UNIVERSIDADE FEDERAL DO PARÁ
CAMPUS UNIVERSITÁRIO DE TUCURUÍ
FACULDADE DE ENGENHARIA ELÉTRICA

KACIA KARINA ROSA DE SOUSA

METODOLOGIA DE APOIO À APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA DETECÇÃO DE PERDAS COMERCIAIS DE ENERGIA ELÉTRICA COM SISTEMA EMBARCADO

TUCURUÍ
2019

KACIA KARINA ROSA DE SOUSA

METODOLOGIA DE APOIO À APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS NA DETECÇÃO DE PERDAS COMERCIAIS DE ENERGIA ELÉTRICA COM SISTEMA EMBARCADO

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia Elétrica, do Campus Universitário de Tucuruí, da Universidade Federal do Pará, como requisito parcial para obtenção do título de Bacharel em Engenharia Elétrica.

Orientador (a): Me. Bernard Carvalho Bernardes.

TUCURUÍ

2019

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S725m Sousa, Kacia Karina Rosa de
Metodologia de apoio à aplicação de técnicas de mineração de dados na detecção de perdas comerciais de energia elétrica com sistema embarcado / Kacia Karina Rosa de Sousa. — 2019.
63 f. : il. color.

Orientador(a): Prof. Me. Bernard Carvalho Bernardes
Trabalho de Conclusão de Curso (Graduação) - Faculdade de Engenharia Elétrica, Campus Universitário de Tucuruí, Universidade Federal do Pará, Tucuruí, 2019.

1. Mineração de Dados. 2. Perdas Comerciais. 3. KDD. 4. Árvore de Decisão. 5. Sistema Embarcado. I. Título.

CDD 620.00285

“Se tu olhares, durante muito tempo, para um abismo, o abismo também olha para dentro de ti.”

Friedrich Nietzsche.

AGRADECIMENTOS

Agradeço a Deus, por ter colocado em meu caminho meus três grandes amigos: Caio Luz, Luan Ephima e Lucas Silva, que permaneceram ao meu lado durante os momentos difíceis, e que possuem minha eterna admiração e carinho.

Agradeço a minhas irmãs Kelly Sousa e Talita Santos pelo apoio, carinho e cuidado comigo. E a minha mãe Dona Célia por ser meu exemplo de força e coragem.

Agradeço ao meu mestre Bernard Carvalho Bernardes, pela paciência, apoio e orientação na construção deste trabalho.

RESUMO

As perdas não técnicas na distribuição de energia elétrica causam grandes prejuízos financeiros, tanto para as concessionárias de energia quanto para seus clientes, além de afetar diretamente na qualidade da energia que chega aos consumidores regulares. Por causa da grandeza do mercado de energia elétrica há uma maior complexidade de se combater as perdas, tornando a resolução dessa problemática de total importância. Isto posto, este trabalho apresenta o desenvolvimento de um sistema embarcado com o microprocessador ESP32 capaz de enviar a cada um minuto para um banco de dados informações de potência ativa, reativa e fator de potência dos consumidores de energia elétrica. Através do padrão de consumo levantado dos consumidores, um banco de dados de curvas de carga residenciais foi formado, construído exatamente para a aplicação da descoberta de conhecimento em bancos de dados (KDD), como forma de aperfeiçoar o processo de identificação de fraudes e furtos na rede de distribuição de energia elétrica. Para isto, foi realizada a coleta de dez curvas de carga diárias residenciais, divididas em três conjuntos de dados de acordo com seu padrão social, e utilizado o módulo scikit-learn da linguagem de programação Python para a mineração dos dados. Para classificar o consumidor em regular ou irregular foi aplicado uma tarefa de classificação supervisionada com algoritmos de árvore de decisão, para cada conjunto de dados. Três modelos de classificadores foram gerados os quais obtiveram taxas de acuracidade de 86, 90 e 86%, respectivamente.

Palavras-chave: Curva de Carga. Fraude. Mineração de Dados. Árvore de Decisão. Classificadores.

ABSTRACT

Non-technical losses in the distribution of electricity cause major financial losses, both to the utilities and their customers, as well as directly affecting the quality of energy reaching regular consumers. Because of the size of the electricity market there is a greater complexity of combating losses, making the resolution of this issue of utmost importance. That said, this work presents the development of an embedded system with the ESP32 microprocessor capable of sending every minute to a database active, reactive power and power factor information of electric consumers. Through the raised consumption pattern of consumers, a database of residential load curves was formed, built exactly for the application of database knowledge discovery (KDD), as a way to improve the fraud and theft identification process. in the electricity distribution network. For this, ten residential daily load curves were collected, divided into three data sets according to their social standard, and the scikit-learn module of Python programming language was used for data mining. To classify the consumer as regular or irregular, a supervised classification task with decision tree algorithms was applied to each data set. Three classifier models were generated which obtained accuracy rates of 86, 90 and 86%, respectively.

Keywords: Load Curve. Fraud. Data Mining. Decision Tree. Classifiers.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
ADC	Conversor Analógico Digital
DAC	Conversor Digital Analógico
FP	Positivo Falso (do inglês, False Positive)
FN	Negativo Falso (do inglês, False Negative)
IDE	Ambiente de Desenvolvimento Integrado
KDD	Descoberta de Conhecimento em Banco de Dados (do inglês, Knowledge Discovery in Databases)
MP	Medidor de Potência
TP	Positivo Verdadeiro (do inglês, True Positive)
TN	Negativo Verdadeiro (do inglês, True Negative)
TC	Transformador de Corrente

SUMÁRIO

1 INTRODUÇÃO

As concessionárias de distribuição de energia recebem a energia do sistema de transmissão para ser repassada aos seus clientes finais. Onde nesse processo parte da energia é perdida. Essas perdas são classificadas como técnicas e não técnicas. As perdas técnicas são próprias do funcionamento da distribuição de energia elétrica, onde uma parcela da energia se perde devido ao processo de transporte, transformação de tensão e medição, já as perdas não técnicas, também conhecidas como perdas comerciais, estão relacionadas às irregularidades no consumo de energia, tais como fraudes no medidor e conexões clandestinas por partes dos usuários (ANEEL, 2019).

1.1 Problemática

De acordo com a ANEEL (2019), o Brasil perdeu no ano de 2018 cerca de 33,3 TWh de energia com práticas ilegais, gerando um custo de aproximadamente R$ 6,6 bilhões, que representa uma volumosa perda na receita das concessionárias, isso proporciona aumento na tarifa de energia e perda da sua qualidade quando chega aos consumidores regulares tornando a resolução dessa problemática de total interesse por parte das próprias concessionárias e para seus clientes. A proposta para a análise e identificação de furtos e fraudes de energia de maior destaque é a utilização de tarefas de Mineração de Dados (do inglês, *Data Mining*) em grandes conjuntos de dados, esta que possui diversos tipos de técnicas como: regressão logística, processamento de linguagem natural, análise discriminante, redes neurais artificiais, árvore de decisão, k-vizinhos mais próximos entre outras, com o objetivo de detectar qualquer anomalia e padrões no sistema de distribuição possuindo a capacidade de prever tendências futuras do comportamento dos consumidores.

Interação com o mundo externo, preparação, transformação, modelagem, processamento e apresentação são etapas básicas para a análise de dados. A etapa de interação com o mundo externo refere-se à leitura e armazenamento usando uma variedade de formatos de arquivos e repositórios de dados e a etapa de preparação trata-se de limpar, manipular, combinar, normalizar, reformatar, tratar e transformar os dados para a análise (MCKINNEY, 2018). A preparação da base de dados é essencial para se obter resultados satisfatórios na Mineração de Dados, este procedimento é indispensável para a detecção de problemas nos

dados e qualquer falha durante o pré-processamento pode acarretar em um processo ineficaz (CASTRO; FERRARI, 2016).

1.2 Justificativa

A maioria das aplicações de tarefas de Mineração de Dados para a detecção de furtos e fraudes no sistema de distribuição de energia é feita utilizando os bancos de dados das próprias concessionárias em busca principalmente por três informações: dados de clientes, dados de inspeções e histórico de consumo.

Castanheira (2008) afirma que a maioria dos bancos de dados hoje são particulares, e que não são direcionados a dar suporte as pessoas responsáveis por extrair o conhecimento da base de dados, isso é o que ocorre com as concessionária de distribuição de energia, seus bancos de dados na maioria das vezes são antigos tendo um grande volume de dados não uniformes e ausentes, incompletos ou omissos, demandando muito tempo dos profissionais na etapa de preparação dos dados. Por conseguinte, a motivação e necessidade de um sistema voltado diretamente para o levantamento do padrão de consumo de energia elétrica, dada a relevância do pré-processamento de dados, com o armazenamento de dados orientados para a utilização de Mineração de Dados visando à elaboração de um banco de dados espera-se assim aumentar a rapidez durante a preparação dos dados e a identificação dos clientes irregulares de energia elétrica com mais eficácia, sendo este modo capaz de lidar com o grande volume de dados que é gerado na rede de distribuição de energia elétrica.

1.3 Objetivos

1.3.1 Objetivo Geral

Desenvolver uma metodologia de apoio à aplicação de técnicas de mineração de dados, de modo que aperfeiçoe o processo de identificação de perdas comerciais de energia elétrica com sistema embarcado. Utilizando um sistema de amostragem de dados, de baixo custo, como ferramenta para o levantamento da curva de carga do consumidor, construir um banco de dados, para a aplicação de KDD.

1.3.2 Objetivos Específicos

- Desenvolvimento do medidor de potência monofásico, utilizando o microcontrolador ESP32;
- Coleta, armazenamento e tratamento das curvas de carga;
- Análise dos dados coletados de cada residência;
- Aplicação da tarefa de Mineração de Dados de classificação Árvore de Decisão e a avaliação do modelo gerado.

2 REVISÃO BIBLIOGRÁFICA

2.1 Perdas de Energia Elétrica na Distribuição

O sistema elétrico de potência é constituído por geração, transmissão e distribuição. A energia elétrica que passa pelo sistema de transmissão e de distribuição, mas que por algum motivo, técnico ou comercial, não chega a ser comercializada é intitulada como perdas de energia. Essas perdas são classificadas como: perdas técnicas e não técnicas.

As perdas técnicas são perdas intrínsecas ao sistema, causadas principalmente pelo efeito joule e perdas nos núcleos dos transformadores. O sistema de distribuição é dividido de acordo com suas características como: segmentos de rede, transformadores, ramais de ligação e medidores. A partir dessas informações é estimado um percentual de perdas técnicas referentes à energia injetada na rede.

As perdas não técnicas, ou comerciais, ocorrem basicamente em função de furto (ligação clandestina, desvio direto da rede) ou fraude de energia (adulterações no medidor), popularmente conhecidos como “gatos” (ANEEL, 2016). Através dessas ações ilegais o consumidor interfere na rede, fazendo com que a energia consumida não seja faturada pela concessionária de energia. As perdas não técnicas reais são definidas pela diferença entre as perdas totais e as perdas técnicas regulatórias, estimadas pela ANEEL.

Os valores regulatórios são aqueles presentes na tarifa de energia, ao passo que os valores reais são os que de fato ocorrem. A diferença entre esses valores é de inteira responsabilidade da concessionária de energia. Em montantes de energia, as perdas técnicas no ano de 2018 corresponderam a cerca de 38,3TWh e as perdas não técnicas 33,3 TWh (ANEEL, 2019).

As perdas não técnicas reais equivaleram em 2018 a um custo de aproximadamente R$ 6,6 bilhões de reais. Porém, as perdas não técnicas regulatórias, custaram cerca de R$ 5 bilhões ao ano, correspondendo a 3% do valor da tarifa de energia elétrica, variando por distribuidora (ANEEL, 2019).

A Figura 1 exibe a porcentagem representada pelas perdas sobre a energia injetada na rede no ano de 2018, em relação aos consumidores do sistema de energia.

Figura 1 - Perdas sobre a Energia Injetada (2018).

Fonte: Relatório de Perdas de Energia Elétrica na Distribuição, ANEEL, 2019, p.2.

O consumidor regular arca parcialmente pelas fraudes e furtos. Entretanto quando a concessionária recupera o consumo irregular, os montantes faturados são incorporados no mercado, sendo dividido de forma proporcional entre todos, diminuindo assim a tarifa de energia. Com a redução das perdas comerciais os consumidores usufruem de redução dos valores regulatórios, redução da tarifa, diminuição do desperdício e aumento na qualidade do fornecimento (ANEEL, 2019).

2.1.1 Padrão de Consumo de Energia

Os valores de consumo de energia elétrica consumidos mensalmente pelos usuários são armazenados pelas concessionárias após a leitura nos medidores. Através da análise dos valores de energia consumida é possível classificar os consumidores por padrões de consumo. As cargas dos consumidores possuem características comuns, como: localização geográfica, tensão de fornecimento, tarifação, dependência da energia elétrica, entre outras. A partir dessas características típicas das cargas o consumidor é classificado.

Seja qual for o conjunto de características que o consumidor possui, o consumo de energia elétrica possui um comportamento sazonal e cíclico, que pode ser verificado através de tipologias de curvas de cargas dos usuários. Esta curva mantém um comportamento regular, denominado padrão ou perfil de consumo.

Através da análise de curvas de consumo de energia elétrica e levando em consideração o comportamento de tais curvas, é possível detectar padrões de consumo que podem ser utilizados para classificar consumidores e detectar anomalias nas redes de distribuição de energia elétrica (QUEIROZ, 2016), isto é, situações em que o consumidor demonstra uma alteração em sua tipologia de carga, o que o diferencia de outros consumidores, indicando uma variação do seu padrão de consumo. Pode indicar um possível consumidor irregular devido à detecção de anomalia no seu padrão de consumo.

2.2 Sistemas Embarcados

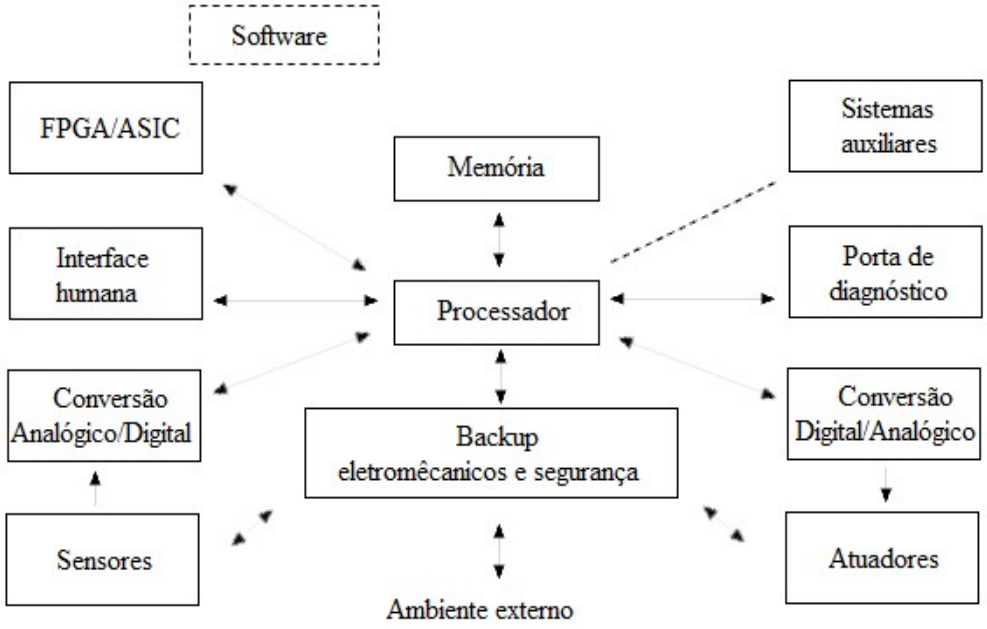
É evidente a evolução eletrônica dos computadores e equipamentos eletrônicos nas últimas décadas. Foi nesse cenário que os sistemas embarcados tomaram destaque e solidificaram sua importância; sendo aplicado em diferentes segmentos como: eletrodomésticos, brinquedos, periféricos de computadores, aparelhos de comunicação, equipamentos médicos, automotivo e industrial.

O termo sistema embarcado remete ao uso de eletrônica e software dentro de um produto, ao contrário de um computador de uso geral, como sistema de laptop ou desktop. É um microprocessador possuindo apenas um chip. O sistema embarcado possui uma combinação de hardware e software de computador, e podendo possuir partes adicionais mecânicas e outras, onde normalmente seu software possui uma função fixa e destinada a uma função específica (STALLINGS, 2010).

Os sistemas embarcados são implementados para um único propósito específico e estes possuem algumas características: é um sistema construído para realizar uma tarefa, completamente ou parcialmente independente da intervenção humana, desenvolvido para realizar ações de forma mais eficiente, tamanho pequeno e pouco peso.

A Figura 2 é uma representação de uma possível organização de um sistema embarcado.

Figura 2 - Sistema Embarcado.



Fonte: William Stallings, 2010.

Basicamente um sistema embarcado possui o processador para executar as instruções do código definidas no software, e memória para o armazenamento do próprio software e dos dados gerados durante o funcionamento do sistema. Para que haja comunicação entre o ambiente externo com o processador é necessário muitas vezes a presença de dispositivos como: sensores, atuadores, conversores analógicos/digitais e digitais/analógicos.

2.2.1 ESP32

O ESP32 é uma série de microcontroladores de baixo custo, é também um sistema em um único chip com microcontrolador integrado. Possui um microprocessador Tensilica Xtensa LX6, foi criado e desenvolvido pela empresa Chinesa Espressif Sistemas. De acordo com sua fabricante o ESP32 é capaz de funcionar de forma confiável com temperaturas que varia de -40°C a + 125°C, possui um consumo de energia ultrabaixo, tem alto nível de integração, possui chip Hybrid, Wi-Fi e Bluetooth. Alguns módulos que integram a série ESP32 são o ESP32-WROOM-32, ESP32-WROOM-32D, ESP32-PICO-D4 e o ESP32-WROVER-IB.

Algumas das principais especificações do modulo ESP32 WROOW-32, Figura 3, de acordo com seu datasheet são mostradas a seguir.

Figura 3 - Placa de desenvolvimento ESP32-WROOM-32.

Fonte: Espressif Sistemas, 2019.

- Wi-Fi e Bluetooh embutidos;
- Cristal oscilador de 40MHz;
- Tensão de operação de 2,7V a 3,6V;
- Corrente de funcionamento de 40mA a 80mA;
- Sensor capacitivo e de temperatura embutidos;
- Memória Flash de 4MB;
- Interface para SD card, UART, SPI, SDIO, I2C, IS2, IR;
- 38 GPIO, sendo 16 canais ADC de 12bits e 2 canais DAC de 8bits.

Diferente de alguns fabricantes de microcontroladores, a Espressif não disponibiliza nenhum IDE junto com a distribuição do conjunto de programas. Podendo assim, ser programada em diferentes IDEs e em diferentes linguagens, dentre elas a IDE Arduino com o Núcleo ESP32 para Arduino, Lua RTOS para ESP32 e MicroPython.

3 MINERAÇÃO DE DADOS

A necessidade de analisar os dados surgiu devido à grande quantidade de dados que são gerados e coletados diariamente. Estamos realmente vivendo na era dos dados. Terabytes ou petabytes de dados são despejados em nossas redes de computadores, o World Wide Web (WWW), dispositivos de armazenamento de dados, negócios, sociedade, ciência e engenharia, medicina e quase todos os outros aspectos da vida diária (HAN; KAMBER; PEI, 2012). Este crescimento explosivo do volume de dados disponível é resultado da informatização da nossa sociedade e o rápido desenvolvimento de poderosas ferramentas de coleta e armazenamento de dados. Ferramentas poderosas e versáteis são extremamente necessárias para descobrir informações valiosas da enorme quantidade de dados e transformar tais dados em conhecimento organizado, esta necessidade levou ao nascimento da Mineração de Dados.

Para Carvalho (2002), *Data Mining* é uma técnica que se aplica a uma grande quantidade de dados e informações que muitas vezes estão escondidas nos bancos de dados das corporações. Essa técnica pode ser aplicada em qualquer segmento (medicina, vendas, marketing, entre outras) que trabalhe com um grande volume de dados armazenados, porém para que essas informações possam ser analisadas de forma coerente, um especialista no assunto não pode ser dispensado.

FAYYAD (1996) ressalta que a Mineração de Dados é um passo do KDD, processo que consiste na aplicação de dados, análise e algoritmos de descoberta que produz uma enumeração particular de padrões (ou modelos) sobre os dados.

Devido à grande disseminação nos últimos anos, do conceito Mineração de Dados causada pela era da informação. Han, Kamber e Pei (2012) apontam que muitas vezes há uma compreensão errônea sobre Mineração de Dados, muitas vezes sendo tratado como algo equivalente a KDD. Porém Mineração de Dados é uma das etapas de KDD. Sendo o que define os sistemas de Mineração de Dados como sistemas que podem automaticamente minerar todos os conceitos valiosos que estão escondidos em um grande banco de dados sem interação ou direcionamento humano. O processo de KDD é uma sequência interativa das seguintes etapas:

1. Limpeza de dados - para remover ruído e dados inconsistentes;

2. Integração de dados - onde várias fontes de dados podem ser combinadas;

3. Seleção de dados - onde os dados relevantes para a tarefa de análise são recuperados da base de dados;

4. Transformação de dados - onde os dados são transformados e consolidados em formulários apropriados para a mineração executando operações de resumo ou de agregação;

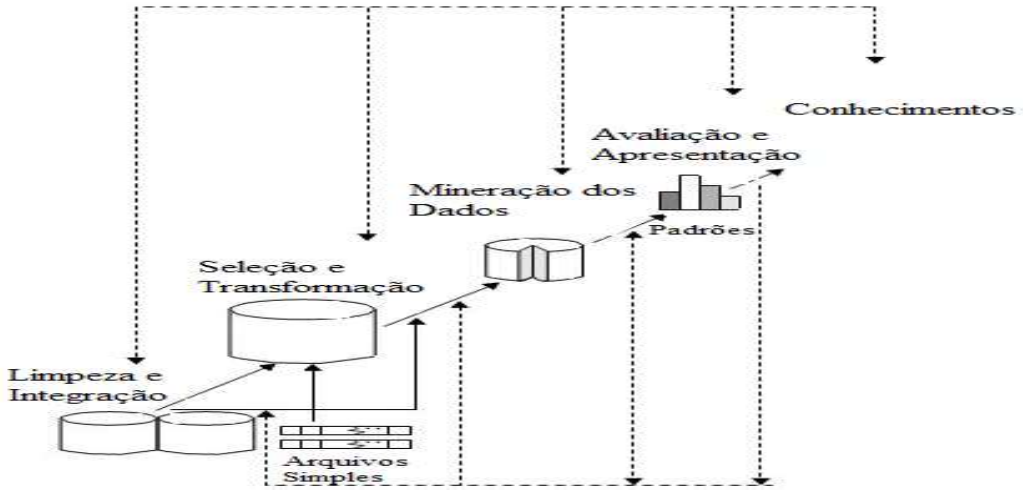
5. *Data Mining* - um processo essencial em que métodos inteligentes são aplicados para extrair padrões de dados;

6. Avaliação de padrões - para identificar os padrões verdadeiramente interessantes que representam o conhecimento baseado em medidas de interesse;

7. Apresentação do conhecimento - onde técnicas de visualização e representação de conhecimento são usadas para apresentar conhecimento minado aos usuários.

A Figura 4 a seguir, uma adaptação, é uma representação das etapas para a KDD.

Figura 4 - Mineração de dados como uma etapa no processo de KKD.



Fonte: Han, Kamber e Pei, 2012.

McKinney (2018), afirma que a fase de pré-processamento é a mais complexa do processo de KDD, chegando a ocupar cerca de 80% de todo o tempo utilizado no processo. Correia (2017) enfatiza a importância da seleção e pré-processamento devido serem a primeira etapa no processo de descobrimento de informação por possuir grande influência na qualidade do resultado final, pois é nessa etapa que é definido os conjuntos de dados contendo todas as possíveis variáveis que farão parte da análise.

O processo de seleção pode ser bastante complexo em razão das várias fontes diferentes como: data warehouses, planilhas e sistemas legados, podendo dispor de diferentes formatos.

3.1 Tarefas de Mineração de Dados

Como mencionado anteriormente a etapa de mineração de dados é uma das etapas essenciais no processo de KDD, pois é nela que ocorre a conversão de um grande volume de dados em padrões. A Mineração de Dados é frequentemente classificada de acordo com os objetivos de uma tarefa, conforme o tipo de informação que se deseja obter através dos dados. Para Larose (2005) as principais tarefas de mineração de dados são:

- Descrição (*Description*): descrevem padrões e tendências contidas nos dados, a tarefa de descrição deve conter padrões claros e aptos a interpretações e explicações intuitivas;

- Classificação (*Classification*): possui uma variável categórica alvo, que é a variável que se deseja prevê, essa tarefa examina um grande conjunto de registros, cada um dos registros contendo informações sobre a variável de destino, bem como um conjunto de entradas ou preditores. O algoritmo primeiro examina o conjunto de dados que contém as variáveis de entrada e a variável alvo, assim o algoritmo é treinado para classificar um novo registro, baseados nas classificações do conjunto de treinamento, o algoritmo atribuiria classificações aos novos registros de dados;

- Estimação (*Estimation*) ou Regressão (*Regression)*: é similar a classificação, porém a variável de destino é numérica e não categórica, os modelos são construídos usando

registros "completos" que fornecem o valor da variável alvo, bem como dos preditores. Então, para novas observações, estimativas do valor da variável de destino são feitos, com base nos valores dos preditores;

- Predição (*Prediction*): é semelhante às tarefas de classificação e estimação, exceto que ela visa descobrir o valor futuro de um determinado atributo;

- Agrupamento (*Clustering*): agrupa os dados que possuem objetos semelhantes, um *cluster* um conjunto de registros semelhantes um ao outro e diferentes dos registros em outros *clusters*, difere da tarefa de classificação, pois não possui variável categórica;

- Associação (*Association)*: descobre quais atributos “combinam”, possui a forma “Se antecedente, depois consequente”.

De acordo com a tarefa de Mineração de Dados escolhida é definido a técnica que será implementada através de algoritmos para atingir os objetivos da tarefa. Os métodos (ou técnicas) de Mineração de Dados normalmente são classificados em supervisionados (preditivos) e não supervisionados (descritivos).

Os métodos de Mineração de Dados são em sua maioria supervisionados, que consistem na existência de uma variável categórica, portanto é fornecido ao algoritmo exemplos de dados em que o valor da variável de destino é especificado, além das variáveis de entrada que caracterizam o objeto, para que o algoritmo possa aprender quais valores das variáveis de entrada estão associados aos valores da variável categórica alvo (variáveis preditoras); tendo como métodos supervisionados mais empregados: classificação e regressão. Contudo, os métodos não supervisionados não precisam de uma pré-categorização, isto é, nenhuma variável categórica alvo é especificada como tal, o próprio algoritmo procura padrões e disposições entre todas as variáveis de entrada dos objetos. Os métodos não supervisionados mais empregados são: associação (que também pode ser supervisionado) e agrupamento (LAROSE, 2005).

3.1.1 Árvore de Decisão

Para a fase de Mineração de Dados desse trabalho foi escolhido uma tarefa de classificação supervisionada denominada Árvore de Decisão (do inglês, *Decision Tree*). Nas tarefas de classificação um modelo (ou classificador) é gerado para prever classes (discretas ou não discretas) dos dados. Elas frequentemente são utilizadas para a identificação de fraudes, markenting, fabricação e diagnóstico médico.

Basicamente o método de classificação é um processo que possui duas etapas. A primeira sendo a etapa de aprendizado, onde o modelo de classificação é gerado a partir de vários atributos considerados, para descrever um conjunto pré-determinado de classes ou conceitos de dados, e a segunda etapa é usada para prever os rótulos ou categorias das classes para os novos dados, com base no modelo de classificação gerado na primeira etapa (HAN; KAMBER; PEI, 2012).

As principais técnicas de classificação atualmente usadas são: Árvore de Decisão, Redes Neurais Artificiais, Classificadores baseados em regras, Bayesianos, Máquina de Vetor de Suporte (SVM) e Sistemas Fuzzy. Ficando a cunho de o profissional qualificado decidir qual técnica aplicar de acordo com o tipo de base de dados disponível.

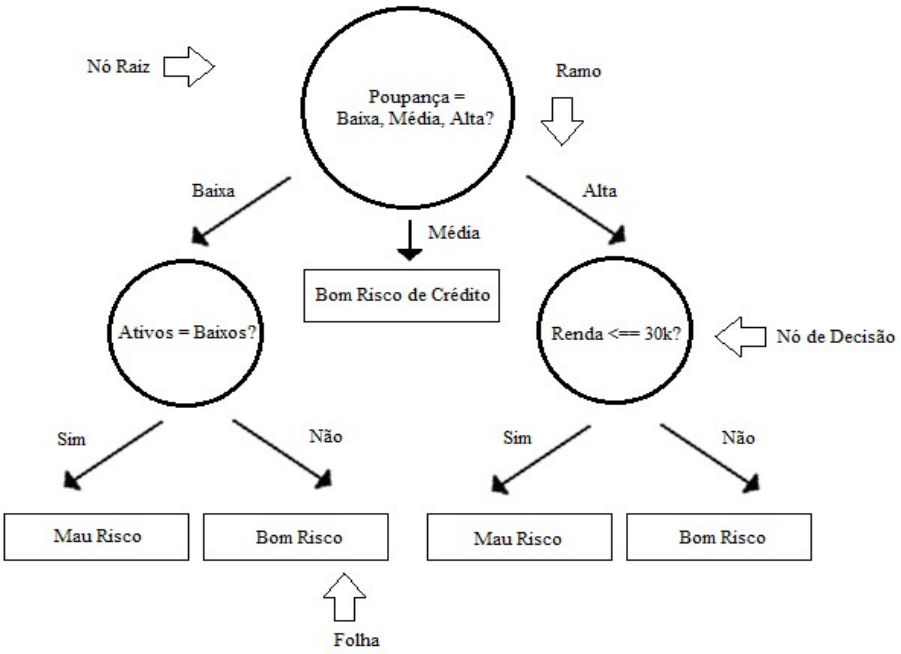
Árvores de Decisão é uma espécie de fluxograma, sendo que os nós internos são os testes feitos sobre cada atributo, as arestas são à saída destes testes e os nós folhas são o rótulo da classe. Quando uma classe desconhecida precisa ser classificada, cada objeto do conjunto de dados é testado na árvore. Então, um caminho é percorrido na estrutura, desde a raiz até o nó folha, que possui o rótulo da classe para o objeto (HAN; KAMBER; PEI, 2012). Deste modo, Árvores de Decisão são estruturas simples, que representam possíveis caminhos de decisão e um resultado para cada trajetória, que não dependem de parâmetros adicionais. São utilizadas para uma análise exploratória dos dados, sendo que é um método bastante popular, devido a sua agilidade e facilidade de interpretação dos resultados, além de lidarem facilmente com dados numéricos e categóricos, no entanto, o seu êxito depende muito dos dados utilizados.

Larose (2005) define o método de classificação por Árvore de Decisão, como uma construção de uma árvore que possui uma coleção de nós de decisão, conectados por ramos estendendo-se para baixo a partir do nó raiz até terminar nos nós folha, conforme ilustrado na Figura 5. Uma adaptação, onde a variável categórica alvo é o risco de créditos, sendo o potencial do cliente classificados como bons ou maus riscos de crédito, as variáveis

preditoras são a poupança (baixa, média, alta), ativos (baixos ou não baixos) e renda (≤ R$30,000 ou > R$30,000).

Os dados com poupança baixa são enviados pelo ramo à esquerda (poupança = baixa) para o outro nó de decisão. Os dados com poupança alta são enviados pelo lado direito, ramificando para um nó de decisão diferente. Os dados com poupança média são enviados a ramificação do meio diretamente para o nó da folha indicando o termina desse ramo. Para os clientes com poupança baixa o próximo nó de decisão verifica se o cliente possui ativos baixos, então aqueles com ativos baixo são classificados em bom ou mau risco de crédito. Para os clientes com poupança alta o próximo nó de decisão verifica se o cliente possui uma renda ≤ R$30,000 ou > R$30,000 e assim são classificados em bom ou mau risco de crédito. Quando nenhuma divisão adicional pode ser feita, o algoritmo para de crescer novos nós.

Figura 5 - Árvore de Decisão simples.



Fonte: Larose, 2005.

- Nó Raiz é o nó no topo da árvore;

- Nós(s) de decisão são definidos como elementos que estão conectados por ramos;
- Ramos é a ligação entres os nós;
- Folha(s) indica uma classe.

Os critérios de seleção para os melhores atributos preditivos são baseados em diferentes medidas, como: impureza, distância e dependência. A maior parte dos algoritmos de indução busca dividir os dados de um nó-pai de forma a minimizar o grau de impureza dos nós-filhos. Existem diversos tipos de critérios de seleção de atributo, sendo este um dos motivos que diferencia os algoritmos de indução de Árvore de Decisão.

O critério usado para a realização dos atributos será o que possui maior utilidade para a classificação, então é determinado o ganho de informação a cada atributo através do critério escolhido. O atributo escolhido para ser o atributo teste pertencente ao nó será o que possuir o maior ganho de informação. A partir desta aplicação inicia-se um novo processo de partição para os próximos nós. Os principais critérios para a seleção dos atributos preditivos utilizados em indução de Árvore de Decisão são fundamentados nos conceitos de entropia e índice de Gini.

Entropia

A Entropia é o cálculo do ganho de informação, utilizada para representar a incerteza ou (im)pureza relacionada aos dados. A entropia é calculada pela equação (1):

$$Entropia\ (\text{nó}) = -\sum_{i=1}^{c} p\left(\frac{i}{\text{nó}}\right) \log_2 \left[p\left(\frac{i}{\text{nó}}\right)\right] \quad (1)$$

Onde:

c é o número de classes;

$p\left(\frac{i}{\text{nó}}\right)$ é a fração dos registros pertencentes à classe i no nó.

Para verificar o quão bom é um atributo de teste realizado é preciso realizar a comparação do grau de entropia do nó-pai, antes da divisão, com o grau de entropia dos nós-

filhos, após a divisão. O atributo que fornece uma maior diferença é escolhido como condição de teste. O ganho é dado pela equação (2), na forma (BASGALUPP, 2010):

$$Ganho = entropia(pai) - \sum_{j=1}^{n}\left[\frac{N(v_j)}{N} entropia(v_j)\right] \quad (2)$$

Onde:

n é o número de valores do atributo, isto é, o número de nós-filhos;

N é o número total de objetos do nó-pai;

$N(v_j)$ é o número de exemplos associados ao nó-filho v_j;

Gini

O Gini mede o grau de heterogeneidade dos dados, sendo utilizado para medir a impureza de um nó. O nó é puro quando o índice for zero e quanto mais próximo do valor um o nó se torna mais impuro, quando há o aumento do número de classes igualmente distribuídas neste nó. O índice Gini foi desenvolvido pelo matemático italiano Corrado Gini em 1912, sendo este índice de um nó determinado pela equação (3):

$$gini_{index}(\text{nó}) = 1 - \sum_{i=1}^{c} p\left(\frac{i}{\text{nó}}\right) \quad (3)$$

Onde:

c é o número de classes.

Para obter o índice Gini é preciso calcular a diferença entre o $gini_{index}$ antes e após a divisão. Essa diferença, Gini, está representada pela equação 04 (BASGALUPP, 2010):

$$Gini = gini_{index}(pai) - \sum_{j=1}^{n}\left[\frac{N(v_j)}{N} gini_{index}(v_j)\right] \quad (4)$$

Onde:

n é o número de valores do atributo, ou seja, o número de nós-filhos;

N é o número total de objetos do nó-pai;

$N(v_j)$ é o número de exemplos associados ao nó filho v_j.

Dessa forma é escolhido o atributo que gerar um maior valor para $Gini$.

Quando em Árvores de Decisão de classificação com divisões binárias, o critério de índice Gini é escolhido tendendo a isolar num ramo os registros que representam a classe mais frequente. Na ocasião em que se utiliza a entropia equilibra-se o número de registros em cada ramo. Durante o processo de construção de uma Árvore de Decisão sempre é importante buscar diminuir a entropia, assim como ser consistente com os dados e ter o menor número possível de nós.

Overfitting

Durante o processo de construção de uma Árvore de Decisão seus ramos podem estar constituídos de anomalias devido a presença de ruídos e outliers (valores discrepantes) dos dados de treinamento, originando o problema de *Overfitting* ou sobreajuste. Isto é, um aprendizado muito específico do conjunto de treinamento, não permitindo ao modelo generalizar.

Para detectar e excluir o sobreajuste são utilizados métodos de poda da árvore para melhorar a taxa de acerto do modelo para novos dados que não foram utilizados no conjunto de dados do treinamento (HAN; KAMBER; PEI, 2012).

Poda

Poda (do inglês, *pruning*) reduz o tamanho da Árvore de Decisão podendo ser dividida de duas maneiras de acordo com a situação. Quando a árvore está em crescimento e deseja-se interromper transformando o nó de decisão em uma folha da árvore representando a classe mais frequente no ramo, é chamada de pré-podagem (ou poda descendente), ou em situações em que a árvore já está completa, onde é removido ramos completos em que tudo que está abaixo de um nó interno é excluído e transformado em folha, representando a classe mais frequente no ramo, chamada pós-podagem (ou poda ascendente).

Apesar de a poda ser um método fundamental, que melhora a taxa de acerto tornando-a mais simples e de fácil interpretação, deve-se ter prudência em sua utilização, para se evitar

um problema denominado *underfitting*, ou sub-ajuste. Ocorre quando a árvore é podada demais fazendo com que os modelos de classificação não aprendam o suficiente sobre os dados de treinamento (BASGALUPP, 2010).

3.1.2 Avaliação dos Modelos

Agora que o modelo foi construindo é necessário estimar a precisão que o classificador possui para prever a variável categórica alvo em que o mesmo não foi treinado. Para isso existem algumas técnicas comuns para avaliar a precisão, ou o quão bom é o seu classificador desenvolvido, com base nas partições amostradas aleatoriamente dos dados de entrada fornecidos, sendo uma delas denominada Matriz de Confusão que mostra o número de classificações corretas em relação às classificações preditivas para cada classe, com base em um conjunto de exemplos de dados.

Dado um conjunto de dados classificados em um modelo preditivo, cada ponto de dados se enquadra em uma das quatro categorias abaixo:

- Positivo verdadeiro (*True Positive* - TP): quando o modelo prevê um caso positivo corretamente;
- Positivo falso (*False Positive* - FP): quando o modelo prevê um caso negativo como positivo;
- Negativo falso (*False Negative* - FN): quando o modelo prevê um caso positivo como negativo;
- Negativo verdadeiro (*True Negative*- TN): quando o modelo prevê um caso negativo corretamente.

Um exemplo de uma matriz de Confusão está presente na Tabela 1, de um classificador de e-mails em spam ou não spam, com partição binária. Onde todos os dados presentes no conjunto são avaliados da seguinte maneira:

- TP: este e-mail é spam e foi previsto corretamente;
- FP: este e-mail não é spam, mas foi previsto que era;

- FN: este e-mail é spam, mas foi previsto que não era;
- TN: este e-mail não é e foi previsto que não era.

Tabela 1 - Matriz de Confusão de um classificador de e-mails.

	Spam	**Não é Spam**
Premissa "Spam"	TP	FP
Premissa "Não é Spam"	FN	TN

Fonte: Joel Grus, 2016, p.145.

É esperado de uma Matriz de Confusão que as taxas de sucesso para Verdadeiro Positivo e Verdadeiro Negativo sejam altas e as outras, baixas. Fazendo uma relação entre VP, FP, VN e FN definem-se uma métrica de desempenho para as taxas de acerto e erro, a acurácia (CORREIA, 2017).

A acurácia avalia a chance do classificador em acertar suas previsões. A acurácia pode ser definida como a fração de casos que foram corretamente classificados, sendo positivos verdadeiros ou verdadeiros negativos. Isso é,

$$Acurácia = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

Outras Métricas também relacionadas à Matriz de Confusão, amplamente usadas na classificação, para a compreensão do modelo de algoritmo gerado em relação à classe preditiva, Verdadeiros Positivos são: precisão, recall e f-score.

Precisão pode ser considerada uma medida de acerto, isto é, quais dados rotulados como positivos são realmente positivos. Ao passo que, o recall é uma medida de totalidade, qual taxa de positivos são rotuladas como tal. A precisão e recall são definidos pela equação (6) e (7), respectivamente.

$$Precisão = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

O f-score faz uma combinação entre a precisão e o recall, e é calculado pela equação (8).

$$F - score = 2\ x\ \frac{precisão\ x\ recall}{precisão + recall} \tag{8}$$

4 MATERIAIS E MÉTODOS

A metodologia utilizada para a realização deste trabalho apresenta caráter explicativo, relacionando teoria e prática, com abordagem majoritariamente quantitativa. A finalidade é a elaboração de um dispositivo de amostragem de dados de potência do consumidor de baixo custo para a classificação do cliente da concessionária de acordo com seu padrão de consumo, como consumidor regular ou irregular, para a formação de um banco de dados especificamente feito para a estimativa de curvas de carga e à aplicação do processo de KDD, na detecção de fraudes na rede de distribuição de energia elétrica.

4.1 Medidor de Potência

4.1.1 Hardware

Para a coleta de dados de potência dos usuários de energia elétrica, em residências monofásicas, foi realizada uma pesquisa de materiais disponibilizados no mercado que atendessem as especificações do projeto possuindo um baixo custo.

Atualmente existem diversos tipos de microcontroladores disponíveis no mercado; o ESP32 foi escolhido por possuir características e funcionalidades necessárias para a composição do medidor de potência como: pequeno porte, baixo custo, wi-fi 802,11 b/g/n com frequência de 2,4GHz, ADCs com resoluções de 12bits, baixo consume de energia e um cristal oscilador de 40MHz.

Os sensores utilizados para o recebimento de dados de tensão e corrente instantânea, em conjunto com o ESP32, foi o sensor de tensão GBK P8 e o sensor de corrente SCT013-000 não invasivo 100A.

O sensor de tensão GBK P8, Figura 6, possui alta precisão e com capacidade de medição de tensões de até 311Vac. Suas principais características estão presentes na Tabela 2.

O sensor de corrente SCT013-000, Figura 7, é um transformador de corrente podendo ser preso a fase ou neutro de uma residência, sem a necessidade de fazer nenhum trabalho elétrico de alta tensão e possui isolamento galvânico. O enrolamento do primário é a fase ou neutro da residência, o secundário fica no enrolamento presente no sensor e necessita de um

resistor de carga para o fechamento do circuito no secundário para fornecer uma tensão proporcional à corrente secundária, suas principais características estão presentes na Tabela 3.

Figura 6 - Sensor de tensão GBK P8.

Fonte: Autocore Robótica, 2019.

Tabela 2 – Sensor de Tensão GBK P8.

SENSOR DE TENSÃO GBK P8	
Tensão de alimentação	5Vdc
Tensão de entrada	127/220Vac
Isolamento de rede	Optoacoplador
Tensão maxima suportável	311Vac
Tensão de saída	0 a 5Vdc ajustável

Fonte: elaborado pela autora.

Figura 7: Sensor de corrente SCT013-000 não invasivo 100A.

Fonte: GearBest BR, 2019.

Tabela 3 – Sensor de corrente SCT013-000 não invasivo 100A.

SENSOR DE CORRENTE SCT013-000 NÃO INVASIVO 100A	
Corrente de entrada	0 a 100A
Corrente de saída	0 a 50mA
Material do núcleo	Ferrite
Dielétrico	6000V AC/1min
Taxa anti-chama	UL94-V0
Temperatura de trabalho	-25° a +70°C
Proteção	Diodo Zener
Não linearidade	±3%

Fonte: elaborado pela autora.

A corrente alternada presente no primário do TC fornece um campo magnético variável no núcleo. A corrente alternada induzida no secundário é proporcional a corrente alternada do primário, devido a relação de espiras, onde o número de espira do primário está para o número de espira do secundário, como mostra a equação (9). Sendo a corrente do secundário calculada pela equação (10).

$$a = \frac{N_1}{N_2} \tag{9}$$

$$I_S = I_P \, x \, a \tag{10}$$

Onde:

I_P – corrente alternada no primário

I_S – corrente alternada no secundário

a – relação de espiras

N_1 - número de espira do primário

N_2 - número de espira do secundário

O resistor de carga (R_c) é necessário para conectar o TC ao microcontrolador ESP32, o sinal de saída do SCT013-000 precisa atender e estar de acordo com o intervalo da tensão positiva exigida pelas entradas analógica do ESP32, o qual possui como tensão de referência ADC de 3.3V. A relação de espira do TC é 2000, ou seja, a corrente no secundário é um fragmento de 2000 da corrente do primário, isto é dado por 100:0,05. Portanto, o valor de resistência de carga deve ser baixo para prevenir a saturação do núcleo SCT013-000.

Todos os componentes eletrônicos, utilizados para a montagem do MP, foram comprados através dos sites Autocore Robótica e GearBest BR no mês de março do ano de 2019, seus respectivos preços estão presentes na Tabela 4.

Tabela 4 - Componentes Eletrônicos Utilizados.

Item	**Quant.**	**Preço**
Microcontrolador ESP32-WROOW-32	1	R$ 37,76
Capacitor Multicamadas de Cerâmica 110nF	3	R$ 01,80
Resistor 22Ω	1	R$ 00,25
Resistor 10kΩ	2	R$ 00,50
Sensor de Corrente SCT013-000 não invasivo 100A	1	R$ 34,89
Sensor de Tensão GBK P8	1	R$ 15,29
Display oLED 0.96”	1	R$ 24,39
Fonte DC 5V – 500mA	1	R$ 05,68
Total		R$ 120,56

Fonte: elaborado pela autora.

Foi utilizado um capacitor do tipo cerâmica de 110nF, por indicação da fabricante Espressif, para minimizar o ruído na porta de leitura do ADC do ESP32, e outros três para os pinos de tensão, para a regulação e estabilidade da tensão de referência. Um display oLEd para a verificação da conexão do ESP32 a um ponto de acesso à internet e a comunicação com a planilha eletrônica no Google *Sheets*.

4.1.2 Calibração Teórica dos Sensores

Sensor de Corrente

O ADC do ESP32 pode converter a tensão de entrada até $2^{12} = 4096$ níveis lógicos, onde 0 se refere a 0V e 4095 a tensão de referência, 3,3V. Determinando aproximadamente uma tensão medida, definida pela equação (11):

$$Val = ADC\ x\ \frac{Vref}{4095} \tag{11}$$

Onde:

Val - valor real medido

ADC - nível lógico de 0 a 4095

$Vref$ - tensão de referência

Para realizar a calibração do sensor de corrente é necessário inserir no código desenvolvido uma constante de calibração. Tendo como resistor de carga de 22Ω, a corrente no primário no TC é convertida em uma tensão proporcional através do resistor de carga, que será medida pela entrada analógica do ESP32.

A tensão na entrada do microcontrolador tende a adicionar uma constante DC às medições, um offset, devido ao circuito com resistores divisores de tensão de polarização. Porém está constante é removida através do filtro passa alta do código fonte do monitor de energia de código aberto - *openenergymonitor.* Deste modo, para identificar a constante de calibração do sensor de corrente é necessário a relação da resolução de amostragem, a corrente máxima no primário, a relação de espira e o valor do resistor de carga, como mostra a equação (12). Para este projeto o valor encontrado da constante foi de 90,9.

$$I_{const} = \frac{\frac{I_P}{a}}{Rc} \tag{12}$$

Sensor de Tensão

O sensor de tensão da GBK modulo P8 possui um opto acoplador que isola a tensão da rede de 127/220V do circuito DC. Ele possui um trimpot de $10k\Omega$, que varia o intervalo DC de saída para a entrada do microcontrolador. Para a conversão de nivel lógico para valores reais de tensão também é utilizada a equação (11). Na definição da constante de calibração é

necessario a utilização da proporção dos valores de tensão de entrada VAC com os valores DC da saida, como definido na equação (13).

$$V_{const} = \frac{V_P}{V_{DC}} \tag{13}$$

Onde:

V_P – tensão AC de pico de entrada

V_{DC} - tensão de saída DC

Medindo a tensão de saída, em uma determinada posição do trimpot foi calculado uma constante de tensão de 245.

4.1.3 Software

Para a programação do microcontrolador ESP32 foi utilizado uma implementação enxuta e eficiente da linguagem de programação Python, que permite a transferência do código do computador para o microcontrolador, denominada MicroPython. Python é uma linguagem de alta produtividade e legibilidade. É uma linguagem limpa, fácil de traduzir o raciocínio em um algoritmo (TELLES, 2008). A uPyCraf foi a IDE utilizada, para a utilização do MicroPython no ESP32.

Para os cálculos dos valores de tensão eficaz, corrente eficaz e potência ativa, foram utilizadas as equações (14), (15) e (16), respectivamente.

$$V_{rms} = \sqrt{\frac{\sum_{n=1}^{N° \, amostras} v^2(n)}{N° \, amostras}} \tag{14}$$

$$I_{rms} = \sqrt{\frac{\sum_{n=1}^{N° \, amostras} i^2(n)}{N° \, amostras}} \tag{15}$$

$$P = \frac{\sum_{n=1}^{N° \, amostras} v(n) \; x \; i(n)}{N° \, amostras} \tag{16}$$

Onde:

$v(n)$ – tensão instantânea

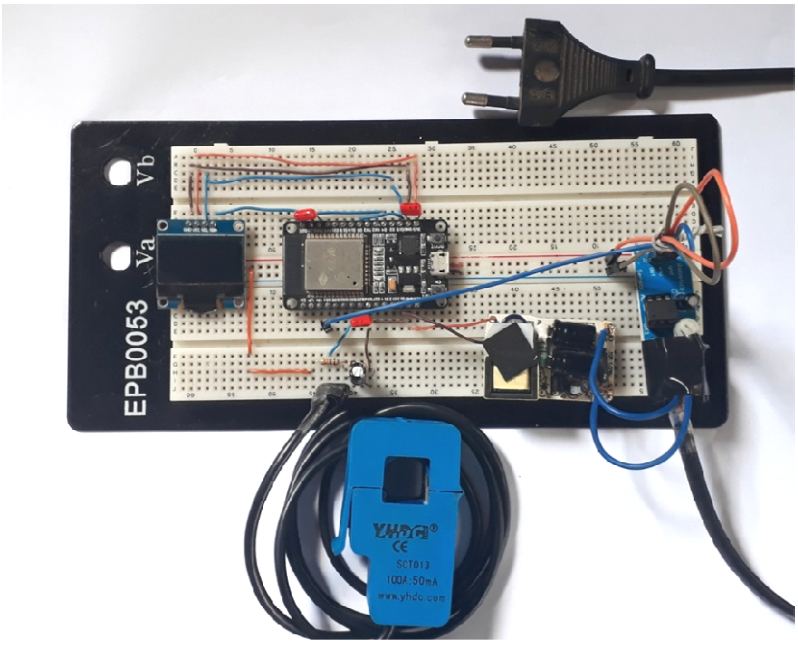
$i(n)$ – corrente instantânea

O cálculo da potência reativa foi feito utilizando a equação (17), através do triângulo das potências, que relaciona as potências aparente e ativa.

$$Q = \sqrt{S^2 - P^2} \tag{17}$$

O MP montado na protoboard pode se visto na Figura 8 a seguir.

Figura 8 – Medidor de Potência.



Fonte: elaborado pela autora.

4.2 Mineração de Dados com Python

À aplicação do processo de KDD para a transformação de dados brutos em informação possui etapas de pré-processamento até o pós-processamento dos resultados da Mineração de Dados. A escolha para a utilização das técnicas deve se levar em consideração quais os objetivos específicos que possui, dependendo do tipo de informação que se deseja extrair dos dados. Por isso, optou-se por aplicar aos dados coletados uma técnica de Mineração de Dados

com tarefa de classificação supervisionada Árvore de Decisão (*Decision Tree*), onde foi adotado como variável categórica alvo a classe 0 para consumidores regulares, e a classe 1 para consumidores irregulares.

Correia Matos (2017) cita que Árvore de Decisão é uma técnica bastante difundida entre os especialistas, pela sua eficácia em trabalhar com um grande volume de dados, ela permite construir classificadores com bons desempenhos em detectar padrões. Possuindo fácil interpretação do resultado.

Python é uma linguagem de programação interpretada, gratuita, orientada a objetos, de script e funcional. Foi criada em 1991, pelo matemático e programador Guido van Rossum no Instituto de Pesquisa Nacional para Matemática e Ciência da Computação (CWI), feita com base na linguagem ABC. Python ficou popularmente conhecida por volta de 2005 para a construção de sites, através de seus vários frameworks web como Django. Porém, nos últimos dez anos, Python além de ser uma linguagem de computação cientifica ela avançou para uma das linguagens mais importantes em ciência de dados e aprendizado de máquinas.

Todo o código de programação desenvolvido para a aplicação de KDD foi produzido no ambiente Jupyter. O tratamento e modelagem dos dados foram realizados principalmente utilizando as bibliotecas fornecidas pela linguagem de programação Python: Matplot, Numpy, Pandas e Scikit-Learn.

4.2.1 Jupyter

Fernando Pérez, em 2014, anunciou o projeto Jupyter. Um ambiente interativo para desenvolver softwares de código aberto, baseado na Web para notebooks, com base para mais de 40 linguagens de programação, entre elas a linguagem Python. Jupyter fornece ao usuário um ambiente que possibilita uma ampla variedade de fluxo de trabalho em ciência de dados, computação científica e aprendizado de máquina. Sendo escolhido como ambiente de desenvolvimento, porque fornece um interpretador de alta qualidade em processamento e análise de dados tendo como principais características um fluxo de trabalho que possibilita a execução, exploração e visualização dos dados.

4.2.2 Matplot

Matplot é uma biblioteca Python bastante popular para plotagem e visualização de dados 2D. Foi criada por John D. Hunter em 2002, principalmente para gerar plotagens adequadas para publicação. Matplot é uma biblioteca bastante semelhante ao Matlab, é capaz de exportar visualização em vários formatos entre eles: PDF, SVG, JPG, BMP etc. É uma biblioteca simples e interativa, tornando muito mais fácil a visualização de informação, etapa primordial para o processo exploratório dos dados (MCKINNEY, 2018).

4.2.3 Pandas e Numpy

McKinney (2018) menciona que nos últimos anos, o suporte melhorado de Python para bibliotecas, como o Pandas e o scikit-learn, o transformou em uma opção para a engenharia de software de propósito geral, é uma excelente opção como uma linguagem principal para a construção de aplicação de dados. A etapa de limpeza, integração, seleção e transformação dos dados foram realizadas através das bibliotecas Numpy e Pandas.

Numpy – *Numerical Python* (Python Numérico) é um pacote de processamento numérico orientados a arrays (matrizes). Possui como principais recursos (MCKINNEY, 2018):

- Possibilidade para trabalhar com narrays que é um vetor multidimensional;
- Várias funções matemáticas para trabalhar com arrays de dados de forma rápida;
- Realiza operações de álgebra linear, transformada de Fourier e geração de números aleatórios
- Relaciona-se com o Pandas no processamento dos dados.

O Pandas disponibiliza ferramentas para trabalhar com dados estruturados ou tabulados. Oferecendo um ambiente eficaz para o pré-processamento dos dados. Os principais objetos do Pandas são (MCKINNEY, 2018):

- Series: é um objeto do tipo array unidimensional, que possui uma sequência de valores, e uma array relacionada aos rótulos (*labels*), denominado de índice, ou seja, possui um índice para cada linha da array;

- DataFrame: é uma estrutura de dados tubular, como uma tabela de dados possuindo linhas e colunas,muito similar a planilhas eletrônicas, com rótulos tanto para linhas como para as colunas. Cada coluna de um DataFrame pode possuir valores numéricos, booleanos, string entre outros, que são armazenados como arrays bidimensionais.

4.2.4 Scikit-Learn

A modelagem dos dados tratados foi feita através da biblioteca Scikit-Learn. O Scikit-Learn é um dos kits de ferramentas Python para aprendizado de máquina mais amplamente utilizado, com suporte simples e eficiente para mineração e análise de dados. Foi construída a base de Numpy, Scipy e Matplot, que são bibliotecas Python.

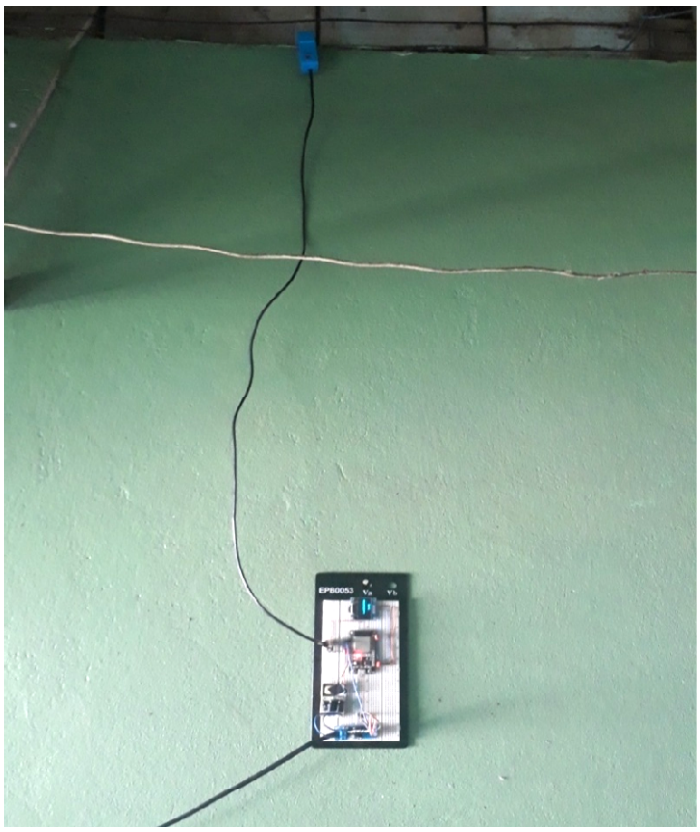
O Scikit-Learn possui diversos métodos padrões supervisionados e não supervisionados de aprendizado de máquina, além de possui ferramentas para a seleção e avaliação dos algoritmos gerados, transformação e carga de dados e persistência de modelos. Essa biblioteca pode ser utilizada em tarefas de classificação, previsão, agrupamento entre outras (MCKINNEY, 2018).

5 RESULTADOS E DISCUSSÕES

O sistema embarcado desenvolvido realiza a aquisição dos dados, através do medidor de potência, Figura 9. O MP possui, aproximadamente, uma frequência de amostragem de $4600Hz$, que significa ter um intervalo de amostragem de $0{,}2ms$. Para a publicação das leituras, foram utilizados os serviços do Google *Forms*. Através desse serviço, é possível coletar informações por meio de um questionário, conectadas automaticamente a uma planilha.

O ESP32 manda os dados coletados como respostas às perguntas estabelecidas no Google *Forms*. O Wi-Fi presente no ESP32 foi programado para funcionar como uma interface de estação, através da biblioteca Network do MicroPython. Isto posto, o ESP32 realiza leituras de tensão e corrente instantâneas, calcula a potência ativa, reativa e fator de potência, se conecta a um ponto de acesso à internet fornecida por um aparelho celular, fazendo uma solicitação HTTP para enviar as leituras ao Google *Forms* que irá salvar os valores em uma planilha no Google *Sheets*.

Figura 9 – Aquisição de dados em uma residência.



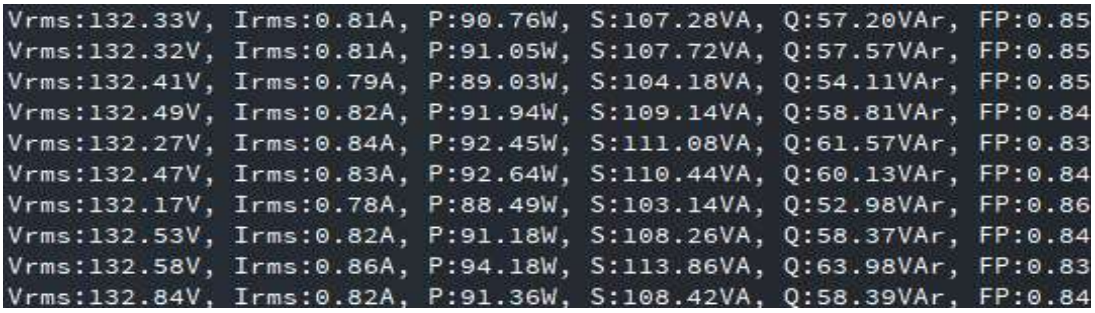
Fonte: elaborado pela autora.

As variáveis medidas pelo MP, potência ativa, potência reativa e fator de potência foram definidas como Real_Power, Reactive_Power e Power_Factor, respectivamente. Assim como a variável referente a data e hora como Timestamp no banco de dados.

5.1 Calibração dos sensores

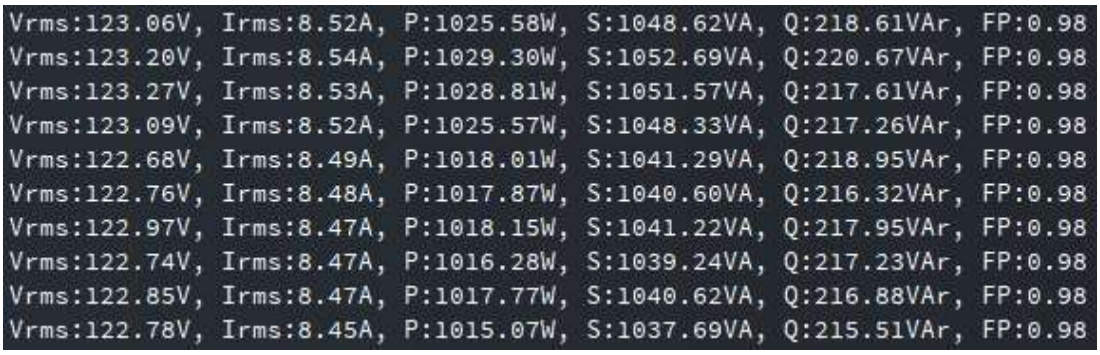
Para verificar a precisão das leituras feitas pelo MP, os sensores de corrente e tensão forma calibrados com um multímetro e um alicate amperímetro, respectivamente. Duas medições foram realizadas para a comparação das leituras: em um eletrodoméstico indutivo (ventilador), Figura 10, e um eletrodoméstico resistivo (ferro elétrico), Figura 11. Em seguida, os sensores foram ajustados.

Figura 10 – Parâmetros elétricos de um ventilador 120W medidos pelo MP.



Fonte: elaborado pela autora.

Figura 11 – Parâmetros elétricos de um ferro elétrico 1200W medidos pelo MP.



Fonte: elaborado pela autora.

Posteriormente, foi efetuada uma comparação aproximada, presente na Tabela 5, entre MP e o medidor eletrônico monofásico da fabricante Eletra Energy Solutions para uma maior confiança nas medições realizadas pelo aparelho desenvolvido.

Tabela 5 – Valores Medidos em kWhrs.

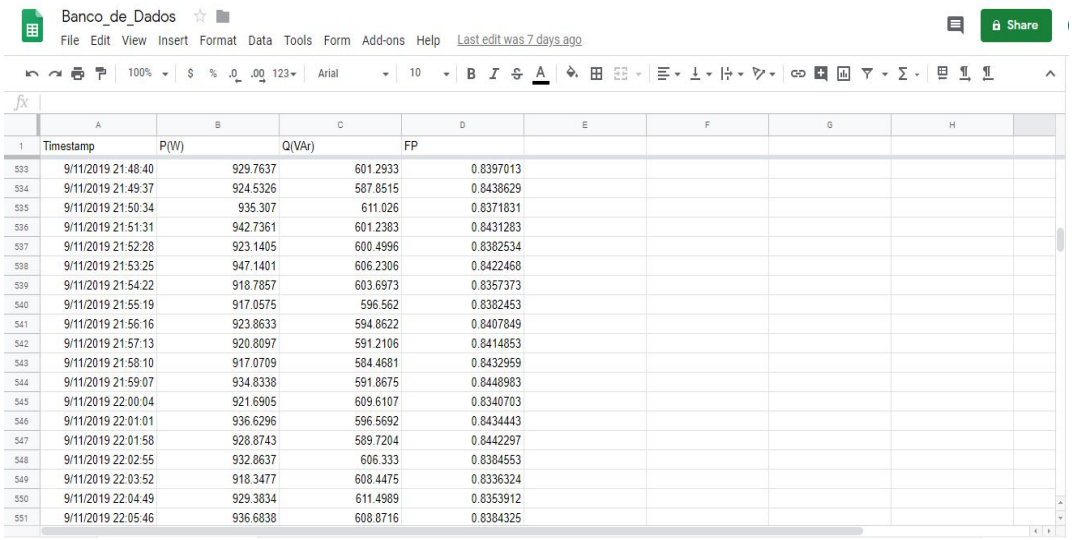
MP (kWhrs)	Medidor Eletrônico (kWhrs)
1,012	1,0
2,039	2,0

Fonte: elaborado pela autora.

5.2 Banco de Dados de Curvas de Carga

Foram feitas coletas de dados em dez residências na cidade de Breu Branco-PA, monofásicas, entre os meses de Junho e Outubro de 2019, para a construção do banco de dados. Os valores da potência diária consumida por cada domicílio foram registrados no Google *Sheets,* como mostrados na Figura 12, onde cada residência gerou 1440 linhas de dados.

Figura 12 – Banco de dados de curvas de carga.



	A	B	C	D	E	F	G	H
1	Timestamp	P(W)	Q(VAr)	FP				
533	9/11/2019 21:48:40	929.7637	601.2933	0.8397013				
534	9/11/2019 21:49:37	924.5326	587.8515	0.8438629				
535	9/11/2019 21:50:34	935.307	611.026	0.8371831				
536	9/11/2019 21:51:31	942.7361	601.2383	0.8431283				
537	9/11/2019 21:52:28	923.1405	600.4996	0.8382534				
538	9/11/2019 21:53:25	947.1401	606.2306	0.8422468				
539	9/11/2019 21:54:22	918.7857	603.6973	0.8357373				
540	9/11/2019 21:55:19	917.0575	596.562	0.8382453				
541	9/11/2019 21:56:16	923.8633	594.8622	0.8407849				
542	9/11/2019 21:57:13	920.8097	591.2106	0.8414853				
543	9/11/2019 21:58:10	917.0709	584.4681	0.8432959				
544	9/11/2019 21:59:07	934.8338	591.8675	0.8448983				
545	9/11/2019 22:00:04	921.6905	609.6107	0.8340703				
546	9/11/2019 22:01:01	936.6296	596.5692	0.8434443				
547	9/11/2019 22:01:58	928.8743	589.7204	0.8442297				
548	9/11/2019 22:02:55	932.8637	606.333	0.8384553				
549	9/11/2019 22:03:52	918.3477	608.4475	0.8336324				
550	9/11/2019 22:04:49	929.3834	611.4989	0.8353912				
551	9/11/2019 22:05:46	936.6838	608.8716	0.8384325				

Fonte: elaborado pela autora.

O Google *Sheets* é um programa de planilhas que faz parte de um pacote de software gratuito baseado na Web, concedido pelo Google em seu serviço de Google Drive. É um programa semelhante ao Microsoft Excel, e compatível com os formatos de arquivo do mesmo. Ele permite que os usuários criem e editem arquivos online, de forma simultânea e em tempo real.

A escolha do Google *Sheets*, para o armazenamento dos dados coletados e por consequência o levantamento do banco de dados, ocorreu devido não haver a necessidade de um computador ou servidor pago para a hospedagem do banco de dados criado.

O medidor de potência a cada um minuto envia para o banco de dados informações sobre potência ativa, potência reativa e fator de potência, a planilha registra a hora e data em que foram feitas as medições durante 24hrs. O tipo de variável presente no banco de dados está presente na Tabela 6.

Tabela 6 – Estrutura do Banco de Dados.

COLUNA	DESCRIÇÃO DA VARIÁVEL
Timestamp	Variável do tipo datatime, contendo os dados de data e hora do registro das informações de potência ativa, potência reativa e fator de potência.
P(W)	Variável do tipo float, contendo os dados de potência ativa.
Q(Var)	Variável do tipo float, contendo os dados de potência reativa.
FP	Variável do tipo float, contendo os dados de fator de potência.

Fonte: elaborado pela autora.

5.3 Perfil dos Consumidores

Para cada medição realizada dos consumidores de energia elétrica foi feito o registro de informações como: quantidade de adultos, quantidade de crianças, padrão social e quantas pessoas trabalham da residência. Cada residência teve seu padrão social definido de acordo com o valor de sua renda.

- Padrão Social Baixo: renda menor que um salário mínimo;
- Padrão Social Médio: renda maior que um salário mínimo e menor ou igual a dois salários mínimos;
- Padrão Social Alto: renda maior que dois salários mínimos.

Essas informações são de grande importância, pois influenciam e justificam o comportamento das curvas de consumo, além de identificar possíveis anomalias. As curvas de

carga dos consumidores das dez residências coletadas estão exibidas na Figura 13, assim como suas características logo abaixo.

Residência 01: família composta por três adultos (dois trabalham) e duas crianças. Com um padrão social médio, possuindo alguns eletrodomésticos, entre eles uma bomba d'agua.

Residência 02: família composta por dois adultos (ambos trabalham), sem crianças. Com um padrão social baixo, possuindo poucos eletrodomésticos.

Residência 03: família composta por dois adultos (ambos trabalham) e uma criança. com um padrão social médio, possuindo alguns eletrodomésticos, entre eles uma bomba d'agua.

Residência 04: família composta de dois adultos (ambos trabalham), duas crianças, com um padrão social médio, possuindo alguns eletrodomésticos.

Residência 05: constituída por um adulto (trabalha), sem crianças. Com um padrão social médio, possuindo alguns eletrodomésticos.

Residência 06: família composta por dois adultos (um trabalha), sem crianças. Com um padrão social médio, possuindo alguns eletrodomésticos.

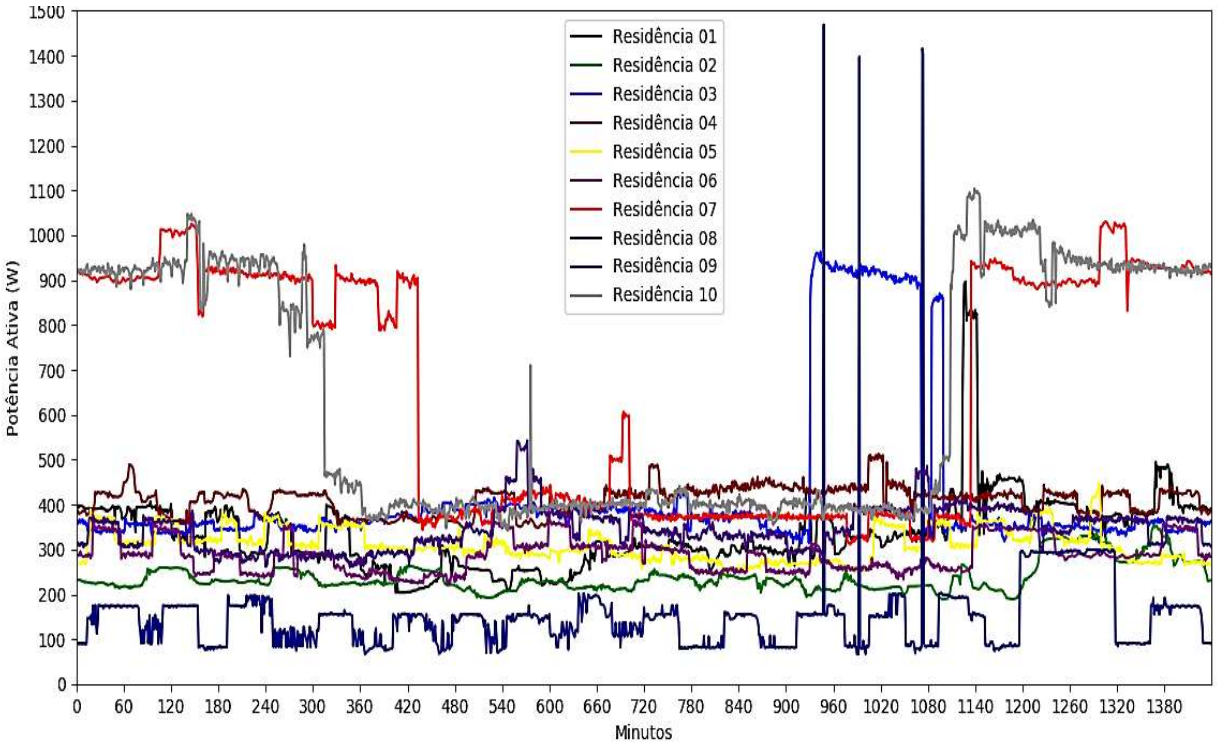
Residência 07: família composta por dois adultos (um trabalha) e uma criança. Com um padrão social alto, possuindo vários eletrodomésticos, entre eles uma bomba d'agua e um ar-condicionado.

Residência 08: constituída por um adulto (universitário), sem crianças. Com um padrão social baixo, possuindo poucos eletrodomésticos, entre eles um micro-ondas.

Residência 09: família composta por dois adultos (um trabalha), uma criança. Com um padrão social médio, possuindo alguns eletrodomésticos.

Residência 10: família composta por dois adultos (ambos trabalham), duas crianças e um adolescente. Com um padrão social alto, possuindo vários eletrodomésticos, entre eles um ar-condicionado.

Figura 13 - Curvas de carga residenciais coletadas.



Fonte: elaborado pela autora.

5.4 Pré-processamento e Modelagem dos Dados

O conjunto de dados, presente no banco de dados de curvas de carga construído com as leituras feitas pelo MP, foi submetido ao KDD com uma tarefa de Mineração de Dados de classificação, através da técnica de Árvore de Decisão, para a identificação de perdas comerciais de energia elétrica. O modelo do algoritmo gerado tem como objetivo predizer a variável categórica alvo, neste caso, definida como a classe a qual o consumidor pertence, através do comportamento das curvas de carga residencial. Considerando que o cliente da concessionária é regular, o mesmo pertence à classe 0, caso seja um cliente irregular (fraudador), este pertence à classe 1. Na etapa de treinamento do algoritmo, dada uma diferença entre as curvas de carga coletadas, devido ao contraste entre o padrão social dos consumidores, o conjunto de dados iniciais para treinamento foi dividido em três:

- Padrão Social Baixo: duas curvas de carga;
- Padrão Social Médio: seis curvas de carga;
- Padrão Social Alto: duas curvas de carga.

As informações de hora, potência ativa, potência reativa e fator de potência coletados através do MP foram definidos de forma proposital para que fossem usados como variáveis de entrada para a etapa de treinamento da Árvore de Decisão, pois são variáveis fortemente ligadas a característica do consumidor de energia elétrica, evitando a necessidade de reformatação dos dados para selecionar as variáveis de entrada.

Os dados coletados e armazenados no banco de dados do Google *Sheets* possuem extensão xlsx, uma extensão compatível com as bibliotecas de análise de dados do Python, evitando qualquer necessidade de conversão para outro tipo de extensão. Após essa verificação foi iniciada o pré-processamento dos dados.

O banco de dados não apresentou dados ausentes, algumas linhas duplicadas foram detectadas devido ao tempo de processamento do código no ESP32 e removidas, a variável data e hora foi convertida de objeto *datatime* para objeto int representando os minutos equivalentes à hora da medição efetuada para uma melhor manipulação dos dados.

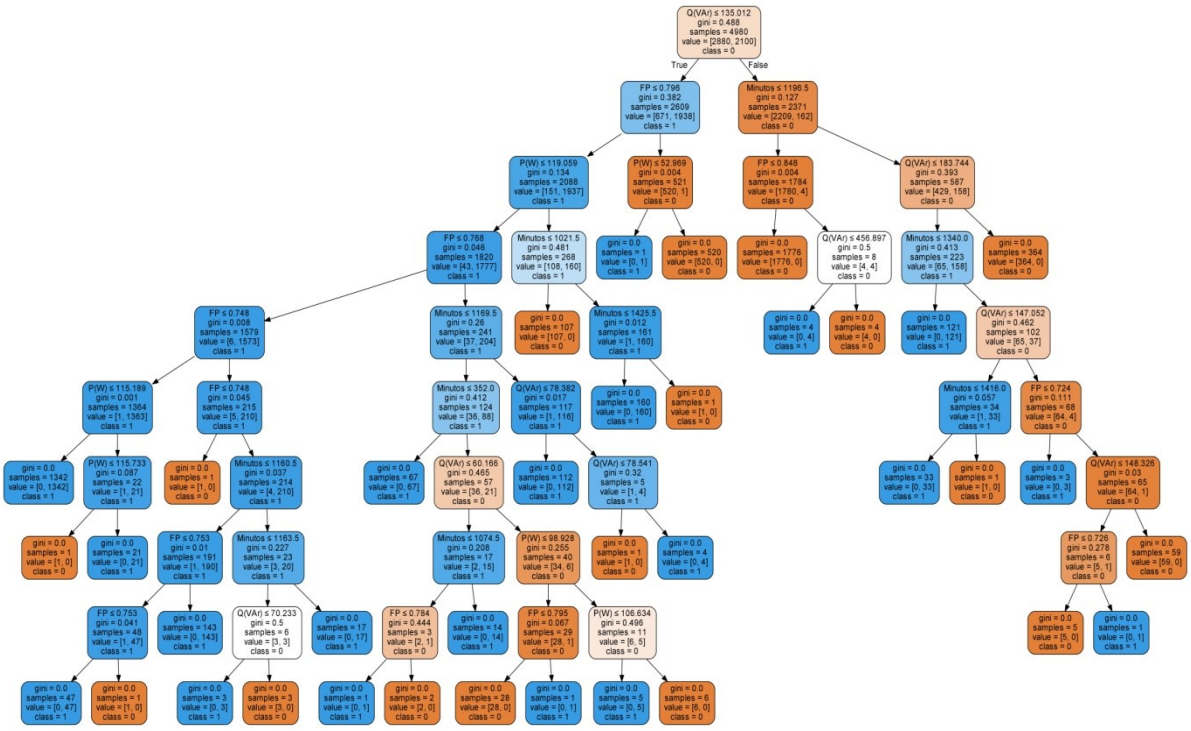
Nos dados coletados pelo MP não houve identificação de fraudes. Em vista disso, foram gerados dados através de simulações consumidores irregulares, a partir da média dos três grupos do conjunto de dados iniciais dos consumidores regulares, foram gerados dois conjuntos de dados para cada um dos três grupos. Possuindo os seguintes tipos de fraude:

1. Subtraído um ou dois desvios padrão das variáveis de entrada dos consumidores regulares.

2. Reduzida pela metade a potência ativa e reativa dos consumidores regulares no período de consumo noturno.

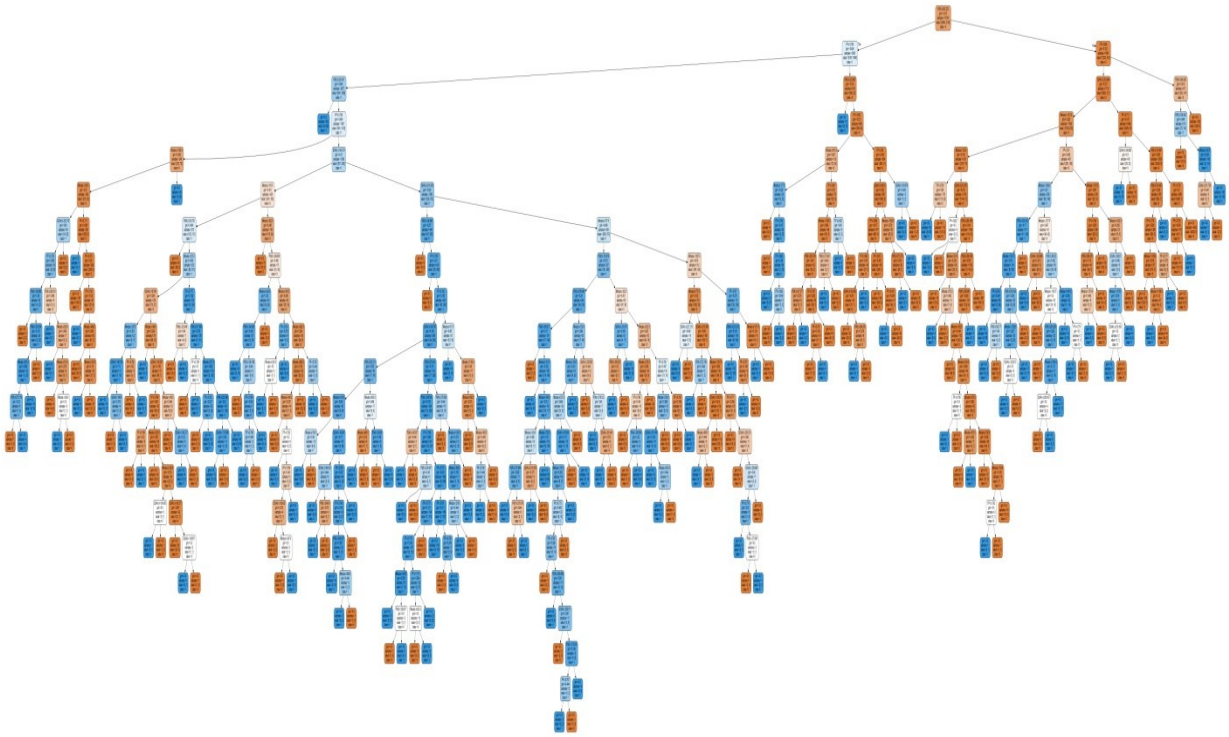
Finalizada a etapa de treinamento dos classificadores foi possível gerar as imagens, como mostrado nas Figuras 14, 15 e 16 das Árvores de Decisão para o padrão social baixo, médio e alto através da biblioteca Scikit-Learn, respectivamente. Que também estão presentes nos Apêndices A, B e C para uma melhor visualização

Figura 14 - Árvore de decisão para consumidores de padrão social baixo.



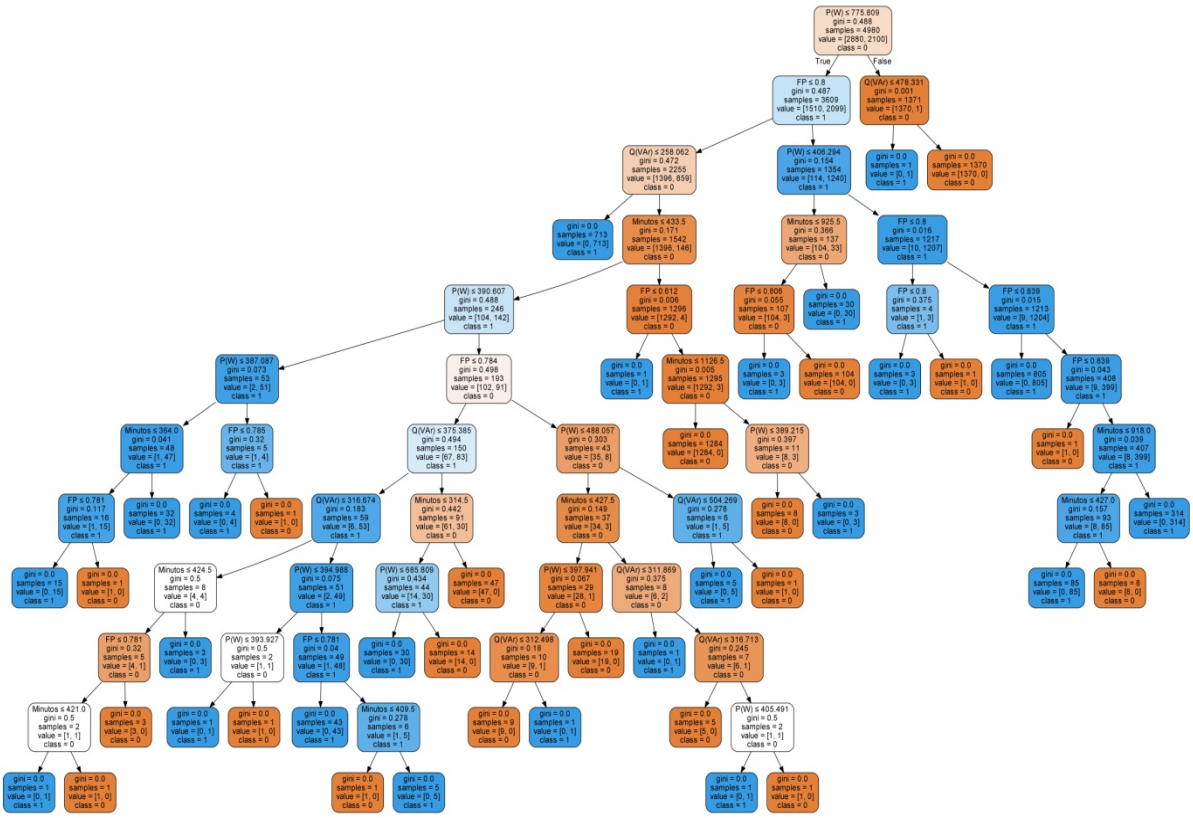
Fonte: elaborado pela autora.

Figura 15 - Árvore de decisão para consumidores de padrão social médio.



Fonte: elaborado pela autora.

Figura 16 - Árvore de decisão para consumidores de padrão social alto.



Fonte: elaborado pela autora.

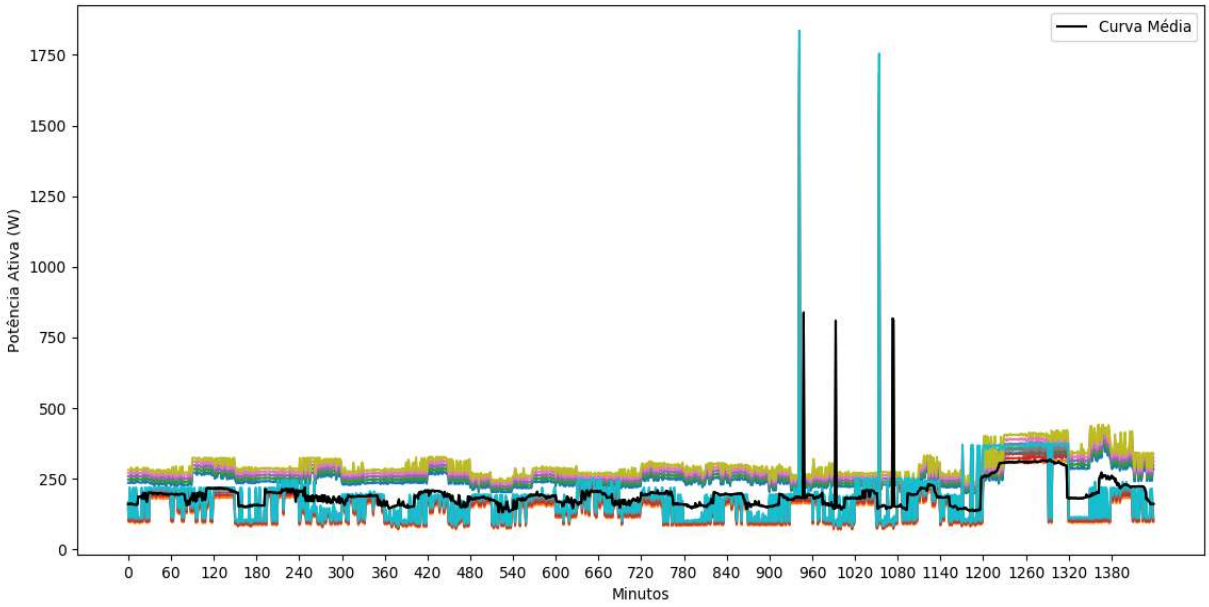
5.5 Validação dos Classificadores

Para verificar a eficácia dos classificadores em predizer a prática dos consumidores, foi aplicada a ferramenta de avaliação Matriz de Confusão para saber como os classificadores estão se comportando, além de métricas como: acurácia, precisão, recall e f-score.

Todos os dados coletados foram utilizados para treinamento de cada Árvore de Decisão na fase de treinamento. Por isso, foram geradas curvas a partir dos dados reais, com escolha de valores aleatórios e um aumento de: 5, 10, 15, 20 e 25%. Sendo considerado como um Verdadeiro Positivo, sequências de indicação da classe 1 de no mínimo um intervalo de 10 minutos consecutivos.

Para o classificador de padrão social baixo foram geradas 10 curvas regulares para teste como mostra a Figura 17.

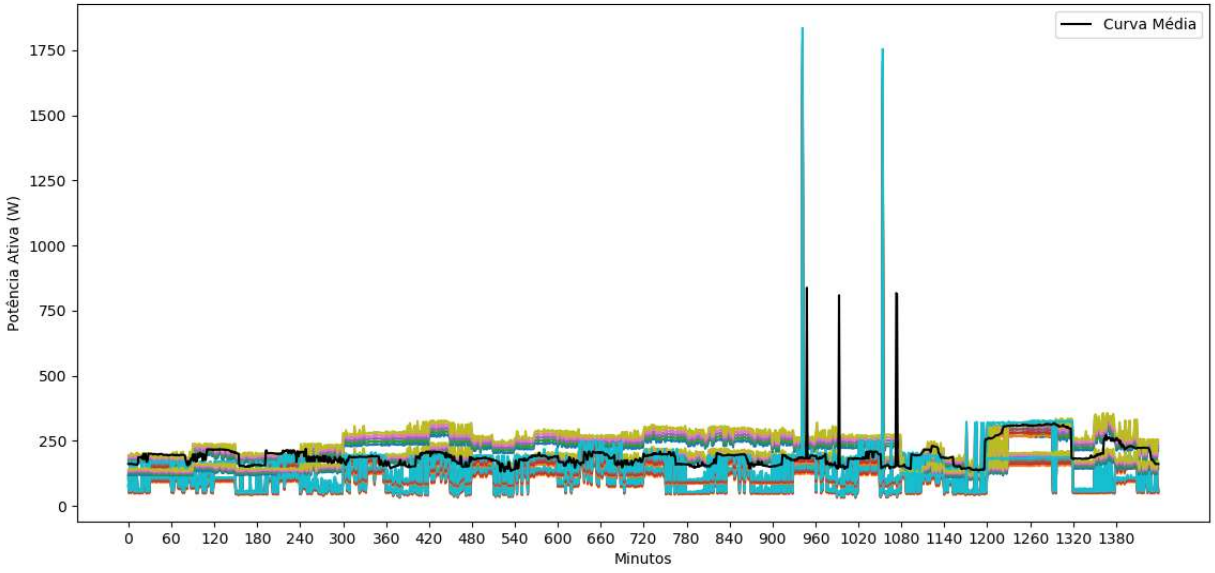
Figura 17 – Curvas regulares geradas para teste do classificador de consumidores com Padrão Social Baixo.



Fonte: elaborado pela autora.

As curvas irregulares foram geradas a partir das curvas regulares, contendo curvas reduzidas em meio e dois desvios padrões, assim como curvas com seus consumos reduzidos pela metade no período noturno, totalizando 20 curvas fraudadas, Figura 18.

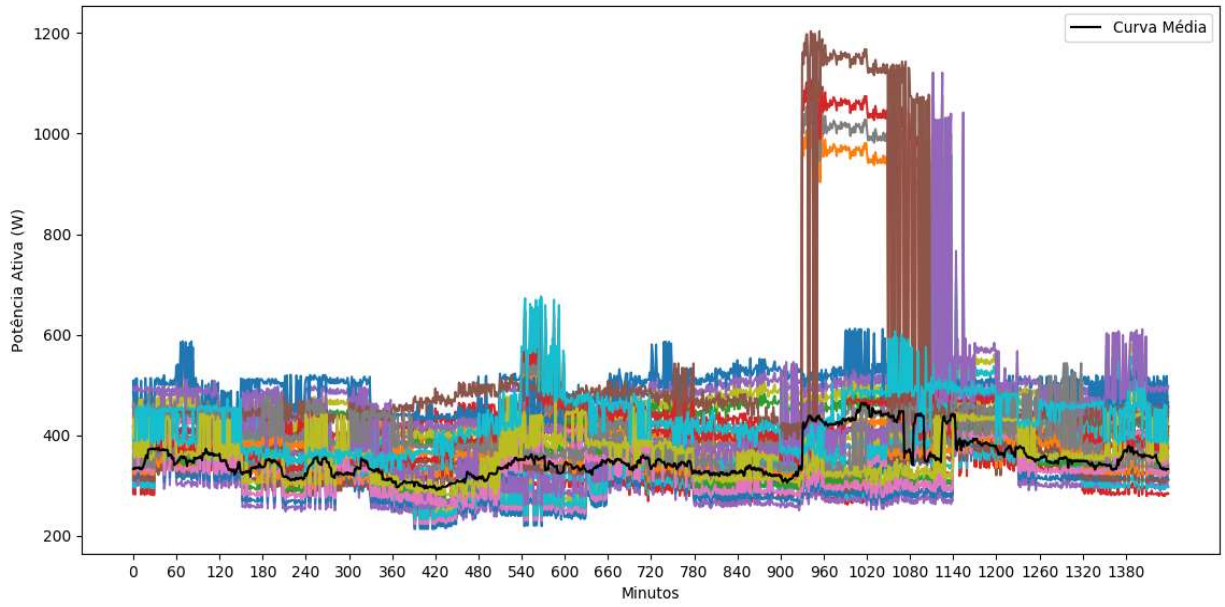
Figura 18 - Curvas irregulares geradas para teste do classificador de consumidores com Padrão Social Médio
.



Fonte: elaborado pela autora.

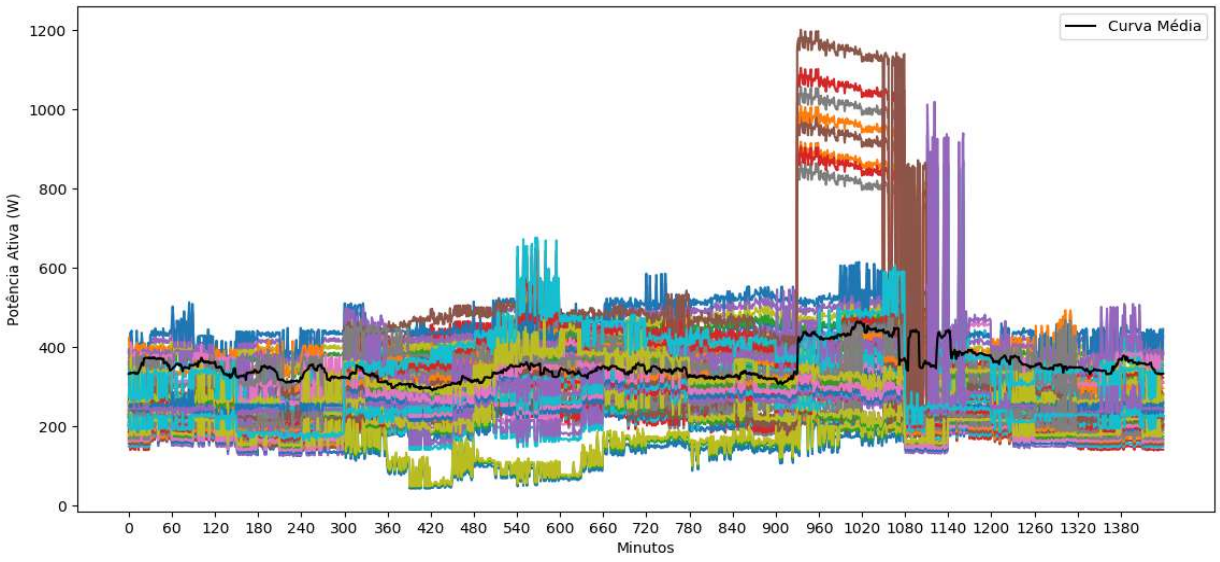
No classificador de padrão social médio foram testadas 30 curvas regulares, Figura 19. E geradas 60 curvas irregulares, entre elas curvas reduzidas com meio, um e dois desvios padrões, assim como curvas reduzidas pela metade no período noturno, Figura 20.

Figura 19 - Curvas regulares geradas para teste do classificador de consumidores com Padrão Social Médio.



Fonte: elaborado pela autora.

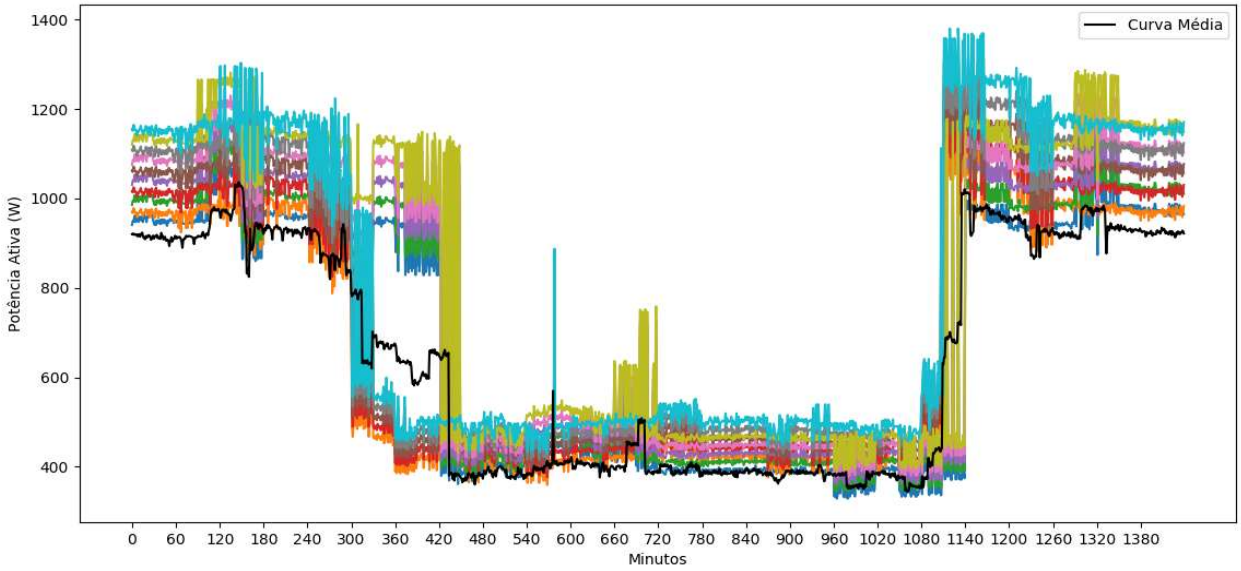
Figura 20 - Curvas irregulares geradas para teste do classificador de consumidores com Padrão Social Médio.



Fonte: elaborado pela autora.

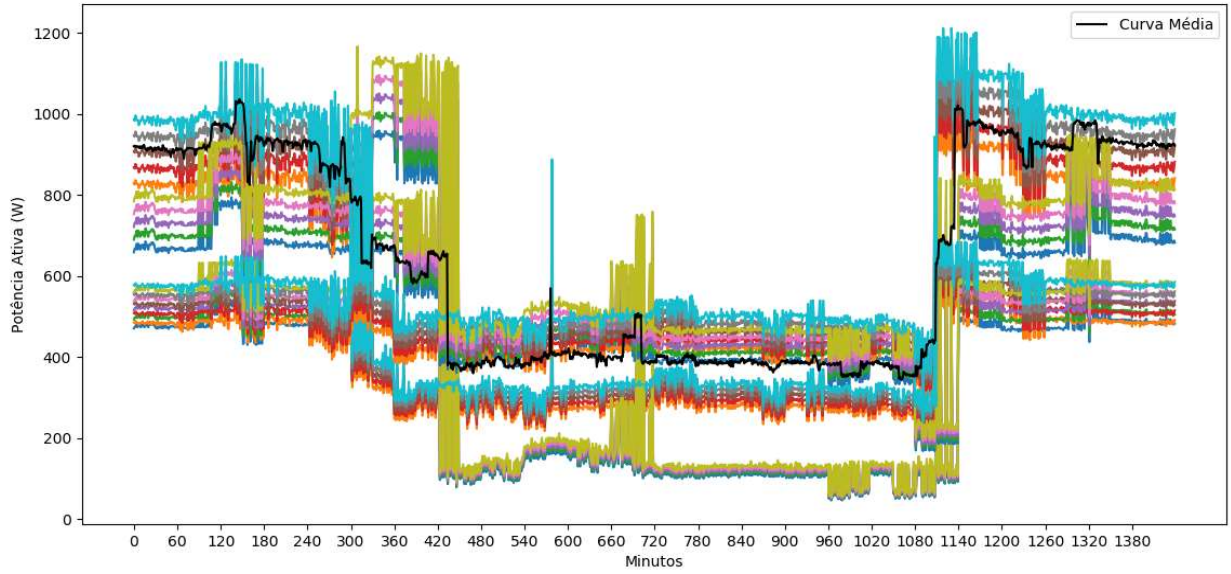
Para o classificador de padrão social alto assim como o de padrão social baixo foram geradas 10 curvas regulares para teste como mostra a Figura 21, curvas irregulares, Figura 22, reduzidas em meio e um desvio padrão, assim como curvas com seus consumos reduzidos pela metade no período noturno, com um total de 20 curvas fraudadas.

Figura 21 - Curvas regulares geradas para teste do classificador de consumidores com Padrão Social Alto.



Fonte: elaborado pela autora.

Figura 22 - Curvas irregulares geradas para teste do classificador de consumidores com Padrão Social Alto.



Fonte: elaborado pela autora.

Cada Árvore de decisão foi treinada, com curvas de furtos geradas a partir da média dos dados reais coletados, com o propósito de ensinar ao classificador que curvas de carga acima da média serão consideradas de consumidores regulares, e curvas abaixo da média um indicador que este consumidor tem uma grande possibilidade de ser um fraudador. Embora os dados fraudados inseridos para treinamento tenham sido com a redução de um ou dois desvios padrões. Como mencionado anteriormente durante a etapa de validação foram geradas curvas irregulares com a diminuição de meio, um e dois desvios padrões, para verificar como o classificador responde a pequenas atenuações no consumo e que estão abaixo da média de cada conjunto de dados.

A quantidade de TN, TP, FN e FP de cada Árvore de Decisão de padrões sociais baixo, médio e alto estão presentes nas Matrizes de Confusão na Tabela 7, 8 e 9, respectivamente.

Na Tabela 10 é possível perceber uma maior acurácia no conjunto de dados de padrão social médio, isso se dá pelo fato que este classificador foi treinado com mais curvas que os outros demais grupos. Pode-se dizer que o classificador compreende melhor diferentes situações de furtos e de consumidores regulares. Porém, todos os três classificadores tiveram ótimos índices de TP e TN, pois suas taxas de acurácia são maiores que 80%. Devido à maioria dos dados classificados estarem representados na diagonal da Matriz de Confusão (HAN; KAMBER; PEI, 2012).

Tabela 7 – Matriz de Confusão da Árvore de Decisão do grupo Padrão Social Baixo.

Classe Real	**Classe Predita**	
	Consumidor Irregular (1)	**Consumidor Regular (0)**
Consumidor Irregular	16	0
Consumidor Regular	4	10

Fonte: elaborado pela autora.

Tabela 8 – Matriz de Confusão da Árvore de Decisão do grupo Padrão Social Médio.

Classe Real	**Classe Predita**	
	Consumidor Irregular (1)	**Consumidor Regular (0)**
Consumidor Irregular	58	2
Consumidor Regular	7	23

Fonte: elaborado pela autora.

Tabela 9 – Matriz de Confusão da Árvore de Decisão do grupo Padrão Social Alto.

Classe Real	**Classe Predita**	
	Consumidor Irregular (1)	**Consumidor Regular (0)**
Consumidor Irregular	20	4
Consumidor Regular	0	6

Fonte: elaborado pela autora.

Tabela 10 – Acurácia das Árvores de Decisão.

Padrão Social	**Acurácia (%)**
Baixo	86
Médio	90
Alto	86

Fonte: elaborado pela autora.

Através da precisão, Tabela 11, é indicada a proporção de identificações positivas, para esta situação, quanto o modelo gerado acertou na predição correta do consumidor irregular. O classificador de padrão social baixo teve uma precisão de 100%, indicando uma ótima predição na classe 1.

Tabela 11 – Precisão das Árvores de Decisão.

Padrão Social	**Precisão (%)**
Baixo	100
Médio	90
Alto	83

Fonte: elaborado pela autora.

O classificador de padrão social alto alcançou uma taxa de 100%, referente à sua proporção de identificação de positivos preditos corretamente, dado pela taxa de Recall, Tabela 12.

Tabela 12 – Recall das Árvores de Decisão.

Padrão Social	Recall (%)
Baixo	80
Médio	96
Alto	100

Fonte: elaborado pela autora.

É importante ressaltar, que uma taxa de precisão de 100% para a classe 1, revela que cada dado do conjunto de teste, rotulado com a classe 1, realmente pertence a classe 1. Entretanto, não se refere a nada sobre o número de dados que o classificador rotulou incorretamente como classe 1 (HAN; KAMBER; PEI, 2012). Do mesmo modo, isso ocorre com a taxa recall de 100%. Podendo existir uma relação inversa entre as duas taxas, como o classificador ter alta precisão identificando o consumidor irregular corretamente, mas em custo de um baixo recall, classificando muitos outros casos incorretamente de consumidores irregulares. Portanto a necessidade da medida f-score que representa uma média entre precisão e recall. Assim podemos visualizar na Tabela 13, que a Árvore de Decisão que apresentou a melhor ponderação entre precisão e recall.

Tabela 13 – F-score das Árvores de Decisão.

Padrão Social	F-score (%)
Baixo	88
Médio	93
Alto	90

Fonte: elaborado pela autora.

6 CONCLUSÕES

6.1 Conclusões Gerais

Neste trabalho foi apresentada uma metodologia como forma de agilizar a identificação de perdas comerciais de energia elétrica. Esta metodologia é constituída por um banco de dados próprio para a aplicação de KDD construído a partir de dados coletados por um Medidor de Potência programado para a leitura de potência ativa, potência reativa e fator de potência dos consumidores de energia elétrica. Utilizou-se a técnica de Árvore de Decisão, com tarefa de classificação, desenvolvida, treinada e testada através da biblioteca Scikit-Learn oferecida pela linguagem de programação Python.

A construção do MP permitiu o levantamento da curva de carga dos consumidores, possibilitando uma melhor estimativa do comportamento do consumidor de energia elétrica, o qual teve como objetivo ser um meio para a construção do banco de dados com poucas variáveis porém, fortemente ligadas com a problemática de perdas comerciais e necessárias para o treinamento dos classificadores. A comunicação via wi-fi foi utilizada por se tratar de uma forma rápida e eficiente de enviar as informações coletadas pelo MP para o banco de dados.

O propósito da construção de um banco de dados próprio para a aplicação do KDD como forma de diminuir o tempo da etapa de pré-processamento dos dados foi alcançado. Durante a esta etapa o único problema identificado foi a presença de dados duplicados referente à hora registrada das leituras do MP. Excluindo qualquer necessidade de preenchimento de dados ausentes, substituição de valores, junção, combinação e reformatação dos dados que são algumas das atividades da etapa de pré-processamento dos dados que demandam cerca de 80% do tempo do processo de descoberta de conhecimento em banco de dados.

As três Árvores de Decisão geradas para os três padrões sociais identificados dos consumidores apresentaram bons desempenhos. Tendo os consumidores de padrão social baixo, médio e alto, taxas de acurácia em torno de 86, 90 e 86% respectivamente, o que mostra que é uma técnica válida e que pode ser aplicada, levando em consideração o comportamento do consumidor para identificar fraude e furtos no sistema de distribuição de energia elétrica.

Dada a importância do assunto abordado neste trabalho e nos resultados obtidos é possível afirmar que a utilização de técnicas de mineração de dados é indispensável sendo capaz de lidar com uma grande quantidade de dados, de grande aplicabilidade destacou-se como sendo uma poderosa ferramenta de apoio para identificar possíveis consumidores irregulares.

6.2 Trabalhos Futuros

O presente trabalho tem como possibilidade algumas propostas de melhoria para aprimorar o número de identificação de perdas comerciais no sistema de distribuição de energia elétrica, além de novas linhas de pesquisas:

1. Desenvolvimento de um medidor de potência bifásico e trifásico;

2. Transmissão dos dados, sem fio, através de Wi-Fi a longa distância como estudo da viabilidade de uma central da concessionária de monitoramento de energia;

3. Aplicação de Redes Neurais aos dados coletados para a comparação da taxa de acurácia entre técnicas de Mineração de Dados;

4. Ampliação dos dados coletados, para o enriquecimento do banco de dados com diferentes características de consumidores de energia elétrica, com o objetivo de aperfeiçoar os classificadores na fase de treinamento.

5. Implementação da tarefa de clusterização, não supervisionada, através da técnica de k- vizinhos mais próximos no banco de dados.

REFERÊNCIAS

BASGALUPP, M.P. **LEGAL-Tree**: Um algoritmo genético multi-objetivo lexicográfico para indução de árvores de decisão. 2010. 116f. Tese de Doutorado, ICMC-USP, São Carlos, 2010.

CARVALHO, Luís A. V. **DATAMINING**: A Mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração. 2a Edição. São Paulo: Érica, 2002.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados**. São Paulo: Saraiva, 2016.

CASTANHEIRA, L. G. **Aplicação de Técnicas de Mineração de Dados e Problemas de Classificação de Padrões**. Dissertação de mestrado, universidade Federal de Minas Gerais – UFMG, 2008.

CÔRTES, S. C. **Mineração de Dados: Funcionalidades, Técnicas e Abordagens**. PUC-RJ. Artigo, 2002 Disponível em: ftp://ftp.inf.puc-rio.br/pub/docs/ techreports/02_10_cortes.pdf.

CORREIA MATOS, Y. C. **Detecção de Fraudes no Consumo de Energia Elétrica Usando Árvores de Decisão**. 2017. 59f. Dissertação de Mestrado em Engenharia Elétrica – Universidade Federal do Pará, Belém, 2017.

ElectricityMonitoring. Disponível em: <learn.openenergymonitor.org>. Acesso em: 06 Abril 2019.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining toKnowledge Discovery in Databases**. American Association for Artificial Intelligence, 1996.

HAN, J.; KAMBER, M.; PEI, J. **Data mining:** conceptsandtechniques. 3ª edição. Whaltman: Morgan Kaufmann, 2011.

KAGAN, N.; OLIVEIRA, C. C. B.; ROBBA, E. J. **Introdução aos sistemas de distribuição de energia elétrica**. São Paulo: Blucher, 2010.

LAROSE, D. T. **DiscoveringKnowledge in Data**: AnIntroductionto Data Mining. John Wileyand Sons, Inc, 2005.

MCKINNEY, Wes. **Python para análise de dados**: tratamento de dados com pandas, numpy e ipython. São Paulo: Novatec, 2018.

Perdas de Energia – Informações Técnicas – ANEEL. Disponível em: <www.aneel.gov.br/informacoes-tecnicas/-/asset_publisher/CegkWaVJWF5E/content/perdas/654800?inheritRedirect=false>. Acesso em: 01 Abril 2019.

QUEIROZ, A. S. **Algoritmos de Inteligência Computacional Utilizados na Detecção de Fraudes nas Redes de Distribuição de Energia Elétrica**.2016. 64f. Dissertação de Mestrado em Engenharia Elétrica e Computação – Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, 2016.

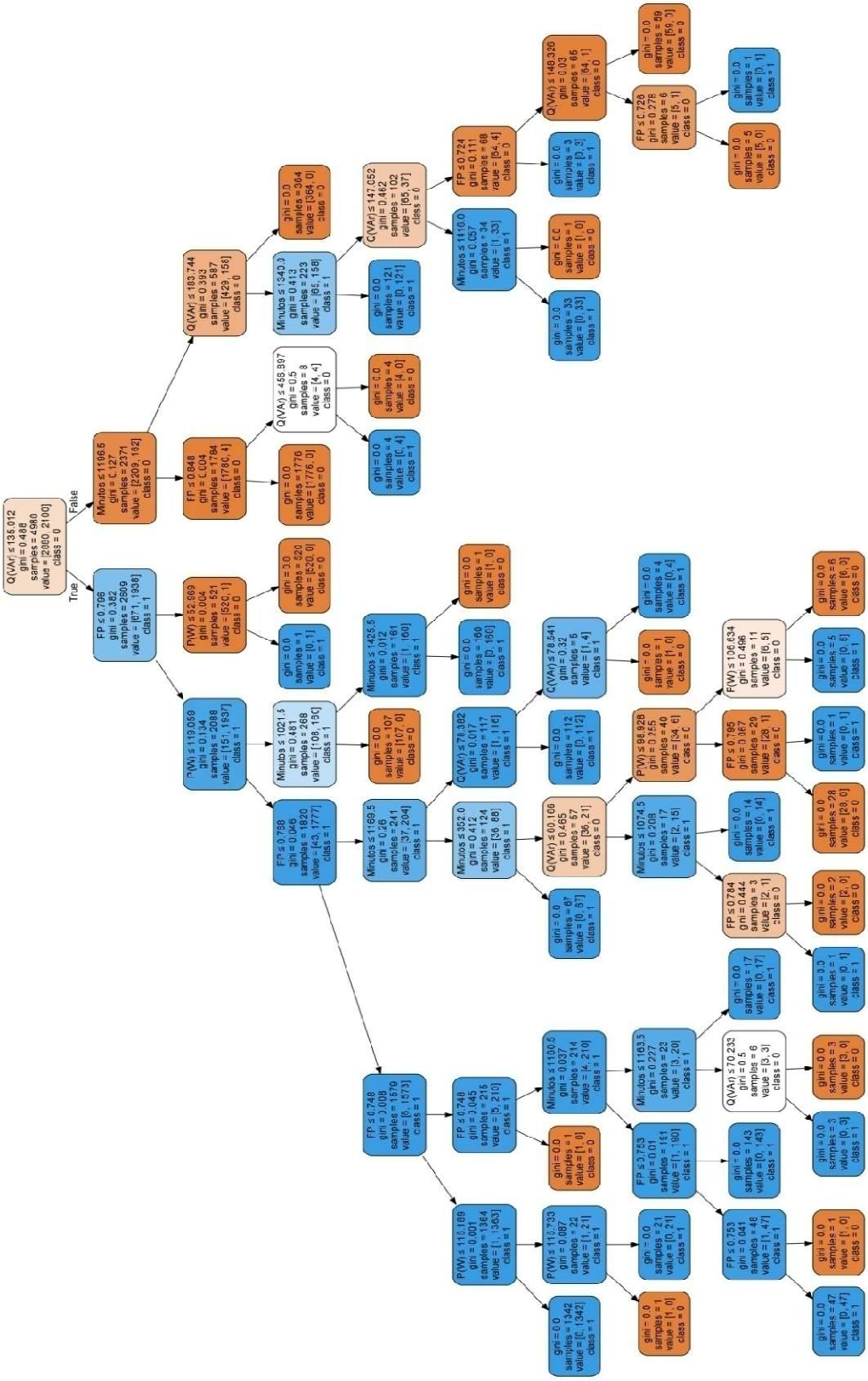
Perdas de Energia Elétrica na Distribuição. Disponível em: <www.aneel.gov.br/documents/654800/18766993/Relat%C3%B3rio+Perdas+de+Energia_+Edi%C3%A7%C3%A3o+1-2019-02-07.pdf/d7cc619e-0f85-2556-17ff-f84ad74f1c8d>. Acesso em: 01 Abril 2019.

TELLES, Matt. **Python Power!**: the comprehensive guide. 1a Edicao. Boston: Editora Cengage Learning PTR, 2008.

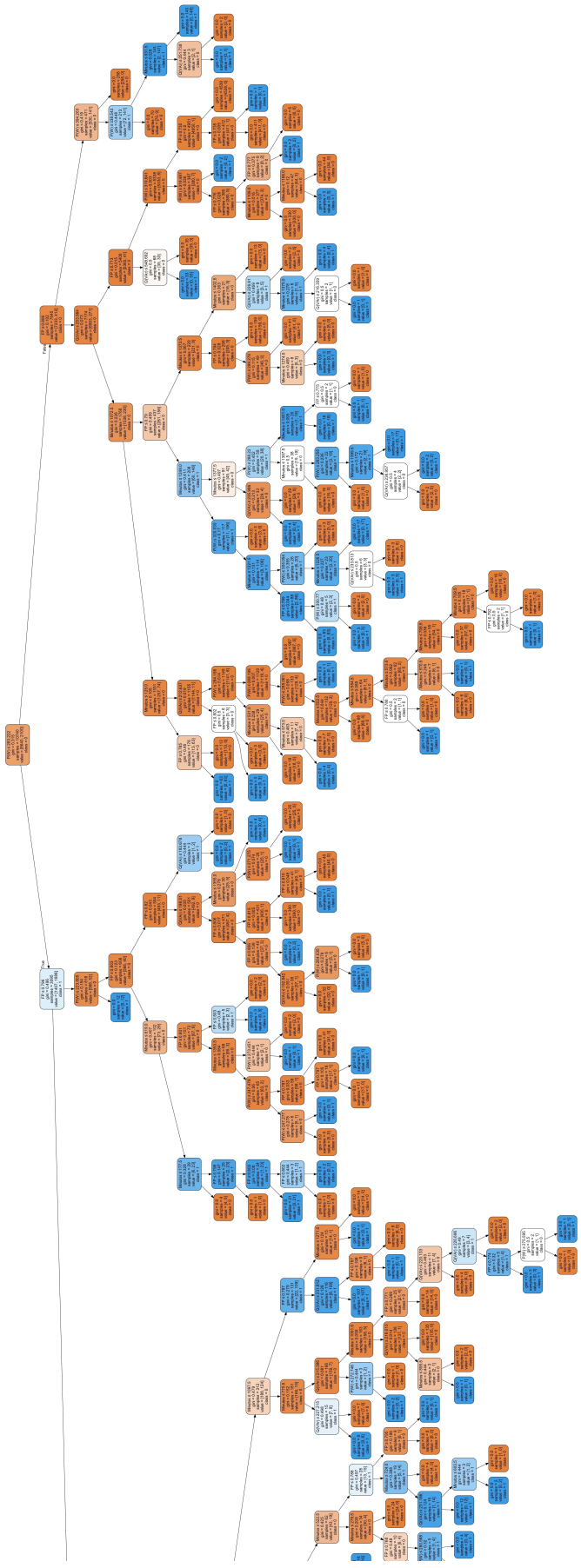
STALLINGS, W. **Arquitetura e Organização de Computadores**. 8ª edição São Paulo: Pearson Pratice Hall, 2010.

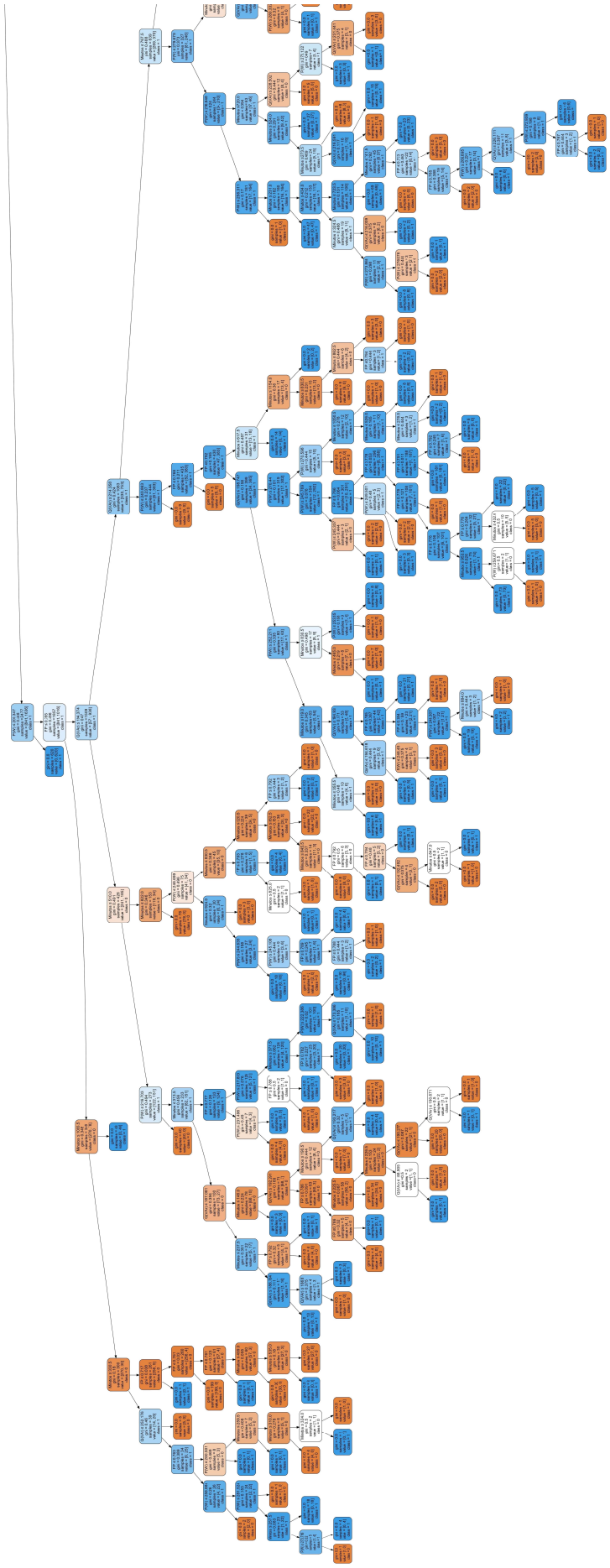
ZUEGE, T. J. **Aplicação de Técnicas de Mineração de Dados para Detecção de Perdas Comerciais na Distribuição de Energia Elétrica**. 2018. 91f. Trabalho de Conclusão de Curso - Centro de Ciências Exatas e Tecnológicas do Vale do Taquari, Porto Alegre, 2018.

APÊNDICE A – Árvore de decisão para consumidores de padrão social baixo.



APÊNDICE B – Árvore de decisão para consumidores de padrão social médio.





APÊNDICE C – Árvore de decisão para consumidores de padrão social alto.