Detecção de Fraude de Energia Elétrica Baseada em Análise de Anomalias

Electricity Fraud Detection Based on Anomaly Analysis

Nathalia A. da Fonseca Alves * **Maria Eusa Alves P. de Oliveira** *
Juan Santos Cavalcanti de Albuquerque *
Héctor Raúl Chávez Arias ** **Yuri Percy Molina Rodriguez** ***
Thainá Matos Santana Delgado *** **Camila Mara Vital Barros** ***

* *Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, João Pessoa, Paraíba, Brasil (e-mail: nathalia.alves@estudante.cear.ufpb.br, maria.oliveira@estudante.cear.ufpb.br, juan.albuquerque@estudante.cear.ufpb.br)*
** *Facultad de Ingenieria Mecânic,a, Universidad Nacional de Ingenieria, Lima, Peru (e-mail: hector.chavez.a@uni.pe)*
*** *Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal da Paraíba, João Pessoa, Paraíba, Brasil (e-mail: thaina.delgado@estudante.cear.ufpb.br, molina.rodriguez@cear.ufpb.br, camila.barros@ci.ufpb.br)*

Abstract: Electricity utilities face substantial financial losses due to energy theft, generating negative impacts for both consumers and the companies themselves. This article explores a methodology that uses supervised learning, namely Decision Trees, to identify anomalies in electricity distribution networks. An analysis was made of the consumption data of around 523,000 people, which was reduced to approximately 11,000 customers. These were assigned a metric classifying customers with a standard deviation of consumption above 30%. Finally, the consumers who scored the highest were effectively proven to be fraudsters after analysis of the training and test data.

Resumo: As concessionárias de energia elétrica enfrentam substanciais prejuízos financeiros devido aos furtos de energia, gerando impactos negativos tanto para os consumidores quanto para as próprias empresas. Este artigo explora uma metodologia que usa aprendizado supervisionado, sendo este Árvore de Decisão, para identificar anomalias em redes de distribuição elétrica. Foi feita uma análise dos dados do consumo de cerca de 523.000 pessoas, que foi reduzido para aproximadamente 11.000 clientes. A estes, foi atribuída uma métrica classificando os clientes que possuíam desvio padrão do consumo acima de 30%. Por fim, os consumidores que mais pontuaram foram efetivamente comprovados como fraudadores após a análise feita com os dados de treino e teste.

Keywords: Machine Learning; Artificial Intelligence; Energy Theft; Decision Tree; Anomaly; Electrical Grid;

Palavras-chaves: Machine Learning; Inteligência Artificial; Furto de Energia; Árvore de Decisão; Anomalia; Rede Elétrica;

1. INTRODUÇÃO

No Brasil, de acordo com a edição 01/2021 do Relatório ANEEL - Perdas de Energia Elétrica na Distribuição, Agência Nacional de Energia Elétrica (ANEEL) (2021), o valor total de perdas na distribuição representa aproximadamente 14,8% do mercado consumidor em 2020. Esse valor, segundo o relatório Agência Nacional de Energia Elétrica (ANEEL) (2021), supera o consumo das regiões Norte e Centro-Oeste do país no ano de 2018. Segundo Glauner et al. (2017), essas perdas podem ser classificadas em dois tipos: as Perdas Técnicas (PTs) e as Perdas Não Técnicas (PNTs). De acordo com o estudo de Park and Kim (2020), as PTs ocorrem no processo de transmissão, transformação e medição da eletricidade ao passo que as PNTs são principalmente causadas por furto, fraude ou problemas de faturamento. Conforme explicitado por Glauner et al. (2017), na maioria dos países, as PNTs representam a parte mais significativa das perdas totais, chegando até 40% da energia total distribuída.

Os métodos para detectar as PNTs descritos na literatura são divididos em duas categorias: sistemas especialistas e aprendizado de máquina. Os sistemas especialistas contêm regras elaboradas manualmente para realizar tomada de decisão. Por outro lado, o aprendizado de máquina proporciona aos computadores a habilidade de adquirir conhecimento a partir de exemplos, sem a necessidade de programação explícita. No decorrer dos anos, Glauner et al. (2017) e Gunturi and Sarkar (2021) enfatizam que o campo do aprendizado de máquina se tornou a direção de pesquisa predominante na detecção de PNTs.

Gunturi and Sarkar (2021) ressalta que dentre os tipos de perdas não técnicas, a questão do furto de energia sempre foi uma preocupação relevante para as empresas de serviços públicos ao redor do mundo, pois representa a maior parte das perdas totais. Aproximadamente 80% destas perdas ocorre por meio de conexões clandestinas.

Park and Kim (2020) ressaltaram que existem diferentes tipos de técnicas de classificação para prever se a energia está sendo usada de forma ilegal, como por exemplo: *Support Vector Machine* (SVM), árvore de decisão e *deep learning*. Neste cenário, há um desafio significativo ao utilizar tais técnicas: a falta de dados sobre possíveis ataques, como explicam Park and Kim (2020) e Himeur et al. (2021). Apesar da facilidade para obter informações sobre o consumo regular de eletricidade, a coleta de dados suficientes sobre casos de furto de energia para treinar um modelo de classificação é uma tarefa complexa. A abordagem de classificação de classes ou a detecção de *outliers* podem ser boas alternativas, pois se baseiam exclusivamente em dados normais, sem a necessidade de informações específicas sobre os ataques. De acordo com Himeur et al. (2021), é possível utilizar técnicas de Inteligência Artificial (IA) para detectar comportamentos anômalos de consumo de energia gerados tanto por usuários finais, falhas de aparelhos ou outras causas potenciais. Park and Kim (2020) afirmam que é fundamental destacar que a identificação de anomalias em amostras individuais de dados, representando o consumo diário de eletricidade, requer uma análise de pós-processamento adequada para confirmar casos de fraude de energia.

Este trabalho realiza detecção de anomalias utilizando IA para encontrar furtos de energia com base na leitura de medidores de consumo. Foi escolhido utilizar técnicas de classificação no conjunto de dados, neste caso, árvore de decisão. A principal contribuição deste artigo é validar a eficácia da técnica de aprendizado de máquina com Árvore de Decisão a fim de localizar furtos de energia. O presente trabalho está organizado da seguinte forma: no Capítulo 2 é detalhada a metodologia do modelo de aprendizado supervisionado. No Capítulo 3 são mostrados os resultados e as discussões sobre o que foi obtido com os testes. Por fim, no Capítulo 4 estão as conclusões e as ideias de possíveis melhorias para trabalhos futuros.

1.1 Revisão Bibliográfica

De maneira geral, Lacerda et al. (2022) e Himeur et al. (2021) ressaltam que existem três tipos de aprendizado de máquina que podem ser utilizados na detecção de outliers: o supervisionado, o não supervisionado e o híbrido.

Stracqualursi et al. (2023) abordam que aprendizado não supervisionado busca detectar padrões de consumo que sejam raros ou até então desconhecidos sem necessitar de conhecimentos prévios. Além disso, os dados de treinamento não são rotulados. Himeur et al. (2021) destacam que nesse tipo de detecção parte-se do pressuposto de que a quantidade de padrões anômalos nos dados de consumo é inferior a 20%. Stracqualursi et al. (2023) e Himeur et al. (2021) analisam que a detecção de anomalias consiste principalmente em modelar o comportamento de consumo normal na maioria dos casos, além de estabelecer métricas específicas nesse contexto para classificar as observações de consumo como normais ou anômalas. As abordagens não supervisionadas são predominantemente baseadas em algoritmos de agrupamento, aprendizado de uma classe e redução de dimensionalidade.

Em contrapartida, Himeur et al. (2021) explicam que a aprendizagem supervisionada de irregularidades no gasto de energia demanda o treinamento de classificadores de *machine learning* (ML) utilizando conjuntos de dados marcados, em que tanto o consumo regular quanto o irregular são identificados. Apesar da detecção supervisionada alcançar resultados de identificação altamente precisos, sua aplicação prática ainda é restrita em comparação com técnicas não supervisionadas, devido à falta de conjuntos de dados marcados relacionados ao consumo de energia.

Por outro lado, segundo Himeur et al. (2021), nenhum dos métodos de detecção de anomalias seria capaz de identificar de forma precisa todas as irregularidades por conta da complexidade dos dados de consumo de energia e fatores que influenciam seu uso em diferentes escalas de tempo, tais como: por hora, por dia, por semana, por mês ou até por estação do ano. Sendo assim, surge o aprendizado

[1] O presente estudo é financiado pelas seguintes instituições e fundações: Conselho Nacional de Desenvolvimento Científico e Tecnológico Desenvolvimento (CNPq) e a Universidade Federal da Paraíba (UFPB).

híbrido, que mescla os dois tipos citados anteriormente. Um exemplo de aplicação consiste em utilizar o conjunto inicial de dados de energia e dividi-lo em múltiplos subconjuntos e diversos modelos são aplicados simultaneamente a esses subconjuntos para identificar possíveis anomalias. Posteriormente, os *scores* de identificação de anomalias são sumarizados, e/ou o mais adequado é selecionado para gerar o resultado efetivo.

2. METODOLOGIA APLICADA

2.1 Pré-Processamento dos Dados

Para realizar o pré-processamento de dados, adotou-se a metodologia descrita por Yip et al. (2018), que detalha a conversão dos dados brutos em um formato padronizado. Este processo envolve três etapas principais: limpeza dos dados, padronização dos dados e extração de características importantes. A etapa de limpeza visa identificar e corrigir inconsistências ou valores ausentes no banco de dados. Na fase de padronização, é crucial identificar *outliers* e convertê-los para formatos aceitáveis usando métodos de normalização. Conforme sugerido por Qiu et al. (2018), este processo permite a separação das características mais relevantes da base de dados, que serão utilizadas no modelo de treinamento. O procedimento completo é ilustrado na Figura 1.

No presente trabalho, a base de dados final foi composta pelos valores de consumo dos medidores, informações sobre o alimentador, a data de medição, o número do cliente e a indicação de se o consumo ocorreu em horário de ponta ou fora de ponta. Para simplificar a análise, foi feita uma divisão da região dos consumidores em várias microrregiões (considerada uma região menor que está inserida dentro de uma região pré-delimitada) com tamanhos iguais, duas delas foram escolhidas para fazer a detecção de *outliers* constando 11.000 usuários na análise de uma base de dados de aproximadamente 523.000 pessoas.

Figura 1. Fluxograma detalhado do processo de pré-processamento

O conjunto de dados utilizado para treinamento e teste compreende informações de consumo mensal ao longo de 60 meses de Unidades Consumidoras (UCs), abrangendo tanto as regulares quanto as irregulares. Nesse conjunto, foram registradas 5.866 UCs regulares e 5.064 UCs irregulares. A identificação de uma UC como irregular é fornecida pela concessionária. As UCs classificadas como regulares foram submetidas a uma análise meticulosa. Tal análise foi fundamental para mitigar possíveis imprecisões na estrutura de decisão dos modelos supervisionados de classificação, visando à minimização da taxa de falsos positivos. Esse procedimento de análise adicionou uma camada de precisão ao processo, garantindo uma abordagem mais confiável na classificação das UCs.

2.2 Modelo de Árvore de Decisão

Uma vez com a base de dados já pré-processada, esta foi aplicada no algoritmo chamado de Árvore de Decisão. De acordo com Géron (2019), tal modelo computacional se caracteriza como um algoritmo poderoso no processamento de grandes bases de dados devido à sua técnica recursiva de classificação.

Em suma, a Árvore de Decisão funciona caracterizando os dados com base em uma série de declarações condicionais e tomando medidas a partir dos resultados dessas declarações. A estrutura da Árvore de Decisão baseia-se na divisão por nós. O dado a ser processado é inserido no nó principal, chamado de raiz. Segundo Cody et al. (2015), a Árvore, então, opera aplicando declarações condicionais. Além disso, dependendo da resposta retornada, o algoritmo se ramifica de forma progressiva, aplicando novas declarações, até que a resposta seja classificada corretamente com base em seu treinamento.

Na Figura 2 é apresentada a estrutura básica da Árvore de Decisão. Neste exemplo, formas geométricas de cores pretas e brancas são os dados de entrada. Diante disso, a Árvore de Decisão busca rotular essas formas de acordo com os parâmetros extraídos em seu treinamento prévio, como explica Filho et al. (2004). Dessa forma, o exemplo evidenciado extraiu algumas categorias de formas geométricas oriundas de seu treinamento e procura classificá-las fundamentado de acordo com essas categorias.

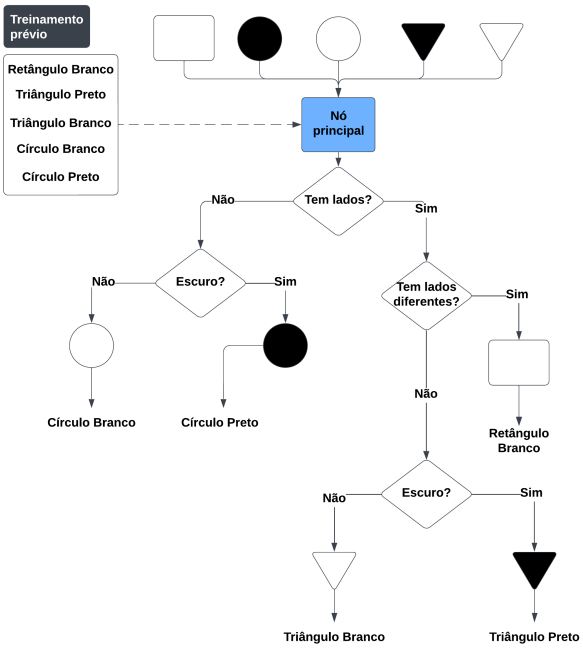

Figura 2. Exemplo de diagrama operacional de Árvore de Decisão - Autoria Própria

Então, para o treinamento da Árvore de Decisão, os dados de consumo foram divididos em dois grupos: "Dados de Treinamento" e "Dados de Validação". O grupo de treinamento foi utilizado para treinar o modelo, enquanto o grupo de validação foi reservado para avaliar seu desem-

penho. A distribuição dos dados está detalhada na Tabela 1.

Tabela 1. Distribuição dos Dados de Consumo

Grupo de Dados	Percentual	Número de UCs
Dados de Treinamento	80%	8.744
Dados de Validação	20%	2.186
Total	100%	10.930

A metodologia utilizada para calcular a pontuação acumulativa com base no desvio percentual (Δ) entre o consumo atual e o consumo médio para cada cliente. Essa pontuação visa identificar clientes com comportamento atípico em relação ao padrão esperado.

Para realizar o cálculo da pontuação, foi utilizado um conjunto de dados contendo informações sobre o consumo de eletricidade de diferentes clientes. Os dados incluem:

- CDC: Identificador único do cliente.
- CONSUMO (kWh): Consumo de energia elétrica em quilowatts-hora.

$$\Delta = \left| \frac{\text{Consumo Atual (kWh)} - \text{Consumo Médio (kWh)}}{\text{Consumo Médio (kWh)}} \right| \times 100$$

A pontuação acumulativa é construída somando os valores de Δ para cada cliente. Se o desvio percentual for superior a um limiar de 30% é atribuída uma pontuação incremental; caso contrário, a pontuação é zero.

O limiar de 30% foi selecionado com base em testes que buscaram equilibrar a sensibilidade e a especificidade da avaliação. Esse valor mostrou-se o mais adequado para identificar desvios significativos no consumo sem gerar pontuações excessivas para variações menores.

3. RESULTADOS PRELIMINARES

Após o treinamento da Árvore de Decisão, utilizou-se o conjunto de dados de validação para avaliar o desempenho do algoritmo. Os resultados dessa avaliação estão apresentados a seguir.

A curva *Receiver Operating Characteristic* (ROC), mostrada na Figura 3, demonstra a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos em vários pontos de corte do modelo. A área sob a curva ROC foi de 1.0, indicando uma capacidade de 100% de discriminação entre exemplos positivos e negativos.

Área Sob a Curva (do inglês, *Area Under Curve* (AUC)) é uma das métricas de classificação por tipo de *ranking* mais populares. Segundo Hossin and Sulaiman (2015), ela é utilizada para construir um modelo de aprendizado otimizado e também para comparar algoritmos de aprendizado. O valor da AUC reflete o desempenho geral de um classificador. Em Hossin and Sulaiman (2015), seu cálculo é feito da seguinte maneira:

$$AUC = \frac{S_p - n_p(n_p + 1)/2}{n_p n_n},$$

em que S_p é a soma de todos os exemplos positivos classificados, enquanto n_p e n_n representam o número de exemplos positivos e negativos, respectivamente.

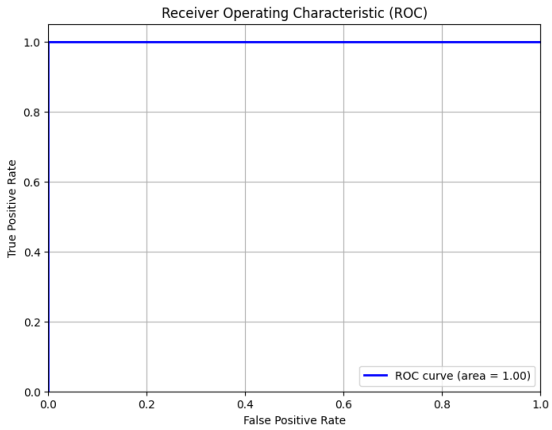

Figura 3. Curva ROC

Hossin and Sulaiman (2015) ressaltam que, embora o desempenho da AUC tenha sido excelente para avaliação e processos de discriminação, o seu custo computacional é alto, especialmente para discriminar um volume de soluções geradas em problemas multiclasse.

Já a curva *Precision-Recall*, ilustrada na Figura 4, destaca a relação entre a precisão e a sensibilidade do modelo em diferentes limiares de classificação. Novamente, a AUC foi de 1.0, que indica um desempenho excepcional na detecção de exemplos positivos, mantendo uma alta precisão.

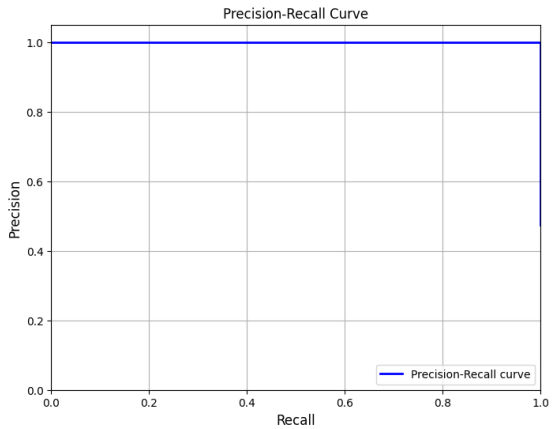

Figura 4. Curva Precision-Recall

Outro instrumento utilizado para medir a exatidão de um modelo de classificação no aprendizado de máquina é a matriz de confusão. Organizada em formato de tabela, essa matriz compila um resumo das previsões acertadas e erradas do modelo, contrastando-as com os dados reais. Nas matrizes de confusão, as linhas indicam as classes previstas pelo modelo, e as colunas, as classes reais observadas. Um exemplo é apresentado na Tabela 2, na qual estão dispostos os dados catalogados no estudo.

Tabela 2. Matriz de Confusão

Parâmetros	Negativos Verdadeiros	Falsos Positivos
Negativos previstos	1151	0
Positivos previstos	0	1035

3.1 Consumo Médio Por Número de Clientes

Para calcular o Desvio Padrão (DP), utilizou-se o consumo médio, conforme mostrado na equação 1. As Figuras 5, 6 e 7 exibem o consumo médio em kWh pelo número do cliente, sendo a Figura 5 uma representação gráfica dos resultados apresentados nas duas microrregiões, os picos apresentados se referênciam as *outliers*.

$$DP = \sqrt{\sum_{n=1}^{N^\circ amostra} \frac{(Amostra - Media)^2}{N^\circ amostras}} \tag{1}$$

Figura 5. Consumo Médio Por Número do Clientes Total

Figura 6. Consumo Médio Por Número do Clientes Microrregião 216

Figura 7. Consumo Médio Por Número do Clientes Microrregião 316

3.2 Boxplot Consumo Médio

A identificação de furtos de energia ocorre em três fases distintas: (1) Construção de um modelo para identificar valores atípicos, utilizando variáveis como corrente, tensão e potência. (2) Conversão dos dados provenientes do medidor inteligente de um cliente específico em uma sequência de dados binários, pro meio de um método de previsão de valores atípicos. (3) Realização da detecção de furtos de energia nos dados binários transformados. Os *boxplots* ilustrados nas Figuras 8, 9 e 10 representam as médias do consumo e a faixa de variação que pode ocorrer com base nas leituras dos clientes. Todos os pontos fora do intervalo de variação são identificados como *outliers*.

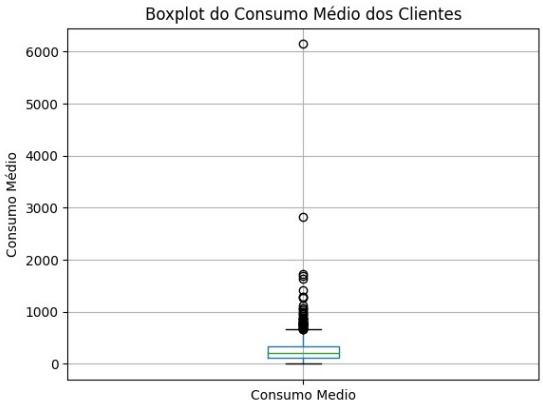

Figura 8. Boxplot do Consumo Médio Total

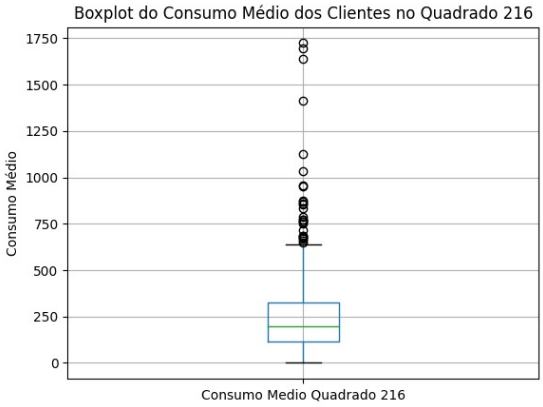

Figura 9. Boxplot do Consumo Médio Microrregião 216

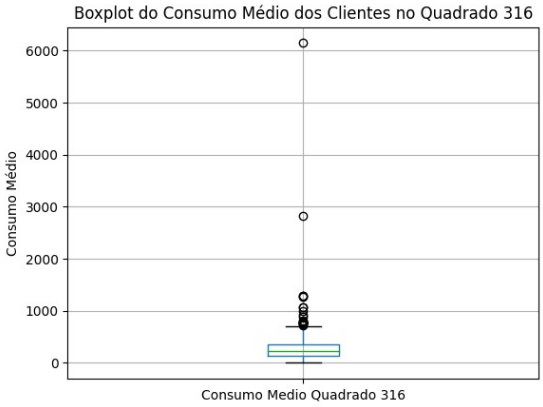

Figura 10. Boxplot do Consumo Médio Microrregião 316

4. CONCLUSÕES E TRABALHOS FUTUROS

O modelo de detecção de furto de energia com base na identificação de anomalias demonstrou um desempenho altamente satisfatório ao comparar os dados de treino e teste. A validação do modelo revelou uma precisão consistente, com todos os indicadores de desempenho alcançando uma acurácia elevada. Isso destaca a capacidade do modelo em classificar corretamente as unidades consumidoras, um aspecto essencial no contexto das perdas não técnicas, especialmente no que se refere ao furto de energia.

Uma estratégia potencialmente eficaz seria levar em consideração a sazonalidade nos padrões de consumo dos clientes residenciais durante o processo de detecção de anomalias. A incorporação desses fatores sazonais pode aprimorar a sensibilidade do modelo às flutuações temporais no consumo de energia, melhorando sua capacidade de identificar comportamentos anômalos. Embora os resultados sejam promissores, existem claras oportunidades para refinamento e desenvolvimento futuro.

Adicionalmente, a inclusão de possíveis falhas no alimentador durante a avaliação das unidades consumidoras pode enriquecer ainda mais o processo de detecção de anomalias.

Em síntese, este estudo representa uma contribuição significativa para o desenvolvimento de soluções avançadas no combate às perdas de energia. As sugestões para futuras pesquisas delineiam um caminho promissor para a continuidade e aprimoramento desta linha de investigação, oferecendo perspectivas valiosas para o avanço do conhecimento e práticas no campo da distribuição de energia elétrica.

AGRADECIMENTOS

Os pesquisadores gostariam de agradecer o Centro de Energias Alternativas e Renováveis da Universidade Federal da Paraíba (UFPB) por fornecer o laboratório para auxílio do estudo e ao colaborador Hyago Fellipe de Oliveira Soares pela contribuição na análise da base de dados.

REFERÊNCIAS

Agência Nacional de Energia Elétrica (ANEEL) (2021). Perdas de energia elétrica na distribuição - edição 1-2021. Technical report, Agência Nacional de Energia Elétrica, Brasil. Dados referentes ao período de janeiro/2008 a dezembro/2020.

Cody, C., Ford, V., and Siraj, A. (2015). Decision tree learning for fraud detection in consumer energy consumption. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, 1175–1179. doi:10.1109/ICMLA.2015.80.

Filho, J., Gontijo, E., Delaiba, A., Mazina, E., Cabral, J., and Pinto, J. (2004). Fraud identification in electricity company customers using decision tree. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, volume 4, 3730–3734 vol.4. doi:10.1109/ICSMC.2004.1400924.

Glauner, P., Meira, J.A., Valtchev, P., State, R., and Bettinger, F. (2017). The challenge of non-technical loss detection using artificial intelligence: A survey. *arXiv*, cs.AI. URL `https://arxiv.org/abs/1606.00626v3`. ArXiv: 1606.00626v3.

Gunturi, S.K. and Sarkar, D. (2021). Ensemble machine learning models for the detection of energy theft. *Electric Power Systems Research*. doi:10.1016/j.epsr.2020.106904. URL `https://www.elsevier.com/locate/epsr`.

Géron, A. (2019). *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn TensorFlow*. Alta Books, Rio de Janeiro.

Himeur, Y., Ghanem, K., Alsalemi, A., Bensaali, F., and Amira, A. (2021). Anomaly detection of energy consumption in buildings: A review, current trends and new perspectives. *arXiv*. URL `https://arxiv.org/abs/2010.04560v5`. ArXiv:2010.04560v5 [cs.CY].

Hossin, M. and Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process*, 5(2), 1–10. doi:10.5121/ijdkp.2015.5201. URL `https://www.researchgate.net/publication/275224157_A_Review_on_Evaluation_Metrics_for_Data_Classification_Evaluations`.

Lacerda, S.L.M., Oliveira, A.C., and Ferreira, A.J.G.A. (2022). Impacto do registro de consumo de energia pela média de consumo na detecção de furto de energia utilizando redes neurais artificiais. In *XXIV Congresso Brasileiro de Automática - CBA 2022*, 3478–3484. Sociedade Brasileira de Automática (SBA), Campina Grande, Paraíba, Brasil. doi:10.20906/CBA2022/3647.

Park, C.H. and Kim, T. (2020). Energy theft detection in advanced metering infrastructure based on anomaly pattern detection. *Energies*, 13(3832). doi:10.3390/en13153832. URL `http://www.mdpi.com/journal/energies`.

Qiu, H., Tu, Y., and Zhang, Y. (2018). Anomaly detection for power consumption patterns in electricity early warning system. In *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*, 867–873.

Stracqualursi, E., Rosato, A., Lorenzo, G.D., Panella, M., and Araneo, R. (2023). Systematic review of energy theft practices and autonomous detection through artificial intelligence methods. *Renewable and Sustainable Energy Reviews*, 184, 113544. doi:10.1016/j.rser.2023.113544. URL `https://www.sciencedirect.com/science/article/pii/S1364032123003341`.

Yip, S.C., Tan, W.N., Tan, C., Gan, M.T., and Wong, K. (2018). An anomaly detection framework for identifying energy theft and defective meters in smart grids. *Electrical Power and Energy Systems*, 101, 189–203. doi:10.1016/j.ijepes.2018.03.025. URL `https://www.sciencedirect.com/science/article/pii/S014206151830025X`.