UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE COMPUTAÇÃO
CURSO DE CIÊNCIA DA COMPUTAÇÃO

THALES HERON PIOTTO DE LIRA

ANÁLISES COMPARATIVAS DE VENDAS EM UM SUPERMERCADO DO MUNICÍPIO DE FOZ DO IGUAÇU POR MEIO DA MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO ANTES E DURANTE A PANDEMIA DE COVID-19

TRABALHO DE CONCLUSÃO DE CURSO

MEDIANEIRA

2020

THALES HERON PIOTTO DE LIRA

ANÁLISES COMPARATIVAS DE VENDAS EM UM SUPERMERCADO DO MUNICÍPIO DE FOZ DO IGUAÇU POR MEIO DA MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO ANTES E DURANTE A PANDEMIA DE COVID-19

Trabalho de Conclusão de Curso apresentado ao Departamento Acadêmico de Computação da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do título de "Bacharel em Ciência da Computação".

Orientador: Prof. Dr. Alan Gavioli

Co-orientador: Prof. Dr. Arnaldo Candido Junior

MEDIANEIRA

2020

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Diretoria de Graduação e Educação Profissional
Coordenação do Curso de Ciência da Computação

TERMO DE APROVAÇÃO

ANÁLISES COMPARATIVAS DE VENDAS EM UM SUPERMERCADO DO MUNICÍPIO DE FOZ DO IGUAÇU POR MEIO DA MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO ANTES E DURANTE A PANDEMIA DE COVID-19

Por

THALES HERON PIOTTO DE LIRA

Este Trabalho de Conclusão de Curso foi apresentado às 14:00h do dia 19 de Novembro de 2020 como requisito parcial para a obtenção do título de Bacharel no Curso de Ciência da Computação, da Universidade Tecnológica Federal do Paraná, Câmpus Medianeira. O candidato foi arguido pela Banca Examinadora composta pelos professores abaixo assinados. Após deliberação, a Banca Examinadora considerou o trabalho aprovado.

Prof. Dr. Alan Gavioli
UTFPR - Câmpus Medianeira

Prof. Dr. Evando Carlos Pessini
UTFPR - Câmpus Medianeira

Prof. Dr. Everton Coimbra de Araújo
UTFPR - Câmpus Medianeira

A folha de aprovação assinada encontra-se na Coordenação do Curso.

RESUMO

LIRA, Thales H. P.. ANÁLISES COMPARATIVAS DE VENDAS EM UM SUPERMERCADO DO MUNICÍPIO DE FOZ DO IGUAÇU POR MEIO DA MINERAÇÃO DE REGRAS DE ASSOCIAÇÃO ANTES E DURANTE A PANDEMIA DE COVID-19. 59 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2020.

Descoberto no final do ano de 2019, o Sars-CoV-19 tem mudado a maneira como vivemos no Brasil, através do fechamento de comércios não essenciais, escolas, e com horário reduzido para atendimento ao público em supermercados e farmácias. Ademais, acontecem demissões em massa por todo o país, que ocorrem com a intenção de reduzir custos, assim, alterando diretamente a renda do consumidor. Nesse sentido, a população passou a ir menos aos mercados e a comprar apenas o essencial. Com isso, surgiu o interesse em iniciar uma pesquisa sobre o comportamento de compras do consumidor, aliado aos fundamentos computacionais, como o desenvolvimento de algoritmos de captura de dados em notas fiscais eletrônicas, por meio de linguagem de programação Python. Posteriormente, utilizou-se os dados formatados e pré-processados em uma suite de ferramentas de inteligência artificial para a realização de mineração de dados por meio de algoritmo de associação, o Apriori. Com a necessidade de poder averiguar os comportamentos de várias frentes diferentes, foi determinado que toda a pesquisa analisaria três meses completos de pico da pandemia, esses meses, por sua vez, segmentados e comparados entre os mesmos e, também, separados por dia da semana. Em paralelo, para o trabalho ser mais fidedigno quanto à mudança de comportamento, para todo viés analisado durante período pandêmico, foram também analisados períodos sem as consequências da pandemia, ou seja, do ano anterior. Após a realização de todas as análises propostas, os resultados de mudanças de comportamentos foram considerados interessantes, afinal, ocorreram mudanças em todos os períodos, como nos meses de abril e junho, em que o consumo de refrigerantes, junto com o de produtos de panificação, dispararam, não só em comparação com 2019, mas também com o mês de maio, em que isso não ocorreu. Foi percebido, ainda, que diferente de todos os outros dias da semana, apenas no domingo houve uma tendência de compra de iogurte com produtos de panificação, em conjunto com um terceiro item que variava entre bolachas, queijos, carne bovina moída, entre outros. Por fim, esta pesquisa se encerrou com o demonstrativo dos relatórios ao especialista, que apontou o que foi considerado relevante e, com isso, poderia ajudar um supermercado a tomar decisões para tentar um aumento no faturamento.

Palavras-chave: mineração de dados, aprendizado não supervisionado, supermercado, pandemia

ABSTRACT

LIRA, Thales H. P.. COMPARATIVE SALES ANALYSIS IN A SUPERMARKET IN THE MUNICIPALITY OF FOZ DO IGUAÇU THROUGH THE MINING OF ASSOCIATION RULES BEFORE AND DURING THE COVID-19 PANDEMIC. 59 f. Trabalho de Conclusão de Curso – Curso de Ciência da Computação, Universidade Tecnológica Federal do Paraná. Medianeira, 2020.

Discovered at the end of 2019, the Sars-CoV-19 has changed the way we live in Brazil, by closing non-essential stores, schools, and with reduced hours for the public in supermarkets and pharmacies. In addition, mass layoffs occur across the country, which occur with the intention of reducing costs, thus directly altering the consumer's income. In this sense, the population started to go less to the markets and buy only the essentials. With that, the interest arose in starting a research on the consumer shopping behavior, allied to the computational fundamentals, such as the development of data capture algorithms in electronic invoices, through the Python programming language. Subsequently, it was intended to use the formatted and pre-processed data in a suite of artificial intelligence tools to perform data mining through an association algorithm, the Apriori. With the need to be able to ascertain the behaviors of several different fronts, it was determined that the entire survey would analyze three full months of pandemic peak, these months, in turn, segmented and compared between them and also separated by day of the week . In parallel, for work to be more reliable in terms of behavior change, for all bias analyzed during the pandemic period, periods without the consequences of the pandemic were also analyzed, that is, the previous year. After carrying out all the proposed analyzes, the results of behavioral changes were considered interesting, after all, there were changes in all periods, such as in the months of April and June, when the consumption of soft drinks, together with that of bakery products , soared, not only in comparison with 2019, but also with the month of May, when this did not happen. It was also noticed that, unlike all other days of the week, only on Sunday there was a tendency to buy yogurt with bakery products, together with a third item that varied between crackers, cheeses, ground beef, among others. Finally, this research ended with the statement of reports to the specialist, who pointed out what was considered relevant and, with that, could help a supermarket to make decisions to try to increase its sales.

Keywords: data mining, unsupervised machine-learning, supermarket, pandemic

Dedico este trabalho à minha família, meu pai Isaias de Lira, minha mãe Cássia Regina Piotto de Lira, meu irmão Thiago Henrique Piotto de Lira e, a nossa joia da família, minha sobrinha Maria Victória V. Lira. Pelo amor, incentivo e apoio para a conclusão desta importante etapa e por me ajudar a buscar meus objetivos com muito mais garra e dedicação.

AGRADECIMENTOS

Ao meu orientador, Dr. Alan Gavioli, e ao meu co-orientador, Dr. Arnaldo Candido Junior, meus agradecimentos, não só durante a indispensável orientação para elaboração deste trabalho, mas por toda a caminhada e ensinamentos durante a graduação.

"Sempre lembrarei dos excelentes guias que tive em forma de professores."

Meus sinceros agradecimentos aos inestimáveis amigos que fiz ao longo do período no curso de Ciência da Computação nesta instituição: Alexandre, André, Angelo, Felipe, Leandro, Maria Vitória, Rafael, Tarlon e Vinicios.

"Nenhum caminho é longo demais quando um amigo nos acompanha."

LISTA DE FIGURAS

FIGURA 1 – Exemplo de um XML e sua estrutura 15
FIGURA 2 – Fases do processo de descoberta de conhecimento em base de dados 17
FIGURA 3 – Variedade de disciplinas que envolvem a mineração de dados 18
FIGURA 4 – Amostra aleatória de vendas de um supermercado 21
FIGURA 5 – Interface gráfica inicial do software de mineração de dados WEKA 24
FIGURA 6 – Fluxograma das etapas desta pesquisa 28
FIGURA 7 – Estrutura XML de uma nf-e 29
FIGURA 8 – Estrutura adaptada de um arquivo texto CSV para leitura pelo algoritmo .. 30
FIGURA 9 – Tela parcial da aba *preprocess* do WEKA 31
FIGURA 10 – Tela parcial da aba *associate* do WEKA 32
FIGURA 11 – Configuração do algoritmo apriori 33

LISTA DE TABELAS

TABELA 1 – Demonstrativo da separação por colunas das classes e seus valores 30
TABELA 2 – Resultado do aglomerado total em pandemia 38
TABELA 3 – Resultado aglomerado total em não pandemia 38
TABELA 4 – Resultado derivado do aglomerado total sem produtos em abundância 39
TABELA 5 – Resultado de abril sem produtos em abundância 41
TABELA 6 – Resultado de maio sem produtos em abundância 41
TABELA 7 – Resultado de junho sem produtos em abundância 42
TABELA 8 – Resultado 15 melhores regras de domingo em pandemia 43
TABELA 9 – Resultado 15 melhores regras de domingo em não pandemia 43
TABELA 10 – Resultado de domingo sobre iogurte e item adverso com prod. de panificação 44
TABELA 11 – Resultado parcial de segunda com produtos do conjunto L(5) 44
TABELA 12 – Resultado Parcial de Sexta com Regras Discrepantes 48

LISTA DE SIGLAS

CSV	Comma-Separated-Values
KDD	Knowledge Discovery in Databases
NF-e	Nota Fiscal Eletrônica
RFB	Receita Federal do Brasil
WEKA	Waikato Environment for Knowledge Analysis
XML	Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO 11
1.1 OBJETIVOS GERAL E ESPECÍFICOS 12
1.2 JUSTIFICATIVA 12
2 FUNDAMENTAÇÃO TEÓRICA 13
2.1 LINGUAGEM PYTHON 13
2.1.1 DINAMISMO DA LINGUAGEM 13
2.1.2 PYTHON MULTIPLATAFORMA E USABILIDADES 14
2.2 XML E SUAS APLICAÇÕES PARA FISCO 14
2.2.1 ESTRUTURA DE UM XML 15
2.2.2 NOTA FISCAL ELETRÔNICA 16
2.3 MINERAÇÃO DE DADOS 16
2.3.1 TÉCNICAS DE MINERAÇÃO 20
2.3.1.1 ASSOCIAÇÃO 20
2.3.1.2 EVOLUÇÃO DO ALGORITMO DE ASSOCIAÇÃO 22
2.4 SOFTWARES DE MINERAÇÃO DE DADOS 22
2.4.1 WEKA 23
3 MATERIAIS E MÉTODOS 25
3.1 SISTEMÁTICA CRONOLÓGICA DA PESQUISA 25
3.1.1 ETAPA 1 25
3.1.2 ETAPA 2 E ETAPA 3 26
3.1.3 ETAPA 4, ETAPA 5 E ETAPA 6 26
3.1.4 DA ETAPA 7 27
3.2 O CONJUNTO DE DADOS 27
3.3 PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS 29
3.4 O PROCESSO DE MINERAÇÃO DE DADOS 31
3.5 ESCOLHA E CONFIGURAÇÃO DO ALGORITMO 32
3.6 DAS CONFIGURAÇÕES PARA CADA PERÍODO 33
3.6.1 AGLOMERADO DOS TRÊS MESES 34
3.6.2 SEGMENTAÇÃO MENSAL 34
3.6.3 POR DIA DA SEMANA 35
4 RESULTADOS E DISCUSSÃO 37
4.1 DOS RESULTADOS DO TODO O PERÍODO 37
4.1.1 ANÁLISE DO TODO 37
4.1.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE 39
4.2 DOS RESULTADOS DA SEGMENTAÇÃO MENSAL 40
4.2.1 ANÁLISE DO TODO 40
4.2.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE 40
4.3 DOS RESULTADOS DA SEGMENTAÇÃO SEMANAL 42
4.3.1 DOMINGO 42
4.3.2 SEGUNDA-FEIRA 44
4.3.2.1 ANÁLISE DO TODO 44

4.3.2.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE 45
4.3.3 TERÇA-FEIRA 45
4.3.3.1 ANÁLISE DO TODO 46
4.3.3.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE 46
4.3.4 QUARTA-FEIRA 46
4.3.5 QUINTA-FEIRA 47
4.3.6 SEXTA-FEIRA 47
4.3.7 SÁBADO 48
5 CONCLUSÕES E TRABALHOS FUTUROS 49
5.1 CONCLUSÔES 49
5.2 TRABALHOS FUTUROS 50
REFERÊNCIAS 51
Apêndice A – CODIGO-FONTE DO MODULO DE CAPTURA DE DADOS DESENVOLVIDO EM PYTHON 3.7 53
Apêndice B – CODIGO-FONTE DO MODULO DE CRIAÇÃO DE CSV PARA MINERAÇÃO DE DADOS DESENVOLVIDO EM PYTHON 3.7 57

1 INTRODUÇÃO

Apesar de ser oficialmente descoberto no final do ano de 2019, o vírus Sars-CoV-2 trouxe suas maiores consequências em 2020 e, dessa maneira, assustou a sociedade como não acontecia desde a última pandemia que atingiu o mundo, há, aproximadamente, 100 anos (GUINANCIO et al., 2020; THULER; MELO, 2020).

Naquela época, bem como atualmente, o consumo diário de alimentos era uma das principais práticas inerentes ao ser humano. Todavia, diferente da pandemia do século passado, a atual atinge um período em que se é possível, por meio do uso de tecnologias computacionais, buscar conhecimento e oferecer pesquisas baseadas nas consequências por ela ocasionadas. O cotidiano da população mudou e, com ele, seus costumes. Hodiernamente, se vê pessoas de máscaras que cobrem do nariz ao queixo, também é observado que cada entrada de estabelecimento comercial, em geral, tem um recipiente com álcool para a higienização das mãos (GUINANCIO et al., 2020).

Sendo assim, tais cuidados diminuíram a incidência de pessoas nas ruas. Logo, é natural imaginar que quando essas necessitam sair, serão mais incisivas em suas compras, para não precisarem retornar tão cedo às ruas. Portanto, a ideia desta pesquisa foi se utilizar de técnicas de mineração de dados, por meio do uso de uma suíte de software de inteligência artificial, para realizar uma análise em cima do comportamento humano sobre um conjunto de dados de vendas de um supermercado, localizado na região de Foz do Iguaçu, no interior do estado do Paraná (VIEIRA; FELIPE, 2001; HAN; KAMBER, 2001).

Em paralelo ao exposto, a mineração de dados poderia acontecer dentro de um *data warehouse*, ou de conjunto de dados mais convencionais, como planilhas ou textos padronizados para serem lidos. Desta forma, o processo pode identificar padrões e descobrir informações relevantes, que cooperam com o comerciante no processo de formação de preços ou combinações de produtos e no comportamento de clientes em relação às compras. Assim, permitindo uma possível tomada de decisão estratégica por parte da empresa-colaboradora dos dados (VIEIRA; FELIPE, 2001).

1.1 OBJETIVOS GERAL E ESPECÍFICOS

Esse trabalho teve como objetivo analisar dados de venda, pré-pandemia e durante a pandemia, de um supermercado localizado em um grande bairro da cidade de Foz do Iguaçu e, com isso, utilizar técnicas de regras de associação de mineração de dados. Esse objetivo principal pode ser dividido nos seguintes objetivos específicos:

- Organizar e analisar conjuntos de dados de vendas do período operacional durante três meses de período pandêmico e período não pandêmico, subdividos em período pandêmico completo, mensal e semanal, bem como em período não pandêmico, completo e semanal;
- Identificar padrões de comportamento de clientes do supermercado analisado, em suas compras;
- Apresentar uma análise de uma auditoria computacional aos especialistas do mercado, para que avaliassem os resultados da análise computacional e fornecessem um *feedback* em relação à novidade e à relevância desses resultados.

1.2 JUSTIFICATIVA

Mesmo sabendo que um período pandêmico é algo momentâneo, com início, meio e fim, o comportamento humano pode sofrer mudanças permanentes, vide a epidemia do H1N1 anos atrás, quando se tornou mais comum o uso do álcool em gel nas bancadas de comércios e nas casas. Nesse sentido, as identificações de padrões de consumo durante esse período podem trazer benefícios estratégicos para o comerciante.

Em razão da problemática relatada, a intenção desta monografia foi analisar mudanças de hábitos de consumo das pessoas durante a pandemia.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 LINGUAGEM PYTHON

Em meados de 1990, foi criada, pelo holanês Guido van Rossum, uma linguagem de programação. Inspirado a nomear a linguagem por conta de um show de comédia do canal britânico BBC, chamado “Python Flying Circus”, que ele e sua equipe gostavam de assistir, a linguagem passou a se chamar Python (LUTZ, 2007).
Esta linguagem evoluiu muito até os dias atuais e hoje se encontra em projetos de softwares empresariais, sistemas embarcados, microsserviços na web, desenvolvimento de interfaces gráficas ao usuário, programação de algoritmos de inteligência artificial, processamentos de cálculos matemáticos e, até mesmo, jogos digitais (PYTHON, 2019).

2.1.1 DINAMISMO DA LINGUAGEM

O Python atende ao paradigma de programação orientada a objetos de tipagem dinâmica, isso é, não havendo a necessidade de declarar o tipo primitivo da variável - que, na realidade, em Python é entendido como objeto - onde o interpretador reconhecerá o tipo primitivo do momento que entender o valor do dado dentro daquela variável. Esse conceito de reconhecimento em tempo de instanciação permite a mudança do tipo primitivo do dado dinamicamente durante a execução em pontos diferentes do algoritmo. Outro motivo que permite esse dinamismo é que a Python é uma linguagem interpretada e interativa (BORGES, 2014).

2.1.2 PYTHON MULTIPLATAFORMA E USABILIDADES

Apesar de inicialmente o Python ter sido criado para manipulação de dados e processamentos flexíveis matemáticos Lutz (2007), outras propostas para a utilização da linguagem surgiram por conta de alguns fatores, como a automatização da manipulação da memória alocada, operação de alto nível com interpolação de tipos primitivos dentro da mesma variável, e a fácil inserção de módulos e/ou frameworks em meio ao desenvolvimento. Esses fatores citados contribuíram para que a linguagem começasse a ganhar espaço em outros ambientes da computação, como empresas ou instituições do ramo da tecnologia, como a Google, com aplicações Web; a Yahoo, também com aplicações Web; e a Microsoft, com a IronPython, que é uma linguagem híbrida de .Net com Python (BORGES, 2014).
A linguagem Python segue uma filosofia para facilitação de recursos, bibliotecas e funções, chamada *battery included*. É feita uma analogia com brinquedos novos que funcionam à bateria, porém,que ao comprar, não vinham com o acessório necessário para seu funcionamento. Já no caso da linguagem com tal filosofia seguida, ela estaria pronta para ser usada assim que instalada. Há uma separação dessas bibliotecas por área de aplicação dos recursos já incluídos, alguns exemplos são (PYTHON, 2019):

- **Web Tasks** XML Parsing, Recuperação de URL com libcurl, Async I/O com o Twisted e importação de gráficos para web com PIL e Chaco;
- **Programação científicia** Manipulação de dados e cálculos matemáticos com NumPy e SciPy;
- **Programação Desktop** Manipulação de arquivos específicos com OpenCSV e ReportLab para PDF;
- **Desenvolvimento de Jogos** Utilização de motores gráficos para jogos com PyGame.

2.2 XML E SUAS APLICAÇÕES PARA FISCO

Derivado da ISO 8879, de 1986, que outorga as diretrizes da padronização das linguagens de marcação, o XML (do inglês *Extensible Markup Language*) foi projetado para desempenhar o transporte de dados em larga escala. Por consequência desse seu propósito, o XML se tornou uma das formas mais importantes na troca de informação na web e sistemas (ISO, 1986; W3C,

2016).

2.2.1 ESTRUTURA DE UM XML

A estrutura de um XML é simples, ocorre de forma hierárquica e separada por marcações. Como sugere sua mantenedora W3Schools (1999), a hierarquia do XML separada pela **raiz** e suas **folhas** respectivas. O que determina a hierarquia são as marcações, e estas são formadas por: sinal de menor, nome da marcação e, por fim, o sinal de maior. O que vier escrito após a marcação é chamado de texto da marcação, que é o conteúdo em si, dentro daquela marcação. Para sinalizar o fim da marcação, os mesmos critérios anteriores de criação de uma marcação devem ser atendidos, mas com um caractere novo, que indicaria o fim da marcação, o sinal barra antes do nome da marcação (W3SCHOOLS, 1999).
Para melhor entendimento, segue um exemplo na Figura 1.

```
<?xml version="1.0"?>
- <Exemplo>
  - <InfoAcademico>
      <Nome>THALES H. P. DE LIRA</Nome>
      <anoIngresso>2015</anoIngresso>
      <semestre>1</semestre>
      <Curso>Ciência da Computação</Curso>
      <Periodo>8</Periodo>
      <Materia>Trabalho de Conclusão de Curso II</Materia>
    </InfoAcademico>
  - <InfoUniversidade>
      <Nome>Universidade Tecnologica Federal do Paraná</Nome>
      <Campus>Medianeira</Campus>
    - <Endereco>
        <Rua>Avenida Brasil</Rua>
        <numRua>4232</numRua>
        <Bairro>Parque Independência</Bairro>
        <Cidade>Medianeira</Cidade>
        <Uf>PR</Uf>
      </Endereco>
    </InfoUniversidade>
</Exemplo>
```

Figura 1 – Exemplo de um XML e sua estrutura

Fonte: Autoria própria

2.2.2 NOTA FISCAL ELETRÔNICA

A NF-e (Nota Fiscal Eletrônica) é um documento de premissa apenas digital, que é emitido e armazenado eletronicamente. O intuito da RFB (Receita Federal do Brasil) com este documento digital, para fins fiscais, é documentar as operações de circulação de prestação de serviço ou de mercadorias que são transitadas entre as partes interessadas. Quanto à validade jurídica, toda NF-e que é emitida contém uma assinatura digital do remetente, que garante a autoria e a integridade, tanto do documento digital, quanto do serviço prestado, ou da mercadoria circulada (FAZENDA, 2020).
De acordo com RFB (2020), a NF-e é um sucesso e trouxe benefícios e vantagens à sociedade, como:

- Melhor intercâmbio e compartilhamento de informações entre os fiscos;
- Fortalecimento do controle e da fiscalização;
- Aumento na confiabilidade da Nota Fiscal;
- Rapidez no acesso às informações;
- Eliminação do papel;
- Redução de custos - que afeta diretamente os cofres públicos - no processo de controle das notas fiscais capturadas pela fiscalização de mercadorias em trânsito;
- Diminuição da sonegação e aumento da arrecadação.

2.3 MINERAÇÃO DE DADOS

A mineração de dados é compreendida como a descoberta de novas informações, em uma grande escala de conjunto de dados, usando regras ou padrões para orientar futuras decisões estratégias ou atividades-fim de uma empresa ou pesquisa. Para isso, técnicas são usadas nas áreas de análise estatística e de inteligência artificial, tal como combinação probabilística, regressão, aprendizado de máquina, redes neurais artificiais, entre outros (VIEIRA; FELIPE, 2001).
Os estudos em cima da área de mineração de dados permeiam logo no início dos anos 90, quando a necessidade de integração de informação digital rápida com o meio corporativo empresarial e industrial estava começando a acelerar e, em paralelo, uma das consequências

disso foram os grandes acúmulos de dados em bancos de dados corporativos. Sendo assim, as pesquisas sobre esses grandes conjuntos de dados emergiram e foram denominadas como KDD - *Knowledge Discovery in Databases*, em português: Processo de Descoberta de Conhecimento em Base de Dados (FAYYAD et al., 1996).

Colocando de uma forma mais técnica de entendimento computacional, o objetivo da mineração de dados é converter os dados em um novo conhecimento. Nesse sentido, para ocorrer a mineração de dados é necessário, previamente, a extração dos dados a serem minerados. Sendo assim, há, basicamente, dois tipos: o simples e o complexo. O conjunto de dados simples é aquele que vem de um banco de dados relacional ou de planilhas geradas por sistemas empresariais. Já o conjunto de dados complexo é aquele que necessita de outras técnicas computacionais aplicadas nessa base de dados para se obter o novo conhecimento pretendido, é nesse segundo caso que se encontra aplicações de inteligência artificial e técnicas de programação (REZENDE, 2005).

Na doutrina acadêmica, há uma divergência entre autores quanto à mineração de dados e ao KDD. Dito isso, para Fayyad et al. (1996) a mineração de dados é um processo dentro do conceito do processo de descoberta de conhecimento em base dados, como dá pra observar na Figura 2, entre os dados formados e o resultado de padrões. Por outro lado, Han e Kamber (2001), que são pesquisadores e autores de grandes contribuições acadêmicas nesta mesma linha de pesquisa, entendem que o KDD e a mineração de dados são sinônimos, em outras palavras, todo o processo de KDD é uma mineração de dados em si.

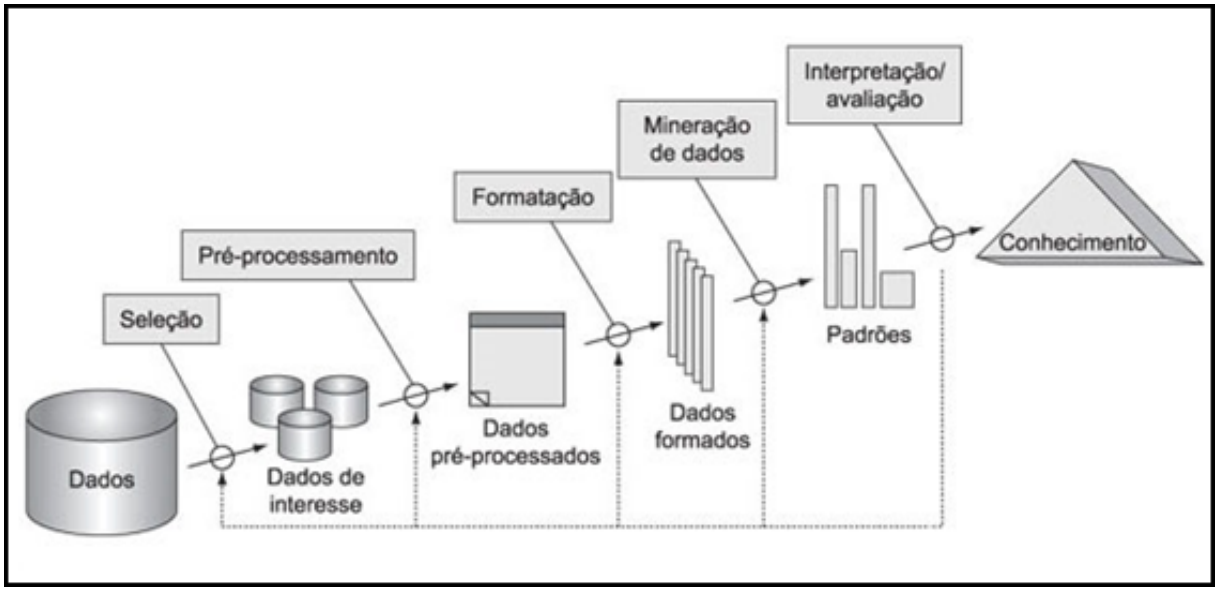

Figura 2 – Fases do processo de descoberta de conhecimento em base de dados

Fonte: Adaptado (FAYYAD et al., 1996).

Esta pesquisa seguirá a linha acadêmica de Fayyad et al. (1996), para o autor, o KDD é dividido em uma sequência de etapas, que são:

1. Seleção de dados, quando ocorre uma especialização direcionada gerando os dados de interesse;
2. Pré-processamento, quando - caso necessário - trata-se os dados, com remoção de ruídos, por exemplo;
3. Formatação, também adota uma forma de tratar os dados, mas no âmbito de como esses serão dispostos;
4. Mineração de dados, quando ocorre o processo matemático computacional por meio de algoritmo que resulta em padrões ou regras;
5. Interpretação por meio de uma avaliação e assimilação de resultados, que, por fim, proverão o novo conhecimento obtido para tomada de decisões.

A mineração de dados continuou evoluindo, o que a motivou, conforme supraticado, foi uma busca pelo conhecimento para a tomada de decisões no meio corporativo e industrial. No entanto, agora, não somente nesse meio, a mineração de dados também vem sendo muito usada nas redes sociais, que hoje dominam o mercado digital da comunicação social da população. Nesse sentido, o que impulsionou tais técnicas a serem usadas nas redes sociais foram os grandes vendedores do mercado tradicional querendo saber, por meio da mineração de dados, os interesses da população, como: o que gostam, o que querem consumir, e sobre o que discutem. Logo, cada clique, cada curtida, cada compartilhamento e cada análise de palavras-chave são de interesse dos grandes varejistas (AGGARWAL, 2015).

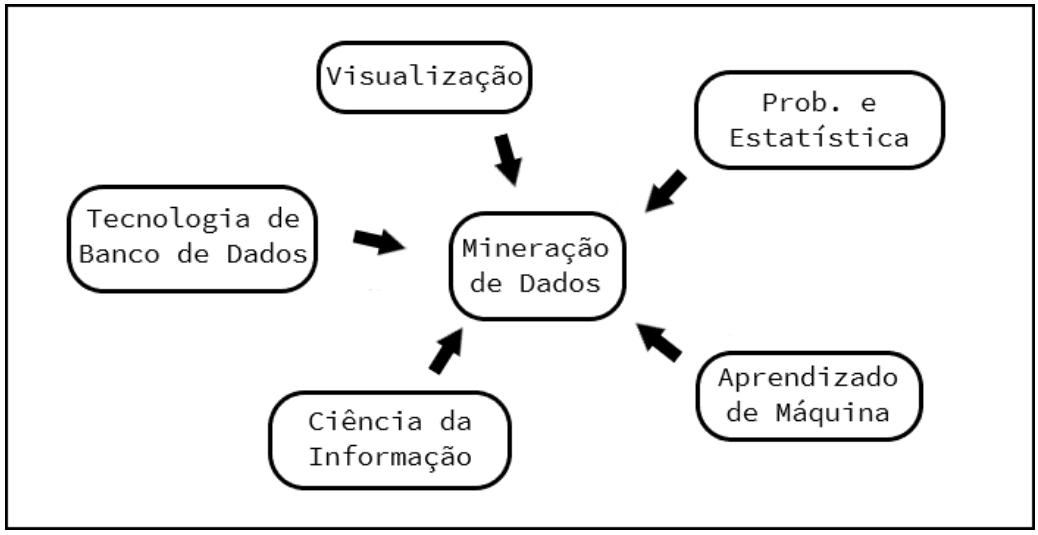

Figura 3 – Variedade de disciplinas que envolvem a mineração de dados

Fonte: Adaptado (HAN; KAMBER, 2001).

Há outros autores que concordam com essa visão de que os grandes varejistas estão mais inseridos no meio digital pelo *e-commerce* emergente durante os últimos anos, sendo assim, esse novo mercado tem trazido novos problemas e desafios para os pesquisadores e especialistas em análise de dados, resultando, dessa forma, em uma boa fomentação por parte da iniciativa privada nesta área acadêmica e, também, em profissionais com conjunto de habilidades novas e necessárias (JAPKOWICZ; STEFANOWSKI, 2016).

Somando ao exposto, a conclusão que define o caminho da mineração de dados é que se tornou algo multidisciplinar, como Han e Kamber (2001) demonstra na figura 3, sendo um tema estudado não só em academias de tecnologia, mas também, nas áreas da administração de empresas e gestão empresarial, para a tomada de decisões importantes por meio da ciência da informação, resultados de probabilidade e estatística, aprendizado de máquina, entre outros (AHLEMEYER-STUBBE; COLEMAN, 2014).

Dando enfoque na etapa 4 do KDD de Fayyad et al. (1996), a parte da mineração de dados em si, que seria o cerne desta monografia, consiste em técnicas de mineração para descobrir regras e padrões. Nesse sentido, nada melhor do que exemplificar o que essas técnicas poderiam fazer (VIEIRA; FELIPE, 2001):

Imagine o banco de dados de um sistema comercial que contenha os dados do cliente, como: nome do cliente, endereço completo, telefone, data de compra, preço do produto comprado e quantidade. Agora, tome como verdade que os produtos dessa empresa sejam produtos eletrônicos em geral, como: sistema para alarme, relógios multiesportivos, periféricos computacionais, dentre outros. Um resultado da mineração de dados para obter um novo conhecimento para os sócios desse comércio poderia se dar por:

- Regras de associação - exemplo: Sempre que o cliente comprar cabos de rede, ele também compra adaptadores RJ-45, ou quase sempre que um cliente compra um teclado pra computador, também compra um mouse;
- Padrões sequenciais - exemplo: Toda vez que o clima está mais quente e/ou o período de vendas está no verão, há um aumento de vendas de relógios multiesportivos;
- Agrupamento - exemplo: Esta regra pode trazer grupos de itens parecidos, como relógios esportivos com o *layout* arredondado, separado dos quadrados, ou, até mesmo, segmentar periféricos voltados para jogos eletrônicos, separados dos que são mais executivos e tradicionais. Tudo isso de acordo com as compras já realizadas e grupos de susposto interesse que o algoritmo encontrou;
- Árvores de classificação e decisão - exemplo: Um cliente que mora em um bairro considerado com alto índice de roubo, tende a comprar um sistema de alarme comparado a um cliente que mora em bairro pacífico.

2.3.1 TÉCNICAS DE MINERAÇÃO

As técnicas de mineração de dados mais usuais são as de Associação, Classificação e Agrupamento. Todas seguem a perspectiva de aprendizado de máquina, podendo ser um aprendizado supervisionado, quando o homem atua auxiliando o algoritmo, ou não supervisionado, que é quando os resultados vêm de padrões que o algoritmo encontrou sozinho (CAMILO; SILVA, 2009; LAROSE; LAROSE, 2015).
A linha de pesquisa deste trabalho utilizou aplicações de técnica de associação.

2.3.1.1 ASSOCIAÇÃO

Esta técnica é utilizada para encontrar similaridades entre um conjunto de objetos em uma tupla (para recordar o que é uma tupla, ver subseção 2.1), também conhecida como instância, no caso de mineração de dados. Dito isso, nos últimos anos ela têm sido utilizada para analisar comportamento de consumo e compra dos clientes, nesse sentido, em sistemas de aprendizagem de máquina não supervisionados, a regra de associação para mineração de dados é uma das tarefas mais comuns e utilizadas (RAO; GUPTA, 2012).
Tradicionalmente, as buscas por padrões pela regra de associação acontecem por meio de transações armazenadas em um conjunto de dados, arquivo que contém essas transações onde a regra de associação acontecerá, usualmente por meio de software especialista (VASCONELOS; CARVALHO, 2004).
Exemplificando o funcionamento básico da regra de associação: Dado um conjunto de dados que representam as vendas realizadas por um comércio, caso dois conjuntos, itens, X e Y, fossem comprados juntos, a representação de associação formaria a seguinte expressão: $\{X\} \Rightarrow \{Y\}$ (leia-se ***X* então *Y***). Sendo assim, de forma intuitiva, compreende-se que para todo X comprado, tende a ter o conjunto item Y junto (VIEIRA; FELIPE, 2001; VASCONELOS; CARVALHO, 2004).
Em complemento ao que foi dito anteriormente, essa regra não mostra estritamente só os conjuntos que tiveram 100% de associação. Nesse sentido, a cada regra encontrada, dois fatores principais são levados em consideração: o suporte e a confiança (VIEIRA; FELIPE, 2001; VASCONELOS; CARVALHO, 2004; AGGARWAL, 2015).

- **Suporte**: indica a incidência de vendas/conjuntos que aparecem os itens *X* e *Y* juntos, comparados ao total de registros. Em outras palavras, o suporte é a porcentagem de instâncias que tiveram todos os itens da união de *X* e *Y* (VIEIRA; FELIPE, 2001; AGRAWAL; SRIKANT, 1994);
- **Confiança**: indica a incidência de vendas/conjuntos que aparecem os itens *X* e *Y* juntos sobre o total de registro do conjunto *X* (VIEIRA; FELIPE, 2001).

Sendo assim, a meta da mineração de dados utilizando regras de associação é descobrir todas as regras que superam os valores mínimos, de suporte e confiança, previamente configuradas pelo usuário (VIEIRA; FELIPE, 2001).

A seguir, encontra-se um exemplo prático (Figura 4) do que foi explicado anteriormente, com base em cestas de supermercado, contendo itens comprados em vendas distintas.

VENDA	HORÁRIO	ITENS DA CESTA
0002	08h21	banana, leite, queijo
0019	10h41	banana, queijo
0055	12h09	banana, pão
0101	17h33	leite, carne, tomate
0197	20h58	uva, refrigerante, alho

Figura 4 – Amostra aleatória de vendas de um supermercado

Fonte: Adaptado (VIEIRA; FELIPE, 2001).

Com base na figura acima, utilizando a teoria explicada nesta seção, escolhe-se a expressão $\{banana\} \Rightarrow \{queijo\}$, e esta regra tem:

- **Suporte**: 40%, afinal $\{banana\} \Rightarrow \{queijo\}$ aparece em 2/5 do conjunto total de vendas;
- **Confiança**: 66,66%, pois, como foi visto, a conta da confiança é feita em cima da quantidade de ocorrência do conjunto de item do lado esquerdo da regra. Em outras palavras, de 3 vendas que contêm o item *banana*, 2 vendas continham o item *queijo*.

Outro critério importante de resultado do pós-procedimento é quanto à soma do comprimento das tuplas resultantes de associação, isto é, o resultado entre quantidade de itens do conjunto de *X*, do lado esquerdo da regra, com a quantidade de itens do conjunto *Y*, que está do lado direito da mesma. Dito isso, o comprimento L(k) é o que determina o tamanho da regra resultante. Tal critério é considerado importante na análise, pois, em primeira análise do algoritmo são gerados conjuntos de tamanho L(1), passando por todos os itens. Logo em

seguida são calculados os conjuntos que resultarão em regras de tamanho L(2), que, por sua vez, têm os itens de L(1) contido neste novo. Nesse sentido, o algoritmo continua até não ter mais subconjuntos de associação a se fazer (AGRAWAL; SRIKANT, 1994).
Tomando como exemplo a figura 4, a linha 1, 4 e 5 seriam de comprimento L(3), enquanto as demais seriam de L(2).
Ainda para Agrawal e Srikant (1994), este critério mencionado do comprimento da regra L(k) é importante, uma vez que aumenta o nível de detalhamento da associação entre os itens e traz uma confiança maior numa associação distinta com muitos itens. Logo, são regras com o comprimento L(k) maiores, que podem significar resultados únicos e satisfatórios.

2.3.1.2 EVOLUÇÃO DO ALGORITMO DE ASSOCIAÇÃO

Por fim, vale ressaltar que, ao longo dos anos, os algoritmos de regras de associação foram sofrendo mudanças e melhorando, dos mais clássicos, como o **Apriori**, o **Fuzzy Apriori** e o **Eclat**, citados e utilizados por Borgelt (2003), e Vasconelos e Carvalho (2004), aos mais atuais, como o **Enhanced Apriori**, que teve sua primeira versão apresentada por Al-Maolegi e Arkok (2014) e, logo em seguida, Tirumalasetty et al. (2015) apresentou melhorias, deixando o mesmo nome para o algoritmo. Recentemente, Ansari et al. (2018) apresentou o **TFI-Apriori** que, por sua vez, apresenta uma requisição de memória menor do que os algoritmos tradicionais baseados em Apriori, que permite a reutilização de fluxos de processamento existentes. Ou seja, a aplicação de um algoritmo Apriori em tempo real em um *big data* constantemente sendo alimentado por novos dados, assim, utilizando os dados mais novos, somados aos mais antigos já analisados uma ou mais vezes para novos resultados.

2.4 SOFTWARES DE MINERAÇÃO DE DADOS

Ao mesmo tempo em que as pesquisas em mineração de dados se desenvolviam, muitas empresas começaram a investir na automatização da utilização de algoritmos de mineração de dados e a oferecer esses serviços. Começaram a desenvolver ferramentas computacionais onde poderiam alimentar e atualizar constantemente as mesmas, ao passo que os algoritmos

melhoravam e a necessidade de uma centralização dos resultados de informação fossem cada vez mais imprescindíveis (MIKUT; REISCHL, 2011; JAPKOWICZ; STEFANOWSKI, 2016). Como exemplos de softwares para mineração de dados, tem o Waikato Environment for Knowledge Analysis, conhecido como WEKA, que começou em 1994, usando, inicialmente, uma biblioteca na linguagem C++, porém reconstruída em Java e até hoje se mantem assim. WEKA é uma ferramenta desenvolvida na Universidade de Waikato, na Nova Zelândia (MIKUT; REISCHL, 2011; WEKA, 2020).
Empresas também começaram a apresentar sua ferramentas, não gratuitas, como a Oracle Data Mining, em 2002, a SAS EnterpriseMiner, em 2007 e o IBM SPSS Modeler, em 2009, com recursos voltados inteiramente para o mundo corporativo (MIKUT; REISCHL, 2011).

2.4.1 WEKA

Para o processo de mineração de dados, foi escolhido o software de inteligência artificial e mineração de dados WEKA, que, por ser de uso gratuito e de uma variedade de algoritmos para mineração e análise de dados, atendeu às necessidades desta monografia (WEKA, 2020). Sendo assim, na Figura 5 se encontra a tela inicial, e nela há as opções de aplicações existentes na ferramenta. Nesse sentido, para esta pesquisa, a aplicação utilizada foi a ***Explorer***. Nesta, encontra-se fases de pré-processamento para arrumar alguns detalhes dos dados, a escolha da regra de mineração e seus algoritmos que serão fortemente detalhados no capítulo 3, referente à Materiais e Métodos. (WEKA, 2020).

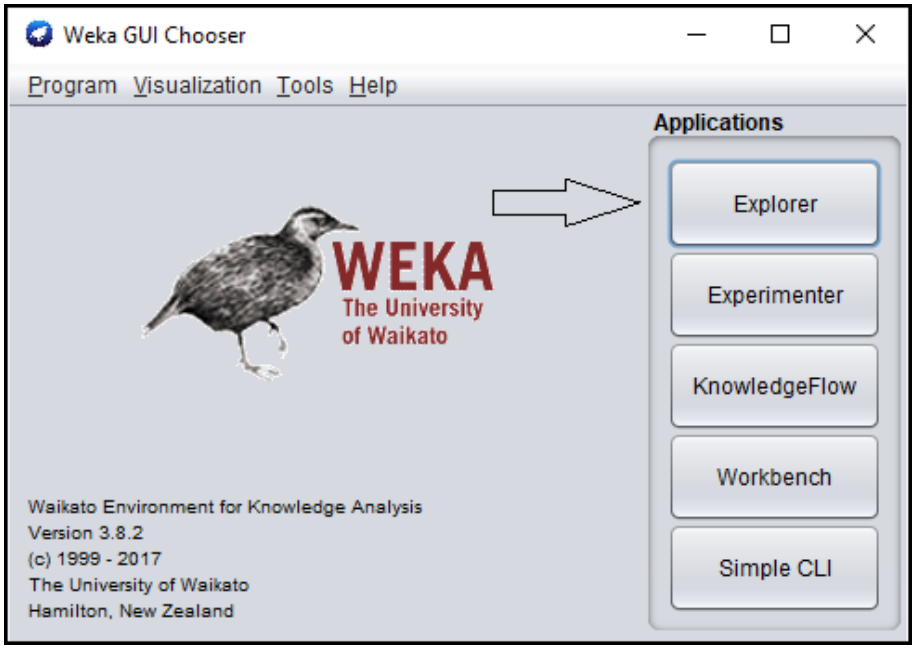

Figura 5 – Interface gráfica inicial do software de mineração de dados WEKA

Fonte: Software WEKA.

3 MATERIAIS E MÉTODOS

Após realizar pesquisas acerca dos fundamentos computacionais necessários para este trabalho, foram levantados softwares e ferramentas para auxiliar nas partes práticas do estudo, tal como o desenvolvimento de *script* para captura, tratamento de dados e informações estatísticas em meio à leitura do conjunto de dados.
Esse capítulo abordará quais dados foram utilizados, quais modificações foram feitas e como foram minerados os mesmos, tal como configurações e técnicas utilizadas em softwares auxiliares.

3.1 SISTEMÁTICA CRONOLÓGICA DA PESQUISA

Antes de iniciar a descrição do que envolve esta pesquisa, foi elaborado um fluxograma que demonstra, do início ao fim, todas as suas etapas.
Vale informar que todas as descrições, a partir daqui, nesta seção 3.1, acompanhará a imagem do fluxograma na Figura 6, na intenção de explicar a imagem.

3.1.1 ETAPA 1

Dando início pela **Etapa 1**, fase em que foi avaliada a viabilidade de poder fazer a pesquisa na área pretendida - comportamento de compras durante período pandêmico -, pois foi preciso um acesso remoto para a obtenção das NF-e que continham os dados necessários para criar, futuramente, o conjunto de dados a ser minerado.
Como será explicado nas próximas seções deste trabalho, dentro de um arquivo XML que representa a NF-e contêm muitos dados que para esta pesquisa não eram interessantes, porém,

isso representava um tamanho de arquivo muito pesado para transferência em um computador local. Dito isso, foi necessário em um total de 26 dias acessando remotamente este servidor para conseguir ter todos os períodos necessários.
Por fim, a etapa se encerra com o estudo e a compreensão de como era a estrutura de um XML e qual linguagem de programação ajudaria na captura dos dados úteis dentro de cada arquivo.

3.1.2 ETAPA 2 E ETAPA 3

Ao fim do estudo da estrutura do XML que compunha os arquivos, deu-se início à **Etapa 2**, na qual consistiu no desenvolvimento de dois módulos: o primeiro de captura de dados entre as *tags* dos XML que alimentavam uma lista de retorno para o segundo módulo, que, ao receber essa lista com os dados, fazia o tratamento e a formatação dos mesmos para gerar um arquivo de saída em CSV, ambos os módulos fazem parte de um mesmo programa desenvolvido que foi executado para todos os períodos a serem pesquisados e minerados posteriormente.
Esta subseção teve a necessidade de trazer a explicação da **Etapa 3**. A terceira parte do fluxograma, como pode ser observado na imagem em que há uma seta vermelha que indica ciclo de retorno, que está diretamente ligada a testes de alimentação do software WEKA com o retorno, caso necessário, no meio da segunda etapa.
Dito isso, levou em torno de duas semanas para compreender como inserir conjuntos de dados em CSV, sendo que o software de mineração de dados aceita, nativamente, outro formato de arquivo. Sendo assim, toda vez que houve divergência na formatação dos dados, voltava-se ao modulo de formatação e criação de arquivo no programa desenvolvido.
Por fim, ao notar que todos os períodos a serem minerados foram lidos corretamente, deu-se o fim da terceira etapa.

3.1.3 ETAPA 4, ETAPA 5 E ETAPA 6

Pode-se dizer que as etapas seguintes são as que representam as atividades-fim desta pesquisa. Afinal, mesmo que se de alguma forma os conjuntos de dados tivessem sido obtidos na formatação correta para mineração de dados, a partir dessas etapas o processo seria o mesmo. Tendo em vista tal conjuntiva, tal como as etapas 2 e 3, a 4, 5 e 6 estão conectadas em ciclos

também, como pode ser percebido pela seta amarela. Porém, ao contrário das duas etapas anteriores, que foram de verificação e formatação, estas tiveram um ciclo obrigatório, pois para cada vez ocorria a **Etapa 4**, iniciava-se, também, a análise sobre um novo período (explicado na seção 3.6) que, em sequência, na **Etapa 5**, escolheria as configurações para o algoritmo e era finalizado com a execução da mineração de dados e análise dos resultados obtidos na **Etapa 6**.

3.1.4 DA ETAPA 7

Finalizando o fluxograma, encontra-se a **Etapa 7**, que consistiu em escrever relatórios dos resultados obtidos e demonstrá-los ao analista responsável, que diria quão pertinentes seriam aquelas informações obtidas.

3.2 O CONJUNTO DE DADOS

Em busca dos objetivos propostos, a empresa colaboradora contribuiu com um acervo de dados que consiste nas vendas de um período correspondente de 01 de abril de 2020 a 30 de junho de 2020, sendo assim, essas são vendas consistidas durante o período pandêmico. Em paralelo, para fins de contraprovas futuras quanto ao comportamento nas compras realizadas fora de período pandêmico, também foram disponibilizados 3 meses diferentes do ano de 2019: Setembro, Outubro e Novembro.
As vendas são registradas em Notas Fiscais Eletrônicas no formato XML, separadas por diretórios diários, em que cada dia contem uma média aproximada de 2750 NFs-e, referente a cada venda. Uma NF-e possui uma considerável quantidade de informação separada por *tags*, tal como a estrutura de um XML, podendo-se notar na figura 7, das quais as principais, segundos os especialistas, para a elaboração deste estudo, são:

- **nNF**: Número da NF-e perante a RFB;
- **dhEmi**: Data e hora da venda, pertinente para determinar o dia da semana e também o turno em que ocorreu a venda;
- **xProd**: Nome do produto vendido;
- **NCM**: Nomenclatura Comum do Mercosul, código importante para identificação por

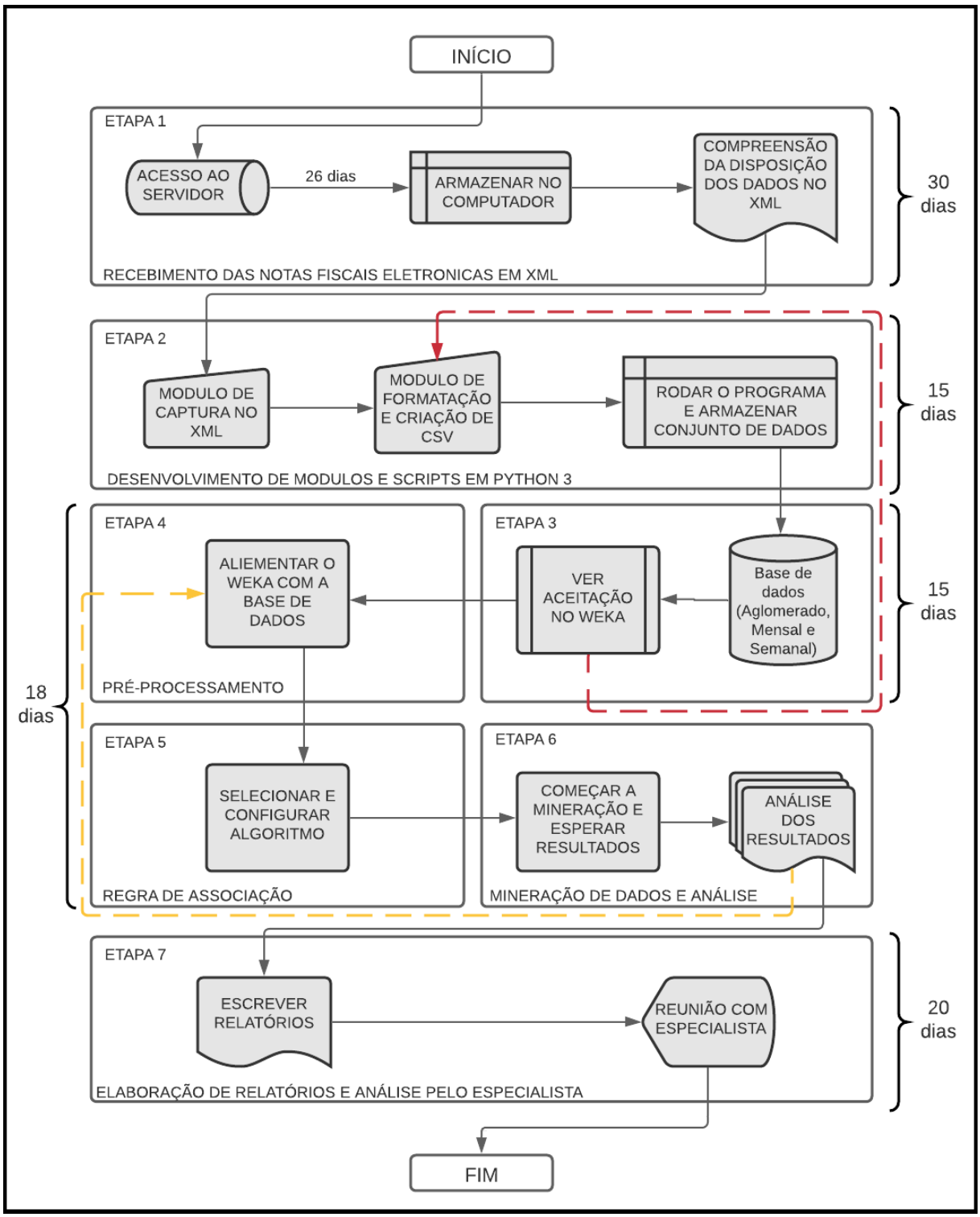

Figura 6 – Fluxograma das etapas desta pesquisa

Fonte: Autoria própria.

uma descrição comum, ignorando marca ou qualquer outra referência.

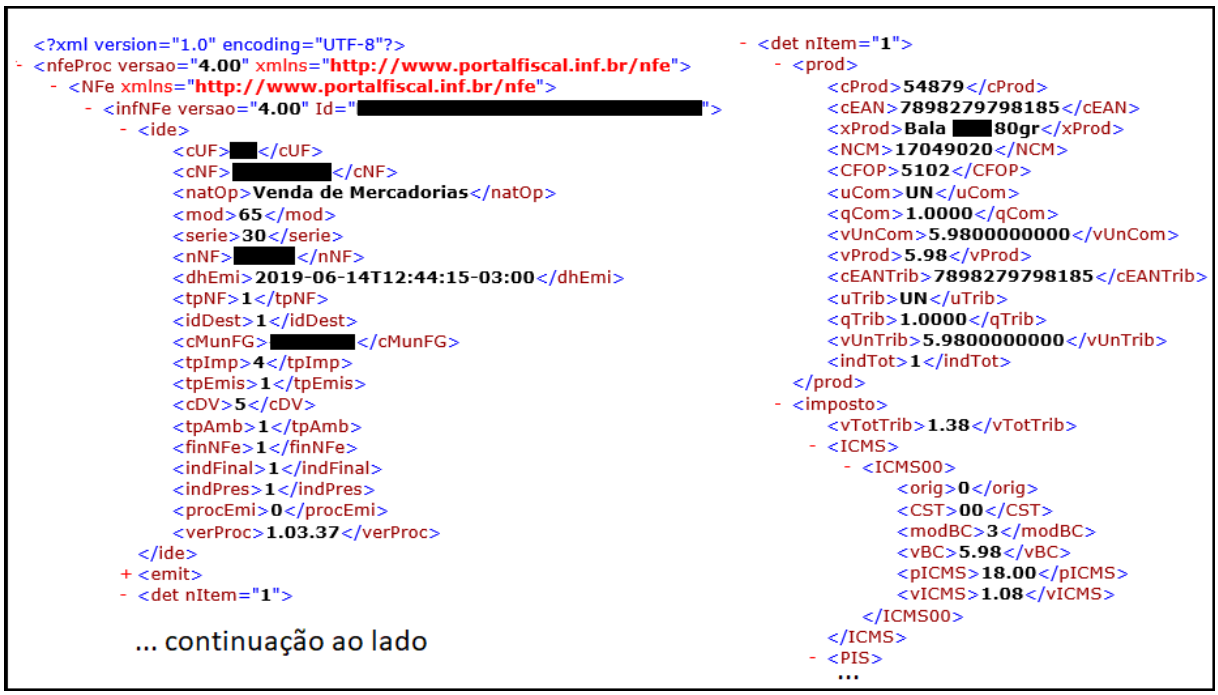

Figura 7 – Estrutura XML de uma nf-e

Fonte: Nota Fiscal Eletrônica.

3.3 PRÉ-PROCESSAMENTO DO CONJUNTO DE DADOS

Nesta fase, foi analisada a quantidade de arquivos, quais os métodos que seriam utilizados para a captura dos dados relevantes de cada arquivo e a quantidade em tempo para esses tratamentos de pré-processamento dos dados.

De início, foi adotada a linguagem Python, em sua versão 3.7, para ser a ferramenta “meio” para todo o período de pré-processamento. Usando uma biblioteca nativa da linguagem, a **xml.etree.ElementTree** em conjunto a outra biblioteca nativa chamada **os**, foi possível desenvolver um código para percorrer os diretórios que estavam divididos em dias por mês, acessando cada arquivo em formato XML, conforme código-fonte deste módulo, descrito no Apêndice A.

Por fim, o resultado final deste pré-processamento foi o desenvolvimento de um módulo de formatação dos dados capturados (Apêndice B) em um arquivo em formato texto, com extensão CSV, para que o WEKA pudesse lê-lo, onde a primeira linha é referente às classes: **NFE**, **DIASEMANA**, **TURNO**, **P-NCM** do primeiro produto encontrado no conjunto de dados até **P-NCM** do último produto distinto do mesmo conjunto.

Já da segunda linha em diante, são os valores de cada classe representando cada venda: **Nº da**

Nota Fiscal Eletrônica, **Dia da semana** (Domingo a Sábado), **Turno em que ocorreu a venda** (Manhã, Tarde e Noite) e, por fim, o campo que determina se teve ocorrência da classe **P-NCM**, determinado por dois valores: **"t"** para verdadeiro, ou **"?"** para falso. É um padrão do software WEKA seguir esses dois valores para verdadeiro e falso em ocorrências.
Para melhor compreensão, pode-se observar a Tabela 1, ou a imagem deste documento em texto na Figura 8.

Tabela 1 – Demonstrativo da separação por colunas das classes e seus valores

NF-e	DIASEMANA	TURNO	P-19059000	P-02102000	P-02071300	P-19059090	P-22021000	...	P-NCM Último Item
00001	SABADO	TARDE	t	t	?	?	t	...	t
00002	SEXTA	NOITE	?	t	t	?	t	...	?
00003	QUARTA	MANHA	?	t	?	t	?	...	?
00004	QUINTA	MANHA	t	t	t	?	?	...	t
00005	TERCA	TARDE	?	t	t	?	t	...	?
00006	SEGUNDA	MANHA	t	?	t	t	t	...	t
00007	DOMINGO	NOITE	t	?	t	?	?	...	t
...	...	...	...	...	...	...	...	...	...
00100	DOMINGO	NOITE	t	?	t	?	t	...	?

Fonte: Autoria própria.

Finalmente, após todo o processo de captura das informações e criação de arquivos CSV preparados para o processo de mineração de dados, os arquivos foram separados em três formas de períodos distintas, todos segmentados de igual forma para o período de pandemia e de não pandemia:

- Aglomerado de todas as vendas de 3 meses respectivos;
- Mensal separado dos respectivos períodos;
- Por dia da semana separado dos respectivos períodos.

```
NFE, DIASEMANA, TURNO, P-19059000, P-02102000, P-02071300, P-19059090, P-22021000, ... P-NCM n-ITEM
000001, SABADO, TARDE, t, t, ?, ?, t, ..., t
000002, SEXTA, NOITE, ?, t, t, ?, t, ..., ?
000003, QUARTA, MANHA, ?, t, ?, ?, t, ..., ?
000004, QUINTA, MANHA, t, t, ?, ?, t, ..., t
000005, TERCA, TARDE, t, t, ?, ?, t, ..., ?
000006, SEGUNDA, MANHA, ?, t, ?, t, t, ..., t
000007, DOMINGO, NOITE, t, t, ?, ?, t, ..., t
...
...
000100, DOMINGO, NOITE, t, ?, t, ?, t, ..., ?
```

Figura 8 – Estrutura adaptada de um arquivo texto CSV para leitura pelo algoritmo

Fonte: Autoria própria.

3.4 O PROCESSO DE MINERAÇÃO DE DADOS

Para o carregamento de dados no WEKA, dentro da aplicação do *Explorer*, encontra-se a aba de pré-processamento de dados. Além de ser o local onde se faz o carregamento dos dados, também é o caso do usuário que queira fazer mais alguns ajustes nos dados antes de serem processados. Dito isso, ocorreram algumas modificações nos dados antes de se iniciar o processo de mineração, como a retirada do atributo correspondente ao dia da semana, pois esse apenas serviu para identificar os dias respectivos das vendas, mas não para ser um atributo em meio ao processo de mineração de dados.

Assim, para todos os casos que serão apresentados, foi decidido que não haveria necessidade de utilizar a classe **TURNO** e, também, que a classe **NFe** não era inerente para os objetivos finais dessa monografia. Sendo assim, tais classes foram retiradas nesta fase. Por fim, ocorreu a retirada da classe **DIASEMANA**, pois esta apenas foi capturada para futura separação por dias da semana do conjunto de dados, como foi dito na seção 3.3.

Para que o carregamento dos dados acontecesse, era necessário selecionar o botão de *Open File* (Figura 9, selecionar o tipo de arquivo que seria importado para o programa e selecionar o arquivo no local em que ele estava originalmente. Uma vez reconhecido como CSV, o WEKA necessitará entender qual é o separador de atributos e itens do arquivo. Por padrão do software, foi utilizada a vírgula, que foi o mesmo seguido para o conjunto de dados a ser analisado.

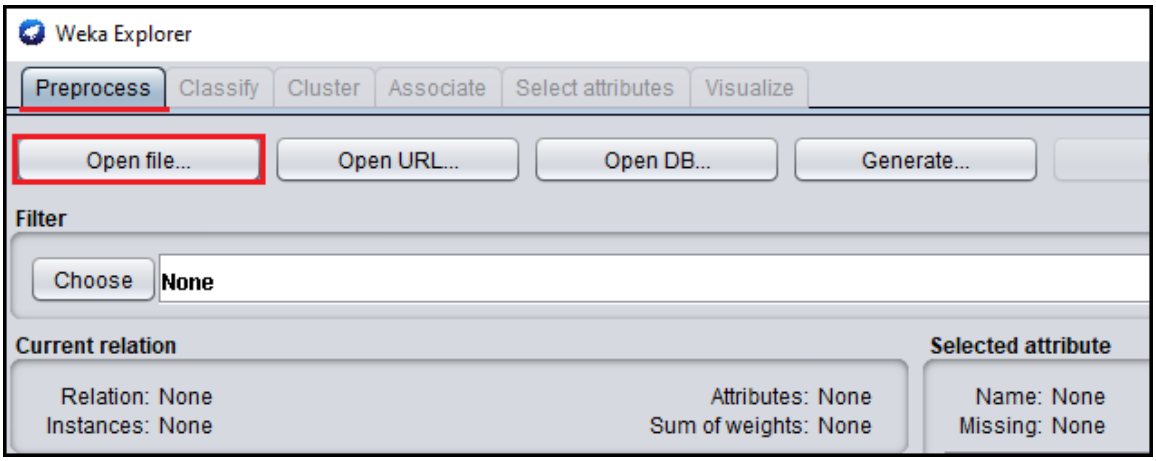

Figura 9 – Tela parcial da aba *preprocess* do WEKA

Fonte: Software WEKA.

3.5 ESCOLHA E CONFIGURAÇÃO DO ALGORITMO

Uma vez realizado o carregamento dos dados para dentro do WEKA, as abas que antes estavam inacessíveis (Classify, Cluster, Associate, etc., na Figura 9), ficaram disponíveis para seleção. Este trabalho consistiu em fazer uma análise de comportamento em cima das vendas de um supermercado, por meio da associação entre produtos de uma mesma cesta de compras, logo, a aba a ser utilizada dentro do software foi o ***Associate***.

Na aba ***Associate***, prioritariamente nesta fase, é utilizado o recurso *Choose*, onde foi escolhido qual algoritmo de associação seria usado, porém, por padrão, já vem o Apriori, que, no caso, é o que foi utilizado nesta pesquisa.

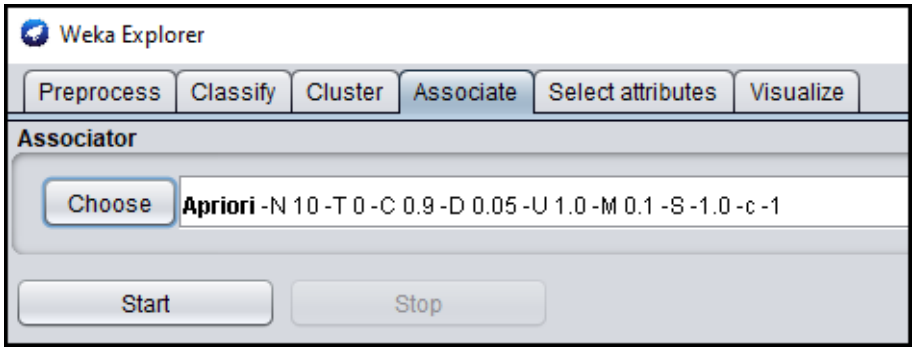

Figura 10 – Tela parcial da aba *associate* do WEKA

Fonte: Software WEKA.

Ainda na mesma aba, foi feita a configuração do algoritmo escolhido - clicando duas vezes em cima da parte branca na Figura 10 - na tela de configuração apropriada (Figura 11)

Nesta aba, o que se considerou interessante para a pesquisa foram os campos: o *lowerBoundMinSupport*, que é o campo onde se estabelece a medida de suporte pretendida, o *minMetric*, que é o campo em que se estabelece a medida de confiança pretendida e, por fim, o campo *numRules*, que é número de regras máximas para retornar nos resultados após o processo de mineração de dados.

Dada as configurações corretas, para começar o processo de mineração de dados, utilizando o algoritimo Apriori, basta clicar em *Start* e esperar o resultado.

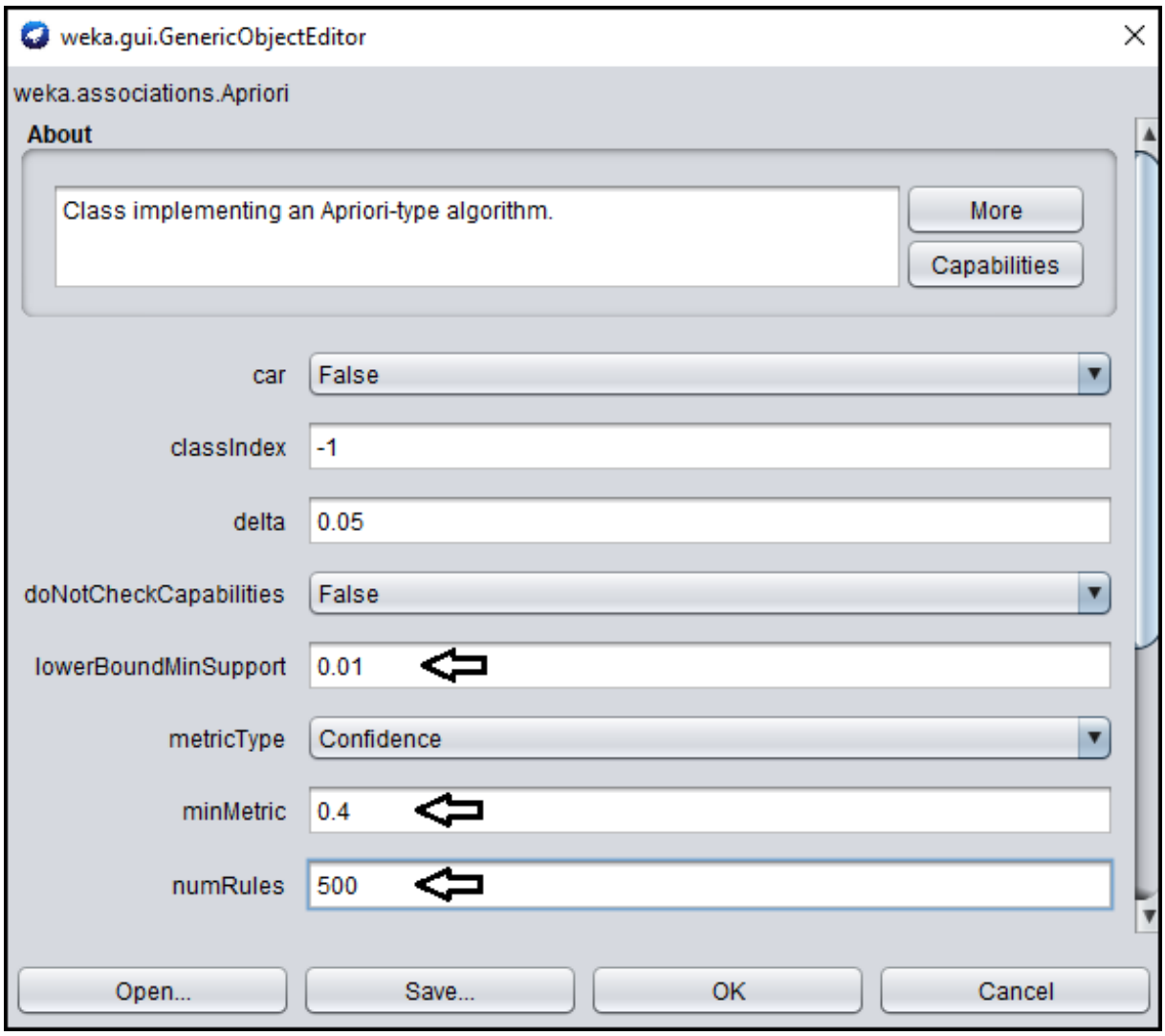

Figura 11 – Configuração do algoritmo apriori

Fonte: Software WEKA.

3.6 DAS CONFIGURAÇÕES PARA CADA PERÍODO

Conforme supracitado (subseção 3.3), a pesquisa foi segmentada em três períodos diferentes. Sendo assim, cada um deles teve configurações diferentes por motivos diversos, como tamanho do conjunto de dados para cada período e, também, a testagem de várias configurações, buscando o melhor resultado possível.

3.6.1 AGLOMERADO DOS TRÊS MESES

Primeiramente, é preciso levantar a quantidade de instâncias que representam as vendas, e a quantidade de produtos diferentes que foram processados durante a mineração. Sendo assim, segue:

- **Período pandêmico em 2020 (Abril, Maio e Junho):**
  196.267 instâncias e 652 produtos diferentes;
- **Período não pandêmico em 2019 (Setembro, Outubro e Novembro):**
  253.718 instâncias e 659 produtos diferentes;

As configurações usadas na mineração de dados para ambos os conjuntos de dados foram:

- Medida de suporte mínimo de 1%, ou seja, os produtos deveriam ser vendidos juntos no mínimo 1.963 vezes, em um total de 196.267 vendas. Sendo assim, abaixo disso foi considerado insignificante pelo algoritmo;
- Confiança mínima de 40%, isso é, a chance de tais produtos incidirem na venda de outro;
- Estaticamente foi escolhido o numero de 2.000 regras como teto máximo de amostragem de regras consideradas significativas pelo algoritmo, para termos uma margem de queda na medida de confiança boa para ser analisada.

Vale ser dito que outras configurações foram usadas no mesmo conjunto de dados, variando o suporte mínimo entre 0.25% e 2%. A configuração de 1% foi dada como a melhor para análise, pois apresentou detalhamento e variedade de produtos similar a um suporte menor, diminuindo a incidência miníma de venda entre os mesmos, porém, com uma queda na medida de confiança tão boa quanto de suportes mínimos superiores a 1%.

3.6.2 SEGMENTAÇÃO MENSAL

A segmentação de todo o período pandêmico em meses teve como objetivo descobrir se houve uma diferença comportamental nas compras sob uma ótica mensal entre esses meses somente, assim, não havendo comparação com 2019.

Primeiramente, é necessário levantar que foram 66.530 instâncias para o mês de abril, 65.940 instâncias para mês de maio e 63.797 para o mês de junho, representando a quantidade de vendas e todos com um montante médio de 603 produtos diferentes minerados pelo algoritmo

Apriori.

As configurações usadas na mineração deste conjunto de dados foram:

- Medida de suporte mínimo de 1%, ou seja, os produtos deveriam ser vendidos juntos, aproximadamente, no mínimo 649 vezes num total de instâncias respectivo de cada um dos meses. Sendo assim, abaixo disso foi considerado insignificante pelo algoritmo;
- Confiança mínima de 40%, isso é, a chance de tais produtos incidirem na venda de outro;
- Estaticamente foi escolhido o numero de 1000, diferente do aglomerado mensal (subseção 3.6), pelo conjunto de dados ser menor, regras como teto máximo de amostragem de regras consideradas significativas pelo algoritmo, para ter uma margem de queda na medida de confiança boa para ser analisada.

Também vale ser dito que outras configurações foram usadas, tal como no aglomerado de todo o período pandêmico, variando o suporte mínimo entre 0.25% e 2%. A configuração de 1% foi dada como a melhor para análise, pois apresentou detalhamento e variedade de produtos similares a um suporte menor, diminuindo a incidência mínima de venda entre os mesmos, porém com uma queda de taxa de confiança tão boa quanto de suportes mínimos superiores a 1%. Era de se esperar, afinal, os meses nada mais são que segmentações do aglomerado integral.

3.6.3 POR DIA DA SEMANA

Já as configurações realizadas para a separação por dia da semana, como foi realizada uma análise comparativa com 2019, este também, tal como o aglomerado dos três meses (subseção 3.6), trará de ambos os anos.

Sendo assim, segue tais informações:

- **Domingo/2019**: 32.441 instâncias como quantidade total de vendas, e um montante de 593 produtos diferentes minerados pelo Apriori;
- **Domingo/2020**: 22.496 instâncias e 560 produtos diferentes;
- **Segunda/2019**: 38.965 instâncias e 606 produtos diferentes;
- **Segunda/2020**: 28.742 instâncias e 583 produtos diferentes;
- **Terça/2019**: 35.071 instâncias e 586 produtos diferentes;
- **Terça/2020**: 29.057 instâncias e 585 produtos diferentes;
- **Quarta/2019**: 36.838 instâncias e 585 produtos diferentes;

- **Quarta/2020**: 29.869 instâncias e 590 produtos diferentes;
- **Quinta/2019**: 36.097 instâncias e 597 produtos diferentes;
- **Quinta/2020**: 26.576 instâncias e 582 produtos diferentes;
- **Sexta/2019**: 37.239 instâncias e 579 produtos diferentes;
- **Sexta/2020**: 26.094 instâncias e 582 produtos diferentes;
- **Sábado/2019**: 38.965 instâncias e 606 produtos diferentes;
- **Sábado/2020**: 33.433 instâncias e 600 produtos diferentes.

As configurações usadas na mineração de dados para ambos os períodos, tal como no aglomerado dos três meses, foram os mesmo, para manter o maior grau de fidedignidade do trabalho:

- Medida de suporte mínimo de 0.50% para todos dias de semana, ou seja, os produtos deveriam ser vendidos juntos, em média, no mínimo 141 vezes num total de, em média, 28 mil vendas pra cada semana. Sendo assim, abaixo disso foi considerado insignificante pelo algoritmo;
- Confiança mínima de 40%, isso é, a chance de tais produtos incidirem na venda de outro;
- Estaticamente, foi escolhido o número de 500, diferente do aglomerado mensal (subseção 3.6) e da segmentação por mês (subseção 3.6) pelo conjunto de dados ser menor ainda, regras como teto máximo de amostragem de regras consideradas significativas pelo algoritmo, para termos uma margem de queda na medida de confiança boa para ser analisada.

Vale ser dito que outras configurações foram usadas no mesmo conjunto de dados, variando o suporte mínimo entre 0.25% e 2%. A configuração de 1% foi dada como a melhor para análise, pois apresentou detalhamento e variedade de produtos similares a um suporte menor, diminuindo a incidência mínima de venda entre os mesmos, no entanto, com uma queda de taxa de confiança tão boa quanto de suportes mínimos superiores a 1%.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, são dispostos todos os resultados obtidos, oriundos dos processos relatados anteriormente no capítulo 3, tal como, as discussões sobre a relevância de tais resultados, se esses foram novidades como um todo, e ainda, se foram pertinentes a ponto de serem usados em estratégias futuras.

4.1 DOS RESULTADOS DO TODO O PERÍODO

Nesta sessão, encontram-se duas análises distintas, por necessidade de uma remoção de regras que continham certa categoria de produtos em abundância, assim, prejudicando possíveis regras relevantes que ficariam menos visíveis nas análises.

4.1.1 ANÁLISE DO TODO

Em primeira análise, foi observada a quantidade de regras em cada conjunto de L(K) (Teoria explicada na subseção 2.3), no aglomerado de três meses em período pandêmico. Diante disso, a quantidade de regras de comprimento L(2), ou seja, que um produto incide na compra de outro, foi de 551 de um total de 2.000 regras. Nesse sentido, é comum e geralmente é o comprimento de K, onde traz resultados não relevantes. Afinal, a associação de dois produtos em grande parte do tempo não traz informações novas ou que permita tirar conclusões satisfatórias.

Diante do exposto, ao dar continuidade na análise de L(K), houve uma quantidade significativa em regras de comprimento L(3), produtos sendo vendidos juntos com a medida de confiança alta (Vide regras: 1, 7, 8, 9, 10, 11, 12, 13, 14, 15 na Tabela 2). Em outras palavras, das 15

regras com maior medida de confiança de venda conjunta, 10 atenderam ao comprimento L(3). E ainda, neste caso em específico, a quantidade de produtos da categoria hortifruti, onde dessas 10, 9 foram da categoria mencionada (Regra: 9).

Não obstante, foram apresentadas 12 regras que corresponderam à L(4), em outras palavras, foram 12 associações, onde produtos tiveram frequência alta de serem vendidos juntos. O mais interessante foi a posição em que tais regras se encontraram no resultado, no qual das 5 melhores - todas com medida acima de 72% de confiança -, 4 regras eram de L(4) e todos da categoria de hortifruti.

Tabela 2 – Resultado do aglomerado total em pandemia

R	L(*K*)	NOMINAL	QUANTITATIVO	CONFIANÇA
01	L(3)	MAÇÃS, MAMÕES ⇒ BANANAS	3087 ⇒ 2247	73%
02	L(4)	LARANJAS, TOMATES, TANGERINAS ⇒ BANANAS	3239 ⇒ 2350	73%
03	L(4)	CEBOLAS, BANANAS, CENOURAS ⇒ TOMATES	2725 ⇒ 1672	72%
04	L(4)	CEBOLAS, BANANAS, BATATAS ⇒ TOMATES	3088 ⇒ 2211	73%
05	L(3)	CENOURAS, MAÇÃS ⇒ BANANAS	2749 ⇒ 1968	72%
06	L(4)	LARANJA, TOMATES, CEBOLAS ⇒ BANANAS	2826 ⇒ 1993	71%
07	L(3)	LARANJA, MAMÕES ⇒ BANANAS	4009 ⇒ 2823	70%
08	L(3)	CEBOLAS, MAMÕES ⇒ BANANAS	3073 ⇒ 2158	70%
09	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, BOLACHAS ⇒ PROD. PANIFICAÇÃO	3825 ⇒ 2669	70%
10	L(3)	BATATAS, MAÇÃS ⇒ BANANAS	2996 ⇒ 2082	69%
11	L(3)	LARANJAS, MAÇÃS ⇒ BANANAS	4300 ⇒ 2985	69%
12	L(3)	LARANJAS, CENOURAS ⇒ BANANAS	3604 ⇒ 2500	69%
13	L(3)	MAMÕES, TANGERINAS ⇒ BANANAS	4489 ⇒ 3109	69%
14	L(3)	MAÇÃS, TANGERINAS ⇒ BANANAS	4369 ⇒ 3015	69%
15	L(4)	LARANJA, CEBOLAS, BANANAS ⇒ TOMATES	2910 ⇒ 1993	68%

Fonte: Autoria própria.

Tabela 3 – Resultado aglomerado total em não pandemia

R	L(*K*)	NOMINAL	QUANTITATIVO	CONFIANÇA
01	L(3)	MAÇÃS, MAMÕES ⇒ BANANAS	4028 ⇒ 2686	67%
02	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, BOLACHAS ⇒ PROD. PANIFICAÇÃO	4871 ⇒ 3226	66%
03	L(3)	BOLACHAS, MUÇARELA ⇒ PRODUTOS DE PANIFICAÇÃO	4098 ⇒ 2709	66%
04	L(3)	BATATAS, MAÇÃS ⇒ LARANJAS	4023 ⇒ 2565	64%
05	L(3)	TOMATES, MAMÕES ⇒ BANANAS	4211 ⇒ 2647	63%
06	L(3)	MAMÕES, LARANJAS ⇒ BANANAS	2826 ⇒ 1993	63%
07	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, MUÇARELA ⇒ PROD. PANIFICAÇÃO	7360 ⇒ 4597	62%
08	L(3)	MAÇÃS, TOMATES ⇒ LARANJAS	6423 ⇒ 3970	62%
09	L(3)	MAÇÃS, LIMÕES ⇒ LARANJAS	5040 ⇒ 3092	61%
10	L(3)	MAÇÃS, OVOS ⇒ BANANAS	4298 ⇒ 2601	61%
11	L(3)	PROD. PANIFICAÇÃO, CARNE SUÍNA CONSERVA ⇒ MUÇARELA	7010 ⇒ 4235	60%
12	L(3)	MAÇÃS, LEITES ⇒ LARANJAS	4382 ⇒ 2636	60%
13	L(3)	MAÇÃS, BANANAS ⇒ LARANJAS	9593 ⇒ 5716	60%
14	L(3)	BANANAS, BATATAS DOCE ⇒ LARANJAS	4283 ⇒ 2551	60%
15	L(3)	BANANAS, CENOURAS ⇒ LARANJAS	5004 ⇒ 2978	60%

Fonte: Autoria própria.

O comportamento nas compras foi em contradição a um período de três meses aglomerados fora do período pandêmico (Tabela 3). Para dar validade à equiparação, foram usadas as mesmas configurações expressas no início das conclusões. Com praticamente 60 mil vendas a mais do que o período pandêmico, inferiu-se deste período menos regras de associação.

A primeira diferença já se deu nos conjuntos L(K). Neste, não houve L(4). Quanto a L(3), foram apenas 128 regras encontradas, contra as 551 das vendas de 2020. Uma diferença de, aproximadamente, 431% de aumento de 3 itens sendo vendidos juntos durante a pandemia do Covid-19.

4.1.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE

Em segunda análise foi decidido que, pela abundância de regras que continuam produtos da categoria de hortifruti, o resultado de mineração passaria por mais um pós-processamento de mineração, agora para limpar qualquer tipo de regra que tenha hortifruti em meio a itens associados. O fato motivador foi que, caso pudessem ser encontradas outras regras significativas, seria melhor ter todas as regras que fossem diferente de uma associação feita com produtos de hortifruti.

Diante disso, foi obtido um total de 50 regras relevantes dentro das regras anteriores já levantadas. Sendo assim, dentre as regras foi notado um novo consumo do grupo de refrigerantes e água gaseificada saborizada, conferindo um total de 7 regras (Tabela: 4) entre as 45 com a melhor medida de confiança. Tal grupo de produtos não aparece na relação de regras de associação de itens do período não pandêmico, em 2019. O consumo do grupo de refrigerantes e água gaseificada, em todas as regras em que apareceu, teve associação com produtos de panificação, apesar desde segundo grupo de produtos ser o segundo em quantidade de regras, apenas sendo inferior à incidência dos produtos de hortifruti. Contudo, os produtos de panificação já apareciam em grande escala de compras em período não pandêmico.

Tabela 4 – Resultado derivado do aglomerado total sem produtos em abundância

R	L(K)	NOMINAL	QUANTITATIVO	CONFIANÇA
094	L(3)	AGUAS GAS. SAB. E REFRIG, BOLACHAS ⇒ PROD. PANIFICAÇÃO	4225 ⇒ 2537	60%
119	L(3)	AGUAS GAS. SAB. E REFRIG, MUÇARELA ⇒ PROD. PANIFICAÇÃO	5130 ⇒ 2961	58%
177	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE BOV. MOÍDA E MIUDEZAS ⇒ PROD. PANIF.	7389 ⇒ 3952	53%
212	L(3)	AGUAS GAS. SAB. E REFRIG, PÃO DE FORMA ⇒ PROD. PANIFICAÇÃO	4553 ⇒ 2346	52%
283	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE BOV. TRAZEIRA ⇒ PROD. PANIFICAÇÃO	5783 ⇒ 2758	48%
293	L(3)	AGUAS GAS. SAB. E REFRIG, MASSAS RECHEADAS ⇒ PROD. PANIFICAÇÃO	4518 ⇒ 2131	47%
407	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE FRANGO MIUDEZAS ⇒ PROD. PANIF.	6475 ⇒ 2720	42%

Fonte: Autoria própria.

Tal comportamento, visto na tabela acima, não foi encontrado após remoção de produtos da categoria de hortifruti no período de três meses juntos, em 2019. Logo, de modo geral, foi percebido=a uma mudança nas compras feitas pelos consumidores.

4.2 DOS RESULTADOS DA SEGMENTAÇÃO MENSAL

Nesta seção, tal como na anterior, o mesmo motivo de uma categoria de produtos aparecerem em grande quantidade, que, para uma segunda análise, houve a necessidade de remover tal categoria, em busca de possíveis regras relevantes.

4.2.1 ANÁLISE DO TODO

Em primeira análise, foi percebido que hortifruti manteve a grande maioria das associações, tal como na análise do aglomerado total do período. Entretanto, nas regras que configuram associações entre os produtos de hortifruti, foi percebida uma diferença entre o interesse do produto “banana”.
Corroborando com o exposto, das 40 regras com maior medida de confiança da categoria de hortifruti no mês de maio, 39 vendas (97,5% das regras) tiveram a “banana” como um dos produtos associados, tanto em conjuntos de regras de comprimento L(3) e L(4). Tal interesse não é semelhante nos outros dois meses de período pandêmico (abril e junho), em que, das 40 primeiras regras com medidas de confiança, semelhantes ao mês anterior citado, corresponderam a 29 regras (72,5% das vendas) com “banana” em meio os hortifrútis associados diferentes.
Sendo assim, houve um aumento de 34,1% da “banana” em meio às cestas de mercado no mês de maio. Já dos outros dois meses, os resultados foram semelhantes ao aglomerado total, apresentado anteriormente.

4.2.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE

Para uma segunda análise, também foi feito, novamente, um pós-processamento para serem avaliadas as regras mensais sem produtos de hortifruti, que se mostravam ser os mais abundantes das regras associadas.

Dando sequência, o primeiro fato que se pode perceber é o quanto teve de representatividade da categoria hortifrúti em cada um dos meses nas primeiras 100 regras totais:

- **Abril**: Das 100 primeiras regras totais com melhores medidas de confiança, apenas 12 não tiveram relação alguma com nenhum produto de hortifrúti;
- **Maio**: Das 100 primeiras regras totais, apenas 2 não tiveram relação com hortifrúti;
- **Junho**: Das 100 primeiras regras totais, apenas 6 não tiveram relação com hortifrúti.

Sendo assim, é possível perceber a força das vendas do hortifruti durante o período pandêmico, em contraponto com o período não pandêmico, observado no aglomerado total de 2019.
Ainda na segunda análise, acompanhando nas Tabelas 5, 6 e 7, sem hortifruti, reiterando o observado da análise no aglomerado pandêmico sobre produtos do grupo de "refrigerantes e água gaseificada saborizada", foi percebido que tais expressivas vendas contendo esses produtos ocorreram, unicamente, nos meses de abril e junho, vide regras: 42, 92, 112, 155, 208, 242, 296 e 303 em abril; e regras 141, 162, 263, 316, 436 e 455 em junho; não sendo apresentada nenhuma regra, das regras associadas pelo algoritmo, contendo este grupo, em específico, no mês de maio. Dito isso, temos que o mês de maio teve um crescimento nas vendas de bananas, porém, praticamente, um interesse mínimo do consumidor pelos produtos de "refrigerantes e água gaseificada saborizada" .
Ainda dentro desta análise, foi notado que nos meses de abril e maio ocorreu uma associação equiparada, tanto em produtos associados, quanto em confiança, no entanto, o mesmo não ocorreu no mês de junho. Neste caso apresentado, são os produtos derivados do molho de tomate, como extratos, molhos e catchup, associados com produtos de panificação e biscoitos industrializados, conforme regra 287, em abril e regra 500, em maio.

Tabela 5 – Resultado de abril sem produtos em abundância

R	L(*K*)	NOMINAL	QUANTITATIVO	CONFIANÇA
042	L(3)	AGUAS GAS. SAB. E REFRIG, BOLACHAS ⇒ PROD. PANIFICAÇÃO	1459 ⇒ 891	61%
092	L(3)	AGUAS GAS. SAB. E REFRIG, MUÇARELA ⇒ PROD. PANIFICAÇÃO	1845 ⇒ 1043	57%
112	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE BOV. MOÍDA E MIUDEZAS ⇒ PROD. PANIF.	2640 ⇒ 1437	54%
155	L(3)	AGUAS GAS. SAB. E REFRIG, PÃO DE FORMA ⇒ PROD. PANIFICAÇÃO	1641 ⇒ 834	51%
208	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE BOV. TRAZEIRA ⇒ PROD. PANIFICAÇÃO	1964 ⇒ 937	48%
242	L(3)	AGUAS GAS. SAB. E REFRIG, MASSAS RECHEADAS ⇒ PROD. PANIFICAÇÃO	1518 ⇒ 694	46%
287	L(2)	CATCHUP E/OU MOLHO EXTR. TOMATE ⇒ PROD. PANIFICAÇÃO	2043 ⇒ 875	43%
296	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE FRANGO MIUDEZAS ⇒ PROD. PANIF.	2241 ⇒ 947	42%
303	L(3)	AGUAS GAS. SAB. E REFRIG, LEITE ⇒ PROD. PANIFICAÇÃO	2070 ⇒ 867	42%

Fonte: Autoria própria.

Tabela 6 – Resultado de maio sem produtos em abundância

R	L(*K*)	NOMINAL	QUANTITATIVO	CONFIANÇA
500	L(2)	CATCHUP E/OU MOLHO EXTR. TOMATE ⇒ PROD. PANIFICAÇÃO	2075 ⇒ 884	43%

Fonte: Autoria própria.

Tabela 7 – Resultado de junho sem produtos em abundância

R	L(*K*)	NOMINAL	QUANTITATIVO	CONFIANÇA
141	L(3)	AGUAS GAS. SAB. E REFRIG, BOLACHAS ⇒ PROD. PANIFICAÇÃO	1350 ⇒ 812	60%
162	L(3)	AGUAS GAS. SAB. E REFRIG, MUÇARELA ⇒ PROD. PANIFICAÇÃO	1577 ⇒ 935	59%
263	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE BOV. MOÍDA E MIUDEZAS ⇒ PROD. PANIF.	2227 ⇒ 1229	55%
316	L(3)	AGUAS GAS. SAB. E REFRIG, PÃO DE FORMA ⇒ PROD. PANIFICAÇÃO	1419 ⇒ 760	54%
436	L(3)	AGUAS GAS. SAB. E REFRIG, LEITE ⇒ PROD. PANIFICAÇÃO	1686 ⇒ 832	49%
455	L(3)	AGUAS GAS. SAB. E REFRIG, CARNE BOV. TRAZEIRA ⇒ PROD. PANIFICAÇÃO	1954 ⇒ 953	49%

Fonte: Autoria própria.

4.3 DOS RESULTADOS DA SEGMENTAÇÃO SEMANAL

Por fim, esta sessão foi subdividida de acordo com os dias da semana e a comparação do comportamento de compras entre as mesmas, a partir dos resultados obtidos.

4.3.1 DOMINGO

Como o primeiro dia da semana a ser analisado, após análises mensais e de aglomerado total, já é logo observado que a supremacia dos produtos de hortifrúti não ocorreram nesta análise segmentada referente ao Domingo. Dito isso, dentre as 15 regras com maiores medidas de confiança, variando de 71% a 84% nas regras associadas, apenas 3 (Regras 6, 10 e 13 na Tabela 8) correspondem a produtos de hortifruti, uma categoria de produtos que vem sendo, nas outras análises, a principal categoria dentre todas. Nesse sentido, foi uma surpresa ver que, já no primeiro dia da semana analisado, não houve tal abundância.

Somando ao exposto, houve uma dominância dessas primeiras 15 regras em produtos de origem animal, como: cortes de carne suína, cortes de carne bovina e queijo mussarela. Logo, esses três produtos distintos corresponderam a 11 das 15 regras totais primárias, como se pode notar na Tabela 8.

Ainda dentro deste dia da semana, outra regra incomum que foi observada foi referente ao produto Iogurte, apesar de apenas aparecer em 6 regras (vide regras: 27, 78, 151, 211, 280 e 313 na Tabela 10) de compras em meio às 500 primeiras, a sua compra foi associada pelo algoritimo em todas as 7 com produtos de padaria e confeitaria. Assim, demonstrando que toda vez que foi comprado Iogurte, haveria chances de 50% a 68% de eles serem comprados com produtos de padaria e confeitaria e outro item diverso, dentre bananas, refrigerantes, carne

Tabela 8 – Resultado 15 melhores regras de domingo em pandemia

R	L(*K*)	NOMINAL	QUANT.	% CONF.
01	L(3)	PERNAS/PATAS/MIUDEZAS SUÍNAS, BOLACHAS ⇒ PROD. PANIFICAÇÃO	139 ⇒ 117	84%
02	L(3)	PERNAS/PATAS/MIUDEZAS SUÍNAS, CARNE BOV. TRAZEIRA ⇒ MUÇARELA	160 ⇒ 129	81%
03	L(4)	CARNE BOV. MOÍDA E MIUDEZAS, MUÇARELA, BOLACHAS ⇒ PROD. PANIF.	126 ⇒ 122	78%
04	L(3)	MUÇARELA, BOLACHAS ⇒ PROD. PANIFICAÇÃO	388 ⇒ 295	76%
05	L(4)	AGUAS GAS. SAB. E REFRIG, PROD. PANIF., PERNAS/PATAS/MIUDEZAS SUÍNAS ⇒ MUÇARELA	154 ⇒ 113	73%
06	L(3)	BANANAS, PERNAS/PATAS/MIUDEZAS SUÍNAS ⇒ MUÇARELA	160 ⇒ 117	73%
07	L(3)	MASSAS RECHEADAS, BOLACHAS ⇒ PROD. PANIFICAÇÃO	333 ⇒ 242	73%
08	L(3)	AGUAS GAS. SAB. E REFRIG, PERNAS/PATAS/MIUDEZAS SUÍNAS ⇒ MUÇARELA	273 ⇒ 172	73%
09	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, BOLACHAS ⇒ PROD. PANIFICAÇÃO	426 ⇒ 307	72%
10	L(4)	BANANAS, ALFACES, CEBOLAS ⇒ TOMATES	169 ⇒ 121	72%
11	L(3)	MUÇARELA, MASSAS RECHEADAS ⇒ PROD. PANIFICAÇÃO	224 ⇒ 160	71%
12	L(3)	MACARRÃO, MUÇARELA ⇒ PROD. PANIFICAÇÃO	189 ⇒ 134	71%
13	L(3)	MAMÕES, MAÇÃS ⇒ BANANAS	206 ⇒ 146	71%
14	L(3)	BOLACHAS, QUEIJOS CREMOSOS ⇒ PROD. PANIFICAÇÃO	199 ⇒ 141	71%
15	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, MASSAS RECHEADAS ⇒ PROD. PANIFICAÇÃO	259 ⇒ 183	71%

Fonte: Autoria própria.

Tabela 9 – Resultado 15 melhores regras de domingo em não pandemia

R	L(*K*)	NOMINAL	QUANT.	% CONF.
01	L(3)	BOLACHAS, MUÇARELA ⇒ PROD. PANIFICAÇÃO	578 ⇒ 406	70%
02	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, MUÇARELA ⇒ PRODUTOS DE PANIFICAÇÃO	912 ⇒ 622	68%
03	L(3)	CARNE BOV. MOÍDA E MIUDEZAS, BOLACHAS ⇒ PRODUTOS DE PANIFICAÇÃO	698 ⇒ 473	68%
04	L(3)	BOLACHAS, MASSAS RECHEADAS ⇒ PRODUTOS DE PANIFICAÇÃO	534 ⇒ 357	67%
05	L(3)	BOLACHAS, PÃO DE FORMA ⇒ PRODUTOS DE PANIFICAÇÃO	784 ⇒ 494	63%
06	L(3)	BANANAS, BOLACHAS ⇒ PRODUTOS DE PANIFICAÇÃO	580 ⇒ 359	62%
07	L(3)	MAMÕES, LARANJAS ⇒ BANANAS	605 ⇒ 374	62%
08	L(3)	MUÇARELA, CARNE SUÍNA CONSERVA ⇒ PRODUTOS DE PANIFICAÇÃO	977 ⇒ 596	61%
09	L(2)	MASSAS RECHEADAS ⇒ PRODUTOS DE PANIFICAÇÃO	1954 ⇒ 1171	60%
10	L(3)	MAÇÃS, TOMATES ⇒ LARANJAS	588 ⇒ 351	60%
11	L(2)	MUÇARELA ⇒ PRODUTOS DE PANIFICAÇÃO	2709 ⇒ 1617	60%
12	L(2)	BOLACHAS ⇒ PRODUTOS DE PANIFICAÇÃO	3409 ⇒ 2018	60%
13	L(3)	MUÇARELA, AGUAS GAS. SAB. E REFRIG ⇒ PRODUTOS DE PANIFICAÇÃO	206 ⇒ 146	58%
14	L(2)	CARNE BOV. MOÍDA E MIUDEZAS, PÃO DE FORMA ⇒ PROD. PANIFICAÇÃO	801 ⇒ 466	58%
15	L(2)	QUEIJOS EM PEDAÇO ⇒ PRODUTOS DE PANIFICAÇÃO	651 ⇒ 376	58%

Fonte: Autoria própria.

bovina moída, pão de forma industrializado, mussarela e bolachas.

Com esses resultados, a mudança de comportamento foi clara com a análise das regras em 2019 (Tabela 9), os dias entre período pandêmico e não pandêmico divergiram muito, a começar pela inexistência de regras formadas de comprimento L(4) entre as 15 melhores no ano de 2019. E, ainda, houve pouca semelhança na consumação dos itens e na medida de confiança das 15 melhores regras entre as duas tabelas.

Em paralelo, quando ocorreu a análise após remoção de produtos da categoria de hortifrúti, não foi constatado esse mesmo alto consumo de iogurte encontrado aos domingos, como foi visto na tabela 10.

Tabela 10 – Resultado de domingo sobre iogurte e item adverso com prod. de panificação

R	L(*K*)	NOMINAL	QUANT.	% CONF.
027	L(3)	IOGURTE, BOLACHAS ⇒ PROD. PANIFICAÇÃO	177 ⇒ 120	68%
078	L(3)	IOGURTE, MUÇARELA ⇒ PROD. PANIFICAÇÃO	192 ⇒ 120	63%
151	L(3)	IOGURTE, PÃO DE FORMA ⇒ PROD. PANIFICAÇÃO	199 ⇒ 115	58%
211	L(3)	IOGURTE, CARNE BOV. MOÍDA E MIUDEZAS ⇒ PROD. PANIFICAÇÃO	159 ⇒ 142	55%
280	L(3)	IOGURTE, AGUAS GAS. SAB. E REFRIG ⇒ PROD. PANIFICAÇÃO	304 ⇒ 156	51%
313	L(3)	IOGURTE, BANANAS ⇒ PROD. PANIFICAÇÃO	308 ⇒ 154	50%

Fonte: Autoria própria.

4.3.2 SEGUNDA-FEIRA

Já na segunda-feira, tal como o período total e a segmentação mensal, também houve a necessidade de separação entre duas análises pelo motivo de uma categoria de produto se sobressair entre as regras de associação apuradas pelo algoritmo.

4.3.2.1 ANÁLISE DO TODO

Analisando o segundo dia da semana, percebe-se que os produtos de hortifrúti retornaram às regras de associação com as maiores medidas de confiança, sendo 100% das 100 primeiras regras (confiança que varia de 75% a 84% de compra conjunta), sendo que a primeira regra que não teve relação com produtos já mencionados foi na regra 103.
Ainda sim, dentro da análise dos produtos de hortifrúti, tiveram alguns itens que foram associados pertencentes ao conjunto de regras de comprimento L(5), em outras palavras, 5 produtos distintos juntos tiverem uma confiança alta (variando de 80% a 84%) sendo comprados juntos entre as primeiras 25 regras (Vide regras: 1, 2, 7, 8, 10, 16 e 17 na Tabela 11).

Tabela 11 – Resultado parcial de segunda com produtos do conjunto L(5)

R	L(*K*)	NOMINAL	QUANT.	% CONF.
01	L(5)	TOMATE, LARANJA, TANGERINA, BATATA DOCE ⇒ BANANA	172 ⇒ 145	84%
02	L(5)	PROD. PANIFICAÇÃO, MAMÔES, LARANJA, TANGERINA ⇒ BANANAS	175 ⇒ 147	84%
07	L(5)	COUVE-FLOR, TOMATES, MAÇAS, TANGERINA ⇒ BANANAS	178 ⇒ 146	82%
08	L(5)	MAMÔES, TOMATES, LARANJA, TANGERINA ⇒ BANANAS	186 ⇒ 152	82%
10	L(5)	COUVE-FLOR, LARANJAS, CENOURAS, TANGERINA ⇒ BANANAS	177 ⇒ 144	81%
16	L(5)	COUVE-FLOR, TAMATES, TANGERINA, BATATA DOCE ⇒ BANANAS	194 ⇒ 156	80%
17	L(5)	TOMATES, CENOURAS, CARNE FRANGO MIUDEZAS, TANGERINA ⇒ BANANA	183 ⇒ 147	80%

Fonte: Autoria própria.

Comparando com o domingo, a relação de iogurte com produtos de padaria e confeitaria não teve associação alguma, tampouco a quantidade de itens derivados de animais, também vistos no primeiro dia da semana, demonstrando outro comportamento diferente entre os dias semanais e, por ora, sendo tal comportamento exclusivo do domingo.

4.3.2.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE

Em uma segunda análise dentro da segunda-feira, foi feito um pós-processamento de resultado para a retirada de produtos da categoria de hortifruti - já que esses apareciam em abundância - na tentativa de conseguir ter uma análise mais limpa quanto a outros tipos de produtos que poderiam ser analisados. Dessa forma, foi obtido apenas 15 regras não correspondentes a hortifruti, com medidas de confiança variando de 66% a 75% de uma compra conjunta acontecer entre os itens das regras. Sendo assim, o único fato a se observar foi que das 15 regras, 13 tiveram relação de associação forte com produtos de padaria e confeitaria.
Por fim, na comparação de mudança de comportamento foi observada, primeiro, para a quantidade de regras para cada conjunto de comprimento L(K), em 2019, apesar de ter quase 10 mil vendas a mais, apresentou combinações de itens bem menores que 2020. Afinal, em 2019, conjuntos L(3) e L(4) tiveram 1033 e 121, respectivamente, em contrapartida com 2020, em que os mesmos tiveram 1676 e 578 respectivamente e, ainda mais, apenas 2020 apresentou 33 regras de conjunto L(5). Logo, podendo-se concluir que durante o período pandêmico, tais regras e avaliações, anteriormente apresentadas, não corroboram com 2019, assim, apresentando comportamentos diferentes.

4.3.3 TERÇA-FEIRA

Também na terça-feira, tal como o período total, segmentação mensal e segunda-feira na análise semanal, houve a necessidade de separação entre duas análises, pelo motivo de uma categoria de produto se sobressair entre as regras de associação apuradas pelo algoritmo.

4.3.3.1 ANÁLISE DO TODO

Partindo para a análise de terça-feira, logo de início foi possível perceber sua semelhança quanto regras com a segunda-feira: a abundância dos produtos de hortifrúti.
Porém, este dia em análise trouxe uma maior precisão, uma vez que ao contrário de segunda-feira, que teve 578 regras de comprimento L(4) e 33 de L(5), a terça-feira teve 1564 regras de L(4), 202 regras L(5) e, de forma inédita, 1 regra pertencente a L(6) (a regra 10), sendo esse último com uma medida de 84% correspondendo a regra: $\{$CEBOLAS, CENOURAS, COUVE-FLOR, TOMATES, TANGERINAS$\} \Rightarrow \{$BANANAS$\}$, que, para cada 190 ocorrências dos itens da mão esquerda, também ocorreram, junto, 160 vezes a compra do item “banana”.

4.3.3.2 ANÁLISE APÓS REMOÇÃO DE CATEGORIA DE PRODUTO ABUNDANTE

Assim como tem sido feito em outras análises, na tentativa de encontrar regras de associação de possível interesse, foi aplicado o pós-processamento nos resultados da terça-feira, para a retirada de produtos de hortifrúti. Porém, entre as 500 melhores regras, apenas uma configurou, a regra 221, com medida de 77% de confiança, sem nenhum produto de hortifrúti relacionado. Tal como ocorrido na segunda-feira, olhando para as quantidades de regras para cada conjunto de comprimento L(K), 2019 apresentou números diferentes: 1951 regras de L(3), 589 regras de L(4) e 23 regras L(5). Enquanto em 2020, de forma respectiva: 2919, 1564 e 202, e também - que não houve em 2019 - foi apresentando uma regra de tamanho 6. Dito isso, corrobora que comportamentos diferentes ocorreram, pois houve uma quantidade bem superiores de associações, que não aconteceram em período não pandêmico.

4.3.4 QUARTA-FEIRA

A análise sobre o dia de quarta-feira apontou muita semelhança com a segunda-feira, tanto as regras do hortifrui em abundância, com medidas de confiança similares, quanto em quantidade de regras para cada conjunto de comprimento L(K), mantendo a mesma média do dia citado.

Também mantendo o seguimento de fidelidade da pesquisa, foi aplicado o pós-processamento nos resultados e não ocorreram regras de associação que não tivesse um produto de hortifrúti relacionado. Logo, a partir disso, não teriam regras a serem analisadas.
Da mudança de comportamento, semelhante ao ocorrido na análise de segunda-feira, as regras, as compras e os comprimentos de conjuntos L(K) tiveram a mesma alteração e mudanças que segunda-feira. Em outras palavras, as segundas-feiras e as quartas-feiras, tanto para 2019, quanto para 2020, eram praticamente iguais em seus respectivos períodos.

4.3.5 QUINTA-FEIRA

Quinta-feira apresentou os mesmos padrões que segunda-feira e quarta-feira quanto à análise de todas as 500 regras mais interessantes, com alta medida de confiança, segundo o algoritmo de apriori. Entretanto, ainda um pouco contrário de quarta-feira, fazendo o mesmo procedimento de pós-processamento, apresentaram-se 11 regras que não eram relacionadas a hortifruti, porém, nenhuma novidade que já não tivesse sido analisada nos dias anteriores, ou que fosse regra em discrepância a ser demonstrada.
Para a análise de mudança de comportamento, as quintas-feiras de 2019 não tinham tantas regras de L(3), com 1179, e de L(4), com 119, quanto tiveram em período pandêmico, em que L(3) teve 2353 regras, L(4) teve 883 e, ainda, neste período, teve 50 regras de L(5). Mostrando, assim, comportamentos e regras diferentes quanto às compras de um período para o outro.

4.3.6 SEXTA-FEIRA

Já na sexta-feira, foram obtidos alguns resultados diferentes dos demais dias semanais. Dito isso, de início se observar, ainda, a grande quantidade de regras de hortifrúti, porém, ao se utilizar do pós-processamento em cima dos resultados para fazer uma análise sem produtos de hortifrúti, o resultado obtido foi diferente dos demais em quantidade de produtos que, segundo o algoritimo, também tiveram mais expressão com maiores medidas de confiança. Nesse sentido, foram obtidas 43 regras que não tinham nenhuma relação com hortifrúti, variando de 60% a 74% de confiança. Em meio desses, foram 7 regras que pertencem ao conjunto de L(4). Uma dessas, a regra 191, é inteiramente de produtos de carne na cesta de compras, os

produtos: carne moída ou carne cortada em cubos/pedaços, carne bovina dianteira desossada, carne bovina traseira desossada e carne de frango em filé ou pedaços, com uma medida de confiança de 67% de serem comprados em conjunto.
Ainda na sexta-feira, uma associação de comprimento L(2), a regra 303 chamou atenção por não aparecer em nenhum dos outros dias, essa consistia em uma compra conjunta de "Hipoclorito de Sódio" (água sanitária ou alvejantes branqueadores) junto de produtos de limpeza e conservação doméstica (desinfetantes e detergentes em geral) com uma medida de 64% de confiança.

Tabela 12 – Resultado Parcial de Sexta com Regras Discrepantes

R	L(*K*)	NOMINAL	QUANT.	% CONF.
191	L(4)	CARNE B. MOÍDA, CARNE B. DIANT., CARNE B. TRAZ. ⇒ CARNE FRAN. MIUDEZAS	195 ⇒ 130	67%
303	L(2)	HIPLOCORITO DE SÓDIO ⇒ PREP. PARA LAVAGEM E LIMPEZA ATIVA	367 ⇒ 234	64%

Fonte: Autoria própria.

Da contraprova de mudança no comportamento de consumo, a sexta-feira de período não pandêmico também se mostrou diferente do correspondente em período de pandemia. Em 2019: L(3): 729, L(4): 20; já em 2020: L(3): 1687, L(4): 332 e teve, também, um L(5) com 1 regra. Corroborando, assim, com comportamentos diferentes nas compras e associações obtidas.

4.3.7 SÁBADO

Sobre as regras e comportamentos de sábado, foi semelhante a quinta, quarta e segunda-feira. Mesmo com o pós-processamento realizado, não ocorreu grande mudança das regras já obtidas anteriormente, apenas por uma única diferente aparente, a venda conjunta de cerveja, carvões de churrasco e carne em linguiça, regra 454, com medida de 63% de confiança.
Por fim, o sábado de 2019 também se mostrou diferente em comparação ao mesmo dia semanal, em período pandêmico.
Em 2019: L(3): 1652, L(4): 213; Já em 2020: L(3): 2446, L(4): 755 e teve, também, um L(5) com 18 regras. Sendo assim, mostrando que mesmo com 6 mil vendas de diferença, existiram mais associações de itens durante período pandêmico.

5 CONCLUSÕES E TRABALHOS FUTUROS

5.1 CONCLUSÔES

Em vista dos resultados e argumentações sobre as informações obtidas, esta seção trata das discussões realizadas com o especialista, que detem o interesse pelos resultados.
Do que se compreende como novidade, foi apontado que já era sabido que os itens mais vendidos eram da categoria hortifruti. No entanto, foi satisfatório saber que os resultados apontaram uma associação mais precisa de quais produtos desta categoria eram mais vendidos juntos, afinal, ocorreram regras de comprimento L(5) e L(6) em alguns dos casos, assim, trazendo uma assertividade da compra conjunta de certos itens. Logo, estas informações seriam, no mínimo, levantadas para reuniões de tomada de decisão estratégica.
Conforme analisado pelo especialista, os fatores considerados novidades foram:

- O aumento no consumo de refrigerantes com produtos de panificação em dois de três meses durante o período pandêmico;
- O aumento de consumo de iogurte com produtos de panificação aos domingos, fato que não ocorreu aos domingos dos períodos analisados no ano de 2019;
- Ainda no domingo, apesar de o especialista não ter como novidade que este dia era o dia que mais vendia carne, a associação foi pertinente para mostrar quais cortes eram mais vendidos juntos;
- A associação de comprimento L(5) no consumo de frutas e verduras nas segundas-feiras, sendo que o “dia de feira” dessa categoria ocorre nas quartas-feiras, que, no caso, apontou semelhança nos resultados;
- O aumento de consumo de produtos de limpeza, de modo geral, ocorreu, exclusivamente, nas sextas-feiras, fato que não era de conhecimento da empresa colaboradora.

Diante de tais apontamentos apresentados, possíveis benefícios podem ser aproveitados pelo comércio, como uma promoção de refrigerantes nos dias da semana que não tiveram o aumento do consumo, ou, até mesmo, trazer uma gôndola de refrigerantes perto do setor de padaria do

mercado. A mesma estratégia pode ser aplicada para o iogurte, que teve seu consumo aumentado aos domingos, assim, podendo criar uma promoção para vendê-lo em outros dias. Ainda nessa linha, ao saber que nas segundas-feiras o consumo de frutas e vegetais é tão fortes quanto nas quartas-feiras, que é o dia de promoção normal desta categoria, a empresa poderia testar um "segundo dia de feira" na mesma segunda-feira, para observar se aumentaria mais a venda, ao ponto de passar a quarta-feira. Dessa forma, possivelmente, descobrindo um dia melhor para sua promoção desta categoria de produtos.
Logo, essas são algumas das várias estratégias que o mercado poderia adotar, caso decidam discutir e levar à pauta de reunião de tomadas de decisão.

5.2 TRABALHOS FUTUROS

- **Utilização de outros algoritmos de associação.**
  Como levantado na fundamentação teórica desta monografia, há outros algoritmos de associação, como os clássicos Fuzzy Apriori e o Eclat, ou, até mesmo os mais novos, como Enchated Apriori ou TFI-Apriori. É possível testar, também, tais algoritmos em um conjunto de dados, tal como o deste trabalho, para uma análise mais abrangente de resultados possíveis. No escopo de tempo para esta pesquisa, não foi possível se utilizar mais de um algoritmo, mas seria uma boa forma de comparar os algoritmos para cada situação.
- **Análise por agrupamento de dias da semana e separados por turnos: manhã, tarde e noite.**
  A ideia inicial deste trabalho também contava com mineração de dados analisados a partir de regras de agrupamento, mas o escopo de tempo não permitiu tal realização. Por conta disso, os dados, no momento do tratamento e da formatação, já estavam preparados com o atributo "turno", com isso, criando a oportunidade de realizar o estudo e trazer resultados ao supermercado acerca da demanda de produtos por turno. Esta analise pode fornecer informações sobre a troca de turno de funcionário ou, até mesmo, para escalar de melhor maneira o turno de folga ante a demanda de compra do setor em que o colaborador trabalha.

REFERÊNCIAS

AGGARWAL, C. C. **Data Mining**: The textbook. [S.l.]: Springer, 2015.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules. **Proceedings of the 20th International Conference on Very Large Data Bases**, Morgan Kaufmann Publishers Inc, 1994.

AHLEMEYER-STUBBE, A.; COLEMAN, S. **A Practical Guide to Data Mining for Business and Industry**. [S.l.]: Wiley, 2014.

AL-MAOLEGI, M.; ARKOK, B. An improved apriori algorithm for association rules. **International Journal on Natural Language Computing Vol. 3, No. 1, February 2014**, 2014.

ANSARI, E. et al. Tfi-apriori: Using new encoding to optimize the apriori algorithm. **Intelligent Data Analysis, vol. 22, no. 4, pp. 807-827**, 2018.

BORGELT, C. Efficient implementations of apriori and eclat. Department of Knowledge Processing and Language Engineering, School of Computer Science - University of Magdeburg, 2003.

BORGES, L. E. **Python para Desenvolvedores**: Aborda python 3.3. [S.l.]: Novatec Editora, 2014.

CAMILO, C. O.; SILVA, J. C. da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. Instituto de Informátia da Universidade Federal de Goiás, 2009.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTHTHOR, P. From data mining to knowledge discovery in databases. Association for the Advancement of Artificial Intelligence - AAAI, 1996.

FAZENDA, M. da Economia e. **NF-e**. 2020. Disponível em: <http://www.nfe.fazenda.gov.br/portal/perguntasFrequentes.aspx?tipoConteudo=E4+tmY+ODf4=>.

GUINANCIO, J. C. et al. Covid-19: Desafios do cotidiano e estratégias de enfrentamento frente ao isolamento social. **Research, Society and Development, v.9, n. 8, e259985474, 2020**, 2020.

HAN, J.; KAMBER, M. **Data Mining**: Concepts and techniques. [S.l.]: Morgan Kaufmann, 2001.

ISO. **ISO 88979**: Standard generalized markup language. 1986. Disponível em: <https://www.iso.org/standard/16387.html>.

JAPKOWICZ, N.; STEFANOWSKI, J. **Big Data Analysis**: New algorithms for a new society. [S.l.]: Springer, 2016.

LAROSE, D. T.; LAROSE, C. D. **DATA MINING AND PREDICTIVE ANALYTICS**. Second. [S.l.]: Wiley, 2015.

LUTZ, M. **Programming Python**. [S.l.]: O'Reilly, 2007.

MIKUT, R.; REISCHL, M. Data mining tools. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, September 2011**, 2011.

PYTHON. **PEP 206**: Batteries included philosophy. 2019. Disponível em: <https://www.python.org/dev/peps/pep-0206/id3>.

RAO, S.; GUPTA, P. Implementing improved algorithm over apriori data mining association rule algorithm. IJCST - International Journal of Computer Science And Technology, 2012.

REZENDE, S. O. Mineração de dados. XXV Congresso da Sociedade Brasileira de Computação, 2005.

RFB. **Sistema Público de Escrituração Digital**: Nf-e. 2020. Disponível em: <http://sped.rfb.gov.br/pagina/show/1328>.

THULER, L. C. S.; MELO, A. C. de. Sars-cov-2/covid-19 em pacientes com câncer. INCA - Instituto Nacional de Câncer José Alencar Gomes da Silva, 2020.

TIRUMALASETTY, S.; JADDA, A.; EDARA, S. R. An enhanced apriori algorithm for discovering frequent patterns with optimal number of scans. **International Journal of Computer Science Issues, 2015**, 2015.

VASCONELOS, L. M. R. de; CARVALHO, C. L. de. Aplicação de regras de associação para mineração de dados na web. Instituto de Informátia da Universidade Federal de Goiás, 2004.

VIEIRA, M. T. P.; FELIPE, J. C. Data warehouse: Data warehousing, olap e data mining. Universidade Federal de São Carlos, 2001.

W3C. **XML**. 2016. Disponível em: <https://www.w3.org/XML/>.

W3SCHOOLS. **XML Tree**. 1999. Disponível em: <https://www.w3schools.com/xml/xml_tree.asp>.

WEKA. **Weka Wiki: Documentation**. 2020. Disponível em: <https://waikato.github.io/weka-wiki/documentation/>.

APÊNDICE A – CODIGO-FONTE DO MODULO DE CAPTURA DE DADOS DESENVOLVIDO EM PYTHON 3.7

```
from datetime import datetime, time, date
import xml.etree.ElementTree as et
import os

def get_periodo(data_nfe):
    hora = int(data_nfe[11:13])
    if hora >= 18:
        return 'NOITE'
    elif hora >= 12:
        return 'TARDE'
    else:
        return 'MANHA'

def get_dia_semana(dia, mes, ano):

    dia_semana_list = ['SEGUNDA', 'TERCA', 'QUARTA', 'QUINTA', '
       SEXTA', 'SABADO', 'DOMINGO']

    data = date(year=int(ano), month=int(mes), day=int(dia))
    num = data.weekday()
    dia_semana = dia_semana_list[num]

    return dia_semana

def run_xml(path):

    diretorios = os.listdir(path)
    lista = []
    lista_return = []
```

```
29
30      for diretorio in diretorios:
31
32          path_autorizazao = f'{path}/{diretorio}/Autoriza  o'
33          xmls = os.listdir(path_autorizazao)
34
35          for xml in xmls:
36
37              ano = diretorio[:4]
38              mes = diretorio[4:6]
39              dia = diretorio[6:]
40              semana = get_dia_semana(dia, mes, ano)
41              path_final = path_autorizazao + f'/{xml}'
42              data = f'{dia}{mes}{ano}'
43
44              xml_tree = et.parse(path_final)
45              root = xml_tree.getroot()
46
47              flag_icms = False
48              periodo = ''
49              nfe = ''
50              serie = ''
51
52              for child in root.iter('*'):
53
54                  if child.tag[36:] == 'total' or child.tag[36:] ==
                        'ICMS' or child.tag[36:] == 'PIS' or child.
                       text is not None:
55                      if child.tag[36:] == 'serie':
56
57                          serie = child.text
58
59                      elif child.tag[36:] == 'nNF':
60
61                          nfe = f'{serie}{child.text}'
62
63                      elif child.tag[36:] == 'dhEmi':
64
```

```
65                          periodo = get_periodo(child.text)
66
67                      elif child.tag[36:] == 'xProd':
68
69                          lista.append(nfe)         # 0 NUM NOTA
                               FISCAL
70                          lista.append(semana)      # 1 DIA DA
                               SEMANA
71                          lista.append(periodo)     # 2 DATA HORA
72                          lista.append(child.text) # 3 DESCR
                               PRODUTO
73
74                      elif child.tag[36:] == 'NCM':
75
76                          lista.append(child.text) # 4 NCM
77
78                      elif child.tag[36:] == 'ICMS':
79
80                          flag_icms = True
81
82                      elif flag_icms is True and child.tag[36:] ==
                           'CST':
83
84                          lista.append(child.text) # 5 CST
85                      elif child.tag[36:] == 'PIS':
86
87                          flag_icms = False
88
89                      elif child.tag[36:] == 'total':
90
91                          break
92
93      pos = 0
94      while pos < len(lista):
95
96          aux = [lista[pos], lista[pos+1], lista[pos+2], lista[pos
                   +3], lista[pos+4], lista[pos+5]]
97          lista_return.append(aux)
```

```
98          pos += 6
99
100     return lista_return
```

APÊNDICE B – CODIGO-FONTE DO MODULO DE CRIAÇÃO DE CSV PARA MINERAÇÃO DE DADOS DESENVOLVIDO EM PYTHON 3.7

```
from src import captura_dados, captura_dados_semanais,
    captura_dados_aglomerado

def criar(path, nome_arquivo, tp):

    lista_dados = []
    if tp == 'm':
        lista_dados = captura_dados.run_xml(path)
    elif tp == 's':
        lista_dados = captura_dados_semanais.run_xml(path)
    elif tp == 'a':
        lista_dados = captura_dados_aglomerado.run_xml(path)

    lista_semana = ['segunda', 'terca', 'quarta', 'quinta', '
        sexta', 'sabado', 'domingo']
    lista_turno = ['manha', 'tarde', 'noite']
    lista_nfe = []
    lista_produtos = []

    for x in lista_dados:

        if f'N-{x[0]}' not in lista_nfe:
            print('Colocando NFE na lista de NFes!')
            lista_nfe.append(f'N-{x[0]}')

        if f'P-{x[4]}' not in lista_produtos:
            print('Colocando produto na lista de produtos!')
            lista_produtos.append(f'P-{x[4]}')

```

```
28      #print(f'{lista_produtos}\n')
29      #print(f'{lista_nfe}\n')
30
31      interpol_nfe = lista_dados[0][0]
32      linha_data = [f'N-{lista_dados[0][0]}', lista_dados[0][1],
           lista_dados[0][2]]
33      for prod in lista_produtos:
34          linha_data.append('?')
35
36      tupla = ''
37      data = []
38
39      lista_coluna_dados = ['nfe', 'diasemana', 'turno']
40      for prod in lista_produtos:
41          lista_coluna_dados.append(prod)
42
43      data.append(lista_coluna_dados)
44
45      for x in lista_dados:
46
47          if x[0] != interpol_nfe:
48
49              tupla = tuple(linha_data)
50              linha_data.clear()
51              tupla = list(tupla)
52              data.append(tupla)
53
54              nfe = x[0]
55              semana = x[1]
56              turno = x[2]
57
58              linha_data = [f'N-{x[0]}', x[1], x[2]]
59
60              for prod in lista_produtos:
61                  linha_data.append('?')
62
63              pos = lista_produtos.index(f'P-{x[4]}') + 3
64              linha_data[pos] = 't'
```

```
65
66              interpol_nfe = x[0]
67
68          else:
69
70              pos = lista_produtos.index(f'P-{x[4]}') + 3
71              linha_data[pos] = 't'
72
73              interpol_nfe = x[0]
74
75      tupla = tuple(linha_data)
76      linha_data.clear()
77      tupla = list(tupla)
78      data.append(tupla)
79
80      arq = ''
81      if tp == 'm':
82          arq = open(f'Arquivos CSV/MENSAL/{nome_arquivo} para
               Apriori.csv', 'w')
83
84      elif tp == 's':
85          arq = open(f'Arquivos CSV/SEMANAL/{nome_arquivo} para
               Apriori.csv', 'w')
86
87      elif tp == 'a':
88          arq = open(f'Arquivos CSV/Aglomerado {nome_arquivo} para
               Apriori.csv', 'w')
89
90      for x in data:
91          str_x = f'{x}'
92          str_x = str_x.replace('[', '').replace(']', '').replace("
               '", '').replace(' ', '')
93          print('Imprimindo linha')
94          arq.write(f'{str_x}\n')
95
96      arq.close()
```