UNIVERSIDADE DO VALE DO RIO DOS SINOS - UNISINOS
UNIDADE ACADÊMICA DE GRADUAÇÃO
CURSO DE ENGENHARIA DE PRODUÇÃO

ANA PAULA KERCKHOFF

MARKET BASKET ANALYSIS UTILIZANDO DATA MINING

São Leopoldo
2018

ANA PAULA KERCKHOFF

MARKET BASKET ANALYSIS UTILIZANDO DATA MINING

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Bacharel em Engenharia de Produção, pelo Curso de Engenharia de Produção da Universidade do Vale do Rio dos Sinos - UNISINOS

Orientador: Prof. Ms. Marcos Leandro Hoffmann Souza

São Leopoldo

2018

AGRADECIMENTOS

Agradeço aos meus pais, Gilmar e Lucineia, por sempre me incentivarem a continuar estudando, prestando todo apoio necessário. Obrigada por serem tão presentes na minha vida e demonstrarem orgulho nos momentos de sucesso e fracasso, indicando o caminho a seguir.

Ao meu namorado, Jean Lucas, obrigada por todo suporte durante essa caminhada, tanto físico como emocional. Sempre paciente nos momentos difíceis, obrigada por todo carinho.

Ao meu orientador Marcos Hoffmann, obrigada por todo auxílio prestado durante este trabalho. Obrigada pela dedicação em corrigir e conduzir o caminho, sempre incentivando em buscar algo melhor. Sinto-me feliz por ter dividido contigo este processo de evolução.

Aos meus amigos Daniel, Emely, Julia, Fernanda e Pamela, obrigada por todos esses anos de amizade e por todas as palavras de motivação, compreendendo também minha ausência em alguns momentos.

À amizade construída dentro da Unisinos, obrigada Bruna, Gabriela, Fernanda e Priscila por compartilharem esses anos comigo, dividindo também as alegrias, angústias e comemorações. Vocês foram essenciais durante o último ano.

A todos que contribuíram para a minha formação, muito obrigada!

“Qualquer conquista começa com a decisão de tentar. ”

Gail Devers

RESUMO

Uma imensa quantidade de dados vem sendo coletada e armazenada pelas empresas ano após ano. No entanto, a maioria deles não está sendo analisada de forma a potencializar o negócio. O *Data Mining* é uma ferramenta que tem por objetivo analisar bases de dados e extrair informações úteis para gerar *insights* e auxiliar os gestores na tomada de decisões. Relacionada a este tema, esta monografia teve como objetivo encontrar regras de associação entre itens frequentes, analisando uma base de dados de um supermercado, a partir da Análise da Cesta de Compras. Para isso, estruturou-se uma estratégia de extração, tratamento e carregamento de dados, e foi desenvolvido um modelo computacional para ser utilizado no *software* R para aplicação do algoritmo *apriori*. Seguindo o modelo, foram encontrados os itens mais frequentes da base de dados, bem como as regras de associação e, a partir dos resultados obtidos, foram propostos quatro combos de venda para o supermercado. O estudo comprovou que a ferramenta pode se tornar uma grande aliada para tomar decisões mais assertivas para o negócio. Este trabalho pode ser utilizado como base para outras pesquisas e, além disso, disponibiliza um modelo que pode ser replicado em diferentes segmentos.

Palavras-chave: Data Mining. Market Basket Analysis. Regras de Associação. Algoritmo Apriori. Supermercado.

ABSTRACT

A huge amount of data has been collected and stored by companies year after year. However, most of them are not being analyzed in order to boost the business. Data Mining is a tool that is a tool that aims to analyze databases and extract useful information to generate insights and assist managers in decision making. Related to this topic, this monograph had as objective to find rules of association between frequent items, analyzing a database of a supermarket, from the Market Basket Analysis. For this, a strategy of extraction, treatment and loading of data was structured, and a computational model was developed to be used in software R for apriori algorithm application. Following the model, the most frequent item from the database were found, as well as the association rules and, from the results, four sales combos were proposed for the supermarket. The study has proven that the tool can become a great ally to make more assertive decisions for the business. This work can be used as a basis for further research and, in addition, provides a model that can be replicated in different segments.

Palavras-chave: Data Mining. Market Basket Analysis. Association Rules. Apriori Algorithm. Supermarket.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

DM	Data Mining
DW	Data Warehouse
ETL	Extract, transform, load
KDD	Knowledge Discovery in Databases
MBA	Market Basket Analysis
RSL	Revisão Sistemática da Literatura
TI	Tecnologia da Informação
TID	Transaction ID

SUMÁRIO

1 INTRODUÇÃO

Devido a globalização e o progresso da tecnologia nas últimas décadas, as organizações vêm buscando aprimorar e atualizar seus processos e conhecimentos, a fim de manter a competitividade. Portanto, deter o controle das informações sobre o negócio torna-se crucial para o seu desenvolvimento. (SACILOTTI, 2011).

Vantagem competitiva, segundo Barney e Hesterly (2011), se dá quando uma empresa tem a capacidade de gerar maior valor econômico do que suas concorrentes. Como fator fundamental para o diferencial de cada empresa, os gestores devem identificar as vantagens e reforçá-las para alavancar os negócios. Sendo assim, tomar as decisões corretas nos momentos certos, com as informações disponíveis, é um elemento-chave dentro das organizações. (SEMAAN; GRAÇA; DIAS, 2006).

Diariamente são coletadas imensas quantidades de dados, e analisá-las é uma necessidade importante. (HAN; KAMBER; PEI, 2012). Freitas (1993) afirma que as informações e o conhecimento formam um recurso estratégico para o sucesso da empresa no meio competitivo. O autor afirma ainda que a empresa precisa aproveitar as oportunidades disponíveis, como tomar uma ação com base na informação e conhecimento.

Com o avanço da tecnologia, é possível obter informações de qualidade em um menor espaço de tempo. Almeida (1995) alega que as organizações podem se tornar mais competitivas tirando proveito das novas tecnologias. Dispondo de determinadas informações, é possível que a empresa eleve o valor agregado de seu produto ou reduza seus custos em relação àquelas que não possuem tais informações. (ALMEIDA, 1995).

O crescente progresso da tecnologia tornou possível a coleta de dados, bem como aumentou a capacidade de armazená-los. Tais avanços impulsionaram a indústria da informação, possibilitando o gerenciamento e análise de dados. (HAN; KAMBER; PEI, 2012). Porém essa imensa capacidade de coletar e armazenar dados superou a competência humana em analisá-los. Por este motivo, se faz necessária a utilização de técnicas e ferramentas que transformem estes dados em conhecimento e informações que possam ser utilizadas para suporte a tomada de decisão. (CASTRO; FERRARI, 2016).

Ao efetuar uma compra por meio de um dispositovo móvel, por exemplo, são armazenados em bancos de dados informações como número de telefone, número

do cartão de crédito, endereço de entrega, entre outros, que permanecem disponíveis e podem ser de grande valia se analisados. (CARVALHO, 2001). Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de descoberta de conhecimento em bases de dados, KDD (*Knowledge Discovery in Databases*), refere-se ao processo formado de várias etapas, para identificação de padrões válidos, úteis e compreensíveis, em grandes bases de dados. Entre as etapas do KDD, existe a etapa de mineração de dados (*Data Mining*), que consiste em aplicar algoritmos para extrair os padrões aparetemente ocultos e analisá-los. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Dentre as tarefas de mineração está a regra de associação, que tem por objetivo encontrar padrões de itens frequentes em bases de dados e é comumente aplicada em supermercados. (GONÇALVES, 2005). As regras de associação formam a base para a chamada análise da cesta de compras (*Market Basket Analysis*), método da mineração de dados com a finalidade de encontrar padrões de compra por meio da extração de associações encontradas nas transações de supermercados. (MAINALI, 2016).

Com as definições apresentadas, este trabalho tem o intuito de aplicar o *Data Mining* em um supermercado, utilizando a regra de associação, a fim de descobrir padrões de compra.

1.1 OBJETO E PROBLEMA DE PESQUISA

De acordo com Porter (1996), uma empresa conseguirá superar a concorrência apenas quando obter e mantiver um diferencial sobre seus concorrentes. No meio dos negócios, a informação vem se mostrando um diferencial competitivo, com o intuito de contribuir para que as organizações alcancem seus objetivos estratégicos. (DIAS, 2002).

Porter e Millar (1985) afirmam que a revolução da informação está assolando a economia e que as empresas não escaparão de seus efeitos. O baixo custo de obter, processar e transmitir informação está mudando a maneira de fazer negócios. Atualmente a Tecnologia da Informação (TI) está se alastrando por toda a cadeia de valor e vem realizando funções de otimização e controle. Além disso, a TI está transformando os processos físicos das atividades, pois apresenta maior agilidade, flexibilidade e rigorosidade. (PORTER; MILLAR, 1985). Da Silva Lima *et al* (2017)

afirmam que os gestores devem utilizar as ferramentas de TI para suportar as decisões e alcançar melhores desempenhos no ambiente organizacional.

A Tecnologia da Informação tem como um dos objetivos agrupar todas as informações presentes na organização e torná-las compreensíveis pelos gestores para auxiliar no momento de tomada de decisão. (FERREIRA; SILVEIRA, 2007). A utilização da TI se mostrou indispensável no setor supermercadista devido a necessidade de maior eficiência no desempenho operacional, exigências fiscais, legais e tributárias, redução de custos e integração com a cadeia logística. No atual cenário competitivo, a TI deve ser utilizada não apenas como ferramenta operacional, mas sim como ferramenta estratégica, uma vez que possibilita a extração de informações de qualidade e significativas para o negócio. (CARVALHO *et al*., 2006; GOMES, 2013).

A redução dos custos de adquirir novas tecnologias proporcionou a adesão de sistemas de informação até mesmo para os pequenos varejistas. Aliar as tecnologias às estratégias de vendas é um dos desafios dos supermercadistas para os próximos anos. Nuno Fouto (diretor vogal do Instituto Brasileiro de Executivos de Varejo e Mercado de Consumo - Ivebar) afirma que é possível mapear o comportamento dos consumidores para determinar os produtos que terão ofertas e melhorar sua exposição a fim de incentivar a compra por impulso. (ITO, 2018).

As companhias vêm acumulando informações em seus bancos de dados ano após ano. Informações estas que podem ser utilizadas para melhorar os processos internos, sendo possível detectar tendências e, dessa forma, antever contratempos futuros. Porém, a maioria das organizações não consegue se beneficiar dos dados por possuir sistemas de gerenciamento de bancos de dados convencionais. O desenvolvimento da tecnologia mostrou como coletar e armazenar os dados, mas ultrapassou a capacidade humana em interpretá-los. (FIGUEIRA, 1998 apud GONÇALVES, 1999; TECNOLOGIA..., 2018).

Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), existe uma necessidade imediata de novas ferramentas que possam auxiliar na extração de informações úteis através das grandes bases de dados. O método convencional de transformar dados em conhecimento depende de análise e interpretação manual. O que se mostra um processo lento, caro e subjetivo. Com os grandes volumes de dados gerados a partir de fontes diferentes nas empresas, esse método está se tornando impraticável. A necessidade em fornecer recursos para analisar a grande quantidade de dados

coletados é também de contexto econômico. Diante disso, alguns supermercados já contam com apoio de *softwares* e serviço de *business analytics* para transformar dados em conhecimento e descobrir novas oportunidades. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; TECNOLOGIA..., 2018).

Assim como a tecnologia permitiu coletar mais dados do que pode-se compreender, é natural que se recorra às suas técnicas para auxiliar na descoberta de padrões e estruturas significativas decorrentes dos grandes volumes de dados. Para isso, o KDD surgiu como uma tentativa de abordar o problema da sobrecarga de dados. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

De acordo com o conteúdo evidenciado nesta seção, a questão que norteia esta monografia é: É possível identificar associações de compra em um supermercado e propor combos de venda com a aplicação da ferramenta *Data Mining*?

1.2 OBJETIVOS

Nessa seção será apresentado o objetivo geral e o objetivo específico deste trabalho.

1.2.1 Objetivo Geral

Avaliar as associações de compra encontradas em um supermercado, utilizando a ferramenta *Data Mining* e, a partir disto, propor combos de venda.

1.2.2 Objetivos Específicos

Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

a) Estruturar uma estratégia de extração, tratamento e carregamento de dados não estruturados, oriundos das transações de vendas do supermercado;
b) Identificar um algoritmo de *Data Mining* que possa ser empregado no estudo;
c) Identificar as associações de compra utilizando uma técnica de *Data Mining*;

d) Propor combos de vendas de produtos, a partir das associações identificadas.

Na sequência serão apresentadas as justificativas que sustentam esta monografia.

1.3 JUSTIFICATIVAS

Essa seção apresenta as duas justificativas para a realização deste trabalho, que são: acadêmica e gerencial.

1.3.1 Justificativa acadêmica

A execução deste trabalho é justificada academicamente a fim de oferecer e complementar estudos relacionados ao tema abordado. Em vista disso, foi realizada uma revisão sistemática da literatura, baseada em Morandi e Camargo (2015), onde foram definidos os termos de busca, os critérios de inclusão e exclusão, bem como as bases de dados, seguindo o protocolo indicado no APENDICE A – PROTOCOLO DE REVISÃO SISTEMÁTICA DA LITERATURA.

Cumprindo o protocolo definido, as pesquisas apresentaram um total de 131.137 trabalhos, dentre os quais 706 títulos foram lidos, com intuito de identificar alguma relação com o tema desta monografia, considerando os critérios de exclusão, e, seguindo essa abordagem, 48 *abstracts* foram analisados. Destes, 24 artigos foram incluídos para posterior leitura completa. Avaliados tais artigos, identificou-se que 8 apresentam alguma relação com o tema regras de associação ou a aplicação de algoritmos utilizados em KDD em supermercados. Os resultados das pesquisas estão apresentados na Tabela 1.

Tabela 1 – Resultados das pesquisas

(continua)

Fonte	Termos de busca	Resultados	Títulos lidos	Resumos lidos	Incluídos
EBSCOHost	Mineração de Dados	101	0	0	0
EBSCOHost	Mineração de dados AND Regras de associação AND Análise da Cesta de Compras	0	0	0	0
EBSCOHost	Mineração de dados AND Regras de associação AND Supermercado	0	0	0	0
EBSCOHost	Mineração de dados AND Regras de associação AND Algoritmo Apriori	1	1	1	0
EBSCOHost	Regras de Associação AND Análise da Cesta de Compras AND Supermercado	0	0	0	0
EBSCOHost	Regras de Associação AND Análise da Cesta de Compras AND Algoritmo Apriori	0	0	0	0
EBSCOHost	Data Mining	65291	0	0	0
EBSCOHost	Data Mining AND Association Rules AND Market Basket Analysis	50	50	9	4
EBSCOHost	Data Mining AND Association Rules AND Supermarket	1	1	0	0
EBSCOHost	Data Mining AND Association Rules AND Apriori Algorithm	206	100	4	3
EBSCOHost	Association Rules AND Market Basket Analysis AND Supermarket	7	7	1	1
EBSCOHost	Association Rules AND Market Basket Analysis AND Apriori Algorithm	6	6	0	0
Emerald	Mineração de Dados	0	0	0	0
Emerald	Mineração de dados AND Regras de associação AND Análise da Cesta de Compras	0	0	0	0
Emerald	Mineração de dados AND Regras de associação AND Supermercado	0	0	0	0
Emerald	Mineração de dados AND Regras de associação AND Algoritmo Apriori	0	0	0	0
Emerald	Regras de Associação AND Análise da Cesta de Compras AND Supermercado	0	0	0	0
Emerald	Regras de Associação AND Análise da Cesta de Compras AND Algoritmo Apriori	0	0	0	0
Emerald	Data Mining	3050	0	0	0
Emerald	Data Mining AND Association Rules AND Market Basket Analysis	22	22	6	3
Emerald	Data Mining AND Association Rules AND Supermarket	6	6	2	0
Emerald	Data Mining AND Association Rules AND Apriori Algorithm	28	28	5	1
Emerald	Association Rules AND Market Basket Analysis AND Supermarket	5	5	1	1
Emerald	Association Rules AND Market Basket Analysis AND Apriori Algorithm	8	8	0	0

(conclusão)

Fonte	Termos de busca	Resultados	Títulos lidos	Resumos lidos	Incluídos
ProQuest	Mineração de Dados	68	0	0	0
ProQuest	Mineração de dados AND Regras de associação AND Análise da Cesta de Compras	0	0	0	0
ProQuest	Mineração de dados AND Regras de associação AND Supermercado	0	0	0	0
ProQuest	Mineração de dados AND Regras de associação AND Algoritmo Apriori	1	1	0	0
ProQuest	Regras de Associação AND Análise da Cesta de Compras AND Supermercado	0	0	0	0
ProQuest	Regras de Associação AND Análise da Cesta de Compras AND Algoritmo Apriori	0	0	0	0
ProQuest	Data Mining	60777	0	0	0
ProQuest	Data Mining AND Association Rules AND Market Basket Analysis	309	100	10	8
ProQuest	Data Mining AND Association Rules AND Supermarket	209	100	3	1
ProQuest	Data Mining AND Association Rules AND Apriori Algorithm	860	100	4	2
ProQuest	Association Rules AND Market Basket Analysis AND Supermarket	71	71	1	0
ProQuest	Association Rules AND Market Basket Analysis AND Apriori Algorithm	161	100	1	0

Fonte: Elaborado pela autora.

Observa-se que, quando pesquisado pelos termos de busca em português, apenas dois artigos retornaram, e não foram incluídos por não estarem alinhados com os objetivos deste trabalho. No entanto, ao pesquisar pelo termo específico “Mineração de Dados”, as pesquisas retornam diversos artigos tratando do assunto de forma geral, porém sem enfatizar os outros termos pesquisados. Percebe-se então a escassez de trabalhos referentes a este assunto, indicando que, dentro do universo definido no protocolo de pesquisa, não foram identificados uma quantidade significante de estudos com esta abordagem.

Ao realizar a pesquisa pelos termos na língua inglesa, é possível visualizar um aumento nos resultados e, consequentemente verificar que este tema está sendo abordado por outros países. Foram identificados diversos artigos referentes à mineração de dados e com propostas e objetivos distintos, assim foi possível constatar que a abordagem possibilita seu uso para diversos fins no setor supermercadista. Os trabalhos apresentaram a descoberta de regras de associação, porém com propostas diferentes. No artigo de Papavasileiou e Tsadiras (2013), as regras encontradas foram analisadas de acordo com a variabilidade de tempo. Já em outro trabalho, proposto por Liao, Chen e Wu (2007) o objetivo foi de analisar as regras a fim de verificar a possibilidade de criação de extensões das linhas de produtos da marca própria do supermercado. Os oito trabalhos selecionados estão apresentados no Quadro 1.

No trabalho realizado por Zekić-Sušac e Has (2015), um algoritmo é implementado, com posterior execução da ferramenta redes neurais, objetivando identificar padrões de itens frequentes a fim de agregar conhecimento e suportar as estratégias de marketing. Análogo a este artigo, Yang e Hao (2010) utilizam as regras de associação encontradas, com a finalidade de prover soluções no que se refere a escolha de produtos para destacar em promoções de supermercados.

No estudo selecionado de Annie e Kumar (2012), foi executada a tarefa de análise de cesta de compras, porém não com o mesmo objetivo da presente pesquisa. No trabalho, o objetivo de analisar a cesta de compras foi de verificar tais itens para uma melhoria de layout, aproximando os itens de maior frequência. Por outro lado, a presente pesquisa tem a finalidade de analisar a cesta de compras, verificando os itens frequentes e, a partir disto, propor combos de vendas.

Quadro 1 – Trabalhos incluídos

Título/Resumo	Autores	Fonte/Origem
Evaluating time variations to identify valuable association rules in market basket analysis O artigo busca identificar quais as regras de associação mais valiosas para a cadeia de supermercados, analisando também indicadores de variabilidade de tempo. Como por exemplo, quanto o comportamento do consumidor varia para uma determinada regra de associação encontrada, analisando-a mês a mês.	Vasilios Papavasileiou, Athanasios Tsadiras	EBSCOHost Grécia
Mining customer knowledge for product line and brand extension in retailing O artigo utiliza o algoritmo Apriori a fim de analisar as regras encontradas e seu consumo para, a partir delas, propor extensões de linhas de mercadorias da marca própria do supermercado.	Shu-Hsien Liao, Chyuan-Meei Chen, Chung-Hsin Wu	EBSCOHost Taiwan
Market basket analysis insights to support category management O artigo apresenta uma abordagem a fim de estudar o comportamento dos consumidores e detectar correlações entre categorias de produtos, com base na análise de cesta de compras.	Andres Musalem, Luis Aburto, Maximo Bosch	Emerald Chile
Applied Research on Client Identification Based on Association Rules Os autores discutem detalhes da implementação de regras de associação combinado ao processo de identificação de cliente, com intuito de apoiar as decisões gerenciais.	JunFeng Tian, liXian II	ProQuest China
Association Rule Mining for Supermarket Sale Analysis Os autores apresentam a utilização do algoritmo Apriori no conjunto de dados de um supermercado, para encontrar regras de associação. São encontrados dois produtos que apresentam maior suporte.	R. R. Shelke, R. V. Dharaskar, V. M. Thakare	ProQuest índia
Data Mining as Support to Knowledge Management in Marketing O artigo propõe a integração das ferramentas redes neurais e regras de associação, com o objetivo de descobrir padrões de itens frequentes e o perfil de clientes para suportar as decisões de marketing.	Marijana Zekić-Sušac, Adela Has	ProQuest Croácia
Market Basket Analysis for a Supermarket based on Frequent Itemset Mining O artigo apresenta a utilização do algoritmo Apriori em um supermercado, o qual se beneficia do método para melhoria de layout, a fim de melhorar a relação com os clientes.	Loraine Charlet Annie M.C., Ashok Kumar D	ProQuest Índia
Product selection for promotion planning Os autores utilizam ferramentas da mineração de dados com o objetivo de beneficiar os gestores de marketing, selecionando os produtos certos para as promoções.	Yinghui Yang, Chunhui Hao	ProQuest Londres

Fonte: Elaborado pela autora.

A RSL evidenciou que há diversos trabalhos relativos ao *Data Mining*, porém ao direcionar para o tema central desta monografia, os resultados são limitados. Embora todos os artigos selecionados apresentem o conceito do DM e, ainda que todos os trabalhos tenham encontrado padrões de compra por meio da aplicação da

ferramenta, nenhum deles propôs montar combos de vendas com os itens frequentes. Sendo essa a proposta desta monografia, o presente trabalho traz benefícios para academia, uma vez que agregará referências na literatura acadêmica.

1.3.2. Justificativa gerencial

A crescente quantidade de dados coletados e armazenados faz do momento atual a era dos dados, e é necessário o uso de ferramentas com capacidade para transformar tais dados em conhecimento dentro das organizações. (HAN; KAMBER; PEI, 2012). Conforme afirmação de Freitas (1993), a informação se tornou um recurso valioso para as empresas. Tendo em vista a alta competitividade atual, deter informações relevantes acerca dos consumidores é fator determinante para se destacar.

O *Data Mining* vem sendo utilizado por empresas que mantenham seu foco direcionado para o consumidor, principalmente organizações de varejo e marketing, que buscam aprimorar o conhecimento através dos dados e tomar decisões rápidas e assertivas para o negócio. A ferramenta disponibiliza ao gestor identificar dados acerca do consumidor, suas preferências de compra, quais as suas necessidades, bem como analisar tendências de consumo. (DATA..., 2017).

A grande quantidade de informações advindas da técnica, bem como a velocidade em que as mesmas são disponibilizadas, garantem ações direcionadas e especificadas a um público alvo. Os dados estão acima da intuição de gestores e profissionais de marketing, assegurando que a análise sobre os consumidores seja baseada em conhecimento personalizado, apresentando suas preferências de consumo e necessidades. (O..., 2017).

O varejo está enfrentando um momento de desafios e possibilidades, de forma que é necessário compreender as mudanças tecnológicas e aplicá-las ao negócio para obtenção de resultados. (MORGADO, 2017). O supermercadista não deve tratar tecnologia como despesa, mas sim um investimento. (MONTEIRO, 2017). O *Data Mining* apresenta informações, geralmente desconhecidas, que podem e devem ser utilizadas para estimar vendas, personalizar ofertas, promover ações de marketing direcionadas, melhorar o relacionamento com o cliente e ainda impulsionar as vendas. (O..., 2017).

O Grupo Pão de Açúcar (GPA) aderiu às tecnologias e está obtendo benefícios da mineração de dados. Aliado ao seu programa de fidelidade, a empresa lançou o aplicativo “Meu Desconto”, que oferece promoções personalizadas para os clientes cadastrados nos dois programas. Por meio do histórico de compras dos clientes, o aplicativo direciona ofertas exclusivas de acordo com o perfil identificado. Se um cliente costuma comprar um item de determinada marca, ele receberá ofertas deste item específico, mas também poderá receber ofertas de outros itens da mesma marca. (VIRI, 2017).

A empresa se beneficia de maneira que, além de elevar as vendas e a rotatividade de clientes na loja física, aumenta a assertividade das promoções. O grupo expõe ainda que estes dados foram disponibilizados aos fornecedores, porém mantendo a sua identidade preservada. O GPA identificou uma grande oportunidade, visto que tradicionalmente as ofertas de produtos são negociadas com os fornecedores em contratos com quantidades expressivas, acarretando na redução da margem. Porém, neste modelo, o fornecedor também analisa os dados e tem o poder de decidir ofertar seus produtos para apenas cinquenta mil consumidores, por exemplo. Consequentemente, o impacto de descontos sob os custos é reduzido. (VIRI, 2017).

É evidente o número de ações possíveis com o auxílio das ferramentas advindas do desenvolvimento da tecnologia. A aplicação de *Data Mining* traz inúmeros benefícios para o setor varejista e pode ser utilizada de formas diferentes, como apresentado pelo Grupo Pão de Açúcar. O GPA afirma ainda que, analisando as promoções efetuadas com base na análise de dados, foi possível verificar que o grupo retomou clientes que haviam direcionado suas compras aos atacados. Além disso, o grupo declara continuar investindo em inovações tecnológicas para estar à frente dos concorrentes. (VIRI, 2017).

Nesse contexto, o presente trabalho se mostra relevante visto que apresentará uma abordagem que tem um potencial de elevar as vendas, além de disponibilizar informações sobre tendências de consumo até então desconhecidas pelo administrador, as quais podem ser analisadas e utilizadas para aprimorar a gestão do estabelecimento.

1.4 DELIMITAÇÕES DO TRABALHO

Neste trabalho de conclusão de curso será apresentado um modelo para encontrar padrões e regras de associação na base de dados, a qual se refere às transações do supermercado.

Os dados coletados abrangem o período de um ano, a iniciar na data de 01/07/2017 à 31/06/2018, isso porque antes disso as transações não eram armazenadas em um servidor de banco de dados.

Com o intuito de atingir os objetivos, serão propostos combos de vendas a partir da análise dos resultados. No entanto, optou-se por não implementar e analisar tais combos, sem excluir sua aplicação futura em uma possível continuação deste trabalho.

Para a construção do modelo, será o utilizado o método proposto por Law e Kelton (1991), porém neste trabalho não se fez necessário seguir todas as suas etapas, como por exemplo a aplicação do modelo com cenários diferentes.

1.5 ESTRUTURA DO TRABALHO

Este trabalho apresenta-se em seis capítulos. No capítulo de número um está a introdução, na qual o tema deste trabalho está contextualizado. Neste capítulo apresenta-se ainda o objeto, os objetivos e as justificativas que norteiam esta pesquisa.

O capítulo de número dois é dedicado ao referencial teórico, o qual busca detalhar os principais itens abordados nesta monografia. No terceiro capítulo é apresentada a metodologia, a qual está subdividida em método científico, método de pesquisa, método de trabalho, bem como apresenta a maneira como se conduziu a coleta e análise de dados.

O capítulo quatro apresenta os principais resultados obtidos na pesquisa, onde são evidenciadas as avaliações sobre o modelo. As discussões dos resultados obtidos contemplam o capítulo de número cinco. E por fim, o sexto capítulo apresenta as conclusões e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresentará o referencial teórico que suporta este trabalho. Inicialmente será apresentado o contexto no qual a ferramenta será aplicada, em seguida será abordado o conceito de *Data Mining* com uma sucinta explicação de cada uma de suas tarefas e, por fim, a seção 2.3 é dedicada a tarefa que norteia este trabalho.

2.1 VAREJO

A existência do varejo advém de épocas remotas, onde países como Atenas, Alexandria e Roma eram conhecidos como áreas comerciais. No entanto, o início do varejo no Brasil ocorreu no final do século XIX, quando também se iniciava a industrialização e surgiam os meios e vias de transporte. Ressalta-se que o desenvolvimento do varejo ocorreu após a Segunda Guerra, com o enfraquecimento do setor atacadista, o qual comandava o setor produtivo e distributivo. (LAS CASAS; BARBOZA, 2007).

Donato (2012) define varejo como o processo de compra de itens em grandes quantidades, de grandes atacadistas, para posterior venda de tais itens em pequenas quantidades ao consumidor final. Assim sendo, o varejo envolve quaisquer atividades de venda de bens e serviços a fim de atender as necessidades dos consumidores finais. O autor afirma ainda que o varejo exerce papel importante na cadeia de suprimentos, posicionando-se estrategicamente entre fornecedores e consumidores. (DONATO, 2012).

Conforme Romero (2012), o varejo é uma figura de suma importância para a sociedade, de modo que não se tornaria vantagem aos fabricantes que estes efetuassem a distribuição ao consumidor final, pois esta prática implicaria em custos elevados, refletindo negativamente no preço final. As atividades que integram o varejo colaboram no sistema de distribuição de produtos e ainda facilitam o dia a dia dos produtores e consumidores. (GIANOTTI, 2013).

O varejo tem apresentado crescente importância para a economia brasileira, uma vez que as atividades do setor têm se mostrado estáveis, ganhando destaque como algumas das maiores empresas do país. A expansão das empresas varejistas traz à tona a necessidade de adotar tecnologias e sistemas de informação e gestão

avançados, que permitam a modernização no sistema de distribuição. O setor varejista é uma das atividades empresariais que vem enfrentando maior ritmo de transformação tecnológica, o que proporciona melhorias na gestão, redução de custos e tende a melhorar o relacionamento com o consumidor. (PARENTE, 2011).

De acordo com Gianotti (2013), o grande desafio dos varejistas está fortemente relacionado com a habilidade de tomar decisões de maneira que consiga satisfazer as necessidades dos clientes. Parente (2011) afirma que os varejistas precisam estar atentos aos desafios da concorrência e às oportunidades de mercado. O autor afirma ainda que para isso é necessário estar em constante processo de melhoria, buscando sempre ferramentas avançadas de tecnologia, gerenciais e mercadológicas, alcançando assim a satisfação do consumidor e minimizando custos operacionais. (PARENTE, 2011).

2.1.1 Supermercado

Segundo Gianotti (2013), o conceito de supermercado instalou-se no Brasil no final de 1950, quando alguns pequenos comerciantes buscaram apresentar o conceito de autosserviço, porém sem ter sucesso. O primeiro supermercado instalado no país foi em 1953, ofertando além de mercearia, carnes, frutas e legumes. O estabelecimento já apresentava na época o *layout* conhecido hoje, com divisões de seções, com propagandas e promoções de produtos, inclusive nas pontas de gôndolas. (VAROTTO, 2006).

De acordo com Fingerl (1996), o supermercado se caracteriza por ser o local das compras rotineiras, abranger os clientes do bairro em que está instalado, bem como apresentar a sua estrutura de acordo com o perfil dessa área. O autor apresenta ainda que o *mix* de produtos dos supermercados é predominantemente composto por itens alimentícios, e que em sua maioria, a gestão do estabelecimento é familiar. (FINGERL, 1996).

Segundo Fingerl (1996), utilizar ferramentas de automação e tecnologia nos supermercados é fundamental para sua operação. Além de auxiliar na gestão das compras, estoques, preços e vendas, as tecnologias podem auxiliar na tomada de decisões. Nos supermercados são comumente adotadas tecnologias de automação comercial, que envolvem o uso do código de barras dos produtos e a utilização do *scanner* para fazer a leitura dos mesmos para integrar as vendas com o estoque. Há

ainda a utilização das balanças e esteiras rolantes de produtos acopladas nos caixas. (CAVALCANTI, 2014).

Dentro do conjunto de tecnologias adotadas nos supermercados destaca-se a implementação do Intercambio Eletrônico de Dados – EDI (*Electronic Data Interchange*), ferramenta que permite a troca eletrônica de informações entre uma empresa e outra, proporcionando melhorias nos resultados operacionais e estratégicos. O EDI permite que as empresas se comuniquem com toda a cadeia de suprimentos, reduzindo as taxas de erros e aumentando a eficiência dos processos. A partir disso, o supermercado tem disponibilidade de produtos e melhora sua relação com o cliente. (PORTO; BRAZ; PLONSKI, 2000; O..., 2016).

Conforme Bravo (2017), os supermercados devem entender de que forma a tecnologia pode afetar e beneficiar o negócio. Neste contexto, a disseminação da internet, aliado ao desenvolvimento dos dispositivos móveis, trouxe um novo desafio para os varejistas, visto que atualmente os consumidores procuram uma maneira diferente de se relacionar. (MORGADO, 2017).

O e-commerce já se tornou uma realidade para os supermercados, mostrando-se como um benefício ao empresário. A ferramenta permite conhecer e entender as necessidades dos consumidores, pois nesse modelo de compra o cliente sempre é identificado, tornado possível que nesse ambiente os anúncios das ofertas sejam segmentados de acordo com o perfil do usuário. (BRAVO, 2017). A tecnologia disponibiliza outros benefícios aos clientes como a possibilidade de não ter que enfrentar filas para realizar o pagamento, o que pode ser efetuado de forma segura com um aplicativo e dados do cartão de crédito. Há também os aplicativos que auxiliam o consumidor em suas compras, deixando a seu dispor uma lista de compras baseada no histórico de compras anteriores. (ALBUQUERQUE, 2018).

Os hábitos de compra dos consumidores estão mudando e o varejo não pode permanecer fora desse movimento e, com inovações e mudanças no formato de gestão é que os supermercados serão capazes de manter e fidelizar seus clientes, aumentando as vendas e mantendo a rentabilidade do negócio. (BRAVO, 2017).

2.2 ARMAZÉM DE DADOS

Para Han, Kamber e Pei (2012), um armazém de dados, comumente conhecido como *Data Warehouse*, refere-se a um repositório de dados, fornecendo uma

plataforma sólida de dados históricos para análise. Os autores apresentam ainda que o DW contempla dados referentes a clientes, fornecedores, produtos e vendas, armazenando as informações que as empresas necessitam para fazer decisões estratégicas. Para manter um *DW* é necessário que se realize a limpeza, integração e consolidação dos dados. Considera-se necessário também que tecnologias de suporte a decisão sejam disponibilizadas, permitindo que os analistas obtenham uma visão geral dos dados, e possam tomar decisões sólidas a partir das informações encontradas. (HAN; KAMBER; PEI, 2012).

Os dados gerados diariamente em diversos setores das empresas estão sendo extraídos a fim de alimentar um *Data Warehouse* (armazém de dados) e consequentemente, criar uma memória para as mesmas. Esse processo de armazenamento de dados tem o objetivo de reunir dados distintos coletados em toda a organização. (BARRY; LINNOF, 1997). Para isso, existe o processo ETL, que consiste na extração, transformação e carregamento de dados para a construção de um *DW*. (PAULA, 2012).

A primeira etapa consiste em extrair dados de outras bases de dados e ficheiros, como por exemplo arquivos de texto. O passo seguinte refere-se ao tratamento e limpeza dos dados coletados, padronizando-os quanto ao tamanho, tipo, correção de erros de digitação, entre outros. Este processo tem grande importância uma vez que, a partir dessas informações coletadas e transformadas, será criado um *Data Warehouse* sólido e confiável. Na última etapa é que são de fato efetuadas as atividades de carga para que os dados sejam unidos em um *DW*. (PAULA, 2012; FERREIRA; MIRANDA; ABELHA; MACHADO, 2010).

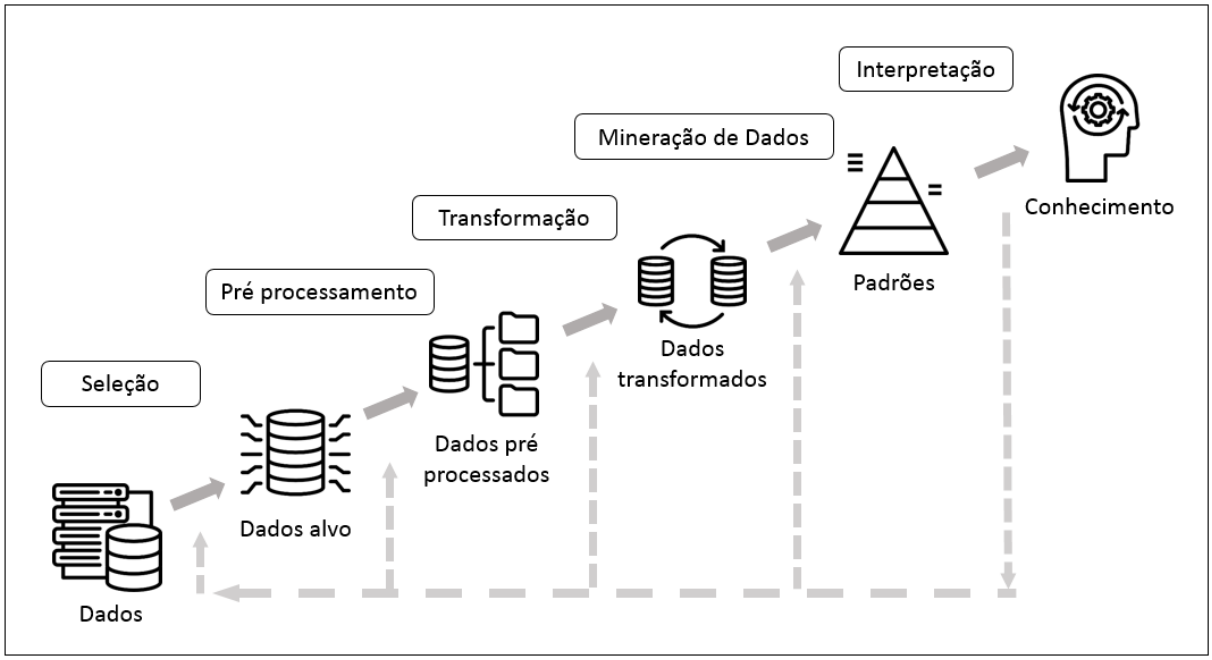
A fim de suportar a tomada de decisões, as organizações têm utilizado as informações contidas no *Data Warehouse* para analisar os padrões de compras dos consumidores, avaliar o desempenho de promoções realizadas entre os meses, semestres e anos; e refletir sobre as operações a fim de buscar novas fontes de lucro. (HAN; KAMBER; PEI, 2012).

Os autores Fayyad, Piatetsky-Shapiro e Smyth (1996) referem-se ao *DW* como sendo o processo de coleta e limpeza de dados transacionais, a fim de disponibilizá-los para análise e suporte à decisão, evidenciando ainda que este processo antecede e auxilia a etapa intitulada descoberta de conhecimento em bancos de dados (KDD).

2.2.1 Descoberta de Conhecimento em Bancos de Dados

Conforme Fayyad, Piatetsky-Shapiro e Smyth (1996), Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*), é caracterizado como o processo de selecionar, pré-processar e qualquer transformação necessária a ser efetuada em um banco de dados a fim de descobrir informações contidas nos dados. Os autores apresentam os passos necessários para completar o processo, representado na Figura 1.

Figura 1 - Etapas do processo de Descoberta de Conhecimento em Bancos de Dados



Fonte: Fayyad, Piatetsky-Shapiro e Smyth (1996).

Inicialmente, em uma base de dados, selecionam-se os dados a qual se deseja aplicar o método para encontrar informações desconhecidas. A partir disto, é realizada a limpeza e o pré-processamento no conteúdo, o que vem a ser a remoção de dados desnecessários para o procedimento, conhecidos como ruídos. A próxima etapa consiste em transformar os dados, a fim de remover os possíveis ruídos encontrados no passo anterior. Com a base de dados remodelada, verifica-se qual o método de *Data Mining* e o algoritmo a ser aplicado, de acordo com o objetivo. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

As informações encontradas na etapa anterior devem ser analisadas e interpretadas de acordo com o contexto da empresa a qual os dados pertencem. Nessa etapa é possível que seja necessário retornar para algum passo anterior, a fim de reavaliar e remodelar os dados novamente. O último passo do processo consiste em atuar sobre a descoberta realizada. É possível que estas informações sejam utilizadas diretamente, como também é possível utilizá-las para montar um plano de ação estratégico, ou reportá-las às partes interessadas. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

2.2.2 Mineração de Dados

Mineração de dados, do termo em inglês *Data Mining*, é uma etapa dentro do processo de KDD, que consiste na aplicação de métodos e algoritmos, com o objetivo de extrair padrões através dos dados. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Segundo Groth (2000), *Data Mining* é o processo de encontrar tendências e padrões em dados, tendo como objetivo classificar grandes quantidades de dados e descobrir novas informações. Para Hand, Mannila e Smyth (2001), o termo se caracteriza por ser a análise de conjuntos de dados que tem por objetivo encontrar modelos e padrões desconhecidos que possam ser, ao mesmo tempo, úteis e compreensíveis.

A mineração de dados é dividida em tarefas, que são classificadas em duas categorias: descritivas e preditivas. As tarefas descritivas caracterizam-se por apresentar os dados de uma forma compreensível pelo ser humano. Já as tarefas preditivas são caracterizadas por induzirem os dados a apresentar previsões. (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996; HAN; KAMBER; PEI, 2012).

Descrição (*description*) é a tarefa que tem o objetivo de simplesmente descrever o que os dados coletados apresentam. É possível encontrar explicações para comportamentos em categorias, como por exemplo dados coletados sobre pessoas, produtos e processos. Ou ainda, a tarefa é capaz de realizar comparações entre categorias diferentes, como identificar os consumidores que compram um produto regularmente, e aqueles consumidores que compram o mesmo produto raramente. (BERRY; LINOFF, 1997; HAN; KAMBER; PEI, 2012).

A tarefa classificação, ou *classification*, é o processo de encontrar um modelo que descreva e faça distinção entre classes ou conceitos. O modelo consiste em

examinar as características de um novo objeto, encontrado nos dados, e atribuí-lo à uma categoria predefinida. Em um banco, por exemplo, os clientes podem ser classificados por apresentarem um risco diferente à instituição, como baixo, médio ou alto. (BERRY; LINOFF, 1997; CARVALHO, 2001; HAN; KAMBER; PEI, 2012).

Estimation, ou estimativa, tem por objetivo avaliar dados coletados para estimar um valor, como por exemplo, determinar o saldo de um cartão de crédito: através da utilização de dados de entrada definidos, disponíveis no banco de dados de uma instituição financeira, é possível estabelecer um valor até então desconhecido para conceder o crédito. (BERRY; LINOFF, 1997; CARVALHO, 2001).

Similar às tarefas de classificação e estimativa, existe a tarefa de previsão, do termo em inglês *prediction*, com o propósito de classificar os dados de acordo com uma previsão de comportamento ou estimativas futuras. São utilizados dados históricos para construir um modelo que explique o comportamento atual, e este modelo é utilizado para prever o comportamento futuro. Como por exemplo, prever as compras futuras de um cliente específico, utilizando como base o histórico de compras. (BERRY; LINOFF, 1997; CARVALHO, 2001).

Clusterização (*clustering*) é a tarefa que busca agrupar os dados de acordo com sua similaridade, sem haver nenhuma categoria predefinida. Cabe ao analista determinar um significado ao cluster formado pelos dados. (BERRY; LINOFF, 1997).

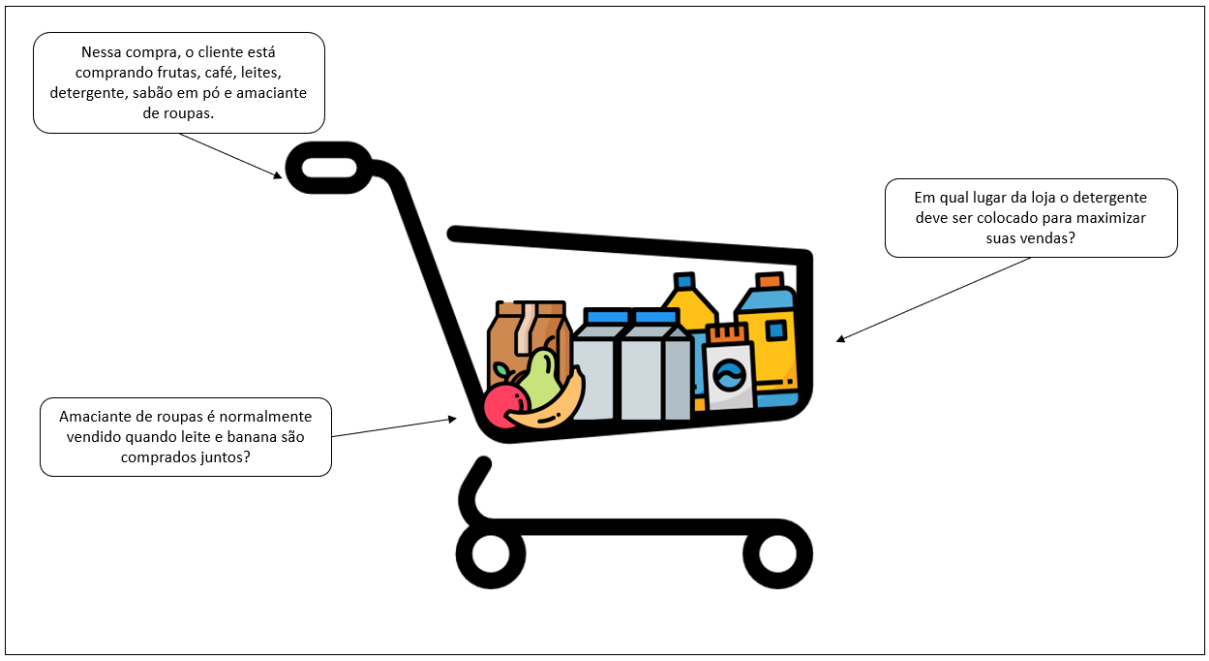
Padrões frequentes (*frequent patterns*), são padrões que ocorrem frequentemente em um banco de dados. Um exemplo dessa tarefa é a análise da cesta de compras (*market basket analysis*), que visa identificar os hábitos de consumo dos clientes, encontrando associações entre diferentes itens que, constantemente são comprados juntos, pelos clientes. (BERRY; LINOFF, 1997; HAN; KAMBER; PEI, 2012).

2.3 ANÁLISE DA CESTA DE COMPRAS

Análise da cesta de compras (*market basket analysis*) utiliza as informações sobre o que o cliente está comprando, com o objetivo de fornecer *insights* sobre os clientes e suas preferências de compras. A Figura 2 representa as compras realizadas por um indivíduo em um supermercado, o carrinho contém diversos itens diferentes, onde é possível visualizar o que este cliente específico está comprando. Através da ferramenta, é possível analisar todas as compras de todos os clientes, obtendo assim

um maior número de informações. (BERRY; LINOFF, 1997; HAN; KAMBER; PEI, 2012).

Figura 2 – Carrinho de compras



Fonte: Adaptado de Berry e Linoff (1997, p.125).

Aplicando a tarefa no banco de dados dos supermercados, é possível analisar quais itens são frequentemente vendidos juntos e quais são mais propícios a promoções, respondendo à questionamentos como: “se os clientes estão comprando leite, qual a probabilidade de também comprarem pão?”. Através das análises, pode-se construir um plano de marketing como também definir estratégias de propaganda. Como por exemplo, os itens que frequentemente são vendidos juntos, podem ser expostos em proximidade um com o outro, incentivando a venda combinada de tais itens. (BERRY; LINOFF, 1997; HAN; KAMBER; PEI, 2012).

Os resultados da análise da cesta de compras são apresentados em forma de regras de associação, as quais são consideradas as representações mais populares de padrões descobertos em bases de dados, e são definidas em três características diferentes: útil, trivial e inexplicável. As regras úteis são aquelas as quais é possível tomar decisões rápidas e agir sobre as informações. Quando se descobre que um determinado item de higiene costuma ser vendido com alguma cerveja, pode-se reposicionar um dos produtos de forma que os dois fiquem próximos, induzindo que a compra aconteça. (BERRY, LINOFF, 1997; HAND; MANNILA; SMYTH, 2001).

As regras triviais são aquelas já conhecidas de certa forma pelos gerentes ou administradores, como a regra de que, quem comprar carne de hambúrguer, comprará também pães para hambúrguer. Esta regra também pode apresentar o resultado de possíveis campanhas realizadas anteriormente, em que a empresa faz uma promoção na compra de dois itens da mesma seção. Nessa situação, a mineração de dados não estará apresentando uma regra desconhecida, e sim apresentando que a promoção obteve resultados positivos. (BERRY; LINOFF, 1997).

Regras inexplicáveis parecem não ter alguma explicação e não sugerem uma possível ação, como por exemplo, descobrir que refrigerantes e vassouras tendem a ser comprados juntos. Diante dessa regra, não é possível identificar o comportamento do consumidor ou sugerir uma ação futura. Através da regra é possível investigar se havia um desconto considerável nas vassouras, ou se elas estão posicionadas em uma área próxima aos refrigerantes. Independente da causa, nessas situações normalmente não se encontra uma justificativa apenas analisando os dados. (BERRY; LINOFF, 1997).

2.3.1 Conjuntos de Itens Frequentes

Considere I = {i1, i2, …, in} um conjunto de itens. Considere D uma base de dados contendo todas as transações, onde cada transação T é um conjunto de itens tal que $T \subseteq I$. Cada transação está associada à um identificador TID. Seja A um conjunto de itens, uma transação T contém A se $A \subset T$. Uma regra de associação se refere à forma $A \Rightarrow B$, onde $A \subset I$, $B \subset I$, $A \neq \emptyset$, $B \neq \emptyset$ e $A \cup B$. (HAN; KAMBER; PEI, 2012).

Os padrões de itens encontrados na análise são representados na forma de regras de associação, as quais são compostas por dois conjuntos de itens, denominados lado esquerdo (LHS) e lado direito (RHS). O primeiro representa os itens antecedentes, enquanto que o segundo representa os itens consequentes, interpretados na forma: se há a compra do item antecedente, então há a compra do item consequente. (ANSELMO, 2017). Um exemplo disto é apresentado na regra abaixo, onde a informação de que a compra de leite implica na compra de pão é representada pela seguinte regra de associação:

$$\{\text{Leite} \Rightarrow \text{pão}\}$$

Para definir as regras de associação significantes, existem as medidas de interesse, ou seja, existe um limite mínimo de suporte e confiança a considerar, que

são especificados pelo analista. Suporte de uma regra representa a porcentagem de transações na base de dados que contém os itens A e B, ou seja, ele apresenta o número total de registros que contém tais itens nas transações. O suporte da regra {A} ⇒ {B} é dado pela seguinte equação:

$$Suporte\ (A\ \Rightarrow\ B) = \frac{Frequencia\ de\ A\ e\ B}{Total\ de\ transações} \quad (1)$$

O numerador refere-se ao número de transações em que A e B ocorreram simultaneamente e o denominador trata do número total de transações da base de dados. (BERRY; LINOFF, 1997; HAN; KAMBER; PEI, 2012; ANSELMO, 2017).

A confiança, por sua vez, é representada pela equação de número dois, onde o numerador refere-se às transações em que A e B ocorreram simultaneamente e o denominador se refere à quantidade de transações em que o item A ocorre, demonstrando a probabilidade de ocorrer B, dado a ocorrência de A. (BERRY; LINOFF, 1997; HAN; KAMBER; PEI, 2012; ANSELMO, 2017).

$$Confiança\ (A\ \Rightarrow\ B) = \frac{Suporte\ (A \cup B)}{Suporte\ (A)} \quad (2)$$

A Tabela 2 representa algumas transações de um supermercado hipotético para exemplificar as medidas apresentadas acima.

Tabela 2 – Exemplo de transações de um supermercado

TID	Itens
1	pão, leite
2	pão, bolacha, presunto, queijo
3	carne, tomate, cebola, brócolis
4	leite, café, bolacha, chocolate, pão
5	requeijão, pão, café, leite

Fonte: Adaptado de Tan, Steinbach e Kumar (2009, p. 390)

É possível verificar que três das cinco transações contém os produtos leite e pão. Sendo assim, o suporte para a regra {Leite ⇒ Pão} é calculado por meio da seguinte equação:

$$Suporte\ (Leite\ \Rightarrow\ Pão) = \frac{3}{5} = 60\% \quad (3)$$

Para a confiança, utiliza-se a informação de que três delas contém os itens leite e pão juntos, e quatro transações que contém o item leite. A confiança da regra {Leite ⇒ Pão} é calculada conforme a seguinte equação:

$$Confiança\ (Leite\ \Rightarrow\ Pão) = \frac{0.60}{(\frac{4}{5})} = 75\% \quad (4)$$

O valor de confiança indica que 75% dos consumidores que compram pão, também compram leite. (TAN; STEINBACH; KUMAR, 2009).

A medida de suporte tem o papel de auxiliar o analista a identificar as regras mais importantes para o negócio, visto que algumas regras de baixo suporte não se tornam interessantes, pois raramente os itens são comprados juntos pelos clientes. A confiança demonstra a confiabilidade da relação que a regra apresenta, mostrando que, maior a confiança da regra A ⇒ B, então maior será a probabilidade de B estar presente nas transações que contenham A. (TAN; STEINBACH; KUMAR, 2009).

Verifica-se que em bases de dados reais, o número de regras geralmente é muito alto, dificultando a análise. Para isso, pode-se adotar outras medidas de interesse para identificação das regras de associação. Uma das medidas mais utilizadas junto do suporte e confiança é o *lift,* que indica o quanto mais frequente se torna B, dado a ocorrência de A. Tal medida é representada pela equação de número cinco:

$$Lift\ (A\ \Rightarrow B) = \frac{Confiança\ (A \Rightarrow B)}{Suporte\ (B)} \quad (5)$$

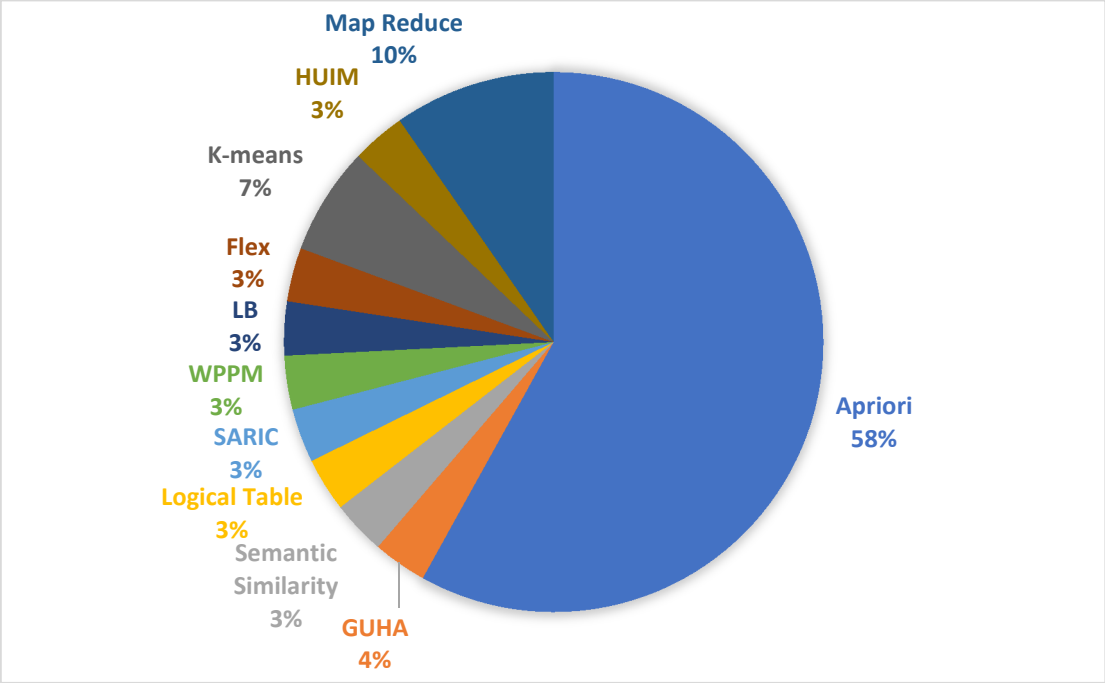
O *Lift* é uma medida utilizada para avaliar a dependência entre um item e outro. Se $Lift\ (A\ \Rightarrow B)$ = 1, então os itens A e B são independentes. Se $Lift\ (A\ \Rightarrow B)$ > 1, então A e B são positivamente dependentes, isto é, são vendidos juntos com maior frequência. Porém se $Lift\ (A\ \Rightarrow B)$ < 1, os itens são negativamente dependentes. Assim sendo, quanto maior o valor do *lift*, mais interessante é a regra, pois o item A aumentou B em maior taxa. (GONÇALVES, 2005).

2.4 ALGORITMOS DE KDD

Existem diversos algoritmos para as tarefas de mineração de dados e, da mesma forma existem diferentes algoritmos para a análise da cesta de compras, como: Apriori, FP-growth, Basic, Genex, DHP, PHP, entre outros. (HAN, KAMBER, PEI, 2012; ANSELMO, 2017). Para a definição do algoritmo a ser utilizado neste trabalho, foi realizada uma breve busca na literatura, onde se priorizou a utilização do algoritmo como fator de escolha. A Tabela 3 apresenta as bases de dados para a pesquisa, os termos de busca e os resultados. Em sua totalidade, a pesquisa retornou 82 trabalhos, dos quais 31 apresentaram relação com os termos buscados, onde foram excluídos trabalhos não escritos em Português e Inglês, e aqueles trabalhos cuja abordagem difere da aplicação de um algoritmo para o estudo.

A busca possibilitou a identificação dos algoritmos utilizados com a abordagem do *Market Basket Analysis*, onde destaca-se o algoritmo Apriori, estando presente em 18 dos 31 trabalhos relacionados ao tema. O Gráfico 1 apresenta a relação dos algoritmos encontrados com a pesquisa.

Gráfico 1 – Relação dos algoritmos encontrados



Fonte: Elaborado pela autora.

Tabela 3 – Resultado da pesquisa para escolha do algoritmo

Fonte	Termos de busca	Resultados	Em concordância	Apriori	GUHA	Semantic Similarity	Logical Table	SARIC	WPPM	LB	Flex	K-means	HUIM	Map Reduce
EBSCOHost	Mineração de Dados AND Análise da Cesta de Compras AND Algoritmo	0	0	0	0	0	0	0	0	0	0	0	0	0
EBSCOHost	Data Mining AND Market Basket Analysis AND Algorithm	36	11	4	1	1	1	1	1	1	1	0	0	0
Emerald	Mineração de Dados AND Análise da Cesta de Compras AND Algoritmo	0	0	0	0	0	0	0	0	0	0	0	0	0
Emerald	Data Mining AND Market Basket Analysis AND Algorithm	26	13	10	0	0	0	0	0	0	0	2	1	0
ProQuest	Mineração de Dados AND Análise da Cesta de Compras AND Algoritmo	0	0	0	0	0	0	0	0	0	0	0	0	0
ProQuest	Data Mining AND Market Basket Analysis AND Algorithm	20	7	4	0	0	0	0	0	0	0	0	0	3

Fonte: Elaborado pela autora.

Em consequência da busca realizada, foi definida a utilização do algoritmo Apriori neste trabalho, o qual foi primeiramente proposto por Agrawal e Srikant em 1994, com o objetivo de minerar conjuntos de itens frequentes para encontrar regras de associação e permitir a extração de regras que contenham mais de um item no consequente da mesma. O algoritmo parte do princípio de que, se um conjunto de itens é frequente, então todos os seus subconjuntos são frequentes. A parametrização do Apriori é baseada em um valor de suporte e confiança mínimos, especificados pelo analista. (ANSELMO, 2017).

Primeiramente o algoritmo avalia o suporte de cada item dentro do conjunto de dados proposto, eliminando aqueles que não satisfazem o suporte mínimo estipulado. A partir disso, a cada passagem do algoritmo pelos dados, ele inicia com o conjunto de itens já encontrados na primeira passagem. Em cada iteração, o algoritmo avalia o suporte dos dados novamente, acrescentando os itens que apresentam suporte igual ou superior ao estabelecido, no conjunto pré-definido na primeira passagem. As iterações continuam até que o algoritmo encontre um conjunto de itens vazio, ou seja, até que não se encontre mais os valores mínimos para o suporte estabelecido. (TAN; STEINBACH; KUMAR, 2009; HAN; KAMBER; PEI, 2012; ANSELMO, 2017).

Cada iteração gera novos conjuntos de itens, a partir dos conjuntos frequentes encontrados na iteração anterior. O algoritmo utiliza essa técnica com a finalidade de controlar o crescimento exponencial das regras a serem encontradas, apresentando assim, somente os itens de maior relevância dado o suporte definido previamente. (TAN; STEINBACH; KUMAR, 2009).

Encontrados os conjuntos de itens frequentes através do suporte, a próxima etapa do algoritmo consiste em gerar as regras de associação considerando um valor mínimo de confiança estipulado. Dessa forma observa-se que o suporte tem o objetivo de gerar os conjuntos de itens frequentes, enquanto que a confiança gera regras de associação através dos conjuntos frequentes encontrados inicialmente. (ANSELMO, 2017).

Expostos os conteúdos e aplicações do tema Data Mining, o próximo capitulo abordará os procedimentos metodológicos utilizados nesta monografia.

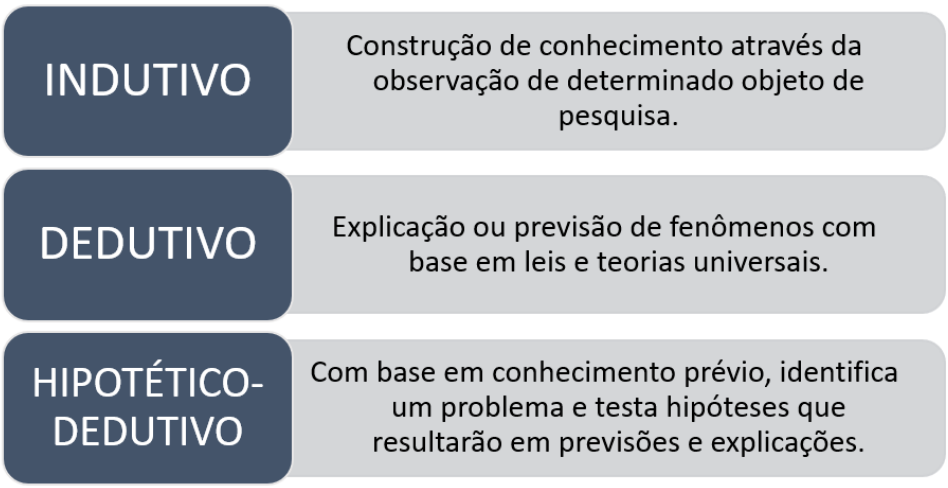
3 PROCEDIMENTOS METODOLÓGICOS

Este capítulo aborda os procedimentos metodológicos utilizados para a condução dessa monografia, que engloba o método científico, o método de pesquisa, o método de trabalho, bem como as técnicas de coleta e análise de dados.

3.1 MÉTODO CIENTÍFICO

De acordo com Dresch, Lacerda e Antunes JR (2015) os métodos científicos apresentam como o conhecimento foi construído e, para a sua escolha, deve-se levar em consideração o ponto de partida da pesquisa, bem como seus objetivos. Os métodos científicos estão explanados na Figura 3.

Figura 3 – Métodos científicos



Fonte: Dresch, Lacerda e Antunes JR (2015)

Para essa monografia, o método científico a ser considerado é o dedutivo, uma vez que se caracteriza por utilizar raciocínio lógico, além de se basear em leis e teorias para a construção do conhecimento.

3.2 MÉTODO DE PESQUISA

Segundo Gil (2017), a pesquisa é definida como um procedimento que utiliza métodos e técnicas a fim de encontrar soluções e respostas para problemas. O autor afirma ainda que as pesquisas podem ser tanto de caráter teórico quanto prático, e

que as mesmas podem favorecer uma a outra. (GIL, 2017). Conforme Silva e Menezes (2005) a pesquisa é classificada de pontos de vista diferentes, que podem ser observados no Quadro 2.

Quadro 2 – Classificações de pesquisa

Ponto de Vista	Classificação
Natureza	Básica
	Aplicada
Abordagem	Quantitativa
	Qualitativa

Fonte: Adaptado de Silva e Menezes (2005).

Em relação ao ponto de vista da natureza, a pesquisa pode ser básica, quando tem por objetivo gerar novos conhecimentos sem aplicação prática. Já a pesquisa aplicada visa gerar conhecimentos para aplicação prática. (SILVA; MENEZES, 2005). Além disso, Gil (2017) afirma que a pesquisa aplicada também se caracteriza por ser aplicada a uma situação específica. Nesse contexto, a presente pesquisa classifica-se como aplicada, uma vez que objetiva gerar conhecimentos para aplicação da ferramenta DM em um supermercado.

Quanto a abordagem, as pesquisas quantitativas traduzem as informações em números a fim de classificá-las e analisá-las. Por outro lado, as pesquisas qualitativas se caracterizam como sendo descritivas, com o objetivo de interpretar os fenômenos, bem como atribuir significados aos mesmos. (SILVA; MENEZES, 2005). Assim sendo, essa pesquisa é quantitativa, sendo que dados numéricos serão apresentados e analisados.

No que se refere ao método, Dresch, Lacerda e Antunes JR (2015) destacam quatro utilizados nas pesquisas, descritos no Quadro 3.

Quadro 3 – Métodos de pesquisa

Procedimentos Técnicos	Descrição
Estudo de Caso	Busca melhor compreender os fenômenos em seu contexto real, através de descrições detalhadas baseada em fontes de dados.
Pesquisa-ação	Busca resolver e explicar problemas, além de produzir conhecimento tanto para a prática quanto para a teoria.
Survey	Objetiva desenvolver conhecimento em uma área específica, através de coleta e análise de dados, para avaliar o comportamento das pessoas e ou o ambiente em que se encontram.
Modelagem	Busca o melhor entendimento dos problemas, através de representações simplificadas da realidade, permitindo compreensão do ambiente estudado.

Fonte: Dresch, Lacerda e Antunes JR (2015)

Conforme Neto e Pureza (2012), na engenharia de produção, a gestão da produção de bens e serviços abrange diversas decisões acerca das atividades desenvolvidas no planejamento estratégico, tático ou operacional. Dentre as atividades dos gestores, esta pode ser considerada a mais desafiadora, visto que as decisões tomadas refletem diretamente no sistema, e o mesmo deve operar da melhor forma possível. Nestas situações, utilizar modelos permite a melhor compreensão do ambiente, identificação de problemas, além de formular estratégias para apoiar o processo de tomada de decisões. (NETO; PUREZA, 2012). Isto porque a modelagem quantitativa parte da premissa de que é possível construir modelos que expliquem ao menos parte do comportamento de processos reais, ou que é possível capturar ao menos parte dos problemas encontrados nas tomadas de decisões dos processos operacionais. (BERTRAND; FRANSOO, 2002).

Neto e Pureza (2012) classificam os modelos em concretos ou abstratos. O primeiro caracteriza-se por utilizar protótipos, maquetes, materiais físicos que representem a realidade. Por outro lado, modelos abstratos se caracterizam por utilizar gráficos, plantas, tabelas, entre outros, para avaliar o modelo que representa a situação em questão. (NETO; PUREZA, 2012). Os autores descrevem modelos abstratos como aqueles que utilizam técnicas analíticas, com aplicação da matemática e estatística, como também técnicas experimentais, como a simulação do modelo real.

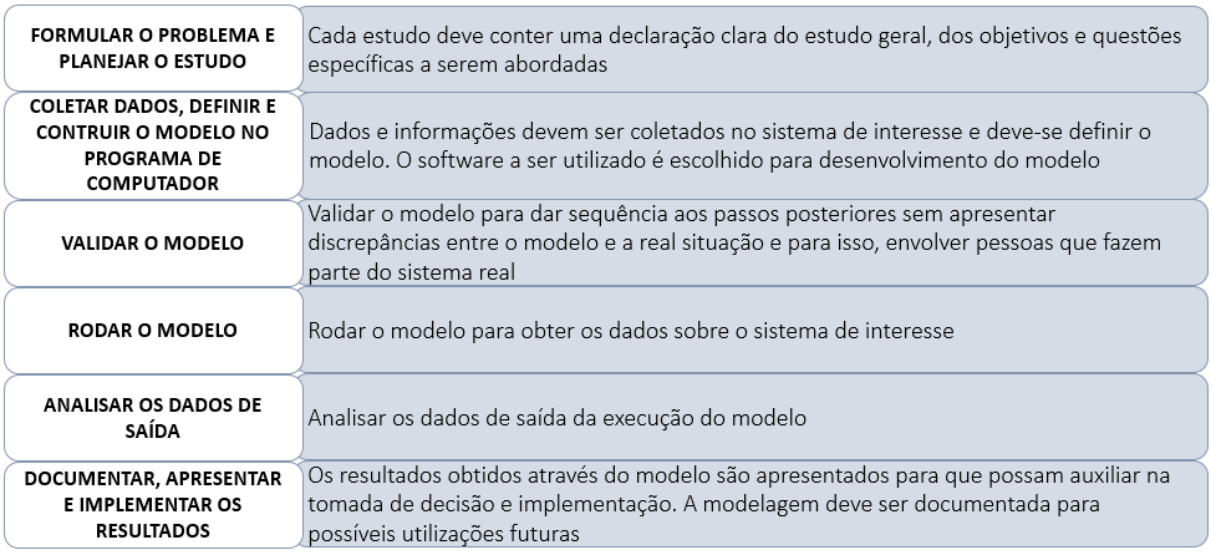
(NETO; PUREZA, 2012). Portanto, neste trabalho será adotado o modelo abstrato, uma vez que sua abordagem é quantitativa e utilizará técnicas analíticas para avaliação do modelo.

De acordo com Bertrand e Fransoo (2002) e Neto e Pureza (2012), as pesquisas baseadas em modelos quantitativos podem ser classificadas em axiomáticas ou empíricas. As pesquisas axiomáticas são dirigidas a problemas idealizados, onde procura-se obter soluções para o modelo. Por outro lado, a pesquisa empírica é dirigida por descobertas e medidas interpretativas, com foco em aplicar modelos científicos, obtidos através de pesquisas teóricas, em processos reais. (BERTRAND; FRANSOO, 2002; NETO; PUREZA, 2012). Assim sendo, este trabalho se enquadra na classe empírica, uma vez que objetiva modelar um sistema para entender as relações existentes entre as compras efetuadas, utilizando o algoritmo *Apriori*.

As pesquisas empíricas podem ser classificadas em descritivas ou normativas. Enquanto que a primeira objetiva criar modelos para compreensão dos processos, a segunda tem o intuito de desenvolver políticas, estratégias e ações para melhorar a real situação, baseando-se em modelos de programação matemática. (BERTRAND; FRANSOO, 2002; NETO; PUREZA, 2012). Logo, este trabalho se enquadra na pesquisa empírica normativa.

Neste trabalho optou-se por utilizar o método proposto por Law e Kelton (1991) para a condução da modelagem. Os autores apresentam uma série de passos a seguir, os quais são apresentados na Figura 4.

Figura 4 – Método para a construção do modelo



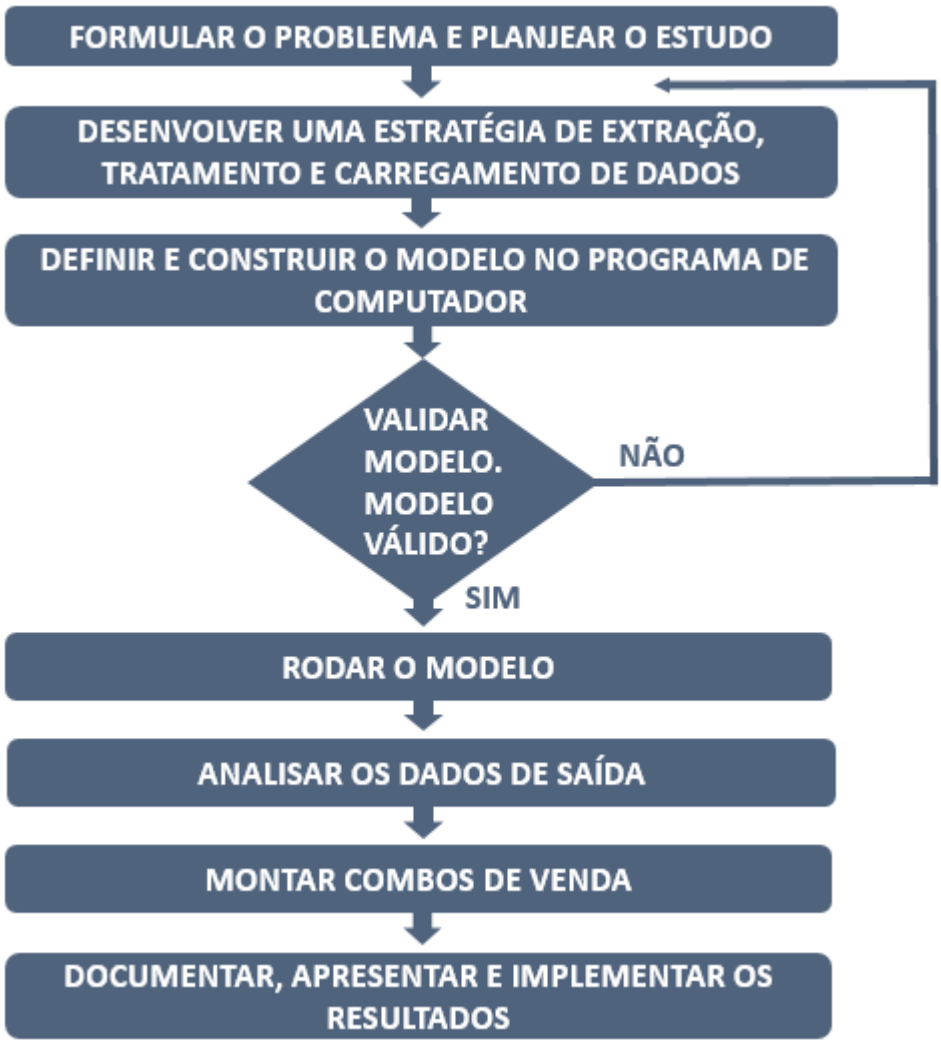
Fonte: Law e Kelton (1991).

Na próxima seção será detalhada a condução desse trabalho de modelagem seguindo as etapas de Law e Kelton (1991).

3.3 MÉTODO DE TRABALHO

Após a definição do método de pesquisa, deve-se estabelecer o método de trabalho, o qual determina a sequência de passos necessários para alcançar os objetivos previamente definidos. (DRESH, LACERDA, ANTUNES JR, 2015). Com o método escolhido para a condução do modelo, segue o detalhamento de cada etapa do modelo.

Figura 5 – Método de trabalho



Fonte: Elaborado pela autora.

Na etapa de número um deve-se esclarecer o contexto do problema e definir os objetivos do sistema. No contexto organizacional foi identificada a constante necessidade de apoio à tomada de decisões no que se refere a promoções de itens de venda. Essa tarefa normalmente é conduzida informalmente, por meio de uma simples análise de produtos considerados importantes pelo gestor do supermercado, e produtos que, ao fazerem parte de alguma promoção, poderiam chamar a atenção dos clientes, induzindo-os à compra.

A fim de entender sobre o conceito de promoções de itens e buscar possíveis ferramentas que poderiam auxiliar o gestor nessa tarefa específica, realizou-se pesquisas em sites confiáveis e nesta busca identificou-se o *Data Mining*, que apresentava como um de seus benefícios a descoberta de itens frequentes,

evidenciando o quanto a ferramenta poderia auxiliar o gestor a fazer escolhas mais assertivas ao definir os itens da promoção. Por meio das informações encontradas formulou-se a justificativa empresarial e os objetivos, detalhados no capítulo um.

A etapa número dois consiste na coleta dos dados do supermercado, que são as transações ocorridas no período entre 01 de julho de 2017 a 31 de junho de 2018. Para isso, foi desenvolvida uma estratégia de extração, tratamento e carregamento de dados, os quais foram coletados junto ao administrativo da empresa. Em seguida, os dados foram tratados para importar no software, cumprindo com a etapa de carregamento.

Esta etapa também aborda a definição e construção do modelo, a qual iniciou com a definição da tarefa de *Data Mining* a ser utilizada e a busca por seus conceitos e aplicações. Esta busca foi realizada no início deste trabalho em artigos, dissertações e vídeos, a fim de compreender como utilizar a ferramenta. Em seguida foi determinado o software a ser utilizado que, de acordo com Law e Kelton (1991), deve-se considerar a facilidade de acesso e o conhecimento para utilização da ferramenta. Portanto, optou-se por utilizar o *software* R para rodar o modelo, pois o mesmo apresenta versão gratuita e há diversos estudos disponíveis que foram utilizados de apoio para esta pesquisa. Em sequência, o modelo foi construído de acordo com as especificações do *software*.

A próxima etapa é a validação do modelo, que tem a finalidade de encontrar possíveis inconsistências e *bugs* e corrigi-los, se necessário. De acordo com Zumel e Mount (2014), ao construir um modelo, é preciso verificar se o mesmo funciona com dados de teste. Para isso, o modelo foi utilizado com um conjunto de dados de transações para análise de cesta de compras, que está disponível no R, chamado *Groceries*, onde foram utilizados os comandos (modelo) no *software* a fim de validá-los, para posterior utilização com os dados coletados, e analisar os resultados gerados.

No que tange à sua aplicabilidade, o modelo foi submetido à avaliação de um grupo focal. De acordo com Bruseberg e McDonagh-Philip (2002), os grupos focais consistem em grupos de pessoas reunidas para discutir um assunto específico. Os autores afirmam que grupos focais podem ser utilizados junto a outras técnicas com a finalidade de suportar as discussões do grupo, auxiliar a triangulação dos dados e colaborar com novas ideias sobre um determinado problema. (BRUSEBERG; MC-

DONAGH-PHILIP, 2002). O grupo focal foi realizado com três participantes que atuam diretamente na escolha de itens para as promoções do supermercado em questão.

Nas etapas de número quatro e cinco, o modelo validado foi utilizado no R com as transações de compra do supermercado, aos quais o algoritmo *Apriori* foi aplicado para gerar as regras de associação.

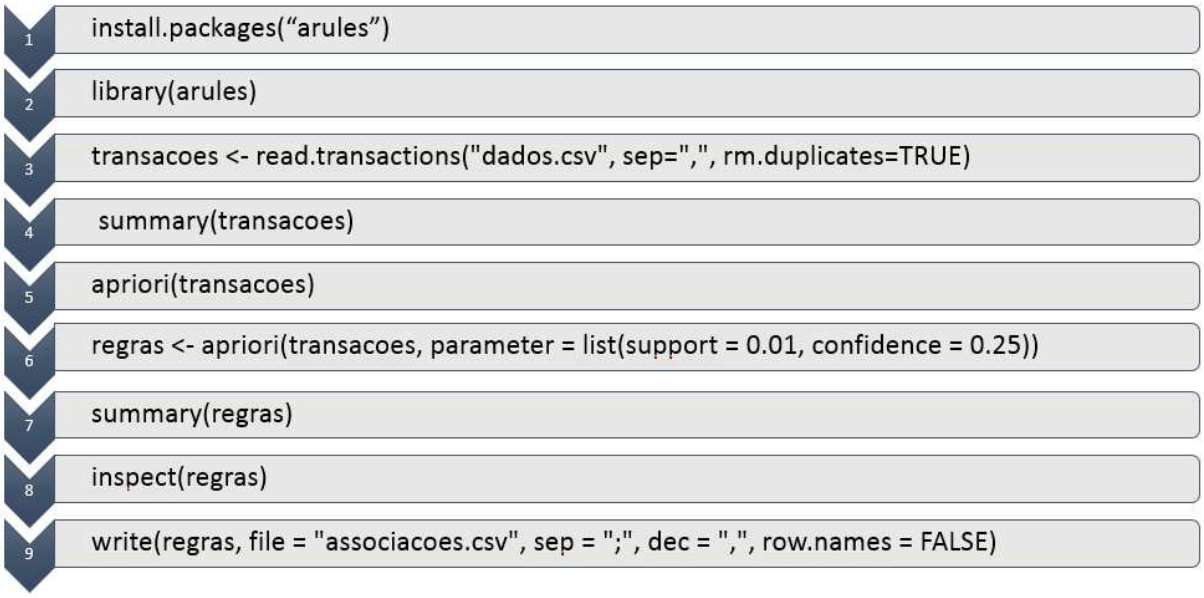
A última etapa proposta por Law e Kelton (1991) consiste em apresentar os resultados obtidos por meio do modelo para que possam auxiliar na tomada de decisão e implementação. Os autores recomendam que o a modelagem seja documentada para possíveis utilizações futuras. O presente trabalho foi documentado em forma de monografia e apresentado à banca avaliadora como trabalho de conclusão do curso de Engenharia de Produção. Posteriormente, os resultados obtidos serão apresentados a empresa que foi objeto de estudo deste trabalho.

3.4 CONSTRUÇÃO DO MODELO

Para desenvolver o modelo foi escolhido o *software* R, que está disponível gratuitamente e fornece uma grande variedade de técnicas gráficas e estatísticas para análise de dados. Essas técnicas são possíveis por meio de pacotes disponibilizados pelo R. No presente trabalho, o pacote a ser utilizado é o *arules*, o qual fornece a infraestrutura necessária para analisar conjuntos de dados e conta também com a função do algoritmo *Apriori,* com o objetivo de encontrar padrões frequentes e regras de associação. (HAHSLER; GRÜN; HORNIK; BUCHTA, 2005).

No R são necessários digitar comandos para que o *software* importe os dados e explore as transações para entregar os resultados. Portanto, o modelo foi construído na forma destes comandos, que são os passos necessários para se obter as regras de associação. Dessa forma, o modelo construído no *software* é apresentado na Figura 6.

Figura 6 – Modelo proposto



Fonte: Elaborado pela autora.

O primeiro comando, *install.packages(“arules”)* é utilizado para instalar o pacote *arules*, pois o mesmo não faz parte das funções iniciais do *software*. Em seguida, utiliza-se a função *library(arules)* para utilizar sua estrutura sobre os dados importados posteriormente.

O próximo comando, *read.transactions()* é utilizado para importar dados de um arquivo. O arquivo com os dados estava em formato CSV (*Comma Separated Values*), o qual faz uso de vírgulas para separar os caracteres de arquivos de texto, e no R é preciso especificar o formato dos dados. Neste primeiro comando também foi necessário utilizar *rm.duplicates* para remover itens duplicados em uma mesma linha. Isto porque nas transações o mesmo item pode aparecer mais de uma vez, e aqui essa informação não fará diferença, pois conforme abordado por Anselmo (2017), neste trabalho também se deseja verificar os padrões de compra adotados pelos clientes e, por isso, interessa apenas contabilizar o número de transações em que o item se encontra face ao número total de transações existentes.

O comando *summary()* apresenta uma visão geral do conjunto de dados, onde informa a quantidade total de transações, a quantidade de itens, bem como a informação de quais são os itens mais frequentes dentro do conjunto de dados.

A função *apriori()* especifica o uso do algoritmo sobre os dados, a qual está sob uma configuração padrão e não há especificação das medidas de interesse suporte e

confiança. Ao utilizar este comando, o *software* também apresenta uma visão geral dos dados, visto a aplicação do algoritmo.

No comando de número cinco, aplica-se novamente o algoritmo aos dados, porém com medidas de suporte e confiança já estabelecidas. Tais medidas definem quais regras devem ser mantidas e quais devem ser descartadas. De acordo com Han, Kamber e Pei (2012), devem-se estabelecer valores altos de suporte e confiança, caso contrário são geradas muitas regras desinteressantes. Neste trabalho foram utilizados valores similares ao trabalho realizado por Anselmo (2017), onde foram estabelecidos valores de 1% para suporte e confiança.

Novamente utiliza-se o comando *summary()*, agora porém para visualizar de maneira geral as regras geradas com aplicação do algoritmo e suas medidas de interesse. Nessa etapa também é possível visualizar a quantidade de regras geradas com as medidas estipuladas. Em seguida, utiliza-se o comando *inspect()* para visualizar todas as regras geradas no ecrã, mostrando inclusive os valores das medidas de interesse para cada regra.

Por fim, a função *write()* é utilizada para exportar os dados para um arquivo CSV, onde os resultados obtidos podem ser visualizados no Excel.

3.5 COLETA DE DADOS

Com o intuito de analisar a cesta de compras e, a partir disso, descobrir regras de associação, os dados coletados foram as transações do supermercado, ou seja, todas as vendas efetuadas no período de 01/07/2017 à 31/06/2018. Tais transações estão arquivadas no banco de dados da empresa e foram solicitadas ao administrativo, para enviar em formato Excel e via e-mail.

Han, Kamber e Pei (2012) afirmam que os dados normalmente contêm inconsistências e ruídos, de forma que necessitam de pré-processamento para análise, pois dados que apresentam baixa qualidade consequentemente apresentarão resultados de má qualidade. Para isso, o autor sugere que se faça a limpeza dos dados, que consiste na tentativa de preencher valores ausentes, suavizar ruídos e corrigir dados inconsistentes. (HAN; KAMBER; PEI, 2012).

A etapa de pré-processamento dos dados faz parte do processo da Descoberta de Conhecimento em Bancos de Dados (KDD) e, de fato, os dados coletados continham diversas informações que não eram necessárias para a aplicação do

algoritmo. O pré-processamento dos dados foi efetuado na própria planilha, onde foram excluídas informações do cupom fiscal, como dados do estabelecimento, código de barras, bem como a quantidade de cada item vendido, pois como já abordado anteriormente, pretende-se contar o número de transações em que um determinado item se encontra a fim de verificar sua frequência perante o total de transações. Desta forma, permanece na planilha apenas o nome do item nas transações. A transformação dos dados também se fez necessária devido à utilização do *software* escolhido, visto que o mesmo não aceita todos os formatos.

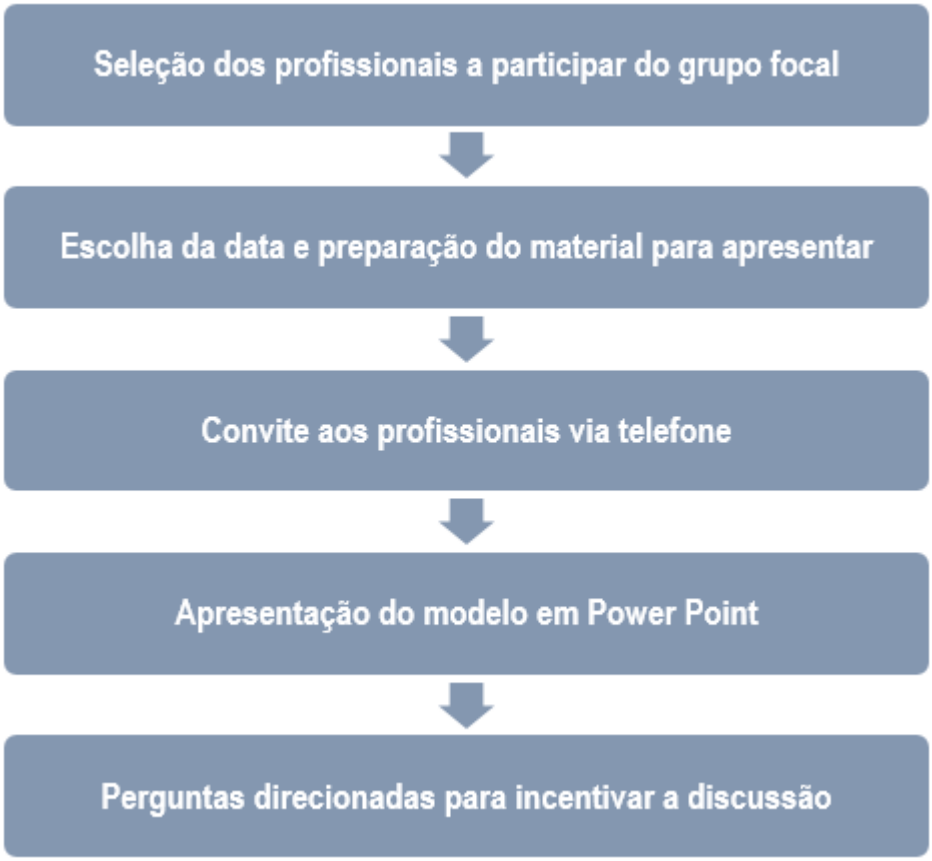
3.5.1 Coleta de dados para validação do modelo

Os dados utilizados na validação do modelo estão disponíveis no *software* R, no pacote de dados denominado *Groceries*. O pacote consiste em transações de um supermercado, o qual não é mencionado no mesmo.

Estes dados foram utilizados a fim de validar o modelo estruturalmente, de modo a corrigir possíveis inconsistências. Para isso, buscou-se a documentação do *dataset* a fim de verificar o que o mesmo disponibiliza, o qual apresenta como resultado as regras de associação. Dessa forma, uma vez replicado o resultado esperado, o modelo estava validado.

Além da coleta das transações, será realizado um grupo focal, que objetiva promover um debate sobre um tema específico e intensifica o acesso às informações a respeito de um fenômeno, possibilitando gerar novas concepções ou analisar profundamente uma problematização. (BACKES, *et al*., 2011). A coleta de dados por grupo focal foi realizada no supermercado, com três participantes que atuam diretamente na elaboração das promoções: os dois sócios e a gerente do estabelecimento. Os passos para a realização da coleta de dados via grupo focal podem ser visualizados na Figura 7.

Figura 7 – Etapas do grupo focal



Fonte: Elaborado pela autora.

Os profissionais foram selecionados de acordo com o papel que exercem na empresa e a participação dos mesmos na escolha dos itens dispostos nas promoções. Neste último quesito, os sócios atuam diretamente, visto que avaliam as vendas dos itens separadamente e selecionam os mesmos de acordo com critérios internos. A gerente, por sua vez, atua da mesma forma, mas avalia também a disposição dos itens na loja, montando pontas de gôndolas e melhor posicionando os produtos, garantindo seu destaque perante aqueles não promocionais.

A escolha da data foi realizada de acordo com a disponibilidade dos participantes, visto que o encontro foi nas dependências do supermercado. A preparação do material foi efetuada buscando expor de forma mais clara possível os conceitos da mineração de dados e análise de cesta de compras, para que os profissionais pudessem avaliar a ferramenta.

A apresentação do modelo ao grupo focal foi realizada em Power Point, com os detalhes do mesmo. Em seguida, foram feitas perguntas direcionadas com o objetivo de incentivar a discussão sobre o modelo:

I. Sobre o tema:
   - O tema “Data Mining” é relevante para o varejo?
   - A ferramenta se mostra útil, considerando a competitividade atual, como diferenciação entre os concorrentes?

II. Sobre a aplicação:
   - O modelo apresenta vantagens para o supermercado?
   - Acredita que é possível obter maior assertividade na escolha dos itens para promoções?
   - A utilização do modelo poderia ser visível na receita do supermercado?

III. Sobre o *software*:
   - Na sua opinião, o *software* é de fácil manuseio?
   - Há interesse em utilizá-lo?

A aplicação e a análise dos resultados oriundos desta etapa foram descritas no subitem 4.2.2 Validação do modelo por meio do Grupo Focal.

3.6 ANÁLISE DE DADOS

A etapa de análise de dados objetiva dar sentido ao conjunto de informações abordadas. (DRESCH; LACERDA; ANTUNES JR, 2015). Os dados a serem analisados neste trabalho são: resultados da validação do modelo, as regras de associação oriundas das transações do supermercado e, a partir do resultado, a montagem dos combos.

O pacote de dados utilizado na avaliação experimental foi submetido à mesma análise a qual os dados reais do problema. A eles foi aplicado o algoritmo *Apriori* e suas regras, com o intuito de certificar que tal procedimento confere o resultado esperado: regras de associação. Nesta validação também foi possível verificar se o modelo apresentou algum erro ou se haveria a necessidade de alterar algum parâmetro.

O grupo focal foi conduzido pelo mediador, o qual foi responsável por fazer os apontamentos durante as discussões, observando as respostas das perguntas. Para análise dos apontamentos foi utilizada a técnica de análise de conteúdo, a qual é muito

utilizada em analises qualitativas. A técnica consiste em três fases: pré-exploração, a seleção das unidades de análise e tratamento dos dados e interpretação. (CAMPOS, 2004; SOUZA JÚNIOR; MELO; SANTIAGO, 2010).

Segundo Campos (2004), a fase de pré-exploração consiste em realizar uma leitura “flutuante” do material coletado para que o pesquisador consiga assimilar e visualizar pistas de forma não estruturada. Optou-se por gravar toda a reunião com o grupo focal, a fim de evitar perdas de apontamentos importantes durante a discussão. Assim sendo, este áudio foi analisado para cumprir a etapa de pré-exploração.

A fase seguinte, denominada seleção das unidades de análise, consiste em selecionar frases, sentenças, parágrafos de acordo com as questões de pesquisa que necessitam ser respondidas. (CAMPOS, 2004). Com o objetivo de encontrar as unidades de análise, nesta etapa foram selecionados os principais comentários gerados na discussão a respeito do tema. De acordo com Souza Júnior, Melo e Santiago (2010), para o tratamento dos dados e interpretação deve-se ressaltar os resultados oriundos da análise. Para isso, um quadro foi desenvolvido com os principais apontamentos, bem como os pontos de concordância entre os profissionais.

O arquivo com as transações do supermercado foi importado no *software* R para aplicação do modelo, objetivando a descoberta de itens frequentes, ou seja, aqueles itens que costumam ser vendidos juntos com maior frequência. Dessa forma, os comandos propostos para explorar as transações foram executados um a um e seus resultados apresentados.

Conforme Han, Kamber e Pei (2012), ao utilizar apenas suporte e confiança como medidas de interesse, não é possível filtrar apenas as regras interessantes. Para isso, recomenda-se utilizar uma medida de correlação, e neste trabalho foi utilizado o *lift*, medida que verifica as dependências entre os itens, conforme explicitado no item 2.3 ANÁLISE DA CESTA DE COMPRAS.

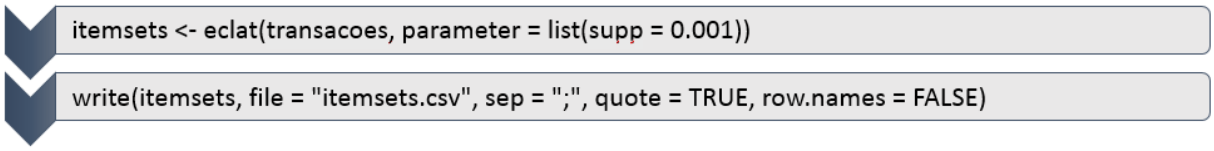
Ao executar o último comando, o qual tem a função de exportar os dados gerados para visualização, os mesmos foram filtrados de acordo com o *lift* apresentado, e listados em ordem decrescente, a fim de verificar os itens que apresentam maior dependência. Conforme exposto por Gonçalves (2005), as regras geradas que apresentam *lift* de valor um são independentes, ou seja, tais itens não apresentam relação entre si. Portanto estas regras foram desconsideradas.

Para a análise dos resultados foram considerados os itens que apresentaram maior frequência dentro da base de dados, bem como a sua relação com as regras

de associação encontradas. Avaliou-se também à qual segmento do supermercado as regras predominam e a variação do suporte entre elas.

Após a aplicação do modelo, os resultados foram analisados para atender a um dos objetivos específicos, a proposição de combos de vendas e, para a montagem dos mesmos, foram considerados os itens mais frequentes da base de dados. Contrastando com estes, foram escolhidos outros itens de menor frequência para compor o combo e, para isso, foram utilizados os dois comandos apresentados na Figura 8.

Figura 8 – Comando *itemsets*



Fonte: Elaborado pela autora.

O comando *itemsets* objetiva gerar conjuntos de itens conforme os parâmetros determinados e, nesta etapa buscou-se identificar os itens com baixa frequência de vendas, portanto, a medida de interesse utilizada foi o suporte, visto que o mesmo tem o propósito de apresentar a frequência dos itens dentro da base de dados. Em sequência o comando *write* foi utilizado para exportar o conjunto em Excel, onde os dados foram filtrados em ordem crescente, ou seja, do menor ao maior suporte. Em seguida, foi utilizada a função ALEATORIOENTRE do Excel para determinar os itens a compor o combo. Os detalhes deste procedimento estão explanados no capítulo de número cinco.

Expostos os procedimentos metodológicos, no próximo capítulo é apresentada a aplicação do modelo proposto.

4 DESENVOLVIMENTO DO MODELO

Este capítulo apresenta o contexto da empresa a qual o modelo foi aplicado, a execução da validação com dados de teste, os apontamentos do grupo focal e o detalhamento da utilização do modelo.

4.1 APRESENTAÇÃO DA EMPRESA E CONTEXTO DO PROBLEMA

As atividades do Super Cidade Nova foram iniciadas no ano de 1997, na cidade de Ivoti, no segmento de comercio alimentício, oferecendo até então alimentos não perecíveis, limpeza e pequenos utensílios. Ao longo dos anos, a estrutura do estabelecimento passou por diversas reformas e ampliações, sendo possível o aumento do *mix* de produtos.

Atualmente o estabelecimento conta com diversas seções de produtos, sendo elas: alimentos, açougue, padaria, frios e congelados, hortifrutigranjeiros, bebidas, utilidades, entre outras. O supermercado conta com quatro *checkouts* e 20 funcionários aproximadamente.

Todas as atividades relacionadas à venda, compra e estoque de produtos é realizada por meio de um *software*, permitindo que os administradores tenham o controle sobre os mesmos. Por meio deste controle é que são estabelecidas as promoções, porém não existe um procedimento padrão para isso.

Observa-se que não existe uma análise profunda sobre os hábitos dos consumidores ou quais produtos costumam ser vendidos juntos. Portanto, a proposta de aplicar a ferramenta *Data Mining* neste banco de dados auxilia o administrador a tomar tais decisões de forma mais assertiva, visto que se propõem a avaliar as vendas de forma integrada e entender as relações entre os itens.

4.2 VALIDAÇÃO DO MODELO

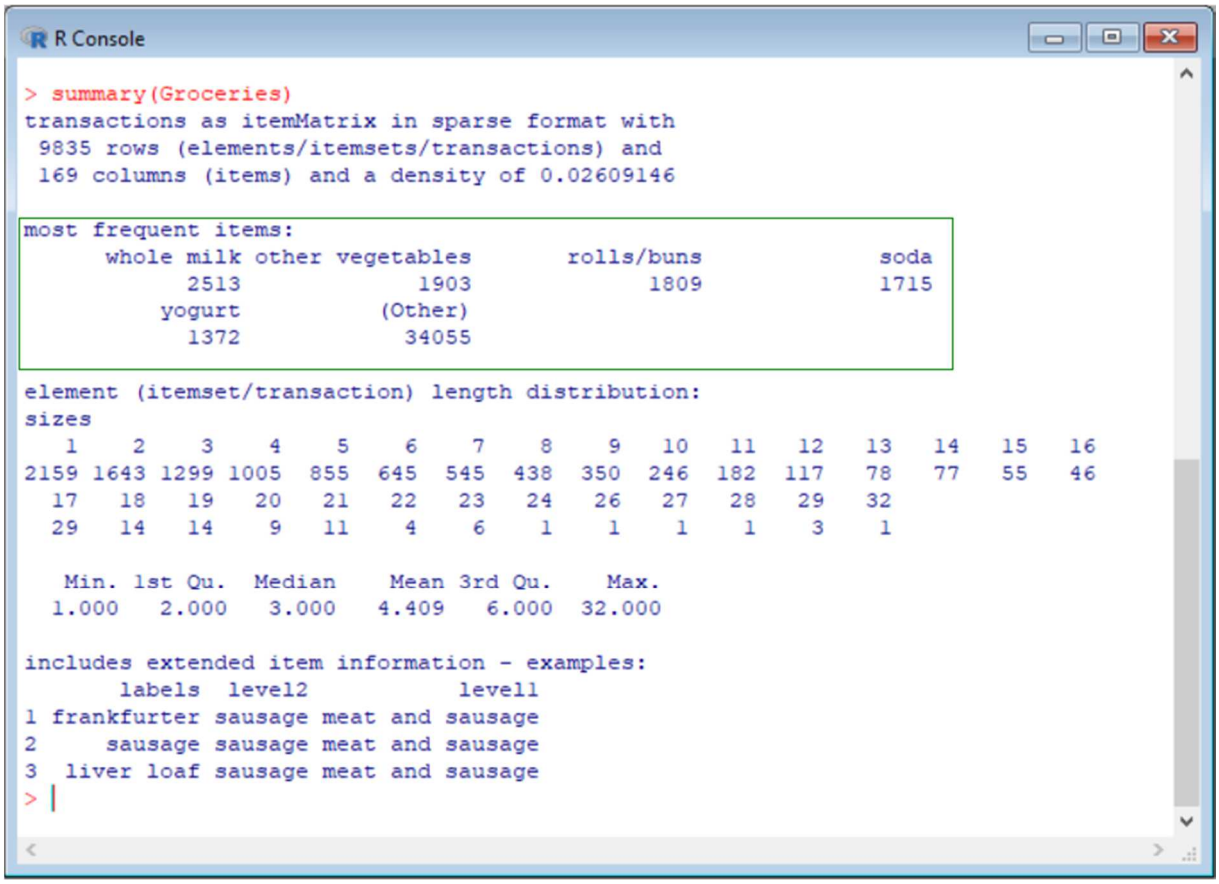
O modelo proposto neste trabalho passou por duas formas distintas de validação. Inicialmente foi validado quanto à possíveis erros de modelagem estrutural, bem como erros de sintaxe na utilização do R. Posteriormente, o modelo foi apresentado ao grupo focal, com a intenção de verificar a sua aplicabilidade na empresa. Tais validações estão explanadas nesta seção.

4.2.1 Validação do modelo com dados de teste

A fim de validar o modelo quanto à sua estrutura, o mesmo foi utilizado com um pacote de dados de transações disponíveis no *software* R, o pacote *Groceries*. Para isso, ao pacote foram aplicados os comandos propostos na seção 3.4 CONSTRUÇÃO DO MODELO.

Durante toda a etapa não foram reportados erros de sintaxe ou *bugs* pelo *software*, afirmando que os comandos estão corretos, não havendo a necessidade de correções antes de utilizar os modelos com os dados reais. Com a aplicação dos primeiros comandos é possível encontrar os itens mais frequentes da base de dados analisada. Na Figura 9 pode ser visualizado que estes itens correspondem à: *whole milk*, contabilizado 2.513 vezes, ou seja, ele está presente em 2.513 transações, sendo o item mais frequente, seguido de *other vegetables*, contabilizado em 1.903, e os demais itens. O número 34.055 que representa *Other,* diz respeito a quantidade de vezes que todos os outros itens aparecem na base de dados.

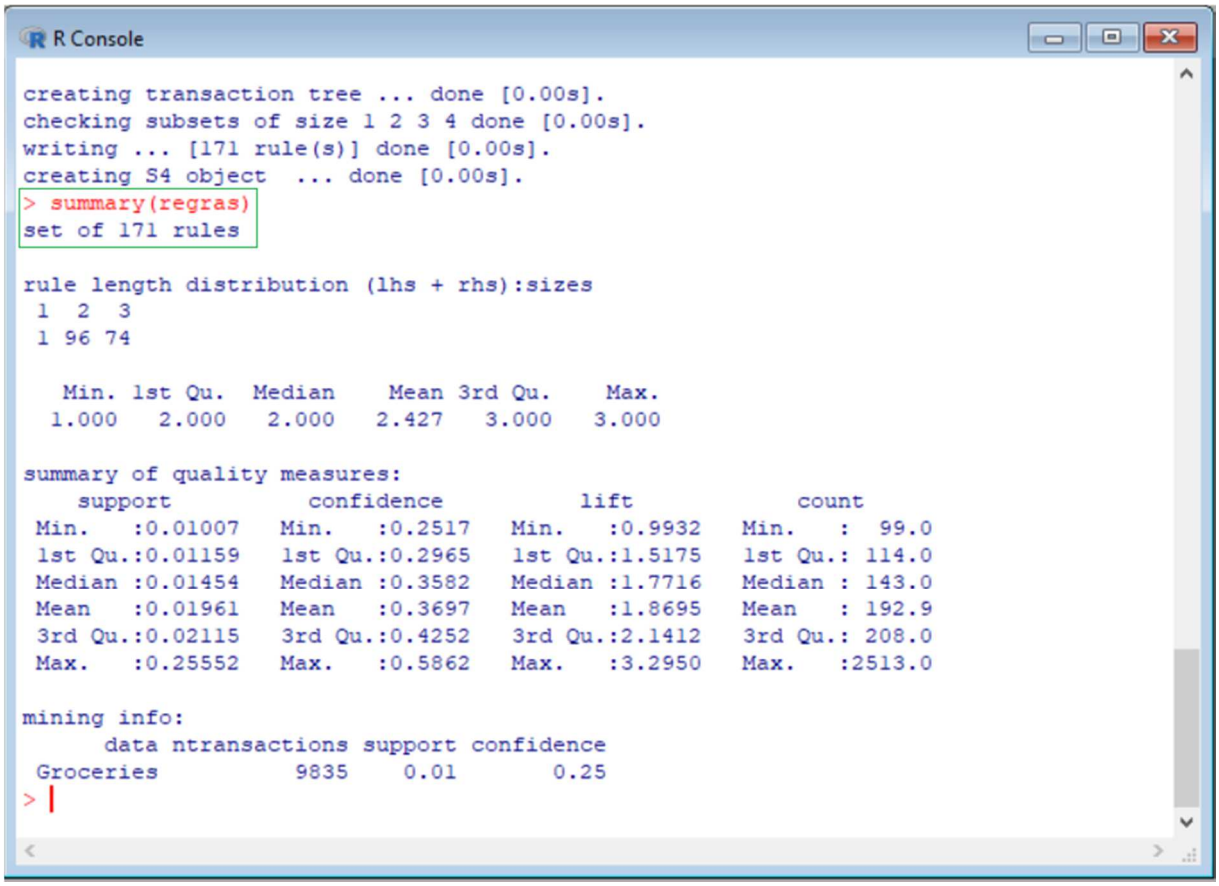
Figura 9 – Comando *summary(Groceries)*



Fonte: Elaborado pela autora.

Após aplicar o comando *apriori()* com suas medidas de interesse à base de dados, ou seja, valores de suporte e confiança, aplicou-se o comando *summary()* novamente para verificar a quantidade de regras geradas. Conforme a Figura 10, o algoritmo gerou 171 regras de associação, onde o item antecedente e o consequente da regra apresentam alguma dependência.

Figura 10 – Comando *summary(regras)*



Fonte: Elaborado pela autora.

Para analisar a dependência entre os itens, as regras foram exportadas em arquivo Excel, onde foram filtradas de maneira decrescente para apresentar o maior *lift*. Na Tabela 4 é apresentada uma amostra das regras de associação descobertas com a aplicação do *Apriori*.

Tabela 4 – Resumo das Regras de Associação com dados de teste

#	Regras	Suporte	Confiança	*Lift*
1	{citrus fruit,other vegetables} => {root vegetables}	0,010371124	0,35915493	3,295045459
2	{other vegetables,tropical fruit} => {root vegetables}	0,012302999	0,342776204	3,144779819
3	{beef} => {root vegetables}	0,017386884	0,331395349	3,040366843
4	{citrus fruit,root vegetables} => {other vegetables}	0,010371124	0,586206897	3,029608422
5	{root vegetables,tropical fruit} => {other vegetables}	0,012302999	0,584541063	3,020999134
6	{other vegetables,whole milk} => {root vegetables}	0,023182511	0,309782609	2,842082049
7	{curd,whole milk} => {yogurt}	0,01006609	0,385214008	2,761355515
8	{other vegetables,yogurt} => {root vegetables}	0,012913066	0,297423888	2,7286977
9	{other vegetables,yogurt} => {tropical fruit}	0,012302999	0,283372365	2,700549625
10	{other vegetables,rolls/buns} => {root vegetables}	0,012201322	0,286396181	2,627524668
11	{tropical fruit,whole milk} => {root vegetables}	0,011997966	0,283653846	2,602365277
12	{rolls/buns,root vegetables} => {other vegetables}	0,012201322	0,50209205	2,594889813
13	{root vegetables,yogurt} => {other vegetables}	0,012913066	0,5	2,584077772
14	{whole milk,yogurt} => {tropical fruit}	0,015149975	0,270417423	2,577088521
15	{pip fruit} => {tropical fruit}	0,020437214	0,27016129	2,574647568
16	{tropical fruit,whole milk} => {yogurt}	0,015149975	0,358173077	2,567516189
17	{whipped/sour cream,yogurt} => {other vegetables}	0,010167768	0,490196078	2,53340958
18	{other vegetables,whipped/sour cream} => {yogurt}	0,010167768	0,352112676	2,524073009
19	{other vegetables,root vegetables} => {tropical fruit}	0,012302999	0,259656652	2,474537961
20	{other vegetables,tropical fruit} => {yogurt}	0,012302999	0,342776204	2,457145748
21	{root vegetables,whole milk} => {other vegetables}	0,023182511	0,474012474	2,449770195
22	{whipped/sour cream,whole milk} => {yogurt}	0,010879512	0,337539432	2,419606644
23	{citrus fruit,whole milk} => {yogurt}	0,010269446	0,336666667	2,41335034
24	{whole milk,yogurt} => {root vegetables}	0,014539908	0,259528131	2,381025341
25	{onions} => {other vegetables}	0,014234875	0,459016393	2,372268119

Fonte: Elaborado pela autora.

Na segunda coluna estão as regras geradas e cada linha apresenta uma delas, com o item antecedente e o consequente. A terceira, quarta e quinta coluna apresentam respectivamente as medidas de interesse suporte, confiança e *lift* de cada uma das regras.

Ao finalizar esta etapa de validação verificou-se que o modelo não apresentou nenhum erro e, ao executar o último comando, entregou o resultado esperado, as regras de associação, estando apto para utilização com as transações do supermercado.

4.2.2 Validação do modelo por meio do Grupo Focal

O grupo focal foi realizado com o intuito de validar o modelo quanto à sua aplicabilidade no supermercado. Conforme já expressado anteriormente, o intuito da utilização do *Data Mining* neste trabalho é demonstrar de que forma a ferramenta pode contribuir na tomada de decisão quando se trata da escolha de itens para compor promoções. Sendo assim, foram convidadas a participar do grupo focal as pessoas que de alguma forma estão envolvidas nesta operação.

O grupo se encontrou nas dependências do supermercado, no dia 10 de setembro de 2018, às 20h. Inicialmente a mediadora fez a apresentação do conteúdo em Power Point aos participantes, onde foi apresentado o conceito de *Data Mining* e sua utilidade, evidenciando o *Market Basket Analysis*. Após a teoria foram apresentados o *software* e o modelo, o qual foi aplicado aos dados do R para mostrar aos participantes como trabalhar no *software* e como os resultados são gerados e analisados.

Ao visualizar as regras, os participantes fizeram diferentes comentários interessantes sobre como montar as promoções. Um deles foi o seguinte: “sabendo que dois itens são vendidos juntos frequentemente, eles ofertariam apenas um deles, podendo até aumentar o preço do outro item, pois os dois teriam grande probabilidade de serem vendidos juntos”. Reforçaram também que, provavelmente um item se destacaria, e com ele poderiam aparecer outros 100 itens associados, cada um em uma regra diferente. Para isso, sugeriram trabalhar em cada um destes itens em meses diferentes.

Os participantes também se manifestaram quanto ao número de regras geradas, afirmando que para eles seria interessante analisar as 50 primeiras regras,

após filtrar os dados de acordo com o maior *lift*. Quanto à montagem dos combos, sugeriram associar um item que é vendido frequentemente com um item de baixa venda.

Quadro 4 – Análise do Grupo Focal

Abordagem	Resultados oriundos da análise
Tema	Percebem que a ferramenta é relevante e aplicável ao varejo, tanto que varejistas de maior porte já estão aderindo, inclusive a rede Smart, da qual o supermercado faz parte. Porém alertam para a dificuldade de utilizar a ferramenta em um estabelecimento menor, pois alegam que seria necessário ter um funcionário responsável para utilizar a ferramenta e analisar os dados gerados, gerando maior custo operacional.
Aplicação	Concordam quanto à assertividade na escolha das promoções, pois verificam que é possível identificar os itens que vendem frequentemente juntos, não sendo necessário ofertá-los juntos. E analisando desta forma, consequentemente poderiam verificar um aumento na receita.
Software	Os participantes alegam interesse em utilizar o *software*, visto que o mesmo é gratuito, mas concordam que a ferramenta não é de fácil manuseio. No entanto há interesse em utilizar o modelo proposto, já que os comandos não seriam alterados.

Fonte: Elaborado pela autora.

De maneira geral os participantes concordam que a ferramenta apresenta benefícios para o varejo, evidenciando que a mesma pode trazer resultados para a receita do estabelecimento, pois impactaria diretamente na escolha de itens para promoções, ou então na montagem de pontas de gôndolas. Destacam ainda que seria possível elevar as vendas de um item que não é vendido frequentemente, associando este à um item frequente. Os resultados da análise podem ser visualizados no Quadro 4 – Análise do Grupo Focal.

4.3 EXTRAÇÃO, TRATAMENTO E CARREGAMENTO DOS DADOS

Devido a condição do *software* R, os dados foram solicitados ao supermercado em arquivo Excel e, em razão da quantidade de informações, foram originados três arquivos com as transações. Para a etapa de tratamento, inicialmente foi realizada a limpeza em cada um dos arquivos separadamente, no qual foram removidas as informações desnecessárias para o estudo, conforme já evidenciado. A Figura 11

apresenta os dados em seu formato original, enquanto que a Figura 12 apresenta os dados tratados.

Figura 11 – Dados originais

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	COMERCIAL ALIMENTICIA KERCKHOFF LTDA					02.171.219/0001-89									
2															
3	Relatorio Cupons detalhe por periodo														
4	Data Inicial: 01/07/2017 - Data Final: 31/10/2017														
5															
6															
7	DATA		CAIXA	CUPOM	COD	PRODUTO					QTD		PRECO	TOTAL	
8	01/07/2017		2	214372	7892840266943	SALGADINHO ELMA DORITOS 86G MYSTERYPRETO					1		6	6	
9	Subtotal	6													
10															
11	01/07/2017		2	214373	7890300168585	GRANULADO FRITZ&FRIDA 80G					1		2,69	2,69	
12	01/07/2017		2	214373	7891000100103	LEITE COND.MOÇA 395GR LATA					1		6,99	6,99	
13	01/07/2017		2	214373	7891025107392	BEBIDA LACTEA DANONINHO 80G MORANGOPRECO S					1		1,49	1,49	
14	01/07/2017		2	214373	7891042091384	MARG.BECEL 250GR S/SAL					1		4,99	4,99	
15	01/07/2017		2	214373	7896022011321	BOLACHA MARIA ISABELA 400G					1		5,99	5,99	
16	01/07/2017		2	214373	7896079500151	LEITE L ELEGE INTEGRAL					1		2,49	2,49	
17	01/07/2017		2	214373	7,89648E+12	ALGODAO SMART 25G HIDROFILO 201893					1		2,29	2,29	
18	Subtotal	26,93													
19															
20	01/07/2017		2	214375	201	CACETINHO					1		7,99	3,14806	
21	Subtotal	7,99													
22															
23	01/07/2017		2	214376	27	GADO 2° CARNE MOIDA II					1		15,5	10,912	
24	01/07/2017		2	214376	219	NATA AGRANEL					1		17,5	3,325	
25	01/07/2017		2	214376	247	BOLACHA CASEIRA NATAL					2		6,5	13	
26	01/07/2017		2	214376	261	CACETINHO 30G ADORMECIDO					1		0,2	2	
27	01/07/2017		2	214376	279	PAO CASEIRO LEITE 380G					1		3,95	3,95	

Fonte: Super Cidade Nova

Figura 12 – Dados tratados

	B
1	SALGADINHO ELMA DORITOS 86G MYSTERYPRETO
2	GRANULADO FRITZ&FRIDA 80G,LEITE COND.MOÇA 395GR LATA,BEBIDA LACTEA DANONINHO 80G MORANGOPRECO SUGERIDO,MARG.BECEL 250GR S/SAL,BOLACHA MARIA ISABELA 400G,LEITE L ELEGE INTEGRAL,ALGODAO SMART 25G HIDROFILO 20189
3	CACETINHO
4	GADO 2° CARNE MOIDA II,NATA AGRANEL,BOLACHA CASEIRA NATAL,CACETINHO 30G ADORMECIDO,PAO CASEIRO LEITE 380G,PRESUNTO KG SADIA FATIADO,HF BATATA BRANCA,HF BERGAMOTA COMUM,QUEIJO K MUSSARELA SADIA,QUEIJO K MUSSARE
5	JORNAL N.H SEMANAL
6	CHOC.GRANULADO SADORO 75G COLORIDO,LEITE COND.TIROL 395GR TRADICIONAL,BOLO ORQUIDEA 400GR CHOCOLATE,BOLO ORQUIDEA 400GR BAUNILHA
7	CACETINHO,MORTADELA KG S/GORDURA,ACHOC.PO NESCAU 800G SACHE,OLEO DE SOJA PRIMOR 900ML PET,MASSA ISABELA BOM GOSTO 500GRPARAFUSO,WAFER ISABELA 145GR CHOCOLATE BRANCO
8	PASTEL GRANDE CASEIRO
9	CERV.BRAHMA LATAO 473ML
10	PAO CASEIRO ADORMECIDO 2,00,PAO DE QUEIJO DE ONTEM
11	LINGUICA CALABRESA,HF VERDES,FEIJAO PRETO FRITZ&FRIDA 1KG T.1
12	EXTRATO T ELEFANTE 340G LATA NOVA
13	CACETINHO,PAO DE QUEIJO,HF BANANA CATARINA,HF CEBOLA,HF REPOLHO UNIDADE,HF ALFACE CRESPA,CR.D COLGATE 70G LUMINOUS WHITEESMALTE BRILHANTE,DET.YPE 500ML NATURAL CLEAR,SABAO PO GIRANDO SOL 1KG DIA DIA SACHEAZUL FL
14	HF BROCOLIS BANDEJA,CHOC.B NEUGEBAUER 120G AO LEITE
15	WAFER PARATI 115GRCHOCOLATE&BAUNILHA,RECH.ISABELA 160G TORTINHAS CHOCOLATE
16	FR CORACAO FR FRANGO,HF PIMENTAO AMARELO&VERMELHO,CACHORRINHO MINI,QUEIJO K LANCHE LAC LELO FATIADO,QUEIJO K LANCHE LAC LELO FATIADO,QUEIJO K LANCHE LAC LELO FATIADO,QUEIJO K LANCHE LAC LELO FATIADO,PAO DE QUEIJC
17	GADO 2° CARNE MOIDA II,PORCO COSTELA DEF,CHARQUE,MONDONGO COZIDO,BOLACHA DOCE MAGIA 250G NATAL,BOLACHA DOCE MAGIA 250G BRIGADEIRO
18	GADO 2° AGULHA C/OSSO
19	FR COXA SOBRE COXA,FR COXA SOBRE COXA,CACETINHO,HF TOMATE LONGA VIDA,HF BATATA BRANCA,BALA UNIDADE,BALA UNIDADE
20	GADO 2° PALETA C/OSSO,FR COXA SOBRE COXA,GADO COSTELINHA DIANTEIRO,NATA PIA 300G
21	FR FRANGO CORTADO,HF PIMENTAO,ACUCAR UNIAO 1KG ORGANICO,REQUEIJAO LAC LELO 200G ZERO LACTOSE
22	CERV.POLAR 473ML LATAO
23	GADO 2° AGULHA C/OSSO,GADO 2° AGULHA C/OSSO,PAO CASEIRO ADORMECIDO 2,00,HF MORANGA VERDE CABUTIA,MOLHO CAJAMAR 340G TRADICIONAL,LEITE L ELEGE INTEGRAL,MARG.QUALY CREMOSA 500GR C/SAL,DOCE B.PRINCIPIO 400GR MAÇ
24	PAO DE XIS,PAO DE XIS,PAO DE XIS,PAO DE XIS,PAO DE XIS,PAO DE XIS
25	GADO 1° PATINHO,HF KIWI,HF CEBOLA,PLETS&BUBALOO,HF TOMATE ITALIANO,BATATA P YOKI 120G EXTRA FINA PREMIUMHOT DOG,MASSA ISABELA BOM GOSTO 500GRMACARRAO,REQUEIJAO PIA 200G TRADICIONAL,REQUEIJAO PIA 200G LIHGT,BATAT
26	JARRA VIDRO CHA 750ML,PORTA FILTRO MELITTA 103,CERA BRILHO FACIL 750ML INCOLOR,FILTRO CAFE MOKA 103 P/CAFE
27	GADO 1° COXAO DE DENTRO,GADO CARNE DE GADO MISTA,HF BERGAMOTA COMUM,PAO CASEIRO 400GR AIPIM,BOLACHA DOCE MAGIA 250G MANTEIGA,PE DE MOLEQUE DACOLONIA 200G,BALA FLORESTAL 100G BRAZILIAN COFFEE

Fonte: Elaborado pela autora.

Após excluídas as informações desnecessárias dos arquivos separadamente, os dados foram reunidos em um único arquivo, salvo em formato CSV para posterior carregamento no *software*.

4.4 UTILIZAÇÃO DO MODELO

Nesta etapa foram seguidos os passos propostos na seção 3.4 CONSTRUÇÃO DO MODELO, a fim de alcançar os objetivos propostos no capítulo de número um desta monografia.

O primeiro comando foi utilizado para instalar os pacotes do *software,* já o segundo comando foi utilizado para buscar o pacote *arules*, que conta com a estrutura necessária para analisar conjuntos de dados, bem como aplicar o algoritmo *apriori*. E em seguida, o terceiro comando foi utilizado para importar os dados no R. Os três primeiros comandos podem ser visualizados na Figura 13.

Figura 13 – Comandos de número um, dois e três

Fonte: Elaborado pela autora.

Após a importação dos dados do supermercado foi utilizado o quarto comando, que apresenta uma visão geral do conteúdo.

Figura 14 – Comando de número quatro

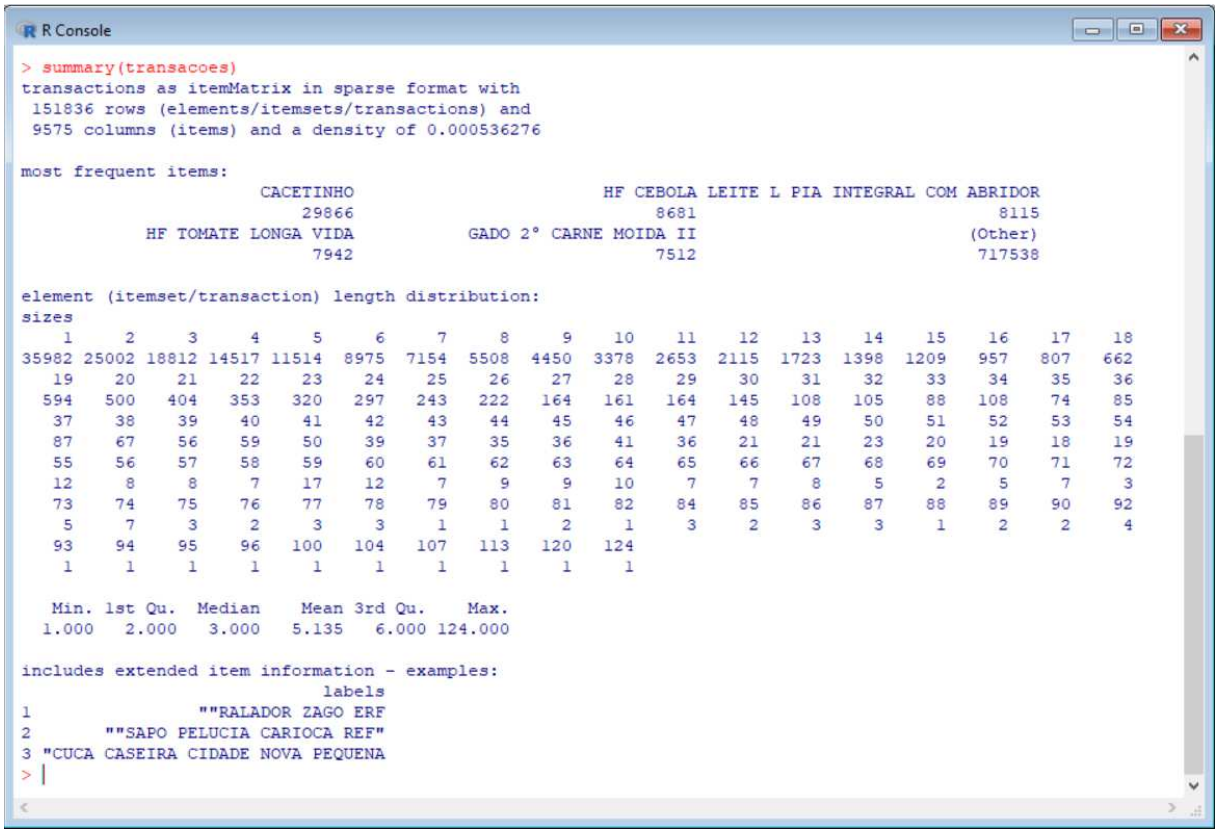
Fonte: Elaborado pela autora.

Algumas informações devem ser destacadas na Figura 15, que representa os resultados obtidos após aplicar o quarto comando. É possível visualizar que na base de dados existe um total de 151.836 transações e 9.575 itens, ou seja, durante o período de um ano analisado, foram vendidos esta quantidade de itens diferentes.

Este comando também apresenta os itens mais frequentes, isto é, os itens mais vendidos da base de dados. O item em destaque é o cacetinho, ou pão francês, aparecendo 29.866 vezes nas transações, ou seja, este item está presente em 20% das transações analisadas. Em segundo lugar está o item cebola, contabilizado em 8.681 transações, seguido do item leite piá integral com abridor, que está presente em

8.115 transações. E os dois últimos itens destacados são tomate e gado 2ª carne moída II, contabilizados em 7.942 e 7.512 transações, respectivamente.

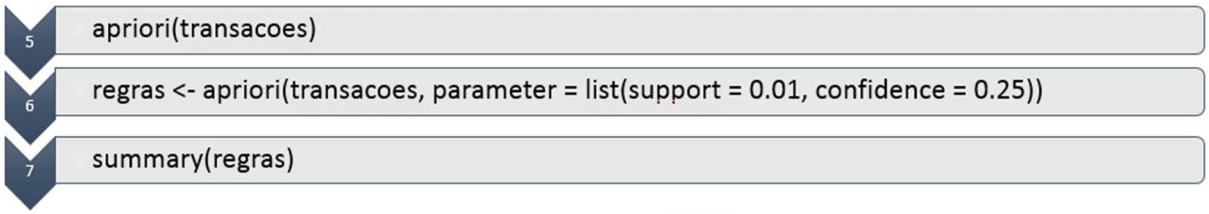
Figura 15 – Comando *summary(transacoes)*



Fonte: Elaborado pela autora.

O próximo comando, de número cinco, especifica o uso do algoritmo aos dados, porém não especifica as medidas de interesse. Por isso, o sexto comando foi utilizado com os valores mínimos de suporte e confiança estabelecidos. Tais medidas são utilizadas para gerar as regras e, segundo Han, Kamber e Pei (2012), quanto maiores estes valores, maior a qualidade das regras geradas, eliminando regras desinteressantes e que não apresentam correlação entre os itens.

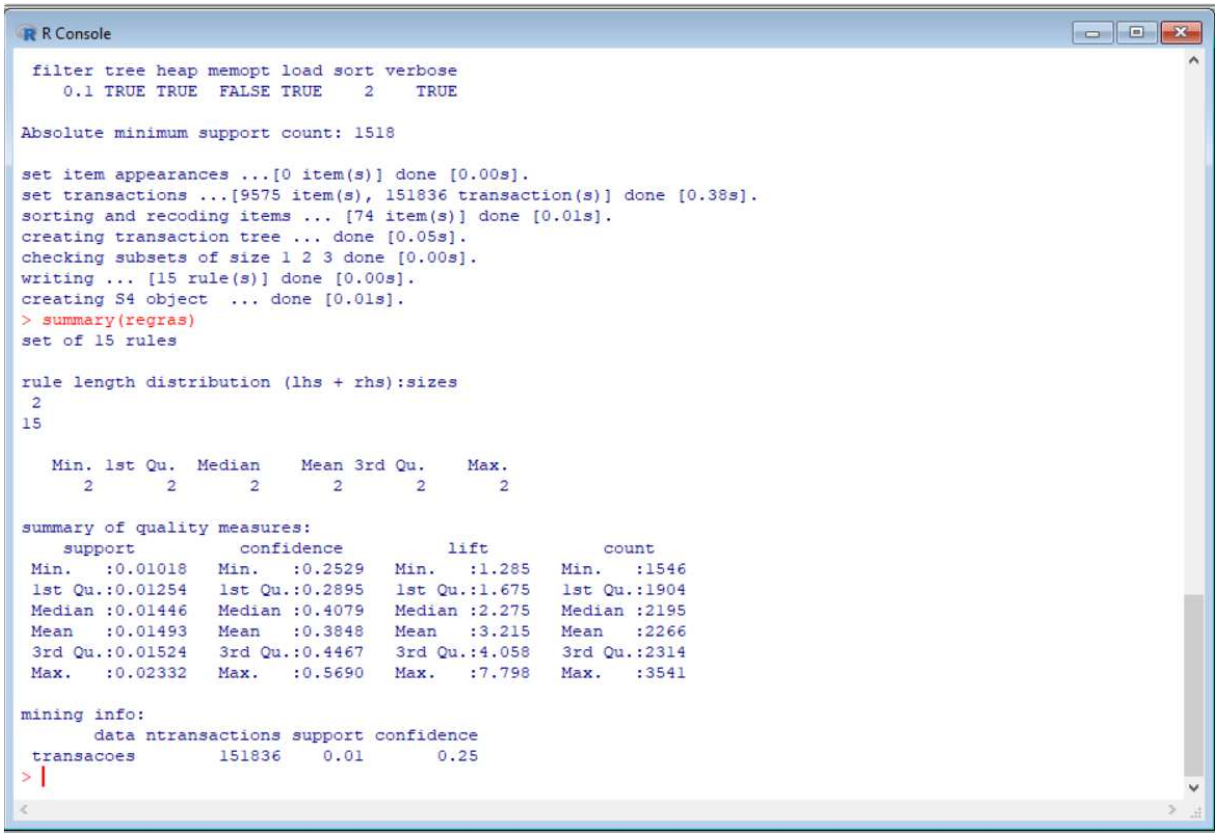
Figura 16 – Comandos de número cinco, seis e sete



Fonte: Elaborado pela autora.

Conforme apresentado no item 2.3.1 Conjuntos de Itens Frequentes, o suporte de uma regra representa a frequência dos itens face o total de transações. Já a confiança mede a probabilidade de ocorrer B, dado a ocorrência de A. Neste trabalho o suporte mínimo foi definido em 1% e a confiança mínima foi definida em 25%, valores similares ao utilizado por Anselmo (2017) em sua pesquisa. Na Figura 17 podemos verificar o resultado da aplicação do sétimo comando, o qual apresenta a quantidade de regras geradas do comando anterior. Neste caso, foram geradas quinze regras de associação e cada uma delas apresenta dois itens, ou seja, um item antecedente e um item consequente.

Figura 17 – Comando *summary(regras)*



Fonte: Elaborado pela autora.

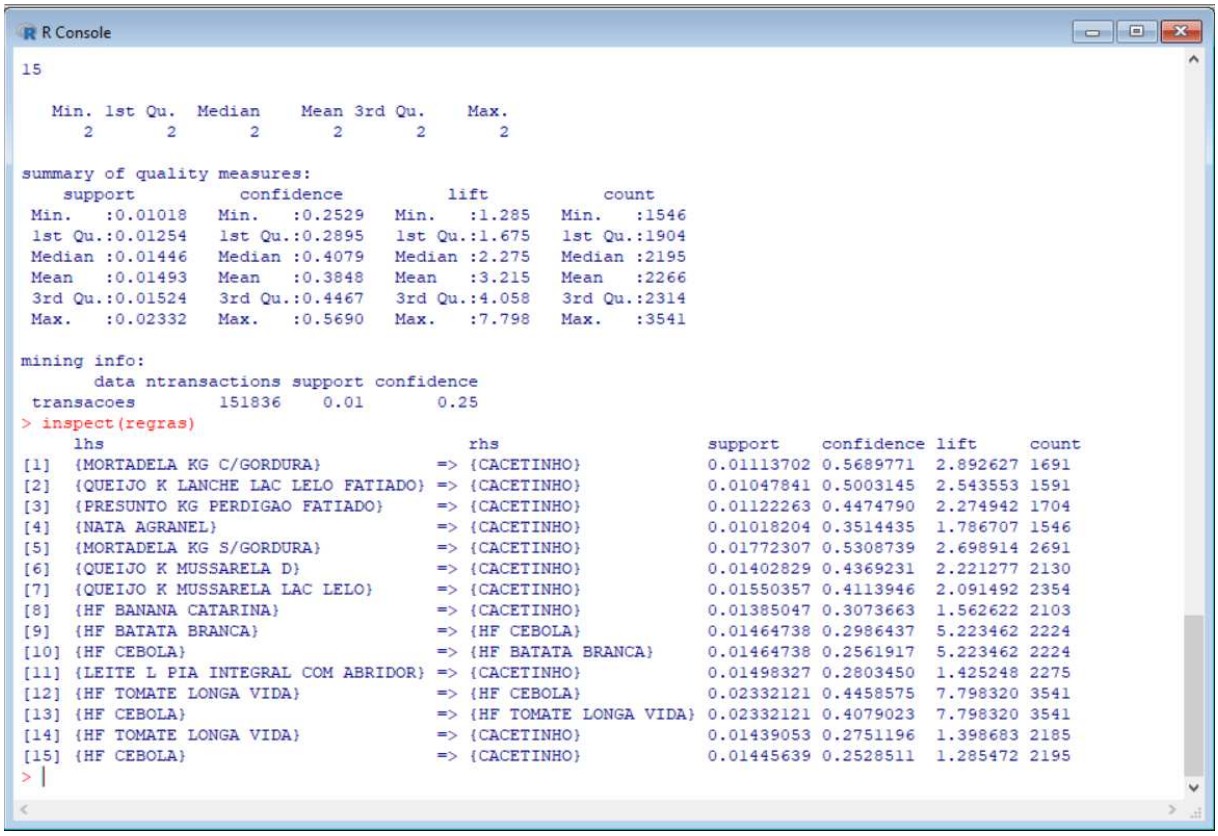
O comando de número oito foi utilizado para mostrar no ecrã do R as regras geradas anteriormente, o que pode ser visualizado na Figura 19.

Figura 18 – Comandos de número oito e nove

```
8  inspect(regras)
9  write(regras, file = "associacoes.csv", sep = ";", dec = ",", row.names = FALSE)
```

Fonte: Elaborado pela autora.

Figura 19 – Comando *inspect(regras)*



Fonte: Elaborado pela autora.

Por fim, o nono comando foi utilizado para exportar as regras em Excel e analisá-las. Seguindo o método proposto, as regras foram filtradas em ordem decrescente da medida de interesse *lift* e serão detalhadas no próximo capítulo.

5 ANÁLISE DOS RESULTADOS

Este capítulo apresenta as regras de associação encontradas por meio da aplicação do modelo às transações do supermercado e, por fim, apresenta os combos de vendas.

5.1 ANÁLISE DAS REGRAS DE ASSOCIAÇÃO

Conforme a Tabela 5 – Regras de associação, as regras foram filtradas de acordo com o maior *lift* para priorizar as regras de acordo com seu maior interesse. Dentre as quinze regras geradas não foram encontrados valores de *lift* = 1, onde os itens não apresentam dependência entre si, ou valores de *lift* < 1, cujos itens representam dependência negativa. Apenas foram encontrados valores de *lift* > 1, apresentando que nas regras os itens têm dependência positiva.

Os resultados apresentam que a associação com maior *lift* (7,798319814) é entre os itens tomate e cebola, apontando que os itens estão presentes em aproximadamente 2% do total de transações, referindo-se ao suporte. Já a confiança indica que, a cada 100 transações contendo o produto tomate, é possível que em 45 delas ocorra também a venda do item cebola. Analisando a regra inversa, regra número dois, verifica-se que os valores das medidas de interesse permanecem praticamente iguais, salvo a exceção da confiança, onde o item tomate está presente em 40% das transações que contém o item cebola.

Conforme apresentado anteriormente, o item mais frequente é o cacetinho, e o mesmo ocorre ao analisar as regras, visto que o mesmo está presente em onze de um total de quinze regras. Os itens tomate e cebola também foram apontados como itens frequentes e, ainda que ocorram em menor proporção que o item cacetinho, os mesmos ainda se mantêm como itens frequentes, uma vez que estão presentes em três e cinco das regras encontradas, respectivamente.

Ainda que o item leite piá integral com abridor tenha sido considerado como o terceiro item de maior frequência, ele está presente em apenas uma associação encontrada. Já o item gado 2ª carne moída II não está presente em nenhuma associação.

É possível verificar que o suporte das regras não apresenta significativa alteração do valor estabelecido no modelo. Mesmo nas regras mais frequentes, o

suporte é relativamente baixo, concentrando-se em 1% e 2%, o que reforça a ampla diversidade de produtos vendidos no estabelecimento.

Tabela 5 – Regras de associação

#	Rules	Suporte	Confiança	*Lift*
1	{HF TOMATE LONGA VIDA} => {HF CEBOLA}	0,023321215	0,445857467	7,798319814
2	{HF CEBOLA} => {HF TOMATE LONGA VIDA}	0,023321215	0,407902315	7,798319814
3	{HF BATATA BRANCA} => {HF CEBOLA}	0,014647383	0,298643749	5,223461847
4	{HF CEBOLA} => {HF BATATA BRANCA}	0,014647383	0,256191683	5,223461847
5	{MORTADELA KG C/GORDURA} => {CACETINHO}	0,011137016	0,56897712	2,892627401
6	{MORTADELA KG S/GORDURA} => {CACETINHO}	0,01772307	0,53087394	2,698914334
7	{QUEIJO K LANCHE LAC LELO FATIADO} => {CACETINHO}	0,010478411	0,500314465	2,543552775
8	{PRESUNTO KG PERDIGAO FATIADO} => {CACETINHO}	0,011222635	0,447478992	2,27494208
9	{QUEIJO K MUSSARELA D} => {CACETINHO}	0,014028294	0,436923077	2,22127678
10	{QUEIJO K MUSSARELA LAC LELO} => {CACETINHO}	0,01550357	0,411394617	2,091492436
11	{NATA AGRANEL} => {CACETINHO}	0,010182039	0,35144351	1,786706515
12	{HF BANANA CATARINA} => {CACETINHO}	0,01385047	0,307366267	1,562621862
13	{LEITE L PIA INTEGRAL COM ABRIDOR} => {CACETINHO}	0,014983271	0,280379591	1,425424079
14	{HF TOMATE LONGA VIDA} => {CACETINHO}	0,014390527	0,275119617	1,398682857
15	{HF CEBOLA} => {CACETINHO}	0,014456387	0,252851054	1,285471527

Fonte: Elaborado pela autora.

De modo geral, pode-se constatar que as regras encontras concentram-se principalmente em dois setores do supermercado: hortifrúti e padaria. Como mencionado no subitem 2.3 existem três formas de classificar as regras de associação: úteis, triviais e inexplicáveis. A primeira diz respeito aquelas regras utilizadas para tomar decisões e promover ações, enquanto que a segunda se refere às regras já conhecidas de alguma forma. Já a terceira se não haver uma explicação lógica, dificultando a tomada de decisão.

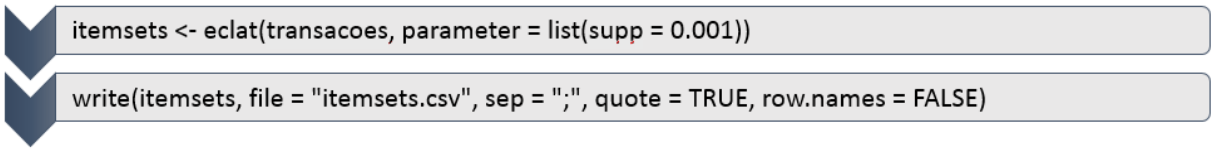
As regras geradas são classificadas em triviais, uma vez que se concentram em setores com maior procura pelo consumidor no decorrer da semana. Itens de padaria são consumidos diariamente, bem como itens de hortifrúti e, por isso, são percebidos pelo gestor como itens de alta rotatividade. O que pode ser comprovado nas regras de número 12, 14 e 15, as quais possuem itens de hortifrúti e padaria, é que os mesmos estão juntos nas compras dos consumidores.

As associações ainda podem ser classificadas em úteis, visto que é possível tomar decisões e ações sobre as mesmas. Conforme proposto pelo grupo focal, analisando as regras pode-se atentar para não coincidir as promoções destes itens: ao ofertar o item cacetinho, não poderiam ser ofertados os itens a ele associados. Isto porque aqui verificou-se que a probabilidade destes itens venderem juntos é alta, então não há a necessidade de ofertá-los juntos em um mesmo dia. Além disso, há a oportunidade de compensar os descontos das promoções nos itens que apresentam maior probabilidade de serem vendidos juntos e dessa forma, aumentar a receita.

5.2 COMBOS SUGERIDOS AO SUPERMERCADO

Cumprindo com um dos objetivos específicos desta monografia, o qual se propõem a elaborar combos de vendas por meio da análise das regras de associação, foram elaborados quatro combos. Para a sua montagem foram utilizados os quatro primeiros itens mais frequentes, incluindo cada um deles em um dos combos. A definição dos itens presentes nos combos foi realizada a partir da análise de um conjunto de itens, objetivando identificar itens que não se mostraram relevantes na base de dados. Este conjunto foi gerado no R a partir dos comandos na Figura 20.

Figura 20 – Comando para gerar o conjunto de itens



Fonte: Elaborado pela autora.

O conjunto de itens foi gerado por meio do suporte, que verifica a frequência dos itens dentro da base de dados e, para obter o maior número possível de conjunto de itens, optou-se por utilizar um suporte igual a 0,1%, ainda menor do que o suporte utilizado inicialmente para gerar as associações. É importante ressaltar que este suporte está apenas delimitando novamente a quantidade de conjuntos de itens, o que não exclui a possibilidade de haver outros itens cujo suporte seria ainda menor. O segundo comando, apresentado na Figura 20, exportou os dados em arquivo Excel para sua análise. Foram geradas 2.167 regras no total e a Tabela 6 apresenta uma amostra dos conjuntos gerados.

Tabela 6 – Amostra do conjunto de itens

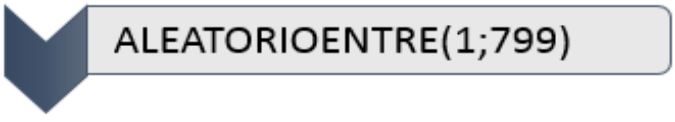
#	*Itemsets*	Suporte	*Count*
1	{GADO 1° COSTELA ESPECIAL,GADO 1° RIPA DA CHULETA}	0.00100108011275323	152
2	{FARINHA T ORQUIDEA 5KG,GADO 2° CARNE MOIDA II}	0.00100108011275323	152
3	{ACUCAR R ALTO ALEGRE 1KG,FARINHA T ORQUIDEA 5KG}	0.00100108011275323	152
4	{HF BATATA BRANCA,LEITE L TIROL 1LT INTEGRAL}	0.00100108011275323	152
5	{GADO 2° CARNE MOIDA II,HF VERDES}	0.00100108011275323	152
6	{CREME DE LEITE TIROL 200G TRADICIONAL,HF BATATA BRANCA}	0.00100108011275323	152
7	{FR FILE DE PEITO FRANGO SEARA 1KG BDJ,QUEIJO K MUSSARELA D}	0.00100108011275323	152
8	{CACETINHO,HF CEBOLA,MORTADELA KG C/GORDURA}	0.00100108011275323	152
9	{HF BATATA BRANCA,HF BROCOLIS BANDEJA,HF CENOURA}	0.00100108011275323	152
10	{HF BANANA CATARINA,OLEO DE SOJA PRIMOR 900ML PET}	0.00100108011275323	152
11	{HF BATATA BRANCA,HF CEBOLA,QUEIJO K MUSSARELA D}	0.00100108011275323	152
12	{FR COXA SOBRE COXA,HF BANANA CATARINA,HF TOMATE LONGA VIDA}	0.00100108011275323	152

Fonte: Elaborado pela autora.

A primeira coluna foi utilizada para enumerar as regras geradas, a segunda apresenta as associações, a terceira seu suporte e a quarta, a quantidade de vezes em que essas associações estavam presentes nas transações analisadas. Os dois itens da primeira associação da tabela apareceram juntos em 152 transações do total analisado, por exemplo.

Primeiramente os dados foram filtrados em ordem crescente de suporte, justamente para verificar os itens menos frequentes. A escolha dos itens a compor os combos se deu por meio da função do Excel, apresentada na Figura 21, onde também foi delimitado um intervalo de valores. Para isso, foi definido que as regras que apresentassem um valor de *count* acima de 800 não seriam utilizadas na função aleatória para compor o combo. Isto porque aqui se deu preferência para os itens menos frequentes, sendo assim, o número aleatório foi estabelecido entre 1 e 799.

Figura 21 – Função aleatória do Excel



Fonte: Elaborado pela autora.

A Tabela 7 apresenta as associações escolhidas para a montagem dos combos com a função aleatória do Excel, onde a primeira coluna representa o número aleatório gerado com a função. Foi necessário utilizar cinco vezes a função, pois uma das associações não pôde ser considerada no combo. A regra de número 724 não foi escolhida por apresentar dois itens perecíveis, sendo produtos que necessitam de refrigeração, não podendo ser expostos em algum outro ponto do supermercado.

Tabela 7 – Escolha dos conjuntos de itens para montar os combos

#	*Itemsets*	Suporte	*Count*
369	{ARROZ ROZCATO 1KG BRANCO}	0.00113938723359414	173
28	{TORRADA ISABELA 160G MULTIGRAOS}	0.00100766616612661	153
724	{CACETINHO,FR COXA SOBRE COXA,GADO 2° CARNE MOIDA II}	0.00133038278142206	202
587	{HF CENOURA,HF VERDES}	0.00124476408756817	189
126	{CACETINHO,SUCO PO TANG 25G LIMAO}	0.00104718248636687	159

Fonte: Elaborado pela autora.

Observando a tabela de conjuntos de itens nota-se que todos apresentam o suporte mínimo estabelecido e não estão presentes em muitas das transações, sendo que o *count* ficou abaixo de 200, afirmando que estas associações não são frequentes. Porém, o item cacetinho foi encontrado em uma delas, associado ao suco. Neste combo seria possível avaliar qual a frequência de dois itens considerados os mais vendidos, se associados à um item não frequente. Os combos sugeridos estão apresentados no quadro abaixo.

Quadro 5 – Combos sugeridos

Combo 1	
1kg	CEBOLA
1un	ARROZ ROZCATO 1KG BRANCO
Combo 2	
1un	LEITE L PIA INTEGRAL COM ABRIDOR
1un	TORRADA ISABELA 160G MULTIGRAOS
Combo 3	
1kg	CACETINHO
1kg	CENOURA
1un	HF VERDES
Combo 4	
1kg	TOMATE
1kg	CACETINHO
1un	SUCO PO TANG 25G LIMAO

Fonte: Elaborado pela autora.

Cada combo contém um dos quatro primeiros itens mais frequentes da base de dados, que são: o cacetinho, o tomate, a cebola e o leite piá integral. Optou-se por não incluir nos combos itens que necessitam de refrigeração para facilitar a disposição dos mesmos e, por isso, o item também frequente, gado 2ª carne moída II, foi desconsiderado.

Para elaboração dos preços dos combos, sugere-se diminuir apenas alguns centavos do valor do item frequente para chamar a atenção dos clientes. Destaca-se ainda que, durante o período de exposição dos combos, deve-se atentar para que estes itens não façam parte de outras promoções para não haver mais um incentivo de venda associado aos mesmos.

Contata-se que a ferramenta proporciona grande suporte na tomada de decisão, garantindo uma análise aprofundada das vendas, fornecendo informações que podem auxiliar na exposição dos produtos, melhorando o *layout* e favorecendo também a escolha de itens para as promoções. Completando o estudo de Annie e Kumar (2012), o qual propôs a melhoria de *layout* por meio do MBA, aproximando as associações encontradas nas prateleiras, a definição dos combos aqui proposta, após

período de exposição, também poderia sugerir uma melhoria na exposição dos produtos, considerando uma análise pós-venda.

No próximo capítulo serão abordadas as considerações finais sobre o estudo, bem como suas limitações e sugestões para trabalhos futuros.

6 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo geral avaliar regras de associação em uma base de dados de um supermercado utilizando a ferramenta *Data Mining* e uma de suas tarefas, o *Market Basket Analysis*. Para isto, aplicou-se o algoritmo Apriori às transações, o qual tem ampla relevância dentre os demais algoritmos, a fim de identificar padrões de consumo e, a partir disto, propor combos de venda.

No caso desta pesquisa, foram identificados os quatro itens mais frequentes de uma base de dados composta por 9.575 itens, além das associações mais relevantes, segundo as medidas de interesse especificadas, a partir de mais de 150.000 transações. Considerando os fatos apresentados, é incontestável a afirmação de que a mineração de dados entrega vantagens significativas, dado que dificilmente estes resultados seriam encontrados com simples análises, comumente utilizadas nas empresas, e em um pequeno espaço de tempo.

Ainda que seja possível identificar algumas informações com simples analises, os resultados alcançados com a aplicação da ferramenta apresentam diversas informações desconhecidas acerca dos produtos, como o destaque dos cinco itens mais frequentes de uma grande base de dados, bem como as regras de associação encontradas. Tais descobertas geram *insights* para os gestores que, consequentemente conseguem traduzir essas informações em decisões mais assertivas para o negócio.

E para alcançar resultados satisfatórios, a adoção de algumas práticas se tornam fundamentais, as quais Han, Kamber e Pei (2012) recomendam, onde além de trazerem a definição de um *Data Warehouse*, os autores afirmam que para manter estes dados armazenados é importante que se realize a limpeza, integração e consolidação dos mesmos. E mais importante do que isso, que sejam disponibilizadas ferramentas de TI para que seja possível realizar a análise destes dados e, a partir disso, auxiliar na tomada de decisões. Porém, foi observado que essa prática não costuma ser realizada dentro de empresas e, principalmente as de pequeno porte, como o caso da presente pesquisa.

Neste trabalho fica evidente o quanto a falta de um tratamento adequado dos dados influencia negativamente, pois para cumprir o objetivo geral, um dos objetivos específicos foi o desenvolvimento de uma estratégia de extração, tratamento e carregamento de dados não estruturados para obter as associações. O que colabora

com a afirmação apresentada no início deste trabalho, onde se enfatizou que as empresas coletam e armazenam os dados gerados diariamente, mas a maioria delas não sabe de que forma utilizá-los efetivamente.

Ainda em relação aos objetivos específicos, neste trabalho foi realizada uma busca nas bases de dados para definição do algoritmo a ser aplicado. Conforme os termos utilizados, a pesquisa indicou que o algoritmo Apriori foi utilizado na maioria dos estudos, porém não foram identificados trabalhos na língua portuguesa seguindo esta abordagem. Por isso este algoritmo foi escolhido para aplicação neste trabalho, também com o intuito de acrescentar referências na literatura.

A proposta aqui apresentada, de montar combos de vendas a partir de associações encontradas, revelou outro benefício obtido com a aplicação do DM, onde verificou-se que é possível encontrar também aqueles itens de menor frequência, por meio da abordagem escolhida. Ainda que não tenham sido avaliados, os combos sugeridos podem trazer impactos positivos, uma vez que estes itens associados podem apresentar aumento de vendas. Importante ressaltar que a proposta também pode revelar impacto negativo no momento que a venda casada não satisfaz o cliente e a compra não é realizada. E a partir destas respostas de venda, tanto a positiva quanto a negativa, é que o gestor consegue identificar os hábitos de consumo, aproximando-o do seu consumidor.

Pode-se dizer que o grande desafio das empresas está em reconhecer a devida importância das informações ocultas nos seus dados, que na grande maioria são não estruturados, e com isso buscar soluções para mudar o cenário atual e investir em ferramentas auxiliares. O mercado continua com rápidas mudanças e para o varejo, se tornar mais sensível e atento às expectativas do consumidor, pode impulsionar a estratégia do negócio. Entender de que forma a ferramenta *Data Mining*, e aqui evidenciada a tarefa *Market Basket Analysis*, deve ser utilizada proporciona inúmeros benefícios às organizações, sendo que a principal vantagem é fornecer informações aprofundadas sobre produtos e clientes, de maneira que proporcione uma diferente percepção que deve ser utilizada para suportar as decisões gerenciais, proporcionando também vantagem competitiva frente à concorrência.

Conforme já exposto, do ponto de vista teórico o presente trabalho tem a contribuir com o repositório brasileiro, dado que a RSL, segundo os critérios utilizados, realçou que ainda não dispõem de relevantes trabalhos com a mesma abordagem.

6.1 LIMITAÇÕES

Os dados analisados neste trabalho são referentes ao período de um ano, isto se deve ao fato de que anteriormente as transações não eram arquivadas em um servidor, portanto, não se teve acesso as transações de anos anteriores.

Han, Kamber e Pei (2012) afirmam que os valores de suporte e confiança devem ser altos, para que não sejam geradas regras desinteressantes, porém não especificam o que consideram valores altos. Desta forma, no presente trabalho foram utilizados valores similares ao trabalho realizado por Anselmo (2017).

De acordo com os termos pesquisados, não foi localizado um procedimento para o processo de montagem dos combos, portanto esta etapa foi desenvolvida sem uma metodologia que a sustentasse. No entanto, fica como um possível método com potencial de teste e validação em estudos futuros. No contexto desta limitação, não há evidencias que fragilize o estudo, haja vista que o resultado mostrou-se passível de ser implementado dentro do ambiente do estudo.

Ainda como limitação, expõe-se o fato de não haver tempo hábil para a montagem e exposição dos combos para análise.

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

Primeiramente, sugere-se dar continuidade a este trabalho, colocando em prática no estabelecimento mencionado, os combos sugeridos para sua posterior avaliação.

Indica-se ainda que esta pesquisa seja replicada e que sejam analisados os impactos de outras variáveis que também influenciam nas vendas, como por exemplo os dias da semana e as estações do ano. Sugere-se também identificar o perfil de compra do consumidor, considerando aspectos como faixa etária, classe social, forma de pagamento, possibilitando a criação de ofertas direcionadas para cada perfil.

O modelo proposto neste trabalho pode ser utilizado como base e também aplicação direta em diversos outros segmentos, como lojas de confecções, farmácias, lojas de cosméticos, entre outros, melhorando consequentemente a disseminação e conhecimento acerca do tema.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. **Fast Algorithms for Mining Association Rules**. In: 20th INTERNACIONAL CONFERENCE ON VERY LARGE DATABASE. Santiago Del Chile: Morgan Kaufmann, 1994, p. 487-499.

ALBUQUERQUE, Juliana. Supermercado: tecnologias transformam experiência do consumo. **Folha Pe,** [S.I.], 25 março 2018. Disponível em: <https://www.folhape.com.br/economia/economia/economia/2018/03/25/NWS,63056,10,550,ECONOMIA,2373-SUPERMERCADO-TECNOLOGIAS-TRANSFORMAM-EXPERIENCIA-CONSUMO.aspx>. Acesso em: 26 mai. 2018.

ALMEIDA, Fernando C. De. Desvendando o uso de redes neurais em problemas de administração de empresas. **Revista de Administração de Empresas**, São Paulo, v. 35, n. 1, p. 46–55, 1995. Disponível em: < http://www.scielo.br/pdf/rae/v35n1/a07v35n1.pdf>. Acesso em: 24 mar. 2018.

ANNIE, Loraine Charlet MC; KUMAR, Ashok D. Market basket analysis for a supermarket based on frequent itemset mining. **International Journal of Computer Science Issues (IJCSI)**, v. 9, n. 5, p. 257-264, 2012. Disponível em: < https://search.proquest.com/docview/1270319005/fulltextPDF/8CB8425BFA5D4F5APQ/1?accountid=26688>. Acesso em: 29 abr. 2018.

ANSELMO, Filomena Clara Gouveia. **Regras de Associação** - Market Basket Analysis - Itens Frequentes e Itens Raros. 2017. 72 f. Tese (Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão) - Mestrado em Economia e Gestão, Universidade do Porto (FEP), Portugal, 2017. Disponível em: < https://sigarra.up.pt/fadeup/pt/pub_geral.show_file?pi_gdoc_id=1032093>. Acesso em: 14 ago. 2018.

BACKES, Dirce Stein. et al. Grupo focal como técnica de coleta e análise de dados em pesquisas qualitativas. **O mundo da saúde**, v. 35, n. 4, p. 438-442, 2011. Disponivel em: <http://bvsms.saude.gov.br/bvs/artigos/grupo_focal_como_tecnica_coleta_analise_dados_pesquisa_qualitativa.pdf>. Acesso em: 18 jun. 2018.

BERRY, Michael J. A; LINOFF; Gordon. **Data Mining Techniques.** For Marketing, Sales, and Customer Support. New York: John Wiley & Sons, 1997.

BERTRAND, J. Will M; FRANSOO, Jan C. Operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v. 22, n. 2, p.241-264, 2002. Disponível em: <https://www.emeraldinsight.com/doi/abs/10.1108/01443570210414338>. Acesso em: 21 jul. 2018.

BRAVO, Fernando. Como o supermercadista pode aproveitar a tecnologia no supermercado. **Info Varejo,** [S.I.], 30 agosto 2017. Disponível em: <https://www.infovarejo.com.br/tecnologia-no-supermercado-aproveitar/>. Acesso em: 26 mai. 2018.

BRUSEBERG, Anne; McDONAGH-PHILP, Deana. Focus groups to support the

industrial/product designer: a review based on current literature and designers' feedback. **Applied Ergonomics**, v. 33, p. 27-38, 2002. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0003687001000539?via%3Dihub>. Acesso em: 18 jun. 2018.

CAMPOS, Claudinei José Gomes. Método de Análise de Conteúdo: ferramenta para a análise de dados qualitativos no campo da saúde. **Revista Brasileira de Enfermagem**, v. 57, n. 5, p. 611-614, 2004. Disponível em: <http://www.scielo.br/pdf/reben/v57n5/a19v57n5>. Acesso em: 19 jun. 2018.

CARVALHO, Alexey. *et al*. Modelo de Maturidade de Utilização da Tecnologia da Informação: Um enfoque para supermercados. In: CONGRESSO BRASILEIRO DE SISTEMAS, 2., 2006. São Paulo. **Anais eletrônicos...** São Paulo: Programa de Pós-Graduação - Centro Paula Souza – CEETEPS, 2006. Disponível em: < https://www.resea rchgate.net/publication/268445216_Modelo_de_Maturidade_de_Utilizacao_da_Tecnologia_da_Informacao_um_enfoque_para_supermercados>. Acesso em: 21 abr. 2018.

CARVALHO, Luís Alfredo Vidal de. **Datamining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. 1. ed. São Paulo: Érica, 2001.

CASTRO, Leandro Nunes de. **Introdução à mineração de dados** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016. Livro eletrônico.

CAVALCANTI, Herodes Beserra. Intensificação do trabalho nos supermercados Extra e Pão de Açúcar. **Pegada**, v. 15, n. 1, p. 50-69, 2014. Disponível em: <http://revista.fct.unesp.br/index.php/pegada/article/view/2678/2622>. Acesso em: 25 jun. 2018.

DATA mining é essencial para as empresas que querem se destacar. **Big Data Business HEKIMA**, [S.I.], 2017. Disponível em: < http://www.bigdatabusiness.com.br/por-que-a-mineracao-de-dados-e-essencial-para-as-empresas-que-querem-se-destacar/>. Acesso em: 02 jun. 2018.

DA SILVA LIMA, Thalles *et al*. A influência da Tecnologia da Informação no Desempenho de Micro e Pequenas Empresas em Belém do Pará. In: Congresso Latino Americando de Varejo: **Engaging and Interactive Shopper Experience - CLAV**, 11., 2017, São Paulo. Disponível em: <http://bibliotecadigital.fgv.br/ocs/index.php/clav/clav2017/paper/view/6061/1822>. Acesso em: 04 nov. 2018.

DIAS, Cristiano Araujo. **Descoberta de Conhecimento em Banco de Dados para Apoio a Tomada de Decisão**. 2002. 63 f. Trabalho de Conclusão de Curso de Especialização (Especialista em Informática Empresarial) - Curso de Especialização em Informática Empresarial, Universidade Estadual Paulista, Guaratinguetá, 2002. Disponível em: < https://s3.amazonaws.com/academia.edu.documents/35235909/CEIE0206.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1524153280&Signature=Td4vKOT0UbB8EAmDmVGU8sgbNWs%3D&response-content-

disposition=inline%3B%20filename%3DCEIE0206.pdf>. Acesso em: 30 mar. 2018.

DONATO, Claudio. **O conceito do varejo e a importância da tomada de decisão.** São Paulo, 19 novembro 2012. Disponível em: <http://www.administradores.com.br/artigos/economia-e-financas/o-conceito-do-varejo-e-a-importancia-da-tomada-de-decisao/67341/>. Acesso em: 19 mai. 2018.

DRESCH, Aline; LACERDA, Daniel Pacheco; ANTUNES JUNIOR, José Antônio Valle. DESIGN SCIENCE RESEARCH: Método de Pesquisa para Avanço da Ciência e Tecnolgia. **Gestão Produção**, v. 20, n. 4, p. 741–761, 2013. Disponível em: <http://www.scielo.br/pdf/gp/v20n4/aop_gp031412.pdf>. Acesso em: 09 jun. 2018.

DRESCH, Aline; LACERDA, Daniel Pacheco; ANTUNES JR, José Antônio Valle. **Design Science Research:** Método de pesquisa para avanço da ciência e tecnologia. Porto Alegre: Bookman, 2015. Livro eletrônico.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. Knowledge Discovery and Data Mining:Towards a Unifying Framework. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996, Portland. **Anais eletrônicos…** Portland: Association for the Advancement of Artificial Intelligence, 1996. Disponível em: <https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf>. Acesso em: 31 mar. 2018.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, Providence, v. 17, n. 3, p. 37-54, 1996. Disponível em: < https://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1230/1131>. Acesso em: 28 mar. 2018.

FERREIRA, João; MIRANDA, Miguel; ABELHA, António; MACHADO, José. O Processo ETL em Sistemas Data Warehouse. In: INForum 2010, [S.I.], 2010, Braga. II Simpósio de Informática. **Anais eletrônicos...** Braga: Universidade do Minho, 2010. p. 757-765. Disponível em: < http://inforum.org.pt/INForum2010/papers/sistemas-inteligentes/Paper080.pdf>. Acesso em: 24 out. 2018.

FERREIRA, Nilson Gessoni Sapata Aguilar; SILVEIRA, Marco Antonio Pinheiro Da. Impactos da Informatização na Gestão de Supermercados. **RAM - Revista de Administração Mackenzie**, São Paulo, v. 8, n.1, p. 108-132, 2007. Disponível em: <http://www.redalyc.org/html/1954/195416699006/>. Acesso em: 21 abr. 2018.

FINGERL, Eduardo Rath (Dir.). **Comércio Varejista:** Supermercado. Rio de Janeiro, 1996. Disponível em: <https://www.bndes.gov.br/SiteBNDES/export/sites/default/bndes_pt/Galerias/Arquivos/conhecimento/relato/supmerca.pdf>. Acesso em: 26 mai. 2018.

FREITAS, Henrique. **A informação como ferramenta gerencial:** um telessistema de informação em marketing para o apoio à decisão. 1. ed. Porto Alegre: Ortiz, 1993.

GIANOTTI, Renata Cabral. **Marketing de serviços e varejo.** São Leopoldo: Unisinos, 2013.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 6. ed. São Paulo: Atlas, 2017. Livro eletrônico.

GOMES, Dayane. Gerenciando a Tecnologia da Informação nos Negócios: Estudo de caso em um supermercado. **Profissionais TI**, [S.I.], 15 janeiro 2013. Disponível em: <https://www.profissionaisti.com.br/2013/01/gerenciando-a-tecnologia-da-informacao-nos-negocios-estudo-de-caso-em-um-supermercado-2/>. Acesso em: 04 nov. 2018.

GONÇALVES, Eduardo Corrêa. Regras de associação e suas medidas de interesse objetivas e subjetivas. **INFOCOMP: Jounal of Computer Science**. Rio de Janeiro, v. 4, n. 1., p. 27-36, 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/index.php/INFOCOMP/article/view/79>. Acesso em: 30 mar. 2018.

GONÇALVES, Lóren Pinto Ferreira. **Mineração de Dados em Supermercados:** O Caso do Suoermercado "Tal". 1999. 36 f. Dissertação (Mestrado em Administração) - Programa de Pós-Graduação em Administração, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, 1999. Disponível em: < http://www.ufrgs.br/gianti/files/orientacao/mestrado/proposta/pdf/24_mest_proposta_goncalves.pdf>. Acesso em: 31 mar. 2018.

GROTH, Robert. **Data Mining:** Building Competitive Advantage. 1. ed. New Jersey: Prentice Hall PTR, 2000.

HAHSLER, Michael; GRÜN, Bettina; HORNIK, Kurt; BUCHTA, Christian. **Introduction to arules** – A computational environment for mining association rules and frequent item sets. Disponível em: <https://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>. Acesso em: 17 ago. 2018.

HAND, David; MANNILA, Heikki; SMYTH, Padhraic. **Principles Of Data Mining (Adaptative Computation and Machine Learning)**. 1. ed. Cambridge: MIT Press, 2001.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data Mining:** Concepts and Techniques. 3. ed. San Francisco: Morgan Kaufmann, 2012.

ITO, Vivian. Tecnologia com estratégia vira arma para estimular compras por impulso. **DCI Diário Comércio Indústria & Serviços**, São Paulo, 22 fevereiro 2018. Disponível em: <https://www.dci.com.br/comercio/tecnologia-com-estrategia-vira-arma-para-estimular-compras-por-impulso-1.685471>. Acesso em: 21 abr. 2018.

LAS CASAS, A. L.; BARBOZA, V. A. Marketing no Varejo. In: ESTRATÉGIAS DO MARKETING PARA VAREJO: Inovações e diferenciações estratégicas que fazem a diferença no marketing de varejo. Novatec. 2007. p. 5 a 23. Disponível em: <http://www.martinsfontespaulista.com.br/anexos/produtos/capitulos/250037.pdf>. Acesso em: 19 mai. 2018.

LAW, Averill M; KELTON, W. David. **Simulation Modeling and Analysis**. 2. ed. Singapore: McGraw- Hill, Inc., 1991.

LIAO, Shu-Hsien; CHEN, Chyuan-Meei; WU, Chung-Hsin. Mining customer

knowledge for product line and brand extension in retailing. **Expert systems with Applications**, v. 34, n. 3, p. 1763-1776, 2008. Disponível em: < https://www.sciencedirect.com/science/article/pii/S095741740700053X>. Acesso em: 01 mai. 2018.

MAINALI; Sanjeev. **Market Basket Analysis**. 2016. 45 f. Trabalho de Conclusão de Curso (Bacharel em Ciência da Computação e Tecnologia da Informação) - Department of Computer Science and Information Technology, Deerwalk Institute of Technology, Tribhuvan University, Nepal, 2016.

MONTEIRO, Ivan. APASNEXT: a importância da tecnologia e da inovação no supermercado. **APAS SHOW**, [S.I.], 18 fevereiro 2017. Disponível em: < http://apasshow.com.br/blog/index.php/2017/09/18/apasnext-importancia-da-tecnologia-e-da-inovacao-no-supermercado/>. Acesso em: 02 jun. 2018.

MORANDI, M. I. W. M.; CAMARGO, L. F. R. **Revisão Sistemática da Literatura**. In: Design Science Research: método de pesquisa para avanço da ciência e tecnologia. Porto Alegre: Bookman, 2015. p. 141–172. Livro eletrônico.

MORGADO, Maurício. Três grandes desafios para o varejo do futuro. **GV-executivo,** São Paulo, v. 16, n. 1, p. 33-35, 2017. Disponível em: <http://rae.fgv.br/sites/rae.fgv.br/files/ce6_3.pdf>. Acesso em: 26 mai. 2018.

MUSALEM, Andres; ABURTO, Luis; BOSCH, Maximo. Market basket analysis insights to support category management. **European Journal of Marketing**, 2018.

NETO, Reinaldo Morabito; PUREZA, Vitoria. **Modelagem e Simulação**. In: Metodologia de pesquisa em engenharia de produção e gestão de operações. 2. ed. Rio de Janeiro: Elsevier, 2012. p. 169–198.

O EDI no setor varejista: um antes e depois. **Edicom Connecting Business**, [S.I.], 07 junho 2016. Disponível em: <https://www.edicomgroup.com/pt_BR/news/8302-o-edi-no-setor-varejista-um-antes-e-um-depois>. Acesso em: 28 jun. 2018.

O que Big Data muda na realidade dos profissionais de marketing?. **Big Data Business HEKIMA**, [S.I.], 2017. Disponível em: < http://www.bigdatabusiness.com.br/o-que-o-big-data-muda-na-realidade-dos-profissionais-de-marketing/>. Acesso em: 02 jun. 2018.

PAULA, Welington Lourenco Melo de. Extract, Transformation and Load (ETL) - Ferramentas BI, **DEVMEDIA**, [S.I.], 2012. Disponível em: <https://www.devmedia.com.br/extract-transformation-and-load-etl-ferramentas-bi/24408>. Acesso em: 24 out. 2018.

PAPAVASILEIOU, Vasilios; TSADIRAS, Athanasios. Evaluating time variations to identify valuable association rules in market basket analysis. **Intelligent Decision Technologies**, v. 7, n. 1, p. 81-90, 2013. Acesso em: < http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=4&sid=fe7ef26e-8cbc-4503-b3dc-d6e83cf746e7%40pdc-v-sessmgr06>. Acesso em: 02 mai. 2018.

PARENTE, Juracy. **Varejo no Brasil**: gestao e estratégia. São Paulo: Atlas, 2011.

PORTER, Michael E. O que é estratégia. **Harvard Business Review**, v. 74, n. 6, p. 61-78, 1996.

PORTER, Michael E; MILLAR, Victor E. How information Gives You Competitive Advantage. **Harvard Business Review,** v. 63, n. 4, p. 149-160. 1985.

PORTO, Geciane S.; BRAZ, Reinaldo N.; PLONSKI, Guilherme Ary. O intercâmbio eletrônico de dados - EDI e seus impactos organizacionais. **FAE**, v. 3, n. 3, p. 13-29, 2000. Disponível em: <https://revistafae.fae.edu/revistafae/article/view/512/407>. Acesso em: 28 jun. 2018.

ROMERO, Cláudia Buhamra Abreu. **Gestão de marketing no varejo:** conceitos, orientações e práticas. 1. ed. São Paulo: Atlas S.A, 2012. Livro eletrônico.

SACILOTTI, Adaní Cusin. **A Importância da Tecnologia da Informação nas Micro e Pequenas Empresas :** Um Estudo Exploratório na Região de Jundiaí. 2011. 116 f. Dissertação (Mestrado em Administração) - Programa de Mestrado em Administração, Faculdade Campo Limpo Paulista (FACCAMP), São Paulo, 2011. Disponível em: <http://www.faccamp.br/new/arq/pdf/mestrado/Documentos/producao_discente/2011/04abril/AdaniCusinSacilotti/dissertaCAo.pdf> Acesso em: 31 mar. 2018.

SEMAAN, Gustavo Silva; GRAÇA, Andrei De Alencastro; DIAS, Carlos Rodrigo. Extração De Associações Em Bases De Dados De Varejo. Simpósio Brasileiro de Pesquisa Operacional, 38., 2006, Goiânia. **Anais eletrônicos...** Juiz de Fora: Faculdade Metodista Granbery, 2006. p. 1312–1322. Disponível em: < http://www.din.uem.br/sbpo/sbpo2006/pdf/arq0197.pdf>. Acesso em: 31 mar. 2018.

SHELKE, R. R.; DHARASKAR, R. V.; THAKARE, V. M. Association Rule Mining for Supermarket Sale Analysis. **International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)**, v. 4, n. 4, p. 20-22, 2017.

SILVA, Edna Lúcia; MENEZES, Estera Muszkat. **Metodologia da Pesquisa e Elaboração de Dissertação**. 4. ed. Florianópolis: UFSC, 2005. p. 138p, 2005. Disponível em: <https://projetos.inf.ufsc.br/arquivos/Metodologia_de_pesquisa_e_elaboracao_de_teses_e_dissertacoes_4ed.pdf>. Acesso em: 03 jun. 2018.

SOUZA JÚNIOR, Marcílio Barbosa Mendonça; MELO, Marcelo Soares Tavares de; SANTIAGO, Maria Eliete. A análise de conteúdo como forma de tratamento dos dados numa pesquisa qualitativa em Educação Física escolar. **Movimento**, v. 16, n. 3, p. 31-49, 2010. Disponível em: <http://www.seer.ufrgs.br/Movimento/article/viewFile/11546/10008>. Acesso em: 19 jun. 2018.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining** Mineração de Dados. Tradução de Acauan P. Fernandes. 1. ed. Rio de Janeiro: Ciência Moderna Ltda, 2009.

TECNOLOGIA amplia interpretação de dados para o varejo. **APAS SHOW**, [S.I.], 14 junho 2018. Disponível em: <

http://apasshow.com.br/blog/index.php/2018/06/14/tecnologia-amplia-interpretacao-de-dados-para-o-varejo/>. Acesso em: 11 nov. 2018.

TIAN, Jun Feng; LI, Li Xian. Applied Research on Client Identification Based on Association Rules. In: **Applied Mechanics and Materials**. Trans Tech Publications, p. 432-435, 2011.

VIRI, Natalia. Pão de Açucar descobre um tesouro nos algoritmos. **Brazil Journal**, [S.I.], 28 julho 2017. Disponível em: < http://braziljournal.com/pao-de-acucar-descobre-um-tesouro-nos-algoritmos>. Acesso em: 15 jun. 2018.

VAROTTO, Luís Fernando. História do varejo. **GV-executivo**, [S.I.], v. 5, n. 1, p. 86-90, 2006. Disponível em: <http://rae.fgv.br/sites/rae.fgv.br/files/artigos/4224.pdf>. Acesso em: 26 mai. 2018.

YANG, Yinghui; HAO, Chunhui. Product selection for promotion planning. **Knowledge and Information Systems**, v. 29, n. 1, p. 223-236, 2010. Disponível em: <https://link.springer.com/article/10.1007/s10115-010-0326-8>. Acesso em: 06 mai. 2018.

ZEKIC-SUSAC, Marijana; HAS, Adela. Data Mining as Support to Knowledge Management in Marketing. **Business Systems Research**, v. 6, n. 2, p. 18-30, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923600001238>. Acesso em: 06 mai. 2018.

ZUMEL, Nina; MOUNT, John. **Practical Data Science With R**. 2. ed. Shelter Island: Manning, 2014.

APÊNDICE A – PROTOCOLO DE REVISÃO SISTEMÁTICA DA LITERATURA

Framework conceitual	Revisão sistemática da literatura para identificar trabalhos sobre aplicação de *Data Mining* em supermercados.
Contexto	Aplicação da ferramenta *Data Mining* com a finalidade de encontrar padrões de itens frequentes em supermercados.
Idiomas	Português e Inglês.
Horizonte	Sem restrição temporal.
Critérios de inclusão	Artigos que analisam a aplicação da ferramenta na gestão das empresas.
Critérios de exclusão	Artigos pagos que estão em desacordo do objetivo dessa pesquisa.
Termos de busca	Mineração de dados AND Regras de associação AND Análise da Cesta de Compras
	Mineração de dados AND Regras de associação AND Supermercado
	Mineração de dados AND Regras de associação AND Algoritmo Apriori
	Regras de Associação AND Análise da Cesta de Compras AND Supermercado
	Regras de Associação AND Análise da Cesta de Compras AND Algoritmo Apriori
	Data Mining AND *Association Rules* AND *Market Basket Analysis*
	Data Mining AND *Association Rules* AND *Supermarket*
	Data Mining AND *Association Rules* AND *Apriori Algorithm*
	Association Rules AND *Market Basket Analysis* AND *Supermarket*
	Association Rules AND *Market Basket Analysis* AND *Apriori Algorithm*
Fontes de busca	EBSCOHost
	Emerald
	ProQuest