produto (comprou em compras anteriores).

Para mostrar como se interligam de forma relacional os diferentes ficheiros, coloca-se na Figura 8 o modelo relacional realizado com recurso ao *software Power BI*.

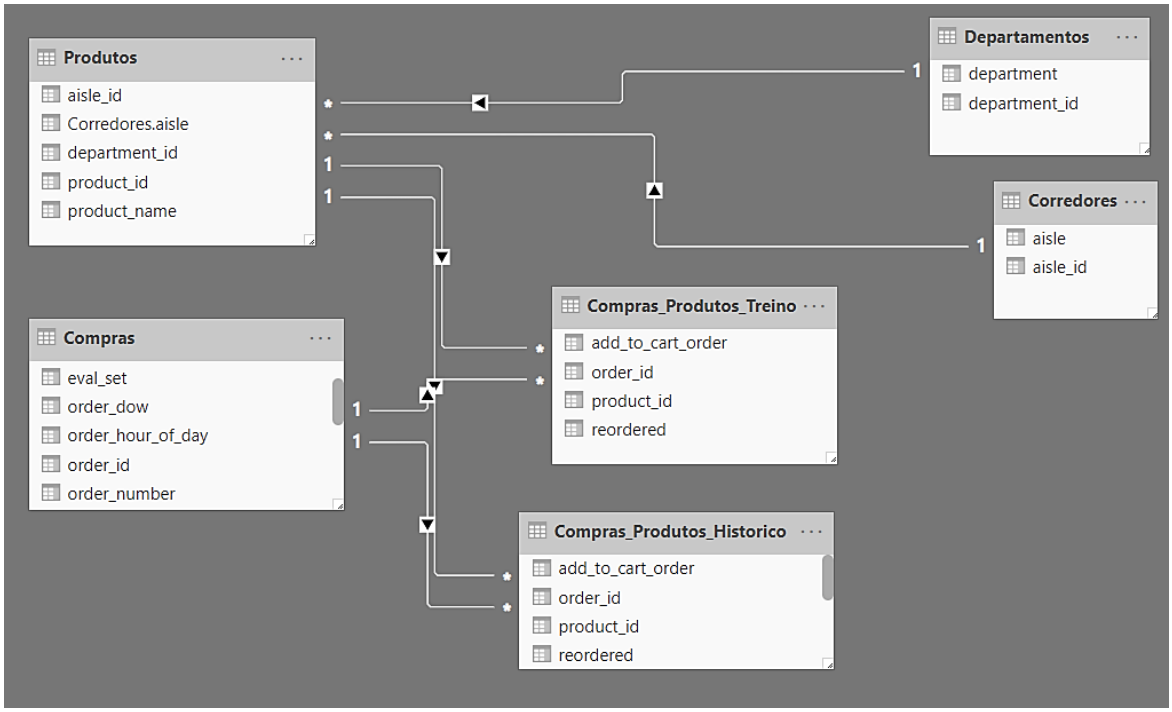

Figura 8 - Modelo Relacional do Conjunto de Dados a analisar

4.1.1 Descrição das Variáveis

Posteriormente identificaram-se os tipos de variáveis presentes no conjunto de dados. A Tabela 7 descreve todos os tipos de variáveis.

Tabela 7- Descrição dos tipos de variáveis do Dataset analisado

Variável	Tipo de Variável	Significado/Propósito
product_id	Nominal	Permite a identificação unívoca do produto
product_name	Nominal	Identifica o nome do produto
aisle_id	Nominal	Permite a identificação unívoca do corredor
aisle_name	Nominal	Identifica o nome do corredor
department_id	Nominal	Permite a identificação unívoca do departamento
department	Nominal	Identifica o nome do departamento
order_id	Nominal	Permite a identificação unívoca da compra
user_id	Nominal	Permite a identificação unívoca do utilizador
eval_set	Ordinal	Identifica a que conjunto de dados pertencem a compra
order_number	Ordinal	Indica o número de compras feito pelo utilizador incluindo a presente compra
order_dow	Ordinal	Identifica numa escala de 0 a 6 o dia da semana a que foi realizada a compra.
order_hour_of_day	Ordinal	Identifica a que hora do dia foi realizada a compra
days_since_prior_order	Discreta	Indica quantos dias passaram desde a última compra
add_to_cart_order	Ordinal	Indica a ordem pela qual os produtos foram adicionados ao carrinho de compras para a compra em questão
reordered	Nominal	Indica se o produto em questão já tinha sido adquirido em compras anteriores pelo mesmo utilizador

Este conjunto de dados é composto por 3 subconjuntos: o subconjunto *prior* que contem o histórico das compras realizadas e que é o conjunto de maior dimensão, o subconjunto *train* que tem uma dimensão menor e o subconjunto *test* que é normalmente utilizado para avaliar o modelo construído aquando a aplicação de algoritmos de *data mining* referentes a *machine learning*.

A análise exploratória dos dados serve para validar a utilidade dos dados presentes no conjunto de dados, identificar possíveis necessidades de limpeza e transformação de

dados e também para adquirir *insights* através da análise dos dados que permitam ficar a conhecer melhor o negócio.

4.2 Principais Métricas e Indicadores

Recorrendo ao *software Power BI*, começou-se por identificar a totalidade de compras efetuadas em todo o conjunto de dados. Entende-se por uma compra uma encomenda efetuada com sucesso no *site* da *Instacart* contendo um ou mais produtos. Criou-se uma *measure* (medida) que conta o número distinto de valores da coluna *order_id* presente na tabela Compras e criou-se um gráfico do tipo *card* com essa informação conforme se pode verificar na Figura 9.

Figura 9 – Número Total de Compras

Posteriormente analisou-se qual o período do dia em que se realizam mais compras. Verifica-se que a hora do dia em que são efetuadas mais compras são as 10 da manhã. E que o período em que são feitas mais compras é o período entre as 10 da manhã até as 16 horas. Isto significa que as compras são efetuadas maioritariamente dentro do horário laboral. Abaixo na Figura 10 coloca-se o gráfico que espelha esta informação. Para uma melhor visualização deste gráfico o mesmo foi disponibilizado no Anexo 8.1.

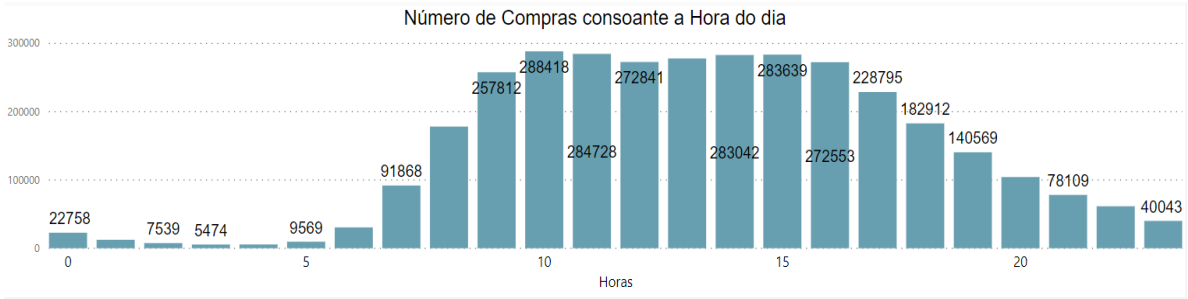

Figura 10 - Número de Compras consoante a Hora do Dia.

Considerou-se relevante verificar também o número de compras consoante o período do dia. No conjunto de dados estava disponível a que horas as compras eram realizadas, assim sendo, no *Power BI* criou-se um grupo onde se agruparam as horas consoante o período do dia, as 24 horas foram divididas em períodos de 6 horas, sendo que a madrugada é composta pelo período das 0 às 5 da manhã, a manhã é composta pelo período das 6 às 11, a tarde é constituída pelo período das 12 às 17 e a noite pelo período das 18 as 23.

Verifica-se que o período da tarde é o que regista mais compras realizadas conforme informação presente na Figura 11.

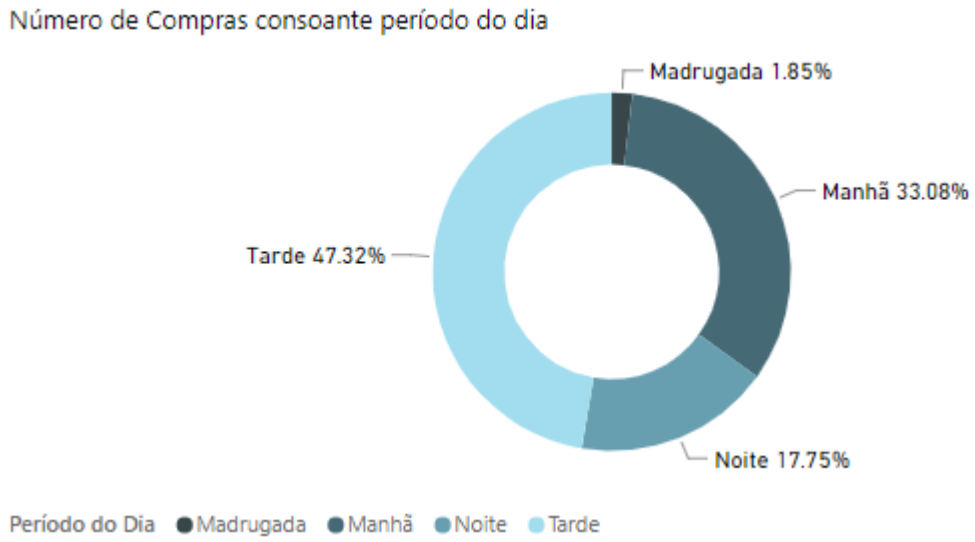

Figura 11 - Número de Compras consoante período do dia

Foi considerado importante analisar também o número de compras por dia da semana. No conjunto de dados não existe indicação de como é feita a correspondência entre os dias da semana e os números 0 a 6 presentes na coluna *order_dow* da tabela Compras. Constata-se que as compras são realizadas preferencialmente nos dias 0 e 1, conforme se verifica na Figura 12.

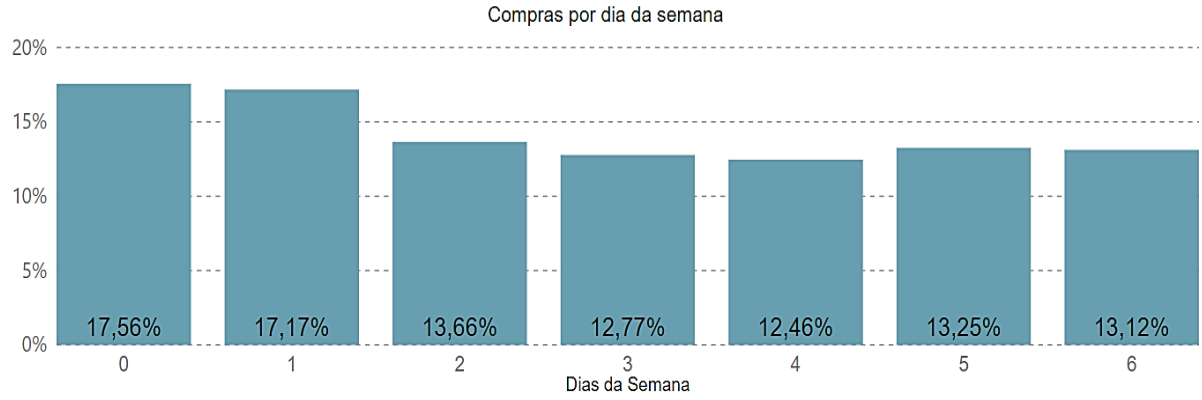

Figura 12- Compras por dia da semana

De seguida analisou-se a relação entre o número de utilizadores e as compras realizadas. As 3421083 compras foram realizadas por 206209 utilizadores. Foi criado um gráfico tipo *card* no *Power BI* com esta informação que está também presente na Figura 13.

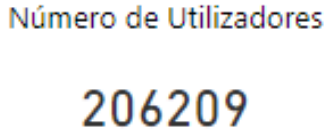

Figura 13 - Número de Utilizadores

A próxima análise verifica a quantidade de compras realizadas por utilizador. Como temos 206209 utilizadores são muitos e condensados os valores apresentados na Figura 14. No entanto, verifica-se claramente que são vários os utilizadores que apresentam 100 compras (o número máximo de compras por utilizador neste conjunto de dados).

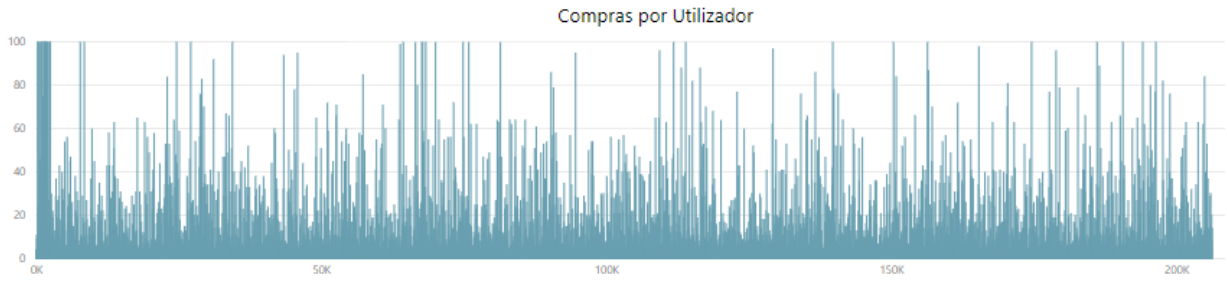

Figura 14 - Compras por Utilizador

Para ter uma melhor perceção acerca dos utilizadores e se os mesmos fazem muitas ou poucas compras agregou-se o total de utilizadores por número de vezes que realizaram uma compra no site da *Instacart*, tendo sido obtido o gráfico ilustrado na Figura 15.

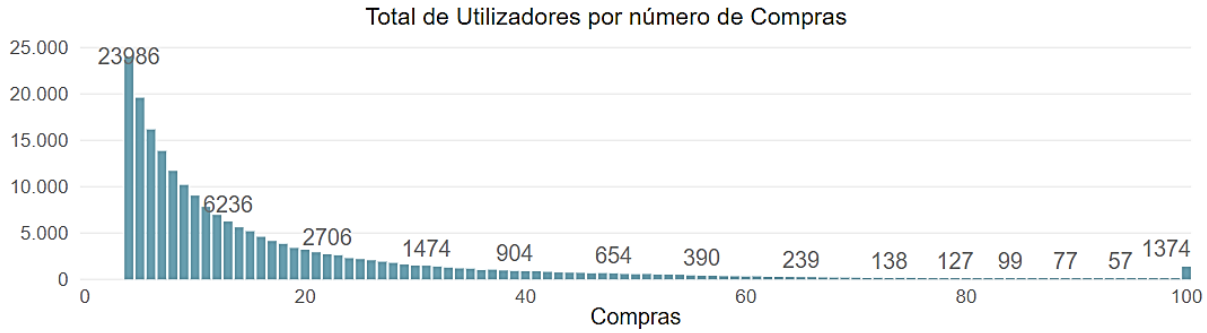

Figura 15 - Total de Utilizadores por número de Compras

Constata-se que o maior número de utilizadores tem registado entre 4 a 10 compras no *site* da *Instacart*. Contudo, verifica-se um valor anormal (*outlier)* nas 100 compras. Não temos mais de 100 compras por utilizador nem temos utilizadores com 1, 2 ou 3 compras efetuadas porque a empresa ao providenciar os dados limitou as compras entre 4 a 100 por cada utilizador.

Como se verificou que existiam muitos utilizadores a realizar entre 4 a 10 compras pretendeu-se averiguar qual o peso percentual que esses utilizadores tinham sobre o total de compras. Para analisar o peso percentual dos utilizadores em termos de número de compras foi criado o gráfico ilustrado na Figura 16

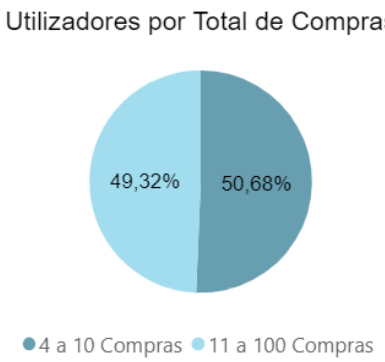

Figura 16- Utilizadores por Total de Compras

Verifica-se então que 50,68% dos utilizadores realizaram entre 4 a 10 compras, ou seja, mais de metade dos utilizadores realizou neste *site* entre 4 a 10 compras. Sendo que em média cada utilizador, neste conjunto de dados, realizou 17 compras e como mediana temos 10 compras por utilizador, conforme se verifica na Figura 17 e na Figura 18 . Optou-se por calcular não só a média como também a mediana para esta situação devido à existência de *outliers*.

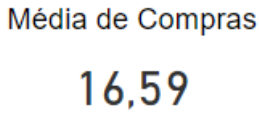

Figura 17 - Média de Compras por Utilizador

Figura 18 - Mediana de Compras por Utilizador

Posteriormente analisou-se o número de dias decorridos entre a atual compra e a anterior e constatou-se que a maior parte das compras são realizadas com uma diferença de 7 dias ou de 30 dias, o que faz sentido pois normalmente adquirimos produtos ou mensalmente ou semanalmente. Além disso, verificou-se também que existiram 206209 compras em que não foi registado o número de dias decorridos desde a compra anterior. Ao analisar esta

situação verifica-se que temos 206209 utilizadores e para a primeira compra efetuada pelos mesmos não temos ainda nenhum registo de compras anteriores. O número máximo de dias decorridos desde a compra anterior é de 30 dias. Esta informação pode ser verificada na Figura 19 ou com maior detalhe no anexo 8.2.

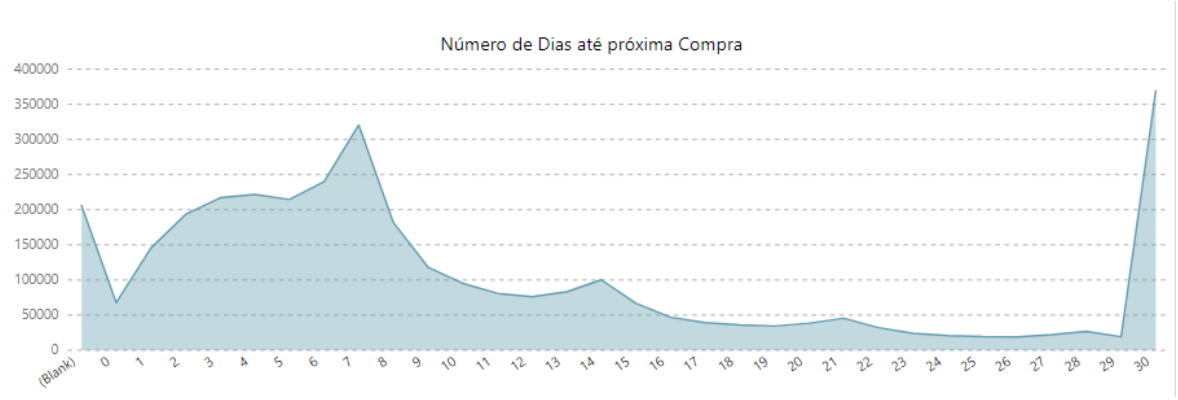

Figura 19 - Número de Dias até próxima Compra

Considerou-se também pertinente analisar os Produtos que são alvo de compras e que compõem o subconjunto de dados. O conjunto de dados apresenta 49688 produtos que estão repartidos por 134 corredores que fazem parte de 21 departamentos.

Constata-se que os departamentos de cuidados pessoais e de *snacks* são os que apresentam mais variedades de produtos, logo seguidos dos departamentos de bebidas e de despensa (constituído principalmente por especiarias, preparados para a comida e itens de confeção de bolos). Verifica-se também que alguns produtos não estão associados a corredores nem a departamentos, estando colocados numa categoria "*Missing*". O gráfico em questão poderá ser consultado no Anexo 8.3.

Analisando os produtos mais adquiridos, em primeiro lugar encontram-se Bananas e em segundo lugar Bananas Orgânicas, seguidas de Morangos Orgânicos e Espinafres Bebés Orgânicos. Ou seja, verifica-se claramente uma tendência para a aquisição de frutas e vegetais em particular de origem orgânica. O gráfico correspondente está disponível para visualização no anexo 8.4. Na Figura 20 poderá ser visualizado outro gráfico que mostra apenas os 10 produtos mais requisitados.

Figura 20 - Produtos mais adquiridos (Top 10)

Foi considerado interessante verificar a que corredores e departamentos pertenciam os produtos mais requisitados. Agrupando os produtos adquiridos por corredor e por departamento obteve-se o gráfico disponibilizado no anexo 8.5. Este gráfico apresenta o Top 10 dos Departamentos com mais produtos adquiridos.

Verifica-se que o departamento *Produce* é o que apresenta mais produtos adquiridos o que é consistente com o top de produtos adquiridos que foi verificado anteriormente, pois é no departamento de *Produce* que estão situados as frutas e os vegetais. Posteriormente temos o Departamento *Dairy Eggs* onde se encontram os lacticínios e os ovos. No top 5 dos departamentos com mais produtos requisitados pelos utilizadores temos também o departamento de *Snacks* onde temos Aperitivos, Batatas fritas, pipocas entre outros, o departamento de *Beverages* (Bebidas) e o departamento de *Frozen* (Congelados).

Por outro lado, o top 5 dos Departamentos que apresentam menos produtos adquiridos são o departamento das bebidas alcoólicas, o departamento de produtos para os animais de estimação, o departamento *missing* que já foi anteriormente referenciado e que são todos os produtos que não têm um valor associado tanto a nível de corredor como a nível de departamento, o departamento Outros e o departamento *Bulk* (a granel). Esta informação pode ser visualizada no anexo 8.6.

O conjunto de dados tem uma coluna denominada *reorder* quando o valor é 1 significa que o utilizador numa compra anterior adquiriu o mesmo produto e quando é 0 significa que o produto que está a ser comprado por aquele utilizador ainda não tinha sido adquirido por ele em compras anteriores. Considerou-se interessante verificar quais os produtos que foram mais vezes recomprados e verificou-se que os produtos que foram mais vezes recomprados são também que foram mais vezes comprados, o que revela coerência nos

hábitos de consumo. Na Figura 21 podemos visualizar o gráfico referente ao Top 10 de produtos que são mais vezes recomprados.

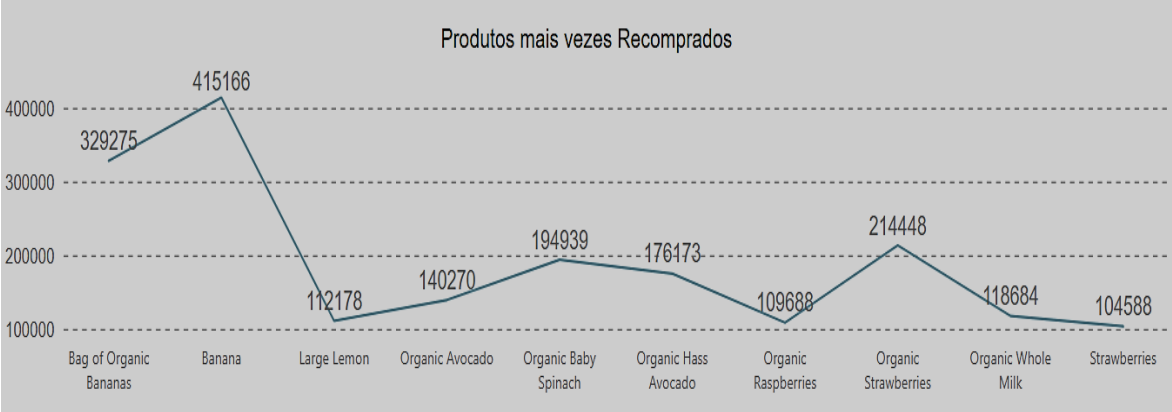

Figura 21 - Produtos mais vezes recomprados

Verifica-se também que 59% das compras apresentam produtos que já tinham sido comprados anteriormente, conforme se pode constatar na Figura 22.

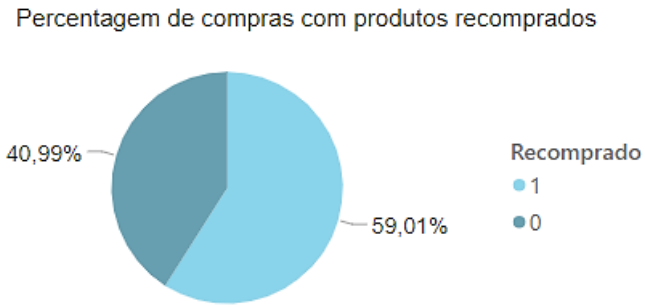

Figura 22 - Percentagem de compras com produtos recomprados

Analisando o número de itens adquiridos em cada compra verifica-se que o maior número de compras é composto por 5 produtos, logo seguida de compras com 6 e 4 produtos. Verifica-se também que na maior compra foram adquiridos 145 produtos e que a média de produtos adquiridos por compra é de 10 produtos. Na Figura 23 temos o gráfico referente ao Top 10 de número de produtos mais adquiridos por compra, tendo sido colocado no anexo 8.7 o gráfico detalhado com o número de produtos por compra. Na Figura 24 temos o *card* resultante do cálculo da média de produtos adquiridos por compra.

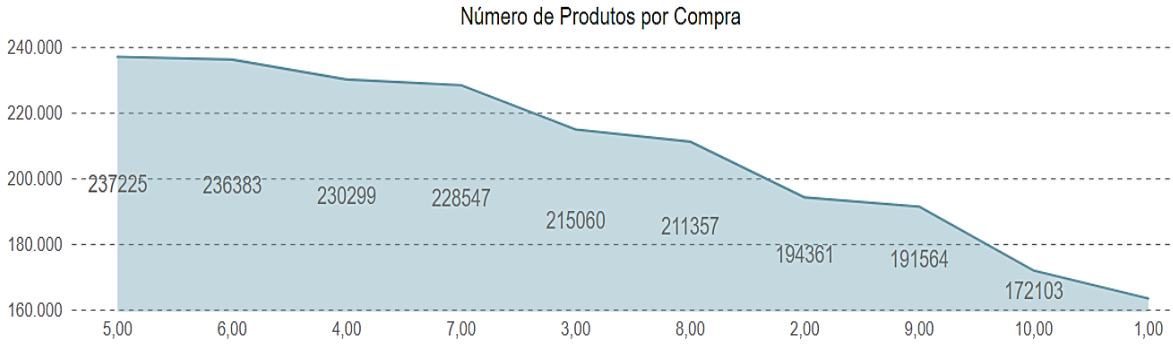

Figura 23 - Top 10 Número de Produtos por Compra

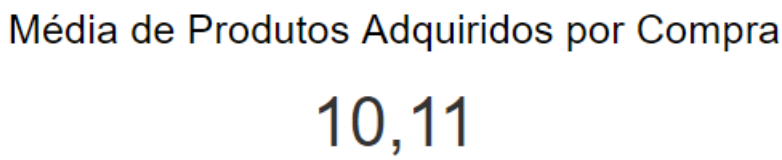

Figura 24 - Média de Produtos Adquiridos por Compra

Uma vez que existem *outliers* no número de produtos adquiridos por compra foi considerado também relevante efetuar o cálculo da mediana uma vez que esta medida não é influenciada pela presença de *outliers*. Assim sendo, como mediana temos 8 produtos adquiridos por compra, conforme se pode visualizar no *card* da Figura 25.

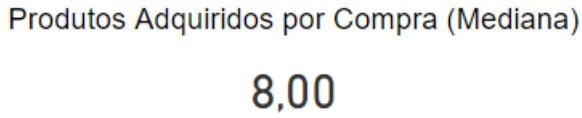

Figura 25 - Mediana de Produtos adquiridos por Compra

Uma vez que no conjunto de dados existia informação acerca da ordem pela qual os produtos eram colocados no carrinho de compras, considerou-se relevante analisar quais os produtos que eram escolhidos em primeiro lugar pelos utilizadores. Esta informação acerca da ordem pela qual os produtos eram selecionados para o carrinho de compras está apenas disponível para os conjuntos de dados *prior* e *train*, pelo que o total de compras que obtivemos foi de 3346083. Para o conjunto de dados *test* temos 75000. A soma destes valores origina o valor 3421083, o valor que foi indicado no início desta análise como o total de compras.

Pretendeu-se identificar os produtos que o utilizador escolhe mais vezes em primeiro, segundo e terceiro lugar quando está a efetuar uma compra. Verifica-se que muitos dos produtos mais vezes adquiridos são também escolhidos em primeiro lugar, mas que as percentagens de escolha em primeiro lugar nem sempre são as mais altas para os produtos mais vezes comprados. No caso do produto Banana, este produto foi adquirido em 491291 compras, desse total de compras foi o primeiro item a ser adicionado ao carrinho de

compras apenas 115521 vezes tendo uma percentagem de apenas 23,51%. Verifica-se também que apesar de ser o produto mais vezes adquirido nem sempre foi o primeiro, segundo ou terceiro produto a ser adicionado ao carrinho de compras, tendo uma percentagem de 52, 91 % para ser selecionado num dos primeiros 3 produtos a colocar no carrinho de compras. Na Figura 26 encontra-se uma tabela que permite a visualização dos 10 produtos mais vezes comprados e o número de vezes e a percentagem em que foram escolhidos como primeiro, segundo ou terceiro item no carrinho de compras.

Produtos e ordem de adição ao carrinho de compras

Nome do Produto	Compras	Selecionado em 1 Lugar	% 1 Lugar	Selecionado em 2 Lugar	% 2 Lugar	Selecionado em 3 Lugar	% 3 Lugar
Banana	491291	115521	23,51%	83631	17,02%	60832	12,38%
Bag of Organic Bananas	394930	82877	20,99%	67204	17,02%	50923	12,89%
Organic Strawberries	275577	28875	10,48%	30836	11,19%	28908	10,49%
Organic Baby Spinach	251705	24412	9,70%	26071	10,36%	25346	10,07%
Organic Hass Avocado	220877	24913	11,28%	27442	12,42%	25029	11,33%
Organic Avocado	184224	23393	12,70%	24513	13,31%	21091	11,45%
Large Lemon	160792	12891	8,02%	14975	9,31%	14998	9,33%
Strawberries	149445	17073	11,42%	17687	11,84%	16342	10,94%
Limes	146660	10092	6,88%	11953	8,15%	12551	8,56%
Organic Whole Milk	142813	32071	22,46%	21630	15,15%	16156	11,31%

Figura 26 - Ordem de adição de itens ao carrinho de compras

Analisou-se também produtos que apresentavam uma elevada percentagem como primeiro produto a ser adicionado ao carrinho de compras e descobriu-se que os produtos que tinham poucas compras eram os que apresentavam maiores percentagem em escolhas de determinado produto em primeiro lugar. Produtos como contracetivos de emergência tinham sido o primeiro item a ser selecionado para o carrinho de compras em 79,25% das vezes, a bebida energética Orangeade foi escolhida 78,79% das vezes como primeiro produto e a bebida alcoólica Vodka de Pêssego foi escolhida 70,83% das vezes. Estas percentagens são elevadas nestas situações visto que são calculadas considerando apenas as compras onde o produto é incluído e não o conjunto total das compras. A Figura 27 apresenta a tabela com a informação acima referida.

Nome do Produto	Compras	Selecionado em 1 Lugar	% 1 Lugar
Emergency Contraceptive	53	42	79,25%
Rehab Energy Iced Tea Orangeade	66	52	78,79%
California Champagne	18	14	77,78%
Blue Label Year of the Ram	8	6	75,00%
Cristal Champagne	4	3	75,00%
Large Picture Hanging Strips	4	3	75,00%
Plus Lens Solution	4	3	75,00%
Quart Bags	4	3	75,00%
Rose Aura Glow Body & Massage Oil	4	3	75,00%
Tip and Coupler Set	4	3	75,00%
Wicked Apple Ale	4	3	75,00%
Optive Advanced Lubricant Single Use Eye Drops	7	5	71,43%
Flavored Vodka, Peach	72	51	70,83%

Figura 27 - Produtos com elevada percentagem em serem selecionados como primeiro produto a colocar na cesta de compras

Verificou-se também por departamentos quais eram os departamentos que tinham produtos a serem adicionados em primeiro lugar mais vezes por compra, e tal como aconteceu com os produtos, nem todos os departamentos que apresentam mais compras são os que têm maior percentagem de escolha de produtos em primeiro lugar para o carrinho de compras, conforme se visualiza na Figura 28.

Departamentos e ordem de adição ao carrinho de compras

Departamento	Compras	Selecionado em 1 Lugar	% 1 Lugar
produce	2506247	937376	37,40%
dairy eggs	2264738	676270	29,86%
beverages	1518833	427869	28,17%
snacks	1448749	233966	16,15%
frozen	1232089	189927	15,42%
pantry	1165491	160846	13,80%
bakery	917980	122330	13,33%
deli	802581	91839	11,44%
canned goods	710721	65587	9,23%
dry goods pasta	623738	53627	8,60%

Figura 28 - Departamentos com mais compras e percentagem de seleção de primeiro item no carrinho de compras

Departamentos como bebidas alcoólicas, Produce (vegetais e frutas) e lacticínios são os que apresentam maior percentagem de selecionarem o produto como o primeiro produto a ser colocado no carrinho de compras. Isto poderá indicar um nicho de mercado que poderá

ser explorado através de estratégias de marketing para *upsell* e *cross sell* de produtos. Esta informação poderá ser visualizada na Figura 29.

Departamentos e ordem de adição ao carrinho de compras

Departamento	Compras	Selecionado em 1 Lugar	% 1 Lugar
alcohol	87794	36118	41,14%
produce	2506247	937376	37,40%
dairy eggs	2264738	676270	29,86%
beverages	1518833	427869	28,17%
pets	62011	12894	20,79%
household	492427	88929	18,06%
personal care	333463	54172	16,25%
snacks	1448749	233966	16,15%
frozen	1232089	189927	15,42%
babies	184074	25688	13,96%

Figura 29 - Departamentos com maior percentagem de produtos selecionados em primeiro lugar no carrinho de compras.

Através da análise exploratória dos dados passou-se a ter um conhecimento mais profundo acerca do negócio e dos dados que o constituem. Neste capítulo começou-se por identificar as relações entre os diversos ficheiros e as variáveis que os constituíam. Confirmou-se que o formato dos dados era idêntico para todos os ficheiros e estabeleceram-se relações entre os vários ficheiros. Utilizou-se o método da sumarização descritiva onde foram identificados *outliers* e aplicadas medidas de tendência central. Identificaram-se também valores a nulo, mas que estavam devidamente fundamentados como o caso do campo dias decorridos desde a última compra apresentar valores a nulo aquando a primeira compra do utilizador.

Adicionalmente para ajudar na análise dos dados e ter uma melhor perceção acerca do comportamento dos utilizadores foram realizadas transformações gerando novas medidas para análise no *software Power BI*, foram agregados e agrupados valores por determinados campos e separados determinados valores em grupos de modo a termos outras perspetivas acerca dos dados.

5 Análise aos carrinhos de compras

Neste capítulo será feita uma análise aos carrinhos de compras (*Market Baskets*) presentes no *dataset* em estudo com o intuito de identificar padrões de compra, o que pode, como já explicado anteriormente, ajudar em diversas tomadas de decisão na disposição de produtos e definição de campanhas promocionais.

Nesta investigação serão igualmente aplicadas as fases da metodologia CRISP de preparação dos dados, modelação e avaliação. A tarefa de preparação de dados consistirá em transformar e limpar os dados presentes no *dataset* de modo a obter os dados necessários como *input* para a fase de modelação e apresentados no formato esperado peloa algoritmos. Por sua vez, a tarefa de modelação consiste em selecionar e aplicar a técnica de *data mining* que permite extrair as regras de associação dos dados em estudo, sendo utilizado em concreto, nesta investigação, o algoritmo Apriori, que foi apresentado na secção 3.3.1.1. Por último, face aos resultados obtidos com o modelo selecionado, nas diversas configurações e *datasets* utilizados, serão analisados e discutidos os resultados.

5.1 Tecnologias utilizadas

Para executar as tarefas inerentes às fases de preparação dos dados e posterior modelação, nesta fase da investigação, optou-se por recorrer à linguagem *Python*, que é uma linguagem muito utilizada em aplicações relacionadas com *data mining* e para as quais existem diversos recursos (bibliotecas) desenvolvidos e testados por uma vasta comunidade que desenvolve investigação/trabalho nesta área (Grus, 2019).

De modo a ter um ambiente dotado das bibliotecas habitualmente utilizadas em projetos de *data science*, incluindo o próprio *Python* (na versão 3), optou-se por utilizar o *software Anaconda*, que além dos recursos de base que integra, disponibiliza ainda um *IDE* (*Integrated Development Environment*) que facilita o acesso aos componentes e outras ferramentas gráficas, como é o caso do *Jupyter Notebook* que permite a criação de páginas *Web* com linhas de comando integradas e que podem ser executadas.

Para completar o ambiente, foi necessário instalar uma biblioteca que implementa o algoritmo Apriori, tendo sido instalada a biblioteca “apyori 1.1.1” disponível em https://pypi.org/project/apyori/ (Mochizuki, 2016). Procurou-se obter uma biblioteca que fosse relevante dentro da comunidade científica, mas ao mesmo tempo fosse de simples implementação e fácil utilização. Efetuando uma breve pesquisa no site https://pypi.org/ e procurando por Apriori verifica-se que, no momento da pesquisa, a biblioteca “apyori 1.1.1” ocupava o primeiro lugar na dimensão *Trending* e o quinto lugar na dimensão *Relevance*.

Adicionalmente, realizaram-se outras pesquisas de bibliotecas que implementassem o algoritmo Apriori em Python e constatou-se que a biblioteca escolhida aparece por inúmeras vezes referenciada no *GitHub*, plataforma mundialmente conhecida por armazenar código fonte e que possibilita a partilha de código e informações tanto para projetos pessoais como também para projetos comerciais.

5.2 **Preparação dos dados**

Nesta investigação foram utilizados dois *datasets* de entre os dados disponibilizados pela Instacart, concretamente os dados referentes às compras registadas, sendo a primeira amostra de menor dimensão e designada por dados de teste (*train*), com 131209 compras e obtida do ficheiro *order_products_train* (disponível em formato CSV). A segunda amostra corresponde ao histórico de compras completo, com 3214874 compras efetuadas e o ficheiro utilizado foi o *order_products_prior.*

A primeira tarefa consistiu na limpeza dos dados, a qual passou pela eliminação das colunas desnecessárias para a execução do algoritmo e que pode permitir obter melhor performance na aplicação do mesmo. Foram eliminadas as colunas *add_to_cart_order* e *reordered* cujo conteúdo não era relevante para o que se pretendia investigar.

A segunda tarefa consistiu na transformação dos dados de modo a que estes estivessem num formato adequado à utilização do algoritmo de *data mining*. O formato dos dados originais é semelhante ao que é utilizado nos modelos relacionais, sendo utilizada uma linha para registo de cada produto que integra uma compra, como exemplificado na Figura 30.

```
order_id,product_id,add_to_cart_order,reordered
 1,49302,1,1
 1,11109,2,1
 1,10246,3,0
 ...
```

Figura 30 – Formato original dos dados das compras

O formato pretendido para a utilização do algoritmo de análise dos carrinhos de compras define que cada compra deve ser representada apenas por uma linha que inclua a lista dos diversos produtos que integram essa compra, o que, considerando o exemplo da compra com id 1 (order_id) da Figura 30, deve ser representado apenas numa linha como o exemplo da Figura 31.

```
order_id,{product_id+}
1,{49302,11109,10246}
```

Figura 31 – Representação das compras numa só linha

O script utilizado para a transformação dos dados pode ser consultado no anexo 8.8.

5.3 **Modelação**

Estando os dados no formato necessário passou-se a utilização do algoritmo sobre os mesmos. A implementação do algoritmo *Apriori* da biblioteca utilizada (apyori) exige que o conjunto de dados esteja na forma de uma lista de listas, onde o conjunto das compras é uma grande lista e cada transação no conjunto de dados é uma lista interna da grande lista externa com os diversos produtos dessa compra.

Para otimizar o algoritmo, nomeadamente na quantidade de memória necessária, os dados previamente transformados para a representação de uma compra por linha foram lidos do ficheiro para um *dataframe* do *pandas e* dado que a quantidade de produtos presente em cada compra pode variar entre um e 145, a representação dos dados dos produtos comprados em formato tabular resulta numa tabela esparsa (com muitos vazios), sendo ainda assim necessário reservar esse espaço de memória. A não-eliminação dos valores vazios, numa fase inicial, resultou em testes preliminares com um desempenho muito inferior. Para poder ter apenas os dados necessários, sem valores nulos, representados internamento como uma lista de listas, foi utilizado o script da Figura 32.

```
# Importação das bibliotecas necessárias
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from apyori import aprior

# Ficheiro já sem "" e sem [], operação efetuada via Notepad++ através da funcionalidade replace.
#Identificação do range das colunas (valor máximo) para não dar erro ao processar ficheiro pois variam de compra para compra.
col_names=[i for i in range(0,79)]
#Carregamento do ficheiro CSV para formato dataframe (Pandas)
Dados = pd.read_csv('F:\\Tese\\TrainGB.csv', header=None, names=col_names)
# Colocação de todos os dados como float para melhor performance
Dados1 = Dados.astype(float)
# Conversão dos dados do dataframe para uma lista de listas eliminando a grande quantidade de células com valores NaN do dataframe
para melhor performance e ter os dados necessários para o algoritmo Apriori.
records = []
#Percorrer todas as linhas do dataframe (correspondem às compras/cestos)
for i in range(0, 131208):
   rec=[]  #criar nova lista para adicionar  produtos da compra atual
   #Percorrer todas as colunas do dataframe
   for j in range(0, 79):
      if not np.isnan(Dados1.values[i,j]) :
         rec.append(Dados1.values[i,j]) #adiciona apenas os valores não-nulos (identificador do produto)
   records.append(rec) #No final adiciona a última compra à lista de compras/cestos

# Utilização do algoritmo com a definição dos parâmetros mínimos de suporte, confiança e lift e visualização das regras geradas
association_rules = apriori(records, min_support=0.05, min_confidence=0.1, min_lift=1.1)
association_results = list(association_rules)
# Visualização das regras geradas
print (association_results)
```

Figura 32 – Carregamento dos dados para algoritmo pesquisa de Regras de Associação

Depois de criada a lista de transações no formato adequado ao algoritmo *Apriori* adotado, a mesma pode ser utiliza pelo mesmo, como demonstrado no script completo colocado no anexo 8.9, sendo apenas necessário definir os parâmetros adicionais de configuração do algoritmo, nomeadamente o valor mínimo de suporte, confiança e *lift*. Numa fase inicial de avaliação do algoritmo e do seu comportamento nos *datasets* utilizados, foram testados valores de suporte elevados, contudo, com esses valores não era gerada nenhuma regra de associação pelo que, foi necessário reduzir esse valor até começarem a ser identificadas algumas regras de associação. Apesar de se ter baixado quer o valor do suporte quer o valor da confiança, o *lift* mínimo nunca foi inferior a 1.1 pois pretendiam-se regras fortes onde os produtos estivessem correlacionados positivamente.

Os resultados iniciais foram os seguintes:

- Mínimo Suporte 0,05; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,04; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,03; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,02; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Gerada 1 Regra de Associação
- Mínimo Suporte 0,01; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Geradas 16 Regras de Associação
- Mínimo Suporte 0,01; Mínimo Confiança 0,2 e mínimo *Lift* 1,1 → Geradas 12 Regras de Associação

Apresentam-se na Tabela 8 as 12 regras de associação geradas com o valor mínimo de suporte de 0,01 e 20% de mínimo de confiança. Nesta primeira análise dos resultados obtidos, pretende-se obter regras fortes, pelo que se optou pelo conjunto de regras que apresentava maior confiança.

Verifica-se que apenas com valores de suporte mínimo 0,02 é que foram geradas regras de associação, tendo sido obtida apenas uma.

Analisando a primeira regra colocada na Tabela 8 temos como antecedente Bananas Biológicas e como consequente Morangos Biológicos com o suporte de 0,023. Isto significa que 2,3 % das compras têm no carrinho Bananas Biológicas (o antecedente). Como confiança temos o valor de 0,282 o que significa que 28,2% das vezes que os clientes compraram Bananas Biológicas também compraram Morangos Biológicos. Por último o

valor do *lift* sendo de 2,392 significa que os itens estão correlacionados positivamente e são dependentes.

Tabela 8 - Regras de Associação para o conjunto de dados Train

Regra	Suporte	Confiança	*Lift*
Bananas Biológicas → Morangos Biológicos	0,023	0,282	2,392
Bananas Biológicas → Espinafre Bebé Biológico	0,017	0,228	1,937
Bananas Biológicas → Framboesas Biológicas	0,014	0,321	2,704
Bananas Biológicas → Abacate Hass Biológico	0,018	0,332	2,813
Banana → Morangos	0,015	0,30	2,102
Morangos Biológicos → Framboesas Biológicas	0,013	0,301	3,627
Morangos Biológicos → Abacate Biológico	0,012	0,211	2,546
Banana → Espinafre Bebé Biológico	0,015	0,204	1,432
Limas → Banana	0,010	0,221	1,546
Limão → Banana	0,016	0,265	1,859
Banana → Abacate Biológico	0,017	0,30	2,096
Limas → Limão	0,012	0,264	4,264

Para os parâmetros previamente testados não foi gerada nenhuma regra de associação do tipo A, B → C. Esta situação pode estar relacionada com o facto de a média de produtos por compra ser de 10 e a mediana de 8. Constata-se que são várias as compras que apresentam entre 4 a 6 produtos o que reduz a probabilidade de serem geradas regras do tipo A, B → C pois se existirem poucos produtos na cesta de compras para valores de suporte mais elevados a probabilidade de encontrar regras A→ B é muito mais elevada do que encontrar regras A, B → C.

Visto não se ter encontrado regras A,B→C neste conjunto de dados para o suporte, confiança e *lift* definidos e como o conjunto de dados apresentava maioritariamente compras com número de produtos de 1 a 5 , foi analisado se para o mesmo conjunto de

dados, mas apenas compras com 11 ou mais produtos eram geradas regras de associação A, B→ C. A modificação necessária de efetuar para que fossem consideradas apenas as compras com 11 ou mais produtos foi realizada com o *script* disponível no anexo 8.10.

Após restringir o conjunto de dados para analisar apenas compras com 11 ou mais produtos, o mesmo ficou reduzido a 52848 compras, a estas aplicaram-se os parâmetros definidos anteriormente (que geraram as 12 regras de associação). Verificou-se que para este conjunto de dados modificado foram geradas 90 regras, sendo geradas 2 regras A, B→ C. O aumento das regras geradas para os mesmos valores de suporte, confiança e *lift* deve-se ao facto de se ter reduzido a amostra e passado a considerar compras com 11 ou mais produtos, existindo uma maior probabilidade de geração de regras A→B pois estamos perante compras com um número de produtos mais significativo. Ao mesmo tempo que a probabilidade de encontrar regras A→ B aumentou, também aumentou a probabilidade de encontrar A, B→C pois num carrinho de compras com muitos produtos a probabilidade de se comprar A e B também compra C aumenta pois são mais produtos por compra.

As 2 regras de associação do tipo A, B→ C encontram-se identificadas na Tabela 9.

Tabela 9 - Regras de Associação A, B --> C para o conjunto de dados Train considerando 11 ou mais produtos por compra

Regra	**Suporte**	**Confiança**	***Lift***
Bananas Biológicas, Morangos Biológicos → Abacate Hass Biológico	0,012	0,269	2,656
Bananas Biológicas, Morangos Biológicos → Framboesas Biológicas	0,011	0,24	2,304

Analisando a primeira regra temos que em 1,2 % das compras temos clientes que compraram Bananas Biológicas e Morangos Biológicos, sendo que cerca de 27% das vezes que compraram Bananas Biológicas e Morangos Biológicos também compraram Abacate Hass Biológico. O *Lift* de 2,656 garante que é uma regra forte em que os produtos estão correlacionados.

Procurou-se encontrar mais regras, quer no formato A → B ou A, B → C pelo que se fez uma alteração nos parâmetros para:

- Mínimo Suporte 0,006; Mínimo Confiança 0,4 e mínimo *Lift* 1,1 → Geradas 11 Regras de Associação

Embora o valor de suporte mínimo tenha sido reduzido, o valor da confiança foi aumentado de modo a obter uma seleção de regras em que a presença em comum dos itens que

constituem a regra é mais elevada (próximo da metade das compras onde o antecedente está).

A seguir encontra-se a Tabela 10 onde são apresentadas as regras geradas.

Tabela 10 - Regras de Associação com Mínimo Suporte 0,006; Mínimo Confiança 0,4 e mínimo Lift 1,1 para o conjunto Train com 11 ou mais produtos por compra

Regra	Suporte	Confiança	*Lift*
Bananas Biológicas → Laranja Biológica	0,011	0,404	2,245
Bananas → Maça Biológica Fuji	0,017	0,408	1,759
Peras de Bartlett → Bananas	0,007	0,418	1,8
Bananas Biológicas, Abacate Hass Biológico→ Limão Biológico	0,006	0,5	2,784
Bananas Biológicas, Morangos Biológicos→ Framboesas Biológicas	0,011	0,414	2,304
Bananas Biológicas, Morangos Biológicos→ Pepino Biológico	0,008	0,434	2,413
Bananas Biológicas, Morangos Biológicos →Abacate Hass Biológico	0,012	0,478	2,656
Bananas Biológicas, Abacate Hass Biológico→ Espinafre Bébe Biológico	0,008	0,419	2,336
Bananas Biológicas, Abacate Hass Biológico→ Framboesas Biológicas	0,009	0,533	2,964
Bananas Biológicas, Abacate Hass Biológico→ Pepino Biológico	0,006	0,468	2,603
Limas, Limao → Abacate Biológico	0,007	0,4	3,533

Posteriormente analisou-se o ficheiro com o histórico de compras que apresentava 3214874 compras efetuadas. O ficheiro foi processado da mesma forma que o ficheiro com os dados *Train*.

Mais uma vez, começou-se por escolher valores de suporte mais elevado, mas como não foi possível obter regras de associação, os valores tiveram que ir diminuindo. O *lift* mínimo nunca apresentou valores inferiores a 1.1 pois pretendia-se regras fortes onde os produtos estivessem correlacionados.

Os resultados foram os seguintes:

- Mínimo Suporte 0,05; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,04; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,03; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,02; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,01; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Geradas 14 Regras de Associação
- Mínimo Suporte 0,01; Mínimo Confiança 0,2 e mínimo *Lift* 1,1 → Geradas 11 Regras de Associação

De seguida colocam-se na Tabela 11 as 11 regras de associação geradas. Pretendeu-se obter regras fortes pelo que se optou pelo conjunto de regras que apresentava maior confiança (mais de 20%), tendo sido descartadas apenas três regras face ao nível de confiança de 10% (que gerou 14 regras).

Tabela 11- Regras de Associação para o conjunto de dados Histórico (Prior)

Regra	Suporte	Confiança	Lift
Bananas Biológicas → Morangos Biológicos	0,019	0,233	1,972
Bananas Biológicas → Espinafres Bebés Biológicos	0,016	0,208	1,764
Bananas Biológicas → Framboesas Biológicas	0,013	0,296	2,504
Bananas Biológicas → Abacate Hass Biológico	0,019	0,292	2,472
Bananas→ Morangos	0,013	0,288	1,962
Morangos Biológicos→ Bananas	0,017	0,212	1,443
Morangos Biológicos → Framboesas Biológicas	0,010	0,247	3
Bananas → Espinafres Bebés Biológicos	0,016	0,212	1,445
Banana → Maça Fuji Biológica	0.011	0,379	2,576
Limão → Bananas	0,013	0,268	1,822
Bananas → Abacate Biológico	0,017	0,302	2,055

Tal como tinha acontecido no conjunto de dados *train,* as regras obtidas estão relacionadas com frutas e vegetais e a maior parte das regras existente num conjunto também está presente no outro. O que significa coerência nos hábitos de consumo dos clientes e que o conjunto de dados *train* poderia ser considerado como uma amostra aceitável para utilização da técnica de redução, caso não fosse possível trabalhar com o conjunto de dados com o histórico das compras.

Tentou-se encontrar outro tipo de regras (A, B → C) e para isso alteraram-se os parâmetros para:

- Mínimo Suporte 0,001; Mínimo Confiança 0,5 e mínimo *Lift* 1,1 → Geradas 2 Regras de Associação nesse formato.

Tabela 12 - Regras de Associação Mínimo Suporte 0,001; Mínimo Confiança 0,5 e mínimo Lift *1,1 conjunto de dados histórico*

Regra	Suporte	Confiança	Lift
Iogurte grego de Morango, Iogurte grego de Pêssego → Iogurte grego de Mirtilo	0,0012	0,594	63,908
Bolachas chocolate glúten free, Água com gás de Limão → Água com gás de Lima	0,0014	0,519	21,980

Verifica-se pela primeira vez a existência de regras de associação que não contemplam nem frutas nem vegetais e que são fortemente correlacionadas. Contudo, apesar de apresentarem valores elevados de confiança, acima de 50%, apenas um analista que estivesse dentro do negócio poderia verificar se as mesmas seriam úteis, uma vez que apresentam valores de suporte bastante baixos.

A mesma interpretação pode ser efetuada para a Tabela 13.

- Mínimo Suporte 0,0025; Mínimo Confiança 0,4 e mínimo *Lift* 1,1 → Geradas 2 Regras de Associação

Tabela 13 - Regras de Associação Mínimo Suporte 0,0025; Mínimo Confiança 0,4 e mínimo Lift 1,1 conjunto de dados histórico

Regra	Suporte	Confiança	Lift
Iogurte grego de Morango → Iogurte grego de Mirtilo	0,0029	0,449	48,343
Bananas Biológicas, Abacate Hass Biológico → Framboesas Biológicas	0,0035	0,442	3,748

Uma vez que ao processar este ficheiro também não se encontrou regras do tipo A, B → C com os parâmetros utilizados, fez-se o mesmo que anteriormente tinha sido feito com o ficheiro *train*. Processou-se novamente o ficheiro, mas teve-se em consideração apenas as compras com 11 ou mais produtos. Ficamos com 1212743 compras.

Aplicaram-se várias combinações de parâmetros obtendo-se as seguintes regras:

- Mínimo Suporte 0,05; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Não foram geradas Regras de Associação
- Mínimo Suporte 0,04; Mínimo Confiança 0,1 e mínimo *Lift* 1,1 → Gerada 1 Regra de Associação
- Mínimo Suporte 0,03; Mínimo Confiança 0,2 e mínimo *Lift* 1,1 → Geradas 5 Regras de Associação, mas apenas Regras A→ B
- Mínimo Suporte 0,02; Mínimo Confiança 0,2 e mínimo *Lift* 1,1 → Geradas 14 Regras de Associação, mas apenas Regras A→ B
- Mínimo Suporte 0,01; Mínimo Confiança 0,2 e mínimo *Lift* 1,1 → Geradas 66 Regras de Associação, uma das regras A, B →C

A regra obtida referente a A, B → C encontra-se refletida na Tabela 14.

Tabela 14 - Regra de Associação A, B --> C para conjunto de dados histórico considerando compras com 11 ou mais produtos

Regra	Suporte	Confiança	*Lift*
Bananas Biológicas, Morangos Biológicos→ Abacate Hass Biológico	0,012	0,288	2,196

Baixaram-se os parâmetros para verificar a existência de mais regras A, B→ C

- Mínimo Suporte 0,008; Mínimo Confiança 0,2 e mínimo *Lift* 1,1 → Geradas 116 Regras de Associação, 5 das regras são A, B →C , tal como se podem verificar na Tabela 15

Tabela 15- Regras de Associação A, B--> C Mínimo Suporte 0,008; Mínimo Confiança 0,2 e mínimo Lift 1,1 para conjunto de dados histórico

Regra	Suporte	Confiança	Lift
Bananas Biológicas, Morangos Biológicas → Espinafres Bebés Biológicos	0,008	0,206	1,721
Bananas Biológicas, Morangos Biológicos → Framboesas Biológicas	0,008	0,218	2,11

Bananas Biológicas, Morangos Biológicos → Abacate Hass Biológico	0,011	0,288	2,196
Bananas Biológicas, Abacate Hass Biológico → Espinafres Bebés Biológicos	0,009	0,286	2,117
Bananas Biológicas, Abacate Hass Biológico → Framboesas Biológicas	0,008	0,314	2,644

- Geraram-se também regras para os seguintes parâmetros Mínimo Suporte 0,002; Mínimo Confiança 0,5 e mínimo *Lift* 1,1 → Geradas 5 Regras de Associação. Os testes onde estas regras foram obtidas são apresentados na secção 5.5.

Tabela 16 - Regras de Associação Mínimo Suporte 0,002; Mínimo Confiança 0,5 e mínimo Lift 1,1 para conjunto de dados histórico com mais de 10 produtos por compra

Regra	Suporte	Confiança	Lift
Iogurte Grego de Morango, Iogurte Grego de Pêssego → Iogurte Grego Mirtilo	0,0026	0,60	33,44
Banana, Morangos → Maça Biológica Fuji	0,0025	0,51	2,16
Maça Honeycrisp, Banana → Morangos	0,0027	0,52	2,20
Água com gás Toranja, Água com gás Limão → Água com gás Lima	0,0022	0,50	13,24
Bananas Biológicas, Morangos Biológicos, Abacate Biológico Hass → Framboesas Biológicas	0,003	0,50	2,80

Para estes valores verificam-se fortes dependências entre produtos, sendo de salientar valores de *lift* obtidos de 33 e 13. Constata-se também uma forte aquisição frequente e em conjunto de produtos biológicos e verifica-se uma regra de A, B, C → D.

5.4 Comparação de Resultados obtidos com os dados de Treino e com os dados completos

Constata-se, como seria de esperar, que os produtos mais adquiridos pelos clientes são os que surgem em mais regras de associação, tanto no conjunto de dados *train* como no histórico. Não foi encontrada nenhuma regra inesperada nas regras analisadas, do género

Fraldas → Cerveja. Contudo, os resultados revelam coerência nos hábitos de consumo. As regras obtidas são fortes, estando os produtos fortemente correlacionados. Relativamente aos valores de suporte e confiança e se a geração das regras será útil para a tomada de decisão, apenas um analista experiente e com forte conhecimento do negócio poderá afirmar o mesmo.

- Na Tabela 17 colocam-se em comparação as regras obtidas para o conjunto de dados Train (12 Regras) e os dados Histórico (11 Regras) com os parâmetros Mínimo Suporte 0,01; Mínimo Confiança 0,2 e mínimo Lift 1,1.

Tabela 17 - Comparação Regras de Associação obtidas Conjunto Train e Conjunto Histórico

Regras Conjunto *Train*	Suporte	Confiança	Regras Conjunto Histórico	Suporte	Confiança
Bananas Biológicas → Morangos Biológicos	0,023	0,282	Bananas Biológicas → Morangos Biológicos	0,019	0,233
Bananas Biológicas → Espinafre Bebé Biológico	0,017	0,228	Bananas Biológicas → Espinafres Bebés Biológicos	0,016	0,208
Bananas Biológicas → Framboesas Biológicas	0,014	0,321	Bananas Biológicas → Framboesas Biológicas	0,013	0,296
Bananas Biológicas → Abacate Hass Biológico	0,018	0,332	Bananas Biológicas → Abacate Hass Biológico	0,019	0,292
Banana → Morangos	0,015	0,30	Bananas→ Morangos	0,013	0,288
Morangos Biológicos → Framboesas Biológicas	0,013	0,301	Morangos Biológicos → Framboesas Biológicas	0,010	0,247
Morangos Biológicos → Abacate Biológico	0,012	0,211	Morangos Biológicos→ Bananas	0,017	0,212
Banana → Espinafre Bebé Biológico	0,015	0,204	Bananas → Espinafres Bebés Biológicos	0,016	0,212
Limas → Banana	0,010	0,221	Banana → Maça Fuji Biológica	0.011	0,379
Limão → Banana	0,016	0,265	Limão → Banana	0,013	0,268
Banana → Abacate Biológico	0,017	0,30	Banana → Abacate Biológico	0,017	0,302
Limas → Limão	0,012	0,264			

Assinalaram-se noutra cor as regras idênticas obtidas para ambos os conjuntos. Verificam-se ligeiras flutuações nos valores de suporte e confiança, mas nada de muito significativo. Das 11 regras, nove são as mesmas tanto para o conjunto *train* como para o conjunto histórico. Mais uma vez, se verifica a coerência nos dados.

Para validação do modelo, existindo coerência tanto entre os resultados obtidos para ambos os conjuntos como na geração de regras em que os produtos mais vezes

comprados são os que geraram mais regras de associação, pode-se considerar que o modelo gerado apresenta valores confiáveis e pode ser utilizado em futuras implementações.

5.5 **Avaliação**

Além dos resultados indicados anteriormente, pretendeu-se também avaliar, através de um conjunto mais alargado de testes, os resultados obtidos, quantificados em número de regras de associação geradas, com configurações distintas do modelo. Pretende-se com esta avaliação descrever o comportamento do modelo com o aumento e diminuição da quantidade de compras, das características das compras consideradas (quantidade de produtos por compra) e do nível de confiança pretendido.

Considerando a dimensão significativa do *dataset* em estudo (*Prior*), nos testes apresentados nesta secção foram utilizados valores de suporte mínimo entre 1% e 0,1%, com intervalos de 0,1%. Para cada um desses valores foi ainda testada a utilização do parâmetro da confiança mínima de 20%, 25% e 50%, continuando a manter-se a aplicação do *lift* mínimo de 1,1, como anteriormente.

Para cado um dois dois *datasets* (Train e Prior), foram criados subconjuntos com algumas compras filtradas, nomeadamente apenas compras com mais do que dez produtos (identificado como “>10”) e apenas compras com mais do que vinte produtos (identificado como “>20”). As dimensões dos conjuntos de dados *train* e *prior* na sua totalidade, considerando apenas compras com mais de dez produtos e considerando compras com mais de vinte produtos podem ser visualizadas na Tabela 18.

Tabela 18 - Número de compras dos conjuntos de dados

Train			Prior		
Todas	>10	>20	Todas	>10	>20
131 209	52 848	14 439	3 214 874	1 212 743	303 410

Como foi visto anteriormente (na secção 5.4), com a alteração (melhoria) dos resultados resultante da seleção das compras com mais do que dez produtos, pretende-se também explorar a informação obtida com uma granularidade mais fina (nas diferentes configurações) a qual permitirá revelar novas regras, menos evidentes e sobre produtos de menor consumo que podem também ser importantes.

Antes de apresentar os resultados obtidos nos diversos conjuntos de testes, apresentam-se na Tabela 19 as contagens mínimas (contagem a partir da qual é considerado frequente), em valor absoluto, que cada *itemset* deve ter para cada um dos valores de suporte mínimo testados. Pretende-se mostrar que, em especial no *dataset Prior*, estas pequenas percentagens correspondem a contagens significativas de compras, que podem ser relevantes em produtos de menor circulação ou segmentos de produtos.

Tabela 19- Número de Compras a que corresponde o Suporte Mínimo em cada subconjunto

Suporte Mínimo	Train			Prior		
	Todas	>10	>20	Todas	>10	>20
1%	1312	528	144	32148	12127	3034
0,9%	1180	475	129	28933	10914	2730
0,8%	1049	422	115	25718	9701	2427
0,7%	918	369	101	22504	8489	2123
0,6%	787	317	86	19289	7276	1820
0,5%	656	264	72	16074	6063	1517
0,4%	524	211	57	12859	4850	1213
0,3%	393	158	43	9644	3638	910
0,2%	262	105	28	6429	2425	606
0,1%	131	52	14	3214	1212	303

Na Tabela 20 apresentam-se os resultados do primeiro conjunto de testes onde se utilizou uma confiança mínima de 20%.

Tabela 20- Contagem de Regras geradas com confiança mínima de 20%

Suporte Mínimo	Train			Prior		
	Todas	>10	>20	Todas	>10	>20
1%	12	90	379	11	66	299
0,9%	16	105	431	14	89	347
0,8%	21	132	516	15	116	412
0,7%	31	160	651	21	139	529
0,6%	42	199	874	27	165	693
0,5%	53	252	1225	37	217	960
0,4%	89	340	1899	69	304	1437
0,3%	128	578	3271	103	476	2468
0,2%	226	1271	7385	187	975	5304
0,1%	863	4622	26539	602	3451	18303

Constata-se que quanto menor for o suporte mais regras de associação serão geradas, como esperado, e se formos restringindo o conjunto de dados, descartando compras que tenham até 10 ou até 20 produtos, o número de regras de associação encontradas aumentará significativamente. Verifica-se que, dado o valor de confiança utilizado nos testes apresentados na Tabela 20, de apenas 20%, quando são descartadas as compras até 20 produtos, o número de regras obtido aproxima-se do milhar para valores de suporte

de 0,5% (sendo maior no *dataset train*) e cresce exponencialmente com o decremento do valor do suporte mínimo, ultrapassando a dezena de milhar no valor de 0,1%.

O mesmo exercício foi efetuado, aumentando ligeiramente a confiança para 25%. Os resultados obtidos encontram-se na Tabela 21.

Tabela 21- Contagem de Regras geradas com confiança mínima de 25%

Suporte Mínimo	Train			Prior		
	Todas	>10	>20	Todas	>10	>20
1%	8	59	265	6	41	216
0,9%	12	70	304	7	54	253
0,8%	15	84	373	8	64	307
0,7%	18	97	493	11	73	401
0,6%	21	124	679	14	91	541
0,5%	26	164	979	18	122	761
0,4%	44	229	1574	30	188	1154
0,3%	65	424	2739	45	306	2051
0,2%	119	980	6242	87	677	4412
0,1%	547	3498	22546	321	2456	15027

Comparando os resultados da Tabela 21 com os anteriores (da Tabela 20), verifica-se que algumas regras previamente encontradas foram filtradas (a contagem desceu) por se encontrarem com um valor de confiança entre 20% e 25%. O decréscimo é mais significativo quando são consideradas todas as compras, com a eliminação de aproximadamente 40% das regras no *dataset train*, em média, considerando todos os níveis de suporte testados, e 50%das regras no *dataset prior*. Nas compras com mais de vinte produtos, foram eliminadas cerca de 20% das regras em ambos os *datasets*, o que indica que nesse conjunto de dados, 80% das regras encontradas previamente têm confiança superior a 25%. No entanto, para este nível de confiança, o número de regras obtido continua a ser significativo.

Para facilitar a visualização da evolução do número de regras obtidas nos diversos *datasets*, a Figura 33 e a Figura 34 mostram, para o *dataset train* e *prior*, respetivamente, a evolução no número de regras geradas, para os vários níveis de suporte (entre 1% e 0,1%). Os gráficos mostram que embora o crescimento seja exponencial com a redução do nível mínimo de suporte, esse crescimento é proporcional dentro de cada *dataset* (todo e os subconjuntos “>10” e “>20”). Destacam ainda o “salto” no número de regras obtidas quando se muda entre cada um dos subconjuntos de compras.

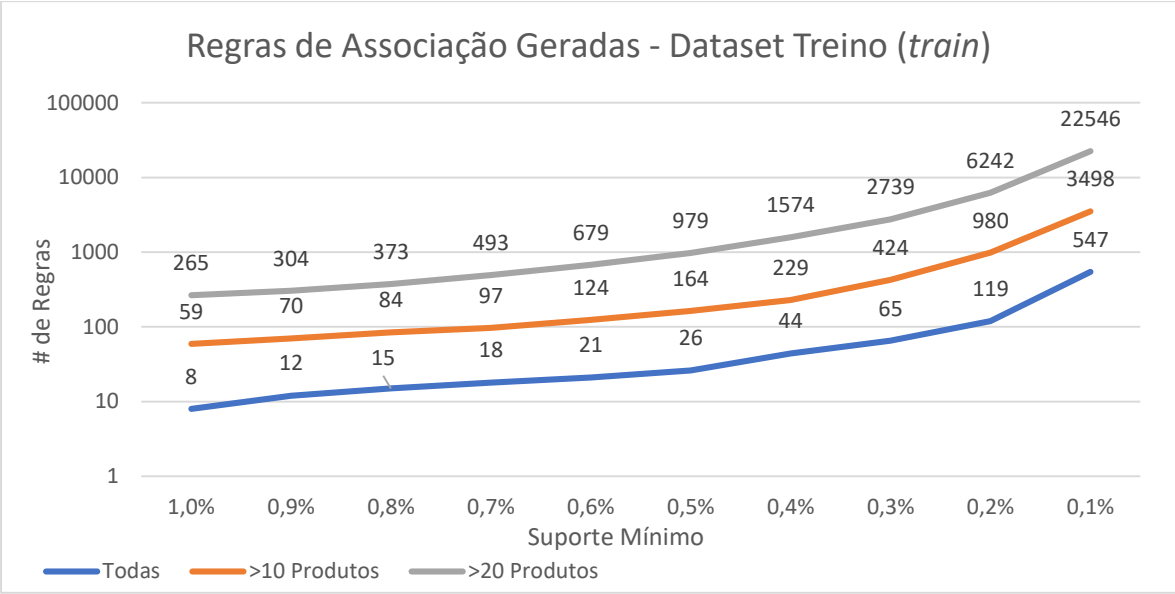

Figura 33 - Gráfico Regras de Associação Geradas - Dataset Treino (confiança=25%)

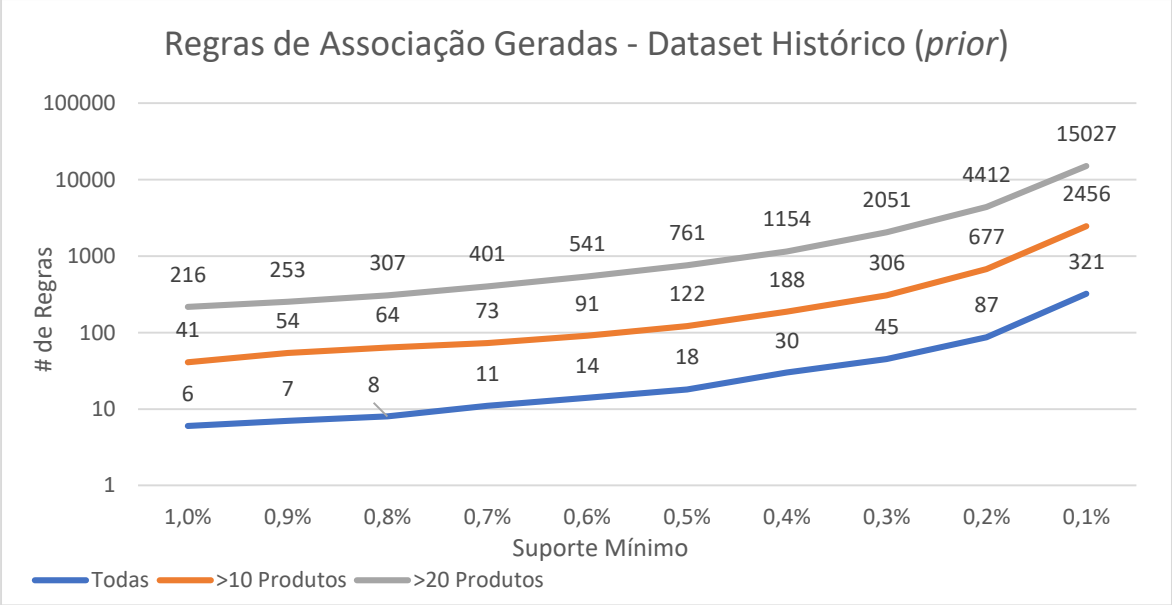

Figura 34 - Gráfico Regras de Associação Geradas - Dataset Histórico (confiança=25%)

Procurou-se então obter os resultados do modelo em que a confiança mínima fosse superior, estabeleceu-se um mínimo de 50% para a confiança. Os resultados obtidos são apresentados na Tabela 22.

Considerando a confiança mínima de 50%, o número de regras obtidas desce drasticamente, sendo quase erradicadas no *dataset prior* com todas as compras, onde apenas são obtidas duas regras. Estes resultados mostram que a quase totalidade das regras obtidas previamente (considerando todas a compras e apenas aquelas com mais

de dez produtos), tem uma confiança inferior a 50%, sobrando apenas cerca de 2% das regras no *dataset train* (>10) e 0,3% *no dataset prior* (>10). Nos *datasets* com as compras de maior dimensão (>20), mantêm-se cerca de 10% das regras encontradas no nível de confiança anterior no *dataset train* e cerca de 3% no *dataset prior*.

Tabela 22 - Contagem de Regras geradas com confiança mínima de 50%

Suporte Mínimo	Train			Prior		
	Todas	>10	>20	Todas	>10	>20
1%	0	0	13	0	0	4
0,9%	0	1	18	0	0	6
0,8%	0	1	25	0	0	6
0,7%	0	1	43	0	0	9
0,6%	0	2	57	0	0	13
0,5%	0	2	86	0	0	20
0,4%	1	5	140	0	0	33
0,3%	1	11	297	0	1	67
0,2%	1	28	796	0	5	175
0,1%	11	169	4477	2	51	885

O gráfico da Figura 35 permite visualizar a evolução do número de regras obtido nos subconjuntos de compras com mais de vinte produtos em ambos os *datasets*. Os números aqui apresentados já tornam mais viável fazer uma exploração das regras obtidas (expecto no conjunto de treino, para o suporte mínimo de 0,1).

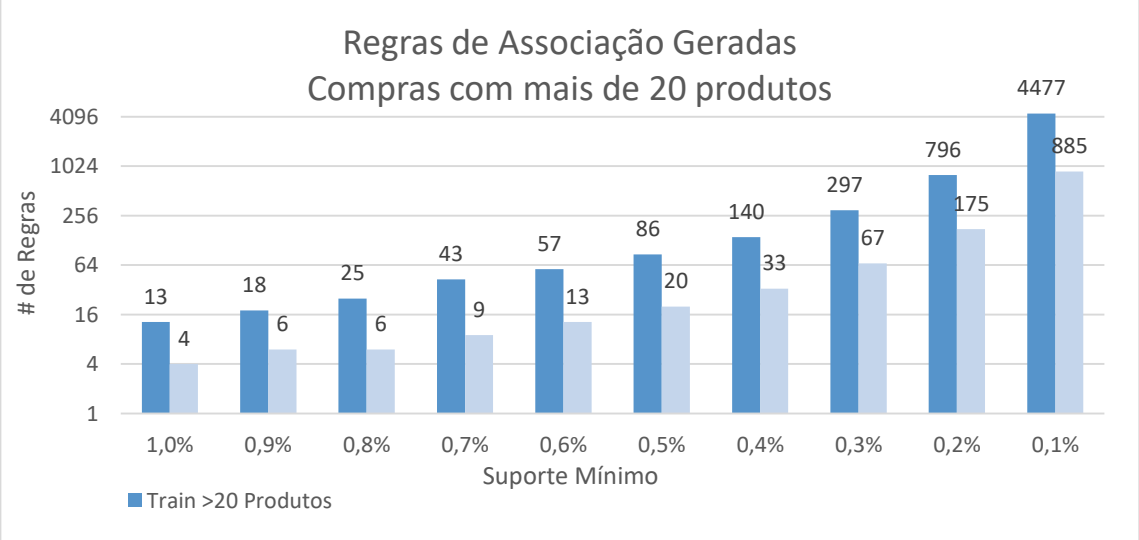

Figura 35 - Gráfico Regras de Associação Geradas – Compras com mais de 20 produtos (confiança=50%)

O facto do conjunto de dados *train* apresentar mais regras geradas estará relacionado com a menor dimensão da amostra, sendo uma amostra mais reduzida consegue-se gerar mais regras do que numa amostra de maiores dimensões, para o mesmo nível mínimo de

suporte (percentual), como comprova a evolução do número de regras quando se descartam as compras até dez produtos e depois até vinte produtos. No gráfico da Figura 35, por exemplo, é atingido o valor de 4477 regras porque, para aquele nível de suporte, um item é considerado frequente quando surge em apenas 14 compras (ver Tabela 19).

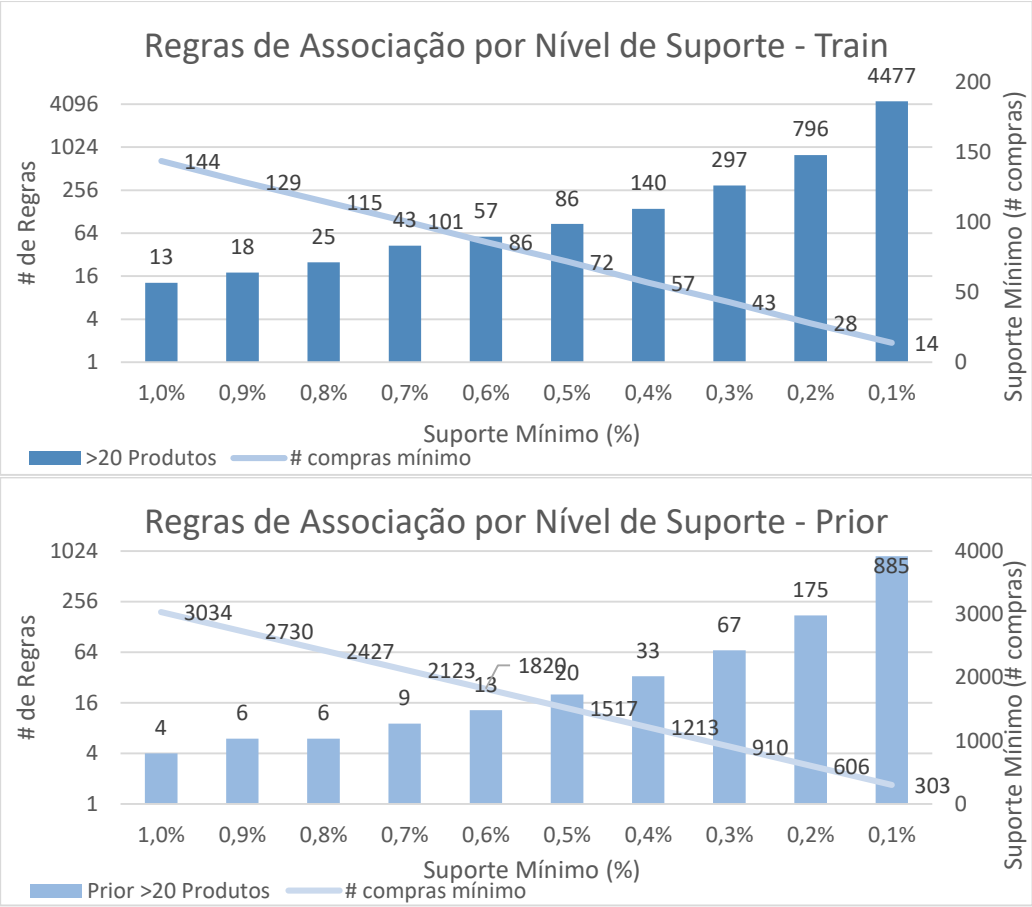

Figura 36 – Comparação do número de regras obtidas com o valor real de suporte mínimo

A Figura 36 apresenta dois gráficos com a representação do número de compras mínimo a que corresponde cada nível de suporte (em percentagem) e o número de regras obtidas em cada *dataset* (baseado no subconjunto com mais de vinte produtos por compra e confiança de 50%).

Seria útil saber mais informação sobre o *dataset* em estudo e/ou segmentar as regras obtidas por departamento, ou outras dimensões, de modo a decidir qual a granularidade que faz sentido explorar. Contrariamente ao que foi feito na secção 5.3, o conteúdo das regras resultantes deste conjunto de testes não foi explorado nesta investigação. Pretendeu-se avaliar o comportamento do modelo com a variação dos parâmetros e dos conjuntos de carrinhos de compras em que é aplicado. Os resultados dos diversos testes realizados foram apresentados e comentados nesta secção permitindo caracterizar o

output esperado do modelo nas diversas combinações de nível de suporte mínimo, confiança, dimensão e características do conjunto de compras fornecido como *input*. Ainda assim, de modo a facilitar futuros desenvolvimentos desta investigação, foi criado um *script* para guardar as regras geradas e outro para ler esse output e criar uma representação legível das mesmas. Os *scripts* e dois exemplos de outputs dos testes realizados nesta secção são apresentados no anexo 8.11.

A falta de informação adicional sobre os dados, aliada ao prazo para conclusão desta dissertação, impediram a exploração mais aprofundada dos *outputs* dos testes apresentados nesta secção, os quais podiam, pelo menos ser explorados, eventualmente com a aplicação de outras técnicas de *data mining*, como o *clustering*, para agrupar as regras obtidas pela sua semelhança. Seria também interessante ter conseguido utilizar algoritmos alternativos (referidos em 3.3.1.4) e comparar a sua performance porque, em algumas das configurações, o algoritmo Apriori, na implementação utilizada, aproximou-se da uma hora de tempo total de processamento para obter o conjunto de Regras de Associação.

6 Conclusões

O principal objetivo desta investigação era demonstrar a aplicação da técnica de *Market Basket Analysis* de modo a conseguir identificar os produtos que os clientes compravam frequentemente em conjunto através da geração de regras de associação. Identificando esses produtos, outro objetivo do projeto era fornecer sugestões/ ideias para ações de *marketing*. Denota-se claramente, quando considerando os resultados obtidos com os valores dos parâmetros mais elevados, que os produtos mais adquiridos e que geraram melhores regras de associação são as frutas e os vegetais, particularmente os de origem biológica.

Assim sendo, a nível de *marketing* propõe-se a realização de campanhas e descontos promocionais tendo como principal foco os produtos biológicos. Através de campanhas de descontos pode-se promover outros tipos de produtos biológicos, pode-se promover compras por impulso lançando campanhas por períodos limitados tendo em consideração que o dia 0 e 1 da semana são os dias em que se registam mais compras e o período da tarde é a altura em que mais compras se realizam. Ao selecionar um determinado produto aparecer a indicação que os utilizadores que adquirem esse produto frequentemente adquirem também o produto B e apresentar algum desconto. Ao selecionar por exemplo o produto Banana aparecer ao lado como escolha o produto Banana biológica. Outra estratégia de *marketing* a considerar será a flutuação dos preços, se a empresa está ciente que o produto Bananas Biológicas é frequentemente adquirido em conjunto com o produto Morangos Biológicos pode escolher aumentar moderadamente o preço de um dos produtos ou aplicar um desconto na aquisição, consoante o tipo de estratégia que pretende implementar no momento. Da mesma forma, pode tirar-se proveito deste conhecimento sobre os produtos que costumam ser comprados em conjunto e criar companhas promocionais que juntem a esses conjuntos os produtos com menos procura, ou novos no mercado, aumentando as suas vendas e/ou para dar a conhecer esses produtos aos clientes.

Relativamente aos objetivos propostos para este projeto considera-se que foram alcançados com sucesso pois:

- Foi feita uma alargada pesquisa sobre os algoritmos e metodologias associadas a projetos de *Data Analytics*, com a aquisição de novo conhecimento que nunca tinha sido abordado no percurso académica da estudante;
- Foi abordado um problema relevante, com expressão e atualidade na comunidade científica;

- O conjunto de dados utilizado, embora não seja novo, ou obtido em particular para esta investigação, permitiu, ainda assim, fazer uma extensa análise exploratória na sua caracterização permitindo alargar o conhecimento sobre o comportamento do negócio da empresa que disponibilizou os dados;
- A estudante expandiu o seu conhecimento pela exploração de algumas tecnologias no desenvolvimento desta investigação;
- Foram analisados e preparados os dados de modo a poder utilizar um algoritmo de *data mining* para obter Regras de Associação que mostram quais os produtos que são habitualmente adquiridos em simultâneo;
- Foram testadas diversas configurações do algoritmo Apriori e documentadas as alterações nos resultados obtidos, mostrando que, dependendo da granularidade pretendida e do conjunto de dados sobre os quais o algoritmo é aplicado, os resultados obtidos podem sofrer alterações significativas, neste caso em concreto, no número de Regras de Associação que são geradas;
- Embora sem conhecimento adicional sobre o negócio, foram sugeridas, neste capítulo, algumas formas de rentabilizar o conhecimento gerado por este trabalho;
- Foi feita uma comparação das Regras de Associação com maior expressão, obtidas do *dataset* de treino e do *dataset* completo, o que permitiu verificar que, não sendo necessária uma granularidade fina, podem ser aplicados os algoritmos de *data mining* apenas numa seleção dos dados, desde que representativos da realidade completa, evitando assim os problemas computacionais associados ao crescimento dos dados.

O conhecimento adquirido ao longo do trabalho desenvolvido nesta investigação, ou mesmo apenas a parte que está descrita nesta dissertação, permitem que a mesma metodologia possa ser replicada pela estudante, ou quem consulte esta dissertação, noutros problemas semelhantes propostos por outras organizações, especialmente locais, para as quais os resultados podem ser ainda mais valiosos caso não tenham ainda adotado a utilização de tecnologias de *data analytics* como suporte ao negócio.

7 Limitações e Recomendações para trabalhos futuros

Este projeto teve como propósito final a geração de regras de associação de uma forma generalista e abrangente, sem qualquer objetivo específico além da identificação desses padrões de compra, optando-se por analisar a totalidade das compras e compras com mais de 10 produtos de modo a evitar que essas compras de menor dimensão afetassem negativamente os resultados do algoritmo. Outras transformações e análises poderão vir a ser realizadas no futuro, optando, por exemplo, por gerar regras que refletissem comportamentos em determinado dia da semana, determinado período temporal, estudando apenas os clientes que fazem mais compras ou inclusive excluindo determinados produtos como frutas e vegetais, ou outros tipos de segmentação, por forma a verificar a existência de novas regras em segmentos concretos da totalidade dos dados estudados.

Caso existissem dados adicionais acerca dos produtos, retalhistas e clientes considera-se que seria interessante, no futuro, expandir a exploração dos dados das compras realizadas utilizando também essas dimensões. Por exemplo, investigar se existem diferenças nos padrões de compra por retalhista ou segmento de clientes ou extrair outro tipo de informação com outros algoritmos. A empresa que divulgou os dados utilizados nesta dissertação propôs como desafio a criação de um modelo de previsão (*machine learning*) que permitisse prever se na próxima visita um cliente vai, ou não, adquirir um determinado produto. A exploração deste tipo de algoritmos seria certamente desafiante, não só pela sua atualidade como pelo impacto que esse tipo de modelos pode ter nas organizações.

No trabalho desenvolvido nesta investigação o principal desafio foi a aprendizagem de uma nova linguagem de programação – *Python*. Tendo pouco contacto com linguagens de programação ao longo do percurso escolar e laboral, aprender as particularidades do *Python* e das diversas bibliotecas disponíveis para manipulação e processamento de dados (*Pandas, NumPy*, etc) por forma a conseguir efetuar as transformações pretendidas nos *datasets* utilizados para aplicar o algoritmo para a geração das regras de associação (Apriori) foi o mais desafiante.

Ao nível de limitações, uma das mais relevantes foi a performance dos *scripts* utilizados. O tempo de processamento do *script* que converte a matriz (Pandas) numa lista de listas com os mais de 3 milhões de compras, depois de algumas otimizações, gastava cerca de 2 horas, sendo que a aplicação do algoritmo Apriori demorava entre alguns segundos e alguns minutos, dependendo do valor de suporte mínimo definido. Esta demora condicionou significativamente os testes realizados com o *dataset* de maior dimensão, tendo a mesma sido ultrapassada apenas no final da escrita desta dissertação através da

criação de um *script* para a leitura direta dos dados do ficheiro de *input* para a lista de listas, sem recorrer à biblioteca *Pandas*. Uma outra limitação ao trabalho que se pretendia desenvolver nesta investigação foi a impossibilidade de obter dados de uma rede de supermercados nacional, a qual esteve desde o início do projeto em negociação, mas acabou por não se viabilizar atempadamente.

A condicionante referida anteriormente impediu que fossem exploradas versões avançadas (otimizadas) do algoritmo Apriori, não porque o desempenho do algoritmo em si fosse mau globalmente, mas nos conjuntos de maior dimensão o tempo computacional aumentou consideravelmente e permitiria a validação dos resultados. Poderiam ainda ter sido explorados outros algoritmos, os quais apresentam vantagens de performance relativamente ao Apriori, como por exemplo o Eclat ou FP-Growth (Heaton J. , 2016).

Os estudos acima referidos poderiam ser incluídos numa nova iteração do modelo, seguindo a abordagem da Metodologia CRISP. Importa ressalvar que este novo estudo deveria de ser aprovado por um analista com profundo conhecimento acerca do negócio, uma vez que ao procurarmos vamos sempre encontrar padrões escondidos nos dados, importando discernir os que efetivamente são confiáveis e trazem mais valias para o negócio dos que podem levar a conclusões erradas e consequentemente a más decisões de negócio acarretando prejuízos para o mesmo.

Referências Bibliográficas

Agência Lusa. (08 de 05 de 2018). *https://observador.pt/2018/05/08/volume-de-vendas-do-retalho-sobe-38-em-2017-e-supera-patamar-de-20-mil-milhoes-de-euros/*. Obtido de https://observador.pt: https://observador.pt/2018/05/08/volume-de-vendas-do-retalho-sobe-38-em-2017-e-supera-patamar-de-20-mil-milhoes-de-euros/

Aggarwal, C. C. (2015). *Data Mining The Textbook.* Springer.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proc. 20th int. conf. very large data bases, VLDB*, *1215*, pp. 487-499.

Azevedo, A., & M.F.Santos . (2008). KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW. *IADIS European Conference Data Mining 2008*, (pp. 182-185).

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

Borgelt, C. (2005). An Implementation of the FP-growth Algorithm. *OSDM05* (pp. 1-5). ACM New York, NY, USA.

Campomar, M. C. (1991). Do uso de "estudo de caso" em pesquisas para dissertações e teses em Administração. *Revista de Administração*, p. 95-97.

Chakrabarti, S., Cox, E., Frank, E., Güting, R., Han, J., & Jiang, X. (2009). *Data Mining Know It All.* Morgan Kaufmann Publishers.

Data-Driven Science. (31 de Janeiro de 2018). *https://medium.com/@data_driven/python-vs-r-for-data-science-and-the-winner-is-3ebb1a968197*. Fonte: https://medium.com.

Deloitte. (2018). *Global-powers-of-retailing-2018.*

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data Myths, Misconceptions and Methods (Business in the Digital Economy).* Springer.

García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining.* Springer.

Gayathri, B. (2017). Efficient Market Basket Analysis based on FP-Bonsai. *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)*, (pp. 788-792).

Gayathri, B. (Fevereiro de 2017). International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud). *IEEE*, pp. 788-792.

Giudici, P., & Figini, S. (2009). *Applied Data Mining for Business and Industry (Second Edition).* John Wiley & Sons.

Goethals, B. (2004). Memory issues in frequent itemset mining. *ACM Symposium on Applied Computing*, (pp. 530-534).

Grus, J. (2019). *Data Science from Scratch: First Principles with Python* (2 ed.). O'Reilly Media, Inc.

Guohua, W., & Francis, T. E. (2017). Data Minig: Concepts, Applications and Techniques. *ASEAN - Journal on Science and Technology for Development*, pp. 77-86.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts And Techniques 3 edição .* Morgan Kaufmann.

Han, J., Pei, J., & Y. Y. (2000). Mining frequent patterns without candidate generation. *ACM sigmod record, 29*(2), pp. 1-12.

Hayes, B. (Janeiro de 2019). *Programming Languages Most Used and Recommended by Data Scientists.* Acesso em Outubro de 2019, disponível em businessoverbroadway.com: https://businessoverbroadway.com/2019/01/13/programming-languages-most-used-and-recommended-by-data-scientists/

Heaton, J. (2016). Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms. *SoutheastCon 2016* , (pp. 1-7). Norfolk, VA.

Huber, S., Wiemer, H., Schneider, D., & Ihlenfeldt, S. (2018). DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model. *Procedia CIRP*, *79*, pp. 403-408. doi:https://doi.org/10.1016/j.procir.2019.02.106

Instacart. (2017). *https://www.kaggle.com/c/instacart-market-basket-analysis/overview*. Fonte: https://www.kaggle.com/.

Kabir, M. F., Ludwig, S., & Abdullah, A. (2018). Rule Discovery from Breast Cancer Risk Factors using Association Rule Mining. *2018 IEEE International Conference on Big Data (Big Data)*, (pp. 2433-2441).

Kaggle. (1 de 2019). *Kaggle's second annual Machine Learning and Data Science Survey*. Acesso em 10 de 2019, disponível em Kaggle: https://www.kaggle.com/kaggle/kaggle-survey-2018

Kantardzic, M. (2011). *DATA MINING Concepts, Models, Methods, and Algorithms SECOND EDITION.* John Wiley & Sons, Inc.

Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. *International Conference on Computational Modeling and Security (CMS 2016)* (pp. 78-85). Procedia Computer Science.

Lai, C.-P., & Lu, J.-R. (23 de 02 de 2018). Evaluating the efficiency of currency portfolios constructed by the mining. *Asia Pacific Management Review*, pp. 11-20.

Larose, D. T., & Larose, C. D. (2014). *DISCOVERING KNOWLEDGE IN DATA An Introduction to Data Mining (Second Edition).* John Wiley & Sons, Inc.

Maciag, T., Hepting, D., Ślęzak, D., & Hilderman, R. (2007). Mining Associations for Interface Design. Paper presented at the , Berlin, Heidelberg. *Rough Sets and Knowledge Technology. LNCS, volume 4481*, pp. 109-117. Berlin: Springer.

Mochizuki, Y. (11 de 04 de 2016). *https://pypi.org/project/apyori/*. Fonte: https://pypi.org.

Monem, R. I., El-Bastawissy, A., & M. Elwakil, M. (2016). DIRA : A FRAMEWORK OF DATA INTEGRATION USING DATA QUALITY. *International Journal of Data Mining & Knowledge Management Process*, 37-58.

Moro, S., Laureano, R., & Cortez, P. (2011). Using data mining for bank direct marketing: An application of the crisp-dm methodology. *Proceedings of European Simulation and Modelling Conference-ESM'2011* (pp. 117-121). EUROSIS-ETI.

Nidhi, T., D. ,., H. C., & Sunita, N. (2018). Estimating Frequent Products in Shopping Cart Using Data Mining. *2018 IEEE International Conference on Big Data (Big Data)*, (pp. 1560-1564).

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques.* Springer.

Petersen, R. (07 de 11 de 2016). *https://barnraisersllc.com/2016/11/companies-data-mining-business-better/*. Fonte: https://barnraisersllc.com: https://barnraisersllc.com/2016/11/companies-data-mining-business-better/

Piatetsky, G. (2014). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects.* Acesso em 5 de outubro de 2019, disponível em KDnuggets.

Piatetsky, G. (Maio de 2019). *Top Data Science, Machine Learning platforms: Trends and Analysis*. Acesso em 10 de 2019, disponível em kdnuggets.com: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html

Prodanov, C. C., & Freitas, E. C. (2013). *Metodologia do Trabalho Científico: Métodos e Técnicas 2ª Edição.* Feevale. Novo Hamburgo - Rio Grande do Sul.

Provost, F., & Fawcett, T. (2015). *Data Science for Business What you need to know about data mining and data-analytic thinking.* O'reilly .

SAS. (30 de 8 de 2017). *https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en*. Fonte: www.sas.com.

Sharma, V., Stranieri, A., Ugon, J., Vamplew, P., & Martin, L. (2017). An Agile Group Aware Process beyond CRISP-DM: A Hospital Data Mining Case Study. *Proceedings of the International Conference on Compute and Data Analysis* (pp. 109-113). ACM.

Singh, S., Garg, R., & Mishra, P. K. (2018). Performance Optimization of MapReduce-based Apriori Algorithm on Hadoop Cluster. *Computers & Electrical Engineering*, 348-364.

Solnet, D., Boztug, Y., & Dolnicar, S. (2016). An untapped gold mine? Exploring the potential of market basketanalysis to grow hotel revenue. *International Journal of Hospitality Management, 56*, 119-125.

Stanley, J. (03 de 05 de 2017). *https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2*. Fonte: https://tech.instacart.com.

Torgo, L. (2017). *Data Mining with R - Learning with Case Studies.* CRC Press.

Tripathi, N., Darshana, V., Himanshu, C., & Sunita, N. (2018). Estimating Frequent Products in Shopping Cart Using Data Mining. *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies*, (pp. 1560-1564).

Wirth, R., & Jochen Hipp. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining.*

Yuan, X. (2017). An Improved Apriori Algorithm for Mining Association Rules. *AIP Conference Proceedings 1820*, (pp. 080005-1 - 080005-6).

Zaki, M., Parthasarathy, S., Ogihara, M., & Li, W. (1997). *New Algorithms for Fast Discovery of Association Rules.* University of Rochester.

8 Anexos

8.1 Gráfico: Número de Compras consoante a Hora do Dia

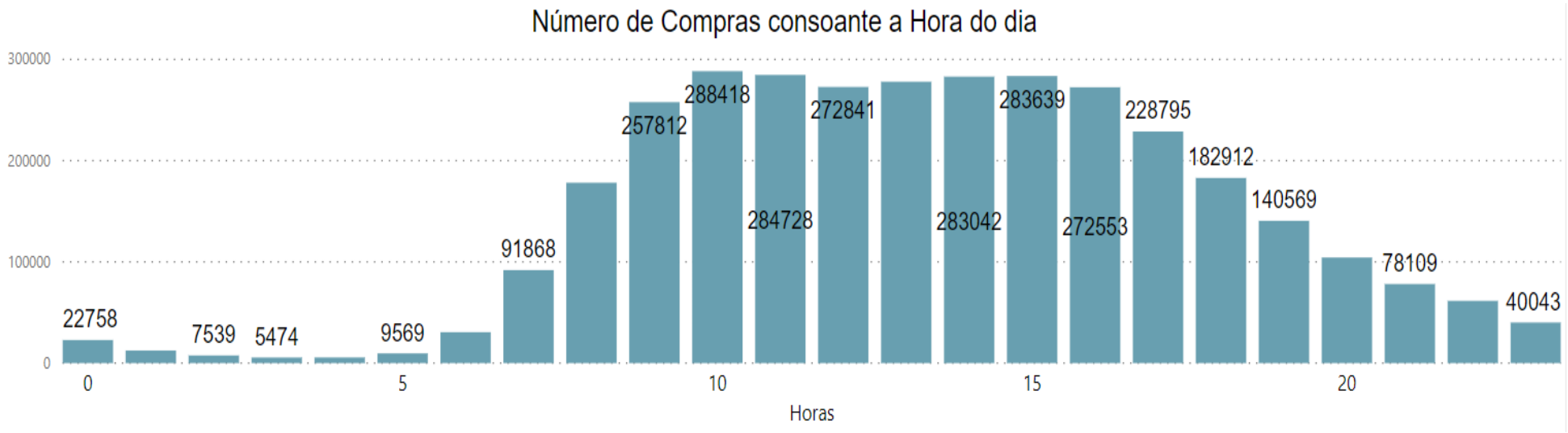

Figura 37 - Anexo 1 Número de Compras consoante a Hora do Dia

8.2 Gráfico: Número de Dias até à próxima Compra

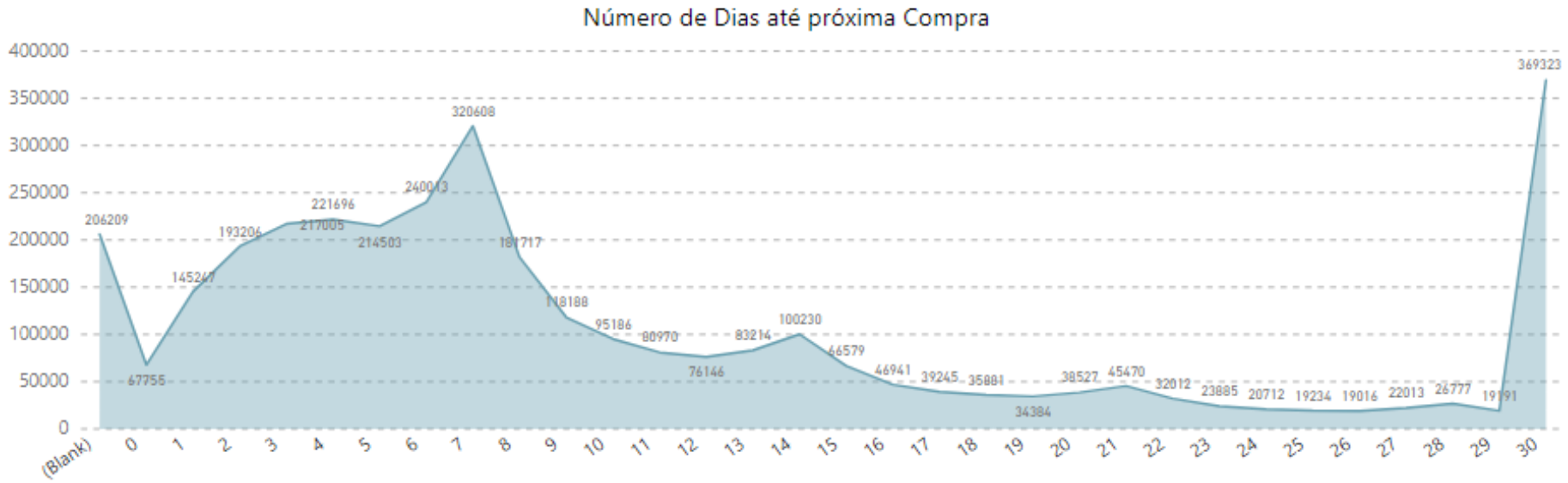

Figura 38- Anexo 2 Gráfico Número de Dias até próxima Compra

8.3 Gráfico: Contagem de Produtos por Departamentos e Corredores

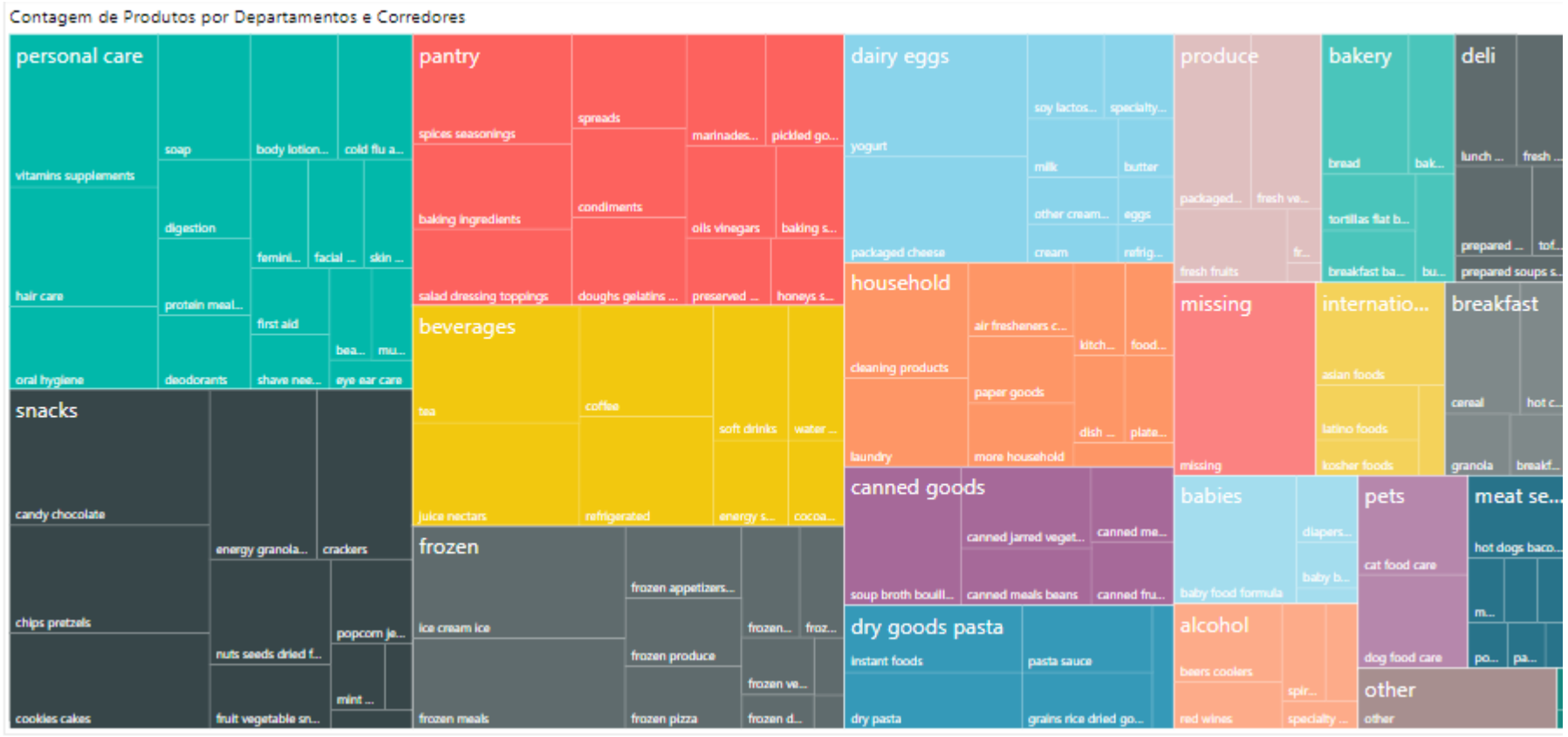

Figura 39 - Anexo 3 Gráfico Contagem de Produtos por Departamentos e Corredores

8.4 Gráfico: Produtos mais adquiridos

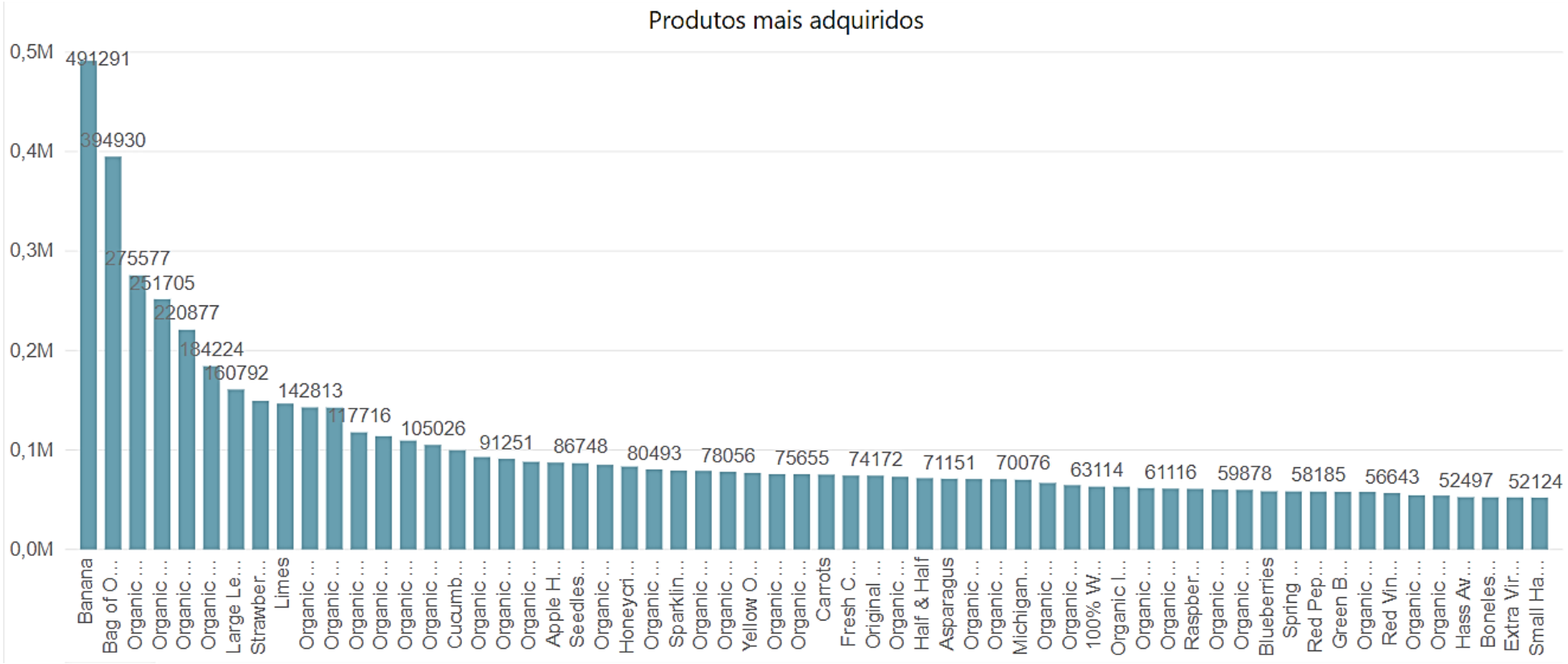

Figura 40 - Anexo 4 Produtos mais adquiridos

8.5 Gráfico: Produtos mais requisitados por Departamentos e Corredores (Top 10)

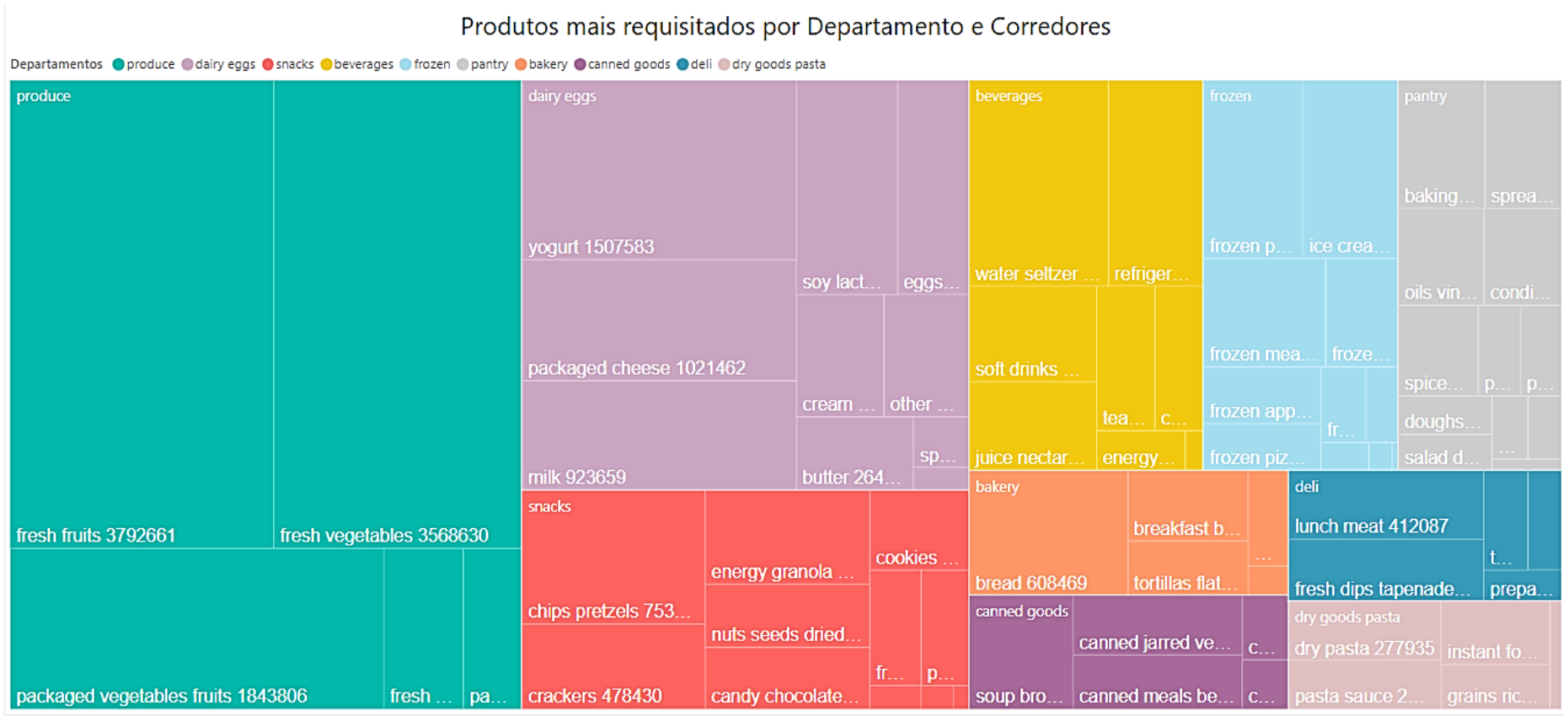

Figura 41- Anexo 5 Produtos mais requisitados por Departamentos e Corredores

8.6 Gráfico: Produtos menos requisitados por Departamentos e Corredores (Top 5)

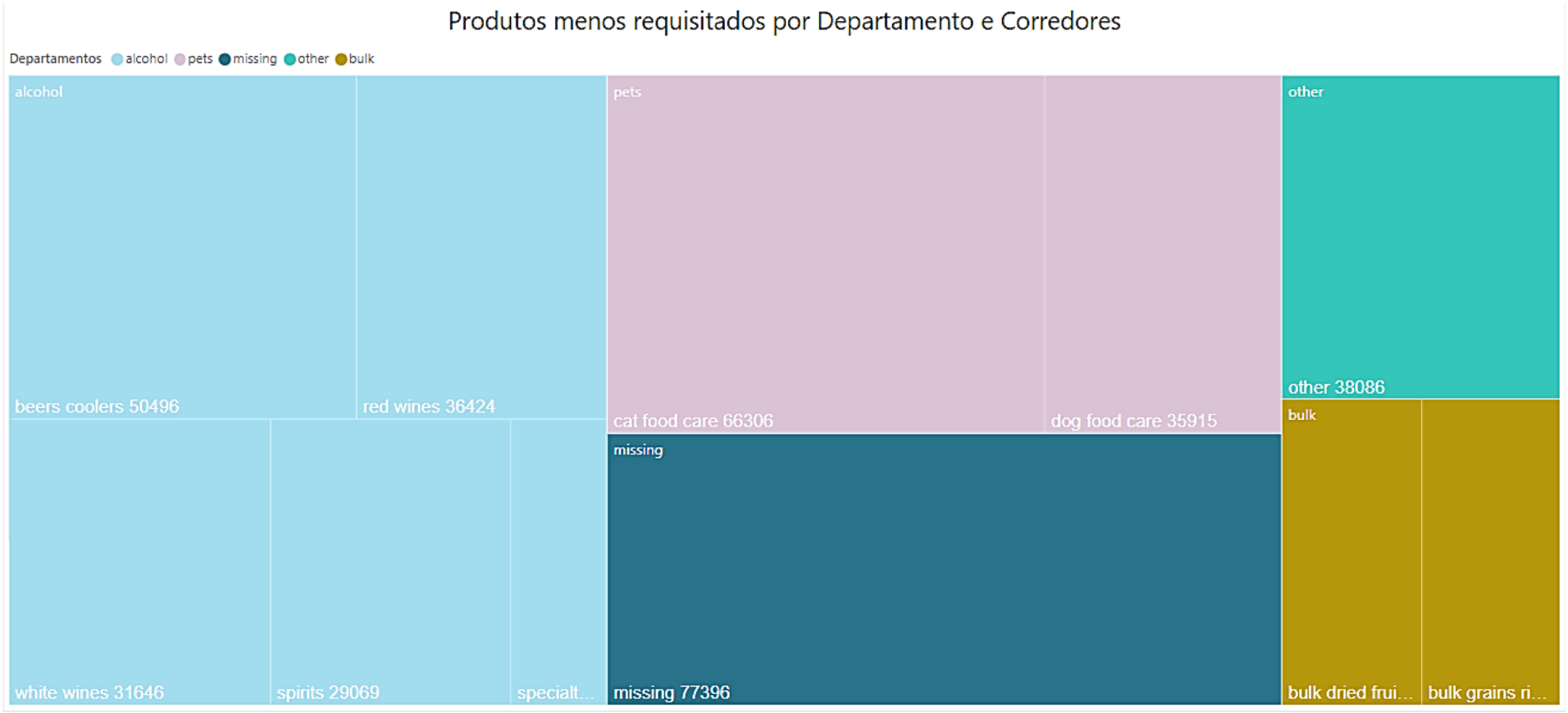

Figura 42 - Anexo 6 Produtos menos requisitados por Departamentos e Corredores

8.7 Gráfico: Número de Produtos por Compra

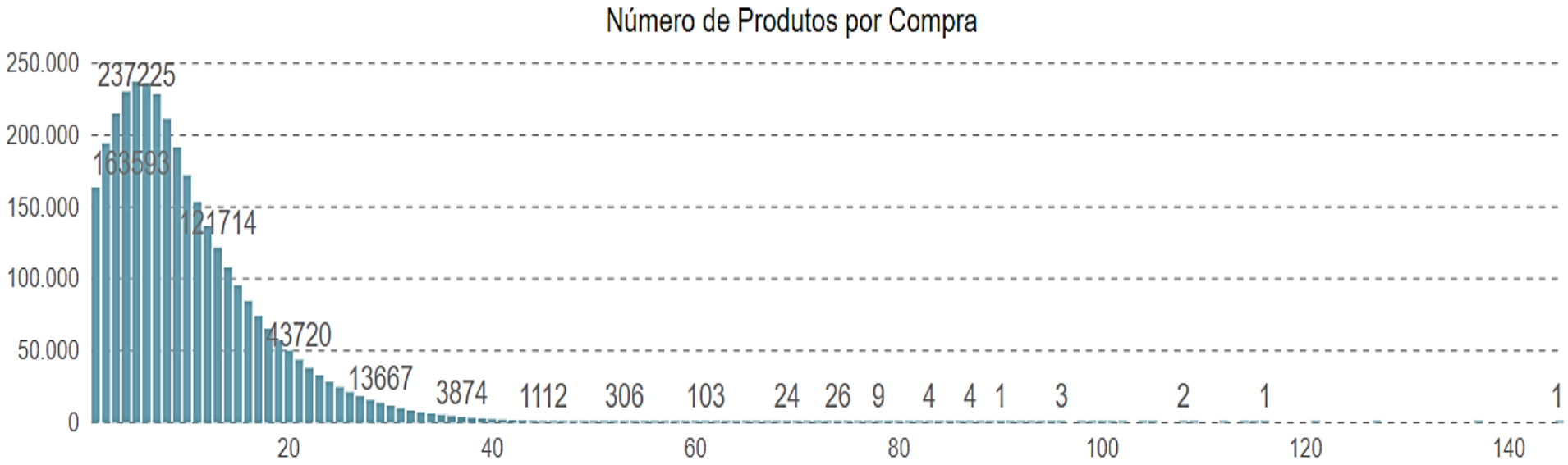

Figura 43 - Anexo 7 Número de Produtos por Compra

8.8 Código Tratamento do Ficheiro

```
# Importação das bibliotecas necessárias

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

from apyori import apriori

# Carregamento do ficheiro CSV para formato panda dataframe e visualização dos dados

store_data =
pd.read_csv('C:\\Users\\JOANA\\Desktop\\TESEBD\\DATA\\order_products__prior.csv')

store_data.head(10)

#Eliminação das counas que não são necessárias

del store_data['add_to_cart_order']

del store_data['reordered']

#Confirmação que colunas foram removidas

store_data.head(10)

# Group By Order_Id e visualização dos dados

df = store_data.groupby(['order_id'])['product_id'].apply(list)

df.head()

# Gravar ficheiro transformado com outro nome sem índice e sem cabeçalho

store_data.to_csv("""C:\\Users\\JOANA\\Desktop\\TESEBD\\DATA\\Remove.csv""",      index=False,
header= False)
```

8.9 Script de Aplicação do Algoritmo Apriori

```
# Importação das bibliotecas necessárias

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from apyori import aprior

# Ficheiro já sem "" e sem [], operação efetuada via Notepad++ através da funcionalidade replace.

#Identificação do range das colunas (valor máximo) para não dar erro ao processar ficheiro pois variam de compra para compra.
col_names=[i for i in range(0,79)]

#Carregamento do ficheiro CSV para formato dataframe (Pandas)
Dados = pd.read_csv('F:\\Tese\\TrainGB.csv', header=None, names=col_names)

# Colocação de todos os dados como float para melhor performance
Dados1 = Dados.astype(float)


# Conversão dos dados do dataframe para uma lista de listas eliminando a grande quantidade de células com valores NaN do dataframe para melhor performance e ter os dados necessários para o algoritmo Apriori.
records = []

#Percorrer todas as linhas do dataframe (correspondem às compras/cestos)

for i in range(0, 131208):
    rec=[]  #criar nova lista para adicionar  produtos da compra atual
    #Percorrer todas as colunas do dataframe
    for j in range(0, 79):
        if not np.isnan(Dados1.values[i,j]) :
            rec.append(Dados1.values[i,j]) #adiciona apenas os valores não-nulos (identificador do produto)
    records.append(rec) #No final adiciona a última compra à lista de compras/cestos


# Utilização do algoritmo com a definição dos parâmetros mínimos de suporte, confiança e lift e visualização das regras geradas
association_rules = apriori(records, min_support=0.05, min_confidence=0.1, min_lift=1.1)

association_results = list(association_rules)

# Visualização das regras geradas

print (association_results)
```

8.10 Aplicação do Algoritmo Apriori para compras com 11 ou mais produtos

```
# Importação das bibliotecas necessárias

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from apyori import aprior


# Ficheiro já sem "" e sem [], operação efetuada via Notepad++ através da funcionalidade replace.


#Identificação do range das colunas (valor máximo) para não dar erro ao processar ficheiro pois variam de compra para compra.
col_names=[i for i in range(0,79)]

#Carregamento do ficheiro CSV para formato dataframe (Pandas)
Dados = pd.read_csv('E:\\3MFinal.csv', header=None, names=col_names)

# Colocação de todos os dados como float para melhor performance
Dados1 = Dados.astype(float)

# Conversão dos dados do dataframe para uma lista de listas eliminando a grande quantidade de células com valores NaN do dataframe para melhor performance e ter os dados necessários para o algoritmo Apriori.

records = []

#Percorrer todas as linhas do dataframe (correspondem às compras/cestos)

for i in range(0, 3214873):

    rec=[]  #criar nova lista para adicionar  produtos da compra atual
    #Percorrer todas as colunas do dataframe
    for j in range(0, 144):
        if not np.isnan(Dados1.values[i,j]) :
            rec.append(Dados1.values[i,j])
    records.append(rec)
records

# Eliminação das compras com menos de 11 produtos

result = []

for record in records:
    if len(record) >= 11:
        result.append(record)

# Definição dos parâmetros mínimos de suporte, confiança e lift  e visualização das regras geradas

association_rules = apriori(result, min_support=0.05, min_confidence=0.1, min_lift=1.1)
association_results = list(association_rules)
print (association_results )
```

8.11 Scripts para guardar regras de associação e outputs

```
import json

# Código adicionado após a execução do algoritmo Apriori

result=[]
for item in association_results:
    rec=[]
# Retira a lista de items que definem o antecedente e consequente da regra e que se encontra na estrutura do tipo RelationRecord, nomeadamente do elemento OrderedStatistic (corresponde ao item[2]). Nesse elemento estão os itens antecedentes e consequentes e ainda a confiança e o lift
#Ex:
#RelationRecord(items=frozenset({13176, 21137, 5876}), support=0.013435833506475517, ordered_statistics=[OrderedStatistic(items_base=frozenset({21137, 5876}), items_add=frozenset({13176}), confidence=0.5606936416184971, lift=2.3263952561291608)])
    antecedente = [x for x in item[2][0][0]]
    consequente = [x for x in item[2][0][1]]
    rec.append(antecedente)
    rec.append(consequente)
    #O valor de suporte é o item com índice 1 no RelationRecord
    rec.append(item[1])
    rec.append(item[2][0][2])
    rec.append(item[2][0][3])
    result.append(rec)
with open("outputFile.txt", "w") as f:
#Será guardada em ficheiro uma lista de regras com o formato:
#[[antecedente - lista de itens], [consequente - lista de itens],suporte, confiança, lift]
#Ex: [[4957], [33754], 0.01017764740779803, 0.502195478939665, 18.36903318445856]
    json.dump(result, f)
```

Figura 44 – Script para guardar regras em ficheiro

```
import json
import pandas as pd
import sys

file_name = sys.argv[1] #abre ficheiro cujo nome é um parâmetro do script
with open(file_name) as f:
    regras = json.load(f)
#Lê a tabela de produtos para consultar os nomes
products = pd.read_csv("products.csv",index_col=0)
#Lê de novo a lista de regras do ficheiro
for item in regras:
    #em cada item das listas de IDs, substitui pelo nome do produto
    antecedente = {products.loc[x,'product_name'] for x in item[0]}
    consequente = {products.loc[x,'product_name'] for x in item[1]}
    suporte= item[2]
    confianca= item[3]
    lift = item[4]
# ”escreve” a regra em formato amigável
    print("Regra: " + str(antecedente) + " -> " + str(consequente) +"\tSuporte: " +
str("{:5.4f}".format(suporte))+"\tConfidence: " + str("{:5.4f}".format(confianca))+"\tLift: " +
str("{:5.4f}".format(lift)))
```

Figura 45 – Script para converter regras do ficheiro para formato legível

Regras resultantes do teste ao *dataset Prior*, com confiança de 50% e suporte mínimo de 0,5% (20 regras)

Regra: {'Total 2% Lowfat Greek Strained Yogurt With Blueberry'} -> {'Total 2% with Strawberry Lowfat Greek Strained Yogurt'} Suporte: 0.0102 Confidence: 0.5022 Lift: 18.3690

Regra: {'Nonfat Icelandic Style Strawberry Yogurt'} -> {'Icelandic Style Skyr Blueberry Non-fat Yogurt'} Suporte: 0.0053 Confidence: 0.5097 Lift: 28.1529

Regra: {'Bartlett Pears'} -> {'Banana'} Suporte: 0.0147 Confidence: 0.5042 Lift: 1.7402

Regra: {'Non Fat Raspberry Yogurt'} -> {'Icelandic Style Skyr Blueberry Non-fat Yogurt'} Suporte: 0.0079 Confidence: 0.5041 Lift: 27.8438

Regra: {'Organic Yellow Squash'} -> {'Organic Zucchini'} Suporte: 0.0059 Confidence: 0.5100 Lift: 5.3392

Regra: {'Organic Strawberries', 'Organic Navel Orange'} -> {'Bag of Organic Bananas'} Suporte: 0.0052 Confidence: 0.5169 Lift: 2.1900

Regra: {'Organic Hass Avocado', 'Organic Navel Orange'} -> {'Bag of Organic Bananas'} Suporte: 0.0062 Confidence: 0.5519 Lift: 2.3382

Regra: {'Apple Honeycrisp Organic', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0108 Confidence: 0.5233 Lift: 2.2169

Regra: {'Broccoli Crown', 'Organic Avocado'} -> {'Banana'} Suporte: 0.0054 Confidence: 0.5372 Lift: 1.8543

Regra: {'Organic Large Extra Fancy Fuji Apple', 'Organic Raspberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0063 Confidence: 0.5273 Lift: 2.2341

Regra: {'Organic Large Extra Fancy Fuji Apple', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0093 Confidence: 0.5083 Lift: 2.1533

Regra: {"Organic D'Anjou Pears", 'Organic Raspberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0057 Confidence: 0.5129 Lift: 2.1731

Regra: {"Organic D'Anjou Pears", 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0077 Confidence: 0.5268 Lift: 2.2319

Regra: {'Organic Hass Avocado', 'Organic Raspberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0183 Confidence: 0.5214 Lift: 2.2089

Regra: {'Strawberries', 'Honeycrisp Apple'} -> {'Banana'} Suporte: 0.0055 Confidence: 0.5805 Lift: 2.0037

Regra: {'Cucumber Kirby', 'Strawberries'} -> {'Banana'} Suporte: 0.0060 Confidence: 0.5136 Lift: 1.7729

Regra: {'Organic Fuji Apple', 'Organic Avocado'} -> {'Banana'} Suporte: 0.0056 Confidence: 0.5139 Lift: 1.7739

Regra: {'Honeycrisp Apple', 'Organic Avocado'} -> {'Banana'} Suporte: 0.0066 Confidence: 0.5121 Lift: 1.7677

Regra: {'Cucumber Kirby', 'Organic Avocado'} -> {'Banana'} Suporte: 0.0094 Confidence: 0.5259 Lift: 1.8151

Regra: {'Organic Strawberries', 'Organic Hass Avocado', 'Organic Raspberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0073 Confidence: 0.5321 Lift: 2.2542

Regras resultantes do teste ao *dataset Train*, com confiança de 50% e suporte mínimo de 0,8% (25 regras)

Regra: {'Organic Large Green Asparagus'} -> {'Bag of Organic Bananas'} Suporte: 0.0091 Confidence: 0.5432 Lift: 2.2539

Regra: {'Organic Lemon', 'Organic Strawberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0134 Confidence: 0.5607 Lift: 2.3264

Regra: {'Organic Lemon', 'Organic Baby Spinach'} -> {'Bag of Organic Bananas'} Suporte: 0.0093 Confidence: 0.5253 Lift: 2.1795

Regra: {'Organic Lemon', 'Organic Raspberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0109 Confidence: 0.6434 Lift: 2.6697

Regra: {'Organic Lemon', 'Organic Cucumber'} -> {'Bag of Organic Bananas'} Suporte: 0.0081 Confidence: 0.5707 Lift: 2.3680

Regra: {'Organic Lemon', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0143 Confidence: 0.5881 Lift: 2.4400

Regra: {'Apple Honeycrisp Organic', 'Organic Strawberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0082 Confidence: 0.5152 Lift: 2.1374

Regra: {'Organic Hass Avocado', 'Organic Red Bell Pepper'} -> {'Bag of Organic Bananas'} Suporte: 0.0082 Confidence: 0.5613 Lift: 2.3290

Regra: {'Organic Large Extra Fancy Fuji Apple', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0095 Confidence: 0.6171 Lift: 2.5605

Regra: {'Organic Cucumber', 'Organic Strawberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0188 Confidence: 0.5094 Lift: 2.1136

Regra: {'Organic Kiwi', 'Organic Strawberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0103 Confidence: 0.5382 Lift: 2.2330

Regra: {'Organic Hass Avocado', 'Organic Strawberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0261 Confidence: 0.5440 Lift: 2.2572

Regra: {'Organic Raspberries', 'Organic Baby Spinach'} -> {'Bag of Organic Bananas'} Suporte: 0.0131 Confidence: 0.5053 Lift: 2.0968

Regra: {'Organic Baby Spinach', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0182 Confidence: 0.5208 Lift: 2.1608

Regra: {"Organic D'Anjou Pears", 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0082 Confidence: 0.5777 Lift: 2.3968

Regra: {'Organic Raspberries', 'Organic Yellow Onion'} -> {'Bag of Organic Bananas'} Suporte: 0.0083 Confidence: 0.5195 Lift: 2.1554

Regra: {'Organic Yellow Onion', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0136 Confidence: 0.5552 Lift: 2.3038

Regra: {'Organic Cucumber', 'Organic Raspberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0121 Confidence: 0.5469 Lift: 2.2691

Regra: {'Organic Raspberries', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0208 Confidence: 0.5925 Lift: 2.4584

Regra: {'Organic Cucumber', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0152 Confidence: 0.5530 Lift: 2.2946

Regra: {'Organic Grape Tomatoes', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0088 Confidence: 0.5248 Lift: 2.1774

Regra: {'Organic Tomato Cluster', 'Organic Hass Avocado'} -> {'Bag of Organic Bananas'} Suporte: 0.0080 Confidence: 0.5800 Lift: 2.4065

Regra: {'Organic Hass Avocado', 'Organic Zucchini'} -> {'Bag of Organic Bananas'} Suporte: 0.0109 Confidence: 0.5097 Lift: 2.1147

Regra: {'Strawberries', 'Organic Avocado'} -> {'Banana'} Suporte: 0.0093 Confidence: 0.5172 Lift: 1.8296

Regra: {'Organic Raspberries', 'Organic Hass Avocado', 'Organic Strawberries'} -> {'Bag of Organic Bananas'} Suporte: 0.0096 Confidence: 0.6150 Lift: 2.5519