MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

UM ESTUDO SOBRE O CONSUMO DE ENERGIA ELÉTRICA UTILIZANDO ALGORITMOS CLASSIFICADORES

por

Mariane Coelho Amaral

Dissertação para obtenção do Título de Mestre
em Modelagem Computacional

Orientadora: Graçaliz Dimuro
Coorientador: Eduardo N. Borges

Rio Grande, 2018

AGRADECIMENTOS

Ao meu esposo, Anderson Garcia Silveira, por todo amor, carinho e paciência em todos os momentos da minha vida.

À Myrna Martinato Coelho, minha mãe, por todo esforço que sempre realizou para que eu obtivesse bons resultados em meus estudos.

À Graçaliz Pereira Dimuro, minha orientadora, que me recebeu de braços abertos logo em minha chegada na FURG e me auxiliou desde o início.

Ao meu coorientador, Eduardo Nunes Borges, por todo empenho e dedicação em me ensinar conteúdos relevantes à realização deste trabalho.

Aos demais que, de alguma maneira, contribuíram para que fosse possível o término desta dissertação.

RESUMO

O estudo do correto balanço energético é de extrema importância para a economia de um país. Quando a demanda é superior à oferta, faltará eletricidade para os consumidores. Já quando a oferta é muito maior que a demanda por eletricidade, as empresas geradoras e distribuidoras de energia sofrem prejuízos. As técnicas de mineração de dados consistem em descobrir conhecimentos em banco de dados, podendo ser de fundamental importância para revelar informações vitais às pesquisas. Assim, este trabalho intencionou, através de algoritmos classificadores, estimar o valor da conta mensal de energia elétrica com base nos atributos investigados, identificando perfis de consumo residencial por meio de um instrumento científico aplicado aos estudantes de graduação da Universidade Federal do Rio Grande (FURG) e da Universidade Federal de Pelotas (UFPel). A análise do valor da conta de energia elétrica a partir de predições realizadas pelos classificadores pode ser considerada o primeiro passo na melhoria em sistemas elétricos de potência, através do planejamento de geração e distribuição de energia.

Palavras-chaves: Energia Elétrica; Perfis de Consumo; Mineração de Dados; Classificação.

A Study on eletrical energy consumption using classication algorithms

ABSTRACT

The study of the correct energy balance is extremely important for a country economy. When the demand is higher than supply, electricity will be lacking for consumers. When the supply is much greater than the demand for electricity, the generating companies and energy distributors will be suffering losses. The data mining techniques consist of discover knowledge in database, and may be fundamental for reveal vital information for researches. Thus, this work intends, through classification algorithms, calculating the value of the monthly electric energy bill based on the attributes about consumption and by a scientific instrument applied to undergraduate students from the Federal University of Rio Grande (FURG) and Federal University of Pelotas (UFPel). The analysis of the value of the energy account based on predictions made by the classifiers can be considered the first step in the improvement in power systems, through generation planning and distribution power.

Keywords: Electrical Energy; Energy Demand Profiles; Data Mining, Classification.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

A	Ganho de informação
A_c	Acurácia
d	Número de dimensões
e	Energia elétrica consumida
E	Especificidade
k	Número de vizinhos mais próximos
K	Número *Kappa*
P	Precisão
P_e	Taxa hipotética de aceitação
P_o	Taxa de aceitação relativa
R	*Recall*
s	Potência consumida
S	Sensibilidade
t	Tempo
u	Unidades de amostras

LISTA DE ABREVIATURAS

ANEEL	Agência Nacional de Energia Elétrica
ARFF	*Attribute-Relation File Format*
BEN	Balanço Energético Nacional
CSV	*Comma-separated values*
DOE	*Design of Experiments*
EE	Energia Elétrica
EPE	Empresa de Pesquisa Energética
FURG	Universidade Federal do Rio Grande
IA	Inteligência Artificial
ID	Identificação de Aluno
ID3	*Induction Decision Tree*
IP	Instrumento Preliminar
KDD	*Knowledge Discovery in Databases*
KWh	Quilowatt-hora
OIE	Oferta Interna de Energia
PDEf	Plano Decenal de Eficiência Energética
PPGMC	Programa de Pós-Graduação em Modelagem Computacional
PPEHU	Pesquisa de Posse de Equipamentos e Hábitos de Uso
Procel	Programa Nacional de Conservação de Energia Elétrica
tep	Tonelada Equivalente de Petróleo
UC	Unidade Consumidora
UFPel	Universidade Federal de Pelotas
MD	Mineração de Dados
SMO	*Sequential Minimal Optimization*
SVM	*Support Vector Machine*
Weka	*Waikato Environment for Knowledge Analysis*

1 INTRODUÇÃO

O planejamento do adequado balanço energético impacta diretamente a economia, visto que a energia é fator essencial na geração de riquezas em todos os setores da indústria. Dessa maneira, quando a demanda de energia elétrica é superior à oferta, faltará eletricidade na residência dos consumidores. Quando, no entanto, a oferta é muito maior que a demanda de eletricidade, as empresas geradoras de energia sofrem prejuízos, pois as empresas distribuidoras não precisam de toda energia gerada pelas unidades geradoras. Segundo Santos (2016), há um grande desafio nesta indústria desregulada por parte das companhias geradoras para acompanhar o consumo real de energia, sem que haja uma perda monetária por parte das empresas e sem que uma parcela da população fique descoberta de energia em sua residência.

De acordo com a Lei 9.427 do ano de 1996, é criada a Agência Nacional de Energia Elétrica, a ANEEL. Essa Agência tem por finalidade regular e fiscalizar a geração, transmissão e distribuição de energia elétrica no país (Brasil, 1996). Compete, ainda, à ANEEL implementar as políticas e diretrizes do governo federal para a exploração da energia elétrica e o aproveitamento dos potenciais hidráulicos e estabelecer tarifas de energia elétrica no âmbito nacional.

A ANEEL define um consumidor como uma pessoa física ou jurídica que assume a responsabilidade pelo pagamento de faturas de energia elétrica. Uma Unidade Consumidora (UC), por sua vez, é definida como um conjunto de instalações de equipamentos elétricos que recebem energia, como exposto por ANEEL (2010). No cenário atual, os consumidores brasileiros são classificados como consumidores residenciais, comerciais, industriais ou outros.

No setor residencial, a energia é consumida praticamente em sua totalidade por eletrodomésticos, equipamentos e iluminação, havendo uma grande diversidade de equipamentos e sistemas sendo utilizados em quantidades cada vez maiores.

De acordo com o Balanço Energético Nacional (BEN) do ano de 2016, e como vem sendo há algum tempo, a principal fonte de geração de energia elétrica no Brasil é a hidráulica. Porém, houve uma queda de 3,7% em comparação com 2015. Isso mostra que, mesmo com a geração hidráulica sendo a principal responsável pela geração, deve haver uma preocupação com esse dimensionamento (EPE, 2016).

A Figura 1.1 mostra a geração hidráulica dominante na matriz energética, representando 64% da oferta interna de energia elétrica no país. Segundo EPE (2015), na oferta interna de energia (OIE) é considerada a soma dos montantes referente à produção nacional mais as importações.

A qualidade do atendimento aos consumidores no Brasil é determinada pela ANEEL, que também fiscaliza esses serviços. Em virtudes de multas e sanções ligadas à qualidade da eletricidade recebida pelos consumidores, é preciso gerenciar a rede de distribuição a fim de evitar ou diminuir a frequência e a duração das interrupções no fornecimento de energia decorrente de um mau dimensionamento de geração.

O BEN do ano de 2015 informa que o consumo brasileiro de energia em 2014 atingiu 265,86

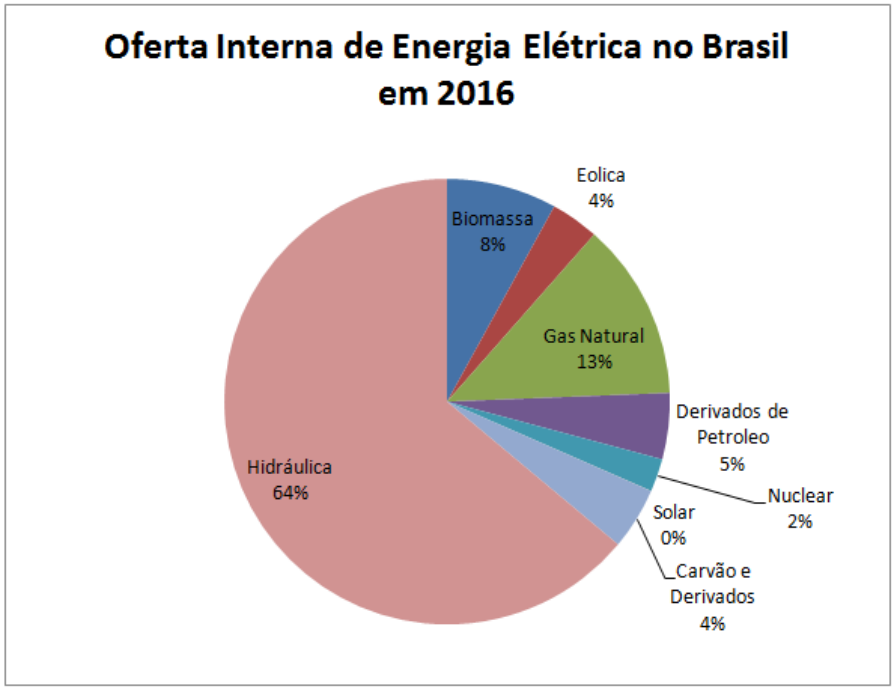

Figura 1.1: Oferta interna de Energia no Brasil em 2016
Fonte: elaborada pela autora.

milhões de toneladas equivalentes de petróleo (tep). Considerando-se as projeções e fazendo-se o uso da taxa de crescimento populacional de 4%, chega-se a 737,10 milhões de tep em 2040. Em 2014, o consumo de energia por habitante no Brasil foi de 1,311 tep. Os 265,86 milhões de tep consumidos no Brasil, em 2014, correspondem a 87% da oferta interna de energia, sendo um consumo 4,28 vezes superior ao verificado em 1970. Esta situação é ilustrada pelo gráfico da Fig. 1.2.

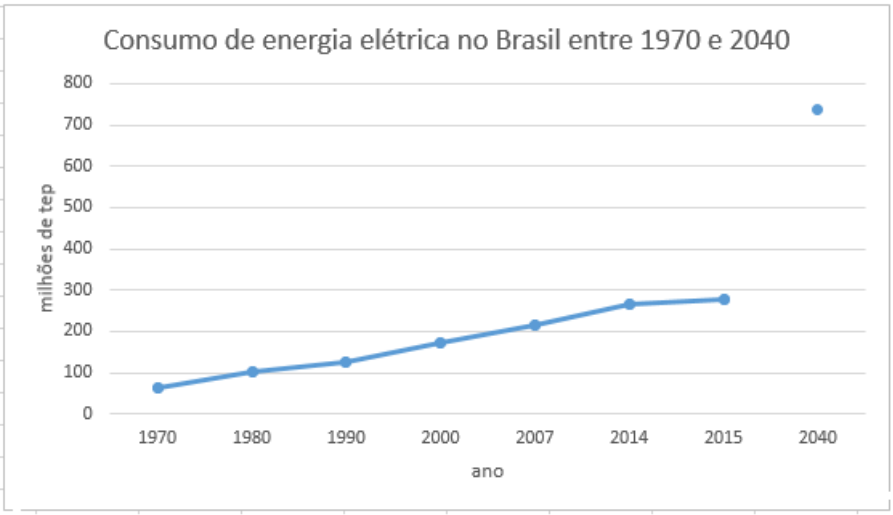

Figura 1.2: Consumo de energia no Brasil entre 1970 e 2040
Fonte: elaborada pela autora.

É necessária uma preocupação cada vez maior com a qualidade de energia, visto que o consumo de energia elétrica residencial no Brasil apresentou crescimento de 5,4% entre 2012 e

2013, segundo a ANEEL e exposto por Hansen (2000). No entanto, melhorias no setor elétrico vêm atreladas a um aumento de custo dos serviços ao consumidor.

O que torna difícil a medição de dados de consumo de energia no momento é o fato de as concessionárias contarem com aparelhos analógicos nas residências, registrando o consumo total entre as medições. Assim, por ainda haver uma indisponibilidade de medidores digitais nas residências, dificultando uma análise diária dos hábitos de consumo, recorre-se a uma pesquisa de hábitos de consumo com uma amostra de usuários de energia elétrica, como realizado atualmente nos trabalhos científicos.

Em virtude da pouca quantidade de dados fornecidos e disponíveis nas distribuidoras e geradoras, para entendimento e descoberta dos hábitos de consumo da população brasileira, ainda é utilizada a pesquisa de posse de eletrodomésticos, por exemplo. Também pode ser utilizada a pesquisa de hábitos de consumo, aplicados em pesquisas de campo. Cabe salientar que, de acordo com Santos (2016), muitas vezes essas pesquisas não são realizadas, dificultando uma criação de um banco de dados confiável dos clientes e causando um maior problema no estudo da demanda de energia.

O maior desafio do setor elétrico é, portanto, em relação ao correto dimensionamento da demanda de energia: pessoas com hábitos semelhantes possuem hábitos de consumo de energia similares em alguns aspectos, causando picos de consumo de energia em certos momentos do dia. Em outros momentos, a energia disponível pode não ser utilizada, causando prejuízo às empresas fornecedoras. O setor elétrico do país, então, necessita estar preparado para lidar com os perfis de consumidores de energia.

Neste contexto, este trabalho propõe analisar e descobrir perfis de consumidores através da predição do valor mensal pago em energia elétrica. Os dados coletados passam por uma série de operações de processamento, denominadas de técnicas de mineração de dados.

Essas técnicas contemplam múltiplas áreas, incluindo inteligência artificial, aprendizagem de máquina, estatística e bancos de dados.

De acordo com Baker e Yacef (2011), os primeiros estudos sobre mineração de dados foram apresentados em meados dos anos 80, motivados pelo início da dificuldade em administrar e controlar o grande volume de dados que estava surgindo. Os estudos, inicialmente, tinham a pretensão de apenas reconhecer simples padrões em bancos de dados. No ano considerado pelos autores, 2009, havia a consciência de que a mineração de dados crescia rapidamente, havendo um grande número de publicações em eventos.

Logo, com a evolução da ciência da computação, o processo de mineração passou a se interessar por sanar as necessidades dos usuários em um mais alto nível: analisando os dados antes, durante e após o processo. Assim, segundo Baker, Isotani e Carvalho (2011), minerar dados significa extrair conhecimento de dados que, aparentemente, não possuem ligações explícitas. Essa descoberta de conhecimento em banco de dados tornou-se conhecida por *KDD*, de *Knowledge Discovery in Databases*.

Dessa maneira, a fim de auxiliar no planejamento do balanço energético, é preciso que haja uma correta oferta de energia, de acordo com a demanda necessária. Para isso, uma ferramenta

útil é a análise de perfis de consumo através de técnicas de descoberta de conhecimento em conjuntos de dados que, aparentemente, não possuam ligações. Pode-se dizer que o primeiro passo a ser dado é a investigação do valor da conta de energia elétrica paga mensalmente pelas habitações do setor residencial.

Neste contexto, as próximas seções evidenciam a motivação e os objetivos do trabalho realizado.

1.1 Justificativa

Este trabalho torna-se de fundamental importância ao ser capaz de fornecer um primeiro estudo às concessionárias, pesquisadores e interessados sobre diversas possibilidades de comparação para predição do valor gasto mensalmente com energia elétrica, podendo contribuir para aprimorar os sistemas elétricos e sendo de valor como aporte para estudos sobre demanda de energia elétrica, bem como fornecer informações sobre seus perfis aos próprios consumidores residenciais.

Dessa maneira, o estudo de perfis de usuários através do cálculo do valor gasto em energia elétrica é útil para auxiliar estudos sobre o setor residencial, industrial e comercial de todo o país, bem como impactar diretamente em sua economia. Este trabalho foca apenas em habitações residenciais.

1.2 Objetivos

Este trabalho intenciona, através da mineração de dados, estimar o valor da conta mensal de energia elétrica com base nos atributos investigados, identificando perfis de consumo residencial e tendo seu espaço amostral definido entre os alunos de cursos de Graduação da Universidade Federal do Rio Grande (FURG) e da Universidade Federal de Pelotas (UFPel). Ainda, o trabalho pretende atingir os objetivos específicos:

- Estudar instrumentos científicos de coleta de dados;
- Aprimorar os conhecimentos nas áreas de mineração de dados e sistemas de energia;
- Construir uma base de dados para usos futuros em outras pesquisas;
- Fornecer às concessionárias e demais interessados informações iniciais sobre estudos relacionando algoritmos classificadores e a demanda residencial de energia elétrica, para futuros estudos mais complexos.

2 REFERENCIAL TEÓRICO

Este capítulo apresenta um conjunto de conceitos essenciais para o entendimento da proposta de trabalho descrita nesta dissertação. A Seção 2.1 define medidas de consumo de energia e mostra a distribuição de eletrodomésticos realizada por pesquisas sobre consumo residencial. A Seção 2.2 apresenta a utilização de questionários dentro do método científico. Por fim, na Seção 2.3 são evidenciados o conceitos sobre análise de dados através de técnicas de mineração, explorando a utilização no Software *Weka*, bem como evidenciando alguns exemplos de classificadores.

2.1 O Consumo de Energia Elétrica

Cada aparelho eletrônico ou eletrodoméstico presente nas residências e indústrias necessita de eletricidade para ser utilizado. Então, o consumo de energia elétrica, definido pela Equação 2.1, é fator de preocupação e estudo de pesquisas energéticas. A variável e representa a energia consumida, em unidade de kWh. A potência de cada equipamento considerado (s) é expressa em *Watts* e, por sua vez t é expresso em horas e representa o tempo que o equipamento permanece ligado.

$$e = s.t \tag{2.1}$$

A preocupação com o consumo de energia está cada vez mais presente na realidade brasileira: políticas e campanhas auxiliam os consumidores a entender suas faturas de energia elétrica, bem como a economizar e consumir conscientemente.

O Programa Nacional de Conservação de Energia Elétrica (Procel), por exemplo, é um programa brasileiro executado pela Eletrobras que teve seu início no ano de 1985 com a intenção de reduzir o desperdício de energia elétrica e conscientizar os usuários a respeito de economia de energia.

Assim, o Procel utiliza-se se um selo para identificar os equipamentos que são mais eficientes no consumo e que agridem menos o meio ambiente. Nesta lista, é possível encontrar equipamentos conhecidos, como fornos, lâmpadas e refrigeradores. O Relatório de Resultados do Procel lançado em 2017 apresenta os resultados globais e específicos em cada área de atuação do programa, explicitando suas metas e missões a partir da análise dos dados de 2016.

Dessa maneira, este programa aponta que as ações fomentadas ajudaram a reduzir a demanda em 8,375 MW, representando no consumo residencial de energia elétrica 11,40% no ano de 2016 em relação ao ano anterior. Como meta para 2017 e 2018, o Procel determina que: "Entre os estudos, destacam-se a realização de uma nova pesquisa de posse de equipamentos e hábitos de uso, o desenvolvimento de procedimentos para ensaios de equipamentos e de metodologias de medição e verificação, o apoio para a elaboração do Plano Decenal de Eficiência Energética (PDEf) e a elaboração de modelos para substituição de equipamentos obsoletos, por

exemplo"(Procel, 2016).

O Procel também disponibiliza em seu endereço virtual resultados de algumas pesquisas, dentre elas a pesquisa de posse de eletrodomésticos. Sabe-se que o setor residencial do Brasil é constituído por um grupo bastante heterogêneo de consumidores, principalmente no que se refere ao perfil de posse e uso de eletrodomésticos e que isso pode ser, em parte, explicado pelas variações de renda familiar, que exercem grande influência nos hábitos de consumo de energia elétrica das residências, e pelo fator climático.

A Figura 2.1 mostra a utilização de eletrodomésticos mais comuns na região sul do país. Nota-se que os itens com maior participação na fatura de energia elétrica são o chuveiro, o ar condicionado e a geladeira responsáveis por, respectivamente, por 25%, 32% e 16% do total. Os dados se referem ao ano de 2005, quando foi realizada uma pesquisa pelo Procel para investigação sobre aparelhos eletrônicos.

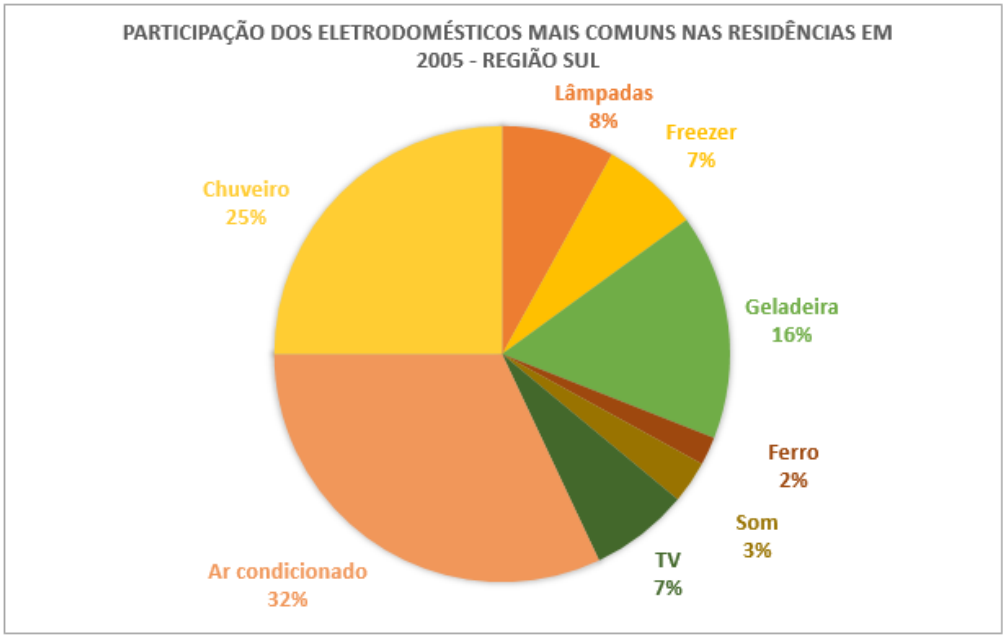

Figura 2.1: Participação de eletrodomésticos na Região Sul - ano 2015
Fonte: elaborada pela autora.

Já a Figura 2.2 mostra a utilização de eletrodomésticos mais comuns na região nordeste do Brasil. Nota-se que os itens com maior participação na fatura de energia elétrica são: as lâmpadas, o ar condicionado e a geladeira responsáveis por, respectivamente, por 11%, 28% e 30%.

Diferentemente da região sul, os moradores da região nordeste, portanto, não contam com uma grande parcela na participação do chuveiro elétrico na fatura mensal de energia. No entanto, as lâmpadas passam a ter um maior papel de participação entre os gastos dos nordestinos.

Salienta-se também a grande diferença entre as porcentagens de participação da utilização do chuveiro na região sul (25%) e na região nordeste do país (9%). Dessa maneira, é evidente a grande diferença entre regiões com climas diferentes na utilização de alguns eletrodomésticos e eletrônicos.

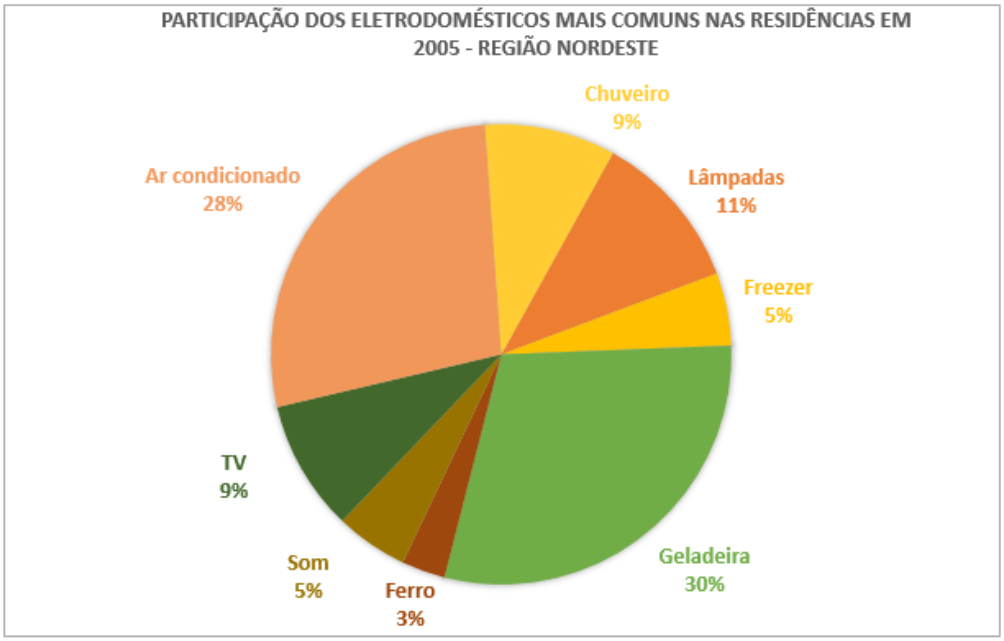

Figura 2.2: Participação de eletrodomésticos na Região Norte - ano 2015
Fonte: elaborada pela autora.

A Pesquisa de posse de equipamentos e hábitos de uso (PPEHU) do ano de 2005 é uma pesquisa de campo realizada pelo Procel através de questionários que tem como finalidade investigar, além da posse e utilização de equipamentos em nível nacional, algumas análises socioeconômicas e qualidade do fornecimento de energia.

Segundo a PPEHU do ano de 2005, 87% do consumo de eletricidade de uma residência padrão do Brasil poderia ser representada por um ar condicionado, congelador, refrigerador, lâmpadas, chuveiro elétrico, máquina de lavar roupas e televisão.

Essa pesquisa contou com a aplicação do questionário, instrumento utilizado em 9.847 pessoas e no item relativo aos refrigeradores conta com questionamentos a respeito da idade do eletrodoméstico, a intensidade do uso e se o usuário preocupa-se com o selo de referência do Procel, por exemplo.

A Pesquisa de posse de equipamentos e hábitos de uso serve de base para a construção de questionários mais simples, como o instrumento utilizado nesta dissertação.

2.2 Elaboração de Questionários Científicos

Atualmente, o interesse por conhecer opiniões e informações de grupos de indivíduos está ligado a diversas áreas do conhecimento, como Biologia, Medicina, Engenharia e uma série de outros setores. Através de pesquisas, portanto, é possível analisar indicadores de uma ampla gama de campos do conhecimento, que vão desde a detecção de uma doença até mesmo a qualidade e consumo de energia elétrica, por exemplo.

A pesquisa se mostra, assim, uma atividade voltada para a solução de problemas teóricos e práticos. Estudos exploratórios, por sua vez, são designados muitas vezes como o passo inicial

no processo de pesquisa pela experiência e um mecanismo auxiliar que traz a formulação de hipóteses para futuras pesquisas.

Utilizando-se de questionários para pesquisa científica, é imprescindível considerar que não basta apenas realizar a coleta de dados de maneira randômica sobre temas e questões de interesse, mas sim saber analisar adequadamente os resultados, com base em análises estatísticas.

Assim, questões como formas de análise dos dados, margem de erro e tamanho da amostra, por exemplo, devem ser levados em conta no momento da formação do banco de dados. A pesquisa, então, passa a ser enquadrada como de cunho qualitativo ou quantitativo, de acordo com Rutter e Sertório (1994).

As pesquisas qualitativas geralmente não podem ser mensuradas estatisticamente, mas mesmo assim atuam de maneira importante no auxílio às pesquisas quantitativas, que são utilizadas para tomada de opiniões, hábitos ou atitudes de uma população escolhida. A ação de planejar experimentos - chamada de *Design of Experiments* (DOE) segue, por sua vez, o princípio da circularidade do método científico.

As fases iniciais do princípio da circularidade do método científico dizem respeito às ações de determinar o problema e planejar uma população amostral. Após, deve-se desenvolver a e aplicar o questionário, bem como realizar a pesquisa de campo para a obtenção dos dados. Por fim, o ciclo se encerra com a análise estatística e com a análise dos resultados. A Figura 2.3 ilustra as ações do princípio da circularidade do método científico.

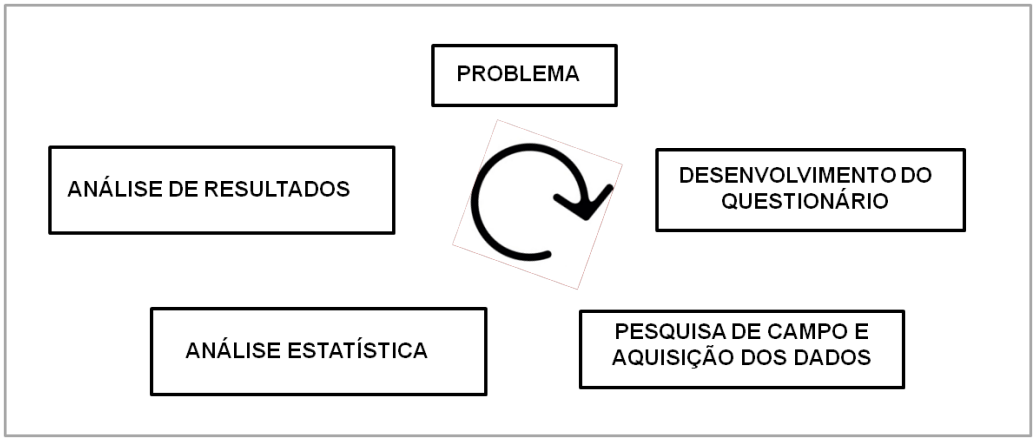

Figura 2.3: Princípio da Circularidade do Método Científico
Fonte: elaborada pela autora.

Em suma, a pesquisa teve início com a definição do problema a ser investigado demonstrando dos objetivos de maneira clara para que os experimentos ofereçam resultados claros e úteis e só então passar para a etapa de elaboração do questionário que será aplicado. Ao elaborar o questionário, é importante que o experimentador escolha as variáveis que irão compor o instrumento, incluindo todas as questões que acredita ser de importância para a pesquisa do problema de maneira simples e direta.

Na escolha das variáveis de resposta, ainda, é preciso escolher variáveis que realmente possam fornecer informações úteis sobre os resultados dos experimentos, evitando redundâncias ou

informações relapsas, bem como avaliar se a variável é de possível mensuração nas condições de realização do experimento.

Já na condução do experimento, deve haver um experimento de teste antes da aplicação dos questionários, para que haja a certeza de que o instrumento fornece as respostas desejadas. A condução de experimentos de teste são imprescindíveis para pesquisadores sem uma grande experiência em coleta de dados ou sem uma vasta experiência sobre os elementos a serem estudados.

A etapa de análise dos dados é importantes para uma melhor visualização e entendimento gráfico dos experimentos, mesmo quando os resultados são claros. A análise matemática e estatística auxilia os pesquisadores a conectar resultados e compreender melhor os elementos estudados. Após a análise dos dados, pode-se então inferir ou constatar resultados, o que é o principal objeto de estudo da pesquisa.

Neste trabalho, o experimento de teste foi realizado e se mostrou de grande valia para ajustes de possíveis erros e detalhes no questionário.

2.3 Mineração de Dados

Esta Seção apresenta um panorama sobre mineração de dados, seus passos, e algumas técnicas aplicáveis neste processo.

A mineração de dados, também conhecida como *data mining*, consiste no processo de descoberta conhecimentos de interesse em banco de dados, podendo ser de fundamental importância para revelar informações vitais às pesquisas. Segundo Witten, Frank e Hall (2005), essa descoberta é ainda mais valiosa quando se trata de uma descoberta em grandes conjuntos de dados, tratando de um campo de estudos que, combinando métodos tradicionais para analisar os dados com algoritmos, possibilita exploração e análise de inúmeros tipos de conjunto de dados.

Ademais, mineração de dados, como o próprio nome sugere, trata-se de uma exploração de dados de interesse em uma grande quantidade de dados em busca de padrões não perceptíveis por uma simples análise manual. Segundo Tan, Steinbach e Kumar (2009), padrões úteis que poderiam ser ignorados são descobertos através da aplicação de técnicas de mineração de dados.

O processo de descoberta de conhecimento costuma ser dividido em algumas etapas, a fim de aprimorar os resultados. De acordo com Witten, Frank e Hall (2005), esse processo se divide em três passos: exploração (onde estão incluídas todas as pré-transformações necessárias, como limpeza, seleção e integração dos dados), construção do modelo e validação do modelo.

As etapas de limpeza e seleção de dados são vistas como as fases em que os ruídos ou dados sem consistência com o conjunto são eliminados do banco.

Também é preciso que haja uma unificação de diferentes fontes em um único banco de arquivos, etapa conhecida como integração dos dados, não esquecendo da realização de uma seleção de atributos interessantes ao processo e da transformação dos dados em um formato apropriado. Esta fase, em sua totalidade, é conhecida como o pré-processamento dos dados.

A etapa de mineração, então, é precedida de diversas ações que contribuem para o sucesso da

aplicação de técnicas específicas. Esta etapa diz respeito a filtrar e encontrar padrões, levando em conta os atributos e objetos de interesse do conjunto de dados que será formado.

A fase de pós-processamento para validação do modelo gerado é o estágio em que os dados podem ser visualizados e são capazes de instigar a interpretação de padrões no conjunto de dados de entrada. Essas três fases juntas formam, basicamente o processo de descoberta de conhecimento em banco de dados que é chamado de maneira geral como *Knowledge Discovery in Databases* (*KDD*). Esse processo é esquematizado na Figura 2.4.

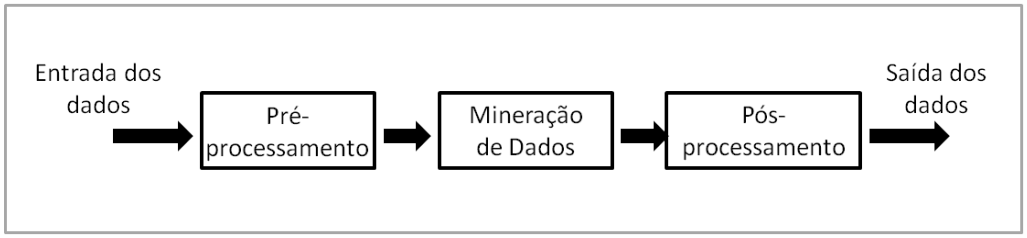

Figura 2.4: Simplificação do processo de *KDD*
Fonte: elaborada pela autora.

O *KDD* refere-se, assim, ao processo de transformar dados sem qualquer tratamento e relação em um conjunto de dados com informações úteis, através de uma seleção, preparação, pré-processamento, transformação, redução e adequação dos dados do conjunto.

A redução das dimensões do banco de dados, por exemplo, é dada pelo número de atributos que este conjunto de dados possui e durante o processo *KDD* de descoberta de conhecimento, modelos computacionais são construídos para a fim de tentar encontrar relações entre os dados ou atributos. Esse processo é realizado de maneira interativa por algoritmos.

As principais etapas do processo *KDD*, segundo Goldschmidt e Passos (2005), são descritas logo abaixo. Para melhor ilustrar estas etapas, tem-se a Figura 2.5, proposta por Fayyad, Piatetsky-Shapiro e Smyth (1996).

1. Seleção dos dados: consiste em analisar os dados coletados a fim de selecionar quais serão utilizados no processo.

2. Pré-processamento dos dados: consiste em realizar um tratamento no conjunto de dados para que possam ser interpretados pelos algoritmos. É nesta etapa que devem ser identificados e tratados valores faltantes ou inválidos.

3. Transformação ou formatação dos dados: utilização, quando necessário, de alguma transformação nos dados, de forma a encontrar aqueles mais relevantes para o problema que está sendo investigado.

4. Mineração dos dados: busca por padrões a partir da aplicação de algoritmos escolhidos.

5. Interpretação do resultado: análise dos resultados da mineração, a fim de aplicá-los no objeto de pequisa.

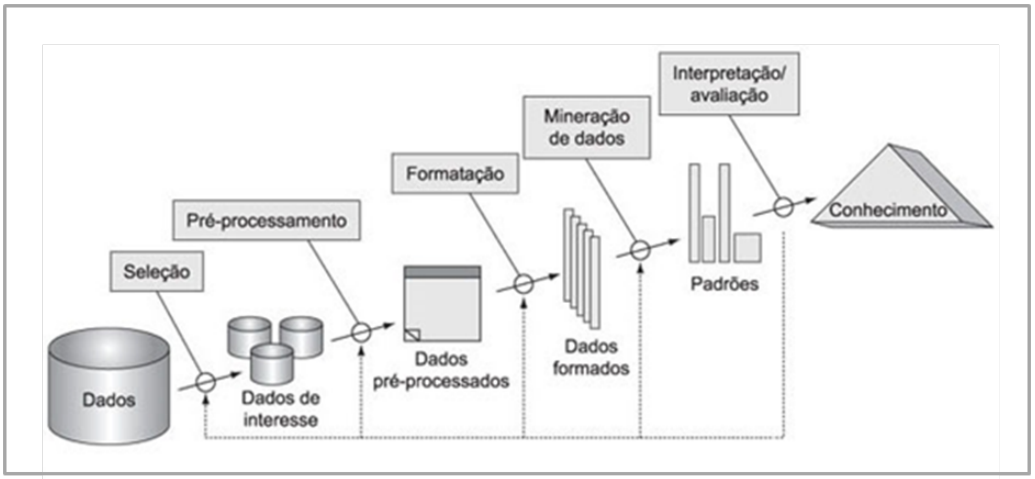

Figura 2.5: Etapas do processo de *KDD*
Fonte: Traduzido de Fayyad, Piatetsky-Shapiro e Smyth (1996).

A etapa de mineração é responsável pela descoberta de novas relações entre os dados por meio de procedimentos da área de aprendizado de máquina, através da análise sistemática sobre instâncias contidas em uma base de dados.

Assim, cabe dizer que as técnicas de mineração servem não somente para descobrir padrões já existentes, mas também podem fornecer previsão de observações futuras nos dados.

2.3.1 O Software Weka

Este trabalho utiliza a plataforma livre *Waikato Enviroment for Knowledge Analysis*[1] (*Weka*), que foi desenvolvida na Universidade de Waikato, na Nova Zelândia, e é um mecanismo composto por grupos de implementações e algoritmos de técnicas de mineração de dados.

O *Weka*, que foi desenvolvido na linguagem de programação *Java* e pode ser utilizado na maioria dos sistemas operacionais. A tela inicial do software é mostrado na Figura 2.6. Para dar início à mineração de dados, deve-se escolher a opção *Explorer*.

Segundo University of Waikato (2010), no *Weka* podem ser utilizados métodos de classificação como árvores de decisão, regras de aprendizagem, Teorema de Bayes, regressões, tabelas de decisão, entre outros. Este trabalho utilizou esta plataforma por ser livre, de fácil manuseio e por permitir a execução dos algoritmos de mineração de dados de forma interativa. Ainda, os resultados podem ser visualizados graficamente. Porém, poderiam ser obtidos os mesmos resultados em outras plataformas.

De acordo com Hall et al. (2009), a organização dos dados a serem analisados pode ser feita em planilha ou banco de dados. Assim, o *Weka* possui um formato com extensão *Attribute-Relation File Format* (ARFF), que é utilizado para que os dados sejam organizados e o arquivo possa ser carregado no software novamente.

[1] *Weka 3: Data Mining Software in Java* - `http://www.cs.waikato.ac.nz/ml/weka/`

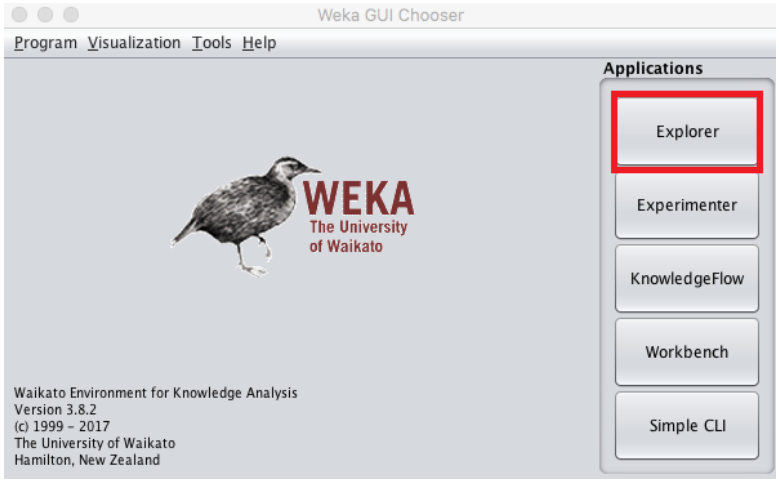

Figura 2.6: Tela inicial do *Weka*
Fonte: elaborada pela autora.

2.3.2 Conjuntos de Dados

A quantidade de dados gerada no mundo tomou proporções gigantescas com o crescente desenvolvimento da computação e, inclusive, como exposto por Witten, Frank e Hall (2005), a estimativa é que a cada 20 meses a quantidade de dados provenientes de diversas fontes de informação e campos de estudo seja duplicada.

Os conjuntos de dados (*datasets*) são formados por objetos, que podem representar seres animados ou inanimados e, assim, cada objeto tem sua representação descrita por uma série de atributos de entrada, também chamados de vetor de características por Faceli et al. (2011), determinando que cada atributo descreve uma propriedade dos objetos. Cabe informar que, na maioria das vezes, os dados não estão prontos após a coleta para utilização em algoritmos de mineração de dados, sendo necessária a aplicação de técnicas eficientes de pré-processamento.

Essas técnicas podem ser, por exemplo, de eliminação manual de atributos ou redução de dimensionalidade. A primeira consiste em excluir dados incompletos ou inconsistentes dos *datasets*, a fim de evitar problemas no desempenho dos algoritmos utilizados. Já a segunda diz respeito à diminuição de atributos associados a cada objeto, auxiliando no desempenho computacional e na clareza dos resultados.

A Tabela 1 mostra o exemplo de um conjunto de dados socioeconômicos. Nota-se que para cada aluno é associada uma série de atributos, como a identificação única (*ID*), o sexo, o ano de ingresso na universidade e o número de pessoas na residência. Os atributos podem ser quantitativos (numéricos) ou qualitativos (categóricos).

Atributos quantitativos podem passar por operações matemáticas, como é o caso do número de residentes na casa do estudante universitário. Esses atributos ainda podem ser contínuos ou discretos. Já os atributos qualitativos não são passíveis de operações aritméticas e podem representar, por exemplo, o sexo de uma pessoa. A Tabela 2 mostra a classificação dos tipos de atributos considerados no conjunto de dados.

A exploração de dados é processo no qual informações úteis podem ser extraídas de *datasets*.

Tabela 1: Conjunto de dados de universitários com seus atributos

ID	**Sexo**	***Ano de Ingresso***	**Número de residentes**
A1	F	2015	2
A2	M	2015	5
A3	M	2016	4
A4	F	2014	4

Fonte: elaborada pela autora.

Tabela 2: Classificação dos tipos de atributos utilizados

Dado	**Tipo**
ID	Qualitativo
Sexo	Qualitativo
Ano de Ingresso	Quantitativo discreto
Número de Residentes	Quantitativo discreto

Fonte: elaborada pela autora.

Essa análise é de fundamental importância para que se tenha um rumo sobre qual técnica de mineração de dados poderá ser escolhida. Com isso, a estatística descritiva que, segundo Faceli et al. (2011), assume que os dados são gerados por um processo estatístico, tem a função de resumir de maneira a quantificar os principais atributos do *dataset* utilizado e também classificar os modelos de algoritmos gerados.

2.4 Metologias e Técnicas

Segundo Carvalho (2005), a etapa de *data mining* pode ser realizada de três diferentes maneiras: descoberta não-supervisionada, teste de hipótese ou modelagem de dados. A descoberta não-supervisionada, para Carvalho (2005), diz respeito a deixar que os algoritmos encontrem as relações existentes entre os dados.

Em se tratando de tarefas de aprendizado, por exemplo, é abordado por Faceli et al. (2011) que podem ser divididas em tarefas preditivas e descritivas. Assim, é objetivo que em tarefas de previsão seja encontrada uma função que descreva o modelo a partir dos dados de treinamento e, assim, ser capaz de prever a classificação de um novo objeto, de acordo com seus atributos. Este tipo de tarefa é considerado aprendizado supervisionado. De acordo com Tan, Steinbach e Kumar (2009), o atributo que será previsto é denominado de atributo alvo ou, ainda, variável alvo ou dependente.

Em tarefas de aprendizados relacionadas à descrição, o principal objetivo pode ser encontrar características em comum para um agrupamento, ou ainda relacionar itens de informação que ocorrem juntos em múltiplas transações do banco de dados. Por exemplo, é possível descobrir quais produtos impactam na venda de outros. Este tipo de tarefa pode ser visto como um aprendizado não supervisionado, para Tan, Steinbach e Kumar (2009).

Já o teste de hipótese é utilizado quando já existe algum tipo de conhecimento sobre o

que devemos buscar através dos dados. Assim, pode ser definida uma hipótese inicial. A modelagem de dados, por sua vez, emprega-se esta metodologia quando se tem um conhecimento maior ainda a cerca da relação entre os dados. As técnicas de mineração de dados podem ser divididas basicamente em classificação (descritiva ou preditiva), análise de associação ou análise de agrupamentos.

Para Carvalho (2005), a classificação consiste na tarefa de organizar objetos em uma entre diversas categorias que já existem, sendo uma das técnicas mais utilizadas em mineração de dados e que envolve modelos de classificação de um conjunto de dados de entrada, podendo este modelo ser descritivo ou preditivo.

A abordagem descritiva evidencia os relacionamentos entre os dados, favorecendo pesquisas de caráter exploratório. Já a abordagem preditiva tem o intuito de prever o valor ou a característica de um atributo (chamado de variável dependente), baseando-se no valor de outros (variáveis independentes). Para a predição, podemos utilizar uma classificação, se as variáveis alvo forem discretas com argumento categórico ou uma regressão, se as variáveis alvo forem contínuas, como aborda Tan, Steinbach e Kumar (2009).

A associação é utilizada para encontrar padrões que descrevam características associadas dentro do conjunto de dados. As informações resultantes são mostradas ao usuário na forma de regras de associação ou implicação, de acordo com Tan, Steinbach e Kumar (2009).

Já a análise de agrupamento busca grupos, também chamados de *clusters*, a fim de agrupar os elementos que são mais semelhantes. Este processo é denominado de clusterização.

2.4.1 Construção de Modelos de Classificação e Classificadores

Cada técnica de classificação se baseia em diferentes algoritmos que podem se adaptar ou não ao conjunto de dados em questão e à classe de saída dos dados. Assim, segundo Tan, Steinbach e Kumar (2009), dizemos que a eficiência de um modelo gerado por um algoritmo depende de quão acertivo ele é ao identificar e predizer a qual classe pertencem instâncias desconhecidas.

Dessa forma, a principal meta, ao utilizarmos um algoritmo de aprendizagem, é que seja construído um modelo capaz de predizer rótulos de instâncias desconhecidas. Para exemplificar este conceito, Tan, Steinbach e Kumar (2009) propõe um mapeamento de atributos, conforme a Figura 2.7.

Em problemas envolvendo a técnica de classificação, deve ser utilizado um conjunto de treinamento, onde os rótulos sejam conhecidos pelo algoritmo. Este conjunto de treinamento é responsável pela construção de um modelo de classificação e, posteriormente, será utilizado em outro conjunto, denominado conjunto de teste.

Submete-se, então, o conjunto de treinamento a uma das técnicas de classificação, onde o modelo será construído e validado no conjunto de teste. Como mencionado anteriormente, dependendo do conjunto de dados inicial e do modelo, alguns algoritmos de aprendizado podem ser mais ou menos eficientes. Com isso, diferentes métodos trazem diferentes resultados, mais

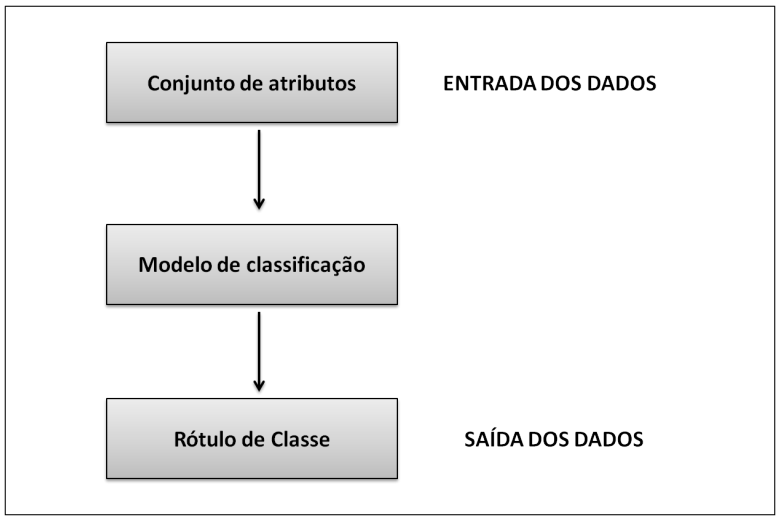

Figura 2.7: Construção de um modelo através de um conjunto de atributos
Fonte: adaptada de Tan, Steinbach e Kumar (2009).

ou menos satisfatórios para o que é desejado na pesquisa. Tan, Steinbach e Kumar (2009) aborda alguns principais tipos de classificadores.

Os classificadores baseados em árvores de decisão (*decision trees*) são determinísticos e evidenciam a construção de uma árvore conforme extrai informações dos conjuntos de dados. Como principais exemplos, temos o algoritmo *C4.5* implementado no classificador *J48* da plataforma *Weka*.

Os classificadores que se baseiam em redes neurais artificiais trabalham com as estatísticas dos dados e, por sua vez, os classificadores bayesianos são baseados em uma inferência probabilística. Existem, ainda, os classificadores baseados em regras, que realizam a classificação a partir de um conjunto de regras específico. Os classificadores de vizinho mais próximo trabalham com noções de proximidade e com o fato de que os elementos do conjunto de treinamento sejam parecidos com os de teste. Por fim, pode-se destacar como um último principal tipo de classificador os baseados em vetores de suporte. Estes classificadores baseiam-se fundamentalmente em teoria de aprendizado estatístico.

Para este trabalho, foram escolhidos quatro tipos de classificadores para os primeiros testes: baseados em árvores, baseados em vizinho mais próximo, baseados em Teorema de Bayes e em vetor de suporte. A seguir serão apresentados com mais detalhes estes quatro tipos.

2.4.2 Classificadores baseados em árvores de decisão e o Classificador C4.5

Para Holsheimer e Siebes (1991), uma árvore de decisão é constituída de nodos (representando os atributos), arestas e de nodos folha. As arestas fazem a ligação entre um nodo e um nodo folha, sendo responsáveis por carregar os possíveis valores do atributo. A Figura 2.8 mostra um exemplo de árvore de decisão simples, onde o atributo alvo é a classe sobre o hábito de comprar eletrodomésticos pela internet. Ainda, temos a idade sendo o fator mais discriminatório, ou seja, o fator que melhor separa as instâncias e, por isso, foi utilizado como

atributo raiz da árvore.

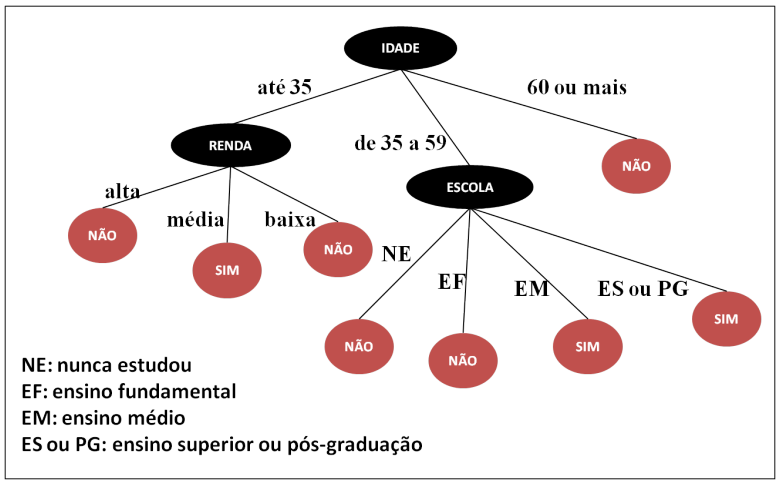

Figura 2.8: Modelo de árvore de decisão
Fonte: elaborada pela autora.

Note que, neste exemplo, se o indivíduo possui até 35 anos de idade, o caminho até as folhas depende ainda da renda. Já para o indivíduo com 60 anos ou mais, é afirmado que não há o hábito de compras pela internet. Para os indivíduos de 35 a 59 anos de idade, a resposta depende da escolaridade.

Por esta razão, diz-se que esta técnica é realizada através de particionamentos sucessivos. Dessa maneira, partindo-se de informações que são encontradas nos dados (conjunto de dados de entrada), o algoritmo procura o atributo mais relacionado à variável de saída até seu último nível, chamado de folha.

Os algoritmos baseados em árvores de decisão geram, além da árvore, um conjunto de regras associadas, porém Holsheimer e Siebes (1991) abordam que não existe uma árvore perfeita, apenas árvores consideradas mais ou menos adequadas de acordo com o que se deseja pesquisar.

Existe ainda o conceito de poda, que ocorre quando o algoritmo é interrompido antes da árvore resultante final, se a árvore gerada até o momento satisfizer os critérios do conjunto de treinamento (pré-poda), ou quando uma subárvore é substituída por um nó folha (pós-poda). Os algoritmos mais conhecidos que trabalham com esse conceito é o *Induction Decision Tree* (*ID3*) e o *C4.5*, como exposto por Witten, Frank e Hall (2005).

Segundo Han e Kamber (2006), o algoritmo *ID3* utilizou-se por muito tempo do critério de ganho de informações, que consiste em uma redução de entropia que é causada pelo particionamento mencionado. A entropia está relacionada ao grau de desordem de um sistema, neste caso interpretado como o grau de impureza das amostras do conjunto.

É importante salientar que o algoritmo *C4.5* trabalha com o critério de taxa de ganho de informações. O valor de ganho de informação (A) causa uma redução na entropia geral do sistema. Diz-que o ganho A pode ser determinado pela variação dos valores da entropia esperada e real para um ramo.

Para representar o modelo de classificação baseado em árvores de decisão, este trabalho

utilizará o classificador *J48*, implementado na plataforma *Weka*. De acordo com Witten, Frank e Hall (2005), os parâmetros para configuração e execução deste algoritmo estão simplificados na Tabela 3.

Tabela 3: Configuração dos parâmetros do *C4.5* no *Weka*

Parâmetro	**Descrição**
ConfidenceFactor	Fator de confiança a ser utilizado para poda da árvore
Unpruned	Utilizado para desativar a poda
MinNumObj	Número mínimo de instâncias por folha da árvore

Fonte: elaborada pela autora.

2.4.3 Classificadores baseados em Teorema de Bayes e o Classificador Naïve Bayes

Para alguns exemplos, o rótulo da classe de uma instância que está sendo testada pode não ser determinada com total certeza, mesmo que o conjunto de atributos do teste seja igual ao do treinamento.

Isso pode acontecer, por exemplo, quando há ruídos que afetam o processo de classificação e não foram detectados no pré-processamento de dados. Esta situação pode ser amenizada ou resolvida pelos algoritmos baseados no Teorema de Bayes, que classificam os elementos de acordo com a probabilidade de eles pertenceram a determinada classe.

É importante apresentar o Teorema de Bayes, que relaciona probabilidades de X e Y com suas probabilidades condicionais. A probabilidade condicional pode ser rapidamente explicada como a probabilidade de um evento ocorrer, dada uma certa condição. O Teorema de Bayes calcula a probabilidade *a posteriori* de acordo com a Equação 2.2, onde *P(X)* e *P(Y)* são probabilidades *apriori*, ou seja, que dizem respeito a ocorrência de eventos independentes.

$$P(X|Y) = \frac{P(X)P(Y|X)}{P(Y)} \tag{2.2}$$

Classificadores baseados no Teorema de Bayes rotulam as instâncias de um conjunto de dados com a classe que maximiza a probabilidade *a posteriori* calculada. O classificador Naïve Bayes é dito ingênuo porque supõe que os atributos são independentes.

2.4.4 Classificadores baseados em vizinho mais próximo e o Classificador K*

Os classificadores baseados em exemplos (ou instâncias) classificam os exemplos de teste de acordo com as classes dos exemplos de treinamento supondo que exemplos similares compartilham a mesma classe. Os algoritmos de vizinho mais próximo são baseados em exemplos e utilizam as mais variadas funções de distância, tais como Manhattan ou Euclidiana.

Cover e Hart (1967) consideram os algoritmos de vizinhos mais próximos os mais simples algoritmos baseados em instâncias e, dessa maneira, utilizam uma função de distância para atribuir uma classe à instância em questão.

Segundo Tan, Steinbach e Kumar (2009), os classificadores baseados na técnica de vizinho mais próximo podem ser analisados em dois passos: o indutivo e o dedutivo. O passo indutivo consiste na construção do modelo a partir dos dados e o passo dedutivo, por sua vez, na aplicação do modelo a exemplos de teste. Assim, tais classificadores encontram todos os exemplos de treinamento que sejam semelhantes aos exemplos de teste, sendo cada exemplo um ponto de dado em um espaço de dimensão d, correspondendo d ao número de atributos.

A classificação realizada por esses algoritmos é baseada nos rótulos de classe dos seus vizinhos mais próximos, cujo número é denominado por k. Dessa maneira, quando k=3, estamos nos referindo à seleção de 3 vizinhos mais próximos. Se a zona delimitada pelos vizinhos mais próximos for composta de mais de um rótulo de classe diferente, a classificação é realizada com base na classe majoritária.

Se tivermos um número de vizinhos mais próximos k muito pequeno pode ocorrer um *overfitting*, que é um superajuste do modelo aos dados utilizados. Também é preciso considerar que, se o número de vizinhos mais próximos é muito grande, a instância de teste pode ser classificada erroneamente, visto que serão incluídos pontos longe do que seria uma verdadeira vizinhança (*underfitting*).

A Figura 2.9 mostra um exemplo em que o número de vizinhos k=3 e a instância na cor preta seria classificada como pertencente à classe dos círculos de cor laranja.

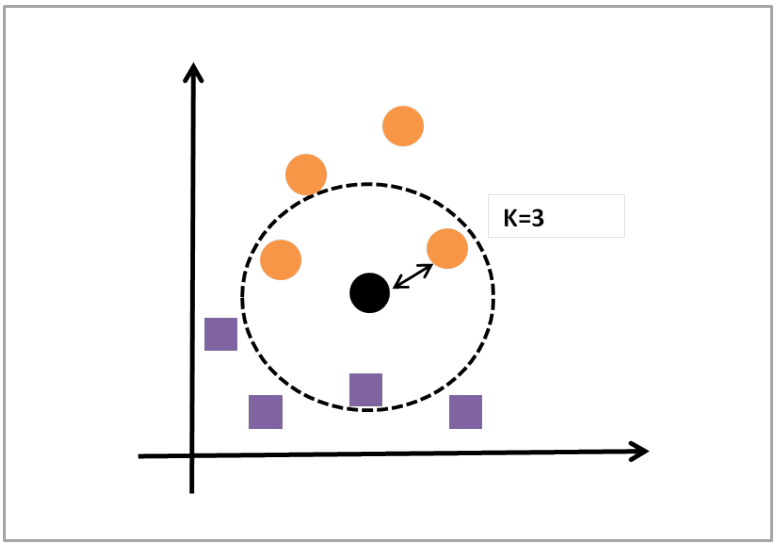

Figura 2.9: Exemplo de classificação por vizinhos mais próximos
Fonte: elaborada pela autora.

Como mencionado, é visível que para avaliar a proximidade ou distância dos elementos, são necessárias funções de distância. O classificador K* (k-estrela) utiliza a medida de distância com base na entropia, que mede o grau de desordem do sistema. Quando considerada a utilização no *Weka*, a Tabela 4 mostra a configuração dos parâmetros do K*.

Tabela 4: Configuração dos parâmetros do K* no *Weka*

Parâmetro	Descrição
Globalblend	Relacionado ao uso de cálculos de entropia, de 0 a 100 (B=20)
Missingmode	Como os valores de atributos faltantes são tratados, utilização padrão

Fonte: elaborada pela autora.

2.4.5 Classificadores baseados em vetor de suporte e o classificador SMO

Máquinas de vetores de suporte, do inglês *Support Vector Machine* (SVM), são redes neurais artificiais que elevam as dimensões dos dados de entrada em busca de um espaço linearmente separável entre duas classes. A predição é baseada em um hiperplano de uma dimensão a menos, cuja a margem de separação que divide efetivamente as duas classes está apoiada sobre as instâncias denominadas vetores de suporte. O algoritmo busca encontrar o hiperplano que maximiza essa margem.

A Figura 2.10 apresenta um exemplo com um espaço bidimensional, mostrando a margem e o hiperplano (reta).

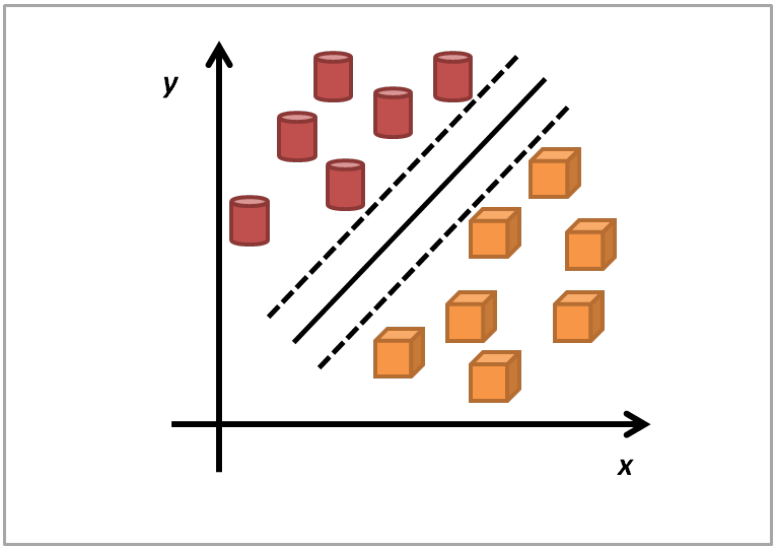

Figura 2.10: SVM aplicado a dados bidimensionais linearmente separáveis
Fonte: elaborada pela autora.

Sequential Minimal Optimization (SMO) é um classificador disponível no Weka que implementa um caso específico de SVM que utiliza menos recurso computacional para elevar as dimensões, ou seja, ele possui complexidade linear em relação a quantidade de memória.

2.4.6 Validação Cruzada

Para tarefas de classificação, é necessário realizar a escolha dos conjuntos de dados de teste e conjunto de dados de treinamento. O conjunto de dados de treinamento diz respeito aos dados que são utilizados para a construção do modelo de classificação. Já o conjunto de dados de teste é utilizado validar o modelo e apontar acertos ou erros.

A funcionalidade de testes de classificadores do *Weka* está no menu *test options*. Segundo Kirkby (2004), a opção *use training set* utiliza o mesmo conjunto para treinamento e teste. O teste pode ser realizado com um conjunto de dados externo selecionado *supplied training set*. Se o usuário desejar que o conjunto de dados seja dividido aleatoriamente em duas partes (treino e teste), *percentage split* permite ao usuário fazer uma escolha em termos de porcentagem. Já a validação cruzada (*cross-validation*) se baseia em um algoritmo que particiona o *dataset* em diversos *folds* de forma aleatória. Após, seleciona uma das partições para teste, enquanto as demais são utilizadas para treinamento. O processo é repetido a fim de garantir que todas as instâncias sejam usadas no teste, mas que nunca estejam no treinamento. O número padrão de partições do software *Weka* é 10, mas este valor pode ser alterado.

A Figura 2.11 mostra o menu do *Weka* com as funcionalidades mencionadas.

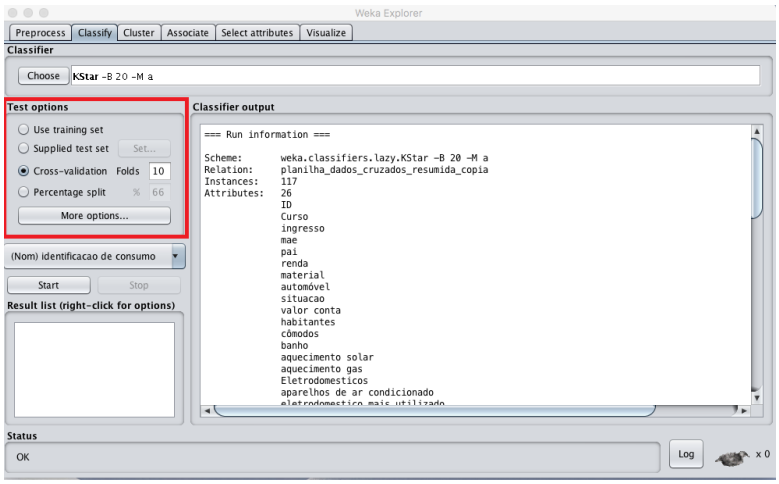

Figura 2.11: Tela do *Weka* com as opções mencionadas
Fonte: elaborada pela autora.

2.5 Avaliação do desempenho de modelos de classificação

Desde o início das civilizações, já eram registrados dados importantes para a humanidade, como o número de nascimentos e óbitos. Mais tarde esse controle de variáveis seria conhecido como Estatística, tornando-se uma ciência utilizável em todas as áreas do conhecimento e imprescindível para a medição de interações coletivas, entendendo como o coletivo se relaciona e qual o comportamento ele descreve, através de uma pesquisa científica.

Assim, a estatística deve ser vista como uma ferramenta essencial para análises e tomadas de decisão, auxiliando a interpretação de dados. A estatística é usualmente dividida em duas grandes áreas: descritiva e indutiva (inferencial).

O objetivo da Estatística Descritiva é resumir as principais características de um conjunto de dados por meio de tabelas, gráficos e resumos numéricos. Descrever os dados pode ser comparado ao ato de tirar uma fotografia da realidade.

A análise estatística deve ser extremamente cuidadosa ao escolher a forma adequada de resumir os dados. Podem ser consideradas tabelas de frequência, que servem para agrupar

informações de modo que estas possam ser analisadas. As tabelas podem ser de frequência simples ou de frequência em faixa de valores. O imprescindível para a utilização da estatística é saber o que se está procurando para então aplicar as análises.

A estatística descritiva também pode contar com gráficos, cujo objetivo é dirigir a atenção do analista para alguns aspectos de um conjunto de dados. Alguns exemplos de gráficos são: diagrama de barras, diagrama em setores, histograma, ramo-e-folhas, diagrama de dispersão, gráfico sequencial.

Em alguns casos, também são utilizados resumos numéricos, de onde podemos levantar importantes informações sobre o conjunto de dados tais como: a tendência central, variabilidade, simetria, valores extremos, valores discrepantes, etc.

Já a Estatística Inferencial, também chamada de Estatística Indutiva, utiliza informações incompletas para tomar decisões e tirar conclusões. O alicerce das técnicas de estatística inferencial está no cálculo de probabilidades. Duas técnicas de estatística inferencial são as mais conhecidas: a estimação e o teste de hipóteses.

A técnica de estimação consiste em utilizar um conjunto de dados incompletos, ao qual iremos chamar de amostra, e nele calcular estimativas de quantidades de interesse. Estas estimativas podem ser pontuais (representadas por um único valor) ou intervalares. O Teste de hipóteses tem como fundamento levantar suposições acerca de uma quantidade não conhecida e utilizar, também, dados incompletos para criar uma regra de escolha.

A Figura 2.12 representa um esquema explicativo sobre os dois principais campos da estatística.

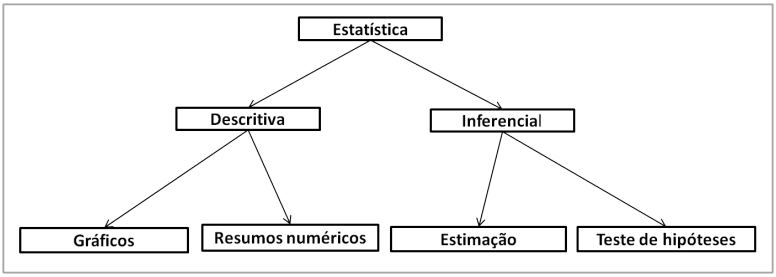

Figura 2.12: Principais campos da estatística
Fonte: elaborada pela autora.

Realizada esta análise preliminar de conceitos estatísticos, é importante para esta dissertação focar em alguns termos utilizados por esta ciência para quantificar e avaliar classificadores. Esses termos são utilizados para medidas de desempenho de classificação e podem dizer respeito a investigação de uma característica, uma doença, um estado emocional e infinitas outras variáveis.

2.5.1 Matriz de Confusão

A matriz de confusão, segundo Story e Congalton (1986), é uma matriz quadrada que é utilizada para demonstrar a quantidade de classificações corretas e classificações incorretas que

um algoritmo efetuou. Assim, é possível visualizar se o classificador está tendo uma baixa acurácia por causa de erros associados a cada classe individualmente.

A matriz mostra ao usuário na horizontal a classe real de um objeto e na vertical a classe predita pelo classificador. A Tabela 5 mostra uma matriz de confusão que investiga se um animal é um gato pelas suas características físicas, por exemplo.

Tabela 5: Matriz de confusão para identificação de gatos

Gato	Cão	←Classificado como
15	10	Gato
20	30	Cão

Fonte: elaborada pela autora.

Sendo assim, observa-se que:

1. O modelo classificou corretamente 15 instâncias como gatos, pois realmente eram gatos;
2. O modelo classificou incorretamente 10 instâncias como cães, pois eram gatos;
3. O modelo classificou incorretamente 20 instâncias como gatos, pois eram cães;
4. O modelo classificou corretamente 30 instâncias como cães, pois realmente eram cães.

Considerando a classe Gato como a de interesse, as 15 instâncias classificadas como gato que realmente eram gato são chamadas de verdadeiros positivos (*True Positives – TP*). Os 20 cães incorretamente classificados como gatos são falsos positivos (*False Positives – FP*). Os 30 cães identificados corretamente são verdadeiros negativos (*True Negatives – TN*) e os 10 gatos identificados incorretamente como cães são falsos negativos (*False Negatives - FN*).

Dessas conclusões, é possível montar uma matriz confusão genérica, conforme a Tabela 6. Nota-se que a soma de todos os elementos da diagonal principal é o número de instâncias classificadas corretamente pelo classificador.

Tabela 6: Matriz de confusão genérica

Sim	Não	←Classificado como
TP	FN	Sim
FP	TN	Não

Fonte: elaborada pela autora.

Com base nas frequências da matriz de confusão diversas métricas de avaliação de resultados podem ser derivadas. A próxima seção especifica as utilizadas nesta dissertação.

2.5.2 Métricas de avaliação

A Acurácia (A_c) diz respeito à porcentagem de acerto do classificador. Para seu cálculo, basta dividir quantidade de instâncias classificadas corretamente pela quantidade total, conforme Equação 2.3.

$$A_c = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

A Precisão (*Precision – P*) identifica qual a porção de classificações positivas estão corretas e pode ser representado pela Equação 2.4. Essa medida é dada pela razão entre a frequência de verdadeiros positivos e todos os positivos (verdadeiros e falsos). Quanto maior a precisão do modelo gerado, melhor o resultado da classificação.

$$P = \frac{TP}{TP + FP} \tag{2.4}$$

A Revocação (*Recall – R*) ou Sensibilidade (S) identifica qual a porção das instâncias que realmente são positivas foram identificados corretamente e é representado pela Equação 2.5.

$$R = S = \frac{TP}{TP + FN} \tag{2.5}$$

A Especificidade (E) é análoga ao *Recall* mas para a classe Negativa, conforme equação 2.6.

$$E = \frac{TN}{TN + FP} \tag{2.6}$$

A Medida F, (*F-Measure*) combina valores de Precisão e *Recall* de modo a trazer um número para apontar a qualidade geral do modelo de classificação. Sua expressão é dada pela Equação 2.7.

$$F = 2.\frac{P.R}{P + R} \tag{2.7}$$

2.5.3 Estatística *Kappa*

A Estatística *Kappa* (K), ou número *Kappa*, é uma maneira de medir a concordância nos algoritmos e é utilizada em modelos de classificação. Para Castro (2010), a estatística *Kappa* é fundamental para, além de medir a qualidade dos modelos de classificasção gerados, também medir quanto a resposta se afasta do que era esperado, desconsiderando o acaso e tornando as interpretações legítimas.

Segundo Pellucci et al. (2011), essa métrica é de fundamental importância ao analisar-se o modelo de classificação, pois avalia o nível de concordância dos dados dentro da base. O número *Kappa* pode variar de 0 até 1 e quanto mais próximo de 0 ele for, maior será o nível de discordância. Já quando o número *Kappa* vai se aproximando de 1, é indicada uma maior concordância entre os dados.

O valor do coeficiente de *Kappa* pode ser calculado pela Equação 2.8, onde P_o é a taxa de aceitação relativa, onde P_e é a taxa hipotética de aceitação, ou taxa de concordância randômica. Para o cálculo do P_o, divide-se o número de concordâncias pelo número total de instâncias.

$$K = \frac{P_o - P_e}{1 - P_e} \tag{2.8}$$

Dessa maneira, os valores de *Kappa* podem ser interpretados de acordo com a Tabela 7, proposta por Landis e Koch (1977). Nota-se que, conforme mencionado, quanto mais próximo de 1 o valor de K, maior o nível de concordância.

Tabela 7: Avaliação da concordância de acordo com o número de *Kappa*

Número *Kappa*	Nível de concordância
0,00	Sem concordância
0,01 - 0,20	Leve
0,21- 0,40	Aceitável
0,41- 0,60	Moderada
0,61- 0,80	Substancial ou boa
0,81- 1,00	Quase perfeita

Fonte: Adaptado de Landis e Koch (1977).

3 METODOLOGIA

Este capítulo descreve as etapas do processo de descoberta de conhecimento em bancos de dados utilizadas no desenvolvimento deste trabalho. É abordada primeiramente a coleta e a seleção de dados através da aplicação de um questionário desenvolvido, abrangendo todas as modificações essenciais nos dados. Também é utilizada a aplicação da técnica de mineração de dados conhecida como classificação.

3.1 Pré-processamento: Seleção e formatação dos dados

Conforme ilustrado por Fayyad, Piatetsky-Shapiro e Smyth (1996) e abordado na Seção 2.3, as etapas de seleção e formatação de dados compõem a etapa de pré-processamento no processo de *KDD*. Essas etapas consistem em elencar os atributos que serão minerados, a fim de apenas considerar dados relevantes e também em como adequar o banco de dados, excluindo instâncias com informações faltantes e monitorando o formato das variáveis, por exemplo.

Com base nos questionários já mencionados na Seção 2.2, foi dado início ao estudo das variáveis que apontavam forte relação com a varável alvo dessa dissertação: o consumo de energia elétrica. Assim, a elaboração do instrumento foi embasada na pesquisa de posse de eletrônicos e equipamentos domésticos, passando por lapidações para validação do instrumento.

Após a montagem do questionário, este passou pela etapa de validação de constructo, explicada com mais detalhes na Seção 2.2. Nesta fase, dois profissionais da área de Engenharia Elétrica identificaram, no questionário de teste, quais atributos poderiam indicar uma relação com o valor da conta de energia elétrica mensal.

Nesse sentido, foi elaborado o questionário piloto, que foi testado diversas vezes e, no momento em que encontrava-se pronto, foi submetido online no *Googleforms*[2]. O *Googleforms* é uma ferramenta livre que permite que, através de um endereço *web*, um formulário seja preenchido e as respostas sejam enviadas ao administrador responsável pela coleta dos dados. O administrador do questionário pode inserir as questões de maneira fácil e interativa, escolhendo o tipo de resposta que deseja obter.

Com o auxílio de professores, o endereço *web* foi divulgado para os alunos de diversos cursos da FURG e da UFPel, sendo solicitado que as perguntas fossem respondidas com a maior exatidão possível. Foram gerados dois questionários exatamente iguais para alunos das duas instituições e, após coleta, as informações foram unificadas em uma única base de dados.

Uma funcionalidade importante desta ferramenta online é a opção de exportar os dados dos respondentes em formato de planilhas eletrônicas. Isso permite que seja mais fácil o manuseio dos dados.

No momento da elaboração do instrumento, foi necessário escolher o tipo de resposta: única, múltipla, caixa de texto, ou outros. Na caixa de texto o usuário tem a liberdade de digitar sua resposta.

[2] *Googleforms* - `https://www.google.com/forms`

A Figura 3.1 mostra duas questões que necessitaram de caixas de texto como resposta, sendo elas o valor pago, em reais, na conta de energia elétrica no último mês e também o número de cômodos da residência. Em ambos os casos, essa é a melhor opção por se tratar de uma resposta que pode incluir uma grande diversidade de valores.

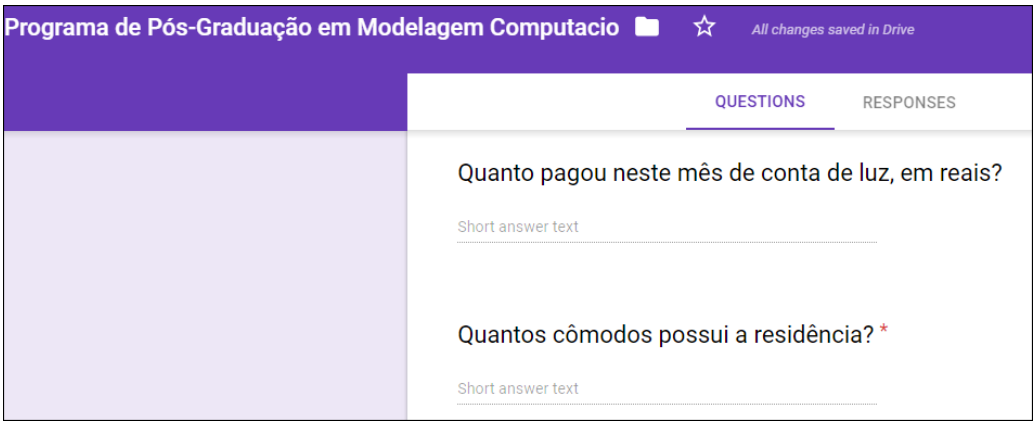

Figura 3.1: Perguntas que requerem caixa de texto
Fonte: elaborada pela autora.

Na escolha de *radio button* só é possível marcar uma alternativa já existente no instrumento. Essa modalidade de resposta foi a mais utilizada. Optou-se, sempre que possível, por esse formato de captura por apresentar uma facilidade e rapidez no momento de responder: o usuário pode, facilmente, apenas clicar na resposta correta. A Figura 3.2 mostra uma das perguntas do questionário com alternativas de resposta única, utilizando *Radio Button*.

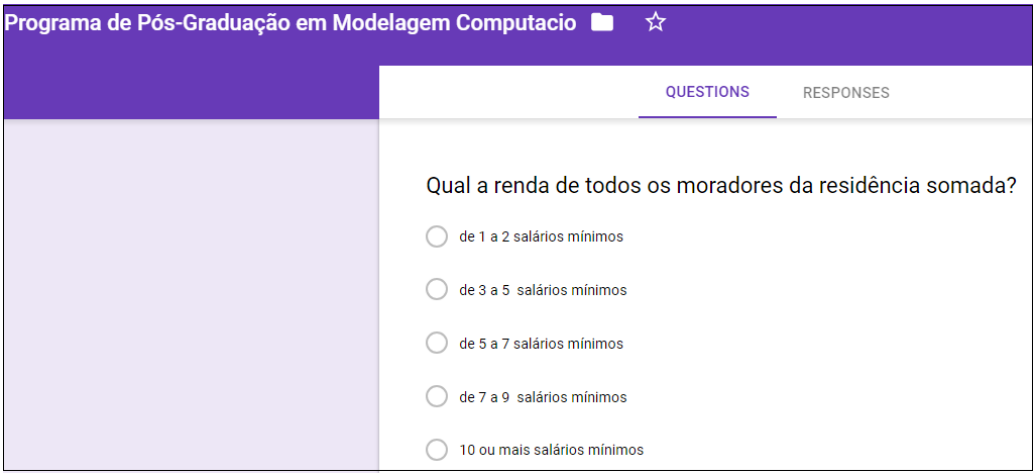

Figura 3.2: Perguntas que requerem *radio button*
Fonte: elaborada pela autora.

Cabe salientar que foi utilizado outro recurso do questionário em uma das questões envolvendo captura por escolha única: a inserção de imagens para exemplificar a resposta. Este recurso adicional é mostrado na Figura 3.3, onde é possível exemplificar o que é uma etiquieta do equipamento e o que é o selo Procel.

Ainda foi utilizada a captura das respostas por *checkbox*, que partindo de alternativas já

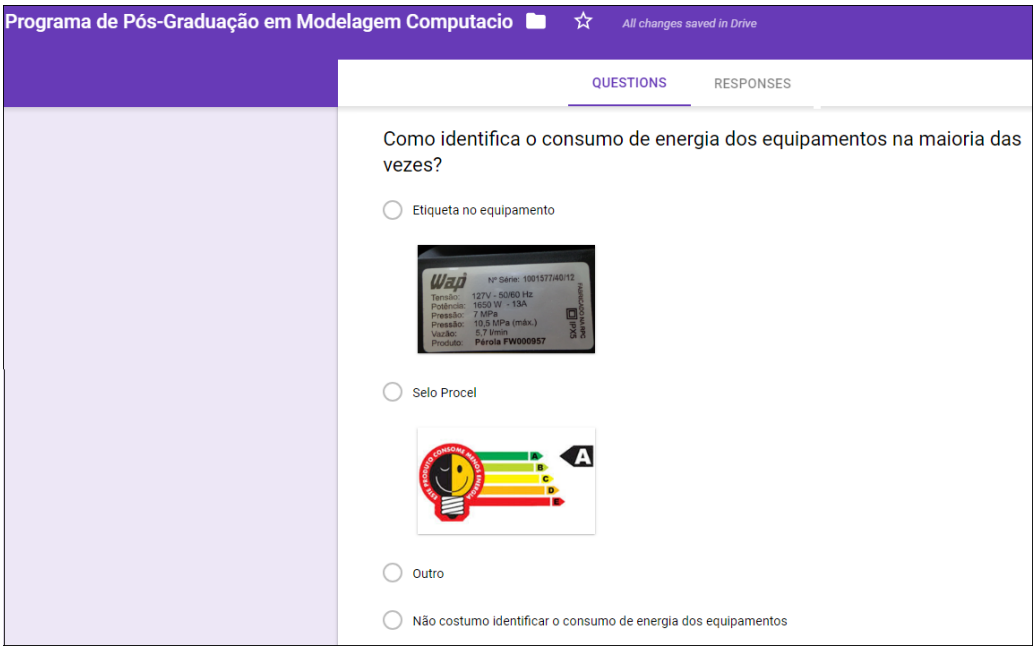

Figura 3.3: Representação gráfica no questionário
Fonte: elaborada pela autora.

pré-estabelecidas, permite que o respondente selecione múltiplas respostas corretas. Este caso foi utilizado para que os respondentes marcassem todos os eletrodomésticos ou eletrônicos utilizados pelo menos uma vez no período de 15 dias. A Figura 3.4 mostra a utilização deste tipo de resposta no questionário.

A Tabela 8 mostra os 26 atributos investigados e uma breve descrição de cada, informando se a resposta foi capturada pelo questionário através da utilização de uma caixa de texto, em formato de escolha única (utilizando-se o *Radio Button*) ou em formato de múltipla escolha (*checkbox*). Foi definido também o tipo de variável que foi tratada no questionário. As variáveis analisadas foram classificadas como qualitativas ou quantitativas.

As variáveis qualitativas podem ser qualitativas nominais (QN) ou qualitativas ordinais (QO). Já as variáveis quantitativas podem ser quantitativas contínuas (QC) ou quantitativas discretas (QD). O questionário pode ser conferido na íntegra no Anexo 8.1 deste trabalho.

Nota-se que para a primeira variável, denominada *ID*, tem-se uma captura nomeada manual. Resolveu-se assim indicá-la pois este atributo não estava originalmente neste formato ao serem constituídas as instâncias do questionário conforme preenchido pelos respondentes.

Em princípio, a plataforma utilizada cria um campo inicial com a data e horário da resposta. Para fins de formatação, essas informações foram, então, substituídas por um número de identificação (*ID*).

É importante observar que esses 26 atributos podem estar relacionados com o consumo de energia elétrica porém, nesta fase, ainda não é possível ter certeza da influência de cada um deles sobre a variável investigada. Como a descoberta de conhecimento é um processo iterativo, dependendo do resultado da etapa de mineração, os atributos podem ser removidos ou selecionados de diferentes formas.

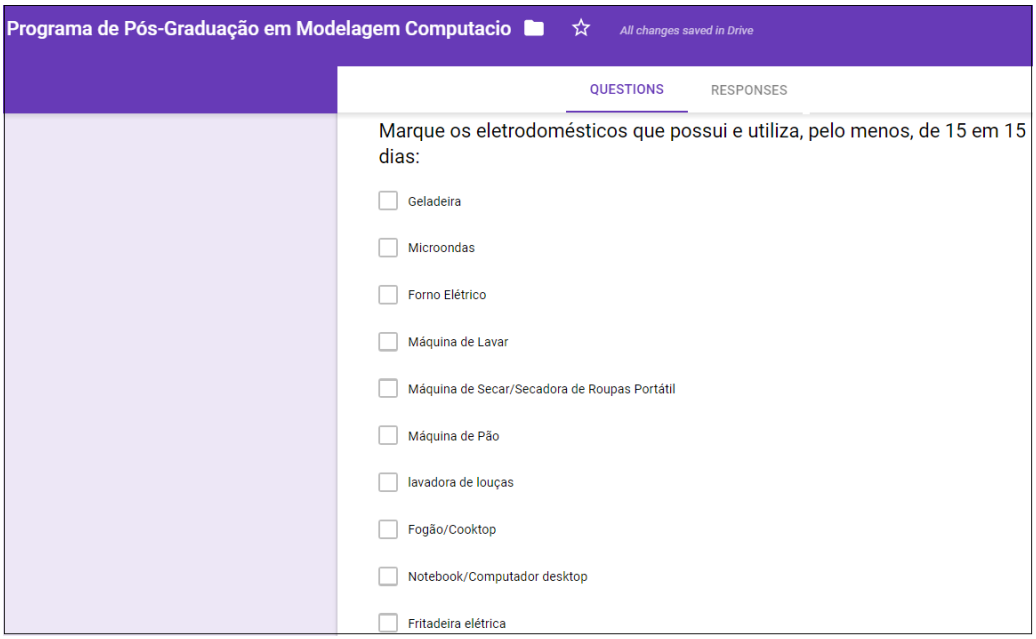

Figura 3.4: Perguntas que requerem *checkbox*
Fonte: elaborada pela autora.

Após o período de coleta das respostas, os dados foram exportados em formato de planilha para que fosse dado início à formatação dos mesmos. Ao final deste processo, foram obtidas 117 respostas. Logo em uma primeira análise, já foi possível perceber que os alunos escreveram de diferentes maneiras os mesmos cursos, por exemplo. Assim, quando um campo do questionário é coletado em formato de *string*, a atenção ao passo de formatação dos dados deve ser redobrada.

Também foram perceptíveis as diferentes maneiras que os usuários informaram o valor aproximado da conta de energia elétrica. Alguns escreveram palavras juntamente com os valores em reais, como “aproximadamente” , “mais ou menos” e “reais”. Também houve divergência no mecanismo separador: em alguns casos os alunos separaram os reais dos centavos através de uma vírgula e, em outros, através de um ponto. Pode-se dizer ainda que o cifrão demonstrativo de reais (R$) apareceu em alguns casos e em outros não.

Devido a esses fatores, após a seleção e coleta dos dados, foi necessária uma exaustiva etapa de formatação a fim de preparar os dados para o processo de processamento. Esta etapa, primeiramente, retirou a primeira coluna do formulário, que consistia em uma informação a respeito do dia e horário em que o respondente preencheu o instrumento e adicionou um campo para identificação de usuário. Ainda, foram padronizados os nomes dos cursos dos estudantes respondentes, aderindo à norma da língua portuguesa e eliminando abreviações pois, por exemplo, o fato de alunos do curso de Engenharia de Controle e Automação identificarem seus cursos apenas por “controle” ou “ECA”.

O próximo passo foi a padronização do valor da conta de energia mensal dos respondentes. A formatação foi realizada através da retirada dos valores de centavos e retirada do cifrão ou de palavras juntamente ao valor informado. Já para a coluna referente à frequência de utilização do eletrodoméstico mais utilizado, adotou-se o padrão de 0 a 7, correspondente ao número de

Tabela 8: Informações sobre as variáveis interrogadas

Variável	**Descrição**	**Resposta**	**Tipo**
ID	Identificação atribuída para cada aluno	Manual	QN
Curso	Curso do respondente na Universidade	Texto	QN
Ano	Ano em que foi realizada a matrícula no curso	Texto	QO
Mãe	Grau de instrução da mãe do respondente	Única	QO
Pai	Grau de instrução do pai do respondente	Única	QO
Renda	Número de salários mínimos de todos os moradores	Única	QO
Material	Tipo de material da residência do respondente	Única	QN
Automóveis	Existência ou não de automóveis na residência	Única	QN
Situação	Se a residência é alugada, cedida, própria ou outro	Única	QN
Valor	Valor da última conta de energia elétrica (reais)	Texto	QC
Residentes	Número de residentes da habitação	Única	QD
Cômodos	Número de cômodos da residência	Única	QD
Banho	Tempo médio de banho dos residentes	Única	QO
Solar	Existência ou não de aquecimento solar na residência	Única	QN
Gás	Existência ou não de aquecimento a gás na residência	Única	QN
Eletro	Eletrodomésticos utilizados na residência	Múltipla	QN
Ar	Número de aparelhos de ar condicionado na residência	Única	QD
Eletromais	Eletrodoméstico mais utilizado	Texto	QN
Frequência	Periodicidade em que é utilizado (campo acima)	Texto	QD
Bairro	Bairro de localização do domicílio	Texto	QN
Maiorfluxo	Turno com maior número de pessoas na residência	Única	QO
Menorfluxo	Turno com menor número de pessoas na residência	Única	QO
Tipo	Se o domicílio é casa, apartamento ou outro	Única	QN
Lâmpadas	Se ainda há lâmpadas incandescentes no domicílio	Única	QN
Peso	Peso da energia elétrica no orçamento da residência]	Única	QO
Identificação	Como o consumo dos equipamentos é identificado	Única	QN

Fonte: elaborada pela autora.

dias na semana que os moradores utilizavam esse eletrodoméstico.

Outro campo que necessitou de formatação foi o campo correspondente ao bairro, que também foi colocando nos padrões da língua portuguesa em alguns casos.

A partir deste momento, considerou-se a base de dados pronta para inserção no *Weka*. É permitido que planilhas sejam importadas no próprio formato, ou em *Comma-separated values* (CSV). Após vinculação da planilha contendo a base de dados ao software, foi realizada uma divisão em classes da variável que corresponde ao valor da conta de energia elétrica, através da aplicação do filtro *discretize*. Assim, as instâncias foram separadas em classes cujos elementos são classificados por um valor mais próximo da conta de energia elétrica. Este filtro permite que o usuário escolha o número de classes em que uma variável será representada.

Ao escolher o número de classes, pode-se optar, por exemplo, por uma distribuição igualitária de frequência em cada classe. Este software não permite que os limites inferiores e superiores das classes sejam editados, sendo estes fornecidos pelo próprio *Weka*. A Figura 3.5

mostra o menu de aplicação deste filtro.

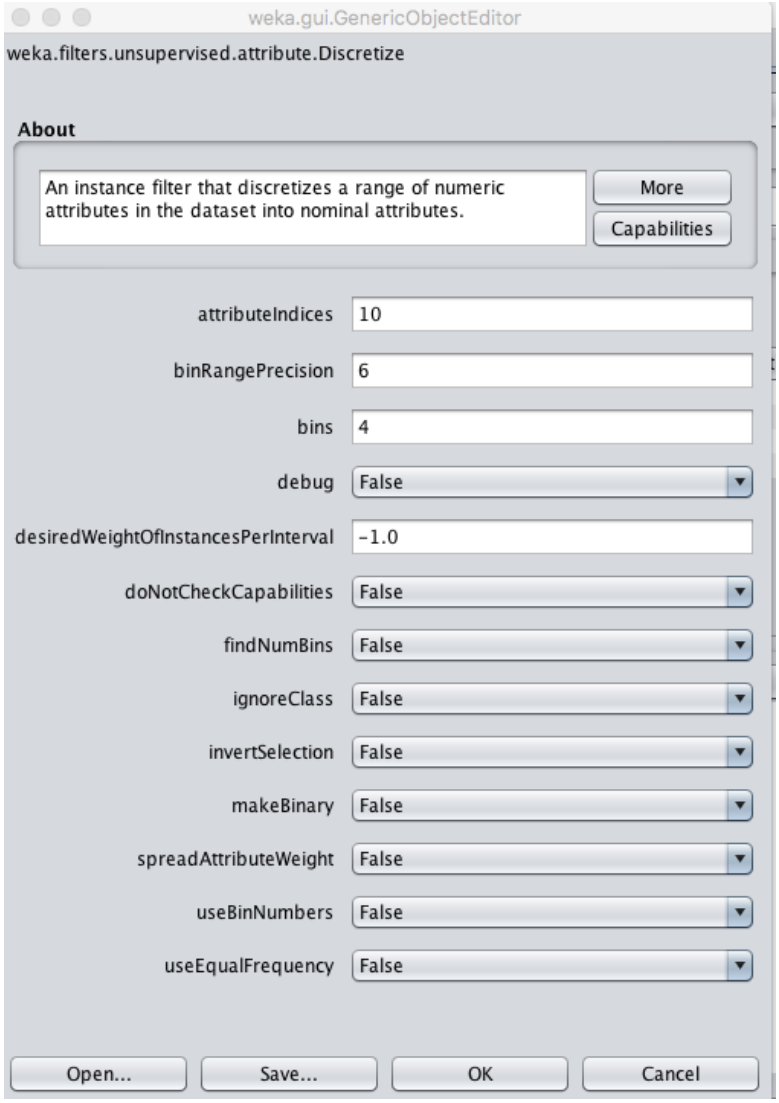

Figura 3.5: Menu para discretização de atributos
Fonte: elaborada pela autora.

Note que é possível atribuir o valor *true* or *false* à *UseEqualFrequency*, atribuindo uma igual distribuição de frequências. Em *bins*, é possível escolher o número de classes e em *atributeindices* é possível selecionar um ou mais índices referentes às variáveis para serem discretizadas.

Primeiramente, alguns algoritmos serão testados de maneira que a predição será realizada para três classes de valores de conta de energia. A Figura 3.6 mostra a tela resultante da aplicação do referido filtro da variável com o valor da conta em 3 classes.

A divisão de frequência mostra que a primeira classe, onde os indivíduos gastam até 100 reais mensalmente em energia, contém 63 indivíduos. A classe em que este valor varia de 100 a 200 reais contém 37 respostas e, ainda, a última classe contém 10 pessoas e corresponde aos indivíduos que pagam mais de 200 reais em energia mensalmente. Dessa maneira, é mostrado na Tabela 9 esta divisão, que será utilizada para a etapa da mineração de dados.

A Figura 3.7 e a Tabela 10 mostram a distribuição do valor da conta em 4 classes. Deve ser salientado que as frequências das classes agora são readequadas, sendo a primeira classe

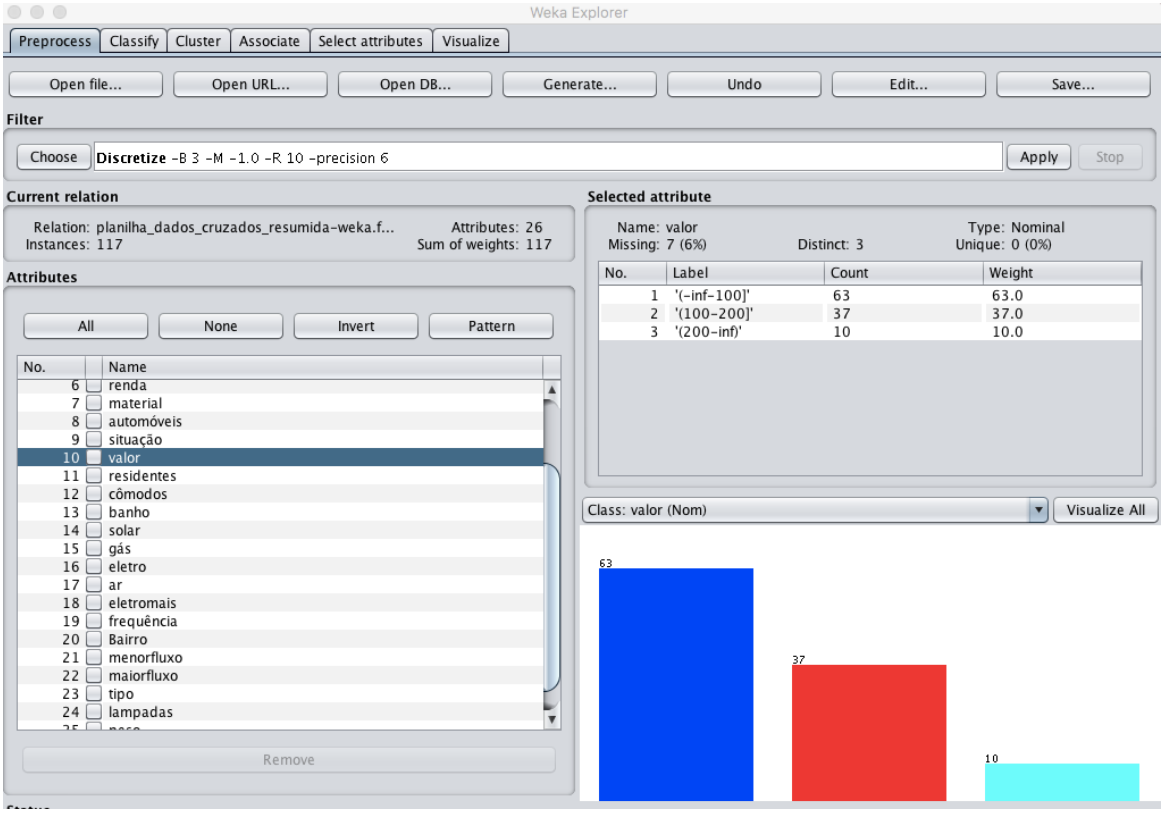

Figura 3.6: Divisão do valor da conta em 3 classes
Fonte: elaborada pela autora.

Tabela 9: Resumo dos valores da conta de energia das residências dos indivíduos representados em 3 classes

Classe	**Valores da energia elétrica, em reais**
1	Menos de 100
2	Entre 100 e 200
3	Mais de 200

Fonte: elaborada pela autora.

representada pelos indivíduos que pagam até 75 reais com 32 respostas, a segunda classe - que corresponde ao valor de 75 a 150 reais - com 55 pessoas, a terceira e classe com intervalo de 150 a 225 reais com 15 pessoas e, por fim, a quarta classe com valores a partir de 225 reais com 8 respondentes.

3.2 O Processo de Mineração de Dados

Para a construção de modelos de classificação deste trabalho foram realizados testes com 4 classificadores, englobando diferentes técnicas de classificação. A Tabela 11 mostra cada técnica de classificação e o respectivo algoritmo selecionado. Cabe salientar que os classificadores utilizaram o conjunto de dados descrito na seção anterior.

O método utilizado para validação dos modelos de classificação obtidos a partir dos dados em questão foi a validação cruzada, conhecida como *cross-validation* para N partições e mencionada

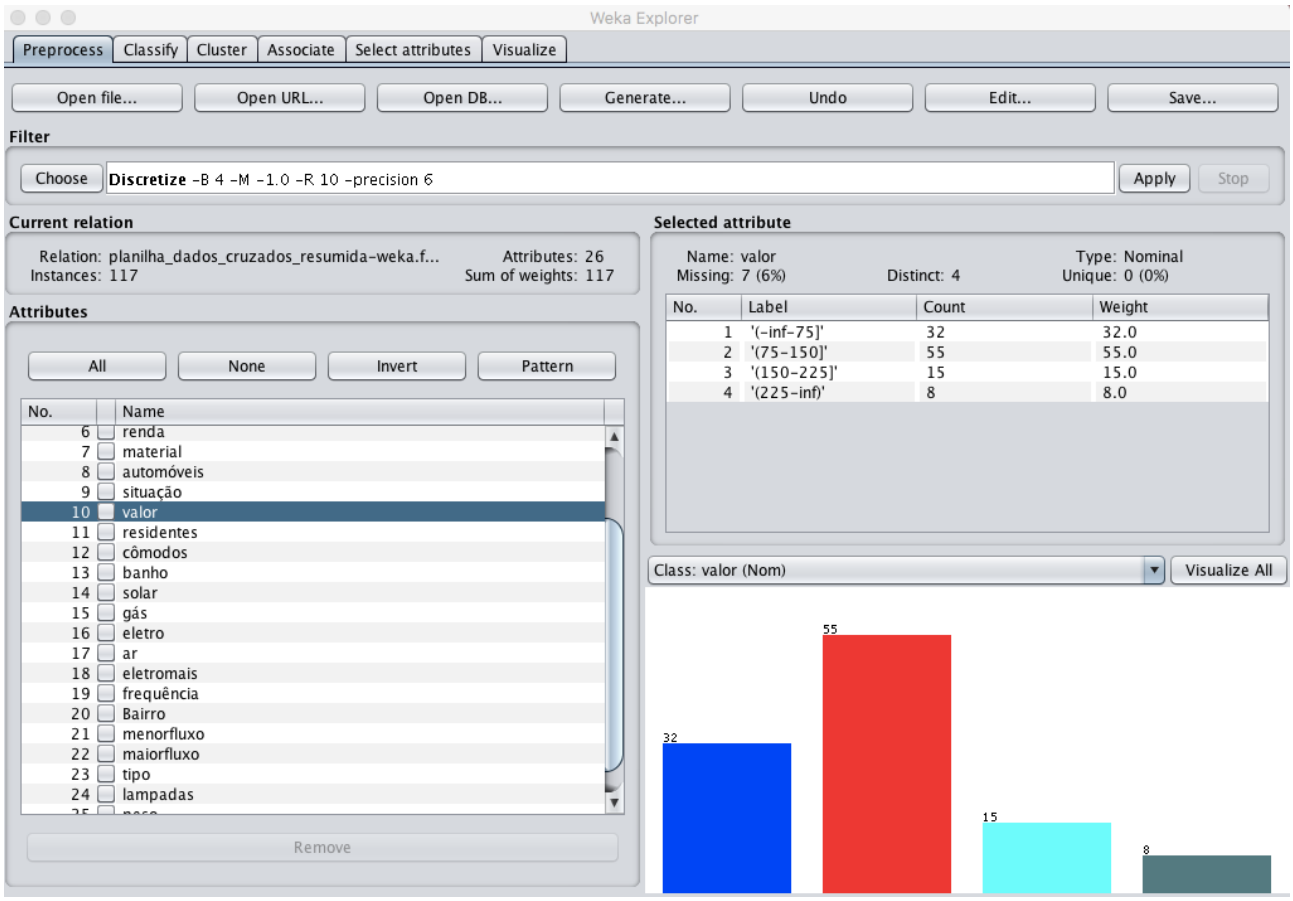

Figura 3.7: Divisão do valor da conta em 4 classes
Fonte: elaborada pela autora.

na Seção 2.4.6. Neste trabalho foram utilizadas dez partições.

Esta metodologia consiste em o algoritmo selecionar, durante cada execução, uma das partições para teste enquanto as outras nove serão utilizadas para treinamento do conjunto. Dessa maneira, este procedimento é repetido 10 vezes de modo que cada partição é utilizada exatamente uma vez, garantindo uma amostragem aleatória.

Foram realizados experimentos variando uma série de parâmetros, de modo a encontrar a melhor configuração dos algoritmos. A partir disso, foi realizada uma análise estatística destes modelos de classificação construídos para a realização de predição do valor da conta de energia elétrica paga na residência dos indivíduos entrevistados.

Esta análise será baseada em métricas de desempenho dos modelos de classificação estabelecidos para a avaliação de qualidade, como já explicado na Seção 2.5. Os valores das métricas são calculados a partir dos dados classificados e tabulados na matriz de confusão resultante.

A avaliação de desempenho de um classificador é baseada na contagem de instâncias testadas que foram classificadas corretamente e incorretamente e será realizada nos resultados.

Tabela 10: Resumo dos valores da conta de energia das residências dos indivíduos representados em 4 classes

Classe	**Valores da energia elétrica, em reais**
1	Menos de 75
2	Entre 75 e 150
3	Entre 150 e 225
4	Mais de 225

Fonte: elaborada pela autora.

Tabela 11: Técnicas de classificação e algoritmos classificadores correspondentes

Técnica	**Algoritmo**	**Implementação no *Weka***
Árvore de Decisão	C4.5	J48
Teorema de Bayes	Naïve Bayes	NaiveBayes
Vetor de Suporte	SMO	SMO
Vizinho mais próximo	K*	KStar

Fonte: elaborada pela autora.

4 RESULTADOS E DISCUSSÕES

Após as etapas de pré-processamento, deu-se início à mineração de dados. Assim, após a divisão em classes do atributo referente ao valor da conta de energia e todas as demais formatações, os algoritmos citados anteriormente foram testados no conjunto de dados em questão. A primeira classificação resultante das instâncias é mostrada na Tabela 12, realizada através da divisão em três classes de valores de energia, conforme a Tabela 9. Cabe salientar que, neste primeiro momento, foram consideradas todas as variáveis investigadas no questionário. Os algoritmos, por questões de valores faltantes, consideraram 110 instâncias das 117 para a classificação.

Tabela 12: Classificações das instâncias para três classes de valores de energia elétrica considerando todos as variáveis investigadas

Classificador	**Acurácia (%)**	**Erro (%)**
C4.5	70,90	29,10
K*	**74,36**	**25,64**
NaiveBayes	74,24	25,76
SMO	70,08	29,92

Fonte: elaborada pela autora.

Nota-se que o melhor classificador, ou seja, com o maior número de instâncias corretamente classificadas, foi o K*, com um percentual de 74,36% de acerto na classificação. Isso resulta em uma taxa de 25,64% de instâncias classificadas incorretamente. Nesta etapa, o classificador menos eficiente foi o *C4.5*, classificando de maneira acertiva 70,90% das instâncias.

O segundo teste a ser mostrado neste trabalho consiste na avaliação dos classificadores de acordo com a Tabela 10, onde os indivíduos são separados em quatro classes. A tabela de classificação, considerando a nova divisão de valores, é mostrada na Tabela 13.

Tabela 13: Classificações das instâncias para quatro classes de valores de energia elétrica considerando todos as variáveis investigadas

Classificador	**Acurácia (%)**	**Erro (%)**
C4.5	64,54	35,46
K*	**69,45**	**30,55**
NaiveBayes	67,27	32,73
SMO	64,54	35,46

Fonte: elaborada pela autora.

O classificador que melhor desempenhou o seu papel nesta etapa foi, novamente, o K*, classificando de maneira correta 69,45% das instâncias, onde agora os indivíduos estão classificados de acordo com quatro classes de renda, e não mais apenas três.

Como era esperado, já que o número de classes é superior ao teste anterior - ao aumentar o número de classes de renda, os métodos classificaram erroneamente mais instâncias.

Diante disso e com a pretensão de melhorar a taxa de acerto dos classificadores, mais testes foram realizados. Neste segundo momento, algumas variáveis do questionário começaram a ser testadas, para que fosse explícito que possuíssem relevância no resultado final.

Dessa maneira, para as mesmas três classes de valor da conta de energia mostradas na Tabela 9 que os indivíduos foram classificados e após os testes, foram excluídos os atributos: ano de ingresso na universidade, bairro de residência, curso de ensino superior e a situação da residência, que informa se a mesma é alugada, cedida ou própria. As classificações deste teste é mostrada na Tabela 14.

Tabela 14: Classificações das instâncias para três classes de valores de energia elétrica e sem os atributos mencionados

Classificador	**Acurácia (%)**	**Erro (%)**
C4.5	84,54	15,46
K*	**87,27**	**12,73**
NaiveBayes	84,54	15,46
SMO	83,63	6,37

Fonte: elaborada pela autora.

É visível que, após a exclusão das variáveis mencionadas, houve um aumento no percentual de instâncias classificadas de maneira correta em todos os classificadores testados, quando compararmos com os resultados da Tabela 12. Sendo assim, o melhor classificador continua sendo o K* que obteve aproximadamente 88% de acerto. O classificador *C4.5*, mesmo apresentando uma melhora, continua sendo o menos indicado para utilização nesse conjunto de dados.

Para uma divisão em quatro classes de valores de energia elétrica, sendo os valores os mesmos apresentados na Tabela 10 e excluindo-se os mesmos atributos mencionados anteriormente, os testes realizados são mostrados na Tabela 15.

Tabela 15: Classificações das instâncias para quatro classes de valores de energia elétrica e sem os atributos mencionados

Classificador	**Acurácia (%)**	**Erro (%)**
C4.5	70,90	29,10
K*	**79,89**	**20,11**
NaiveBayes	75,45	24,55
SMO	75,45	24,55

Fonte: elaborada pela autora.

Ao analisar a Tabela 15, percebe-se que, quando consideramos quatro classes de valores de conta de energia, a taxa de acerto de instâncias classificadas corretamente diminui, como o esperado, já que agora há mais classes novamente.

O melhor classificador foi, novamente, o K*, chegando a aproximadamente 80% de acerto a cerca da classe do valor de energia que uma residência pagará por mês. É importante mencionar que estes foram os testes mais satisfatórios ao longo das tentativas, muitos outros testes com os

classificadores foram executados, porém se fossem disponibilizados aqui, tornariam o trabalho extenso.

Após todos os testes, conclui-se que os dois melhores foram os testes representados pela Tabela 14 e 15. Esses testes apontaram, com o classificador K*, um percentual acertivo de aproximadamente 80% e 88%, quando as classes de valores de energia elétrica são divididas em três e quatro intervalos, respectivamente e com a utilização do método de validação cruzada. Escolheu-se, então, os dois últimos resultados para a demonstração das matrizes de confusão e das métricas de desempenho.

Diante disso, para as três classes de valores de conta de energia elétrica representado os resultados acima, tem-se a matriz confusão gerada pelo K* e representada na Tabela 16.

Tabela 16: Matriz de confusão 3 classes

a	**b**	**c**	←Classificado como
57	6	0	a= (-inf, 100]
6	**30**	1	b=(100, 200]
0	1	**9**	c=(200, +inf)

Fonte: elaborada pela autora.

Na diagonal principal, pode ser observado o número de instâncias classificadas corretamente, que corresponde à soma algébrica dos termos. Se quisermos, portanto, confirmar a porcentagem de acerto do algoritmo, dividimos a soma dos elementos da diagonal principal pelo número total das instâncias consideradas pelo algoritmo, multiplicando-se por 100. Este cálculo resultará na porcentagem de acerto de aproximadamente 88%.

Os algoritmos também fornecem as métricas de desempenho apresentadas neste trabalho na Seção 2.5, onde podem facilmente ser demonstrados pelas equações correspondentes. A Tabela 17 mostra as métricas para o K* quando os respondentes foram agrupados em três classes. As medidas *Precision* e *Recall* obtidas são iguais para cada classe porque a matriz é simétrica.

Tabela 17: Métricas de Desempenho 3 Classes

TP Rate	***FP Rate***	***Precision***	***Recall***	***F-measure***
0,904	0,127	0,904	0,904	0,904
0,810	0,095	0,810	0,810	0,810
0,900	0,010	0,900	0,900	0,900

Fonte: elaborada pela autora.

Na Tabela 18 é apresentada a matriz de confusão gerada quando o K* é testado com quatro classes. Pode-se, analogamente à matriz anterior, calcular a soma dos elementos da diagonal principal e dividir-se por 110 para chegar-se ao total de acerto do algoritmo. São mostradas as métricas de desempenho para este teste na Tabela 19. Note que o classificador K* foi capaz de classificar muito bem as classes relacionadas aos valores mais baixos de contas (a e b) considerando a medida F como métrica geral de qualidade. Mas as classes c e d foram mais

difíceis. Especialmente a classe d com valores superiores a 200 reais, onde apenas a metade das instâncias foram identificadas corretamente. Além disso, elas foram confundidas com todas as demais classes.

Considerando a avaliação das métricas de desempenho mencionadas, este classificador pode ser considerado bom para todas as três classes, pois além da Precisão ter um valor bom (90,40%, 81% e 90%), a *F-measure*, que representa o balanceamento entre a precisão e o *Recall*, foi quase igual a Acurácia nas três classes, não apenas na classe que possui mais elementos a serem classificados. Também observa-se que, na maioria dos casos, o classificador errou para a classificação da instância imediatamente ao lado da que seria a correta, o que é o mais desejável quando um erro ocorre.

Tabela 18: Matriz de confusao 4 classes

a	**b**	**c**	**d**	←Classificado como
27	5	0	0	a= (-inf, 75]
6	**44**	5	0	b=(75, 150]
0	2	**12**	1	c= (150,225)
1	1	2	**4**	d=(225, +inf)

Fonte: elaborada pela autora.

Tabela 19: Métricas de desempenho 4 classes

TP Rate	***FP Rate***	***Precision***	***Recall***	***F-measure***
0,843	0,089	0,794	0,843	0,817
0,800	0,145	0,846	0,800	0,822
0,800	0,073	0,631	0,800	0,705
0,500	0,009	0,800	0,500	0,666

Fonte: elaborada pela autora.

Ainda, pode ser analisado nos dois testes escolhidos o valor do Coeficiente *Kappa*, fornecido pelo registro do *Weka*. Para o teste com três e quatro classes aqui demonstrado, os valores de *Kappa*, de acordo com a Tabela 7 proposta por Landis e Koch (1977), são considerados substanciais ou bons. A Tabela 20 mostra esses valores.

Tabela 20: Coeficiente *Kappa* (K) para os testes do classificador K* e sua interpretação

Classes	**Acurácia (%)**	***K***	**Interpretação**
3	87,27	0,6822	Substancial ou bom
4	79,08	0,6775	Substancial ou bom

Fonte: elaborada pela autora.

5 CONCLUSÕES

Este trabalho partiu da aplicação de um questionário em estudantes da Universidade Federal do Rio Grande e da Universidade Federal de Pelotas a fim de investigar os hábitos de consumo e a posse de eletrodomésticos ou equipamentos nas residências desses estudantes. Estes dados foram utilizados para a confecção de uma base de dados inicial a ser utilizada posteriormente.

Após a validação de constructo realizada por profissionais da área da engenharia elétrica, constatou-se que todas as variáveis envolvidas poderiam apontar uma relação com a variável a ser investigada: o valor da conta de energia elétrica no domicílio dos respondentes.

O objetivo principal deste trabalho foi, partindo das informações coletadas no questionário, montar uma base de dados para predizer com a maior certeza possível o valor da conta de energia elétrica no domicílio de estudantes com as mesmas características dos entrevistados através de técnicas de mineração de dados, identificando, assim, perfis de consumidores.

Foi realizada uma pesquisa sobre técnicas de mineração de dados existentes e optou-se pela utilização de métodos de classificação na plataforma livre *Weka*.

A predição do valor da conta de energia baseou-se em análise matemática, onde o caminho encontrado durante a pesquisa e posteriormente aplicado foram as técnicas de mineração de dados. Este primeiro passo auxilia os especialistas na tomada de decisão em relação ao consumo de energia elétrica e deve ser aprimorado para uma maior exatidão.

O trabalho contou com um exaustivo processo de pré-processamento, a fim de deixar os dados mais preparados possíveis para a aplicação da mineração de dados. Esta etapa foi constituída de um preparo manual em planilha de dados e um preparo no próprio *Weka*.

O melhor classificador para o banco de dados deste trabalho foi o K*, apresentando um percentual de instâncias classificadas corretamente de aproximadamente 88% ao dividir-se o valor da conta de energia em três intervalos, sendo eles: menos de 100 reais, de 100 a 200 reais ou mais de 200 reais. Dessa maneira, 88% das instâncias foram classificadas corretamente, acertando a classe a qual pertence o elemento em questão.

O mesmo classificador apresentou um percentual de 80% de acerto ao se deparar com quatro classes de valores de energia elétrica, agora sendo mais específico e acertando 80% dos casos se a instância de teste pertence ao grupo das pessoas que pagam menos de 75 reais, de 57 a 150 reais, de 150 a 225 reais ou mais de 225 reais ao final do mês em energia elétrica.

6 TRABALHOS FUTUROS

A análise do valor da conta de energia elétrica a partir de predições realizadas pelos classificadores pode ser considerada o primeiro passo na melhoria no sistema elétrico de potência, através do planejamento de geração e distribuição de energia.

Após estes testes, espera-se que, como trabalho futuro, possam ser investigadas outras variáveis que possuam uma relação com o valor do gasto energético mensal a fim de elevar a taxa de instâncias corretamente classificadas, assim como um estudo que abranja uma maior população e número de amostras. Poderão ser testados também outros classificadores ou técnicas de mineração de dados.

Espera-se que o conteúdo deste trabalho seja utilizado como marco inicial em prestadoras de serviços de energia, a fim de prever, de acordo com a população, a quantidade de energia gerada e distribuída, procurando um bom balanço energético e evitando problemas na rede nacional.

Serão melhor investigados também modelos de classificadores caixa-branca, que permitem um melhor entendimento de como a classificação é realizada, para tentar descobrir os atributos mais decisivos para o processo.

Ainda, espera-se que este trabalho sirva de incentivo para pesquisas mais abrangentes a cerca de um assunto tão fundamental.

7 REFERÊNCIAS

ANEEL. **Resolução Normativa Nº. 414** . 2010.

ARANGO, H. **Bioestatística Teórica e Computacional**. Rio de Janeiro: Editora Guanabara Koogan, 2008.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o brasil. **Brazilian Journal of Computers in Education,** v. 19, p. 03, 2011.

BAKER, R.; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. **Journal of Educational Data Mining**, v.1, p.3-16, 2009.

BRASIL, Legislação. **Lei nº 9.427, de 26 de dezembro de 1996**. Institui a Agência Nacional de Energia Elétrica-ANEEL, disciplina o regime das concessões de serviços públicos de energia elétrica.

CALLEGARI-JACQUES, S. **Bioestatística: princípios e aplicações**. Porto Alegre: Artmed Editora S.A., 2003.

CARVALHO, L. A. V. **Data Mining: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.** Rio de Janeiro: Editora Ciência Moderna, 2005.

CASTRO, D. Procedimentos de data mining na definição de valores para as análises de multicritérios como apoio à tomada de decisões e análise espaciais urbanas. **XXIV Congresso Brasileiro de Cartografia**. Aracaju, 2010.

COVER, T.; HART, P. Nearest Neighbour Pattern Classification. **IEEE Transactions on Information Theory**, v.13, p. 21-27, 1967.

EPE, 2015. **Balanço energético nacional 2015: Ano base 2015**. Rio de Janeiro. 2015.

EPE, 2016. **Balanço energético nacional 2016: Ano base 2015**. Rio de Janeiro. 2016.

FACELI, K.; LORENA, A.; GAMA, J.; CARVALHO, A. **Inteligência Artificial: Uma abordagem de aprendizado de máquina.**. Rio de Janeiro: Editora LTC, 2011.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From Data Mining to KnowledgeDiscovery in Databases. **Revista American Association for Artificial Intelligence,** v.13, p. 37-54, 1996.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: um guia prático.** Rio de Janeiro: Editora Elsevier, 2005.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. The WEKA Data Mining Software: An Update. **SIGKDD Explorations**, v.11, p.10-17, Issue 1, 2009.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. Burlington: Editora Morgan Kaufmann Publishers, 2006.

HANSEN, A.M.C. **Padrões de consumo de energia elétrica em diferentes tipologias de edificações residenciais.** 2000. Dissertação (Mestrado em Engenharia Civil) - Escola de Engenharia, Universidade Federal do Rio Grande do Sul, 2000.

HOLSHEIMER, M., SIEBES, A. *Data mining: The search for knowledge in databases.* Artigo técnico CS-R9406, Amsterdam, 1991.

KIRKBY, R.. WEKA Explorer User Guide for Version 3- 4-3. University of Waikato, 2004. Disponível em: `http://weka.sourceforge.net/manuals/ExplorerGuide.pdf`. Acesso em: 20 de junho de 2018.

LANDIS, R.; KOCH, G. The measurement of observer agreement for categorical data. Biometrics. **Journal of International Biometric Society**, v.33, p.159-174, 1977.

PELUCCI, R. S.; PAULA, R. R.; SILVA, W. B.; LADEIRA, A. P. Utilização de técnicas de Aprendizado de Máquina no reconhecimento de entidades nomeadas no Português. **Revista E-xacta**, v.4, p. 73-81, 2011.

PROCEL. **Plano de Aplicação de Recursos**. 2016. Disponível em: `https://goo.gl/VFvMrW`. Acesso em: 07 setembro 2017.

RUTTER, M.; SERTÓRIO, A. A. **Pesquisa de Mercado**. 2. ed. São Paulo: Editora Ática S. A., 1994.

SANTOS, B. F. C. **Caracterização de Diagramas de Consumo Doméstico de Eletricidade na Perspetiva de Comercialização**. 2016. Dissertação (Mestrado Integrado em Engenharia Eletrotécnica e de Computadores) - Universidade do Porto, Porto, 2016.

STORY, M., CONGALTON, R. Accuracy Assessment: A User's Perspective. **Journal Photogrammetric Engineering and Remote Sensing**, v.52, p. 397-399, 1986.

TAN, P.; STEINBACH, M.; KUMAR, V. **Introdução do Data Mining**. Rio de Janeiro: Editora Ciência Moderna, 2009.

UNIVERSITY OF WAIKATO. **Weka 3 – Machine Learning Software in Java**. Waikato, 2010. Disponível em: `http://www.cs.waikato.ac.nz/ml/weka`. Acesso em: 20 junho 2018.

WITTEN, Ian; FRANK, Eibe, HALL, Mark. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. São Paulo: Editora MK, 2005.

8 ANEXOS

8.1 Anexo A - Questionário

Pesquisa sobre hábitos de consumo e posse de equipamentos

Este instrumento tem o intuito de realizar uma pesquisa de mestrado em modelagem computacional para identificar diferentes perfis de consumo de energia elétrica entre os estudantes. Não precisa se identificar e leva apenas 5 minutos.

1. Curso: *Escreva aqui a sua resposta*

2. Ano de ingresso na universidade: *Escreva aqui a sua resposta*

3. Escolaridade da mãe:

○ Nunca estudou

◉ Ensino fundamental

○ Ensino médio ou técnico

○ Ensino superior

○ Pós-graduação

4. Escolaridade do pai:

○ Nunca estudou

◉ Ensino fundamental

○ Ensino médio ou técnico

○ Ensino superior

○ Pós-graduação

5. Qual a renda de todos os moradores da residência somada?

○ De 1 a 2 salários mínimos

◉ De 3 a 5 salários mínimos

○ De 6 a 8 salários mínimos

○ De 9 a 10 salários mínimos

○ Mais de 10 salários mínimos

6. A sua residência é feita, predominantemente:

○ De alvenaria

◉ De madeira

○ Mista

7. Possui pelo menos um automóvel entre os moradores?

◉ Sim

○ Não

8. A sua casa de moradia é:

◉ Própria

○ Alugada

○ Cedida

9. Qual o número de pessoas que reside em sua casa, contando com você?

◉ Só eu

○ 2

○ 3

○ 4

○ 5

○ 6

○ Mais de 6

10. Quanto pagou neste mês de conta de luz, em reais? *Escreva aqui a sua resposta*

11. Quantos cômodos possui a sua residência? *Escreva aqui a sua resposta*

12. Qual o tempo médio de banho da maioria dos residentes?

○ De 5 a 10 minutos

◉ De 10 a 15 minutos

○ De 15 a 20 minutos

○ Mais de 20 minutos

13. Em sua residência há aquecimento solar?

◉ Sim

○ Não

14. Em sua residência há aquecimento à gás?

○ Sim

◉ Não

15. Marque os eletrodomésticos que possui e utiliza, pelo menos, de 15 em 15 dias:

- Clique nas caixas e vá selecionando múltiplos equipamentos.

  ☑ Geladeira

  ☐ Microondas

  ☐ Forno elétrico

  ☑ Máquina de lavar roupas

  ☐ Máquina de Secar/Secadora de Roupas Portátil

  ☑ Máquina de Pão

  ☐ Lavadora de louças

  ☑ Fogão/Cooktop

  ☑ Notebook/Computador desktop

  ☐ Fritadeira elétrica

16. Quantos aparelhos de ar condicionado a residência possui?

○ 0

◉ 1

○ 2

○ 3

○ 4 ou mais

17. Em sua opinião, qual o eletrodoméstico mais utilizado na sua residência, além da geladeira? *Escreva aqui a sua resposta*

18. Com qual frequência? *Escreva aqui a sua resposta*

19. Em qual bairro e cidade sua residência está localizada? *Escreva aqui a sua resposta*

20. Em qual período sua casa fica com o mínimo de pessoas?

◉ Manhã

○ Tarde

○ Noite

21. Em qual período sua casa fica com o máximo de pessoas?

○ Manhã

○ Tarde

◉ Noite

22. Em qual tipo de domicílio você reside?

○ Apartamento

◉ Casa

○ Outro

23. Ainda são utilizadas lâmpadas incandescentes em sua residência?

◉ Sim

○ Não

24. Em sua opinião, qual o peso da conta de energia elétrica no orçamento familiar?

○ Leve

○ Mediano

◉ Pesado

25. Como identifica o consumo de energia dos equipamentos na maioria das vezes?

○ Etiqueta no equipamento

◉ Selo Procel

○ Outro

○ Não costumo identificar o consumo de energia dos equipamentos