Este artigo está originalmente publicado nos Proceedings do evento onde foi apresentado:

Eds. Fernandes, Paula O.; Nunes, Alcina; Lopes, Isabel Maria; Pereira, João Paulo; Teixeira, João Paulo; Leite, Joaquim; Alves, Jorge; Ribeiro, Nuno A.; Moutinho, Nuno; Raposo, Mário Lino Barata; Ferreira, João José de Matos; Alves, Helena Maria Batista; Leal Millán, Antonio; Barroso Castro, Carmen; Navarro García, Antonio (2020). XXX Jornadas Luso-Espanholas de Gestão Científica. Bragança: Instituto Politécnico. ISBN 978-972-745-279-8

http://hdl.handle.net/10198/22373

EXTRAÇÃO DE CONHECIMENTO ATRAVÉS DE *DATA MINING*: OBTENÇÃO DE REGRAS DE ASSOCIAÇÃO NUM *DATASET* PÚBLICO

Joana Raquel Carias de Oliveira, joana.carias.oliveira@estudantes.esce.ips.pt, Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal
Vítor Barbosa, vitor.barbosa@esce.ips.pt, Escola Superior de Ciências Empresariais do Instituto Politécnico de Setúbal

RESUMO: Este artigo apresenta um trabalho de investigação que consistiu na Análise de Carrinhos de Compras (*Market Basket Analysis*) através da aplicação de técnicas de *data mining* a um conjunto de dados de dimensão significativa. Pretende-se encontrar conjuntos de itens habitualmente comprados em conjunto e daí gerar Regras de Associação, as quais podem ser valiosas para campanhas promocionais ou sistemas de recomendação. Foi adotada a metodologia de investigação CRISP, especialmente vocacionada para *data mining*, e são descritas as diversas fases da mesma, com foco na análise exploratória dos dados (para os conhecer), a preparação dos dados para aplicação e configuração do algoritmo Apriori e o estudo do comportamento do modelo obtido, aumento e diminuição do número de Regras de Associação geradas, de acordo com a variação dos parâmetros e dos dados utilizados.

PALAVRAS-CHAVE: *Data Mining*, Regras de Associação, Algoritmo Apriori.

ABSTRACT: This paper presents a research that consisted of a Market Basket Analysis through the application of data mining techniques to a large dataset. The aim is to find sets of items commonly purchased together and then generate Association Rules, which can be valuable for promotional campaigns or recommendation systems. The CRISP research methodology was adopted, which is especially designed for data mining research. Its various phases are described, focusing on the exploratory data analysis (to understand better the data), the preparation of data for application and configuration of the Apriori algorithm and the assessment of the model behaviour, measured by the increase and decrease of the number of generated Association Rules, according to the variation of the parameters and the selected data.

KEYWORDS: Data Mining, Association Rules, Apriori Algorithm.

1. INTRODUÇÃO

O grande volume de dados atualmente gerado no dia-a-dia das organizações acarreta grandes oportunidades e desafios às mesmas para utilizar esses dados a favor do seu negócio, isto é, na criação de valor (Bose, 2009).

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 e 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 e 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.

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. 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 deve-se 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.

Na próxima secção deste artigo são apresentados alguns conceitos teóricos para enquadrar o tema e técnicas abordadas na investigação, a terceira secção apresenta a metodologia utilizada, a quarta secção integra as primeiras fases da metodologia utilizada e consiste na apresentação/exploração dos dados utilizados e na preparação dos mesmos para a modelação. A secção cinco descreve a aplicação do algoritmo Apriori aos *datasets* estudados e o ajuste aos parâmetros para obtenção das primeiras Regras de Associação e ainda uma avaliação do comportamento do modelo com a variação dos valores dos parâmetros e da segmentação dos dados considerados como *intput*. O artigo termina com a secção onde são apresentadas as conclusões.

2. ENQUADRAMENTO TEÓRICO

2.1 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 e 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 considerável 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*. A diversidade de técnicas existentes é também considerável (Gama, Carvalho, Faceli, Lorena, & Oliveira, 2017), devendo o cientista de dados ser capaz de adequar a melhor técnica de acordo com o propósito e os dados disponíveis.

2.2 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 em *data mining*, 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 e 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 e 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.

2.3 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, e 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 um cesto completo de produtos, geralmente produtos de diferentes categorias. O uso de informações sobre os cestos de compras permitem que se saiba não só quais os produtos e categorias de produtos que tendem a ser comprados juntos, mas também determinam quais os 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 e 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 1 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 1 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 1: Transações registadas ao longo do dia pelo agricultor

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

Fonte: Adaptado de (Larose & Larose, 2014)

Segundo Larose e 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 e 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 e 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 e 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 (Gama, Carvalho, Faceli, Lorena, & Oliveira, 2017), um dos algoritmos mais utilizado e popular é o algoritmo Apriori.

Segundo Singh, Garg e Mishra (2018), o algoritmo Apriori foi proposto por Agrawal e Srikant (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*.

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.

3. METODOLOGIA

Segundo Azevedo & Santos (2008), 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*).

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 *data mining* 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).

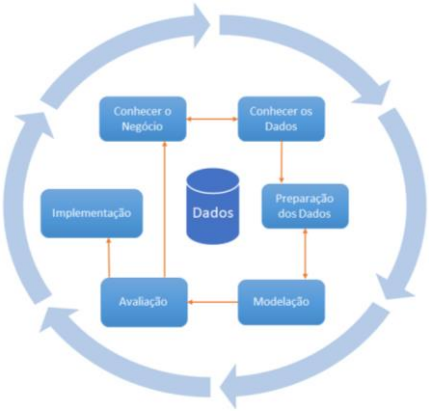

Figura 1: Metodologia CRISP
Fonte: Adaptado de Provost e Fawcett, (2015)

1. **Conhecer o Negócio:** foco no entendimento do problema a ser abordado dado que raramente são pré-definidos os objetivos da *data mining* 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 dos 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 *data mining*, 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 *data mining* 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.

Neste trabalho de investigação foi adotada a metodologia CRISP, particularmente as fases 2 a 5, uma vez que a investigação utiliza um *dataset* público com apenas alguma informação sobre o negócio, não se considerando, portanto, essa fase inicial como uma fase independente. Tratando-se de um trabalho académico, não foi também incluída a última fase, da implementação.

4. CONHECER E PREPARAR OS DADOS

Tendo como orientação 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 2 o modelo relacional realizado com recurso ao *software Power BI*.

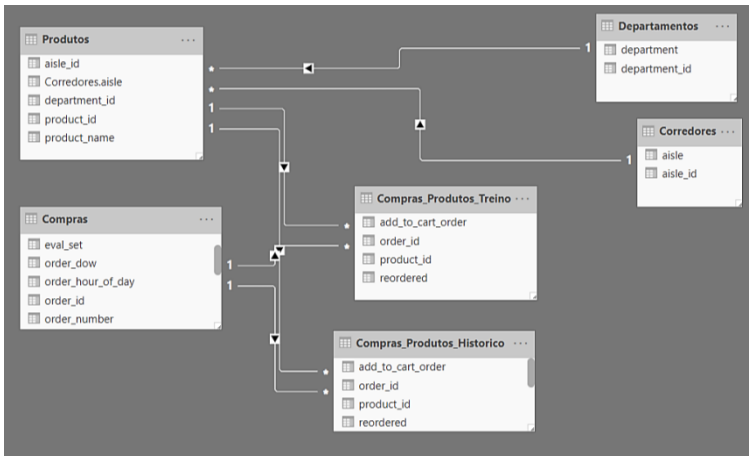

Figura 2: Modelo Relacional do Conjunto de Dados a analisar

Este conjunto de dados é composto por 3 subconjuntos: o subconjunto *prior* que contém 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 ANÁLISE EXPLORATÓRIA

Recorrendo ao *software Power BI* (Microsoft, 2019), começou-se por analisar a totalidade do histórico de compras efetuadas (conjunto *prior*). Entende-se por uma compra uma encomenda efetuada com sucesso no *site* da *Instacart* contendo um ou mais produtos. Criou-se uma medida para contar o número distinto de valores da coluna *order_id* presente na tabela Compras e confirmou-se que foram registadas 3 214 874 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. A Figura 3 mostra o gráfico com esta informação.

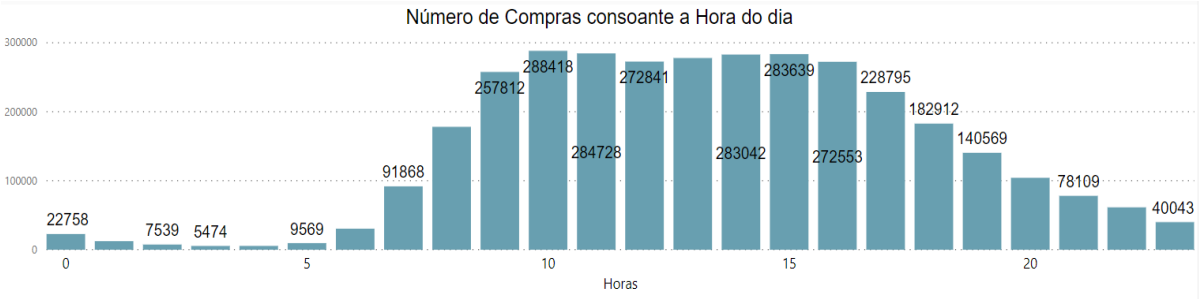

Figura 3: Número de Compras consoante a Hora do Dia

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 dividas 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 4.

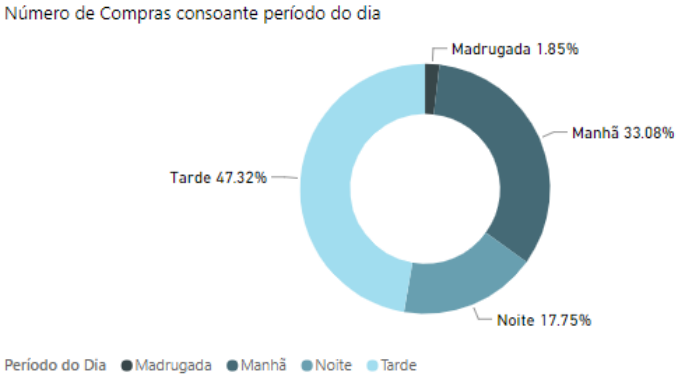

Figura 4: Número de Compras consoante período do dia

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 5, assumindo-se que se pode tratar do fim de semana.

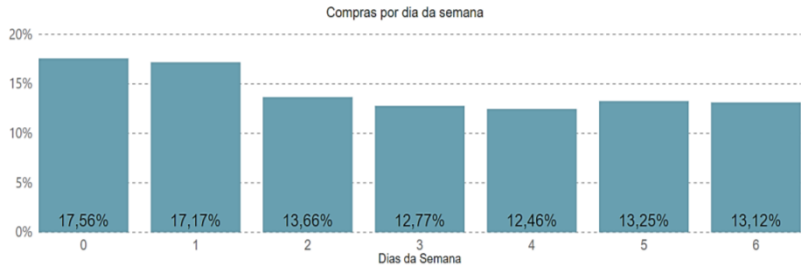

Figura 5: Compras por dia da semana

De seguida analisou-se a relação entre o número de utilizadores e as compras realizadas. As 3421083 compras foram realizadas por 206209 utilizadores. 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 6.

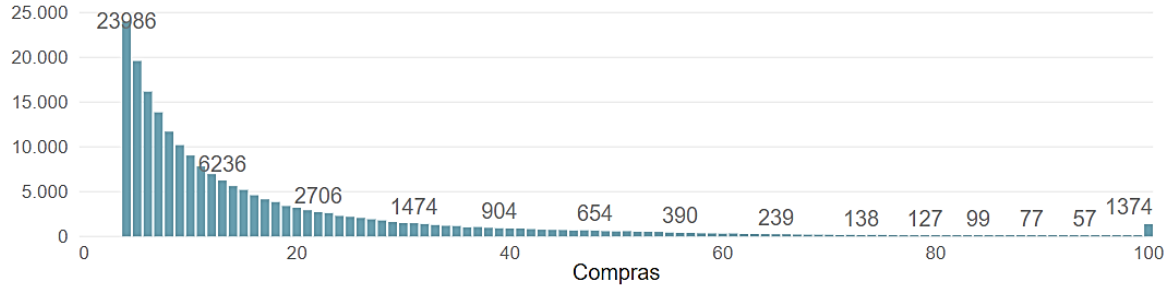

Figura 6: Total de Utilizadores por número de Compras

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. Possivelmente os dados dos clientes com mais do que 100 compras foram truncados em 100, justificando o maior valor. Em média, neste conjunto de dados, cada utilizador realizou 17 compras e como mediana temos 10 compras por utilizador. Optou-se por calcular não só a média como também a mediana para esta situação devido à existência de *outliers*.

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 (uma semana) ou de 30 dias (um mês), o que mostra alguma rotina nos hábitos de compra (compras semanais ou mensais).

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 7 temos o gráfico referente ao Top 10 de número de produtos mais adquiridos por compra.

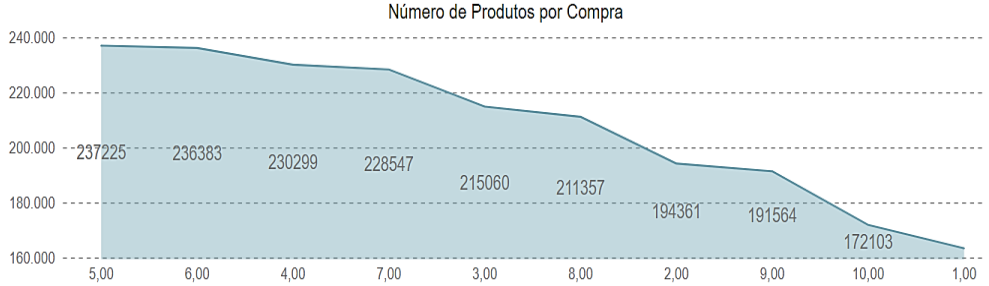

Figura 7: Top 10 Número de Produtos por Compra

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.

4.3 PREPARAÇÃO DOS DADOS

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; Vasconcelos & Barão, 2017).

Nesta investigação foram utilizados dois *datasets* de entre os dados disponibilizados pela Instacart, 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 8.

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

Figura 8: 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 8, deve ser representado apenas numa linha como o exemplo da Figura 9.

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

Figura 9: Representação das compras numa só linha

5. MODELAÇÃO E AVALIAÇÃO

Nesta secção são descritas as etapas de modelação e avaliação do modelo de *data mining* adotado para o problema em estudo. Como o que se pretende é extrair regras de associação dos *datasets* com as compras da Instacart, a modelação acaba por ser uma etapa simples, através da aplicação de um algoritmo pré-existente (de uma biblioteca Pyhton) e o ajuste aos seus parâmetros de configuração de modo a obter as primeiras regras de associação. A fase de avaliação apresenta um estudo sobre o comportamento do modelo com variações nos parâmetros de entrada e nos dados selecionados para analisar.

5.1 MODELAÇÃO

Para completar o ambiente de modelação foi necessário instalar uma biblioteca adicional do Python 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*.

Estando os dados no formato necessário passou-se à 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.

Depois de criada a lista de transações no formato adequado ao algoritmo *Apriori* adotado, a mesma pode ser utiliza pelo mesmo, sendo apenas necessário definir os parâmetros adicionais de configuração do algoritmo, nomeadamente o valor mínimo de suporte, confiança e *lift* (opcional).

Foram feitos testes para avaliar o comportamento do algoritmo com diferentes valores dos parâmetros verificou-se que apenas com um suporte mínimo de 2% foi possível identificar a primeira regra de associação no *dataset train*. A Tabela 2 mostra as regras encontradas com suporte mínimo de 1% (ou mais).

Tabela 2: 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, foi analisado se para o mesmo conjunto de dados, mas considerando apenas compras com mais de 10 produtos eram geradas regras de associação A, B→ C.

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 do tipo 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, existindo uma maior probabilidade de geração de regras A→B pois numa amostra menor, a contagem de compras necessária para o mesmo suporte é inferior. 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, sendo, neste caso 3 produtos idênticos entre compras corresponde a aproximadamente 25%, ou menos, do carrinho de compras se todos eles tiverem 11 ou mais produtos, quando anteriormente podia atingir 75% (para carrinhos de 4 produtos).

As 2 regras de associação do tipo A, B→ C encontram-se na Tabela 3.

Tabela 3: 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

Foram testados os mesmos parâmetros utilizando o *dataset prior*, tendo sido obtidas 11 regras de associação.

Na Tabela 4 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. 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 e que resulta da alteração da dimensão da amostra. Das 11 regras, nove são as mesmas tanto para o conjunto *train* como para o conjunto histórico. Verifica-se a coerência nos dados, mostrando que o *dataset train* é representativo.

Tabela 4: Comparação das Regras de Associação obtidas nos Conjuntos Train e Prior

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			

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. Os testes preliminares mostram também que o modelo/algoritmo adotado em combinação com o problema/*dataset* utilizado, é sensível à granularidade com qual se pretende analisar os dados, sendo que quando são utilizadas *datasets* com muitos registos apenas são obtidas regras com valores bastante baixos de suporte, e, consequentemente, de confiança. Estas observações levaram-nos a querer investigar o comportamento do modelo variando os parâmetros e os dados fornecidos como *input*, segmentando as compras por dimensão (quantidade mínima de produtos no carrinho). A avaliação do modelo perante essas alterações é apresentada na subsecção seguinte.

5.2 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 adotado. 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 5.

Tabela 5: 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, 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 (carecendo de validação posterior).

Antes de apresentar os resultados obtidos nos diversos conjuntos de testes, apresentam-se na Tabela 6 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 6: 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 7 apresentam-se os resultados do primeiro conjunto de testes onde se utilizou uma confiança mínima de 20%. 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 7, 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%.

Tabela 7: 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

O mesmo exercício foi efetuado, aumentando ligeiramente a confiança para 25%. Os resultados obtidos encontram-se na Tabela 8.

Tabela 8: 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 8 com os anteriores (da Tabela 7), 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 10 mostra, para o *dataset train* e *prior*, 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.

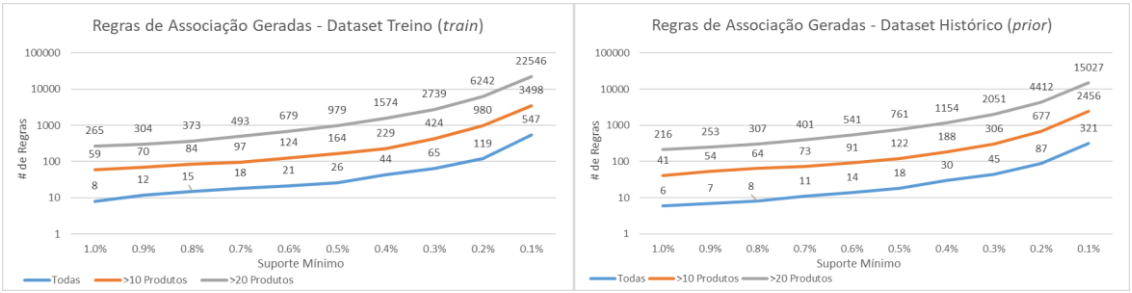

Figura 10: Gráficos com contagem de Regras de Associação Geradas (confiança=25%)

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 9.

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 9: 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

6. CONCLUSÃO

Este artigo apresenta a investigação da aplicação de uma técnica de *market basket analysis* (MBA) a um *dataset* público com mais de três milhões de registos de compras realizadas na aplicação Instacart. Foi adotada a metodologia CRISP para guiar a investigação e são detalhadas neste artigo as diversas etapas previstas nessa metodologia, com especial foco na análise exploratória dos dados, enquadrado na fase de "conhecer os dados", e na avaliação do comportamento do modelo adotado, que consistiu na aplicação do algoritmo Apriori aos registos de compras de modo a extrair Regras de Associação entre os produtos adquiridos pelos clientes.

Nesta investigação não se pretendeu analisar o conteúdo das Regras de Associação obtidas, mas sim confirmar a aplicabilidade da técnica de data mining selecionada ao conjunto de dados em estudo e aferir o comportamento do modelo com a variação dos parâmetros e dos dados utilizados.

O trabalho de limpeza e transformação dos dados originais permitiu a aplicação de uma implementação do algoritmo Apriori com sucesso e a fase de avaliação demonstra que segmentando os dados é possível encontrar Regras de Associação adicionais, eventualmente valiosas em segmentos de produtos e/ou perfis de consumidores.

Os resultados desta investigação abrem novas oportunidades de investigação, explorando novas segmentações dos dados utilizando outras dimensões (classes de produtos, tipos de cliente, etc.) ou aplicar as mesmas técnicas a outros conjuntos de dados, idealmente oriundos de supermercados nacionais. Os dados obtidos das Regras de Associação, em cada contexto de aplicação, podem ser utilizados para definição de campanhas promocionais (packs de produtos) que permitam escoar produtos de menor saída (associando, por exemplo a um par de produtos habitualmente comprado em conjunto), mas em contextos mais dinâmicos, como compras online ou com dispositivos móveis (em loja), podem ser usados para recomendar produtos ao cliente de acordo com os produtos previamente selecionados e/ou de acordo com o perfil de cliente (Bodapati, 2008). A exploração destes mecanismos de recomendação são também oportunidades de investigação futura.

REFERÊNCIAS

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., & Santos, M. F. (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.

Bodapati, A. V. (2008). Recommendation Systems with Purchase Data. *Journal of Marketing Research*, 45(1), 77-93. doi:10.1509/jmkr.45.1.077

Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management & Data Systems*, 109(2), 155-172. doi:10.1108/02635570910930073

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

Gama, J., Carvalho, A. C., Faceli, K., Lorena, A. C., & Oliveira, M. (2017). *Extração de conhecimento de dados: data mining (3ª ed.)*. Lisboa: Edições Sílabo.

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).

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

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

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).

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*, 11-20.

Larose, D. T., & Larose, C. D. (2014). *Discovering Knowledge in Data an Introduction to Data Mining (Second Edition*). John Wiley & Sons, Inc.

Microsoft. (2019). Acesso em 3 de 11 de 2019, disponível em Power BI: https://powerbi.microsoft.com/pt-pt/

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

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).

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

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).

Vasconcelos, J. B., & Barão, A. (2017). *Ciência dos Dados nas Organizações*. FCA.

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).