Universidade Federal do Maranhão
Centro de Ciências Exatas e Tecnologia
Programa de Pós-graduação em Engenharia Elétrica

Mayara Martins Pereira

ÁRVORE DE REGRESSÃO APLICADA À IDENTIFICAÇÃO E PRIORIZAÇÃO DE PERDAS NÃO TÉCNICAS

São Luís
2025

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Martins Pereira, Mayara.
Árvore de Regressão aplicada à identificação e priorização de perdas não técnicas / Mayara Martins Pereira. - 2025.
104 f.

Orientador(a): Shigeaki Leite de Lima.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia Elétrica/ccet, Universidade Federal do Maranhão, São Luís, 2025.

1. Perdas Comerciais. 2. Perdas Não Técnicas. 3. Árvore de Regressão. 4. Aprendizado de Máquina. 5. Distribuição de Energia Elétrica. I. Leite de Lima, Shigeaki. II. Título.

Mayara Martins Pereira

ÁRVORE DE REGRESSÃO APLICADA À IDENTIFICAÇÃO E PRIORIZAÇÃO DE PERDAS NÃO TÉCNICAS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da UFMA, como requisito para a obtenção do grau de MESTRE em Engenharia Elétrica.

Shigeaki Leite de Lima, Dr.

(Orientador)

São Luís

2025

Pereira, Mayara Martins Pereira

ÁRVORE DE REGRESSÃO APLICADA À IDENTIFICAÇÃO E PRIORIZAÇÃO DE PERDAS NÃO TÉCNICAS / Mayara Martins Pereira Pereira - 2025

94f.

Orientador: Shigeaki Leite de Lima

Impresso por computador (fotocópia)

Dissertação (Graduação em Engenharia Elétrica) - Universidade Federal do Maranhão, curso de graduação em Engenharia Elétrica, São Luís, 2025.

1.Sistemas de Potência. 2. Custos 3.Confiabilidade. I. Lima, Shigeaki Leite de ,orientador. II. Título.

CDU 621.314.1

Mayara Martins Pereira

ÁRVORE DE REGRESSÃO APLICADA À IDENTIFICAÇÃO E PRIORIZAÇÃO DE PERDAS NÃO TÉCNICAS

Dissertação apresentada ao Programa de Pós-graduação em Engenharia Elétrica da UFMA, como requisito para a obtenção do grau de MESTRE em Engenharia Elétrica.

Aprovado em 02 de outubro de 2025

BANCA EXAMINADORA

Shigeaki Leite de Lima, Dr.

(Orientador)

André Nunes de Souza, Dr.

(Membro da Banca Examinadora)

Silvangela Lilian da Silva Lima Barcelos, Dr.

(Membro da Banca Examinadora)

Dedico este trabalho à minha família,
grande razão e motivação de tudo.

Resumo

Este trabalho apresenta uma metodologia de priorização de alvos para inspeção de perdas não técnicas em sistemas de distribuição de energia elétrica, fundamentada em modelagem preditiva por árvore de regressão. O objetivo central é identificar instalações com maior propensão a fraudes e irregularidades no consumo, de modo a apoiar ações de normalização e reduzir os impactos técnicos, econômicos e sociais dessas perdas. Para tanto, foi realizada uma revisão bibliográfica e bibliométrica sobre métodos de aprendizado de máquina aplicados ao combate às perdas comerciais, bem como um estudo aprofundado da técnica de árvore de regressão e de sua aplicabilidade neste contexto. O modelo proposto foi desenvolvido com dados reais de uma distribuidora, abrangendo informações de consumo, características cadastrais e registros operacionais, e incluiu etapas de preparação da base de dados, seleção de variáveis preditoras, construção e poda da árvore de regressão por complexidade de custo, além da avaliação do desempenho preditivo. Os resultados obtidos evidenciam a efetividade do modelo na priorização de unidades consumidoras com maior relevância para as perdas não técnicas, demonstrando seu potencial como ferramenta de apoio à tomada de decisão.

Palavras-chave: Perdas Comerciais; Perdas Não Técnicas; Árvore de Regressão; Aprendizado de Máquina; Distribuição de Energia Elétrica.

Abstract

This dissertation presents a target prioritization methodology for the inspection of non-technical losses (commercial losses) in electricity distribution systems, based on predictive modeling with regression trees. The main goal is to identify consumer units with higher propensity to fraud and consumption irregularities, in order to support normalization actions and mitigate the technical, economic, and social impacts of such losses. A bibliographic and bibliometric review of machine learning methods applied to non-technical loss detection was conducted, along with an in-depth study of regression tree techniques and their applicability in this domain. The proposed model was developed using real data from a distribution company, including consumption history, consumer attributes, and operational records, and comprised database preparation, selection of predictive variables, construction and cost-complexity pruning of regression trees, and predictive performance evaluation. Results highlight the effectiveness of the model in prioritizing consumer units with the greatest impact on non-technical losses, reinforcing its potential as a decision-support tool in the electricity sector.

Keywords: Commercial losses; Non-technical losses; Regression trees; Machine learning; Electricity distribution.

Agradecimentos

Neste espaço, venho agradecer a todos que fizeram este trabalho possível e que estiveram ao meu lado neste percurso acadêmico.

A Deus, a Força celestial que guia nossas vidas e faz acontecer coisas inimagináveis.

A minha família e amigos que sempre se fazem presentes, com amor e apoio incondicional em todos os momentos.

Aos meus queridos amigos e mestres do IEE, por todo companheirismo, incentivo e partilha durante estes anos de trabalho.

Aos meus colegas de profissão, pela troca de conhecimento e vivências que se fazem transparecer neste trabalho e para além dele.

"A felicidade só é real quando compartilhada...".

Christopher McCandless

SUMÁRIO

LISTA DE FIGURAS **8**

LISTA DE TABELAS **10**

1 INTRODUÇÃO **13**

1.1 Objetivos 15

1.2 Estrutura do trabalho 16

2 REVISÃO BIBLIOGRÁFICA **18**

2.1 Bibliometria 18

2.2 Estado da arte 22

3 REFERENCIAL TEÓRICO **26**

3.1 Perdas no sistema de distribuição 26

3.1.1 Perdas Técnicas e Não Técnicas 29

3.1.2 Perdas Não Técnicas no Brasil 34

3.1.3 Influências de Aspectos Socioeconômicos nas Perdas Não Técnicas 40

3.1.4 Classificação de Perdas Não Técnicas 47

3.2 Aprendizado de máquina com árvore de regressão 50

3.2.1 Estrutura da Árvore de Regressão 52
3.2.2 Compromisso Viés-Variância em Árvores de Regressão . . . 56
3.2.3 Decomposição do Erro Esperado 57
3.2.4 Estratégias para Balanceamento do Compromisso Viés-Variância 59
3.2.5 *Cost-Complexity Pruning* 61
3.2.6 Determinação do Erro de Previsão do Modelo 63

4 ALGORITMO APLICADO A IDENTIFICAÇÃO E PRIORIZAÇÃO DE PERDAS NÃO TÉCNICAS 65

4.1 Preparação e qualificação da base de dados 67
4.2 Extração de variáveis preditoras relevantes por unidade consumidora 72
4.3 Construção e avaliação do modelo de regressão por árvore de decisão 76
4.4 Modelo de Previsão 78

5 RESULTADOS 82

6 CONCLUSÃO 89

6.1 Trabalhos Futuros 89
6.2 Publicações 90

REFERÊNCIAS 92

LISTA DE FIGURAS

2.1 Distribuição da amostra por tipo de publicação no período 2015–2025. 19

2.2 Evolução anual das publicações sobre *Non-Technical Losses* (2015–2025). 20

2.3 20 maiores produções por país no período 2015–2025. 21

3.1 Variação das perdas no Brasil entre 2008 e 2023. 27

3.2 Perdas sobre a energia injetada por região do Brasil. 28

3.3 Metodologia de cálculo utilizado para obtenção dos percentuais de perdas não técnicas. 31

3.4 Comparação entre Perdas Não Técnicas Reais e Regulatórias. . . . 36

3.5 *Participação das PNT reais do país por concessionárias* 37

3.6 PNT na Região Norte-Nordeste 38

3.7 *PNT real e regulatório do mercado de baixa tensão* 39

3.8 Participação das perdas não técnicas nas tarifas residenciais 41

3.9 *Ranking* de Distribuidores de Grande Escala 43

3.10 *Modelo de uma árvore CART* 54

3.11 Visualização do processo de poda por meio da função *prune* no MATLAB 62

4.1 Fluxograma de Preparação e Qualificação da Base de Dados 71

4.2 Árvore completa antes da aplicação da poda por complexidade de custo 78
4.3 Erro de validação (MSE) por nível de poda (Cost-Complexity Pruning) 79
4.4 Arvore após a poda (Cost-Complexity Pruning) 81

5.1 Dispersão entre SCORE real e SCORE predito pelo modelo podado 83
5.2 Distribuição dos resíduos (SCORE Real – Predito) 84
5.3 Estrutura visual da árvore de regressão com os primeiros nós de decisão 85
5.4 Exibição do Matlab da lista priorizada 87

LISTA DE TABELAS

3.1 Correlação entre variáveis socioeconômicas e Perdas Não Técnicas (PNT)0 46

3.2 Classificação das perdas não técnicas de energia elétrica segundo a origem 48

4.1 Descrição das Variáveis por Classe 68

4.2 Código das Classes de Consumo 73

4.3 Exemplo de Dados por Instalação (Variáveis preditoras) 75

5.1 Desempenho Preditivo do Modelo de Regressão Podado 82

5.2 Comparação de desempenho preditivo entre métodos na literatura e o modelo proposto, tomando como referência o MSE e MAE . . . 82

5.3 Importância relativa das variáveis no modelo de árvore de regressão após poda. 86

5.4 Distribuição dos Alvos priorizados por Classe e Tipo de Irregularidade 87

LISTA DE ABREVIATURAS

ABRACE	*Associação Brasileira de Grandes Consumidores de Energia e Consumidores Livres*
AM	*Amazonas (Unidade Federativa)*
ANEEL	*Agência Nacional de Energia Elétrica*
AP	*Amapá (Unidade Federativa)*
AUC	*Area Under the Curve*
BT	*Baixa Tensão*
CART	*Classification and Regression Trees*
CCEE	*Câmara de Comercialização de Energia Elétrica*
COVID-19	*Coronavirus Disease 2019*
ERGEG	*European Regulators Group for Electricity and Gas*
GD	*Geração Distribuída*
GWh	*Gigawatt-hora*
ID3	*Iterative Dichotomiser 3*
IEA	*International Energy Agency*
IEEE	*Institute of Electrical and Electronics Engineers*
LDCs	*Least Developed Countries*
MAE	*Mean Absolute Error (Erro Absoluto Médio)*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Squared Error (Erro Quadrático Médio)*
NTL	*Non-Technical Losses*
OPF	*Optimum-Path Forest*

PB	*Paraíba (Unidade Federativa)*
PNT	*Perdas Não Técnicas*
PRODIST	*Procedimentos de Distribuição de Energia Elétrica*
PT	*Perdas Técnicas*
REN	*Resolução Normativa*
RMSE	*Root Mean Squared Error*
RN	*Rio Grande do Norte (Unidade Federativa)*
SAMP	*Sistema de Acompanhamento de Informações de Mercado para Regulação Econômica*
STD	*Superintendência de Regulação dos Serviços de Transmissão e Distribuição de Energia Elétrica*
SVM	*Support Vector Machines*
TWh	*Terawatt-hora*
UC	*Unidade Consumidora*
UNCTAD	*United Nations Conference on Trade and Development*
US$	*United States Dollar*

1 INTRODUÇÃO

O setor de energia elétrica, que é a base para indústria, comércio e serviços, vem se remodelando há alguns anos, com especial foco em energias renováveis, descentralização e virtualização dos processos. O aumento na demanda residencial, crescimento exponencial do uso de fontes renováveis na rede, usuários buscando soluções de pagamento, leitura e comunicação on-line desafiam as empresas geradoras, transmissoras e distribuidoras de energia elétrica a suprir e avançar no desenvolvimento de soluções eficientes, seguras e sustentáveis.

Nesse contexto, as concessionárias têm realizado investimentos com o objetivo de ampliar significativamente seu desempenho financeiro e operacional, buscando aumento na produtividade, eficiência e lucratividade. Dessa forma, a gestão de perdas de energia elétrica surge como um dos maiores desafios do setor, uma vez que elas acarretam custo econômico indesejável às empresas e aos consumidores regulares, que são impactados parcialmente por meio das tarifas.

As perdas comerciais refletem nas tarifas aplicadas às contas de energia dos consumidores finais. Apesar desse repasse tarifário ser limitado regulatoriamente pela ANEEL, em ciclos de 3 a 5 anos em cada distribuidora, de modo que os impactos das perdas reais não sejam totalmente repassados aos consumidores regulares e sejam analisadas especificidades, tais como características do mercado e variáveis socioeconômicas da região, o consumidor regular arca de forma parcial pelas irregularidades de outros consumidores, tendo em vista que a ANEEL reconhece valores regulatórios eficazes [1].

O problema é comumente abordado pelas distribuidoras através de campanhas de prevenção e da realização de inspeções técnicas nas unidades consumidoras. Entretanto, além da inspeção de todos os consumidores atendidos ser inviável, os métodos de inspeção tradicionais utilizados se mostram, em geral, ineficientes e tem alto custo de execução, o que motiva o estudo de novas abordagens de combate às perdas comerciais [2].

As inspeções são realizadas *in loco* e por técnicos especializados. A execução é comumente guiada pela seleção de clientes considerados suspeitos ou por meio de técnica de varredura, onde uma área específica é selecionada e percorrida por uma equipe de inspeção com o objetivo de identificar possíveis perdas não técnicas[3].

A partir da necessidade de modernização e alto desempenho nos processos de distribuição de energia elétrica, diversas tecnologias vêm sendo desenvolvidas para realizar a verificação e controle de perdas de modo eficaz e assertivo, identificando alvos em potencial, otimizando assim, a regularização desses alvos com o direcionamento correto de receita.

Os sistemas de tecnologia de informação surgem com o objetivo de auxiliar as concessionárias a realizar uma fiscalização eficiente e simultânea no consumo de energia elétrica, reduzindo as perdas em geral e melhorando os indicadores de continuidade de energia por meio de uma rápida detecção e intervenção de falhas no sistema.

O aprendizado de máquina é um campo de estudo da inteligência artificial que tem como finalidade o desenvolvimento de técnicas computacionais de aprendizado e formação de sistemas com a capacidade de obter conhecimento de maneira automática, isto é, uma ferramenta computacional que é capaz de decidir, baseada em suas experiências adquiridas, uma solução ótima para o problema

proposto [4].

Uma das formas utilizadas para realizar essa identificação de perdas, na gestão dos dados do sistema, são os algoritmos de aprendizado de máquina, tendo em vista o processamento de grandes volumes de dados e a alta confiabilidade na tomada de decisão, através da identificação de um padrão entre os dados dos clientes que indiquem uma possível irregularidade.

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver e aplicar uma metodologia de priorização de alvos com base em modelagem preditiva por árvore de regressão, visando identificar instalações com maior propensão à ocorrência de perdas não técnicas, como fraudes e irregularidades no consumo de energia elétrica. Já os objetivos específicos são:

- Realizar um levantamento abrangente do cenário das perdas de energia elétrica no sistema de distribuição, com ênfase nas perdas não técnicas, suas origens, classificações e impactos técnicos, econômicos e sociais.
- Elaborar uma revisão bibliográfica dos métodos de aprendizado de máquina utilizados no combate a perdas comerciais em redes de distribuição;
- Realizar estudo da técnica de árvore de regressão e sua aplicabilidade na detecção e priorização de perdas comerciais no sistema de distribuição de energia elétrica;
- Desenvolver e avaliar o algoritmo de árvore de regressão com foco na modelagem preditiva de comportamentos atípicos de consumo, priorização de

unidades consumidoras com indícios de perdas comerciais e apoio à tomada de decisão nas ações de normalização.

1.2 Estrutura do trabalho

Neste trabalho, é proposta uma abordagem baseada em aprendizado de máquina supervisionado, utilizando a técnica de árvore de regressão, para auxiliar na identificação e priorização de perdas comerciais no sistema de distribuição de energia elétrica. O modelo desenvolvido utiliza dados reais de uma distribuidora de energia, incorporando informações de histórico de consumo, características cadastrais e técnicas, além de registros operacionais para construção de uma base robusta.

O trabalho está estruturado em quatro capítulos. O primeiro capítulo é apresentado o estado da arte sobre os principais métodos de detecção de perdas comerciais aplicados no Brasil e no exterior, com base em uma pesquisa bibliográfica e bibliométrica realizada entre os anos de 2015 a 2025, destacando metodologias, limitações e oportunidades de aprimoramento.

No Capítulo 2 são explorados os conceitos fundamentais relacionados às perdas no sistema de distribuição, com ênfase nas perdas comerciais, sua classificação e a influência de fatores socioeconômicos. Além disso, é apresentada uma introdução detalhada à técnica de árvore de regressão, abordando suas estruturas, princípios de funcionamento e aplicações em problemas de regressão.

O Capítulo 3 são descritas as etapas práticas da metodologia proposta, iniciando-se com a preparação e qualificação da base de dados, seleção de atributos preditivos, construção do modelo por árvore de regressão com poda por complexidade de custo *cost-complexity pruning* e avaliação do erro de previsão.

No Capítulo 4 são apresentados os resultados obtidos com a aplicação de dados reais da distribuidora no modelo por árvore de regressão com poda por complexidade de custo, incluindo análises estatísticas, desempenho preditivo e análise da efetividade da árvore de regressão na priorização de unidades com maior impacto nas perdas comerciais.

Por fim, nas conclusões são apresentadas as principais contribuições do trabalho, dificuldades encontradas, e sugestões para pesquisas futuras, reforçando o potencial da técnica estudada como ferramenta de apoio à tomada de decisão no combate às perdas comerciais no setor elétrico.

2 REVISÃO BIBLIOGRÁFICA

Com o objetivo de filtrar informações pertinentes ao desenvolvimento deste trabalho, foi realizada uma revisão bibliográfica acerca do estado da arte, com foco em técnicas de aprendizado de máquina e métodos estatísticos aplicados à detecção de perdas não técnicas em sistemas de distribuição de energia elétrica.

As informações foram obtidas por meio de consulta à base de dados do *Institute of Electrical and Electronics Engineers* (IEEE), bem como a outras bases científicas relevantes (Scopus, Web of Science e ScienceDirect), considerando uma amostragem no período de 2015 a 2025.

A busca utilizou como principal palavra-chave o termo *Non-Technical Losses*, resultando em um total de 1.356 publicações, distribuídas da seguinte forma: 892 artigos em periódicos indexados, 412 trabalhos apresentados em congressos internacionais, 32 capítulos de livros e 20 publicações em outros formatos (relatórios e papers).

Essa amostra permitiu identificar tendências metodológicas, com destaque para o crescimento do uso de algoritmos de aprendizado supervisionado (como árvores de decisão, *Random Forest*, SVM e redes neurais) e a incorporação de variáveis socioeconômicas e contextuais como diferenciais nos estudos mais recentes.

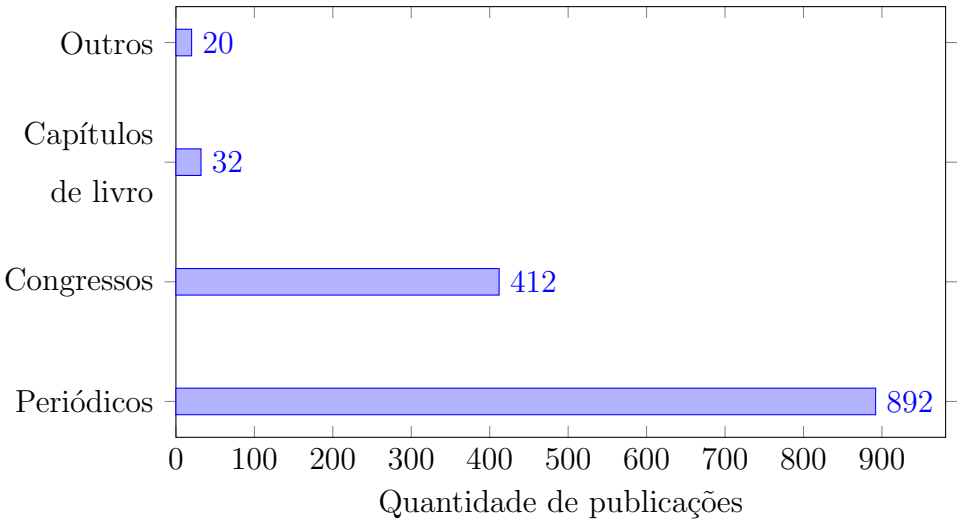
2.1 Bibliometria

Com a finalidade de compreender o panorama das pesquisas sobre perdas não técnicas no setor elétrico, foi realizada uma análise bibliométrica no período de

2015 a 2025. A busca concentrou-se no termo *Non-Technical Losses*, contemplando publicações indexadas nas principais bases científicas, como IEEE Xplore, Scopus, Web of Science e ScienceDirect.

O levantamento resultou em um total de 1.356 publicações, distribuídas entre diferentes tipos de veículos: 892 artigos em periódicos científicos (65,8%), 412 trabalhos apresentados em congressos (30,4%), 32 capítulos de livros (2,4%) e 20 publicações em outros formatos (1,5%). A figura 2.1 ilustra a distribuição da amostra por tipo de publicação.

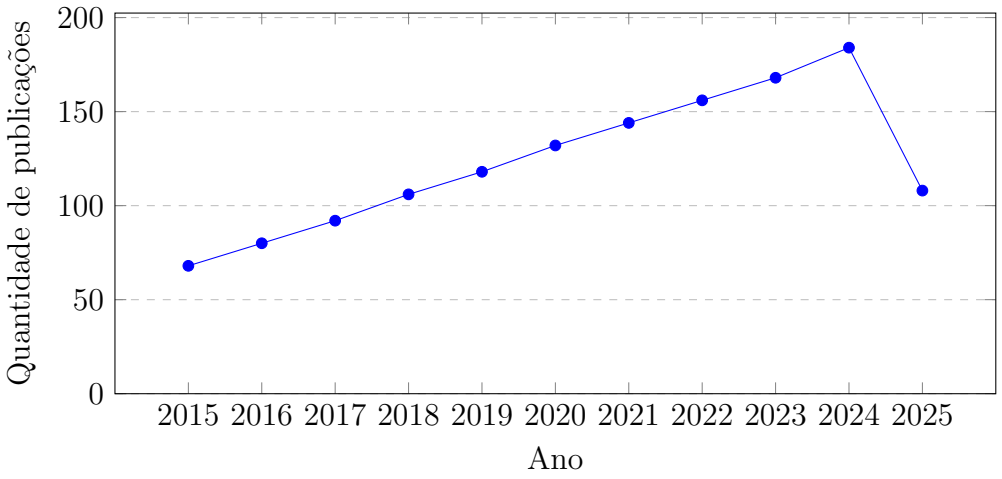
Figura 2.1: Distribuição da amostra por tipo de publicação no período 2015–2025.



Fonte: Elaborado pela autora (2025).

Observa-se crescimento consistente a partir de 2018, impulsionado pela adoção de técnicas de aprendizado de máquina e pelo aumento da disponibilidade de dados de medição inteligente e sistemas comerciais das distribuidoras. A figura 2.2 mostra a evolução anual de publicações no período de 2015 a 2025.

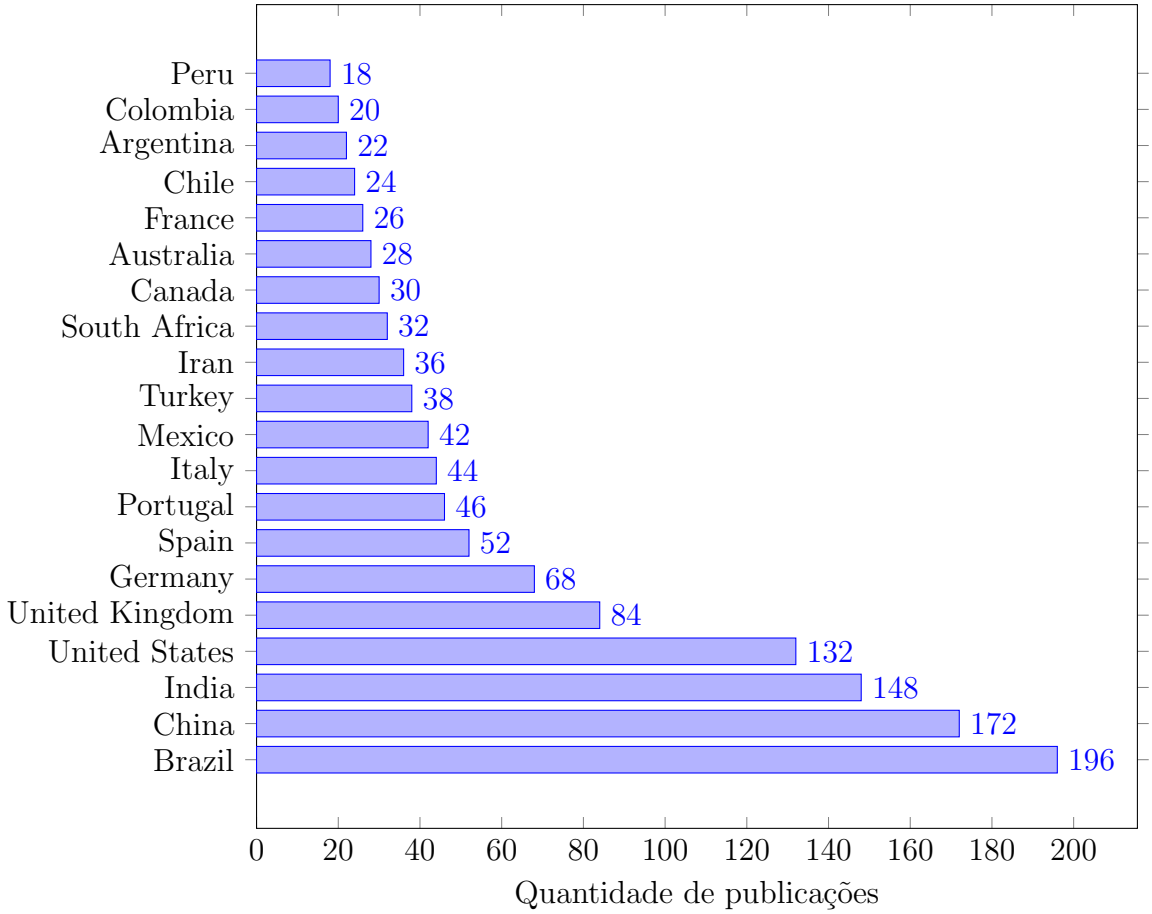
Figura 2.2: Evolução anual das publicações sobre *Non-Technical Losses* (2015–2025).



Fonte: Elaborado pela autora (2025).

Adicionalmente, a análise por país (figura 2.3) evidencia a concentração de publicações nos países Brasil, China, Índia, Estados Unidos e Reino Unido, seguida por um conjunto de países com produção relevante (Alemanha, Espanha, Portugal, Itália, México, Turquia, Irã, África do Sul, Canadá, Austrália, França e países da América do Sul). Essa distribuição sugere a presença de linhas de pesquisas consolidadas em centros acadêmicos e grupos industriais com forte interação com concessionárias.

Figura 2.3: 20 maiores produções por país no período 2015–2025.



Fonte: Elaborado pela autora (2025).

Em síntese, a análise bibliométrica confirma alguns pontos: a predominância de estudos em periódicos de alto impacto, o crescimento contínuo do tema; a diversificação metodológica com ênfase em abordagens supervisionadas (árvores de decisão, *Random Forest*, SVM, redes neurais e métodos de *boosting*) e a incorporação de variáveis contextuais (socioeconômicas e geoespaciais), que têm elevado a capacidade preditiva para a priorização de alvos de perdas.

2.2 Estado da arte

No trabalho desenvolvido por [5] foi utilizada a técnica *Support Vector Machines* (SVM) objetivando a pré-seleção de usuários para serem inspecionados baseados nas irregularidades que apresentaram no comportamento do consumo, com a finalidade de identificar as perdas comerciais do sistema. Uma das vantagens do SVM é a sua resistência com relação ao problema de desequilíbrio de classes. O algoritmo SVM constrói uma função de decisão ótima, de maneira que preveja com precisão dados não identificados em duas classes e minimizando o erro da classificação. Ele encontra um hiperplano entre duas classes de dados, através da atribuição de diferentes pesos a vários tipos de erros de classificação.

As SVMs podem ser consideradas eficazes em casos de identificação de perdas não técnicas, entretanto o estudo ressalta que pode apresentar desempenho difícil e com tempo prolongado quanto se trata de ajustes, aumentando assim o tempo de construção do modelo para os casos com um conjunto de dados de grande porte. Essa característica torna este método uma solução ineficiente para aplicações em tempo real.

Em [6] propõe-se reconhecer o comportamento de roubo de eletricidade por meio de dados de várias fontes. Além do uso de eletricidade pelos usuários registrados, analisa-se o comportamento do usuário por meio de fatores regionais (perda não técnica) e fatores climáticos (temperatura) nas correspondentes áreas do transformador.

Ao conduzir experimentos analíticos, foram descobertos vários padrões interessantes: por exemplo, consumidores ilícitos, com furto de eletricidade, tendem a consumir mais energia elétrica do que usuários normais, especialmente sob temperaturas extremamente altas ou baixas. Motivados por essas observações empíricas, foi projetado ainda uma nova estrutura hierárquica para identificar furtos.

Resultados experimentais baseados em um conjunto de dados do mundo real demonstram que o modelo proposto pode atingir o melhor desempenho em detecção de roubo de eletricidade em comparação com outras metodologias. O trabalho foi aplicado pela State Grid da China e usado para capturar irregularidades em Hangzhou com uma precisão de 15% durante a investigação mensal no local. O conjunto de dados compreende três partes: Dois conjuntos de dados relacionados aos registros de eletricidade fornecidos pela State Grid Zhejiang Power Supply Co. Ltd.4, e os registros de temperatura coletados na rede de internet local [6].

O método empregado por [7] consiste na utilização da técnica Árvore de Decisão (*Decision Tree*) como ferramenta de mineração de dados para detecção de fraudes e defeitos nos medidores. Foram selecionados atributos cadastrais, de faturamento e inspeções fornecidos por uma concessionária de energia elétrica referentes aos consumidores de baixa tensão para analisar potenciais fraudes e definir uma lista de alvos para inspeção.

A avaliação detalhada utilizando a classificação de cada atributo e ainda a variância entre padrões contínuos e discretos são o diferencial desta metodologia, pois permitem gerar diversas situações de classificação que levam a categorização de uma instalação com fortes indícios de fraudes, apresentando diferentes estimativas de taxa de acerto de detecção.

Na abordagem feita por [8] é possível detectar perdas comerciais através do uso de algoritmos baseados em *boosting*. Utilizando o aprendizado por meio de árvore de decisão, associada à técnicas de regressão, em conjunto com um algoritmo denominado Adaboost, possibilitou a identificação de anomalias no consumo dos clientes.

A análise de dados específicos relacionados ao consumo de energia elétrica de cada instalação ocorreu com o objetivo de dividi-los em grupos (clusters)

de acordo com as curvas registradas e aplicar o método estatístico RMSD nos dados para atribuir a classificação em clientes normais e anormais. Apesar do Adaboost apresentar um alto desempenho de classificação e estar entre as melhores técnicas utilizadas para este fim, o resultado obtido se comparado à outras metodologias para a identificação de anomalias foi inferior.

Em [9] foi abordado o comportamento de três técnicas de classificação para detecção de perdas comerciais com atuação sobre bancos de dados de diferentes proporções entre consumidores lícitos e ilícitos. Foram analisados grandes volumes de dados (registro de consumo de mais 100 mil clientes, em quatro anos de registro) por meio das seguintes técnicas: regras booleanas, lógica fuzzy e Máquinas de Vetores de Suporte (SVM). Utilizando para comparação de desempenho a métrica AUC, que demonstrou que o classificador baseado em SVM obteve os melhores resultados.

Em [10] também foram realizadas três metodologias de classificação para sua análise: *Random Forest*, Regressão Logística e Máquinas de Vetores de Suporte, utilizando como métrica comparativa a AUC. O estudo elegeu diversos atributos caracterizados por padrões de localização, similaridade e infraestrutura e definiu-os como entrada para os algoritmos de classificação. Assim, foi possível concluir que os dados selecionados referentes ao consumo dos clientes são suficientes para uma classificação assertiva entre os consumidores regulares e irregulares, apresentando desempenho superior em relação a resultados obtidos de parâmetros designados pelas distribuidoras.

Neste trabalho [11] propõe um sistema de combate a perdas não técnicas utilizando o classificador Floresta de Caminhos Ótimos (OPF). Nele são selecionados dados categóricos das unidades consumidoras, realizando uma normalização desses dados com modificações em relação a métodos encontrados na literatura.

Os testes são desenvolvidos com base em dados de clientes residenciais, diferentemente de outros trabalhos que utilizaram dados de consumidores comerciais e industriais. O método obteve resultados satisfatórios a partir das modificações propostas, otimizando o desempenho do OPF.

Mais recentemente, [12] analisaram a influência de fatores socioeconômicos na ocorrência de perdas não técnicas, integrando variáveis externas — como renda média, densidade populacional, índice de vulnerabilidade social e taxa de inadimplência — aos históricos de consumo. Os autores aplicaram modelos supervisionados, em especial *Random Forest* e Regressão Logística, para avaliar o ganho preditivo dessas informações adicionais. Os resultados indicaram que a inclusão de variáveis contextuais elevou significativamente os índices de acurácia e a métrica AUC, reforçando a relevância de aspectos socioeconômicos na priorização de inspeções e no planejamento estratégico de fiscalização das distribuidoras.

Em complemento, [13] exploraram o uso de árvores de regressão com poda por complexidade de custo (*cost-complexity pruning*) como técnica para reduzir o sobreajuste e aumentar a capacidade de generalização dos modelos aplicados à detecção de perdas comerciais. O estudo comparou a árvore podada com algoritmos como *Support Vector Machines* e *Gradient Boosting*, demonstrando que a abordagem de regressão apresentou desempenho competitivo, além de maior interpretabilidade e facilidade de implementação em ambientes corporativos. A pesquisa obteve métricas robustas de desempenho, incluindo valores de RMSE e MAPE reduzidos em relação à árvore não podada, evidenciando o potencial da técnica como ferramenta prática e transparente para concessionárias no combate às perdas não técnicas.

3 REFERENCIAL TEÓRICO

A energia elétrica gerada nas usinas é transportada inicialmente por linhas de transmissão, que operam em alta tensão, até as subestações de distribuição. Essas subestações desempenham o papel de reduzir a tensão para níveis adequados antes de transferir a energia para as linhas de distribuição. Essas linhas, por sua vez, direcionam a energia em baixa ou média tensão até os grandes centros de consumo [18].

Em um cenário ideal, toda a energia elétrica gerada seria igual à energia consumida. Contudo, na prática, essa equivalência não é alcançada pois o processo de transmissão e distribuição inevitavelmente gera perdas ao longo do sistema [19].

3.1 Perdas no sistema de distribuição

As perdas de energia correspondem à parcela da energia elétrica gerada que transita pelas linhas de transmissão e redes de distribuição, mas que não é efetivamente comercializada, seja por motivos técnicos ou comerciais [20].

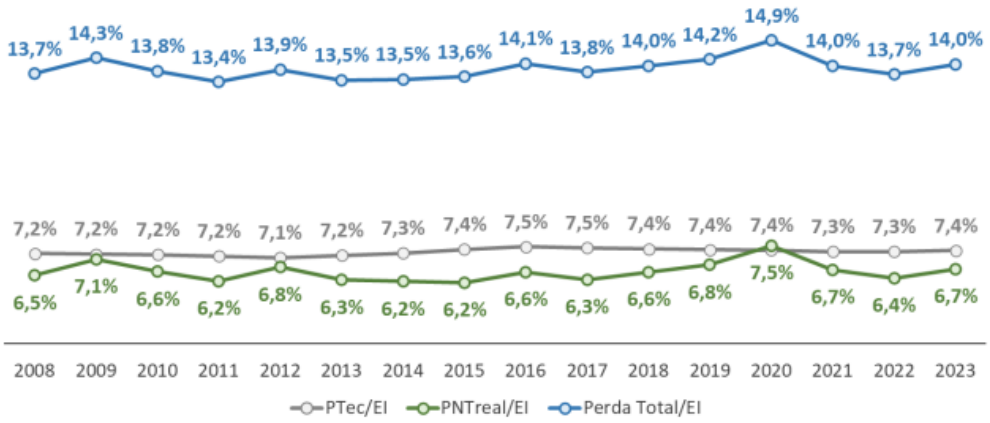
Essas perdas são classificadas em dois tipos principais: perdas técnicas e perdas não técnicas, cuja soma representa as perdas totais no sistema de distribuição. A diferença entre a energia adquirida pelas distribuidoras e a faturada aos consumidores indica as perdas no sistema de distribuição. Já as perdas totais na distribuição, são resultado da soma das perdas técnicas e não técnicas [20]. A média mundial de perdas na distribuição de energia elétrica gira em torno de 8% a 15% . Conforme apontado pelo Grupo Europeu de Reguladores de Eletricidade

e Gás (ERGEG), as definições regulatórias do termo ”perdas de energia”variam consideravelmente entre os países. Em particular, no caso das perdas não técnicas, a falta de uniformidade nas definições dificulta a comparação de porcentagens de perdas entre diferentes nações [21].

No entanto, estima-se que as perdas globais não técnicas possam ascender a 80–100 bilhões de dólares por ano. Para mitigar as perdas de energia, diversas ações podem ser implementadas. A modernização da infraestrutura, com investimentos em tecnologias mais eficientes, como cabos de alta tensão e sistemas que reduzem a resistência elétrica, é fundamental. Outra medida estratégica é a adoção de smart grids, redes inteligentes que permitem monitorar e gerenciar a distribuição de energia com maior eficiência [22].

É ilustrado na figura 3.1 a variação das perdas no Brasil entre o período de 2008 e 2023.

Figura 3.1: Variação das perdas no Brasil entre 2008 e 2023.

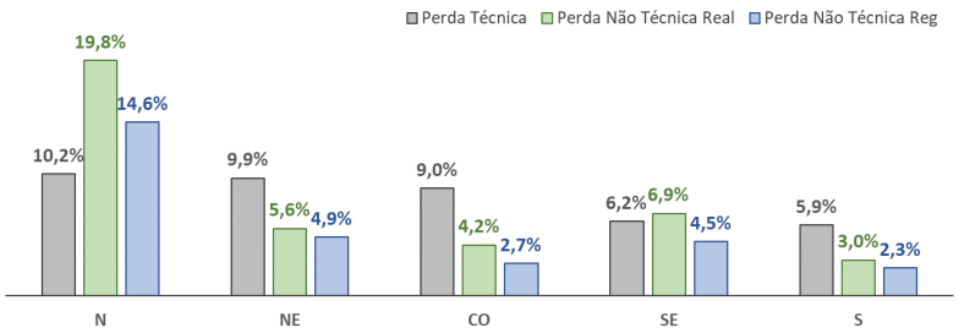


Fonte: [20].

As perdas de energia elétrica no Brasil permaneceram constantes entre 2008 e 2023, tendo um valor mínimo de 13,4 % em 2011 e máximo de 14,9 % em

2020, ano da pandemia da COVID-19. Entretanto, essas perdas não são distribuídas de forma uniforme entre as regiões do país. A figura 3.2 apresenta a variação das perdas entre as regiões do país.

Figura 3.2: Perdas sobre a energia injetada por região do Brasil.



Fonte: [20].

Percebe-se que região Norte do país possui os maiores valores de perdas totais, cerca de 30%, sendo 10,2 % de PT e 19,8 % de PNT. As perdas não técnicas regulamentadas pela ANEEL para a região Norte em 2023 foram de 14,6 %, mas essa meta não foi atingida por essa região e nem pelas outras.

Já a região Nordeste, que é a segunda região com os maiores índices de perdas, apresentou cerca de 15,5 % de perdas totais, sendo 9,9 % de PT e 5,6 % de PNT, dessa forma, tambem ultrapassando as perdas regulatórias estabelecidas que eram cerca de 4,9 %. Os valores regulatórios das PNT são aqueles reconhecidos na tarifa de energia, enquanto os valores reais são os que efetivamente ocorrem.

3.1.1 Perdas Técnicas e Não Técnicas

As perdas técnicas são as partes não faturadas de energia elétrica, inerentes ao processo de distribuição. Essas perdas de energia são causadas, em carga e em vazio, devido à passagem da corrente elétrica nos diversos elementos que compõe a rede de distribuição, tais como cabos e transformadores, dissipando parte da energia em forma de calor [20].

O sistema de distribuição de energia elétrica é segmentado em diferentes componentes, como redes de alta, média e baixa tensão, transformadores, ramais de ligação e medidores. Para cada um desses segmentos, são aplicados modelos específicos que utilizam informações simplificadas das redes e equipamentos existentes, como o comprimento e a bitola dos condutores, a potência dos transformadores e a energia fornecida às unidades consumidoras.

Com base nesses dados, é possível estimar o percentual de perdas técnicas eficientes em relação à energia injetada, que corresponde à energia elétrica introduzida na rede de distribuição para atender aos consumidores, incluindo as perdas no processo.[20].

Essas perdas são calculadas mensalmente pela Câmara de Comercialização de Energia Elétrica (CCEE), e seu custo, estabelecido anualmente nos processos tarifários, é dividido igualmente entre a geração e os consumidores, com cada parte arcando com 50%. O cálculo detalhado das perdas técnicas regulatórias está descrito no Módulo 7 dos Procedimentos de Distribuição de Energia Elétrica (PRODIST) e é realizado pela Superintendência de Regulação dos Serviços de Transmissão e Distribuição de Energia Elétrica (STD) [20].

Os níveis de perdas técnicas tambem podem ser calculados e controlados, ainda que de forma estimada, utilizando parâmetros estocásticos. Essas perdas

podem ser determinadas tanto por meio de simulações computacionais quanto por medições diretas ou avaliações específicas do sistema elétrico em análise [23].

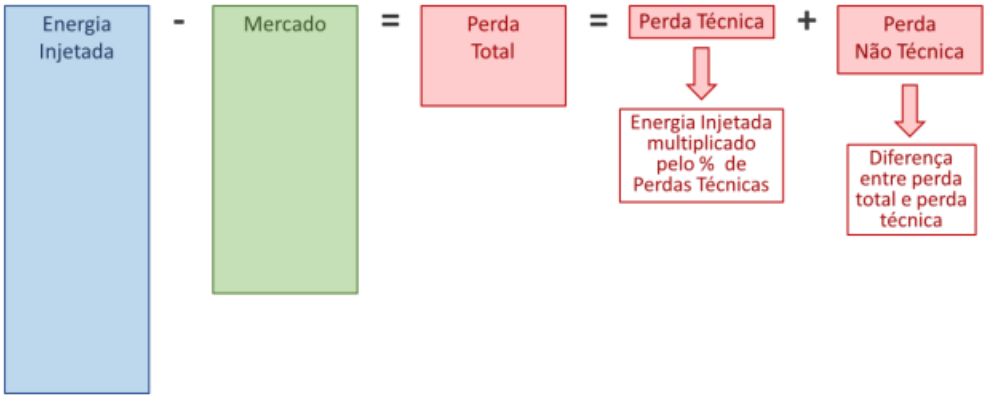
De acordo com [24], o cálculo das perdas técnicas geralmente abrange os seguintes aspectos:

- Perdas resistivas em alimentadores primários (Alta Tensão);
- Perdas associadas aos transformadores de distribuição, incluindo perdas resistivas no núcleo e nos enrolamentos;
- Perdas resistivas em alimentadores secundários (Média Tensão e Baixa Tensão);
- Perdas resistivas nos ramais de ligação, que conectam a distribuidora ao consumidor;
- Perdas nos equipamentos de medição.

Já as perdas não técnicas (PNT) ou perdas comerciais, calculadas como a diferença entre as perdas totais e as perdas técnicas, têm como principais causas furtos (como ligações clandestinas e desvios diretos da rede), fraudes (adulterações de medidores ou desvios), além de erros em leitura, medição e faturamento. Popularmente conhecidas como “gatos”, essas perdas estão amplamente relacionadas à gestão da concessionária e às condições socioeconômicas das áreas atendidas. [20].

A metodologia para o cálculo das perdas e obtenção dos percentuais de perdas não técnicas é ilutrado na figura 3.3.

Figura 3.3: Metodologia de cálculo utilizado para obtenção dos percentuais de perdas não técnicas.



Fonte: Adaptado de [20].

As PNT estão atreladas diretamente a fatores sociais e econômicos do país. Uma vez que os consumidores em situações adversas como, por exemplo, as pandemias que assolam o mundo inteiro diminuem o poder aquisitivo da população, aumentando os indices de pobreza e desemprego, consequentemente induzindo ao consumo irregular de energia.

Estima-se que as empresas de serviços públicos no mundo todo tem um prejuízo anual de mais de US$ 25 bilhões em razão da ocorrência de desvios irregulares de eletricidade. Esse problema tem acarretado diversos efeitos negativos, que vão desde a sobrecarga da unidade de geração, ocasionando sobretensão, e em casos extremos, interrupção no fornecimento de energia para os consumidores, levando as concessionárias a repassar tais extravios aos clientes lícitos na forma de tarifas [16].

As perdas comerciais refletem nas tarifas aplicadas às contas de energia dos consumidores finais. Apesar desse repasse tarifário ser limitado regulatoria-

mente pela ANEEL, em ciclos de 3 a 5 anos em cada distribuidora, de modo que os impactos das perdas reais não sejam totalmente repassados aos consumidores regulares e sejam analisadas especificidades,tais como características do mercado e variáveis socioeconômicas da região, o consumidor regular arca de forma parcial pelas irregularidades de outros consumidores, tendo em vista que a ANEEL reconhece valores regulatórios eficazes [1].

De acordo com a ANEEL, as perdas não técnicas reais no Brasil, considerando a multiplicação dos montantes de energia pelo preço médio das tarifas, desconsiderando tributos, ocasionaram um custo de aproximadamente R$ 9,9 bilhões em 2023. As perdas totais sobre a energia injetada nesse mesmo ano representaram cerca de 14,1% do mercado consumidor. Em quantidade de energia, as perdas técnicas no processo de distribuição corresponderam a aproximadamente 42,0 TWh e as perdas não técnicas 38,2 TWh.

Elevados índices de perdas não técnicas representam desafios econômicos significativos para as concessionárias, especialmente aquelas que atuam em países menos desenvolvidos [25].

Além disso, o consumo irregular em uma região pode elevar a demanda média a níveis que excedam a capacidade dos equipamentos elétricos, como transformadores, devido ao roubo de energia. Isso pode resultar em problemas como colapso de tensão, sobrecarga dos equipamentos e cortes no fornecimento de energia.

Se esses cortes se tornarem frequentes, podem impactar negativamente a relação entre os consumidores e a concessionária. Os clientes podem perder a confiança na capacidade da empresa de fornecer um serviço confiável, o que agrava o problema das perdas não técnicas, já que os consumidores, insatisfeitos, podem se recusar a pagar por um serviço percebido como inadequado e em constante

deterioração [26].

Os Países e regiões com elevados níveis de perdas não técnicas não parecem sofrer com a falta de iniciativa em programas voltados para energias renováveis. O Brasil, por exemplo, enfrenta desafios significativos relacionados às perdas não técnicas de eletricidade, mas seu mix de geração de energia é amplamente composto por fontes renováveis. Nos primeiros dez meses de 2020, fontes como hidrelétrica, nuclear, eólica, solar fotovoltaica e biomassa representaram quase 90% da geração total de energia, um aumento de 2% em relação ao mesmo período de 2019 [27].

Além disso, grandes projetos hidrelétricos geralmente estão situados em áreas remotas, longe dos principais centros populacionais. Essa distância dificulta a promoção de um senso de pertencimento nas comunidades locais, algo que ocorre quando as tecnologias de geração são visíveis e acessíveis. Esse engajamento comunitário é essencial para fomentar a percepção de que elas fazem parte da solução na redução das perdas não técnicas.

Em [28] foi proposto que, no Brasil, a adoção da geração distribuída (GD) por meio de sistemas de energia solar fotovoltaica em telhados, especialmente em áreas com altos índices de perdas não técnicas, poderia ser uma estratégia mais eficiente do que os métodos tradicionais que buscam bloquear o fluxo ilegal de energia da rede de distribuição. A ideia central é transformar consumidores irregulares em prosumidores regulares, indivíduos que atuam tanto como consumidores quanto como produtores de energia.

Essa abordagem reduziria as perdas não técnicas, melhoraria a qualidade do serviço, diminuiria os custos das contas de energia para consumidores em conformidade, reduziria as despesas de manutenção das concessionárias e contribuiria para o avanço socioeconômico de comunidades de baixa renda. Ainda assim,

há desafios significativos associados a isso, como o alto custo inicial de implantação dos sistemas, a necessidade de garantir a manutenção adequada, a integração da geração distribuída em redes elétricas frágeis ou que não foram projetadas para suportar tais configurações, e o tempo necessário para desenvolver competências e capacitar pessoas no uso dessas tecnologias.

3.1.2 Perdas Não Técnicas no Brasil

O desempenho na redução das perdas não técnicas no Brasil está em constante evolução e demanda investimentos crescentes, além de estudos voltados para o desenvolvimento de abordagens mais eficientes, dada a relevância da redução dessas perdas para o mercado de energia. As distribuidoras de energia desempenham um papel crucial na manutenção do fornecimento de energia aos consumidores em níveis adequados, enquanto buscam maximizar os lucros na prestação desse serviço [29].

De acordo com a ANEEL, as concessionárias de grande porte, com mercados superiores a 700 GWh, são responsáveis pela maior parte das perdas não técnicas no Brasil. Isso se deve ao tamanho de seus mercados e à maior complexidade envolvida no combate a essas perdas. Os níveis de perdas não técnicas são influenciados pela gestão das concessionárias, pelas condições socioeconômicas e pelos aspectos comportamentais das áreas de concessão.

Como cada concessionária opera em regiões com características específicas, como peculiaridades de mercado e variáveis socioeconômicas, a comparação entre elas é realizada com base em um ranking de complexidade. Esse ranking é desenvolvido a partir de modelos econométricos e permite avaliar o desempenho das distribuidoras na redução das perdas não técnicas. Por meio do ranking de complexidade, subentende-se que concessões localizadas em áreas de menor com-

plexidade socioeconômica deveriam apresentar menores coeficientes de perdas não técnicas.

A ANEEL utiliza uma metodologia regulatória que define valores de referência para as perdas não técnicas, conhecidos como valores regulatórios. Esses valores são geralmente inferiores às perdas reais registradas pelas concessionárias, pois são baseados em critérios de eficiência. Isso significa que a ANEEL estabelece um limite para o quanto das perdas não técnicas pode ser repassado aos consumidores por meio das tarifas, incentivando as concessionárias a reduzirem suas ineficiências.

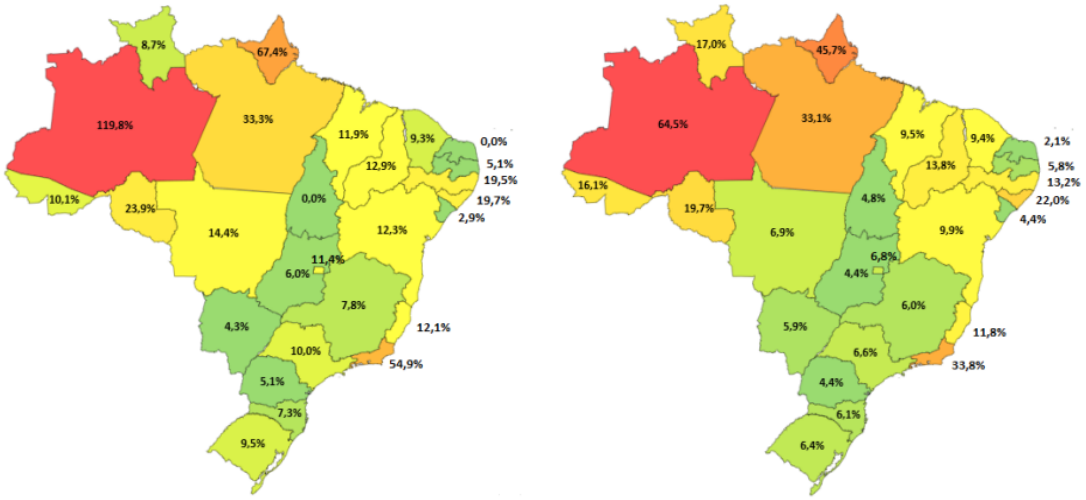
Dessa forma, as distribuidoras são incentivadas a atuar constantemente na redução das perdas, independentemente do nível regulatório estabelecido. Isso ocorre porque, além de mitigarem prejuízos quando as perdas reais excedem as regulatórias, elas também obtêm ganhos financeiros se as perdas reais ficarem abaixo do nível regulatório.

Essa regulação por incentivos protege os consumidores de custos adicionais injustificados, pois, caso as perdas reais de uma distribuidora excedam o nível regulatório estabelecido, os custos adicionais associados a essas perdas devem ser arcados pela concessionária. Assim os clientes não são penalizados por possíveis negligências da empresa.

Nas figuras 3.4a e 3.4b são apresentados os níveis das perdas não técnicas reais e regulatórias sobre o mercado de baixa tensão faturado no Brasil em 2023.

Figura 3.4: Comparação entre Perdas Não Técnicas Reais e Regulatórias.

(a) *Perdas Não Técnicas Reais* (b) *Perdas Não Técnicas Regulatórias*

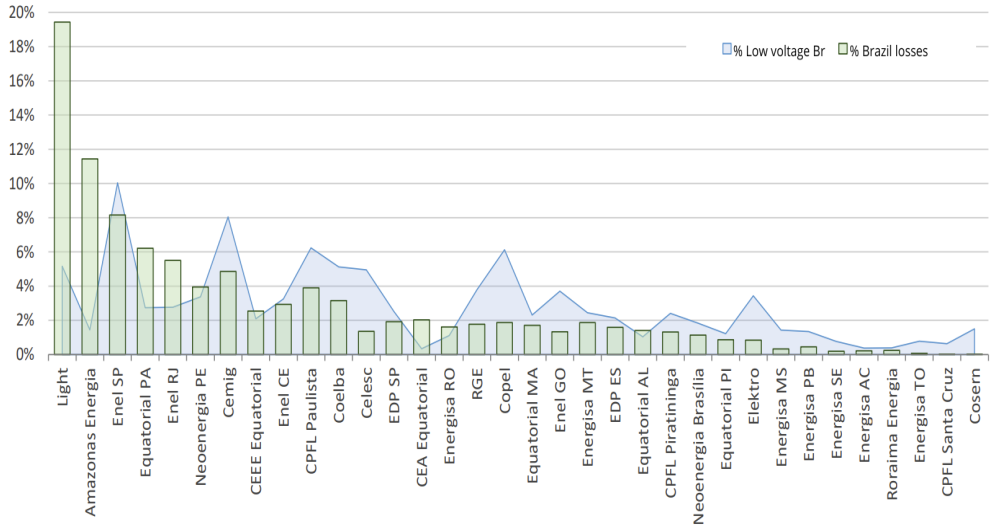


Fonte: [20].

O maior desvio foi registrado no estado do Amazonas, com 55,3 pontos percentuais acima do nível regulatório estabelecido. Em contrapartida, o estado do Rio Grande do Norte destacou-se com 0,0% de perdas não técnicas reais, apresentando o melhor desempenho no cenário nacional.

Na figura 3.5 é mostrada a participação das perdas não técnicas reais do país por concessionárias, em 2022, comparando com a representatividade do mercado BT no Brasil.

Figura 3.5: *Participação das PNT reais do país por concessionárias*

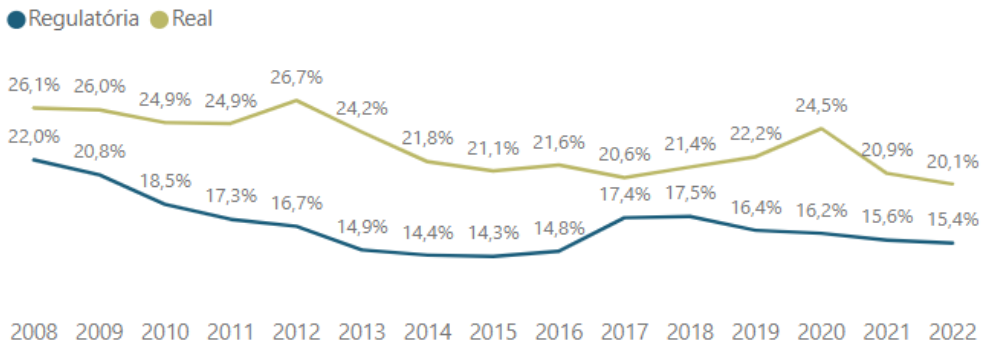


Fonte:[20].

É possível observar na figura 3.5 que as 10 distribuidoras com maiores montantes de perdas respondem por 69 % das perdas não técnicas do país, sendo somente a Light (RJ) e a Amazonas Energia (AM) responsáveis por 31 % dessa parcela.

O histórico de PNT das distribuidoras do Norte e Nordeste, no período de 2008 a 2022 é ilustrado na figura 3.6, comparando os limites regulatórios de repasse com as perdas reais obtidas por meio dos dados informados pelas distribuidoras no SAMP(Sistema de Acompanhamento de Informações de Mercado para Regulação Econômica).

Figura 3.6: PNT na Região Norte-Nordeste

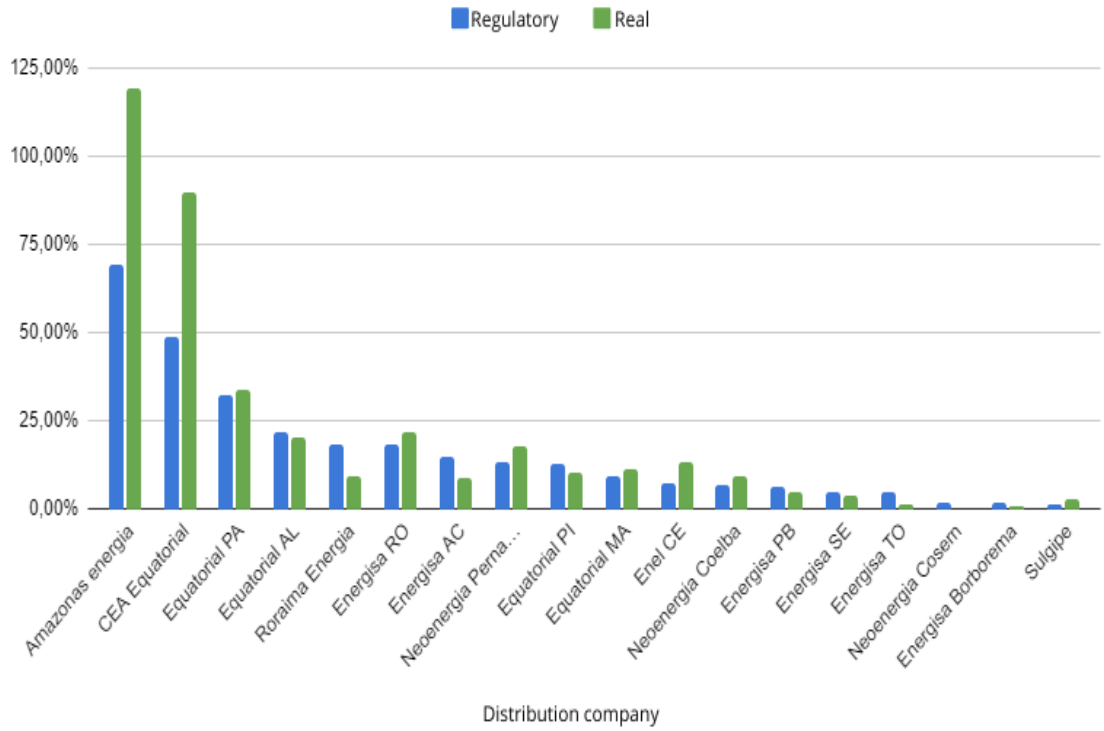


Fonte: [20].

Em 2012 obteve-se a maior diferença entre os valores de perdas reais e regulatórias, nos anos seguintes a diferença foi amenizada chegando ao seu menor valor em 2022. No entanto, durante todo esse período as concessionárias tiveram prejuízos econômicos, pois os valores regulatórios das perdas não técnicas são normalmente inferiores aos valores praticados pelas concessionárias, pois a metodologia adotada pela [30] observa critérios de eficiência, limitando o repasse das perdas não técnicas reais.

A figura 3.7 adaptada de [31] ilustra o desempenho específico de cada concessionária das Regiões Norte e Nordeste em 2022, no qual grande parte das concessionárias de distribuição obteve resultados negativos na redução da diferença das perdas reais em relação as regulatórias.

Figura 3.7: *PNT real e regulatório do mercado de baixa tensão*



Fonte: Adaptado de [31]

A concessionária Amazonas Energia (AM) ultrapassou as perdas regulatórias em 50,3 p.p. apresentando os maiores índices de perdas, seguida pela concessionária CEA Equatorial (AP) que também teve o limite regulatório ultrapassado cerca de 40,9 p.p. A Neoenergia Cosern (RN) obteve redução total nas perdas não tecnicas reais, apresentando o melhor desempenho, seguida por Energisa Borborema (PB), com cerca de 0,75 % abaixo do limite regulatório.

3.1.3 Influências de Aspectos Socioeconômicos nas Perdas Não Técnicas

De acordo com [32], as tarifas de energia elétrica no Brasil, especialmente no segmento de clientes residenciais, estão entre as mais elevadas do mundo. Esse cenário é particularmente preocupante quando se observa que tais patamares tarifários se aproximam e, em alguns casos, se equiparam aos praticados em países cuja renda per capita é significativamente superior à brasileira.

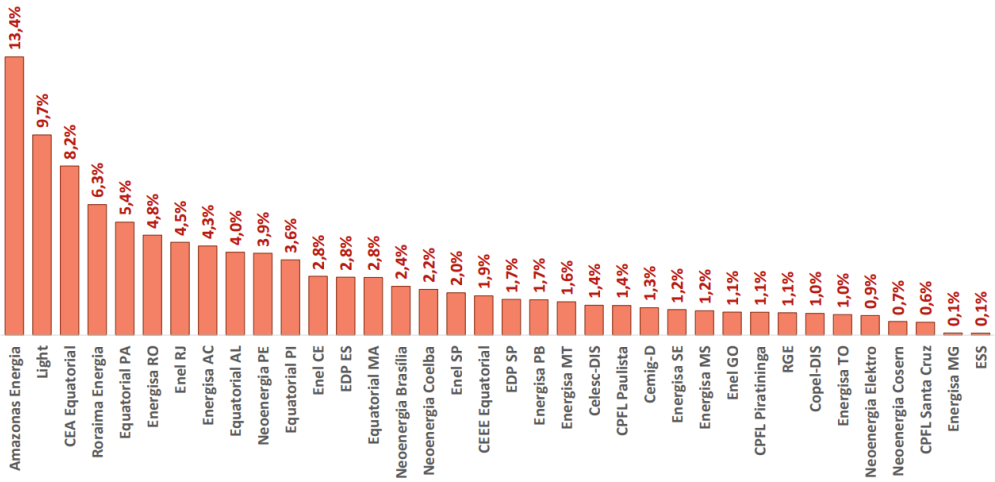
A composição dessas tarifas resulta de diversos fatores, entre os quais se destacam os encargos setoriais, impostos e tributos incidentes sobre a fatura de energia. Esses encargos englobam custos destinados a subsidiar políticas públicas, investimentos em infraestrutura do setor elétrico e programas de incentivo a fontes renováveis, enquanto a tributação inclui impostos federais, estaduais e, em alguns casos, municipais, compondo um percentual expressivo do valor final pago pelo consumidor. Tal estrutura tarifária impacta diretamente o poder aquisitivo das famílias e a competitividade da economia, especialmente em regiões onde a renda média é mais baixa.

Na figura 3.8 é mostrada a participação da componente de perdas não técnicas nas tarifas residenciais, considerando dados provenientes das Resoluções Homologatórias publicadas no período compreendido entre junho de 2022 e maio de 2023. As perdas não técnicas correspondem, essencialmente, à energia consumida, mas não faturada, em decorrência de irregularidades como furtos, fraudes, ligações clandestinas e falhas na medição. Esses custos, embora resultem de práticas ilícitas ou ineficiências operacionais, são repassados aos consumidores regulares, aumentando o valor das tarifas.

No caso da distribuidora Amazonas Energia, por exemplo, tal compo-

nente chegou a representar 13,4% do valor total da tarifa residencial no período analisado, o que evidencia a magnitude do problema e a necessidade de estratégias de mitigação mais eficazes. Esse percentual elevado também indica um desafio estrutural, uma vez que as condições socioeconômicas e geográficas da região amazônica, aliadas à dificuldade de fiscalização e monitoramento, contribuem para a persistência e a intensidade dessas perdas.

Figura 3.8: Participação das perdas não técnicas nas tarifas residenciais



Fonte:[20].

Em decorrência das elevadas tarifas de energia elétrica, os consumidores residenciais brasileiros comprometem parcela significativa do seu orçamento familiar com o pagamento desse serviço essencial. Tal situação leva a diferentes comportamentos de adaptação e reação por parte dos usuários: a redução voluntária do consumo, muitas vezes por meio de restrição do uso de equipamentos eletrodomésticos ou adoção de hábitos mais econômicos, o aumento da inadimplência, decorrente da incapacidade de arcar com os custos dentro dos prazos estabelecidos, e, em casos mais extremos, a adoção de práticas ilegais, como furtos e fraudes no fornecimento de energia elétrica. Essas últimas, embora ilícitas, tornam-se alter-

nativas para parte da população em razão do peso excessivo que a tarifa exerce sobre a renda disponível, especialmente em regiões de baixa renda.

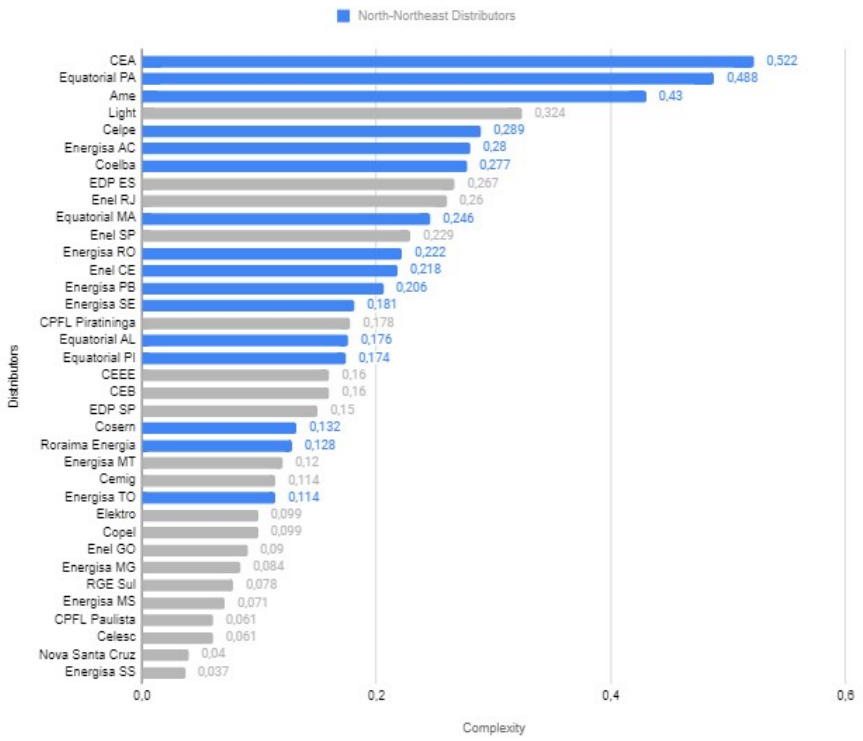
Quando se trata especificamente de furtos e fraudes, observa-se que as chamadas PNT estão fortemente associadas a variáveis socioeconômicas e estruturais das regiões de concessão. Entre os principais fatores relacionados estão o nível de escolaridade da população, as taxas de emprego e renda, a qualidade da infraestrutura urbana, as condições de habitação e até mesmo o grau de efetividade e confiabilidade das instituições governamentais locais. Nessas circunstâncias, distribuidoras que operam em áreas com maiores desafios socioeconômicos tendem a enfrentar níveis mais elevados de PNT, dada a complexidade de monitorar e coibir tais práticas em contextos de vulnerabilidade social.

Segundo a ANEEL, as metas de PNT estabelecidas para cada concessionária consideram não apenas aspectos técnicos do sistema elétrico, mas também características socioeconômicas e de mercado específicas das áreas onde atuam. A comparação entre diferentes concessionárias, no que diz respeito ao cumprimento dessas metas, é viabilizada pelo chamado índice de complexidade socioeconômica, um indicador sintético que reúne e pondera variáveis explicativas da dificuldade de combate às perdas não técnicas em cada área geográfica de concessão.

A metodologia de cálculo desse índice foi aprimorada pela [33], incorporando a aplicação de 138 modelos econométricos para construção do ranking de complexidade e da matriz de comparação entre empresas. O modelo considera múltiplas variáveis, incluindo índice de violência, níveis de pobreza e desigualdade social, precariedade habitacional, condições de infraestrutura e indicadores de inadimplência. Essa abordagem multivariada possibilita uma avaliação mais precisa da realidade de cada concessionária, permitindo que metas de PNT sejam mais justas e condizentes com os desafios enfrentados.

O *ranking* de complexidade das distribuidoras, divulgado em [30], apresenta os valores correspondentes para 36 empresas consideradas de grande porte, das quais 16 estão localizadas nas regiões Norte e Nordeste do país, conforme adaptação apresentada na figura 3.9. A elevada presença dessas regiões no grupo com maiores índices de complexidade reforça a influência determinante de fatores socioeconômicos e geográficos sobre a dificuldade de mitigação das perdas não técnicas, evidenciando que soluções puramente técnicas, sem ações estruturais paralelas, tendem a ter eficácia limitada.

Figura 3.9: *Ranking* de Distribuidores de Grande Escala



Fonte: Adaptado de [30].

O órgão regulador do setor elétrico reconhece que as concessionárias que operam em áreas com maior grau de complexidade socioeconômica e que, mesmo diante desse contexto adverso, conseguem apresentar níveis de PNT infe-

riores à média esperada demonstram maior eficiência operacional e capacidade de gestão. Essas distribuidoras passam a ser consideradas referências para o setor, servindo como exemplo de boas práticas e estratégias bem-sucedidas que podem ser adaptadas e replicadas por outras empresas. Esse reconhecimento não se limita a um mérito simbólico: ele também se traduz em incentivos econômicos, na medida em que o desempenho frente às metas regulatórias pode impactar a receita da concessionária.

No modelo regulatório vigente, quando as perdas reais superam as perdas regulatórias estabelecidas pela Agência Nacional de Energia Elétrica (ANEEL), a distribuidora tende a sofrer prejuízos financeiros, uma vez que não consegue repassar integralmente o custo adicional aos consumidores. Por outro lado, quando as perdas efetivas se mantêm abaixo do limite regulatório, a concessionária obtém ganhos econômicos, pois consegue reter parte da receita associada à energia que, embora comprada, não foi perdida. Esse mecanismo de incentivos busca alinhar os interesses das distribuidoras com os objetivos de eficiência e redução de desperdícios, promovendo, assim, maior comprometimento com o combate às PNT.

Contudo, é importante destacar que a solução para o problema das perdas não técnicas extrapola o âmbito puramente técnico-operacional. O combate eficaz a essas perdas exige a compreensão de sua estreita relação com fatores estruturais, como a situação do mercado de trabalho local, a renda per capita da população, as condições de habitação e a oferta de serviços públicos básicos.

Nesse sentido, este trabalho adota como referência a Síntese de Indicadores Sociais, divulgada pelo [34], para analisar o panorama das condições de vida da população brasileira. Tal abordagem possibilita estabelecer comparações e traçar paralelos entre os indicadores socioeconômicos e o cenário das perdas comerciais de energia elétrica no Brasil, com ênfase especial nas regiões Norte e Nordeste,

onde os índices de PNT tendem a ser mais elevados devido à combinação de fatores sociais, econômicos e geográficos conforme Tabela 3.1

Tabela 3.1: Correlação entre variáveis socioeconômicas e Perdas Não Técnicas (PNT)0

Variável Socioeconômica	**Relação com as Perdas Não Técnicas (PNT)**
Mercado de Trabalho	Altas taxas de desemprego e subutilização da força de trabalho aumentam a vulnerabilidade social, favorecendo práticas como fraudes, ligações clandestinas e inadimplência [35]. A informalidade do trabalho reduz a renda estável e amplia os riscos de PNT, sobretudo nas regiões Norte e Nordeste [34].
Renda Per Capita	Baixa renda domiciliar per capita torna o custo da energia mais pesado no orçamento familiar [36], levando ao aumento da inadimplência e, em alguns casos, ao uso de ligações irregulares para acesso à energia. Estados com maior proporção da população vivendo com até 1 salário mínimo apresentam índices mais elevados de inadimplência e também de PNT [34].
Condições de Moradia	A ausência de titularidade formal e a informalidade habitacional dificultam o controle e a fiscalização por parte das distribuidoras, criando um ambiente propício para furtos de energia. Populações em assentamentos precários e sem documentação apresentam maior propensão às perdas não técnicas.[37]
Índice de Vulnerabilidade Social	Regiões com maiores níveis de vulnerabilidade socioeconômica (Norte e Nordeste) apresentam índices mais altos de PNT, evidenciando a relação entre exclusão social, desigualdade econômica e incidência de perdas comerciais.[31]

Fonte: Adaptado de [31],[34],[35],[36],[37].

3.1.4 Classificação de Perdas Não Técnicas

As PNT de energia elétrica referem-se à energia consumida que não é faturada, geralmente resultante de ações irregulares de clientes ou de deficiências na gestão das distribuidoras. Essas perdas representam um problema relevante para o setor elétrico, pois impactam diretamente a receita das empresas, a sustentabilidade financeira do sistema e podem elevar os custos para os consumidores [38].

Segundo [39], as PNT podem ser classificadas de acordo com sua origem, permitindo identificar tanto a responsabilidade do cliente quanto as falhas internas das distribuidoras. Na Tabela 3.2 é apresentada uma síntese dessa classificação.

Tabela 3.2: Classificação das perdas não técnicas de energia elétrica segundo a origem

Categoria	**Ações do Cliente**	**Ações/Deficiências da Distribuidora**
Fraudes	Manipulação de medidores; ligações clandestinas	Falta de fiscalização; detecção tardia de irregularidades
Furtos de Energia	"Gatos"; desvios diretos da rede sem medição	Ausência de controle sobre áreas de risco
Erros/defeitos de Medição	Uso de dispositivos para reduzir registro de consumo	Medidores descalibrados; equipamentos obsoletos ou queimados
Erro de Leitura	Obstrução de acesso ao medidor ou alteração do visor	Leitura manual incorreta; falha no sistema de leitura remota
Cadastro Irregular	Registro de menor carga ou categoria inadequada	Falhas no cadastro técnico-comercial; falta de atualização
Corrupção/Fraude Interna	Suborno para não registrar irregularidades	Conivência de funcionários; falta de controle interno
Problemas Contratuais	Uso indevido de tarifas especiais ou inadimplência proposital	Políticas comerciais inadequadas ou mal fiscalizadas

Fonte: Adaptado de [38],[39],[40],[41].

Cada categoria de PNT possui características e impactos distintos, tais

como:

- **Fraudes:** envolvem manipulação deliberada dos medidores ou ligações clandestinas, dificultando a medição correta do consumo. Esse tipo de perda exige monitoramento constante e programas de inspeção eficientes [39].
- **Furtos de Energia:** também conhecidos como "gatos", são desvios diretos da rede que não passam pelo medidor. Geralmente ocorrem em áreas sem fiscalização adequada, representando risco à segurança e à confiabilidade do sistema [40].
- **Erros/Defeitos de Medição:** resultam de medidores descalibrados, equipamentos obsoletosou ou queimados e o uso de dispositivos que reduzem o registro do consumo real. A modernização tecnológica e a manutenção preventiva podem reduzir significativamente essas perdas [38].
- **Erro de Leitura:** pode ocorrer devido a obstrução do acesso ao medidor, alteração do visor ou falhas na leitura manual ou remota. A implementação de sistemas automáticos de leitura e a capacitação dos profissionais são medidas corretivas importantes [41].
- **Cadastro Irregular:** inclui registros incorretos de carga ou categoria inadequada do cliente. Essa falha administrativa gera perdas financeiras e prejudica a gestão tarifária, sendo necessária a atualização constante dos cadastros [39].
- **Corrupção/Fraude Interna:** envolve suborno ou conivência de funcionários para não registrar irregularidades. O controle interno rigoroso e auditorias periódicas são essenciais para mitigar esse tipo de perda [40].

- **Problemas Contratuais:** surgem do uso indevido de tarifas especiais ou inadimplência proposital, muitas vezes pela falta de políticas comerciais adequadas ou fiscalização ineficaz [38].

Portanto, a classificação das PNT não apenas permite compreender a origem das perdas, mas também auxilia na definição de estratégias preventivas e corretivas. A integração entre fiscalização, tecnologia e gestão administrativa é fundamental para reduzir esses impactos, melhorar a eficiência do sistema elétrico e garantir a sustentabilidade econômica do setor [40].

Em conclusão, embora as perdas não técnicas englobem diversas categorias, este estudo dará ênfase principalmente à detecção de defeitos de medição, furtos de energia e fraudes. Essas categorias representam as principais fontes de perdas financeiras e técnicas no sistema elétrico, sendo fundamentais para a formulação de estratégias de fiscalização, modernização tecnológica e controle administrativo, com o objetivo de reduzir impactos econômicos e melhorar a eficiência do setor [39, 38].

3.2 Aprendizado de máquina com árvore de regressão

A árvore de regressão é um método de aprendizado de máquina baseado na divisão recursiva de dados em subconjuntos homogêneos. Este método utiliza uma estrutura hierárquica onde os dados são divididos em nós internos, que representam condições baseadas nos valores das variáveis preditoras, e nós folha, que contêm os valores previstos da variável alvo. Este modelo é amplamente utilizado em aplicações que requerem previsão contínua devido à sua interpretabilidade e capacidade de lidar com dados não lineares [42].

Conforme observado por [43], as primeiras iniciativas envolvendo ár-

vores de regressão remontam ao final dos anos 1950, com Hunt conduzindo experimentos voltados à formulação de padrões [44]. Posteriormente, Breiman [42]. introduziu o algoritmo CART (Classification And Regression Trees), enquanto Quinlan [45]. apresentou os algoritmos ID3 (Iterative Dichotomiser 3, 1986) e C4.5 (1993) [46]. Além disso, a aplicação prática de árvores de regressão foi incorporada em soluções mais modernas como o Microsoft SQL Server, conforme destacado por Seidman [47], que utilizou essa abordagem tanto para problemas de classificação quanto de regressão.

Os algoritmos de árvores de regressão citados anteriormente são amplamente reconhecidos como pioneiros na área, tendo inspirado diversas variações que empregam a mesma abordagem baseadas no dividir para conquistar. Essa estratégia baseia-se em uma sequência de instruções condicionais do tipo se-então (if-then, em inglês) e segue uma estrutura hierárquica que progride de forma descendente, partindo do nó raiz até alcançar os nós folha [44]. Sendo uma das principais características da árvore de regressão a sua capacidade de organizar os registros da base de dados e fragmentar o problema original em múltiplos subproblemas menores, possibilitando a identificação de soluções mais simples e específicas para cada um desses subproblemas.

Um dos principais aspectos vantajosos das árvores de regressão é sua habilidade de capturar relações complexas entre as variáveis preditoras e a variável alvo. Diferentemente de modelos lineares, que assumem uma relação linear entre as variáveis, as árvores de regressão particionam iterativamente o espaço de entrada em regiões mais simples, permitindo representar relações não lineares com precisão. Essa flexibilidade torna o método particularmente útil em contextos onde as relações entre os dados são complexas ou desconhecidas [46].

Outro aspecto importante das árvores de regressão é sua interpretabi-

lidade. A estrutura hierárquica da árvore permite que os analistas identifiquem facilmente quais variáveis tiveram maior impacto nas previsões e quais condições foram usadas para dividir os dados. Sendo particularmente relevante em áreas como saúde e energia, onde a compreensão do processo de decisão é fundamental para a confiabilidade do modelo [48].

Apesar de suas vantagens, árvores de regressão possuem algumas limitações. Por exemplo, elas podem ser sensíveis a variações nos dados de treinamento, levando a modelos superajustados que não generalizam bem para novos dados. Estratégias como poda (*pruning*) e validação cruzada são frequentemente utilizadas para mitigar esses problemas, reduzindo a complexidade da árvore e melhorando sua capacidade preditiva [42].

Uma extensão popular das árvores de regressão é o uso de métodos ensemble, como florestas aleatórias (*random forests*) e *boosting*. Esses métodos combinam múltiplas árvores para melhorar a precisão e a robustez do modelo, reduzindo o risco de sobreajuste. Florestas aleatórias, por exemplo, treinam várias árvores de decisão em subconjuntos aleatórios dos dados e combinam suas previsões para gerar uma resposta final mais confiável [48].

Em síntese, foi possível verificar que as árvores de regressão são ferramentas versáteis para a previsão de dados contínuos. Elas oferecem um equilíbrio relevante entre simplicidade, interpretabilidade e capacidade de lidar com dados complexos, tornando-as adequadas para uma ampla gama de aplicações.

3.2.1 Estrutura da Árvore de Regressão

A árvore de regressão é um modelo preditivo que segmenta os dados em regiões de decisão baseadas em valores das variáveis preditoras. Cada divisão, ou *split*, é feita

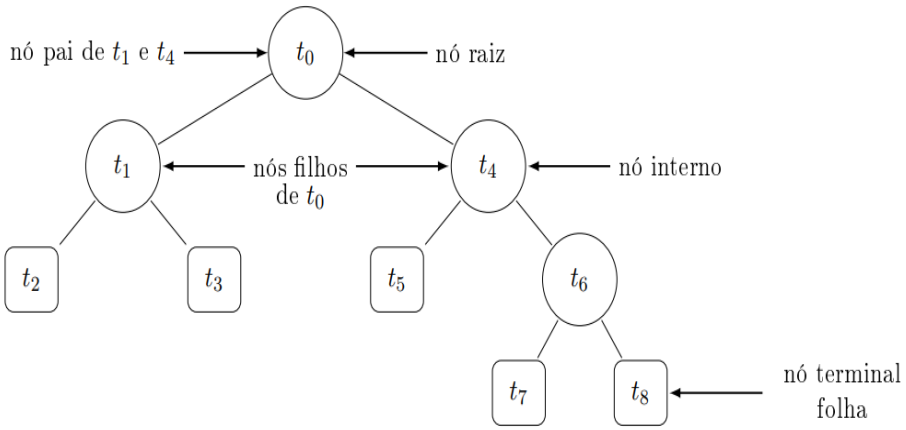
de forma a minimizar o erro de previsão dentro de cada região. A estrutura da árvore consiste em nós internos, que representam decisões baseadas em atributos, e nós folha, que contêm as previsões finais para cada região.

Uma árvore de regressão é treinada recursivamente, dividindo os dados em subconjuntos menores com base nos valores das variáveis preditoras, até atingir um critério de parada, como um número mínimo de observações em cada folha ou uma profundidade máxima da árvore.

Para sua aprendizagem, uma árvore de regressão pode ser estruturada de acordo com o tipo de variável de entrada, sendo estas contínuas, no caso de regressão, ou discretas, no caso de classificação [49]. De acordo com [50], árvores de decisão se destacam pela simplicidade de implementação. Dado um conjunto de dados, o usuário seleciona uma variável-alvo como saída, e o algoritmo determina a variável mais significativa associada à esta saída, definindo-a como o primeiro ponto de divisão ou nó raiz. Em seguida, outras variáveis relevantes são organizadas hierarquicamente em nós subsequentes até que os nós folha, ou terminais, sejam alcançados, representando o resultado final.

A figura 3.10 ilustra, esquematicamente, o gráfico que representa o algoritmo CART, *Classification And Regression Trees*.

Figura 3.10: *Modelo de uma árvore CART*



Fonte: [42]

O modelo CART é baseado em árvores binárias [51]. Este modelo possue as seguintes definições em sua composição:

- Árvore : Uma árvore é um grafo G = (V, E) no qual quaisquer dois vértices (nós) conectam-se exatamente por um único caminho.

- Árvore com raíz: Uma árvore com raiz é aquela em que um dos nós é designado como raiz. Logo, assume-se que uma árvore com raiz é um grafo estruturado a partir do nó raiz.

- Nós pais e nós filhos: Se existe uma aresta que conecta dois nós t1 e t2, onde t1 e $t2 \in \mathbb{E}$, portanto t1 é chamado de nó pai de t2, enquanto que t2 é nó filho de t1.

- Nós internos e nós terminais: Para uma árvore com raiz, um nó é classificado como nó interno quando possue um ou mais filhos e nó terminal quando não há filhos. Um nó terminal também é chamado de folha.

- Árvore Binária: Uma árvore binária é uma árvore com raiz onde estruturalmente todos os nós internos tem exatamente dois nós filhos.

Para construir uma Árvore de Decisão segundo [42], considera-se um conjunto de dados composto por p variáveis preditoras, representadas por $X \in \mathbb{R}^p$, e por uma variável de interesse Y, de modo que $(y_i, x_i),\ i \in \{1, 2, \ldots, n\}$ sintetizem as n observações de (Y, X). Define-se o espaço dos preditores por $\mathbb{R}^p$, assumindo que não há perda de generalidade.

De maneira geral, os modelos baseados em Árvores de Decisão realizam a divisão do espaço dos preditores em J regiões que satisfazem:

$$\mathbb{R}^p = \bigcup_{j=1}^{J} R_j, \quad R_j \cap R_{j'} = \varnothing \ \text{ se } j \neq j'. \tag{3.1}$$

Dessa forma, a previsão para uma nova observação x é obtida pela expressão:

$$\hat{f}(x) = \sum_{j=1}^{J} \hat{y}_{R_j} \cdot \mathbf{1}_{\{x \in R_j\}}, \tag{3.2}$$

onde $\hat{y}_{R_j}$ representa a média dos valores y_i para os quais $x_i \in R_j$, isto é:

$$\hat{y}_{R_j} = \frac{1}{N_j} \sum_{x_i \in R_j} y_i, \tag{3.3}$$

com N_j sendo o número de observações pertencentes à região R_j.

Dado que é inviável, do ponto de vista computacional, considerar todas as possíveis divisões do espaço $\mathbb{R}^p$, o algoritmo CART utiliza uma abordagem heurística denominada Partição Recursiva Binária. Esse método aplica uma estratégia que busca encontrar, em cada passo iterativo, a melhor divisão local do espaço dos preditores. Entretanto, devido à natureza iterativa do processo, não há garantia de que a árvore gerada represente a melhor solução global.

Quando se trata de um problema de regressão, o algoritmo CART tem início com a escolha de um preditor $X_v,\ v \in \{1, 2, \ldots, p\}$, e de um ponto de corte s tais que o espaço dos preditores $\mathbb{R}^p$ se obtém da união das regiões R_1 e R_2, definidas pela seguinte expressão:

$$R_1(v, s) = \{x \in \mathbb{R}^p : x_v < s\}, \quad R_2(v, s) = \{x \in \mathbb{R}^p : x_v \geq s\}. \tag{3.4}$$

As observações localizadas nas regiões R_1 e R_2 atendem às condições $x_v < s$ e $x_v \geq s$, respetivamente. A seleção do preditor e do ponto de corte é realizada de forma a maximizar a redução provocada na função de erro $L(\cdot, \cdot)$. Dessa maneira, o objetivo é determinar os valores v e s que minimizem a expressão:

$$\min_{v \in \{1,\ldots,p\},\, s \in \mathbb{R}} \left[\sum_{x_i \in R_1(v,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{x_i \in R_2(v,s)} (y_i - \hat{y}_{R_2})^2 \right]. \tag{3.5}$$

O processo de construção da árvore prossegue com a subdivisão sucessiva das regiões R_1 e/ou R_2 até que um critério de parada seja atingido. Este critério desempenha um papel essencial no controle da complexidade da árvore e pode ser definido como o valor mínimo de observações num nó final ou como o tamanho máximo da árvore, por exemplo. Os preditores categóricos, assim como os preditores contínuos, são tratados da mesma forma, sendo suas categorias convertidas para o valor médio correspondente da variável resposta.

3.2.2 Compromisso Viés-Variância em Árvores de Regressão

O compromisso viés-variância é um conceito essencial na modelagem estatística e no aprendizado de máquina, sendo amplamente discutido em diversas obras da área, como em [48] e [52]. Esse compromisso reflete a relação entre a capacidade do modelo de capturar padrões complexos nos dados (variância) e sua capacidade de generalizar para novas observações (viés). Modelos excessivamente simples tendem

a apresentar alto viés e baixa variância, enquanto modelos muito complexos podem exibir baixo viés, porém alta variância, levando a problemas de sobreajuste.

As árvores de regressão, apresentadas inicialmente por Breiman et al. em [42], particionam recursivamente o espaço dos preditores em regiões homogêneas, dentro das quais a predição é baseada na média das observações presentes. A principal vantagem desse método está na sua interpretabilidade e na capacidade de modelar relações não-lineares e interações complexas entre variáveis.

No entanto, ao construir uma árvore de regressão, surge o dilema viés-variância. Árvores muito profundas capturam detalhes específicos dos dados de treinamento, levando a alta variância e baixa capacidade de generalização. Por outro lado, árvores muito superficiais apresentam alto viés, pois simplificam excessivamente os padrões presentes nos dados, conforme discutido em [53].

3.2.3 Decomposição do Erro Esperado

O erro esperado de um modelo de aprendizado de máquina pode ser decomposto em três componentes principais, como destacado em [48]:

$$\mathbb{E}[(\hat{f}(X) - f(X))^2] = (\mathbb{E}[\hat{f}(X)] - f(X))^2 + \mathbb{E}[(\hat{f}(X) - \mathbb{E}[\hat{f}(X)])^2] + \sigma^2 \quad (3.6)$$

No qual:

- O primeiro termo representa o viés, que mede o erro devido à simplificação do modelo, levando a predições sistematicamente imprecisas;
- O segundo termo é a variância, que reflete a sensibilidade do modelo a variações nos dados de treinamento;

- O terceiro termo representa o erro irreduzível, inerente aos dados devido a ruído e fatores não modelados.

A compreensão dessa decomposição é fundamental para a construção de modelos eficazes, pois permite identificar as principais fontes de erro e aplicar estratégias adequadas para mitigá-las [54]. Um modelo com alto viés tende a subestimar a complexidade do problema, resultando em predições simplistas e, consequentemente, um desempenho insatisfatório. Em contrapartida, modelos com alta variância se ajustam demasiadamente aos dados de treinamento, capturando ruídos e peculiaridades específicas que não generalizam bem para novos dados.

Para lidar com o compromisso entre viés e variância, diversas abordagens podem ser adotadas, como regularização, uso de conjuntos de treinamento maiores e técnicas de validação cruzada [55]. Métodos como *bagging* e *boosting* também são amplamente utilizados para reduzir a variância, combinando múltiplos modelos fracos para formar um modelo mais robusto. Em particular, o *bagging* reduz a variância ao treinar modelos em diferentes subconjuntos de dados, enquanto o *boosting* minimiza o viés ao focar em observações mal preditas.

A avaliação adequada do erro do modelo envolve a escolha apropriada de métricas, como o erro quadrático médio (MSE) e o erro absoluto médio (MAE) [42]. Essas métricas fornecem uma visão clara do desempenho do modelo em termos de viés e variância, permitindo ajustes iterativos que buscam equilibrar ambos os componentes. A validação cruzada k-dobras é uma técnica amplamente empregada para obter uma estimativa confiável do erro esperado, garantindo que o modelo não esteja superajustado ou subajustado.

Outro aspecto relevante na decomposição do erro esperado é o impacto do pré-processamento dos dados e da engenharia de atributos. Técnicas de normalização, seleção de variáveis e tratamento de valores ausentes podem reduzir a

variância do modelo e melhorar sua capacidade preditiva [53]. A escolha criteriosa dos algoritmos de aprendizado também desempenha um papel crucial, pois modelos mais complexos, como redes neurais profundas, tendem a ter alta variância, enquanto modelos mais simples, como regressão linear, apresentam maior viés.

Portanto, o entendimento da decomposição do erro esperado e suas implicações práticas é essencial para o desenvolvimento de modelos preditivos eficazes. O sucesso de um modelo não reside apenas em minimizar o erro total, mas em encontrar o equilíbrio ideal entre viés e variância, garantindo que ele seja capaz de generalizar para novos dados com um desempenho consistente.

3.2.4 Estratégias para Balanceamento do Compromisso Viés-Variância

Para alcançar um equilíbrio adequado entre viés e variância, diversas estratégias podem ser aplicadas, conforme sugerido em [42] e [52]:

- **Poda de Árvores:** Consiste na remoção de ramos irrelevantes ou redundantes da árvore para reduzir a complexidade e minimizar o sobreajuste.
- ***Ensemble Learning:*** Métodos como *bagging* e *boosting* combinam múltiplas árvores para reduzir a variância sem aumentar significativamente o viés [42].
- **Critérios de Parada:** Definir limites mínimos de observações por nó ou uma profundidade máxima para evitar sobreajuste.
- **Validação Cruzada:** Avaliação de diferentes tamanhos de árvore em subconjuntos de dados para encontrar o nível ideal de complexidade [48].

Uma estratégia fundamental para controlar a variância é a utilização de técnicas de regularização, como a penalização dos parâmetros do modelo. Em

árvores de decisão, isso pode ser implementado por meio da restrição da profundidade máxima da árvore ou da imposição de limites para a divisão de nós [55]. A regularização impede que o modelo se ajuste excessivamente aos dados de treinamento, favorecendo a capacidade de generalização.

Outra abordagem eficiente para equilibrar viés e variância é o uso de dados sintéticos ou aumento dos dados (*data augmentation*), principalmente em conjuntos de dados pequenos [56]. Esse método cria variações artificiais dos dados existentes, permitindo que o modelo aprenda padrões mais generalizados sem se sobreajustar a um pequeno conjunto de amostras.

O ajuste dos hiperparâmetros da árvore de decisão, como o número mínimo de amostras por nó ou a complexidade máxima, desempenha um papel crucial no controle da complexidade do modelo [52]. Técnicas como a busca em grade (*grid search*) ou a busca aleatória (*random search*) são amplamente utilizadas para encontrar os valores ótimos desses parâmetros, equilibrando viés e variância de forma eficiente.

O uso de técnicas de reamostragem, como *bootstrapping*, pode ajudar na criação de múltiplas versões do conjunto de dados de treinamento, reduzindo a sensibilidade do modelo a variações específicas dos dados originais [42]. A combinação das previsões de múltiplos modelos treinados em diferentes amostras melhora a robustez do modelo e diminui sua variância.

Por fim, uma estratégia amplamente empregada para mitigar a variância excessiva em árvores de decisão é a utilização de florestas aleatórias (*random forests*) [42]. Esse método constrói múltiplas árvores a partir de subconjuntos aleatórios dos dados e combina suas previsões, resultando em um modelo mais estável e com melhor desempenho em dados não vistos.

3.2.5 *Cost-Complexity Pruning*

As árvores de decisão, quando construídas sem restrições, podem crescer excessivamente, resultando em modelos que se ajustam muito bem aos dados de treinamento, mas que apresentam baixo desempenho em novos dados devido ao sobreajuste [42]. Para mitigar esse problema, uma técnica amplamente utilizada é a poda por complexidade de custo, conhecida como *Cost-Complexity Pruning*.

Esse método consiste em encontrar um equilíbrio entre o ajuste aos dados de treinamento e a complexidade da árvore, removendo ramos que não contribuem significativamente para a redução do erro [48]. A técnica introduz um parâmetro de regularização, α, que controla o balanço entre o número de nós terminais e a capacidade explicativa da árvore. A função de custo-complexidade é definida como:

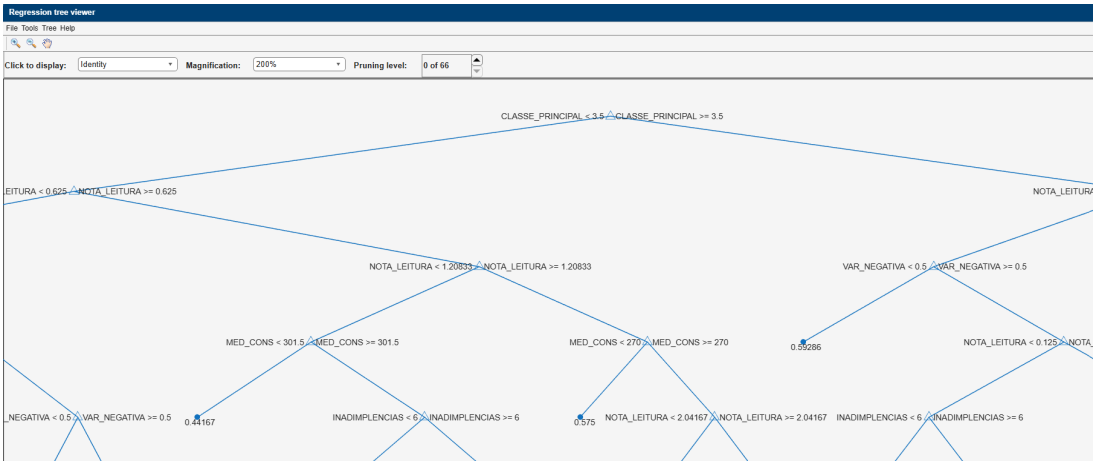
$$C_\alpha(T) = \sum_{m \in T} N_m Q_m(T) + \alpha |T| \tag{3.7}$$

Sendo:

- T representa a árvore gerada,
- N_m é o número de observações no nó m,
- $Q_m(T)$ é a impureza no nó m,
- α é o parâmetro de regularização,
- $|T|$ representa o número total de nós terminais na árvore.

A estratégia de poda segue um processo iterativo, onde, para diferentes valores de α, subárvores são geradas a partir da árvore completa, e aquela que apresenta o menor erro de validação cruzada é selecionada [52]. Esse processo melhora a capacidade de generalização do modelo ao reduzir a complexidade desnecessária. Em MATLAB, esse processo pode ser implementado manualmente com o uso de *prune* e análise da função de perda em diferentes níveis de complexidade [33].

Figura 3.11: Visualização do processo de poda por meio da função *prune* no MATLAB



Fonte: Adaptado de [58].

Estudos demonstram que a poda por complexidade de custo melhora significativamente a interpretação e estabilidade das árvores de decisão, proporcionando um modelo mais compacto e eficiente, sem perda substancial de precisão preditiva [55].

3.2.6 Determinação do Erro de Previsão do Modelo

Avaliar o desempenho preditivo de um modelo de árvore de decisão é essencial para garantir sua eficácia em cenários reais. A determinação do erro de previsão envolve quantificar a discrepância entre as previsões do modelo e os valores reais observados. Essa análise é geralmente realizada utilizando métricas específicas para diferentes tipos de problemas, como regressão e classificação [48].

No contexto de regressão, o erro pode ser medido por métricas como:

- **Erro Quadrático Médio (MSE):**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.8}$$

- **Erro Absoluto Médio (MAE):**

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.9}$$

Para problemas de classificação, as métricas mais comuns incluem a acurácia, a matriz de confusão e a métrica F1-score, que avaliam a proporção de predições corretas e a harmonia entre precisão e revocação [52].

Outro método amplamente utilizado para avaliar o desempenho de árvores de decisão é a validação cruzada k-dobras, na qual o conjunto de dados é dividido em k partes iguais. O modelo é treinado em $k-1$ subconjuntos e avaliado no subconjunto restante. Esse processo é repetido k vezes, garantindo uma estimativa robusta do erro [55].

Além das métricas tradicionais, a importância de variáveis e a análise de resíduos são ferramentas úteis para identificar possíveis problemas no modelo, como

a presença de vieses sistemáticos ou predições inconsistentes [42]. Métodos como *bootstrapping* e conjuntos de dados independentes para teste são frequentemente utilizados para validar ainda mais o modelo. Assim, a escolha da métrica de erro mais adequada depende da natureza do problema e dos objetivos específicos da análise, sendo essencial garantir que a avaliação do modelo reflita com precisão sua capacidade preditiva em novas amostras.

4 ALGORITMO APLICADO A IDENTIFICAÇÃO E PRIORIZAÇÃO DE PERDAS NÃO TÉCNICAS

Este capítulo descreve a metodologia proposta para a priorização e identificação de instalações potencialmente irregulares em sistemas de distribuição de energia elétrica, utilizando algoritmos de aprendizado supervisionado com base em árvores de regressão. A proposta contempla a implementação computacional da técnica em ambiente MATLAB, visando à construção de um modelo preditivo capaz de hierarquizar os alvos de fiscalização com base em características históricas de consumo e comportamento cadastral e técnico.

Inicialmente, apresenta-se uma visão geral da abordagem adotada, fundamentada em técnicas de ciência de dados, onde a modelagem do problema se dá por meio da regressão supervisionada. A metodologia é estruturada em três grandes etapas: (i) preparação e qualificação da base de dados, (ii) extração de variáveis preditoras relevantes por unidade consumidora, e (iii) construção e avaliação do modelo de regressão por árvore de decisão.

Na etapa inicial, realiza-se o carregamento da base de dados com registros mensais de consumo, trocas cadastrais e indicadores de inadimplência e medição. Em seguida, aplica-se uma filtragem de qualidade, restringindo o conjunto de análise apenas às instalações que possuem exatamente 12 medições mensais válidas. Adicionalmente, são descartadas as unidades que apresentaram trocas recentes de titularidade ou medidor, a fim de garantir consistência estatística no

conjunto de treinamento e evitar falsos positivos.

Na segunda etapa, são extraídas variáveis de interesse a partir do histórico de cada instalação, incluindo: (i) a classe principal do cliente, (ii) o histórico e consumo médio mensal, (iii) o histórico das notas de leitura apontadas pelo leiturista, utilizadas como proxy de confiabilidade do dado, e (iv) a contagem de meses com inadimplência. Com base nesses indicadores, calcula-se ainda um *score* composto, atribuído por meio de pesos ponderados de acordo com a relevância da informação, utilizado como variável alvo para o treinamento supervisionado do modelo.

Na etapa final, constrói-se um modelo de árvore de regressão, capaz de estimar o *score* de priorização das instalações a partir das variáveis extraídas. O algoritmo divide iterativamente o espaço de decisão, aprendendo padrões e condições que maximizam a variabilidade explicada do *score* atribuído. Para evitar o sobreajuste (overfitting), é aplicada a técnica de poda por complexidade (Cost-Complexity Pruning), que avalia diferentes níveis de simplificação da árvore com base no MSE. O modelo final é escolhido com base no nível de poda que apresenta o menor erro de validação.

O desempenho do modelo é avaliado com base em métricas estatísticas de erro. O MSE mensura a média dos quadrados das diferenças entre os escores reais e os previstos, penalizando fortemente grandes desvios. Já o MAE indica a média das diferenças absolutas, sendo uma medida mais robusta a outliers. Os valores obtidos (MSE = 0,0002 e MAE = 0,0027) atestam a elevada acurácia do modelo.

As previsões geradas são organizadas em ordem decrescente de prioridade, e o sistema exibe os 100 principais alvos com maior propensão a apresentar irregularidades no consumo e prioridade de regularização. Esta etapa final visa

subsidiar de forma objetiva e automatizada as ações de fiscalização em campo, com base em evidências estatísticas extraídas do histórico de comportamento das instalações.

A metodologia apresentada representa uma evolução em relação à simples aplicação de regras fixas, uma vez que o modelo de árvore de regressão permite identificar combinações não triviais de variáveis e capturar relações não-lineares entre os atributos e o grau de risco associado a cada unidade consumidora.

4.1 Preparação e qualificação da base de dados

A qualidade e consistência da base de dados são fatores críticos para o desempenho de modelos de aprendizado de máquina aplicados à identificação de perdas não-técnicas em sistemas de distribuição. A presente etapa tem como objetivo garantir que as informações utilizadas no treinamento do modelo reflitam o comportamento real das unidades consumidoras, sem a presença de ruídos estatísticos ou registros inconsistentes que comprometam a interpretação dos padrões.

A construção da modelagem da base foi fundamentada na análise do histórico de dados reais dos clientes, levando em consideração tanto as informações cadastrais e técnicas quanto os padrões de consumo de energia registrados fornecidos por uma distribuidora da região Nordeste, totalizando 6.963 unidades consumidoras com registro de consumo de doze meses entre casos de clientes regulares e clientes fraudadores.

Tabela 4.1: Descrição das Variáveis por Classe

Classe	**Variável**	**Descrição**
Dados Cadastrais	INSTALACAO REGIONAL CLASSE_PRINCIPAL PARCEIRO_NEGOCIO TROCA_TITULARIDADE	Dados de cadastro, identificação e localização do cliente
Dados Comerciais	MES_COMPETENCIA STATUS_COMERCIAL VARIACAO_CONSUMO_ATUAL DESVIO_PADRAO_DIF_CONS OUTLIER_DIF_CONS DIF_CONS DIF_CONS% MED_CONS INADIMPLENCIA	Dados referente ao perfil, histórico de consumo e adimplência do cliente
Dados Técnicos	QTD_NOTAS_FISC DISTANCIA_LEITURA MEDIDOR_ATUAL TROCA_MEDIDOR LEITURA_ATUAL NOTA_LEITURA NORMALIZACAO STATUS_AVANCO_LEITURA	Dados relacionados ao medidor, registro de consumo, estado da medição e histórico de fiscalizações e/ou substituições da medição

Fonte: Elaborado pela autora (2025).

Os dados cadastrais dizem respeito as informações de identificação, localização e tipo de cliente em que a unidade consumidora está inserida. Estes dados são de fundamental importância na detecção de mudanças de titular da UC, localização geográfica e tipo de carga declarada por tipo de cliente, fatores que influenciam em possíveis mudanças de consumo, evitando assim, falsos positivos na identificação de perdas não técnicas.

Já os dados comerciais dizem respeito à caracteristicas de consumo e situação do cliente em relação à distribuidora, com o histórico de demanda registrada, variações de consumo, suspensões de fornecimento e adimplência do con-

trato. Tais informações se fazem necessárias e relevantes em casos de priorização de recuperação de receita e identificação de consumos não registrados ou bruscas variações no perfil do cliente.

Os dados técnicos são referente ao estado do medidor instalado na UC, como o código único de identificação do medidor, para detectar possíveis trocas de medição oriundas de fiscalizações recentes e a situação do medidor em campo por meio do apontamento ou leitura registrada, sendo possível identificar possíveis defeitos ou irregularidades, causadoras de perdas comerciais.

Inicialmente, foi necessário compreender a estrutura lógica e relacional da base, verificando-se os tipos de variáveis disponíveis (categóricas, binárias e contínuas), a cardinalidade dos registros por instalação e a ocorrência de valores ausentes ou discrepantes.

A base continha, em sua origem, 15.004 clientes, com dados em diferentes níveis de granularidade temporal e duplicações associadas a modificações cadastrais simultâneas a eventos de leitura, o que motivou a aplicação de uma rotina de deduplicação por chave composta (instalação + mês competência). A primeira verificação consistiu na contagem de registros por instalação, especificando que cada UC possuísse exatamente 12 medições mensais consecutivas, sem lacunas temporais. Essa condição garante uma janela fixa para análise e equaliza o peso de cada observação no modelo. Instalações com menos de 12 registros válidos foram desconsideradas, conforme mostra a figura 4.1, na etapa de validação da série histórica. Essa decisão se alinha a práticas recomendadas de modelagem supervisionada que requerem uniformidade na série histórica por amostra [48].

Após validação da série histórica, a etapa crítica foi a definição de regras de desclassificação automática de instalações cujo histórico compromete a capacidade de inferência do modelo. Três critérios principais foram empregados:

(i) ocorrência de troca de titularidade dentro da janela de análise, (ii) mais de uma troca de medidor, e (iii) existência de visitas técnicas no intervalo temporal registrado.

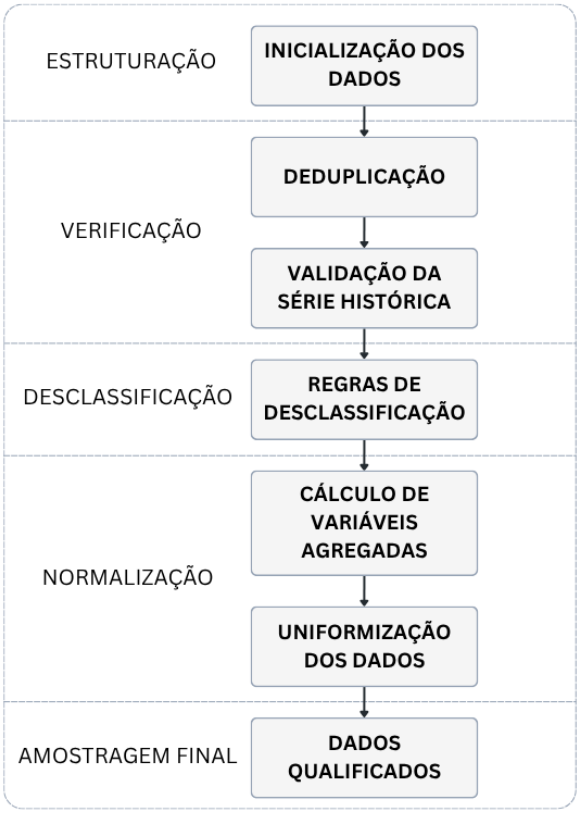
A troca de titularidade implica mudança no padrão de consumo que não reflete necessariamente fraude, mas uma substituição legítima do perfil da instalação. Assim, por precaução metodológica, estas amostras foram excluídas do conjunto de treino. As trocas múltiplas de medidor e registros de visitas de inspeção indicam histórico de intervenção técnica, podendo distorcer o perfil da unidade consumidora. Essas variáveis, embora relevantes para detecção de perdas em contexto mais amplo, foram tratadas como critério de desclassificação para o modelo de regressão supervisionada afim de selecionar unidades consumidoras que ainda não foram identificadas como possíveis alvos e não houveram interferências técnicas recentes.

Para complementar e traçar um perfil analítico do consumidor construi-se variáveis estatísticas agregadas por instalação, como a média, o desvio padrão, a variação mensal e outlier de consumo mensal, que foram utilizadas como preditoras no modelo. Paralelamente, a uniformização da base contemplou também a padronização dos formatos de data, códigos de instalação, tipos de classe consumidoras e codificação binária das variáveis categóricas (ex: troca de medidor = 0 ou 1).

Finalizada a filtragem e normalização, a base passou a conter apenas instalações com série mensal íntegra, dados confiáveis de leitura e histórico sem interferências externas relevantes, totalizando 6.963 clientes. Esse refinamento é essencial para evitar que o modelo aprenda padrões espúrios ou enviesados, como uma aparente queda de consumo motivada por uma troca de titularidade legítima. A figura 4.1 resume essas ações na forma de fluxograma sequencial, evidenciando

a lógica progressiva das etapas de qualificação: estruturação → verificação → desclassificação → normalização → amostragem final.

Figura 4.1: Fluxograma de Preparação e Qualificação da Base de Dados



Fonte: Elaborado pela autora (2025).

Após a etapa de qualificação da base de dados, a próxima fase da metodologia consiste na extração e consolidação de variáveis preditivas relevantes por unidade consumidora. Essas variáveis constituem os atributos de entrada do modelo supervisionado, sendo utilizadas para estimar a propensão à irregularidade no consumo de energia.

4.2 Extração de variáveis preditoras relevantes por unidade consumidora

O objetivo central desta etapa é traduzir o histórico de comportamento de cada instalação em um vetor de características fixas, que represente fielmente suas peculiaridades de consumo, inadimplência e confiabilidade cadastral ao longo do período analisado. A extração de variáveis é conduzida com base na janela de 12 meses consecutivos, previamente filtrada na etapa de qualificação. Todos os atributos extraídos são agregações estatísticas calculadas por instalação.

A primeira variável selecionada é a média de consumo mensal (`MED_CONS`). Ela representa o valor energético típico consumido pela instalação no período, e tem forte relação com a classe econômica e o perfil de uso da unidade. O consumo médio foi escolhido, em vez do valor total acumulado, para garantir invariância à duração da série, além de facilitar a comparação entre consumidores de mesma natureza, independentemente de sua demanda sazonal.

A segunda variável extraída é a média da variável (`CLASSE_PRINCIPAL`), que codifica numericamente o tipo de consumidor (por exemplo: 1 = residencial, 2 = comercial, 3 = rural, etc.). Embora a classe seja, em teoria, uma variável categórica estática, na prática podem ocorrer alterações cadastrais ao longo do ano. Assim, a média ponderada fornece uma aproximação contínua do perfil dominante da instalação, conforme valores atibuídos durante a normalização dos dados, conforme tabela abaixo:

Tabela 4.2: Código das Classes de Consumo

Classe de Consumo	**Código**
Consumo Próprio	1
Iluminação Pública	2
Residencial	3
Rural	4
Poder Público	5
Serviço Público	6
Comercial, Serviços e Outras A	7
Industrial	8

Fonte: Elaborado pela autora (2025).

A classificação se deu seguindo orientação de especialistas da área em relação ao nível de relevância de cada tipo de cliente quando se trata de variação de consumo e detecção de perdas não técnicas. De modo que clientes com maior nível de consumo e complexidade, como clientes industriais, receberam maior peso, enquanto clientes de consumo próprio (como agências de atendimento) receberam menor peso de relevância, visto que já são monitoradas diariamente pelos próprios colaboradores.

Uma terceira variável considerada como preditora é a média das notas de leitura (`NOTA_LEITURA`). Este indicador é extraído da coluna que codifica a confiabilidade da leitura mensal feita no local. Notas elevadas estão associadas a apontamentos de leitura feito pelos leituristas informando a situação irregular do medidor em campo.

Durante a normalização, as notas de leitura normal receberam peso 0, notas indicando problemas cadastrais como duplicidade de cadastro (uma instalação com dois medidores, sendo apenas um medidor funcional) com peso 1, medidores com defeitos técnicos ou displays apagados com atribuição 2 e medido-

res com fraudes ou furtos de energia com peso 3, sendo estes os mais relevantes para a identificação de perdas não técnicas, visto que se trata de um consumo não registrado pela distribuidora e de uma possível perda de receita definitiva.

A quarta variável preditiva é a contagem de meses com inadimplência de pagamento das faturas de energia elétrica dos contratos por instalação representada na coluna `(INADIMPLENCIAS)`. Essa variável é gerada por meio da soma binária dos meses em que a instalação apresentou falta de pagamento no prazo. A inadimplência é reconhecidamente uma variável proxy para risco operacional e comportamental. Instalações com histórico recorrente de inadimplência tendem a apresentar maior probabilidade de fraude, seja como causa ou consequência [14].

A quinta variável preditiva é a presença de variação brusca negativa de consumo mensal de energia elétrica nas unidades consumidoras, representada pela coluna `(VAR_NEGATIVA)`. Essa variável é construída com base na análise da coluna `(DIF_CONS)`, que representa a diferença de consumo entre meses consecutivos. Quando há pelo menos um mês com variação inferior a um limiar negativo definido (neste caso, -100 kWh), a variável é marcada como 1, sinalizando um comportamento atípico de queda abrupta no consumo. Este tipo de padrão pode ser indicativo de irregularidades operacionais, como intervenções na medição, consumo fraudulento mascarado por subutilização proposital ou defeitos no equipamento de medição. A literatura técnica e estudos empíricos apontam que quedas súbitas e injustificadas no consumo podem estar correlacionadas com práticas fraudulentas, sendo, portanto, um forte indício de risco e justificativa para priorização de inspeções. [57].

Além dessas variáveis principais, foram consideradas outras de apoio, como o número de trocas de titularidade, trocas de medidor e quantidade de notas de inspeção realizadas, que foram utilizadas como critérios de desclassificação e

também como descritores qualitativos da amostra. No entanto, como essas últimas variáveis foram utilizadas para excluir registros na etapa anterior, elas não entram diretamente no modelo de regressão como preditoras. Isso evita circularidade ou redundância explicativa.

Todas as variáveis preditivas foram escaladas e formatadas em uma tabela final por instalação, resultando em um dataset com colunas fixas: `INSTALACAO, MED_CONS, CLASSE_PRINCIPAL, NOTA_LEITURA, INADIMPLENCIA e SCORE (alvo)`.

Tabela 4.3: Exemplo de Dados por Instalação (Variáveis preditoras)

INSTALACAO	MED_CONS	CLASSE	NOTA_LEITURA	INADIMPLENCIAS	VAR_NEGATIVA
1001001	428.5	2.0	0.12	5	1
1001055	180.7	1.0	0.08	1	0
1001123	672.3	2.5	0.40	8	1

Fonte: Elaborado pela autora (2025).

A criação dessas variáveis segue os princípios de interpretabilidade e estabilidade, ou seja, devem ser facilmente compreensíveis por especialistas de negócio e estatisticamente consistentes ao longo de diferentes amostras.

Uma vantagem do modelo baseado em árvore de decisão é que ele não requer normalização das variáveis, pois as divisões internas da árvore são feitas com base em limiares absolutos, e não em distâncias métricas. Ainda assim, análises exploratórias foram conduzidas para identificar colinearidade entre os atributos. Nenhuma correlação forte foi detectada entre os preditores principais, o que favorece a construção de um modelo estável.

A seleção das variáveis foi validada empiricamente por meio da importância relativa dos atributos na árvore final treinada, conforme será discutido na próxima seção. A decisão de manter um conjunto enxuto de preditores buscou evitar o problema de "overfitting dimensional", onde um número excessivo de variáveis

pode induzir divisões artificiais na árvore.

Paralelamente, assegurou-se que todas as variáveis utilizadas tivessem significado operacional, ou seja, fossem compreendidas e acionáveis por equipes técnicas e especialistas da distribuidora.

Variáveis como "troca de medidor" e "troca de titularidade", embora potencialmente informativas, foram tratadas como restrições de entrada, e não como variáveis explicativas, justamente para evitar confusão causal. A estrutura final da base para treinamento do modelo consiste, portanto, em uma tabela com N linhas (instalações) e 6 colunas preditoras, além da variável resposta `(SCORE)`. Esta estrutura tabular representa o ponto de partida para a construção do modelo de regressão por árvore de decisão, abordada na próxima seção metodológica.

4.3 Construção e avaliação do modelo de regressão por árvore de decisão

A etapa final da metodologia compreende a modelagem supervisionada do problema, utilizando árvores de regressão para estimar o escore de priorização por instalação, com posterior aplicação da técnica de poda por complexidade (cost-complexity pruning). Essa abordagem foi escolhida por combinar interpretabilidade, desempenho e capacidade de captura de relações não lineares entre variáveis.

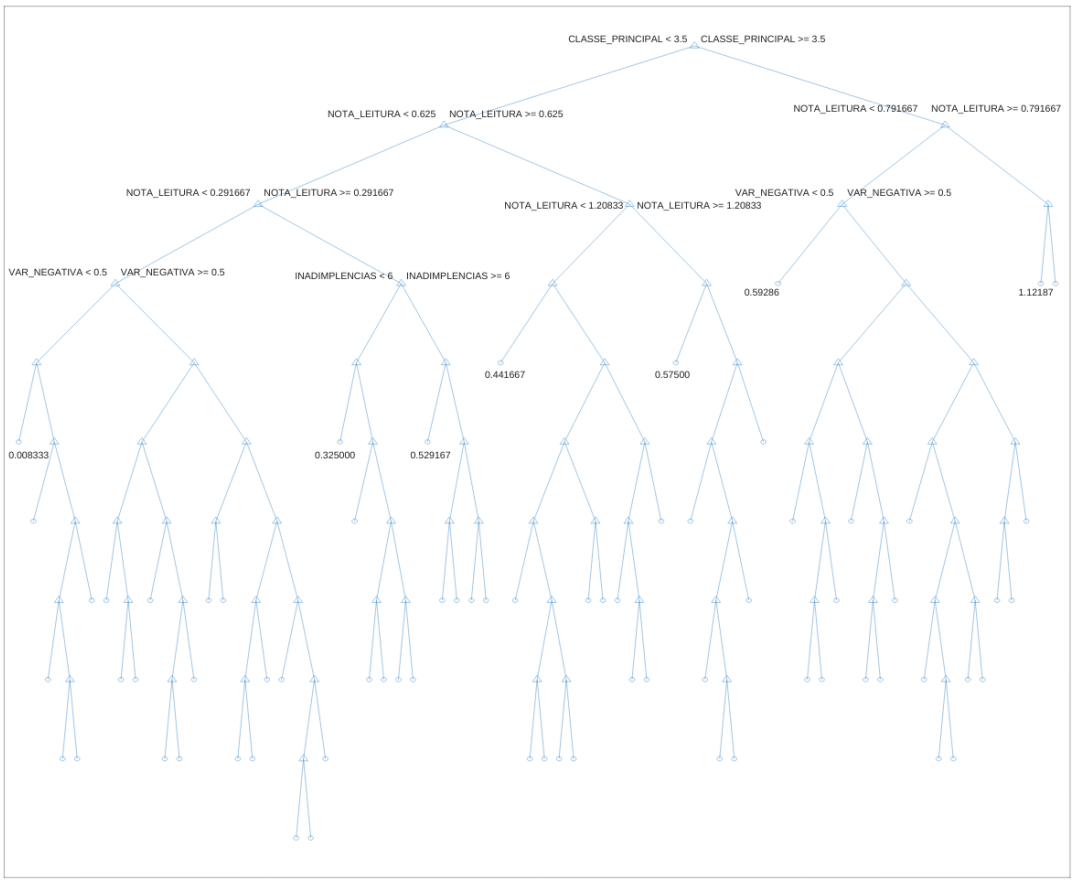
O modelo de aprendizado supervisionado empregado neste estudo é a árvore de regressão (regression tree), implementada via o algoritmo fitrtree da toolbox de estatística do MATLAB [58].

O objetivo da árvore é construir um conjunto hierárquico de regras (divisões binárias) com base nas variáveis preditivas extraídas anteriormente, de

forma a minimizar o erro de predição do *score* (variável alvo). Cada nó da árvore realiza uma divisão do espaço amostral com base em um limiar de uma variável, criando subconjuntos mais homogêneos em relação ao valor da variável resposta.

A árvore completa (não podada) tende a ajustar perfeitamente os dados de treino, criando ramos muito específicos para amostras individuais, o que resulta em sobreajuste (overfitting). Para evitar esse problema, aplicou-se a técnica de poda por complexidade, que busca simplificar a árvore reduzindo seus ramos de forma gradual, avaliando o impacto dessa redução no erro de previsão.

Figura 4.2: Árvore completa antes da aplicação da poda por complexidade de custo



Fonte: Elaborado pela autora (2025).

A estrutura detalhada na figura 4.2 ilustra todos os ramos gerados pelo modelo treinado com base nos dados das instalações.

4.4 Modelo de Previsão

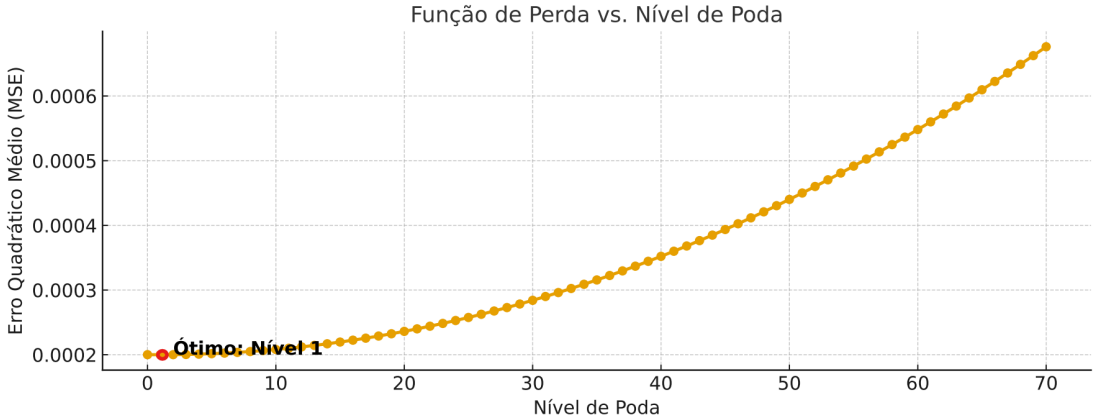
O modelo foi treinado utilizando como entrada a matriz de variáveis preditoras (MED_CONS, CLASSE_PRINCIPAL, NOTA_LEITURA, INADIMPLENCIAS e

VAR_NEGATIVA) e como saída o escore predito atribuído por pesos. O algoritmo dividiu iterativamente os dados em subconjuntos, calculando em cada divisão o ponto ótimo de separação baseado no critério de minimização da soma dos erros quadráticos residuais. O resultado foi uma árvore com profundidade completa, com vários nós terminais altamente especializados em subgrupos da amostra.

A árvore inicial forneceu excelente desempenho nos dados de treino, mas uma análise com dados de validação revelou aumento do erro preditivo, indicando overfitting. Para lidar com esse problema, foi executada a rotina de poda baseada em avaliação do Erro Quadrático Médio (MSE) em diferentes níveis de simplificação da árvore.

O processo de poda consistiu em avaliar sucessivamente a árvore com cortes nos ramos menos relevantes (nós com menor ganho de informação), até atingir o ponto em que o MSE se torna mínimo. A figura 4.3 ilustra o gráfico da função cvLoss, que representa o MSE obtido em validação cruzada para cada nível de poda.

Figura 4.3: Erro de validação (MSE) por nível de poda (Cost-Complexity Pruning)



Fonte: Elaborado pela autora (2025).

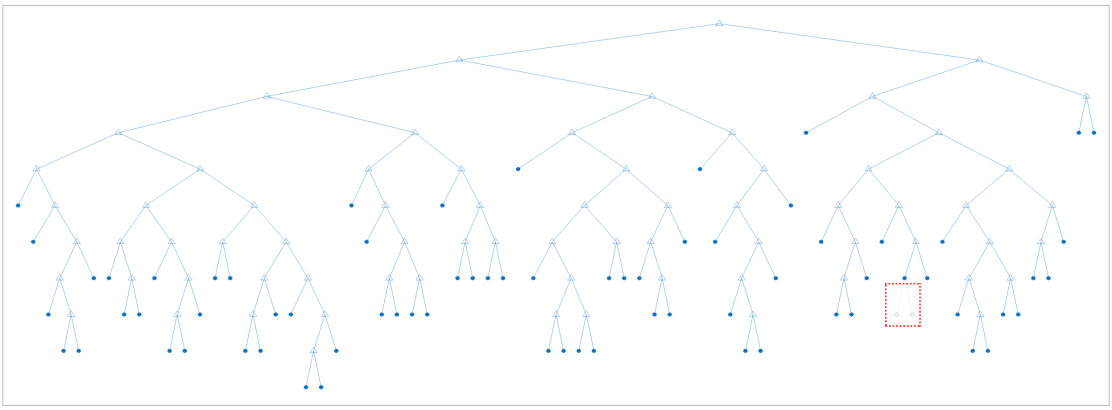
A figura 4.3 foi gerada com base em uma sequência simulada de valores de erro quadrático médio (MSE) obtidos via validação cruzada para diferentes níveis de poda em uma árvore de regressão. O gráfico mostra como o desempenho do modelo varia conforme a complexidade da árvore é reduzida, sendo o nível 0 correspondente à árvore completa, e os níveis seguintes representando modelos podados progressivamente.

O ponto destacado com um círculo vermelho e anotação "Ótimo: Nível 1"indica o nível com menor MSE na validação cruzada, isso significa que a árvore no nível 1 representa o melhor equilíbrio entre complexidade do modelo e capacidade de generalização. Árvores com mais profundidade (nível 0) podem estar superajustadas (overfitting), enquanto árvores muito podadas (níveis altos) perdem precisão (underfitting). A curva tem formato de "U" suave, o que é típico em problemas de modelagem: No início (níveis baixos), o modelo é mais ajustado e comete menos erros. Depois de certo ponto (nível 1), o erro começa a aumentar com a simplificação da árvore. Utilizar o nível de poda ótimo evita tanto o overfitting quanto o underfitting.

Observa-se que o erro inicialmente diminui conforme a árvore é podada, atingindo um ponto ótimo por volta do nível 1, e depois volta a crescer, indicando perda de capacidade preditiva. Esse comportamento é típico de modelos que enfrentam a dicotomia entre viés e variância: árvores muito grandes apresentam baixa variância, mas alto viés; árvores muito pequenas, o oposto. O nível ótimo de poda foi selecionado como aquele que apresentou o menor MSE no conjunto de validação, e a árvore foi reestruturada nesse ponto.

A árvore após processo de poda apresenta a estrutura mostrada na figura 4.4.

Figura 4.4: Arvore após a poda (Cost-Complexity Pruning)



Fonte: Elaborado pela autora (2025).

O nível ótimo de poda, com a remoção das folhas, destacado em vermelho, representa o melhor compromisso entre viés e variância, sendo esse o modelo final adotado para aplicação prática.

5 RESULTADOS

A Tabela 5.1 resume as principais métricas de desempenho do modelo após poda:

Tabela 5.1: Desempenho Preditivo do Modelo de Regressão Podado

Métrica Avaliada	**Valor Obtido**
Erro Quadrático Médio (MSE)	0,0002
Erro Absoluto Médio (MAE)	0,0027
Correlação entre SCORE real e predito	0,9960

Fonte: Elaborado pela autora (2025).

Esses valores indicam altíssima acurácia na predição do escore, com erros médios inferiores a 0,3% da escala total, o que é particularmente satisfatório em modelos de risco. Na Tabela 5.2 é possível verificar o desempenho do método em relação à outros modelos preditivos.

Tabela 5.2: Comparação de desempenho preditivo entre métodos na literatura e o modelo proposto, tomando como referência o MSE e MAE

Método / Referência	**Base de Dados**	**MSE**	**MAE**
Árvore de Regressão Podada (Este trabalho)	6.963 UCs (Brasil)	0,0002	0,0027
Random Forest — Messinis et al. (2018)	50.000 UCs (Grécia)	0,0015	0,0120
Optimum Path Forest (OPF) — Reis Filho et al. (2020)	12.000 UCs (Brasil)	0,0021	0,0150
Deep Learning Autoencoder — Hu et al. (2021)	100.000 UCs (China)	0,0009	0,0085

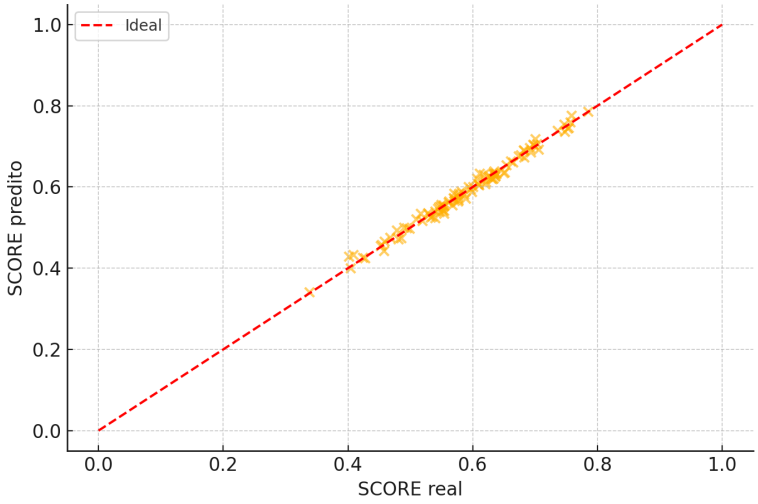
Fonte: Elaborado pela autora (2025) adaptado de [5],[**?**],[6].

Os resultados obtidos com a árvore de regressão podada apresentam valores significativamente menores que os alcançados por técnicas como Floresta

Aleatória, Floresta de Caminhos Ótimos e Autoencoders. Esse desempenho evidencia não apenas a elevada capacidade preditiva do modelo proposto, mas também sua eficiência em capturar padrões de irregularidades com maior precisão e menor erro médio, consolidando-o como uma alternativa competitiva e de fácil interpretabilidade para aplicação prática.

A figura 5.1 apresenta o gráfico de dispersão entre os escores reais e os escores previstos pelo modelo.

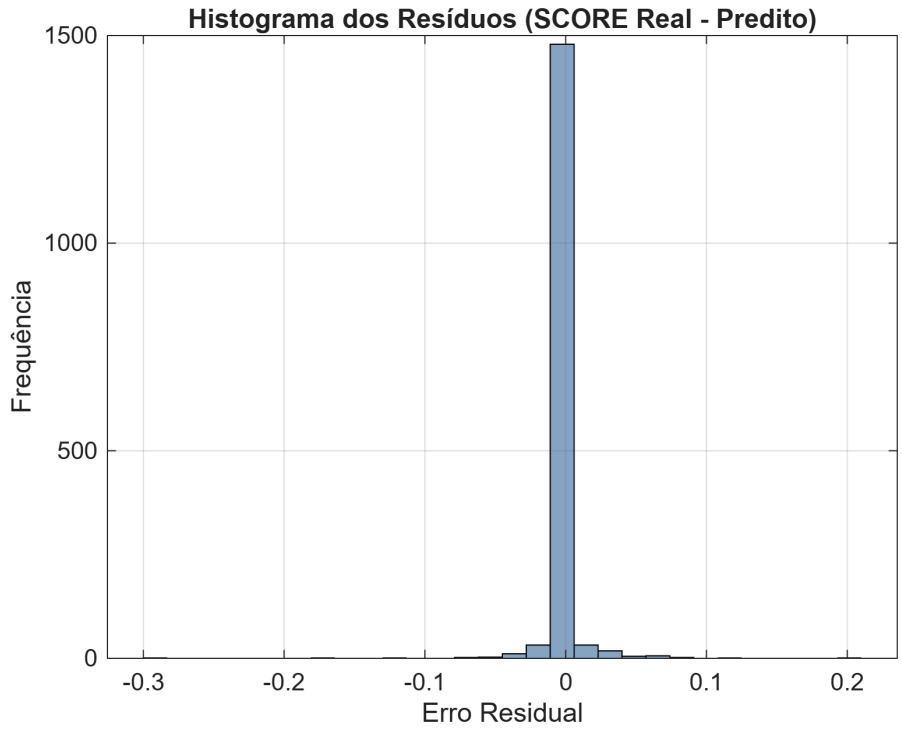
Figura 5.1: Dispersão entre SCORE real e SCORE predito pelo modelo podado



Fonte: Elaborado pela autora (2025).

O alinhamento dos pontos em torno da diagonal indica alto grau de correlação entre os valores previstos e os reais, validando a confiabilidade da árvore podada. Outro indicador relevante foi a análise dos resíduos, ou seja, as diferenças entre os escores reais e os preditos. A figura 5.2 mostra o histograma desses resíduos.

Figura 5.2: Distribuição dos resíduos (SCORE Real – Predito)



Fonte: Elaborado pela autora (2025).

A figura 5.2 apresenta a análise estatística dos resíduos obtidos após a aplicação da técnica de regressão com árvore podada. Esses resíduos são definidos como a diferença entre o SCORE real atribuído a cada instalação e o SCORE estimado pelo modelo de regressão treinado com poda por complexidade de custo. Matematicamente, têm a forma:

$$\text{Resíduo}_i = \text{SCORE}_i^{\text{real}} - \text{SCORE}_i^{\text{predito}}$$

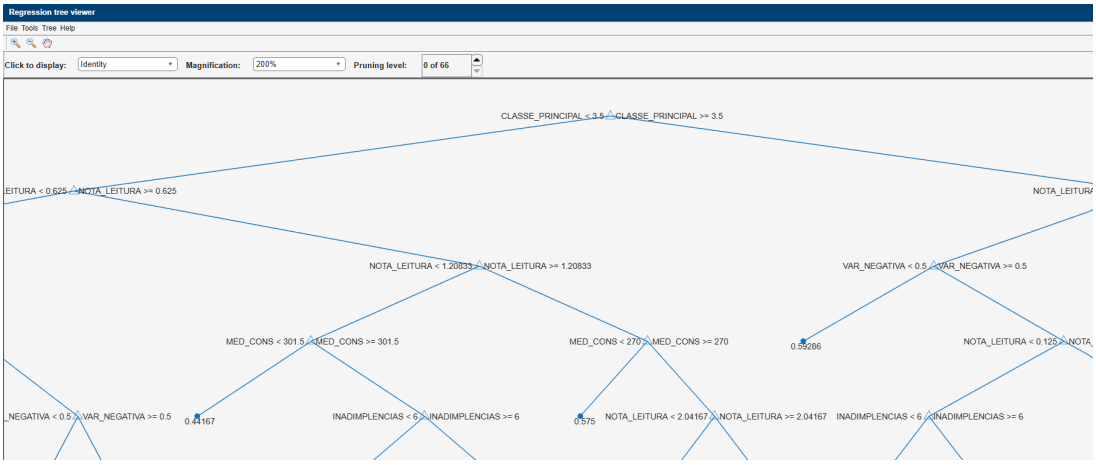
O histograma mostra que a maioria dos resíduos está concentrada em torno do valor zero, o que indica que, para a maioria das instalações, o modelo foi altamente preciso na predição do escore. O pico acentuado no valor zero sugere uma excelente aderência entre o modelo e os dados reais.

Além disso, a forma aproximadamente simétrica e levemente gaussiana da distribuição sugere que os erros são distribuídos de forma homogênea, o que é desejável em modelos de aprendizado supervisionado. A estreita dispersão reforça a qualidade das predições, enquanto a ausência de caudas longas evidencia a ausência de grupos com erros excessivamente altos. A simetria da distribuição indica que não há viés sistemático do modelo, ele não tende a superestimar nem subestimar os escores.

Em síntese, essa análise dos resíduos contribui para validar a consistência do modelo treinado, demonstrando que a estrutura podada da árvore de decisão oferece generalização adequada com baixo erro sistemático.

A visualização da árvore de regressão podada (figura 5.3) permite interpretar as regras principais utilizadas pelo modelo para realizar a seleção e definição dos alvos prioritários com base no score predito.

Figura 5.3: Estrutura visual da árvore de regressão com os primeiros nós de decisão



Fonte: Elaborado pela autora (2025).

Essa abordagem contribui significativamente para a transparência e interpretabilidade do modelo, tornando-o mais confiável e alinhado à expertise de

analistas e gestores operacionais.

A estrutura resultante é facilmente interpretável por especialistas de área, que podem traduzir cada ramo da árvore como uma regra de inspeção operacional. A Tabela 5.3 apresenta a importância relativa das variáveis no modelo final, calculada com base na redução acumulada do erro nos nós em que a variável foi usada.

Tabela 5.3: Importância relativa das variáveis no modelo de árvore de regressão após poda.

Variável	**Importância Relativa**
NOTA_LEITURA	0,4757
CLASSE_PRINCIPAL	0,4254
INADIMPLENCIAS	0,0733
MED_CONS	0,0256

Fonte: Elaborado pela autora (2025).

A situação no momento da leitura e a classe do cliente como variáveis principais é coerente com a literatura sobre perdas não-técnicas, reforçando a validade prática da árvore construída.

Por fim, o modelo foi utilizado para estimar os escores de todas as instalações da amostra. Os 100 alvos com maior escore predito foram priorizados para direcionamento das inspeções. A figura 5.4 ilutra o ranking gerado, com a instalação, o escore predito e ordenada pela classificação de prioridade.

Figura 5.4: Exibição do Matlab da lista priorizada

Top 100 Instalações Priorizadas:

INSTALACAO	CLASSE_PRINCIPAL	MED_CONS	NOTA_LEITURA	INADIMPLENCIAS	VAR_NEGATIVA	SCORE	SCORE_PREDITO
3.9462e+06	7	198	1.75	12	true	1.125	1.2417
1.013e+07	6	40996	1.3333	0	true	1.2	1.2417
3.4662e+07	4	4409	1.3333	12	true	1.3	1.2417
3.8841e+07	4	7726	1.5	0	true	1.25	1.2417
4.3749e+07	4	1645	1.3333	12	true	1.3	1.2417
2.0002e+09	4	1584	1.25	12	true	1.275	1.2417
4.001e+06	3	1248	2.25	0	true	1.175	1.225
3.9538e+07	3	1761	2.0833	12	true	1.225	1.225
4.0933e+07	3	1120	2.25	12	true	1.275	1.225
1.2055e+07	4	2217	0.83333	0	true	1.05	1.1219
3.2071e+07	5	9759	1.1667	0	true	1.15	1.1219

Fonte: Elaborado pela autora (2025).

A árvore de regressão com poda se mostra, portanto, uma ferramenta robusta, interpretável e estatisticamente eficiente para apoiar decisões de fiscalização em campo com base em dados históricos. O total de 6.963 qualificados, 5.118 possuiam consumo comprovadamente regular e 1.845 possuiam perdas não técnicas identificados em campo, o algoritmo conseguiu obter um resultado satisfatório ao identificar e priorizar somente instalações dentro do volume irregular, apresentando a distribuição de alvos conforme Tabela 5.4.

Tabela 5.4: Distribuição dos Alvos priorizados por Classe e Tipo de Irregularidade

Rótulos de Linha	**DEFEITO**	**FRAUDE**	**GESTÃO**	**Total Geral**
Comercial, Serviços e Outras A	6	1	2	9
Poder Público	3	7	0	10
Residencial	16	15	5	36
Rural	30	11	1	42
Serviço Público	1	1	1	3
Total Geral	**56**	**35**	**9**	**100**

Fonte: Elaborado pela autora (2025).

Conforme observado na Tabela 5.4 as ocorrências de perdas comerciais

foram classificadas conforme sua origem, com 56 instalações apresentando defeito técnico na medição, 35 apresentando fraudes ou irregularidades de consumo oriundas de ações de manipulação do clientes e 9 com perdas relacionadas a problemas de gestão no processo de faturamento, sejam erros de leitura ou irregularidades cadastrais.

6 CONCLUSÃO

Este trabalho demonstrou que a utilização de árvores de regressão com poda é uma abordagem eficaz, interpretável e de baixo custo computacional para apoiar a tomada de decisão na priorização de alvos para fiscalizações por perdas não técnicas. A metodologia proposta foi capaz de distinguir padrões relevantes de consumo, irregularidades e variações bruscas a partir de dados históricos, produzindo uma hierarquização robusta de alvos com alta probabilidade de perdas comerciais.

Ao sinalizar os 100 primeiros alvos priorizados, o modelo contribui para a otimização de recursos das equipes de campo, direcionando-as para pontos com maior potencial de recuperação de energia e regularização de receitas. Além disso, a construção do modelo com base em variáveis facilmente extraídas de sistemas comerciais existentes demonstra sua aplicabilidade prática imediata para distribuidoras de energia.

Outro ponto positivo reside na transparência do processo decisório, possibilitada pela visualização da árvore, que facilita a explicação dos critérios utilizados para priorização, aspecto importante em ambientes regulados. Além da versatilidade do modelo para atribuição de pesos e relevâncias às variáveis utilizadas de acordo com os interesses da distribuidora.

6.1 Trabalhos Futuros

Como próximos passos, sugere-se o aprofundamento da análise por meio das seguintes direções:

- **Incorporação de variáveis socioeconômicas e geoespaciais:** O modelo pode ser enriquecido com informações agregadas por setor censitário, como renda média, densidade populacional, índice de vulnerabilidade social, e distância de centros urbanos, aumentando a sensibilidade do modelo a contextos regionais.

- **Modelos híbridos com aprendizado profundo:** Embora as árvores de regressão sejam altamente interpretáveis, a combinação com redes neurais ou autoencoders pode capturar padrões mais complexos de comportamento, especialmente em séries temporais para os casos de clientes telemedidos.

- **Validação cruzada com amostragens rotativas em campo:** A aplicação prática do modelo pode ser acompanhada de campanhas de fiscalização rotativa para retroalimentação do modelo e ajuste contínuo de parâmetros.

- **Construção de dashboards em tempo real:** A integração com plataformas como Power BI pode permitir que a supervisão das prioridades seja feita de maneira dinâmica, por equipe ou região.

A continuidade e evolução dessa abordagem podem consolidar uma frente inteligente e estratégica de combate às perdas não técnicas, agregando valor à gestão comercial das distribuidoras e ampliando a sustentabilidade do sistema elétrico nacional.

6.2 Publicações

PEREIRA, Mayara Martins; OLIVEIRA, Isabelle Gomes de; SOUZA, Hellen Dianne Pereira de; LIMA, Shigeaki Leite. *Analysis of Non-Technical Losses in the Electric Distribution System and the Influence of Socioeconomic Factors.* XXV

Congresso Brasileiro de Automática (CBA 2024), Rio de Janeiro. Disponível em: https://www.sba.org.br/cba2024/papers/paper_3154.pdf. Acesso em: 15 set. 2025.

REFERÊNCIAS

[1] TEIXEIRA, Jordana Ramos Lino. *Perdas não técnicas: uma análise do setor de distribuição de energia elétrica no Estado de Goiás no período de 2009 a 2019*. 2020. Monografia (Graduação em Ciências Econômicas) — Pontifícia Universidade Católica de Goiás, Goiânia, 2020. Disponível em: <https://repositorio.pucgoias.edu.br/jspui/handle/123456789/887>. Acesso em: 20 ago. 2024.

[2] VIEGAS, J. L.; ESTEVE S, P. R.; MELICIO, R.; MENDES, V. M. F.; VIEIRA, S. M. Solutions for detection of non-technical losses in the electricity grid: a review. *Renewable and Sustainable Energy Reviews*, v. 80, p. 1256-1268, 2017. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1364032117308328>. Acesso em: 18 ago. 2024.

[3] Monedero, I., Biscarri, F., León, C., Guerrero, J. I., Biscarri, J., Millán, R. (2012). Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. International Journal of Electrical Power & Energy Systems, 34(1), 90-98.

[4] MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes — Fundamentos e Aplicações*, v. 1, n. 1, p. 32, 2003. Disponível em: <https://dcm.ffclrp.usp.br/ augusto/publications/2003-sistemas-inteligentes-cap4.pdf>. Acesso em: 18 nov. 2024.

[5] Messinis, G. M., & Hatziargyriou, N. D. (2018). Review of non-technical loss detection methods. Electric Power Systems Research, 158, 250-266.

[6] Hu, W., Yang, Y., Wang, J., Huang, X., & Cheng, Z. (2020, April). Understanding electricity-theft behavior via multi-source data. In Proceedings of The Web Conference 2020 (pp. 2264-2274).

[7] REIS FILHO, J. Sistema inteligente baseado em árvore de decisão, para apoio ao combate às perdas comerciais na distribuição de energia elétrica. 2006. 174 f. Dissertação (Mestrado em Engenharia Elétrica) — Universidade Federal de Uberlândia, Uberlândia, 2006. Disponível em: <https://repositorio.ufu.br/handle/123456789/14493>. Acesso em: 18 ago. 2024.

[8] Gueldini, A. V., & Santos, G. A. M. B. D. (2022). Identificação de perdas não técnicas de energia utilizando técnica de regressão baseada em boosting.

[9] Glauner, P., Boechat, A., Dolberg, L., State, R., Bettinger, F., Rangoni, Y., & Duarte, D. (2016, September). Large-scale detection of non-technical losses in imbalanced data sets. In 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT) (pp. 1-5). IEEE.

[10] Meira, J. A., Glauner, P., State, R., Valtchev, P., Dolberg, L., Bettinger, F., & Duarte, D. (2017, February). Distilling provider-independent data for general detection of non-technical losses. In 2017 IEEE Power and Energy Conference at Illinois (PECI) (pp. 1-5). IEEE.

[11] TREVIZAN, Rodrigo Daniel. *Detecção e identificação de perdas comerciais em sistemas de distribuição: metodologia baseada em floresta de caminhos ótimos*. 2014. 90 f. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2014. Disponível em:

<https://lume.ufrgs.br/bitstream/10183/118824/1/000966970.pdf>. Acesso em: 18 ago. 2024.

[12] Pereira, M. M., Oliveira, I. G., Souza, H. D. P., & Lima, S. L. (2024). *Analysis of Non-Technical Losses in the Electric Distribution System and the Influence of Socioeconomic Factors*. Anais do Congresso Brasileiro de Automática (CBA).

[13] Silva, R. A. (2023). *Regression tree models with cost-complexity pruning for non-technical loss detection in Brazilian distribution systems*. Journal of Modern Power Systems and Clean Energy.

[14] Bernardon, D. P., Comassetto, L., Canha, L. N., & Abaide, A. R. (2007). Perdas técnicas e comerciais de energia elétrica em sistemas de distribuição. In VII Conferência Brasileira de Qualidade de Energia Elétrica. Santa Maria: AGEPOC.

[15] LIMA, Fabiana Borges. *Identificação e combate às perdas comerciais em uma distribuidora de energia elétrica*. 2018. 76 f. Monografia (Graduação em Engenharia Elétrica) — Instituto de Ciências Exatas e Aplicadas, Universidade Federal de Ouro Preto, Campus João Monlevade, João Monlevade, 2018. Disponível em: `https://www.monografias.ufop.br/bitstream/35400000/1366/6/MONOGRAFIA_Identifica%C3%A7%C3%A3oCombatePerdas.pdf`. Acesso em: 18 ago. 2024.

[16] Depuru, S. S. S. R., Wang, L., & Devabhaktuni, V. (2011). Electricity theft: Overview, issues, prevention and a smart meter based approach to control theft. Energy policy, 39(2), 1007-1015.

[17] ANEEL, AN d EE. Perdas de Energia Elétrica na Distribuição. Brasília, DF. Disponível em:< https://www. aneel. gov.

br/documents/654800/18766993/Relat% C3% B3rio+ Perdas+ de+ Energia, 2019..Acesso em: 18 ago. 2024.

[18] ENERGES. Perdas comerciais e técnicas no setor elétrico: análise e métodos de mitigação. 2021. Disponível em: https://repositorio.unesp.br/server/api/core/bitstreams/46ebd6e2-8274-48c0-9d90-b82da5b65756/content. Acesso em: 18 ago. 2025.

[19] CANCIAN, Wellington Fazzi. *Análise de perdas comerciais de energia elétrica em redes de distribuição*. 2021. Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Belo Horizonte, 2021. Disponível em: https://repositorio.ufmg.br/handle/1843/45812. Acesso em: 24 abr. 2025.

[20] AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (ANEEL). Relatório de Perdas de Energia Elétrica na Distribuição. Brasília, 2023. Disponível em: https://git.aneel.gov.br/publico/centralconteudo/-/raw/main/relatorioseindicadores/tarifaeconomico/Relatorio_Perdas_Energia.pdf. Acesso em: 28 abr. 2024.

[21] EUROPEAN REGULATORS GROUP FOR ELECTRICITY AND GAS (ERGEG). *Treatment of losses by network operators: ERGEG Position Paper for public consultation*. Bruxelas, 2008. Disponível em: https://www.ceer.eu/wp-content/uploads/2008/01/C07-WPDC-10-03_WP2008-public_17-Jan-08.pdf>. Acesso em: 10 ago. 2024.

[22] LOUW, Quentin E. The Impact of Non-Technical Losses: A South African Perspective Compared to Global Trends. South African Revenue Protection Association (SARPA) Conference Paper, 2019.

[23] CANCIAN, Wellington Fazzi. *Metodologia para identificação e clusterização espacial de perdas não-técnicas em sistemas de distribuição de energia elétrica*. 2013. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Minas Gerais, Belo Horizonte, 2013. Disponível em: https://repositorio.ufmg.br/handle/1843/BUBD-9H7HP3. Acesso em: 22 abr. 2024.

[24] POVEDA, M. A New Method to Calculate Power Distribution Losses in an Environment of High Unregistered Loads. New Orleans: IEEE Transmission and Distribution Conference, 1999.

[25] CONFERÊNCIA DAS NAÇÕES UNIDAS SOBRE COMÉRCIO E DESENVOLVIMENTO (UNCTAD). *Reconhecimento dos Países Menos Desenvolvidos*. Genebra, 2023. Disponível em: <https://digitallibrary.un.org/record/4041784?ln=en>. Acesso em: 18 abr. 2024.

[26] CARR, D.; THOMSON, M. Non-Technical Electricity Losses. *Energies*, v. 15, p. 2218, 2022. Disponível em: <https://www.mdpi.com/1996-1073/15/6/2218>. Acesso em: 17 ago. 2024.

[27] AGÊNCIA INTERNACIONAL DE ENERGIA (IEA). *Electricity Market Report – December 2020*. Paris, 2020. Disponível em: <https://iea.blob.core.windows.net/assets/a695ae98-cec1-43ce-9cab-c37bb0143a05/Electricity_Market_Report_December_2020.pdf>. Acesso em: 17 ago. 2024.

[28] BAFFI, E.; LAMAISON URIOSTE, R. M.; ARAGÜÉS PEÑALBA, M. Potential benefits of distributed generation in the reduction of non-technical losses. *Renewable Energy and Power Quality Journal*, v. 1, p. 39–44, 2018.

[29] CHAVES, A. C. et al. *As perdas não técnicas no setor de distribuição brasileiro: uma abordagem regulatória.* São Paulo: D7 Editora, 2019.

[30] ANEEL. Perdas de energia elétrica na distribuição. 2023. Disponível em:https://git.aneel.gov.br/publico/centralconteudo/-/raw/main/relatorioseindicadores/tarifaeconomico/Relatorio_Perdas_Energia.pdf. Acesso em: 15 abr. 2024.

[31] ANEEL. Relatório de perdas de energia elétrica. 2024. Disponível em: https://portalrelatorios.aneel.gov.br/luznatarifa/perdasenergias. Acesso em: 19 abr. 2024.

[32] TRIBUNAL DE CONTAS DA UNIÃO (TCU). *Sustentabilidade tarifária de energia elétrica: lista de alto risco da Administração Pública Federal*. Brasília, 2022. Disponível em:<https://static.poder360.com.br/2022/11/tcu-lista_de_alto_risco_da_administracao_publica-16-11-22.pdf>. Acesso em: 16 ago. 2024.

[33] ANEEL. Resolução Normativa nº 1.003, de 28 de março de 2022. Estabelece a metodologia do Submódulo 2.6 do PRORET para definição das metas de perdas não técnicas. Disponível em:https://www2.aneel.gov.br/cedoc/ren20221003.pdf. Acesso em: 19 abr. 2024.

[34] IBGE. *Síntese de indicadores sociais: uma análise das condições de vida da população brasileira*. Rio de Janeiro, 2023. Disponível em: <https://static.poder360.com.br/2023/12/sis-ibge-2023.pdf>. Acesso em: 18 ago. 2025.

[35] SERASA. *Perfil e comportamento do endividamento brasileiro 2022*. São Paulo, 2022. Disponível em: https://cdn.builder.io/o/assets%2Fb212bb18f00a40869a6cd42f77cbeefc%2F3737e87997744fea99f21146c9647091?alt=media&

token=0a8ba1e9-f983-4fba-8a35-789113b1da81&apiKey=
b212bb18f00a40869a6cd42f77cbeefc. Acesso em: 22 abr. 2024.

[36] ABRACE ENERGIA, "Brasil tem a conta de luz que mais pesa no bolso da população entre 34 países," *Associação Brasileira dos Grandes Consumidores de Energia*, 2023. [Online]. Available: https://git.aneel.gov.br/publico/centralconteudo/-/raw/main/relatorioseindicadores/tarifaeconomico/Relatorio_Perdas_Energia.pdf. Acesso em: 22 abr. 2024.

[37] INSTITUTO ACENDE. *Perdas comerciais e inadimplência no setor elétrico*. White Paper nº 18, 2017. Disponível em: <https://acendebrasil.com.br/wp-content/uploads/2020/04/2017_WhitePaperAcendeBrasil_18_PerdasInadimplencias.pdf>. Acesso em: 17 ago. 2024.

[38] Agência Nacional de Energia Elétrica - ANEEL, *Relatórios sobre perdas não técnicas*, 2023. Disponível em: https://www.aneel.gov.br/. Acesso em: 17 ago. 2024.

[39] FRANÇA, J. et al. *Perdas comerciais e inadimplência no setor elétrico*. Instituto ACENDE Brasil, White Paper nº 18, 2017. Disponível em: <https://acendebrasil.com.br/wp-content/uploads/2020/04/2017_WhitePaperAcendeBrasil_18_PerdasInadimplencias.pdf>. Acesso em: 16 ago. 2024.

[40] BARROS, Rafael M. R. *Gestão da Perda Não Técnica de Energia Elétrica*. Brasília: Interciência, 2024. ISBN 6589367736. Disponível em: <https://books.google.com.br/books/about/Gest%C3%A3o_da_Perda_N%C3%A3o_T%C3%A9cnica_de_Energi.html?id=B9ceEQAAQBAJ>. Acesso em: 18 jul. 2024.

[41] Oliveira, M., *Análise de perdas não técnicas em distribuidoras brasileiras*, Revista Brasileira de Energia, vol. 22, n. 3, pp. 45–60, 2016.

[42] Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. Wadsworth International Group.

[43] CASTANHEIRA, L. G. Luciana Gomes. *Aplicação de técnicas de mineração de dados em problemas de classificação de padrões*. Dissertação de Mestrado, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte – MG, 2008. Disponível em: `https://repositorio.ufmg.br/bitstream/1843/BUOS-8CDFQK/1/luciana_gomes_castanheira.pdf`.Acesso em: 15 abr. 2024.

[44] HUNT, E. B.; MARIN, J.; STONE, P. J. Experiments in Induction. New York: Academic Press, 1966.

[45] Quinlan, J. R. (1986). Induction of Decision Trees. In Machine Learning (pp. 81–106). Springer.

[46] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

[47] Seidman, D. (2001). Data Mining with Microsoft SQL Server 2000. Microsoft SQL Server Technical Article. Microsoft Corporation.

[48] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media.

[49] Russell, S., & Norvig, P. (2003). Artificial Intelligence: A Modern Approach. Prentice Hall.

[50] PRASS, M. S. *Árvores de Decisão e suas Aplicações*. 2009. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

[51] Souza, Cleber Batista. *Árvores de Decisão: A Evolução do CART ao BART*. Dissertação de Mestrado, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, Brasil, 2021. Disponível em:`https://www.teses.usp.br/teses/disponiveis/45/45133/tde-05042022-095004/`. Acesso em: 17 de janeiro de 2025.

[52] G. James, D. Witten, T. Hastie, and R. Tibshirani.*An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

[53] S. Geman, E. Bienenstock, and R. Doursat. "Neural networks and the bias/variance dilemma."*Neural Computation*, vol. 4, no. 1, pp. 1-58, 1992.

[54] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer, 2013.

[55] J. H. Friedman. "Greedy function approximation: A gradient boosting machine."*Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

[56] ZHANG, H.; CISSE, M.; DAUPHIN, Y. N.; LOPEZ-PAZ, D. "mixup: Beyond empirical risk minimization". *arXiv preprint arXiv:1710.09412*, 2017. Disponível em: `<https://arxiv.org/abs/1710.09412>`. Acesso em: 18 ago. 2024.

[57] Glauner, P. et al., *The challenge of non-technical loss detection using artificial intelligence: A survey*, International Journal of Computational Intelligence Systems, vol. 10, no. 1, pp. 760–775, 2017.

[58] MathWorks. (2024). fitrtree documentation. Disponível em: `https://www.mathworks.com/help/stats/fitrtree.html` Acesso em: 5 de janeiro de 2025.