UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

**JOSUÉ JONATHAN BORGES DE OLIVEIRA**

# Análise Comparativa de Modelos de Machine Learning para Sugestão de Inspeções em Clientes de Distribuidoras de Energia Elétrica

NITERÓI
2023

UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

**JOSUÉ JONATHAN BORGES DE OLIVEIRA**

# Análise Comparativa de Modelos de Machine Learning para Sugestão de Inspeções em Clientes de Distribuidoras de Energia Elétrica

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações.

Orientador:
D. Sc. Vitor Hugo Ferreira

NITERÓI
2023

Espaço reservado para a ficha catalográfica.

JOSUÉ JONATHAN BORGES DE OLIVEIRA

# Análise Comparativa de Modelos de Machine Learning para Sugestão de Inspeções em Clientes de Distribuidoras de Energia Elétrica

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Energia Elétrica.

BANCA EXAMINADORA

gov.br Documento assinado digitalmente
VITOR HUGO FERREIRA
Data: 19/12/2023 20:36:47-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Vitor Hugo Ferreira, D.Sc. - Orientador

Universidade Federal Fluminense - UFF

gov.br Documento assinado digitalmente
HENRIQUE DE OLIVEIRA HENRIQUES
Data: 19/12/2023 21:12:50-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Henrique de Oliveira Henriques - D.Sc.

Universidade Federal Fluminense - UFF

gov.br Documento assinado digitalmente
LEONARDO WILLER DE OLIVEIRA
Data: 19/12/2023 21:32:43-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Leonardo Willer de Oliveira - D.Sc.

Universidade Federal Juiz de Fora - UFJF

Niterói

Dezembro de 2023

*Dedico à minha família.*

# Agradecimentos

Agradeço à Deus pela força e coragem, a minha família pelo apoio e ao meu orientador Vitor Hugo Ferreira por todo o suporte.

Por fim, agradeço a UFF e a PPGEET pela oportunidade e pelo sonho realizado.

# Resumo

Atualmente, com o aumento das perdas não técnicas nas distribuidoras de energia elétrica, as atividades de análise e combate a esse grande ofensor se tornaram cada vez mais essenciais para o setor de distribuição. Os estudos referentes a mitigação das perdas não técnicas têm sido amplamente debatidos na literatura. Com o avanço do poder computacional, cada vez mais autores propõe a utilização de técnicas de inteligência artificial para a identificação de clientes fraudadores. Neste contexto, o presente trabalho possui como objetivo o desenvolvimento de um modelo utilizando técnicas de inteligência artificial, mais precisamente métodos de aprendizagem supervisionada, que irão propor melhorias no processo de seleção de possíveis fraudadores de forma a contribuir para uma maior assertividade e tomada de decisão em campo. O modelo utilizará dados pertencentes à distribuidora de energia elétrica Light S.A. e contribuições de especialistas da área de perdas, assegurando critérios essenciais para uma correta análise e direcionamento das ações. Para isso, foram realizados testes com diferentes algoritmos classificadores e métodos de extração de características, cuja finalidade é assegurar o melhor método de aplicação. Por fim, os resultados alcançados serão avaliados por técnicas apropriadas seguindo as melhores práticas da literatura.

**Palavras-chaves:** aprendizagem supervisionada, algoritmos de classificação, distribuidora de energia elétrica, sistema de distribuição, identificação, fraude, perdas não-técnicas.

# Abstract

Nowadays, with the increasing non-technical losses in electric power distribution companies, activities related to the analysis and combat of this significant offender have become increasingly essential for the distribution sector. In view of this, studies regarding the mitigation of non-technical losses have been extensively debated in the literature. With the advancement of computational power, more authors propose the use of artificial intelligence techniques for identifying fraudulent customers. In this context, this work aims to develop a model using artificial intelligence techniques, specifically supervised learning methods, to propose improvements in the selection process of potential fraudsters, thereby contributing to greater accuracy and decision-making in the field. The model will use data from the electric power distributor Light S.A. and input from loss area experts, ensuring essential criteria for proper analysis and action direction. To achieve this, tests were conducted with different classifier algorithms and feature extraction methods, aiming to ensure the best applicable method. Finally, the achieved results will be evaluated using appropriate techniques following the best practices in the literature.

**Keywords:** supervised learning, classification algorithms, electricity distributors, system distribution, feature extraction methods, identification, fraud reporting.

# Lista de figuras

# Lista de Tabelas

# Lista de Abreviaturas e Siglas

| | |
|---|---|
| AF | Acerto Fraude |
| AR | Acerto Regular |
| CIRED | *International Conference on Eletricity Distribution* |
| FR | Fraude |
| IA | Inteligência Artificial |
| INSP | Inspecionado |
| IR | Irregularidade |
| ML | *Machine Learning* |
| NA | Nada Apurado |
| NI | Não Inspecionado |
| NTL | *Non Technicals Loss* |
| RF | *Random Forest* |
| RNA | Redes Neural Artificial |
| SVM | *Support Vector Machine* |
| TL | *Technicals Loss* |
| VIS | Visitado |

# Sumário

# Capítulo 1 - Introdução

O Brasil tem enfrentado uma crise econômica prolongada que se agravou ainda mais com a pandemia. A situação de desigualdade social tem se acentuado devido ao aumento do desemprego e à redução dos rendimentos reais da população, gerando preocupações em diversos setores, inclusive no setor elétrico. Em consequência disso, o setor de distribuição de energia elétrica vivencia uma adversidade desafiadora, que é o aumento das perdas não técnicas de eletricidade (PNT), além da redução da demanda e do aumento do nível de inadimplência.

O estudo realizado pela ANEEL em 2020 [30] destaca as perdas não técnicas de energia no setor elétrico. Cita que esse problema afeta principalmente as empresas distribuidoras de energia. Nesse mesmo ano, essas perdas representaram um impacto significativo de 37,9 terawatts-hora (TWh). Sua origem está relacionada principalmente com irregularidades oriundas de furtos (ligação clandestina, desvio direto da rede) ou fraude de energia (adulterações no medidor). O setor elétrico juntamente com pesquisadores vem buscando e desenvolvendo diversos métodos para localizar tais irregularidades e reverter esse problema.

Diferentes trabalhos como [10] e [31] revelaram algumas possibilidades de utilização de tecnologias para se combater este grande ofensor. Dentre as mencionadas por eles está a inserção ao *Smart Grid*, um conceito inovador que tem ganhado destaque mundialmente. O principal motivo por trás desse movimento é a necessidade de lidar eficientemente com os desafios decorrentes das perdas não técnicas, otimizando a gestão da distribuição de energia. Além disso, a busca pela modernização através do conceito de *Smart Grid* não apenas representa um avanço tecnológico, mas também uma estratégia fundamental para enfrentar os desafios crescentes no setor de distribuição de energia elétrica.

O progresso notável na tecnologia e na capacidade computacional tem catalisado a integração da inteligência artificial (IA) em variados domínios, capacitando soluções tecnológicas a realizar tarefas previamente reservadas à inteligência humana, tudo isso viabilizado pelas técnicas de IA. Em particular, o aprendizado de máquina, uma abordagem dentro da inteligência artificial, merece destaque, caracterizando-se pela capacidade dos sistemas em aprimorar seu desempenho automaticamente a partir da experiência adquirida.

Na literatura científica, devido à relevância do tema, vários trabalhos têm sido publicados empregando diversas técnicas da área de IA tais como as Redes Neurais Artificiais (RNA) [17], *Deep Learning* [26] e Máquinas de Vetor Suporte [13]. Essas abordagens representam uma ampla gama de ferramentas utilizadas para abordar o problema, enfatizando

a importância desse campo de estudo. Além disso, esses modelos têm a capacidade de oferecer abordagens mais precisas e eficientes para analisar dados complexos, o que pode levar a melhorias significativas na gestão e eficiência do setor energético.

Um exemplo disso é o estudo realizado por [5], que utilizou técnicas de aprendizado de máquina para detectar fraudes e anomalias a partir de dados coletados por medidores inteligentes. Especificamente, o estudo empregou o classificador *XGBoost* para a análise dos dados e identificação de possíveis fraudes ou irregularidades, evidenciando o potencial do aprendizado de máquina no setor de energia elétrica. Com o avanço das técnicas de aprendizado de máquina, é importante destacar que sua utilização tem se mostrado uma ferramenta valiosa no apoio à redução das perdas não técnicas, especialmente no setor de distribuição de energia elétrica. As fraudes e irregularidades são um dos principais responsáveis por essas perdas, especialmente em áreas carentes. Portanto, o uso de técnicas de aprendizado de máquina, aliado à coleta de dados por meio de medidores inteligentes, pode fornecer informações importantes para a detecção precoce dessas anomalias, permitindo que as distribuidoras atuem de forma mais eficaz na prevenção de perdas não técnicas.

No presente estudo, buscou-se a comparação de diversos algoritmos de aprendizado de máquina com o propósito de apoiar a detecção de potenciais fraudadores na área de distribuição de energia elétrica. Testes foram conduzidos com métodos de aprendizagem supervisionada e uma variedade de algoritmos de classificação utilizando a linguagem de programação Python. O objetivo foi alcançar um modelo representativo que pudesse melhorar a assertividade na identificação e localização de possíveis causadores de perdas comerciais.

## 1.1 Motivações e Objetivos

O Brasil enfrenta um grande desafio em relação às perdas comerciais de energia elétrica. Estima-se que as irregularidades representem um prejuízo médio total de R$ 5 bilhões por ano, o que corresponde a cerca de 5% da energia total consumida no país [1]. Essas perdas impactam diretamente o setor elétrico, causando prejuízos financeiros para as empresas e para os consumidores regulares [9].

Entre as empresas que sofrem com esse problema, destaca-se a Light S.A., que atende uma média de 7 milhões de unidades consumidoras em 31 municípios do Rio de Janeiro. Essa distribuidora enfrenta grandes desafios em relação às perdas comerciais, que impactam

negativamente em sua capacidade de investimento e no desenvolvimento de projetos de melhoria da qualidade do serviço prestado aos consumidores [29].

Além dos prejuízos financeiros provocados pelas perdas comerciais, a inadimplência resultante dos furtos de energia elétrica repercute significativamente sobre as tarifas impostas aos consumidores regulares. Isso ocorre em razão de as empresas fornecedoras serem compelidas a absorver os custos associados à produção e distribuição da energia usurpada, custos esses que, inevitavelmente, são redistribuídos entre os consumidores legítimos por meio do incremento das tarifas.

Para enfrentar esse desafio, é fundamental que o setor elétrico brasileiro invista em soluções tecnológicas e em políticas públicas eficientes para combater as perdas comerciais de energia elétrica. É necessário que sejam desenvolvidas medidas de fiscalização mais efetivas e que sejam aplicadas punições mais rigorosas para os consumidores que praticam furtos de energia elétrica. Além disso, a conscientização da população sobre a importância de pagar corretamente suas contas de energia também é um fator fundamental para reduzir as perdas comerciais no setor elétrico.

Dada a complexidade do tema, a referência [29] apresenta um estudo de perdas não técnicas realizado pela distribuidora Light S.A. que mapeou geograficamente as áreas onde há maior índice de perdas. Este estudo concluiu que existe uma enorme relação entre o nível de perdas não técnicas da empresa e as peculiaridades do Rio de Janeiro. Com objetivo de combater esse grande ofensor, a instituição estuda meios para mitigar o alto valor de perda identificado. Para tanto, nota-se que uma maneira objetiva de conseguir esses resultados baseia-se nos apontamentos de clientes suspeitos de fraudes.

O modelo de identificação de clientes com possíveis irregularidades atualmente praticado na empresa se baseia em técnicas computacionais para detecção de possíveis clientes fraudadores, tendo por base regras heurísticas construídas a partir da experiência de cada operador. Por meio deste processo, os indicadores da empresa mostraram que o resultado médio de acerto obtido na identificação de clientes irregulares no ano de 2020 foi inferior a 50%. Dessa forma, verifica-se que o procedimento atualmente praticado apresenta uma oportunidade de melhoria no que diz respeito à sua taxa de assertividade na identificação do cliente com fraude.

Assim, o objetivo geral desta dissertação consiste em apresentar como a aplicação de *Machine Learning* pode contribuir durante uma análise técnica para a seleção de clientes a serem inspecionados (alvos) e direcionados para as ações de campo com objetivo de mitigar este grande ofensor.

Os objetivos específicos terão como foco principal a análise de dados e a identificação de padrões e tendências que possam auxiliar no combate às perdas. É fundamental que sejam identificadas as principais causas das perdas não técnicas e que sejam propostas medidas efetivas para reduzi-las.

Assim, nesta dissertação serão aplicadas técnicas de aprendizado de máquina para identificar possíveis usuários com indícios de irregularidades no consumo de energia elétrica. Para alcançar esse objetivo, serão desenvolvidos diferentes modelos utilizados em aprendizado de máquina e os principais algoritmos considerados na construção da metodologia.

Como insumo, serão utilizados dados da distribuidora de energia Light S.A para estabelecer um cenário de comparação, classificação e identificação de possíveis usuários com indícios de irregularidades. Para isso, será utilizado um modelo de aprendizagem supervisionada, que permitirá comparar diferentes abordagens a respeito do processamento e pré-processamento aplicado, cuja finalidade está em gerar um apontamento final classificatório assertivo.

No segmento dedicado aos resultados, serão expostas as simulações efetuadas com o intuito de validar o modelo proposto. Os resultados, por sua vez, serão apresentados de forma gráfica e tabular, possibilitando uma análise minuciosa das conclusões alcançadas. Destacar-se-ão as principais conclusões decorrentes do estudo, assim como as implicações práticas pertinentes tanto para a Light S.A quanto para o setor elétrico de modo abrangente.

Com a utilização de técnicas de aprendizado de máquina, espera-se que o estudo auxilie no aumento da assertividade e tomada de decisão em campo, contribuindo com a distribuidora de energia elétrica no sentido da adoção de medidas efetivas para reduzir as perdas comerciais e garantir um serviço de qualidade aos seus consumidores.

## 1.2 Justificativa

A pesquisa proposta tem como justificativa a sua possível contribuição para o meio acadêmico e para a sociedade em geral. Por meio da contextualização do problema das perdas comerciais de energia elétrica e da aplicação de técnicas de aprendizado de máquina na identificação de possíveis usuários com indícios de irregularidades, espera-se enriquecer e agregar ao conhecimento existente sobre o tema.

Além disso, a pesquisa tem potencial para ser uma fonte de contribuição para o meio social, por meio da disponibilização de um material coeso e estruturado que possa ser compreendido por leitores não especialistas no assunto.

## 1.3 Produções Científicas Oriundas do Trabalho

Atualmente, um dos maiores ofensores encontrados pelas distribuidoras de energia elétrica, são sem dúvida as perdas não técnicas, oriundas principalmente pelas inúmeras ligações irregulares e pelo grande volume de clandestinos ligados na rede de baixa tensão. Em função disso, este assunto tem sido amplamente debatido na literatura e com o avanço do poder computacional, cada vez mais autores propõe a utilização de técnicas de inteligência artificial para a identificação e seleção de clientes suspeitos de fraudes. Neste contexto, este trabalho apresenta o desenvolvimento de um modelo em *Machine Learning* através do uso do algoritmo de classificação *Random Forest,* capaz de interpretar corretamente os elementos presentes nos dados de uma determinada distribuidora de energia elétrica, indicando alvos suspeitos de fraudes e possibilitando assim, uma melhor eficiência no processo de combate e recuperação de energia.

O fenômeno das perdas não técnicas constitui um desafio persistente para as distribuidoras de energia elétrica, impactando diretamente na eficiência operacional e na sustentabilidade financeira dessas empresas, além de acarretar implicações para a equidade tarifária entre os consumidores.

Oliveira e Ferreira (2022) contribuem para o avanço do conhecimento e da prática no combate às fraudes em distribuidoras de energia, ao propor a aplicação de técnicas de inteligência artificial, mais especificamente, através do desenvolvimento de um modelo de *Machine Learning* baseado no algoritmo de classificação *Random Forest*. Este modelo é capaz de analisar e interpretar os dados fornecidos por uma distribuidora de energia, identificando potenciais fraudadores com uma eficiência notável. Trata-se de uma abordagem que representa um avanço significativo em relação aos métodos tradicionais, permitindo uma seleção mais precisa e eficiente de casos suspeitos para investigação. Assim, o uso do algoritmo *Random Forest,* dentro do contexto de detecção de fraudes em sistemas de distribuição de energia elétrica, destaca-se por sua capacidade de manejar grandes volumes de dados e sua eficácia na classificação de informações complexas, facilitando a identificação de padrões que indicam possíveis fraudes.

## 1.4 Estrutura do Documento

Este documento está estruturado da seguinte forma. No capítulo dois será apresentada a fundamentação teórica sobre os métodos aplicados nessa dissertação. Nesse capítulo ainda será vista uma breve introdução sobre os efeitos oriundos das perdas não técnicas, bem como as tecnologias desenvolvidas para mitigar esse grande ofensor. Por fim do capítulo, serão apresentadas as principais categorias dos diferentes tipos de aprendizado dentre os algoritmos mais utilizados em *Machine Learning*.

O capítulo três apresenta informações inerentes à construção do programa desenvolvido, sobre os dados utilizados para o desenvolvimento do modelo e ao final serão abordadas as técnicas estatísticas utilizadas para avaliação dos resultados alcançados.

No capítulo quatro, serão apresentados e discutidos os resultados obtidos durante a aplicação do modelo proposto nesta dissertação. Serão avaliados três diferentes conjuntos de dados, representados por uma base de 80 mil instalações com e sem apontamentos de indícios de fraude, a fim de verificar a assertividade do estudo.

A análise dos resultados será realizada de forma minuciosa, buscando identificar os pontos fortes e fracos do modelo proposto. Será avaliada a eficácia da abordagem de processamento e pré-processamento utilizada, bem como a precisão na identificação de possíveis usuários com indícios de irregularidades no consumo de energia elétrica.

A avaliação da assertividade do estudo será realizada por meio de métricas de desempenho, que permitirão uma análise quantitativa dos resultados. Serão apresentados gráficos e tabelas que facilitarão a compreensão dos dados e permitirão uma interpretação mais clara dos resultados.

Dessa forma, o capítulo quatro representa uma importante etapa na pesquisa, pois é a partir da análise dos resultados que será possível verificar a eficácia do modelo proposto e identificar possíveis melhorias e ajustes para sua aplicação prática no combate às perdas comerciais de energia elétrica.

Por fim, no capítulo cinco são apresentadas as conclusões do presente trabalho e os desafios futuros propostos.

# Capítulo 2 – Fundamentação Teórica

No capítulo dedicado à fundamentação teórica, é essencial a abordagem dos conceitos-chave e dos métodos aplicados ao estudo, estabelecendo um alicerce para a compreensão da pesquisa realizada. A distinção entre os diversos tipos de perdas enfrentadas pelas distribuidoras de energia elétrica constitui o ponto de partida dessa exploração teórica, sublinhando não apenas a natureza dessas perdas, mas também seu impacto operacional e econômico no setor elétrico.

Assim, a dissertação desdobra-se na exploração das categorias dos paradigmas de *machine learning* empregados no estudo, delineando um panorama das técnicas de inteligência artificial aplicadas. Nesse sentido, damos ênfase na técnica do *Random Forest*, em particular, merece destaque, dada a sua centralidade na construção do modelo de classificação para a identificação de usuários com indícios de irregularidades no consumo de energia elétrica. A escolha desse algoritmo reflete um julgamento metodológico significativo, tendo em vista suas propriedades de eficiência e a capacidade de manejar grandes volumes de dados com alta dimensionalidade, características essenciais para o tratamento de dados complexos como os encontrados nas distribuidoras de energia elétrica.

Além disso, a comparação com outros algoritmos de aprendizado de máquina, tais como *SVM (Support Vector Machines) e XGBoost (eXtreme Gradient Boosting)* são abordados. Este é um aspecto que remete à importância de estudos comparativos entre diferentes técnicas, visando não apenas à identificação da ferramenta eficazes para o contexto específico da detecção de fraudes em sistemas elétricos, mas também ao enriquecimento do diálogo acadêmico sobre as vantagens e limitações de cada método. Essa comparação entre diferentes algoritmos é capaz de fornecer concepções sobre a adaptabilidade e performance das técnicas em cenários variados, contribuindo para um entendimento matizado das capacidades e potenciais de aplicação de cada abordagem.

Portanto, é imperativo não apenas apresentar e discutir os conceitos e técnicas fundamentais, mas também refletir criticamente sobre os métodos utilizados, o que inclui uma consideração cuidadosa das razões por trás da seleção de ferramentas específicas, como o *Random Forest*, e a importância de uma análise comparativa entre diferentes métodos para uma avaliação abrangente das opções disponíveis no campo do *machine learning* aplicado à detecção de fraudes em sistemas de distribuição de energia elétrica.

## 2.1 Classificação das Perdas de Energia

Conforme previamente abordado, as perdas de energia elétrica emergem como um considerável desafio para as distribuidoras de energia, acarretando prejuízos financeiros e uma redução na eficiência operacional. Tais perdas manifestam-se em todas as fases do ciclo de conversão de energia, desde a geração até a distribuição.

Dentro do arcabouço legal do setor elétrico brasileiro, as perdas são categorizadas de maneira precisa e influenciam diretamente a forma como são reconhecidas na tarifa. Primeiramente, existem as perdas totais, que representam a diferença entre a energia injetada na rede pelas usinas geradoras e a energia efetivamente faturada aos consumidores.

As perdas totais são subdivididas em duas principais categorias: técnicas e não técnicas:

As perdas técnicas são aquelas que ocorrem devido às características intrínsecas do sistema elétrico, como resistência dos cabos e transformadores, e são inevitáveis em qualquer sistema de distribuição. Elas são reconhecidas na tarifa e, portanto, refletem diretamente no custo da energia elétrica para o consumidor. Por outro lado, as perdas não técnicas englobam as perdas comerciais e são resultado de fatores como fraudes, erros de medição, faturamento incorreto e outros problemas operacionais e comerciais. Estas também são reconhecidas na tarifa, mas geralmente são alvo de ações regulatórias para minimizá-las.

Além dessas categorias, existe a classificação das perdas regulatórias, que são as perdas impostas pelo próprio arcabouço legal e regulatório do setor elétrico. Isso inclui perdas associadas a políticas públicas, subsídios cruzados e outros mecanismos de custeio do setor elétrico. Estas também afetam a tarifa, mas seu reconhecimento é feito de maneira específica e transparente, como parte da estrutura tarifária.

Portanto, no contexto do setor elétrico brasileiro, é fundamental compreender a diferenciação entre perdas totais, técnicas, não técnicas e regulatórias, uma vez que cada uma delas desempenha um papel crucial na determinação dos custos e tarifas de energia elétrica, impactando diretamente tanto as concessionárias quanto os consumidores.

Embora as perdas sejam inevitáveis em qualquer processo de conversão de energia, é possível adotar medidas para reduzi-las e aumentar a eficiência do sistema elétrico. É nesse contexto que se insere a pesquisa proposta neste trabalho, que busca identificar possíveis usuários com indícios de irregularidades no consumo de energia elétrica por meio da aplicação de técnicas de aprendizado de máquina.

Segundo a ANEEL em [30], o custo das perdas técnicas obtido pela multiplicação dos montantes pelo preço médio da energia nos processos tarifários sem considerar tributos, é da ordem de R$ 8,5 bilhões. Em contrapartida as perdas não técnicas regulatórias no país, representaram um custo de aproximadamente R$ 5,6 bilhões ao ano.

As perdas não técnicas regulatórias, no contexto do setor de energia elétrica, referem-se às perdas de energia que ocorrem devido a fatores não relacionados à operação física dos componentes do sistema elétrico, mas sim a questões regulatórias, comerciais ou administrativas. Isso pode incluir problemas como erros de medição, faturamento incorreto, falhas no processo de cobrança e outras irregularidades que afetam o registro e a cobrança adequada do consumo de energia.

O entendimento e a mitigação das perdas não técnicas regulatórias são cruciais para garantir a eficiência e a sustentabilidade das operações no setor elétrico.

### 2.1.1 Perdas Técnicas

No contexto da distribuição de energia elétrica, as perdas técnicas são consideradas inevitáveis, uma vez que são decorrentes da dissipação de energia em condutores e equipamentos utilizados nas linhas de transmissão e distribuição, bem como das perdas magnéticas em transformadores. Essas perdas são intrínsecas ao funcionamento da distribuidora e são originadas por fatores físicos, tais como a resistência elétrica dos condutores e as perdas no núcleo e nos enrolamentos dos transformadores, além de outras características da rede e do modo de operação. Por essa razão, é necessário considerar as perdas técnicas como parte integrante do processo de distribuição de energia elétrica [35].

Embora sejam consideradas inevitáveis, as perdas técnicas podem ser minimizadas por meio da adoção de estratégias que visam a melhoria da eficiência do sistema elétrico, como a utilização de materiais mais eficientes na construção das redes elétricas e a modernização dos equipamentos utilizados na distribuição de energia elétrica. A maior quantidade de perdas técnicas em um sistema de energia está nas linhas de distribuição primária e secundária, enquanto as linhas de transmissão respondem por cerca de 30% das perdas totais [31].

No que diz respeito às diretrizes para a classificação das perdas técnicas, é possível distinguir dois tipos: as perdas fixas e as perdas variáveis [33]. As perdas fixas são constantes e não dependem do fluxo de energia na rede, sendo causadas principalmente pela resistência elétrica dos condutores, pelas perdas magnéticas nos transformadores e pelos equipamentos

mais antigos e menos eficientes utilizados na distribuição de energia elétrica. Essas perdas só podem ser reduzidas por meio da substituição desses ativos por equipamentos de maior eficiência.

Por outro lado, as perdas variáveis mudam de acordo com o fluxo de potência na rede elétrica, sendo influenciadas pela demanda de energia e pelo carregamento das linhas. Em redes altamente carregadas, essas perdas podem ser significativamente maiores do que as perdas fixas. Por isso, é importante que as distribuidoras de energia elétrica adotem medidas para minimizar tanto as perdas fixas quanto as perdas variáveis, visando a melhoria da eficiência do sistema elétrico e a redução dos custos operacionais. Isso pode ser feito por meio da modernização dos equipamentos utilizados na distribuição de energia elétrica, da manutenção preventiva das redes elétricas e da adoção de estratégias para otimizar o fluxo de energia na rede elétrica.

De acordo com a Associação Brasileira de Distribuidores de Energia Elétrica (ABRADE) [36], as perdas na rede básica são calculadas pela diferença entre a energia gerada e aquela entregue aos consumidores pelas redes de distribuição. Essas perdas são calculadas mensalmente pela Câmara de Comercialização de Energia Elétrica (CCEE), e têm seu custo determinado anualmente durante os processos tarifários. Esse custo é então distribuído entre diversos participantes do setor elétrico, incluindo os geradores, consumidores, empresas transmissoras e distribuidoras de energia elétrica, bem como os próprios consumidores. Isso significa que o impacto das perdas não se limita apenas aos geradores e consumidores, mas é compartilhado por vários atores da indústria elétrica.

O custo das perdas na rede básica é um dos componentes que compõem a tarifa de energia elétrica paga pelos consumidores, sendo definido anualmente nos processos tarifários regulados pela ANEEL (Agência Nacional de Energia Elétrica). Através dos Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional (PRODIST) em seu módulo 7, a ANEEL define uma série de metodologias, parâmetros, indicadores e procedimentos para obtenção dos dados necessários para o cálculo de perdas técnicas.

De acordo com tais informações, em [37] a ANEEL registrou que no ano de 2022 a distribuidora Light S.A. apresentou uma perda técnica de 6,86%, o que representa uma redução em relação aos anos anteriores [15]. Essa redução pode ser atribuída a investimentos em melhorias e modernização da rede elétrica da distribuidora.

## 2.1.2 Perdas Não Técnicas

As perdas não técnicas, também reconhecidas como perdas comerciais, constituem a diferença entre as perdas totais e as perdas técnicas no contexto do fornecimento de energia elétrica. Como observado, essas perdas estão frequentemente associadas a práticas fraudulentas ou irregularidades no consumo, como a identificação de ligações clandestinas ou anomalias na medição, resultando na ausência de registros precisos do consumo real do cliente.

Essas perdas de energia não apenas geram impactos adversos significativos em diversos setores da sociedade, mas também acarretam prejuízos substanciais para a população, as empresas distribuidoras e o governo. Os efeitos negativos estendem-se além do âmbito econômico, afetando a confiabilidade do fornecimento de energia e contribuindo para questões ambientais.

Na próxima seção, serão abordados de maneira mais aprofundada os impactos decorrentes dessas perdas, destacando a necessidade premente de estratégias eficazes para mitigar essas práticas prejudiciais. Adicionalmente, a Figura 1 proporciona uma visão abrangente desse fenômeno em diferentes regiões do mundo, evidenciando a globalidade e a complexidade do desafio enfrentado no combate às perdas não técnicas.

| Região | Destaques |
|---|---|
| Europa | • Índices variam de 2,3% na Suécia até 19% na Turquia. |
| Ásia | • Índia apresenta variações a depender da região, entre 11% e 58%;<br>• Bangladesh apresenta perda superiores a 20%;<br>• Indonésia apresenta perdas de 7%;<br>• Malásia apresenta perda de até 15%;<br>• Tailândia apresenta perda de 11%. |
| América do Norte e Central | • México apresenta índices de 13%;<br>• Estados Unidos apresentam percentuais próximos a zero, mas em algumas regiões são registradas ligações clandestinas relacionadas ao cultivo ilegal de maconha. |

**Tabela 1. Índices de perdas não técnicas [46].**

Sob uma perspectiva econômica, conforme indicado pelo levantamento realizado pelo *Northeast Group* em 2017 [47], o impacto global decorrente das perdas não técnicas de energia atinge a expressiva cifra de US$ 96 bilhões anualmente.

Em virtude disso, a Figura 2 oferece uma visão panorâmica desse cenário, destacando os quatro países com os maiores impactos econômicos resultantes das perdas não técnicas.

Nesse contexto, o Brasil se destaca como um protagonista singular, representando sozinho 11% desse impacto global.

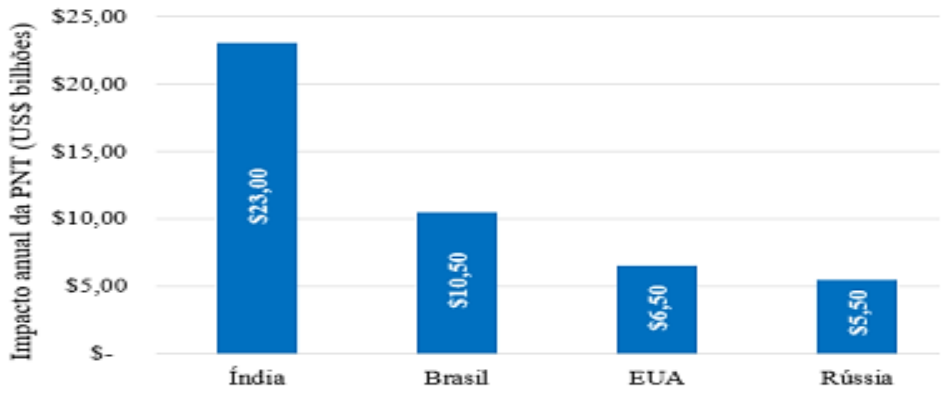


**Tabela 2. Impacto Econômico Causado pela PNT [47].**

Os dados apresentados na figura 2 revelam a dimensão do prejuízo causado pela perda não técnica de energia. Dentro desse aspecto, parte do prejuízo é assumido pelas concessionárias de distribuição e a outra parte é repassada aos consumidores por meio da fatura ou subsídios do governo, a depender do modelo regulatório adotado no país [46]. No caso do Brasil, em média, 3% do custo da tarifa de energia elétrica é relativo ao custo e perda não técnica. Em outros países esse valor pode chegar a 21,3% [48].

Realizada a descrição sobre o prejuízo causado pelas perdas, torna-se válido apresentar as duas formas de furto de energia mais comuns na rede de distribuição, que segundo [30] são:

- **Captação de energia:** A captação ilegal de energia é frequentemente efetuada por meio da derivação não autorizada dos circuitos, redirecionando a energia para usuários fraudulentos. Essa prática pode também ocorrer através de conexões clandestinas estabelecidas diretamente nos postes ou transformadores da rede elétrica, as quais são desconectadas durante os períodos de medição do consumo.

- **Fraude do medidor:** Em algumas áreas em que a leitura do medidor é feita diretamente pelo eletricista, pode ocorrer suborno para fornecer leituras falsas, resultando em valores pagos por uma quantidade de energia menor do que a energia real consumida. Além disso, a adulteração dos medidores pode ocorrer por meio da obstrução do movimento do disco, que geralmente é eletromecânico e consiste em discos que giram lentamente para registrar a energia consumida.

Práticas como essas resultam em impactos significativos no faturamento da distribuidora, comprometendo sua capacidade de investimento e expansão dos serviços oferecidos aos clientes. Por isso, é fundamental que as distribuidoras de energia elétrica adotem medidas efetivas para combater as perdas comerciais e identificar possíveis usuários que estejam agindo de forma irregular.

As perdas não técnicas no contexto da ocupação do espaço urbano [41] englobam um conjunto complexo de fatores econômicos, sociais e territoriais que se desenvolveram historicamente e não podem ser simplificados como uma mera consequência de um ambiente social desfavorável.

Primeiramente, aspectos econômicos desempenham um papel fundamental. A falta de acesso a serviços elétricos confiáveis e a consequente necessidade de recorrer a soluções improvisadas, como conexões clandestinas, muitas vezes é resultado de desigualdades econômicas e da falta de recursos para pagar pela energia de maneira regular [41]. Isso cria um ciclo em que as perdas não técnicas aumentam devido à falta de capacidade financeira dos consumidores. Além disso, aspectos sociais desempenham um papel importante. Comunidades marginalizadas ou em situação de vulnerabilidade muitas vezes recorrem a conexões ilegais de energia como uma forma de suprir necessidades básicas, sem acesso adequado aos serviços elétricos formais. Isso não apenas contribui para as perdas não técnicas, mas também reflete questões sociais mais amplas, como desigualdade e acesso limitado a serviços públicos.

Por fim, a ocupação do espaço urbano desempenha um papel significativo. O crescimento desordenado das áreas urbanas e a falta de planejamento adequado podem levar à instalação de infraestrutura elétrica inadequada, propiciando conexões ilegais e perdas não técnicas.

Portanto, a governança das perdas não técnicas requer uma abordagem holística que leve em consideração não apenas o ambiente governamental, mas também os fatores econômicos, sociais e de ocupação do espaço urbano que moldam esse problema complexo ao longo do tempo. Abordar as perdas não técnicas de modo eficaz envolve ações coordenadas entre o governo, as concessionárias de energia e a sociedade em geral para resolver esses problemas subjacentes.

No âmbito dessas perdas, a implementação de técnicas de aprendizado de máquina, conforme sugerido nesta dissertação, pode constituir um meio eficaz para identificar tais irregularidades. Isso possibilita uma tomada de decisão mais precisa por parte da distribuidora,

visto que as consequências dessas situações podem impactar diretamente a qualidade do serviço prestado, prejudicando a relação entre a empresa e seus clientes [40].

## 2.2 Técnicas para Combate a Perdas Comerciais

As distribuidoras de energia elétrica têm adotado uma série de tecnologias e ferramentas que visam mitigar as perdas não técnicas na rede de distribuição e com isso potencializar a qualidade do seu fornecimento. Adiante serão descritas algumas das principais técnicas utilizadas [42]:

- Sistema de Monitoramento Avançado: Utilizam sistemas de monitoramento que incluem sensores de rede. Isso permite o acompanhamento em tempo real do fluxo de energia e a detecção de anomalias que podem indicar perdas não técnicas.
- Medidores Inteligentes: A implantação de medidores inteligentes permite uma medição mais precisa do consumo de energia e uma detecção mais eficiente de desvios no fornecimento, o que ajuda a identificar perdas não técnicas.
- Tecnologia GIS (Sistema de Informação Geográfica): O GIS auxilia no mapeamento detalhado das redes elétricas, facilitando a localização de possíveis pontos de perdas não técnicas, como conexões clandestinas.
- Análise de Dados e Big Data: As distribuidoras coletam grandes volumes de dados operacionais e de consumo. A análise de big data e técnicas de análise de dados avançados podem identificar padrões suspeitos e comportamentos irregulares que indicam perdas não técnicas.
- Inteligência Artificial (IA): A IA é utilizada para desenvolver modelos preditivos que ajudam na detecção de possíveis perdas não técnicas. Algoritmos de *Machine Learning* podem analisar dados históricos para identificar tendências e comportamentos anômalos.
- Sistemas de Denúncias e Canais de Comunicação: Estabelecem canais diretos de comunicação com os consumidores para que possam denunciar suspeitas de perdas não técnicas, incentivando a colaboração da comunidade.
- Inspeções e Fiscalizações: Realizam inspeções periódicas para identificar irregularidades nas instalações elétricas e combater fraudes de energia.

- Treinamento de Equipe: Investem em capacitação para equipes de campo e de atendimento ao cliente, para que estejam preparadas para identificar e lidar com casos de perdas não técnicas.
- Automatização de Processos: Automatizam tarefas de detecção e gerenciamento de perdas, tornando o processo mais eficiente e preciso.
- Integração de Dados: Integrar dados de várias fontes, como informações de medição, geoespaciais e denúncias de consumidores, permite uma visão mais abrangente das perdas não técnicas.

Essas tecnologias e ferramentas, quando combinadas, possibilitam uma abordagem mais abrangente e eficaz no combate às perdas não técnicas, contribuindo para a eficiência operacional das distribuidoras e para a redução dos custos associados a essas perdas.

Em geral as distribuidoras têm procurado cada vez mais usar novas tecnologias para se evitar roubo de energia e fraudes. Um exemplo disso, é mostrado em [44], onde a violação do medidor de energia pode ser detectada utilizando um arranjo simples de um LED IR mais um fotodiodo.

Este estudo foi desenvolvido exclusivamente para situações em que os medidores de energia são eletromecânicos de padrão convencional. Para isto, um fotodiodo é colocado no eixo do disco giratório do medidor e é iluminado com luz infravermelho do LED. Em operação normal, a saída do fotodiodo fornece um sinal lógico baixo para o microcontrolador. No entanto, se o medidor interfere, ou seja, a rotação do disco é obstruída ou a tampa do medidor é removida, um obstáculo é criado entre o LED e o fotodiodo, resultando em um sinal lógico alto para o microcontrolador. O microcontrolador detecta essa mudança no sinal lógico e com base nisso, envia uma mensagem para o modem GSM através do deslocador de nível Max 232. O modem GSM então envia a mensagem do medidor de energia adulterado para um local específico e uma ação apropriada é tomada em conformidade com o detectado. Uma das ações a serem priorizadas nestes casos corresponde ao corte da alimentação elétrica vinculada à instalação verificada e posteriormente a substituição do contador em caso de avaria [49].

Outro exemplo bastante comum e muito utilizado pelas distribuidoras de energia no combate à mitigação das perdas é o uso da Telemedição. Como visto anteriormente, esta tecnologia trata do monitoramento remoto da medição das unidades consumidoras sem haver a necessidade de intervenção física, pois o processo é feito de forma automatizada.

A comunicação do sistema é realizada por meio de redes GPRS (*General Packet Radio Service*), que utilizam torres sem fio de empresas de telefonia móvel como meio de

comunicação, e por redes *mesh* que utilizam roteadores e medidores com módulos de comunicação capazes de transferir dados entre si. Os dados são transmitidos até chegar ao coletor, onde são disponibilizados para a concessionária gerenciar seus equipamentos.

A Telemedição é composta por:

1. Sistema de medição centralizada (SMC): os medidores de energia são retirados do padrão do cliente e são instalados de forma agrupada nos postes de maneira que evite possíveis fraudes. Nesse tipo de sistema suas informações são transmitidas diretamente para concessionária, sendo capaz de fazer determinadas ações como execução de corte e religamento.
2. Sistema de medição individual (SMI): neste sistema os equipamentos são instalados de forma convencional no padrão do cliente, porém com interface de comunicação direta através de monitoramento em tempo real de energia.

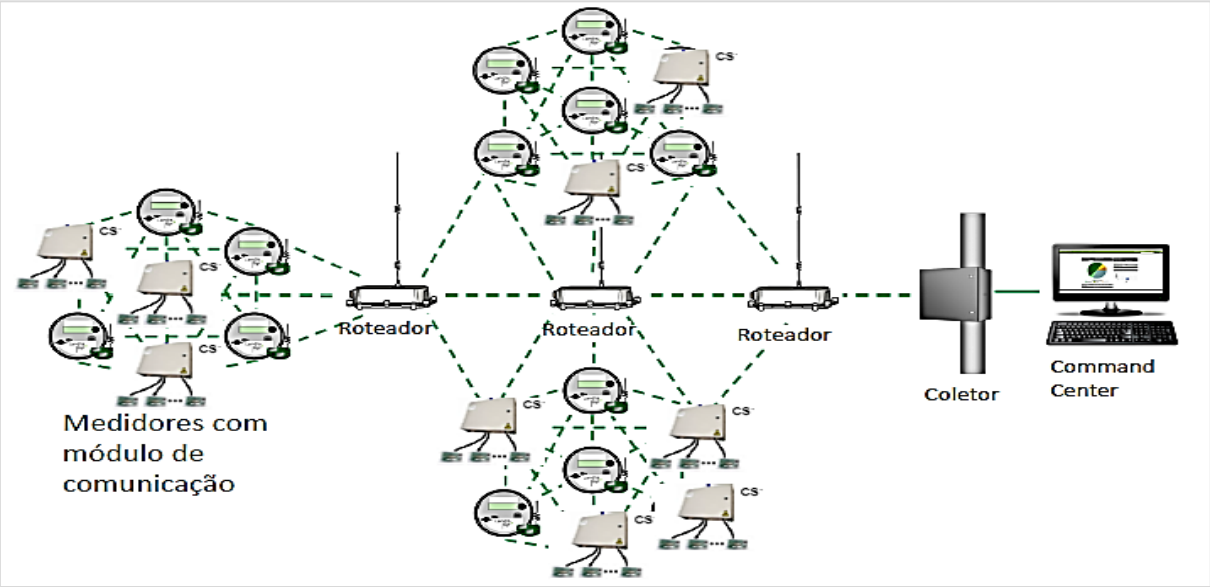


**Figura 1. Modelo de comunicação dos medidores SmartGrid**
Fonte: Própria

Em ambos os processos é possível realizar remotamente aplicações como medição, corte, religação e diagnóstico do fornecimento através de alarmes de monitoramento para controle de qualidade na distribuição. Com relação ao seu princípio de funcionamento, a Figura 1 ilustra resumidamente a hierarquia de sua aplicação. Os medidores inteligentes coletam dados de consumo e se comunicam com roteadores. Os roteadores gerenciam o tráfego de dados na rede, encaminhando informações entre medidores, coletores e o centro de comando. Os

coletores agregam dados de diversos medidores em uma área específica, enquanto o centro de comando monitora, analisa e gerencia a rede como um todo, tomando decisões para otimizar a operação do sistema elétrico.

## 2.3 Redução das PNT nas distribuidoras de energia

A identificação incorreta de um possível roubo de energia pode levar a uma operação ineficiente e a resultados indesejados. Embora o processo de identificação de fraudes possa ser eficaz ao apontar fraudes simples que resultam em recuperação de baixa energia, ele pode ser ineficaz com fraudes mais sofisticadas cometidas por clientes residenciais, comerciais ou industriais de alto padrão, que podem permanecer não detectados por anos e causar perdas financeiras significativas para a concessionária.

Muitas concessionárias de serviços públicos falham em identificar com eficácia o roubo de energia, o que resulta na incapacidade de detectar e capturar os fraudadores de forma rápida, levando a inspeções ineficazes e grandes vazamentos de receita. Isso pode deixar muitos furtos, especialmente os maiores e mais valiosos, sem serem detectados, enquanto frequentemente incomoda clientes inocentes com inspeções infrutíferas. Além disso, as recuperações de energia geralmente dependem da capacidade de pagamento do cliente em casos mais sofisticados de roubo, o que representa outro problema adicional para as concessionárias.

O método tradicional de detecção de fraudes, baseado em monitoramento manual, filtragem e relatórios, traz problemas de integridade em potencial, com alta probabilidade de relatórios imprecisos e sujeitos a erros. Os problemas de integridade referem-se à possibilidade de que os dados coletados não estejam completos ou precisos, devido a diversos fatores, como falhas humanas, técnicas ou mesmo intencionais. Esses problemas podem incluir, por exemplo, omissão de informações relevantes, distorção de dados ou até mesmo a criação de informações falsas.

Os relatórios imprecisos podem ocorrer devido a uma série de fatores, incluindo falhas no processo de coleta e análise de dados, erros humanos na interpretação dos dados, falta de controle de qualidade e problemas técnicos nos sistemas de monitoramento. Além disso, os relatórios também podem ser afetados por fatores externos, como interferências eletromagnéticas ou condições climáticas adversas.

Os erros podem ser de diversos tipos, incluindo erros de medição, erros de registro, erros de transmissão de dados, erros de processamento e erros de interpretação. Esses erros podem

ocorrer em qualquer etapa do processo de detecção de fraudes, desde a coleta até a análise dos dados.

Contudo, percebe-se em [5], por exemplo, que um software utilizado adequadamente pode dobrar a eficácia das inspeções, além de ajudar no processo de validação evitando novos erros. Algumas concessionárias oferecem suporte às atividades anteriores com gerenciamento de fraude ou pacotes estatísticos projetados para uso geral. No entanto, é importante ressaltar que a implementação e personalização de um software para atender às necessidades específicas de uma concessionária pode exigir investimentos significativos em termos de tempo e recursos financeiros. Além disso, embora possa trazer benefícios, esse tipo de solução pode não oferecer todas as funcionalidades necessárias, exigindo a adaptação de outras ferramentas ou o desenvolvimento de recursos personalizados. Uma estratégia comum é contratar equipes terceirizadas para realizar a maior parte das inspeções, e em paralelo dedicar uma equipe interna de especialistas para vistoriar as instalações tecnicamente mais complexas, como industriais e de alto comércio.

Independentemente da situação, a confiabilidade dos resultados obtidos nas inspeções é um elemento central para todo o processo, pois relatórios imprecisos, por exemplo, podem comprometer as análises subsequentes baseadas em tais informações. Dentre as possíveis causas para tal cenário estão: treinamento inadequado, falta dos equipamentos necessários para verificar possíveis irregularidades, comportamentos antiéticos por parte das equipes de campo e internas, tempo insuficiente para realizar uma inspeção precisa ou, ainda, para concluí-la. Essas variáveis destacam a importância de abordagens que visem mitigar tais desafios e assegurar a qualidade das inspeções realizadas.

Dentro desse contexto, é fundamental saber que cada resultado de inspeção será uma nova informação importante para verificar a precisão das previsões e de todos os processos anteriores e, portanto, precisa ser monitorada. Compreender o que aconteceu de acordo com a previsão e o que não aconteceu e, em seguida, aprender sobre as questões problemáticas é fundamental para melhoria contínua do processo.

Para além das precauções anteriormente abordadas, é relevante destacar que verificações e ajustes apoiados por software ocasionalmente podem ser impactados por fatores externos, comprometendo sua operacionalidade. Para ilustrar essa situação, consideremos um cenário em que um sistema de monitoramento identifica uma redução expressiva no consumo de energia de alguns clientes em um edifício recém-construído. Após investigação, constata-se que um problema na atualização do sistema de faturamento inadvertidamente excluiu esses clientes do cálculo de consumo [45].

Outros exemplos comuns incluem registros incorretos de clientes, medidores danificados ou mal calibrados, inserção equivocada de dados de medição no sistema de faturamento, programação de inspeções que não foram efetuadas, entre outras eventualidades.

Frente a esse contexto, torna-se evidente a necessidade imperativa de monitorar indicadores para o controle e gestão eficiente desse processo. Dessa maneira, apresentam-se a seguir os principais tipos [52]:

- **NTL (perdas não técnicas ou comerciais**) - estima a quantidade de energia que não é cobrada dos clientes por motivos não técnicos. É composto de fraude, roubo e problemas de processo (medição, erros humanos e de sistema). Esta análise pode ser realizada através do balanço energético da distribuidora, onde se verifica por meio de cálculos a demanda de energia fornecida na rede com relação à consumida pelos clientes.

- **TL (perdas técnicas):** Este indicador representa as perdas técnicas causadas principalmente pelas imperfeições dos processos físicos de distribuição e impedância elétrica. Como as NTL, este indicador também pode ser verificado através do cálculo de balanço energético através da diferença entre a perda total e a perda não técnica.

- **Efetividade das inspeções de campo** - Percentual calculado como o número total de roubos relatados nas inspeções, dividido pelo número de inspeções realizadas.

- **Produtividade das inspeções de campo** - É calculado pela quantidade de energia que não foi faturada durante o período de fraude, dividido pelo número de inspeções. Seu objetivo é avaliar como um todo a eficiência das normalizações.

Com a compreensão dos indicadores fundamentais para o controle e gestão do processo em destaque, é oportuno explorar as estratégias adotadas pelas distribuidoras de energia elétrica. A próxima seção apresentará algumas das abordagens empregadas com o intuito de mitigar as perdas não técnicas na rede, destacando a importância de práticas eficazes para o setor.

## 2.3.1 Blindagem de Rede

Na seção anterior, foi delineada a relevância dos indicadores no contexto operacional de distribuidoras de energia elétrica, destacando a presença das perdas não técnicas como um desafio significativo. Agora será iniciada a discussão das estratégias adotadas por essas distribuidoras para mitigar tais perdas. Serão exploradas as abordagens implementadas com o objetivo de aprimorar a eficiência, fortalecer a gestão e enfrentar os desafios inerentes ao setor de distribuição de energia.

Um exemplo disso é a blindagem de rede, uma estratégia muito eficaz utilizada por algumas distribuidoras de energia e que tem proporcionado excelentes resultados. Neste cenário, todos os equipamentos e infraestrutura elétrica são blindados para minimizar e evitar o roubo de energia. A blindagem da rede, então, protege do vandalismo os pontos dos componentes da rede elétrica considerados frágeis. Alguns produtos antifurto podem ajudar a proteger a rede e, consequentemente, reduzir o roubo de energia. Esses produtos dificultam o acesso ao interior dos cabos e também evitam que campos elétricos e sinais de alta frequência cheguem aos circuitos próximos ao equipamento blindado. [53].

Como exemplo da efetividade da rede de blindagem [15], a LIGHT apresentou um modelo padrão de rede blindada com telemedição denominada BT ZERO, que teve como local de execução a comunidade Santa Marta, no bairro de Botafogo. De acordo com o projeto [15], houve uma melhora de mais de 95% na perda de energia, reduzindo de um patamar de 90% para apenas 4%. Esse tipo de blindagem conta com um transformador blindado, onde suas conexões ficam contidas numa caixa de aço junto ao transformador e uma caixa conjugada que serve para fazer a associação entre as unidades consumidoras. A Figura 2 ilustra a aplicação dessa estratégia em campo.

**Figura 2. Exemplo de Caixa Blindada (CB)**
**Fonte: própria**

## 2.3.2 Blindagem de medidor

Em virtude das inúmeras fraudes identificadas no equipamento de medição, a estratégia de proteger o equipamento e seus compartimentos internos tornou-se uma solução muito atraente para a distribuidora de energia elétrica. Por este motivo, algumas áreas com grande concentração de perdas na medição tiveram a instalação da caixa blindada nos medidores, limitando o total acesso do usuário ao equipamento e restringindo a operação apenas à concessionária.

A Figura 3 ilustra um grupo de medidores onde anteriormente à execução do projeto o consumidor possuía livre acesso aos respectivos bornes, tornando o sistema vulnerável e propício a fraudes. Após a blindagem, o cliente fica limitado apenas à coleta de leitura conforme regulamentado pela ANEEL [30].

**Figura 3. Medidores convencionais antes e pós blindagem**

**Fonte: própria**

## 2.3.3 Estratégias de Prevenção

Como visto nas seções introdutórias, nos últimos anos, o setor elétrico tem enfrentado um desafio constante e crescente no que diz respeito às perdas não técnicas, particularmente a fraude elétrica. No início, a abordagem predominante para prevenir a fraude no setor elétrico era fortemente baseada em medidas físicas e de segurança. Essas estratégias eram vitais e tiveram um impacto significativo na redução das perdas não técnicas. No entanto, à medida que a complexidade e a sofisticação das práticas fraudulentas aumentaram, tornou-se evidente que abordagens meramente físicas não eram suficientes para acompanhar a evolução do problema.

Foi nesse contexto que a inteligência artificial emergiu como uma ferramenta poderosa no combate à fraude elétrica. A análise de dados e técnicas de aprendizado de máquina passaram

a desempenhar um papel fundamental na identificação de padrões de consumo suspeitos, na detecção de anomalias e na previsão de possíveis casos de fraude. Essa transição representou uma revolução na forma como as empresas de energia elétrica abordam o problema das perdas não técnicas.

Nas seções subsequentes serão exploradas as estratégias emergentes de combate à fraude de energia utilizando *Machine Learning*, bem como os desafios e oportunidades que surgem da integração dessas abordagens.

## 2.4 *Machine Learning*

Com o objetivo de proporcionar uma abordagem mais eficaz e holística para o problema das perdas não técnicas, nesta seção será visto como o *Machine Learning* pode ser útil no aspecto de análise, resultando em benefícios tangíveis para as empresas de distribuição de energia.

No contexto da redução de perdas, o princípio de funcionamento do *Machine Learnin*g destaca-se pela capacidade de aprender e generalizar padrões a partir de dados específicos. Os seus algoritmos, ao serem treinados com conjuntos de dados que refletem padrões de perdas, conseguem identificar regularidades e tendências. Isso possibilita a criação de modelos preditivos capazes de analisar novos dados e prever potenciais casos de perdas não técnicas.

A aplicação dessa abordagem oferece uma ferramenta valiosa para as distribuidoras de energia elétrica, permitindo a antecipação e mitigação proativa de situações que poderiam resultar em perdas indesejadas.

O aprendizado de máquina tem uma gama muito ampla de aplicações possíveis. No geral, suas aplicações abrangem [58]:

- Encontrar, extrair e resumir dados relevantes;
- Fazer previsões com base nos dados analisados;
- Calcular probabilidades para certos eventos;
- Adaptar ao ambiente de forma independente e
- Otimizar processos com base em padrões reconhecidos.

### 2.4.1 Tipos de Aprendizado

Serão exploradas agora diferentes abordagens de aprendizado em *Machine Learning*, cada uma desempenhando um papel distinto no reconhecimento de padrões. Essas abordagens podem ser categorizadas em três principais tipos: Aprendizagem Supervisionada, Aprendizagem Não Supervisionada e Aprendizagem por Reforço [54].

De forma geral, a aprendizagem supervisionada é usada quando há um conjunto de dados rotulados disponíveis para treinar o modelo, enquanto o aprendizado não supervisionado é utilizado quando não há rótulos disponíveis e o modelo deve encontrar padrões seguindo a distribuição e as semelhanças entre os dados. Já o aprendizado por reforço é usado quando um agente precisa aprender a tomar ações em um ambiente incerto visando maximizar uma recompensa associada a cada ação tomada em cada instante de tempo.

Dentro das principais classes de problemas aos quais os modelos de *Machine Learning* podem ser aplicados destacam-se a classificação, regressão, clusterização e otimização, desempenhando papéis cruciais na resolução de diversos desafios em várias áreas [27].

A classificação, uma técnica de *Machine Learning*, é empregada para categorizar dados em diferentes classes ou categorias, atribuindo *labels* com base em características específicas. Um exemplo comum é a classificação de e-mails como spam ou não spam. Por sua vez, a regressão é utilizada para prever valores contínuos a partir de dados de entrada, sendo amplamente aplicada em problemas de previsão, como a estimativa de preços de imóveis com base em características como tamanho e localização. A clusterização, outra categoria importante, agrupa dados semelhantes em clusters ou grupos, identificando padrões intrínsecos nos dados sem a necessidade de rótulos prévios. Um exemplo prático é agrupar clientes com base em seus hábitos de compra. Finalmente, a otimização concentra-se em encontrar o melhor valor possível de uma função de custo ou objetivo, sendo empregada em diversas aplicações, desde a otimização de rotas de entrega até o ajuste de parâmetros em modelos de *Machine Learning* para alcançar um melhor desempenho. Essas categorias oferecem ferramentas essenciais para abordar uma ampla gama de desafios, cada uma adequada para tipos específicos de tarefas e requisitos de dados.

Por fim, o aprendizado de máquina é uma área muito ativa de pesquisa e desenvolvimento, e novas técnicas e algoritmos estão sendo desenvolvidos continuamente para lidar com novos desafios e problemas complexos.

### 2.4.2 Aprendizagem por Reforço

O aprendizado por reforço (ou *Reinforcement Learning* – RL) é uma abordagem de *Machine Learning* onde um agente interage com um ambiente, toma decisões e, com base nessas decisões, recebe feedback na forma de recompensas ou penalidades. O objetivo é que o agente aprenda a realizar ações que maximizem as recompensas ao longo do tempo [55].

Esse tipo de aprendizado é frequentemente aplicado em situações em que um sistema precisa tomar uma série de ações sequenciais para atingir um objetivo final. O agente explora o ambiente, aprende com as consequências de suas ações e ajusta sua estratégia ao longo do tempo para otimizar as recompensas. Isso encontra aplicações em jogos, robótica, navegação autônoma e outras áreas onde a tomada de decisões sequenciais é essencial.

O processo envolve a exploração contínua do agente no ambiente para aprender as melhores ações em diferentes situações, proporcionando uma abordagem dinâmica e adaptativa.

### 2.4.3 Aprendizagem Não Supervisionada

O aprendizado não supervisionado é uma abordagem de *Machine Learning* em que o algoritmo é treinado em um conjunto de dados não rotulado, ou seja, um conjunto em que as saídas desejadas não são fornecidas. O objetivo principal é explorar a estrutura e os padrões subjacentes nos dados sem orientação prévia sobre as saídas esperadas.

Diferentemente do aprendizado supervisionado, onde o modelo recebe exemplos rotulados para aprender a fazer previsões, no aprendizado não supervisionado, o algoritmo tenta identificar padrões intrínsecos nos dados por conta própria. Isso pode envolver a descoberta de agrupamentos naturais, a redução da dimensionalidade ou a identificação de relações entre variáveis.

Um exemplo comum de aprendizado não supervisionado é o algoritmo de agrupamento, como o *k-means*, que tenta agrupar os dados em clusters com base em similaridades. Essa abordagem é útil quando se deseja explorar a estrutura subjacente de conjuntos de dados sem a necessidade de rótulos pré-existentes.

Suas aplicações incluem segmentação de mercado, análise exploratória de dados e compressão de dados.

## 2.4.4 Aprendizado Supervisionado

A aprendizagem supervisionada é um ramo do *Machine Learning* que consiste em estimar uma função que mapeia um conjunto de entradas para uma ou mais saídas desejadas com base em pares de entrada-saída de exemplo. Neste método os conjuntos de dados são rotulados para que haja uma resposta com a qual a máquina possa medir sua precisão [28].

Destaca-se que para total compreensão do tema, é necessário o entendimento de dois tipos de variáveis existentes:

1. Variável independente ou preditora: aquela que será passada para o modelo, tendo influência na variável que se pretende estimar;
2. Variável alvo ou dependente: a variável que se pretende prever, estimar ou projetar.

Neste exemplo, o aprendizado supervisionado é aplicado quando se tenta encontrar a relação entre a variável-alvo e as variáveis independentes. Assim, o objetivo do aprendizado supervisionado é obter, ao final de um processo iterativo de ajuste dos seus parâmetros internos, um modelo matemático treinado para criar associações entre as variáveis independentes e as variáveis de saída [57].

Em modelos de aprendizado de máquina supervisionado, o algoritmo é treinado com dados rotulados para aprender os padrões dos dados e gerar uma função matemática que aproxime a dinâmica da variável dependente em função das variáveis independentes. Essa função é usada para gerar valores de resposta para novos dados de entrada, tornando possível a previsão de resultados futuros com base em dados históricos.

Como por exemplo a função [28]:

$$\hat{y} = f(X) = f(x_{1,}x_{2,}\ldots, x_n) \quad (1)$$

Onde:

- $f(X)$: é a função que o algoritmo irá estimar;
- $x_n$: é o conjunto de variáveis independentes, também conhecidas como atributos;
- $\hat{y}$: a saída projetada, com base na função estimada e no conjunto de variáveis independentes apresentado ao modelo.

Dentro desse contexto, podem ser identificados alguns algoritmos de aprendizagem supervisionada que são mencionadas com mais frequência na literatura [59]:

• KNN (K-Nearest Neighbors): é usado para classificar uma amostra de um conjunto de dados avaliando sua distância e relação com os vizinhos próximos;

• RL (Regressão Linear): é usado para estimar o valor esperado da variável dependente a partir da combinação linear dos valores das variáveis independentes, utilizando um conjunto de exemplos (por exemplo, histórico) para estimar os coeficientes da referida combinação linear;

• MLP (Perceptron de Múltiplas Camadas): funcionam com uma rede de neurônios artificiais, que aprendem a partir dos dados de entrada e são capazes de identificar relações complexas e não-lineares.

• SVM (*Support Vector Machine*): é uma ferramenta de classificação e regressão que constrói hiperplanos de margem máxima em um espaço n-dimensional para classificar ou regredir dados;

• RF (*Random Forest*): é uma técnica de aprendizado baseada em árvores de decisão, em que cada árvore é construída com amostras aleatórias do conjunto de dados.

A literatura é vasta no que diz respeito aos modelos de aprendizado supervisionado e nas próximas seções serão discutidos aqueles selecionados para este trabalho, a saber: SVM, árvore de decisão e *XGBoost*. Adianta-se que, mediante uma análise comparativa criteriosa, o algoritmo *Random Forest (RF)* foi aquele que apresentou eficácia superior. Estes resultados são explorados no capítulo de resultados, evidenciando o papel fundamental desempenhado pelo RF na abordagem e resolução dos desafios propostos.

### 2.4.4.1 *Support Vector Machine*

A máquina de vetor suporte (*Support Vector Machine - SVM*), desenvolvida por VAPNIK et al em [63], é um algoritmo de aprendizado de máquina supervisionado que pode ser usado tanto para problemas de classificação quanto de regressão. Esta tecnologia visa

construir um modelo com máxima capacidade de generalização, que basicamente encontra uma superfície de separação, conhecida usualmente como hiperplano de margem máxima de separação entre dados pertencentes de duas classes distintas.

A Figura 4 apresenta dois exemplos do uso do SVM. A esquerda em um problema de classificação linearmente separável, ou seja, em que as classes podem ser separadas por uma linha reta, e à direita, onde o método é aplicado a problemas não linearmente separáveis. Esta segunda aplicação é viabilizada através do uso de uma técnica chamada *kernel trick* [63]. Essa técnica mapeia os dados para um espaço de dimensão superior, em que as classes tem maior probabilidade de serem linearmente separáveis do que no espaço original. A SVM então encontra a margem máxima que separa as classes nesse espaço de dimensão superior.

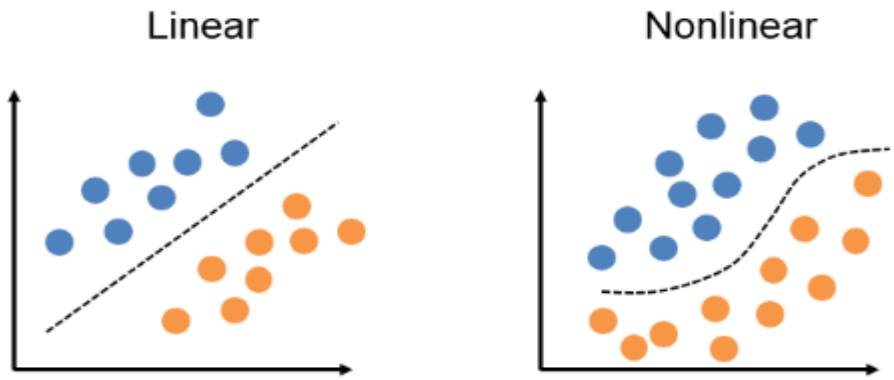


**Figura 4. Classes linearmente e não linearmente separáveis [60]**

A margem de separação é a distância entre a linha de separação e os pontos mais próximos de cada classe, chamados de vetores de suporte. O objetivo da SVM é maximizar essa margem, ou seja, encontrar a linha de separação que está o mais longe possível dos vetores de suporte. Isso torna a SVM mais robusta a ruídos nos dados e a possíveis sobreajustes do modelo (*overfitting*) [65].

Em resumo, a SVM é uma técnica de classificação que busca encontrar a margem máxima que separa as classes em um espaço de dimensão superior, utilizando o conceito de vetores de suporte e a maximização da margem de separação. Esse modelo pode também ser estendido para problemas de regressão e maiores detalhes teóricos podem ser encontrados em [64].

## 2.4.4.2 Árvore de Decisão

A árvore de decisão é um tipo de algoritmo utilizado para modelar processos de decisão baseados em possíveis consequências. Esse modelo leva em consideração eventos disponíveis e, do ponto de vista da tomada de decisão, é representado por um conjunto mínimo de perguntas que avaliam a probabilidade de decisões adequadas à luz das respostas fornecidas nos eventos (registros) disponíveis.

Um exemplo prático de algoritmo baseado em árvore de decisão é o *Random Forest*, que é utilizado em diversas aplicações, incluindo o estudo em questão. Esse algoritmo cria várias árvores de decisão e combina os resultados para obter uma maior precisão na previsão. Mais especificamente, o *Random Forest* é um conjunto de várias árvores de decisão que possuem diferentes nós, gerados aleatoriamente para a classificação desejada.

O uso de árvores de decisão e algoritmos relacionados tem se mostrado muito eficaz em diversas áreas, incluindo a análise de dados e a tomada de decisão. A simplicidade do modelo, combinada com sua capacidade de lidar com múltiplos registros e sua versatilidade em diferentes contextos, torna esses algoritmos ferramentas valiosas para a resolução de problemas complexos em diversas áreas. Ao final, assim que todas as árvores tenham terminadas sua classificação individual, o algoritmo realiza um comitê para validar qual a classificação mais indicada.

Para melhor entendimento do método, pode-se supor que exista um problema simples para classificar o formato e a cor de um objeto, podendo este ser apenas um triângulo ou um quadrado, preto ou branco, e que além disso, cada pergunta feita diante do problema proposto, seja considerada um nó. Sendo assim, todo nó identificado irá dividir a árvore em dois caminhos diferentes que no decorrer do problema não irão se cruzar em nenhum momento.

Nesta situação, a criação do primeiro nó de decisão aconteceria mediante a pergunta de quantos lados possui a figura, como por exemplo, o objeto tem 4 lados? Se sim, a parte relevante da árvore se torna a da direita com a classificação de um “quadrado”, caso a resposta seja negativa a relevância seria a da esquerda classificado como “triângulo”. E assim segue a sequência de perguntas até que o resultado final indique a cor e a forma do objeto em questão. Vale ressaltar que ao fim do modelo serão obtidos 4 registros: triângulo preto, triângulo branco, quadrado preto e quadrado branco, ilustrados na figura 5.

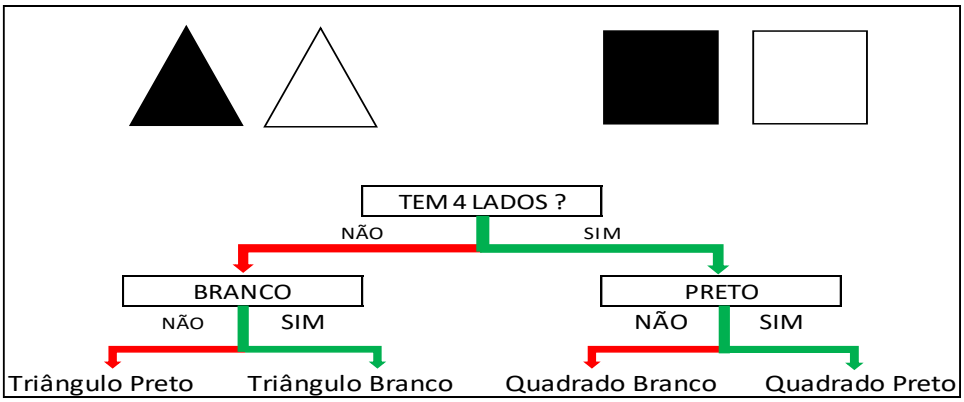


**Figura 5. Exemplo de árvore de decisão**
**Fonte: própria**

### 2.4.4.3 *XGBoost*

Diferente da árvore de decisão mostrada na seção anterior, o *XGBoost* é um algoritmo de conjunto que combina várias árvores de decisão em um modelo mais complexo. Ele cria árvores profundas em uma abordagem iterativa, o que pode resultar em modelos mais sofisticados.

Este algoritmo utiliza como princípio de funcionamento uma estrutura de *Gradient Boosting*. Seu objetivo é aumentar a precisão das previsões, introduzir técnicas de regularização e tornar o algoritmo mais rápido.

Dentre suas principais contribuições, pode-se citar a criação de um novo método para testar o melhor valor para um nó, por meio de pontuação de semelhança e ganho, e técnicas de regularização, como λ (semelhante ao decaimento de peso em redes neurais) e γ (poda).

O *XGBoost* fornece um reforço de árvore paralela que tem apresentado bom desempenho em muitos problemas de ciência de dados. Exemplo disso pode ser visto em [66] e [67], onde os autores utilizam este algoritmo para identificação de possíveis clientes fraudadores. Embora o modelo geralmente alcance maior precisão do que uma única árvore de decisão, ele reduz a interpretabilidade intrínseca das árvores de decisão. Para obter desempenho e interpretabilidade, algumas técnicas de compressão de modelo permitem transformar um *XGBoost* em uma única árvore de decisão que se aproxima da mesma função de decisão.

Sua aplicação envolve a divisão de um conjunto de dados em subconjuntos com base nas características dos dados, com o objetivo de criar uma estrutura hierárquica em forma de árvore. Segue adiante um resumo simplificado do seu processo [67]:

1. Seleção do Atributo: Inicialmente, escolhe-se um atributo do conjunto de dados que seja relevante para a classificação ou decisão que se deseja tomar.
2. Divisão dos Dados: Com base no atributo selecionado, os dados são divididos em grupos menores. Cada grupo contém instâncias de dados que possuem valores específicos para o atributo escolhido.
3. Critério de Divisão: Um critério de divisão é usado para determinar como dividir os dados. Os critérios comuns incluem ganho de informação, índice Gini e erro de classificação. O critério ajuda a escolher a divisão que melhor separa as classes de interesse.
4. Recursão: O processo de seleção de atributos e divisão dos dados é repetido em cada nó filho, criando uma estrutura de árvore hierárquica. Isso continua até que um critério de parada seja atingido, como uma profundidade máxima da árvore ou um número mínimo de amostras em um nó.
5. Classificação ou Decisão: Uma vez que a árvore está construída, é usada para classificar novos dados. Quando um novo dado entra na árvore, ele segue o caminho dos nós de acordo com os valores de seus atributos, até chegar a uma folha da árvore que contém a classe ou decisão final.

Em geral, o *XGBoost* é uma escolha mais comum quando se busca desempenho superior, enquanto as árvores de decisão podem ser preferíveis quando a interpretabilidade do modelo é crucial.

## 2.5 Validação Cruzada

Devido à sua fundamental importância no processo de modelagem, ao longo dos anos, uma variedade de técnicas de validação foi desenvolvida [72]. No entanto, devido à complexidade subjacente a esse tópico, sua discussão tem sido objeto de amplo debate na literatura.

Entre as várias abordagens, quatro técnicas de validação cruzada são amplamente reconhecidas, a saber: *Hold-out, K-fold, Leave-One-Out e Bootstrap*. Neste trabalho foi escolhida a utilização do método de validação cruzada *K-fold* devido à sua aplicabilidade eficaz considerando o conjunto de dados disponível.

Ao empregar a metodologia *K-fold* [70], a amostra é dividida em *K* partes $(d_1, d_2, \dots, d_k)$ de tamanhos uniformes. Este processo implica em *K* iterações, em cada uma das quais uma das

partes da amostra é designada como conjunto de validação, representada por $d_n$, com $n$ variando de 1 até *K*. O conjunto de treinamento, por sua vez, é composto pelas *K*-1 partes restantes, o que significa que em cada iteração, uma parte diferente da amostra atua como conjunto de teste. Essa abordagem oferece uma avaliação robusta e abrangente do desempenho do modelo, tornando-a particularmente adequada às necessidades deste trabalho.

A aplicação da fórmula do *K-fold* em *Machine Learning* segue o seguinte procedimento:

1. Divisão do conjunto de dados em K subconjuntos;
2. Iteração *K* vezes, cada vez utilizando um subconjunto diferente como conjunto de validação e os demais como conjunto de treinamento;
3. Treinamento do modelo em cada iteração;
4. Avaliação do desempenho do modelo utilizando métricas apropriadas;
5. Cálculo da média das métricas obtidas nas *K* iterações.

A estatística média para avaliação de desempenho é representada pela seguinte equação:

$$k\ f\ K = \frac{1}{K}\sum_{n=1}^{K}\frac{1}{m_n}\sum_{i=1}^{m_n} L(y_{in}, f_{(-n)}(x_{in})) \quad (2)$$

Onde:

$n$ = 1 $até\ K$, sendo o modelo avaliado nas observações da amostra de teste.

$f_{(-n)}(x_n)$ é criado como a amostra de treino de $d(K)$.

$L(y_{in}, f_{(-n)}(x_{in}))$ é uma medida de diferença entre a saída desejada e a saída projetada pelo modelo

Destaca-se que o número de padrões em cada conjunto $d(K)$ diminui quanto maior for o valor de *K* [70]. Logo, utilizar um valor de *K* muito elevado acaba aumentando o custo computacional da técnica, além de uma amostra de teste pequena, o que aumenta a variância. Na literatura se discute qual valor de *K* seria o ideal, sendo os mais utilizados, dois, três, cinco e dez [70].

## 2.6 Métricas de Avaliação do Modelo

As métricas de avaliação desempenham um papel crítico na avaliação de modelos de *Machine Learning* pois fornecem insights objetivos sobre o desempenho do modelo. Um exemplo notável é a matriz de confusão, que desempenha um papel fundamental na análise do desempenho de um modelo em tarefas de classificação, permitindo a identificação de falsos positivos e falsos negativos. Essas métricas são de importância crucial para garantir a confiabilidade e a adequação do modelo à sua aplicação específica.

A figura 8 ilustra a representação visual dessa matriz, que sintetiza de forma concisa os resultados de um problema de classificação. Ela oferece uma visão abrangente do desempenho do modelo, ajudando a compreender a relação entre as predições corretas e as incorretas.

| | | Valor Predito | |
|---|---|---|---|
| | | NÃO | SIM |
| Real | NÃO | Verdadeiro Negativo (TN) | Falso Positivo (FP) |
| | SIM | Falso Negativo (FN) | Verdadeiro Positivo (TP) |

**Figura 6. Exemplo de matriz de confusão para um algoritmo de classificação binária [72].**

A seguir serão abordados cada um dos termos apresentados na figura 6:

- Verdadeiro Positivo (*True Positive* - TP): São previsões corretamente identificadas como verdadeiras. Neste contexto, representa quando o modelo acerta ao identificar um cliente suspeito de fraude.
- Verdadeiro Negativo (*True Negative* - TN): São previsões que foram classificadas corretamente como clientes regulares.
- Falso Positivo (*False Positive* - FP): Representam previsões erroneamente identificadas como fraudadores.
- Falso Negativo (*False Negative* - FN): São previsões que foram classificadas incorretamente como clientes regulares.

Para problemas de classificação, diversas métricas comuns desempenham um papel essencial na avaliação e garantia da qualidade do modelo desenvolvido e testado, as quais serão discutidas em detalhes a seguir em suas respectivas representações. A acurácia é a proporção da contagem dos resultados verdadeiros em relação ao número total de casos. Sendo assim, uma

métrica essencial e de fácil compreensão, adequada tanto para problemas de classificação binários como de classificação multi-classe, segundo Taylor [71], é dada matematicamente por:

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

No numerador, estão contemplados todos os exemplos que foram corretamente classificados pelo algoritmo. No denominador, compreende-se todo o conjunto de amostras.

A acurácia, por si só, não proporciona uma avaliação completa da qualidade das previsões do modelo, pois se limita a indicar a probabilidade de previsões corretas, sem aprofundar na qualidade dessas previsões.

A precisão é a taxa de verdadeiro positivos em relação ao total dos positivos (verdadeiros e falsos) que foram preditos corretamente. Essa métrica pode ser utilizada quando os dados se encontram bem equilibrados e não apresentam um viés pronunciado. No entanto, é reconhecido que existem outras métricas mais apropriadas para diferentes tipos de amostras. Sua expressão é dada da seguinte forma:

$$P = \frac{TP}{TP + FP} \tag{4}$$

A sensibilidade é uma métrica de grande relevância quando o objetivo do modelo é identificar um evento que ocorre com uma frequência menor que outro evento, ou seja, quando a base é desbalanceada. Através dela é mostrada a proporção em que o modelo está classificando com precisão os verdadeiros positivos [74]. Matematicamente é representada por:

$$R = \frac{TP}{TP + FN} \tag{5}$$

Por fim, a *F1-score* é uma métrica que representa a média harmônica entre precisão e recall, variando entre 0 e 1. Quanto mais próximo de 1 for o valor do *F1-score*, maior é a qualidade do modelo de aprendizado de máquina. Essa métrica oferece uma avaliação mais confiável da performance do modelo, considerando tanto a precisão quanto a capacidade de recuperação das informações relevantes.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

A *F1-score* foi escolhida para ser utilizada como critério de comparação entre as pipelines geradas. Segundo Taylor em [71], essa medida é a mais confiável para avaliar o desempenho real do modelo, uma vez que leva em consideração a média harmônica entre precisão e sensibilidade.

# Capítulo 3 – Metodologia

No capítulo anterior foi apresentada toda a fundamentação teórica utilizada para o desenvolvimento do modelo proposto, sendo apresentada uma descrição a respeito do impacto causado pelas perdas não técnicas nas distribuidoras, técnicas para combate a este ofensor, modelos de *Machine Learning* e suas aplicações. Contudo, existem necessidades de exposição de como essas técnicas serão utilizadas em conjunto para a construção do modelo proposto.

Diante da necessidade delineada no parágrafo anterior, este capítulo tem como propósito apresentar ao leitor a metodologia elaborada para a construção do modelo, bem como os ajustes e tratamentos dos dados referentes às estratégias adotadas. Nesse contexto, a etapa de processamento ganha destaque, abrangendo o fundamental processo de treino e teste do modelo.

## 3.1 Metodologia Desenvolvida

Esta seção tem como objetivo principal a exploração e apresentação da estrutura do modelo proposto nessa dissertação. Esse novo modelo é concebido como uma resposta às limitações identificadas no sistema correntemente utilizado pela concessionária, visando aprimorar sua eficácia e com isso contribuindo para melhoria das operações e das inspeções conduzidas pela distribuidora de energia elétrica.

Neste contexto, o fluxograma que resume a metodologia desenvolvida é apresentado na Figura 7, com a finalidade de fornecer uma representação visual clara e concisa da organização e funcionamento desse novo modelo.

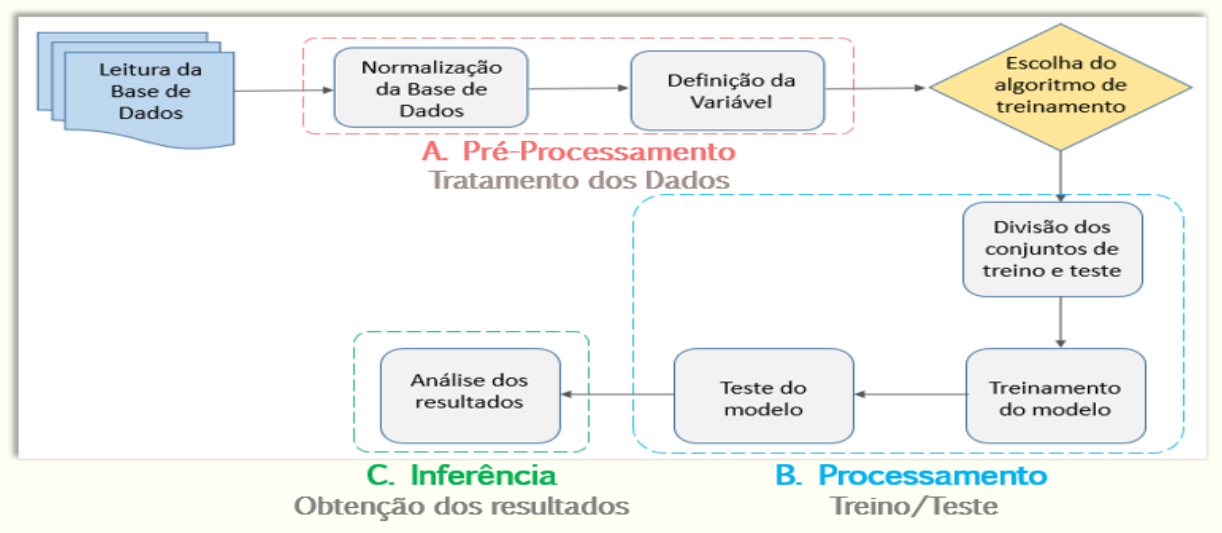


**Figura 7. Fluxograma da metodologia desenvolvida**
**Fonte: Própria**

Observe que o fluxograma da metodologia desenvolvida é estruturado em três etapas fundamentais:

- **Etapa A - Pré-Processamento:** nesta fase, ocorre a leitura das variáveis de entrada e o tratamento dos dados, bem como a extração e caracterização das principais *features* relevantes para a análise.
- **Etapa B - Processamento:** Neste ponto, focaliza-se a fase de execução do treinamento e dos testes.
- **Etapa C – Inferência:** Na fase final, são obtidos os resultados da classificação, consolidando as conclusões e inferências do modelo desenvolvido.

Essa estrutura proporciona uma abordagem clara e sequencial, permitindo uma compreensão precisa do desenvolvimento metodológico adotado nesta pesquisa.

## 3.1.1 Pré-Processamento

Nesta etapa, inicialmente é feita a análise de todos os dados de entrada do modelo. Estes dados abarcam diversos elementos, como a variação do vetor da janela deslizante, o número de visitas calculadas, o total de inspeções realizadas, categorias "Nada Apurado (NA)", "Não Inspecionados (NI)" e informações referentes ao bairro, totalizando seis insumos. Mais informações sobre suas representações e conceitos serão mostrados a seguir:

1. **Variação do vetor da janela deslizante:** refere-se à variação dos valores de consumo mensal de energia elétrica de uma unidade consumidora dentro de um período de três meses (janela) pré-definido. Seu valor é calculado pela fórmula:

$$\text{Variação} = (\text{Máximo} - \text{Mínimo}) / \text{Média} \quad (7)$$

2. **Número de visitas realizadas:** refere-se ao número total de visitas realizadas pelos leituristas em uma unidade consumidora durante um período de janeiro a dezembro de 2020.

3. **Número de inspeções realizadas:** refere-se ao número total de inspeções realizadas pelos leituristas em uma unidade consumidora durante um período de janeiro a dezembro de 2020.

4. **Nada apurado (NA):** refere-se às notas de serviço registradas pelas equipes de campo indicando que não foram encontradas irregularidades ou fraudes na unidade consumidora.

5. **Não inspecionados (NI):** refere-se às notas de serviço registradas pelas equipes de campo indicando que a unidade consumidora não foi inspecionada por algum motivo, como endereço não encontrado, área de risco, impedimento de acesso, entre outros.

6. **Bairro:** corresponde a localização em que a unidade consumidora está instalada. Importante ressaltar que para a abrangência das UCs consideradas no estudo corresponde a 20 municípios pertencentes ao Estado do Rio de Janeiro.

Estas variáveis estão representadas na Figura 8.

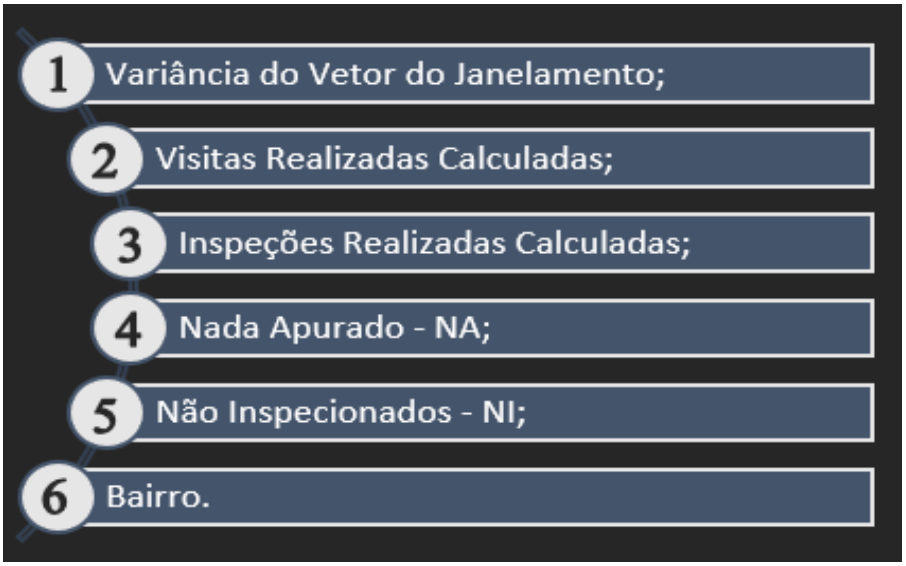


**Figura 8. Variáveis de Entrada do Modelo.**
**Fonte: Própria**

Ao compreender a definição de cada variável utilizada como entrada do modelo, a Figura 9 apresenta o fluxograma do pré-processamento, enumerando claramente as seis variáveis escolhidas.

É importante notar que as três primeiras variáveis, identificadas como NA, NI e Localização das UCS, são extraídas do cadastro interno dos clientes e podem ser facilmente acessadas por meio do fechamento das notas de serviços. A quarta variável, denominada "janela deslizante", representa a variação trimestral agrupada, cuja explicação detalhada ocorrerá mais adiante, onde os dados são tratados e ajustados. Por fim, a quinta e sexta variável, "cálculo – visita e inspeção", passará também por pequenos ajustes com objetivo de evitar que o modelo antecipe informações de fraude ou irregularidade já verificadas anteriormente na etapa de treino.

Esse processo também será detalhado mais a fundo na etapa de processamento.

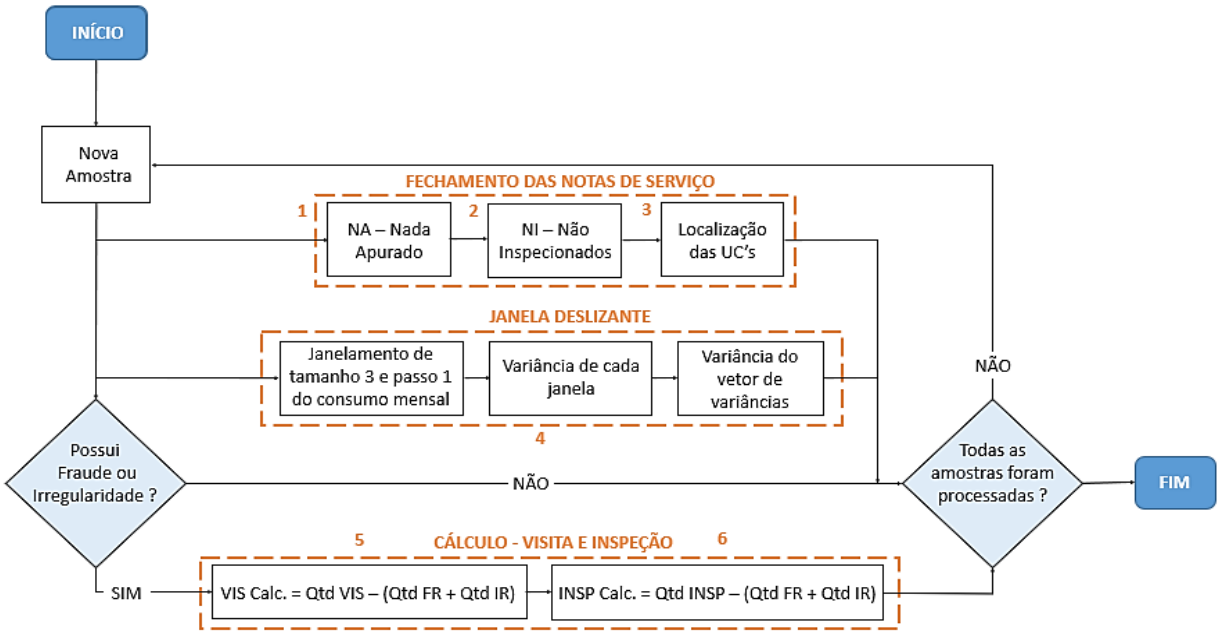


**Figura 9. Fluxograma: Pré-Processamento ou Extração de característica**

**Fonte: Própria**

## 3.1.1.1 Tratamento dos Dados

A etapa de tratamento dos dados desempenha um papel crucial em modelos de *Machine Learning*, exercendo influência direta nos resultados esperados. Durante esta etapa, diversos procedimentos são aplicados, como limpeza de dados, tratamento de valores ausentes, normalização e extração de características relevantes. Essas práticas visam não apenas melhorar

a estabilidade do funcionamento do modelo, mas também garantir que ele seja capaz de identificar padrões significativos nos dados, contribuindo para uma tomada de decisão mais precisa e confiável.

Dentro desse aspecto, duas estratégias são adotadas:

1. a janela deslizante;
2. e a inspeção.

A aplicação dessas duas estratégias visa tratar três das seis variáveis de entrada do modelo (janelamento do consumo, visitas e inspeções calculadas), visando alcançar resultados superiores. Os padrões escolhidos foram submetidos a testes e comparações com outros métodos, revelando um desempenho superior, cujo resultado será detalhado posteriormente neste documento.

Quanto a representação das estratégias comentadas, importante iniciar com um breve comentário sobre o modo de janela deslizante. É importante ressaltar que, antes da determinação da estratégia a ser adotada, foram conduzidos diversos testes distintos visando alcançar resultados satisfatórios.

No âmbito desse desafio, iniciou-se com a aplicação da diferença de consumo. Este método tem como propósito avaliar a discrepância entre o consumo mensal registrado do cliente e o do mês anterior. Caso essa diferença ultrapassasse os 40% do consumo lido, seria sinalizada uma possível fraude. No entanto, essa abordagem apresenta a limitação de requerer uma análise mês a mês, além de considerar uma diferença de 40% como relativamente baixa.

Como parte da análise, também foram realizados testes com a estratégia de janelamento, que aborda o mesmo conceito mencionado anteriormente, mas analisando a variação do consumo por meio de um janelamento, representado por um agrupamento de três meses, cuja variação seja limitada a 60% do consumo médio registrado. A estratégia de janelamento, que analisa a variação do consumo por meio de um agrupamento de três meses, é adotada na análise como um método para capturar padrões sazonais ou cíclicos no uso de energia, que podem ser indicativos de comportamento irregular ou fraudulento.

No exemplo apresentado na Figura 10, a formação da primeira janela ocorrerá pelo somatório do consumo do mês 1 ao mês 3. Posteriormente, o modelo progredirá para o mês subsequente, mantendo a ordem de avanço, e assim formará a segunda janela, composta pelo consumo do mês 2 ao mês 4.

Essa metodologia é aplicada para todos os meses do ano de 2020, utilizando uma janela móvel de 3 meses para calcular o consumo médio do cliente. A premissa subjacente é que a média de consumo não deve variar além do 60% predeterminado, caso ultrapasse essa margem, será considerado um indício de possível fraude.

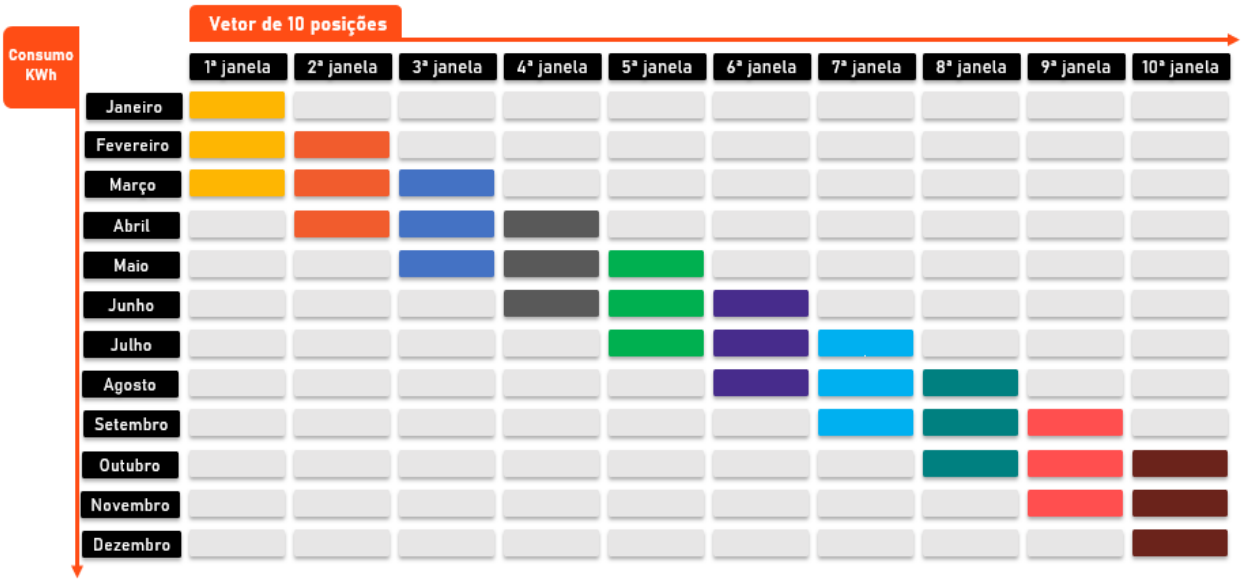


**Figura 10. Representação: Janelamento – 12 meses**
**Fonte: Própria**

Nesse contexto, o modelo avança através da criação de grupos de janelas deslizantes, abrangendo o período completo de estudo até o décimo segundo mês do ano. Esse método, que resulta na formação de 10 janelas deslizantes consecutivas, é crucial para a análise detalhada dos padrões de consumo ao longo do ano. Essa abordagem permite identificar variações sazonais e tendências significativas, contribuindo não apenas para a detecção de possíveis irregularidades, mas também para uma compreensão mais aprofundada do comportamento de consumo dos clientes. Além disso, ao considerar um período extenso, torna-se possível capturar nuances e mudanças sutis que poderiam passar despercebidas em análises mais limitadas.

Essa aplicação estratégica da janela deslizante potencializa a capacidade do modelo em fornecer insights valiosos para a gestão eficaz do consumo de energia elétrica e detecção de possíveis fraudes.

Com base nessa estratégia, o modelo incorporará como uma de suas entradas um vetor único de dez posições, constituído por esse conjunto de dez variâncias construídas. Esse vetor desempenhará um papel fundamental como uma das entradas no modelo de classificação, uma vez que essa abordagem revelou um desempenho superior em comparação com outras

estratégias testadas, como, por exemplo, a utilização de uma variância por trimestre. Além disso, a capacidade de empregar um único vetor para representar o período de doze meses analisados simplifica significativamente o desenvolvimento do modelo, proporcionando eficiência e coerência na análise dos padrões de consumo ao longo do ano.

O cálculo de estimativa de variação do consumo trimestral em torno do seu valor médio é representado pela fórmula a seguir:

$$VAR = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{n} \quad (8)$$

Onde:

- $x_1$, $x_2$ e $x_3$ : representam o consumo mensal;
- $\bar{x}$: consumo médio dentro da janela deslizante.
- $n$: número de elementos (mês).

Durante a fase de preparação dos dados, a segunda estratégia adotada foi o método de inspeção. Este método foi aplicado exclusivamente aos clientes que apresentaram registros de fraudes, estabelecendo uma relação entre duas entradas fundamentais do modelo: Visitas Realizadas Calculadas e Inspeções Realizadas Calculadas.

O propósito central consiste em identificar possíveis correlações entre a quantidade de visitas ou inspeções realizadas e os apontamentos de fraude. Nesse contexto, o resultado de cada cálculo desempenha um papel crucial na avaliação, atribuindo maior propensão à fraude a clientes cuja diferença se aproxima de zero e menor propensão a fraude quando a diferença é superior a três. O cálculo aplicado é o seguinte:

$$VIS\ Calculadas = Tot.VIS - (FR + IR) \quad (9)$$

$$INSP\ Calculadas = Tot.INSP - (FR + IR) \quad (10)$$

Onde:

- Tot. VIS: total de visitas realizadas;
- Tot. INSP: total de inspeções realizadas;
- FR: nº de fraudes encontradas;
- IR: nº de irregularidades encontradas;

A aplicação das estratégias mencionadas anteriormente é demonstrada na Figura 11, após organização e ajustes cuidadosos. Como salientado no início desta seção, é crucial destacar que ambas as estratégias, somadas às outras três variáveis de entrada (NA, NI e localização), compõem o conjunto abrangente de variáveis utilizadas como entrada para o modelo de

classificação. Essa abordagem integrada é essencial para fornecer ao modelo uma visão holística e abrangente dos padrões de consumo e características específicas dos clientes. Essas variáveis, juntas, desempenham um papel crucial na eficácia do modelo, como será explorado em detalhes no próximo capítulo, que abordará a fase de treinamento do modelo e as premissas adotadas para a condução dos testes.

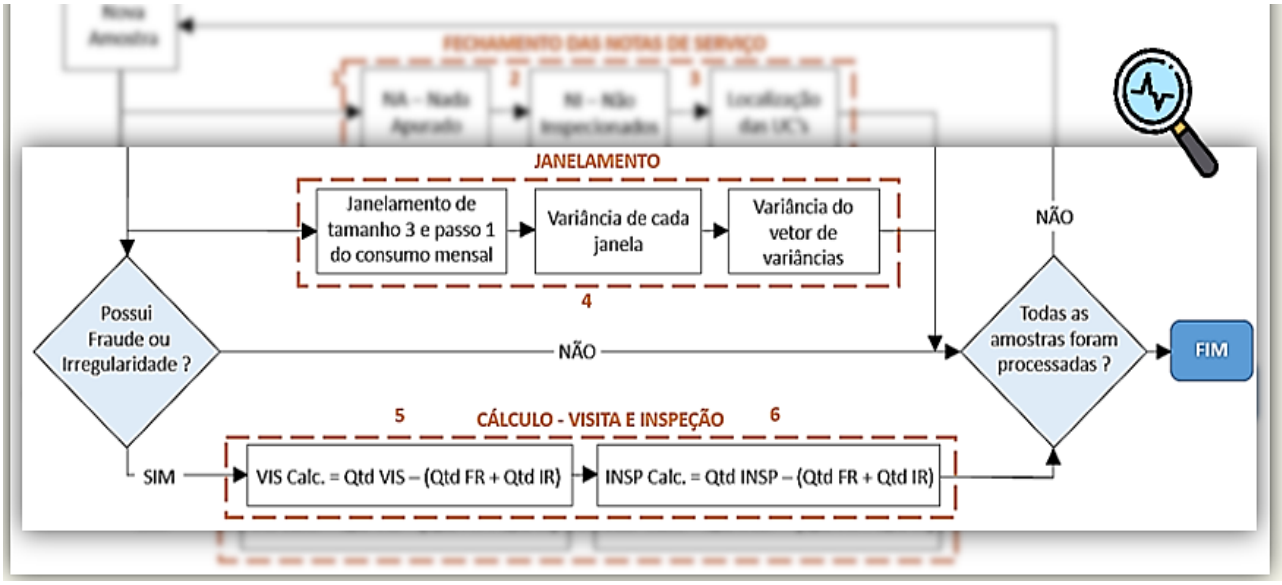


*Figura 11. Fluxograma: Pré-Processamento ou Extração de característica*

**Fonte: própria**

## 3.1.2 Processamento

Na seção anterior, dedicada ao tratamento e limpeza de dados, foram abordadas as estratégias essenciais para preparar o conjunto de dados utilizado no estudo. Agora, neste novo capítulo, será dedicada atenção à etapa de processamento do modelo de *Machine Learning*.

Esta fase desempenha um papel fundamental em moldar a eficácia do modelo, influenciando diretamente a capacidade do modelo em fornecer insights precisos e úteis. Destaca-se que nesta etapa, será abordada a etapa de treinamento e teste do modelo construído. A eficácia dessa etapa é vital para assegurar que o modelo seja robusto, generalizável e capaz de enfrentar desafios no contexto real de detecção de fraudes em instalações elétricas.

### 3.1.2.1 Treinamento do Modelo

Para aplicação desta etapa foi utilizado um processador Intel(R) Core (TM) i5-10210U CPU @ 1.60GHz 2.11 GHz, 64 bits e 8GB RAM e sistema operacional *Windows Feature Experience Pack*, com a metodologia sendo implementada em linguagem de programação *Python*.

Vale ressaltar que para tal análise foi utilizada a técnica de validação cruzada *K-fold*. O conceito central desta técnica é o particionamento do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, o uso de alguns destes subconjuntos para a estimação dos parâmetros do modelo na fase de treinamento, sendo os subconjuntos restantes empregados na validação e teste do modelo [62]. Este método consiste em dividir o conjunto total de dados em *k* subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir daí, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para estimação dos parâmetros, fazendo-se o cálculo da acurácia do modelo[62].

Com isso, definiu-se a priori que o número de grupos a serem utilizados para a divisão dos padrões disponíveis para o treinamento/validação fosse de três. É importante citar que, para cada um dos três *folds* pré-estabelecidos haverá a divisão dos conjuntos de dados em treino e testes. Através desta divisão, os conjuntos de treino passaram a compor 2/3 da base, enquanto o conjunto de teste será definido por 1/3 da mesma.

Desta forma, para cada um *fold*, será gerado um modelo onde será testado e terá o seu valor registrado. Por fim, se terá três resultados de testes diferentes definidos, o que significa dizer que toda a base será verificada. A Figura 12 ilustra esse processo.

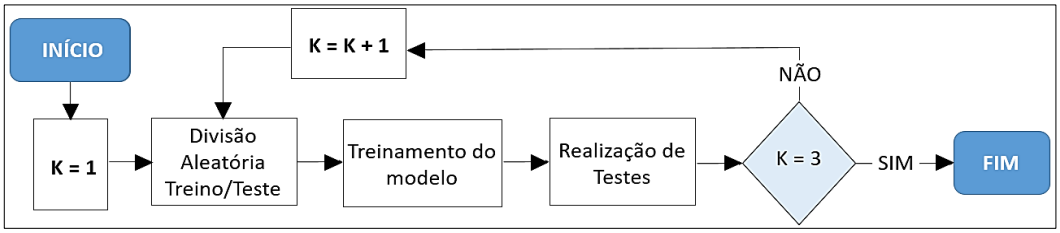


**Figura 12. Fluxograma: Treinamento.**

**Fonte: Própria**

Durante este processo as informações dos consumidores são transmitidas ao modelo, que por sua vez tem seus parâmetros ajustados. A base de dados resultante é empregada no treinamento do módulo de classificação, o qual é responsável por indicar se o consumidor

pertence à classe regular (normal) ou irregular (fraudador), correspondendo a um problema de classificação binária. Ao longo do processo de treinamento, identificou-se uma característica desbalanceada nos dados utilizados, devido à disparidade quantitativa entre as classificações de clientes considerados fraudadores e regulares. Essa discrepância torna desafiante para o algoritmo a tarefa de distinguir eficientemente entre as classes. Como consequência, o modelo pode não desenvolver a capacidade de generalização essencial para lidar com o conjunto de dados de teste, correndo o risco de apresentar o indesejado efeito de sobreajuste (*overfitting).*

A técnica de sobreamostragem minoritária sintética (SMOTE) e o modelo Near-miss são duas abordagens comumente utilizadas para lidar com desbalanceamento de classes em problemas de classificação [62], porém para o trabalho em questão, optou-se pela técnica de subamostragem, por alguns motivos que serão delineados a seguir.

Primeiramente, a base de dados utilizada já possuía um número relativamente grande de exemplos, o que permitiu a utilização da subamostragem sem que houvesse uma perda significativa de informação. Além disso, a subamostragem é uma abordagem mais simples e rápida do que o uso de técnicas de sobreamostragem ou modelos de aprendizado de máquina específicos para lidar com desbalanceamento de classes. No âmbito desse contexto, a base utilizada durante o teste compreende 52.608 clientes regulares e 27.392 clientes com registros de fraude.

## 3.1.2.2 Aplicação do Modelo ao Conjunto de Teste

No âmbito desta seção, destaca-se a relevância dos resultados apresentados durante a análise comparativa, realizada por meio das aplicações individuais dos algoritmos escolhidos, a saber: *Support Vector Machine* (SVM), *Xgboost* e *Random Forest* (RF).

Essa análise revelou-se crucial para a definição do modelo de classificação a ser adotado no estudo, uma vez que proporcionou resultados não apenas assertivos, mas também demonstrou um desempenho interessante quando comparado com o *baseline* correntemente utilizado pela concessionária. A minuciosa avaliação dos resultados de cada algoritmo desempenhou um papel decisivo na escolha do modelo mais eficaz para a detecção de clientes fraudadores. Além disso, ofereceu uma base sólida para compreender as nuances de desempenho de cada abordagem no contexto específico em questão.

Inicialmente, nos testes optou-se por empregar o *Support Vector Machine* (SVM), um algoritmo notório por sua eficácia na classificação de conjuntos de pontos segundo a literatura

[13]. Na tabela 3 serão apresentados os resultados obtidos por meio da matriz de confusão gerada e suas respectivas métricas, enriquecendo ainda mais a análise e permitindo uma avaliação mais aprofundada da capacidade de detecção de clientes fraudadores pelo modelo testado.

**Support Vector Machine:**

| F1-Score | Accuracy | Precision | Recall | Matriz de Confusão |
|---|---|---|---|---|
| 50,150% | 70,138% | 58,528% | 43,870% | [[44093 8515]<br>[15375 12017]] |

**Tabela 3. Resultado Teste Algoritmo SVM.**

Neste segundo momento, a ênfase recai sobre o algoritmo *XGBoost*, uma abordagem fundamentada em árvores de decisão e estruturada com *Gradient Boosting* para otimização [17]. A proposta é avaliar minuciosamente os resultados alcançados por meio desta abordagem, ampliando ainda mais a compreensão sobre o desempenho do modelo. A tabela 4 apresenta algumas estatísticas para esse modelo.

**XGBoost:**

| F1-Score | Accuracy | Precision | Recall | Matriz de Confusão |
|---|---|---|---|---|
| 66,590% | 74,525% | 60,432% | 74,146% | [[44093 8515]<br>[15375 12017]] |

**Tabela 4. Resultado Teste Algoritmo XGBoost.**

Após a realização de testes com os dois algoritmos apresentados, foi aplicado também o algoritmo *Random Forest* (RF), cujo desempenho está descrito na tabela 5:

**Random Forest:**

| F1-Score | Accuracy | Precision | Recall | Matriz de Confusão |
|---|---|---|---|---|
| 77,997% | 84,840% | 75,534% | 80,626% | [[45485 7123]<br>[5535 21857]] |

***Tabela 5. Resultado Teste Algoritmo RF.***

Em virtude dos resultados obtidos, este algoritmo foi selecionado como o modelo classificador adotado no estudo. A escolha se fundamenta na notável eficácia do RF, evidenciada pelos resultados superiores alcançados em comparação com os demais algoritmos

testados. Este desfecho ressalta a importância dessa etapa de análise comparativa para a tomada de decisões fundamentadas na escolha do modelo mais apropriado para a detecção de clientes fraudadores no contexto em questão.

## 3.2 Linguagens de Programação

No que se refere às abordagens previamente discutidas acerca dos diversos paradigmas de aprendizado e suas múltiplas aplicações, surge uma imperiosa necessidade de estabelecer um ambiente de programação apropriado para o desenvolvimento da solução em questão. A sua utilização faculta ao ser humano a capacidade de interagir com a máquina, permitindo a emissão de comandos e a análise dos dados oriundos do software providos pelo sistema, conforme atestam em [69].

Atualmente no mercado encontra-se uma pluralidade de linguagens de programação, a maioria delas versáteis e aplicáveis em diversos domínios. A escolha da linguagem apropriada está intrinsecamente ligada à natureza do problema a ser solucionado, uma vez que cada uma delas possui características particulares, apresentando vantagens e desvantagens distintas. Neste contexto, para a presente dissertação, optou-se pela linguagem Python, reconhecida por estudiosos em [54] como uma linguagem de programação de alto nível, munida de bibliotecas abrangentes em uma ampla gama de áreas de aplicação. A combinação dessas bibliotecas em um ambiente Python proporciona uma base sólida e flexível para a construção, treinamento e avaliação do modelo, garantindo eficiência e precisão nos resultados obtidos. Nesse contexto, adiante segue as principais bibliotecas utilizadas no estudo e discutidas em [76, [77] e [78]:

- *Pandas*: torna o trabalho com dados mais fácil e intuitivo, fazendo uso de sintaxe de alto nível para fazer análises e manipulação dos dados de maneira prática [76].

- *NumPy*: está no centro dos ecossistemas científicos para *Python* e adiciona estruturas de dados a esta linguagem que garantem cálculos eficientes com vetores e matrizes, fornecendo uma gama de funções matemáticas de alto nível. É amplamente utilizada em *Pandas, SciPy, Matplotlib, scikit-learn* e na maioria dos outros pacotes de ciência de dados [77].

- *MatplotLib*: utilizada para criar visualizações estáticas, animadas e interativas [78];

*Scikit-learn*: desenvolvida especificamente para aplicação prática de *Machine Learning*, esta biblioteca dispõe de ferramentas simples e eficientes para análise preditiva de dados, principalmente por utilizar outras bibliotecas eficientes (*NumPy, SciPy e MatplotLib*) nas suas respectivas funções [78].

# Capítulo 4 – Resultados

Neste capítulo, antes de serem apresentados os resultados obtidos, será feita uma breve explanação sobre a abrangência geográfica da distribuidora de energia elétrica Light S.A, bem como os dados fornecidos que serviram de atributos para o modelo, enfatizando a relevância desse tópico.

Em seguida, serão discutidos os resultados alcançados em dois contextos distintos. Primeiramente, serão abordados os resultados obtidos no conjunto de teste e, posteriormente, serão examinados os resultados no contexto do conjunto de dados da prática real.

## 4.1 Análise Exploratória dos Dados

A área de concessão da Companhia abrange cerca de 26% (11.307 mil km²) do Estado do Rio de Janeiro e comporta uma população de 11 milhões de pessoas, representando 64% da população total do Estado. Dos 92 municípios do Estado com um total de 7 milhões de consumidores de energia elétrica, a Companhia atua em 31 municípios, representando 34% dos municípios totais, e possui uma base de cerca de 4,5 milhões de clientes [15].

A abrangência das unidades consumidoras (UCs) consideradas neste trabalho corresponde a um volume de 80 mil instalações escolhidas aleatoriamente pertencentes ao grupo B convencional (baixa tensão) e inspecionados no ano de 2020 (janeiro a dezembro), cujo resultado das inspeções corresponde a 52.608 registros da classe de clientes ditos regulares e 27.392 irregulares.

## 4.2 Atributos Utilizados

Para a elaboração do presente trabalho foram utilizados dados fornecidos pela concessionária de energia elétrica Light S.A. e são originários de diversas bases de dados que contêm atributos gerais dos clientes. Esta extração respeitou integralmente as condições previstas na Lei Geral de Proteção de Dados [68]. A Tabela 6 elenca uma lista com os atributos disponíveis que foram utilizados para desenvolvimento do modelo:

| Atributo | Descrição |
|---|---|

| | |
|---|---|
| Dados da Unidade Consumidora | Localização geográfica da Unidade Consumidora (UC). |
| Dados de Consumo | Histórico de consumo das unidades consumidoras, consumo medido e faturado. |
| Notas de leitura | Apontamento de irregularidade constatada pela equipe de campo |
| Notas de Serviço | Fechamento da nota de serviço executado pela Equipe |

**Tabela 6. Atributos Selecionados na Base de dados.**

Para o melhor entendimento do leitor, torna-se válido apresentar as definições das variáveis supracitadas, que são:

- **Dados da unidade consumidora:** são informações que podem ser obtidas por meio de formulários de cadastro e são úteis para caracterizar melhor a unidade consumidora (UC). Essas informações incluem, por exemplo, se a unidade consumidora está associada a uma pessoa física ou jurídica, sua localização geográfica e a classe de consumo (residencial, comercial, industrial ou público).

- **Dados de consumo:** incluem a data de referência em que ocorreu a leitura, a energia consumida ou medida e o consumo faturado. Essas informações são especialmente importantes em termos tarifários e na definição do modelo de aprendizado de máquina. As informações de leitura e serviço são utilizadas pelas distribuidoras de energia elétrica brasileiras para avaliar unidades consumidoras de baixa tensão em busca de identificação de irregularidades.

- **Notas de leitura:** são apontamentos realizados pelos leituristas durante a verificação em campo e servem como insumo para o sistema de seleção de alvos (possíveis fraudadores) correntemente utilizado pela distribuidora.

- **Notas de serviço:** caracterizam a real condição do cliente vista pela equipe durante a inspeção, indicando irregularidades, fraudes, nada apurado, entre outras situações.

## 4.3 Ferramenta de Seleção da Distribuidora

A aplicação de uma ferramenta voltada para a seleção de clientes destinados à inspeção de campo é um componente essencial na administração de perdas em uma distribuidora de energia elétrica. Seu papel deve ser estratégico na identificação de casos suspeitos de irregularidades, possibilitando a otimização dos recursos operacionais. Contudo, é imperativo exercer cautela durante a análise de seleção, pois um apontamento incorreto para inspeção pode acarretar prejuízos significativos.

Explorando mais detalhadamente o modelo atual de seleção da distribuidora Light S.A, pertinente ao escopo do estudo em análise, é essencial ressaltar que o funcionamento da ferramenta opera por meio de três vertentes fundamentais:

1. Geração de notas via sistema de inteligência;

2. Seleção dos apontamentos do modelo conforme a disponibilidade operacional, para a alocação das notas de serviço às equipes de campo; e

3. Inspeção e normalização dos clientes indicados.

Importa destacar que esse processo é cíclico, visto que os resultados das inspeções realizadas pelas equipes de campo retroalimentam o modelo.

Dentro do processo comentado anteriormente, o foco principal do estudo recai sobre a primeira etapa, que diz respeito aos apontamentos do modelo em relação aos clientes suspeitos de fraude. É crucial ressaltar que esse procedimento incorpora elementos heurísticos, os quais integram conhecimento especializado no problema. Essa abordagem visa refinar as sugestões geradas pelo sistema, visando aprimorar a precisão das inspeções realizadas. Esse refinamento, baseado em expertise, desempenha um papel crucial na otimização do processo de seleção, porém a complexidade de alcançar resultados eficazes foi evidente ao longo do ano de 2020. Este cenário enfatiza a necessidade de uma avaliação crítica e contínua do modelo, visando identificar e corrigir lacunas, promovendo melhorias constantes para ampliar a efetividade na detecção de possíveis irregularidades.

Dentro desse contexto, é essencial frisar que o modelo de seleção atual adotado pela distribuidora de energia Light engloba não apenas as informações previamente discutidas,

provenientes da experiência de especialistas na área, mas também incorpora diversos critérios específicos, os quais foram incorporados no sistema como regras. Vale salientar que, por questões de privacidade, a divulgação de alguns desses critérios não é permitida. No entanto, entre os critérios mencionados, destacam-se dois dos mais comuns utilizados pelas distribuidoras: a aplicação da regra degrau em conjunto com a regra do consumo congelado. A regra degrau identifica quedas abruptas no consumo, atingindo até 60% do registrado no mês anterior pelo cliente, enquanto a regra do consumo congelado indica a ausência de variação no consumo medido. Essas regras, quando combinadas, atuam como indicadores de suspeita de fraude.

Para uma compreensão mais clara dessas regras, a Figura 13 apresenta a interface visual correspondente, onde o eixo X representa os meses do ano, e o eixo Y representa o consumo medido registrado em kWh. É possível observar que, ao identificar uma redução no consumo ao longo de dois meses e a ausência de variação do consumo nos meses posterior a queda, o cliente é encaminhado para uma inspeção de campo, a fim de realizar uma investigação mais aprofundada.

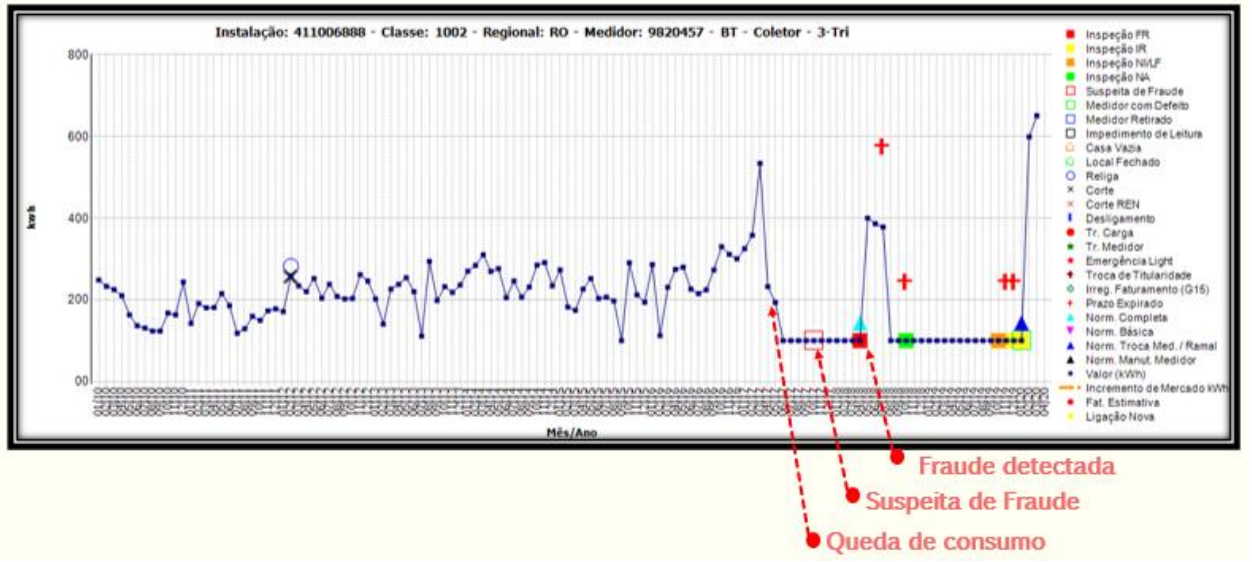


**Figura 13. Exemplo das regras: degrau e consumo congelado**
**Fonte: Própria**

## 4.3.1 Características do Processo de Seleção

Nesta seção, serão abordadas as premissas e conceitos fundamentais do modelo de seleção da distribuidora, destacando sua integração com outros processos internos.

Entre esses processos, o analítico assume um papel essencial, englobando regras criteriosas, a ativação de alarmes e uma variedade de análises minuciosas. Este processo visa aprofundar a compreensão sobre possíveis casos de irregularidades e fraudes. Paralelamente, o processo administrativo, voltado para as perdas administrativas e informações sistêmicas, desempenha um papel complementar, fornecendo uma visão mais ampla do panorama operacional da distribuidora. A interconexão desses processos é crucial para a eficácia global do modelo, permitindo uma abordagem abrangente no gerenciamento de perdas.

Os pontos anteriormente discutidos são ilustrados na Figura 14:

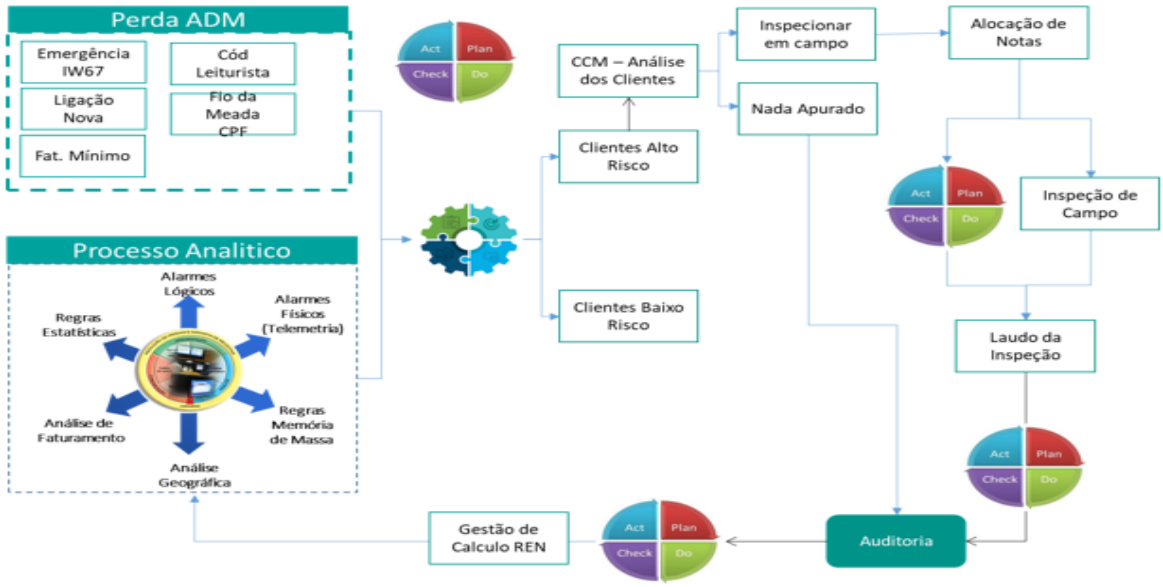


**Figura 14. Macrofluxo do monitoramento de fraudes – BT.**
**Fonte: Própria**

Conforme evidenciado na Figura 19, o modelo atual adotado pela distribuidora é capaz de classificar os clientes em dois grupos distintos. O primeiro grupo é composto por clientes de baixo risco, identificados pelo baixo retorno de faturamento após a normalização. Por outro lado, o segundo grupo engloba clientes considerados de alto risco, caracterizados por apresentarem um retorno significativo de energia após a normalização. Após a análise interna, o processo prossegue com o encaminhamento para a equipe de campo, que executa as inspeções, emite laudos e, em seguida, submete os casos à auditoria e apuração do cálculo da

energia recuperada. Vale ressaltar que o fluxo referente ao grupo de clientes de baixo risco não é postergado devido ao seu baixo retorno, evidenciando a eficiência do modelo nesse contexto.

Em resumo, o processo representado no fluxograma é resultado de um esforço conjunto que envolveu a experiência de especialistas e a análise de dados. Não foi de fácil construção devido à sua complexidade e à necessidade de conhecimento profundo do domínio. Embora possa demandar intervenção do usuário em alguns pontos, o seu objetivo é o uso automatizado, para que assim, torne o sistema mais eficiente e eficaz.

## 4.3.2 Resultados do Sistema de Seleção

Até o momento, foram apresentadas as premissas, conceitos e critérios do processo do modelo de seleção atual da distribuidora. Nesta seção, serão destacados os resultados de sua eficácia, além de abordar como a baixa assertividade do modelo pode impactar no contexto de perda e arrecadação.

A Tabela 7 apresenta os resultados da análise da assertividade das inspeções nas cinco regionais nos anos de 2020 e 2021. Notavelmente, observa-se um crescimento significativo de 26% para 38% na média regional. No entanto, é crucial ressaltar que, mesmo com esse aumento, o modelo ainda apresenta limitações que exigem melhorias substanciais.

| Regionais | Ano (2020) | Ano (2021) | Média (%) |
|---|---|---|---|
| Regional Leste | 21% | 31% | 26% |
| Regional Centro Sul | 18% | 12% | 15% |
| Regional Vale | 17% | 19% | 18% |
| Regional Baixada | 47% | 48% | 48% |
| Regional Oeste | 38% | 42% | 40% |
| Média Regional | 26% | 38% | 32% |

**Tabela 7. Acerto médio do sistema utilizado na Distribuidora**

No âmbito técnico da área de perdas, é imperativo observar atentamente alguns pontos críticos. A Regional Baixada por exemplo, com sua alta consistência de 48%, pode se beneficiar de estratégias específicas para aprimorar ainda mais sua eficácia. A Regional Centro Sul, que registrou uma queda de 18% para 12%, destaca-se como uma área que requer atenção e investigação aprofundada para identificar os motivos por trás dessa redução. Além disso, a média regional de 38% aponta para uma melhoria geral, mas é essencial considerar a

variabilidade entre as regionais e identificar padrões específicos que possam ser direcionados para otimização.

Com base nessas informações, o modelo desenvolvido nesta dissertação visa potencializar os ganhos obtidos, almejando resultados mais robustos e um direcionamento mais preciso para as equipes de campo. Este contexto ressalta a contínua importância do modelo, que busca melhorias significativas no sistema atual.

A necessidade de aumentar a assertividade das inspeções é vital para otimizar as operações de identificação de possíveis fraudes, contribuindo assim para a redução de perdas não técnicas. Apesar dos avanços já percebidos, a busca constante por aprimoramentos permanece como uma prioridade, visando alcançar maior eficiência e eficácia na identificação de casos suspeitos em todas as regionais.

Neste cenário, é essencial destacar a relevância da métrica conhecida como Valor Positivo Preditivo (VPP), que indica a proporção de clientes comprovadamente irregulares entre aqueles classificados como suspeitos. Conforme evidenciado na Figura 20, observa-se que, até o momento, a aplicação dessa metodologia resultou em um VPP inferior a 40%.

É pertinente salientar que quanto menor for o valor preditivo calculado, maior será o risco de não atingimento das metas de energia recuperadas com as programações das ações, uma vez que esse resultado reflete um direcionamento inadequado das equipes de campo, concentrando-se em locais onde não há perdas. Isso ressalta a importância de aprimorar a assertividade do modelo para otimizar os recursos operacionais.

Como exemplificação do mencionado, é importante ressaltar dois dos principais impactos resultantes da baixa assertividade:

1. Financeiro: diminuição na arrecadação, resultante da falta de faturamento dos clientes fraudadores.

2. Físico: redução na energia recuperada durante o período da fraude e diminuição na energia incorporada no mês subsequente à normalização.

Para melhor contextualizar o tema, a tabela 8 apresenta o desempenho geral do processo utilizado pela distribuidora durante o período de janeiro a maio de 2020.

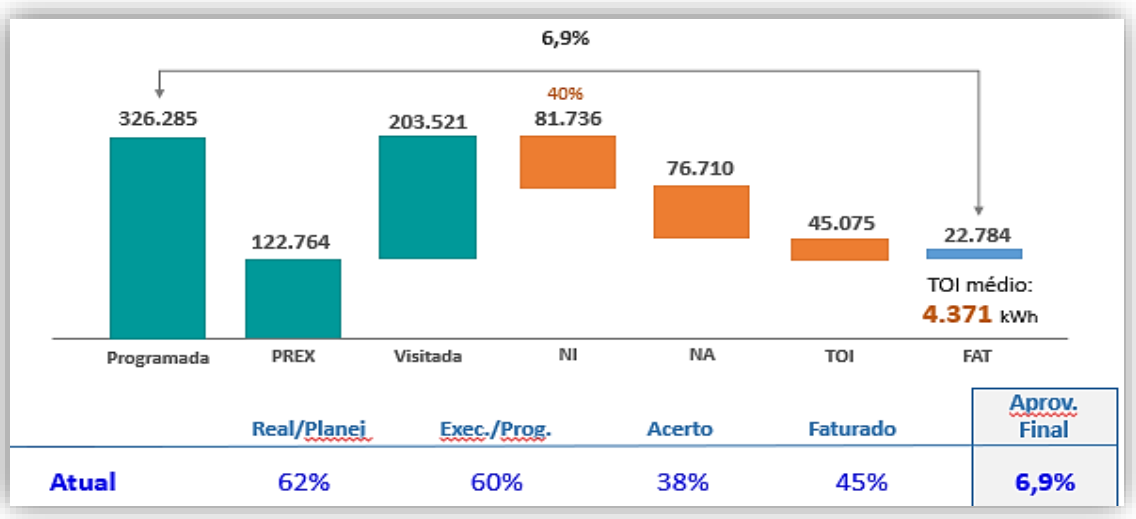


**Tabela 8. Panorama de Efetividade das Notas Geradas**

**Fonte: Própria**

Considerando as informações expostas na tabela 8, que representam a quantidade de notas de serviço programadas no período de janeiro a maio de 2020, provenientes do modelo de seleção de alvos da distribuidora de energia Light S.A., observa-se um cenário desafiador quanto à eficácia do modelo na identificação de clientes suspeitos de fraude.

Das 326.285 notas programadas para inspeção em campo, apenas 62,4% (203.521) foram efetivamente visitadas, destacando uma baixa taxa de cobertura. Dentro desse conjunto inspecionado, 37,7% (76.710) das notas não indicaram indícios de fraude, apontando para limitações significativas na capacidade do modelo de identificar casos suspeitos.

Destaca-se também, que 37,6% (122.764) das notas programadas não foram executadas devido ao prazo expirado, enquanto 25,2% (81.736) não foram inspecionadas. Dentro deste conjunto não inspecionado, 15,7% (32.396) são identificadas como áreas de risco, indicando possíveis desafios relacionados à segurança ou logística. Adicionalmente, 7,6% (15.617) referem-se a medições localizadas internamente, sem acesso à normalização, evidenciando desafios operacionais específicos para essa categoria de casos. Ambos os cenários apontam para ineficiências no agendamento ou na seleção dos locais a serem inspecionados.

Apesar da aplicação de 13,8% (45.075) de TOIs aos clientes identificados com fraude, o valor faturado de 6,9% (22.784) destaca uma possível desconexão entre as ações tomadas e os resultados financeiros associados. O TOI médio, equivalente a 13,4% de energia (4.371 kWh), proporciona uma média ponderada do impacto energético das intervenções realizadas. Esse panorama evidencia a necessidade de aprimoramentos substanciais na estratégia de seleção de clientes suspeitos, visando melhorar a eficiência e a eficácia do modelo para

melhorar a efetividade das operações de inspeção e ampliar o retorno financeiro das ações tomadas pela distribuidora.

## 4.4 Resultados do Modelo Construído

Esta seção tem como objetivo principal apresentar inicialmente a premissa dos parâmetros adotados durante a construção do modelo, bem como os resultados obtidos durante a etapa comparativa entre o modelo construído no decorrer deste estudo e o modelo adotado pela distribuidora. Além disso, será abordada a análise comparativa com a aplicação real de campo.

Essa avaliação é crucial para verificar a eficiência do modelo desenvolvido, destacando possíveis melhorias em relação ao modelo vigente e identificando lacunas que necessitam de ajustes para aprimorar a assertividade do sistema como um todo.

### 4.4.1 Hiperparâmetros

Os parâmetros, representados pelos pesos ajustáveis, são intrínsecos ao processo de treinamento, enquanto os algoritmos subjacentes exigem parâmetros específicos, como a taxa de aprendizagem. Contudo, para moldar a estrutura mais ampla do modelo e guiar efetivamente a busca pelos pesos ideais, torna-se essencial explorar os hiperparâmetros.

Na tabela 9, antecipando a apresentação da tabela, serão delineados os hiperparâmetros fundamentais considerados para cada modelo de *Machine Learning* explorado neste estudo: SVM, *XGBoost* e *Random Forest*.

| Algoritmo | Melhor Ajuste de Hiperparâmetros |
|---|---|
| Random Forest | Número Máximo de Profundidade (`max_depth`): 5<br>Número de Estimadores (`n_estimators`): 164 |
| SVM | Kernel: RBF<br>Custo (`C`): 1.0<br>Gama (`gamma`): auto |
| XGBoost | Número Máximo de Profundidade (`max_depth`): 2<br>Número de Estimadores (`n_estimators`): 453 |

***Tabela 9. Ajuste dos Hiperparâmetros.***

**Fonte: Própria**

Cada um desses modelos é distintivo por um conjunto exclusivo de hiperparâmetros que desempenham um papel crítico em sua eficácia. Esses hiperparâmetros, como taxa de aprendizado, profundidade da árvore, e número de estimadores, são fundamentais para a calibragem adequada do modelo e influenciam diretamente seu desempenho. A escolha apropriada desses hiperparâmetros é crucial para otimizar o modelo em termos de precisão, generalização e capacidade de adaptação a diferentes conjuntos de dados. Portanto, uma compreensão aprofundada desses hiperparâmetros é essencial para a configuração ideal e refinamento de cada modelo.

Para uma compreensão mais abrangente, os critérios de definição utilizados na determinação desses hiperparâmetros, mais especificamente do RF serão explicitados.

| Hiperparâmetro | Valor | Descrição |
|---|---|---|
| n_estimators | 164 | Número de árvores no modelo. |
| Criterion | 'gini' | Critério de divisão em cada nó da árvore. |
| max_depth | 5 | Profundidade máxima de cada árvore. |
| min_samples_split | 2 | Número mínimo de amostras para divisão em um nó. |
| min_samples_leaf | 1 | Número mínimo de amostras em folhas da árvore. |
| Bootstrap | True | Indica se a amostragem é realizada com substituição. |
| ccp_alpha | 0.0 | Parâmetro de complexidade de custo mínimo. |
| class_weight | None | Peso associado a cada classe no caso de classificação. |
| max_features | 'sqrt' | Número máximo de características consideradas para dividir um nó. |
| max_leaf_nodes | None | Número máximo de nós folha. |
| min_impurity_decrease | 0.0 | Valor mínimo necessário para uma divisão. |
| oob_score | False | Indica se deve ser usado o out-of-bag samples para estimar o R^2. |
| random_state | None | Semente para controle de aleatoriedade. |
| Verbose | 0 | Nível de detalhe das mensagens de saída. |
| max_samples | None | Número máximo de amostras a serem usadas para treinamento de cada árvore. |
| **Métrica** | | **Valor** |
| Folds | | 3 |
| Duração da Validação Cruzada (1/3) | | 8.74 min |
| Duração da Validação Cruzada (2/3) | | 19.12 min |
| Duração da Validação Cruzada (3/3) | | 26.86 min |
| Média da Pontuação | | 74.819% |

**Tabela 10. Premissas – Hiperparâmetros - RF**

## 4.4.2 Resultados do Conjunto de Teste Teórico

Com o intuito de realizar uma análise comparativa entre os modelos desenvolvidos neste estudo e o modelo em uso pela distribuidora, conduzimos um teste de eficiência que abrangeu uma base de 80 mil clientes. Para garantir a confiabilidade dos resultados e assegurar a precisão dos apontamentos, todas as 80 mil instalações analisadas por ambos os modelos foram submetidas à inspeção por equipes de campo, resultando em um veredito final para cada cliente inspecionado.

Conforme evidenciado durante a fase de processamento, a base de clientes em análise foi particionada em três grupos. Cada um desses grupos foi subdividido, sendo que 2/3 da base foram alocados para os conjuntos de treino, enquanto o conjunto de teste foi composto por 1/3 do total. Uma representação visual dessa subdivisão é apresentada na Figura 15:

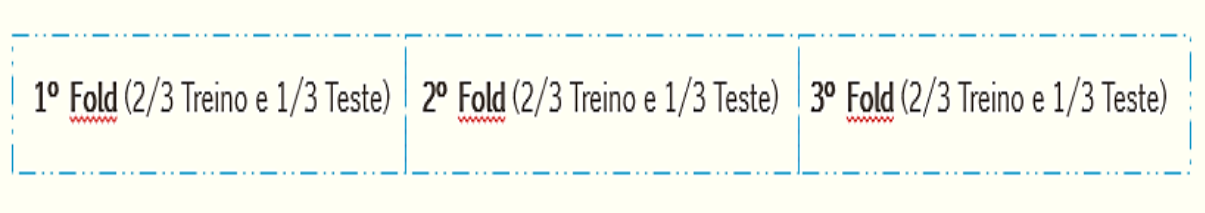


**Figura 15. Divisão entre base em treino e teste – K-fold.**

**Fonte: Própria**

Para análise dos resultados, foram considerados saída “AF” (Acerto Fraude) para os clientes com apontamentos de fraude, saída “AR” (Acerto Regular) para clientes com apontamentos de regulares e “ERRO” para os resultados divergentes do verificado pelas equipes de Campo.

Dado que o modelo de tomada de decisão adotado pela distribuidora Light é cumulativo, capaz de ponderar diversos indícios para um único cliente com base em suas regras e critérios de seleção de alvos, este estudo estabeleceu como parâmetro os clientes que apresentaram um saldo de três ou mais indícios. Esse critério reflete a abordagem da empresa em relação ao envio de inspeções de campo, visando alinhar-se ao processo já estabelecido.

Em relação aos resultados das equipes de campo, dois critérios cruciais foram considerados: apontamentos de clientes com fraude e irregularidades. Essa abordagem visa assegurar uma análise consistente e alinhada com as práticas operacionais da distribuidora.

Para maior clareza dos resultados observados, as tabelas 11 e 12 apresentam um panorama geral do resultado individual das bases verificadas comparado ao observado em campo.

| | | Normal | Fraude | |
|---|---|---|---|---|
| | | N | F | Total |
| Normal | N | 37.908 | 14.700 | 52.608 |
| Fraude | F | 9.102 | 18290 | 27.392 |
| Total | | 47.010 | 32.990 | 80.000 |

| | |
|---|---|
| F1 Score (%) | 60,581% |
| Precisão (%) | 55,441% |
| Recall (%) | 66,771% |

**Tabela 11 Tabela Verdade - Sistema Light.**

| | | Normal | Fraude | |
|---|---|---|---|---|
| | | N | F | Total |
| Normal | N | 46.670 | 5.938 | 52.608 |
| Fraude | F | 5.416 | 21.976 | 27.392 |
| Total | | 52.086 | 27.914 | 80.000 |

| | |
|---|---|
| F1 Score (%) | 79,471% |
| Precisão (%) | 78,728% |
| Recall (%) | 80,228% |

**Tabela 12. Tabela Verdade - Modelo Proposto.**

A Tabela 11 elenca o resultado da ferramenta utilizada atualmente pela distribuidora de energia elétrica Light S.A. ao passo que a Tabela 12 apresenta o resultado do modelo construído, comparado com o retorno de campo, que por sua vez foi usado como referência de estudo.

Ao comparar os resultados de eficiência entre o modelo construído e o atualmente empregado pela distribuidora, destaca-se a notável eficácia do modelo desenvolvido. No modelo construído, a precisão atinge 78,728%, indicando uma significativa melhoria em relação aos 55,441% do modelo da distribuidora. Além disso, o F1 Score, uma métrica que harmoniza precisão e recall, alcança 79,471%, superando os 60,581% do modelo da distribuidora. O recall, que mede a habilidade de encontrar todos os casos positivos, também é superior no modelo construído, atingindo 80,228% em comparação com os 66,771% do modelo da distribuidora.

Esses resultados robustos sinalizam a eficácia do modelo construído na identificação precisa de clientes fraudulentos, indicando um avanço substancial em relação ao modelo atualmente utilizado pela distribuidora.

### 4.4.3 Resultados do Conjunto de Teste Prático

Na seção anterior, foi apresentado o desempenho do modelo construído ao compará-lo ao sistema de seleção utilizado pela distribuidora. Esta análise comparativa revelou o potencial do modelo de superar o desempenho do sistema atual, destacando sua capacidade como ferramenta na identificação de perdas não-técnicas de energia. O modelo agora será aplicado a um novo cenário, onde a aplicação prática e a validação em campo proporcionarão uma compreensão ainda mais profunda de seu desempenho real. A análise cuidadosa dos resultados obtidos nesta fase é fundamental para validar a utilidade e a eficácia da abordagem proposta, pavimentando o caminho para contribuições significativas no campo da eficiência energética.

Desta forma, para efetuar o teste real do modelo construído, foi essencial seguir um procedimento meticuloso na seleção da área de análise, utilizando o balanço energético como critério. Dentro dessa área, foi identificado um grupo específico de clientes vinculados ao transformador em análise, os quais foram cuidadosamente selecionados para fins de teste. Durante o período de avaliação prática, foi possível confrontar os apontamentos do modelo com a realidade operacional em busca de sinais de possíveis fraudes ou irregularidades.

Para melhor compreensão e maior clareza, serão apresentados a seguir os critérios que orientaram a escolha da área de análise, bem como os pormenores da execução da ação destinada a verificar os resultados do modelo em campo.

Para verificação da assertividade do modelo, tornou-se então necessária a criação de uma ação piloto nas abrangências territoriais pertencentes a distribuidora de energia elétrica Light S. A. Para isso, previamente foi realizada uma análise técnica do resultado histórico do Balanço Energético de Transformadores da distribuidora no ano de 2022, com intuito de localizar um transformador com alto percentual de perdas não técnicas, sabendo que o grande desafio para a distribuidora atualmente é a identificação de possíveis clientes fraudadores e não apenas a área a qual estão concentradas essas perdas. Após análise, por conta do alto volume de perda, foram escolhidos para o projeto piloto os clientes vinculados ao transformador de número 930066, localizado no bairro de Jardim América na Cidade do Rio de Janeiro (RJ), pertencente à Regional Leste da Light.

Para a realização das inspeções, contou-se com o apoio de três equipes de normalização: duas equipes de REN – Recuperação de Energia da Light e uma terceira equipe parceira do grupo Light. Além dos citados, a ação contou também com a presença de 4 agentes da polícia civil, dois agentes de negociação, dois supervisores Light e um engenheiro responsável.

As figuras 16 e 17 apresentam algumas imagens registradas no dia da ação:

**Figura 16. Residência direcionada a inspeção.**
**Fonte: própria**

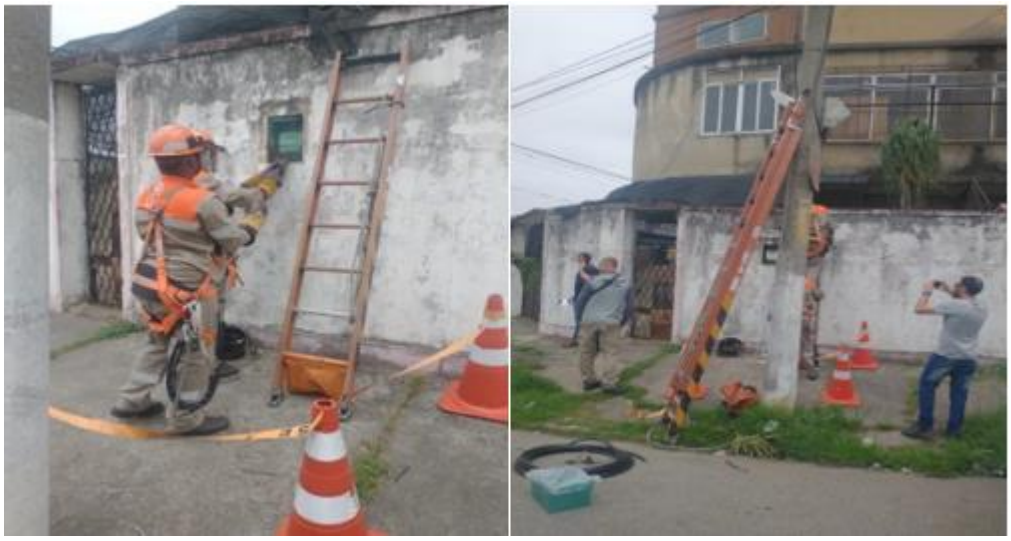

***Figura 17. Normalização de cliente em fraude.***

**Fonte: própria**

Na Figura 17, pode ser observada uma residência qual foi direcionada a inspeção. Considerando os dados desta unidade consumidora (UC), sua situação poderia ser considerada aparentemente normal, visto que possui contrato ativo e com faturamento igual ao mínimo nos três últimos meses. No entanto, durante uma inspeção minuciosa realizada pela equipe de campo, foi identificada uma adulteração no medidor de eletricidade, especificamente com a

desativação da bobina. Tal adulteração foi um flagrante caso de irregularidade que exigiu ação imediata. A gravidade da situação foi evidente o suficiente para justificar o apoio da Polícia Civil, visando evitar transtornos com o cliente envolvido na adulteração e garantir a integridade física dos colaboradores, bem como a legalidade do processo.

É relevante destacar que a instalação em análise não recebeu um alerta de não conformidade por parte do sistema de monitoramento da distribuidora de energia. No entanto, o modelo que foi desenvolvido indicou a necessidade de realizar uma inspeção de campo, considerando-a como um possível caso de fraude.

Como resultado dessa ação, ao verificar a instalação, a equipe de inspeção identificou uma irregularidade no medidor do cliente, gerando assim uma maior confiança nos resultados alcançados. Essa constatação reforça a importância de utilizar abordagens de detecção de fraudes baseadas em modelos de *Machine Learning* para complementar as ferramentas convencionais de monitoramento e garantir a integridade do sistema elétrico.

Destacando o resultado obtido após a aplicação do modelo de *Machine Learning* sobre os clientes identificados como possíveis fraudadores, é relevante ressaltar que a perda histórica na zona fornecida pela distribuidora nos seis meses anteriores à ação era consideravelmente elevada. Após a implementação do modelo, notou-se uma significativa redução nessa perda, evidenciando a eficácia da abordagem adotada.

A Tabela 13 oferece uma visão detalhada do histórico de perdas, juntamente com características específicas dos clientes associados a essa zona. Esses resultados revelam não apenas a capacidade do modelo em identificar possíveis fraudes, mas também seu impacto substancial na mitigação das perdas econômicas, ressaltando a importância prática e a efetividade do modelo de *Machine Learning* implementado.

| | | | | | | | Mês da Atuação |
|---|---|---|---|---|---|---|---|
| | mar/22 | abr/22 | mai/22 | jun/22 | jul/22 | ago/22 | set/22 |
| Perda kWh | 9.479,88 | 10.154,81 | 8.859,09 | 7.071,49 | 8.640,25 | 7.609,43 | 6.148,94 |
| Perda % | 60,87% | 61,07% | 66,49% | 63,81% | 68,84% | 63,66% | 65,70% |

| Equipamento |
|---|
| ZNA930066 |
| 44 Clientes |

| ATIVO | Mínimo | regular | estimado |
|---|---|---|---|
| 36 | 11 | 18 | 7 |

| INATIVO | Suspenso |
|---|---|
| 8 | 8 |

**Tabela 13. Resultado de Perda % anterior a ação de campo.**

Os dados de perdas mostrados na tabela 6 são acompanhados internamente pela coordenação de *SmartGrid* da distribuidora, que além do citado é responsável também pelo suporte técnico das áreas comerciais e operacionais, planejamento operacional, detecção de irregularidades nas áreas telemedidas e fechamento de carga junto à CCEE.

Dentro do escopo da presente dissertação, que se concentra em um modelo de *Machine Learning* dedicado à detecção de potenciais clientes fraudulentos, a avaliação dos resultados, conforme demonstrado na Tabela 13, revela uma média expressiva de perdas não técnicas calculadas pela distribuidora no período anterior à data da intervenção. Constatou-se que essa média ultrapassa os 60%, indicando uma magnitude preocupante de perdas na região analisada. Ademais, destaca-se o número considerável de clientes associados à zona com contratos suspensos, uma condição que surge como uma possível explicação para o elevado índice de perdas evidenciado nos cálculos. Essa associação é atribuída à natureza essencial da energia elétrica, tornando-se praticamente impossível para um domicílio funcionar sem o fornecimento regular. Dessa forma, pode-se inferir que uma residência ocupada com contrato suspenso está, quase certamente, envolvida em alguma forma de irregularidade.

Essas conclusões fundamentam-se em uma análise profunda do contexto da metodologia empregada, centrada no modelo de *Machine Learning*, e corroboram a importância prática dos resultados apresentados neste capítulo, oferecendo insights valiosos para a gestão eficaz e a mitigação de perdas no setor de distribuição de energia elétrica.

Com base nas informações levantadas e análise técnica realizada, foi possível então realizar a aplicação do modelo desenvolvido para todos os clientes envolvidos na ação com intuito de averiguar quais dentre eles havia apontamentos de possíveis fraudadores. Em paralelo a esta situação, foi agendado com o time de campo a etapa de validação da vinculação.

Esta etapa é fundamental para certificar que a base de clientes considerada pelo sistema é a mesma que a encontrada em campo, inclusive os contratos ditos como suspensos. Importante ressaltar também que quanto maior for a diferença entre a vinculação verificada no sistema com a realidade de campo, maior será o erro visto pelo cálculo de balanço. Nesta etapa de validação da vinculação, todos os clientes alimentados pelo secundário do transformador devem ser localizados e confirmados os dados de instalação, número do medidor e endereço.

Após a verificação da associação e considerando a dispensa de reajuste, a operação foi programada para iniciar em setembro, com conclusão prevista para outubro de 2022. Inicialmente, foram geradas 27 notas de serviço destinadas aos clientes identificados pelo modelo, suspeitos de estarem envolvidos em fraude ou em uma situação irregular que poderia afetar a falta de registro de seu consumo real.

A tabela 14 apresenta a situação contratual dos 27 clientes selecionados pelo modelo para inspeções.

| **Situação (TIPO/FATURA)** | **Total de Clientes** | **Status da Leitura** |
|---|---|---|
| Mínimo (MIN) | 8 | Cliente com registro de consumo medido mensal inferior ou igual ao mínimo. |
| Estimados (EST) | 7 | Situações adversas que impossibilitam a coleta de leitura manual. Exemplo (medidor interno). |
| Inativo (INT) | 6 | Cliente sem Parceiro de Negócio (PN) e sem contrato ativo. |
| Suspenso (SUSP) | 6 | Representa suspensão de contrato junto a distribuidora, com a interrupção do fornecimento de energia (Cortado). |

**Tabela 14. Situação Contratual dos clientes Inspecionados**

A fim de fornecer uma visão mais abrangente da capacidade do modelo em identificar possíveis clientes fraudadores, na tabela 15, serão apresentados em detalhes os resultados obtidos pelo modelo desenvolvido, em comparação com os resultados do sistema de seleção da distribuidora e os resultados observados no campo, indicando a quantidade de clientes identificados em situações de fraude ou irregularidade.

É fundamental destacar que, neste exemplo prático, o modelo demonstrou um índice de precisão de 90% de acertos. Esses dados são apresentados na tabela 8 e podem ser considerados de suma importância para a avaliação da eficácia do modelo em comparação com abordagens tradicionais, contribuindo significativamente para a dissertação de mestrado ao realçar a sua eficiência na detecção de perdas não técnicas no setor de distribuição de energia, além de fornecer melhoria para o processo já existente da distribuidora.

| | **QTD. DE CLIENTES** | | |
|---|---|---|---|
| Situação CONTRATO | Apontado pelo modelo desenvolvido como possível fraudador | Apontado pelo sistema de seleção da distribuidora como possível fraudador | Identificado pela equipe de campo em fraude/ irregularidade |
| Mínimo (MIN) | 8 | 5 | 8 |
| Estimados (EST) | 7 | 7 | 7 |
| Inativo (INT) | 6 | 8 | 5 |
| Suspenso (SUSP) | 6 | 6 | 9 |
| Total APONTAMENTOS | 27 | 26 | **29** |

| | | | |
|---|---|---|---|
| Total ACERTO | 26 | 23 | |
| % ACERTO | 90% | 79% | |

**Tabela 15. Resultado dos apontamentos gerados e sua assertividade**

Os resultados apresentados na tabela 15, que se refere à análise comparativa do desempenho da aplicação real, revelam uma proximidade considerável no total de apontamentos entre o modelo construído e o sistema de seleção atual da distribuidora. Entretanto, destaca-se que o total de acertos é inferior no sistema vigente, classificando-o como menos eficiente em comparação ao modelo desenvolvido nesta dissertação. Essa constatação ressalta a importância de aprimorar o sistema de seleção atual, buscando otimizações que conduzam a um aumento na taxa de acertos, refletindo em uma seleção mais assertiva de clientes para inspeção de campo. Pode-se então observar que 90% dos clientes identificados pelo modelo proposto como potenciais fraudadores estavam, de fato, envolvidos em atividades fraudulentas.

Com o objetivo então, de validar a assertividade do modelo em relação aos clientes identificados como fraudadores e submetidos a normalização em campo, realizou-se uma análise do balanço energético na área de atuação no mês subsequente à intervenção. Os resultados são apresentados na tabela 16 e revelam uma redução significativa das perdas, passando de 65% para 14%. Esse notável resultado destaca a eficácia do modelo construído, evidenciando sua capacidade não apenas de identificar clientes fraudulentos, mas também de contribuir substancialmente para a mitigação das perdas não técnicas na distribuição de energia elétrica. Essa redução expressiva das perdas representa um impacto positivo tanto para a distribuidora quanto para o sistema como um todo, reforçando a efetividade e relevância do modelo implementado.

| | set/22 | out/22 | nov/22 | dez/22 |
|---|---|---|---|---|
| Perda kWh | 6.148,94 | 1.948,95 | 1.871,49 | 1.962,31 |
| Perda % | 65,70% | 14,81% | 13,42% | 14,91% |

| Equipamento |
|---|
| ZNA930066 |
| 44 Clientes |

| ATIVO | Mínimo | Regular | estimado |
|---|---|---|---|
| 42 | 3 | 39 | 0 |

| INATIVO | Suspenso |
|---|---|
| 2 | 2 |

**Tabela 16. Resultado de Perda % após ação de campo**

Esses resultados não apenas reforçam a importância da aplicação de técnicas avançadas de *Machine Learning* no contexto das perdas de energia não técnicas, mas também indicam seu

potencial para melhorar a integridade do sistema elétrico como um todo. O desempenho do modelo também pode ser analisado por meio do gráfico fornecido na Figura 25.

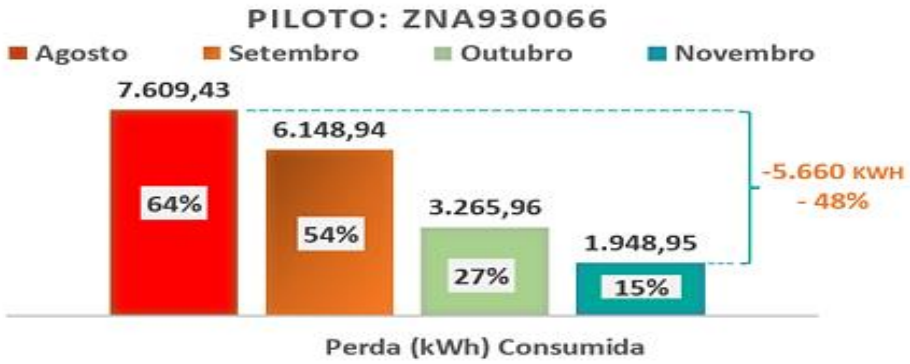


**Figura 18. Redução da Perda em kWh pós atuação.**
**Fonte: Própria**

É relevante destacar que a área de cobrança e faturamento da distribuidora de energia registrou um aumento expressivo no faturamento no mês subsequente à regularização dos clientes, conforme evidenciado na Figura 25. Esse incremento totalizou R$ 4.276,40, refletindo diretamente na recuperação de 5.660 kWh de energia que anteriormente eram contabilizados como perdas não técnicas. Os cálculos foram baseados na tarifa média do Rio de Janeiro estipulada pela ANEEL em [75], a qual foi de R$ 0,754 por kWh.

Os resultados obtidos evidenciam que a capacidade do modelo de direcionar de forma precisa os clientes suspeitos de fraudes, aliada à expertise das equipes de inspeção, cria uma sinergia importante que não apenas aumenta a eficiência das operações, mas também desempenha um papel fundamental na redução das perdas. Essa abordagem integrada representa uma evolução no setor, proporcionando à distribuidora uma ferramenta eficaz para proteger seus ativos e garantir a integridade de seu sistema elétrico, com benefícios tangíveis em termos de eficiência e rentabilidade.

# Capítulo 5 – Conclusão e Trabalhos Futuros

Nesta dissertação delineou-se o desenvolvimento de uma metodologia embasada em *Machine Learning*, com o propósito de aprimorar a precisão na identificação de clientes de baixa tensão em situações irregulares da distribuidora de energia elétrica Light. Mediante a isso, o estudo fundamentou-se em dados reais provenientes de inspeções realizadas em campo, abrangendo uma base de 80 mil instalações. Para isso, foram utilizados atributos de consumo, localização geográfica das unidades consumidoras, apontamentos das equipes durante as inspeções e fechamento de notas de serviço.

No desenvolvimento do modelo, foi adotado também o método de janelamento com tamanho três e passo um para o consumo mensal. Essa abordagem de agrupamento trimestral proporciona uma análise eficiente dos padrões de consumo ao longo do tempo, contribuindo para a construção de um modelo robusto. Além disso, foi considerado um tratamento nas informações de visitas e inspeções realizadas em campo, cujo objetivo era não fornecer ao modelo informações antecipadas e que mascarariam o resultado classificatório final.

O melhor modelo para o sistema de classificação foi aquele utilizando o algoritmo *Random Forest*, aplicado a uma base de dados que representa registros referentes ao ano de 2020, com uma única exceção para os apontamentos de reincidência, que tiveram como referência o ano de 2019. Além disso, foi observado que o tempo de execução do código foi satisfatório, demonstrando assim, uma eficiência no processamento e no tempo de resposta durante sua execução.

Apesar da redução de atributos nos dados originais da empresa, os valores obtidos com o VPP apresentados nesse trabalho foram superiores aos valores verificados considerando o sistema de indicação de inspeções correntemente utilizado pela concessionária. Dessa forma, também pode-se concluir que os resultados apresentados demonstram que o modelo proposto é bastante promissor no problema de identificação de irregularidades em baixa tensão, o que estimula a aplicação dessa nova metodologia para todos os clientes da Light.

Como trabalho futuro, pretende-se incrementar novas técnicas para constante aprendizado do modelo, pois foi observado que o desempenho do modelo varia para diferentes períodos analisados. Ampliar a melhoria dos dados utilizados como variável de entrada no modelo construído, com o propósito de poder selecionar os clientes fraudadores que poderão trazer um retorno de energia melhor de acordo com o seu potencial de consumo também pode ser um projeto futuro.

# Bibliografia

[1] Acende brasil (2017). Perdas Comerciais E Inadimplência No Setor Elétrico. White Paper, Edição nº 18, fev/ 2017. Acesso em 21 abr. 2022.

[2] Angelos, E. W. S., Saavedra, O. R., Cortés, O. A. C., Souza, A. N. (2011). Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. IEEE Transactions on Power Delivery, 26(4): 2436-2442.

[3] Araujo, B., Almeida, H., and Mello, Filho (2019). Computational Intelligence Methods Applied to the Fraud Detection of Electric Energy Consumers IEEE Latin America Transactions, vol. 17, no. 1, january 2019.

[4] Ahamad, T., Hongyu, Z., (2022). Energetics Systems and artificial intelligence: Applications of industry 4.0. Sciencedirect. Volume 8, November 2022, Pages 334-361.

[5] Buzau, M., Tejedor-Aguilera, J., Cruz-Romero, C., and Gómez-Expósito, A., Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning, IEEE Transactions on Smart Grid (Early Accept), 2018. doi: 10.1109/TSG.2018.2807925.

[6] Cristina, I. (2020). Repórter da Agência Brasil - Rio de Janeiro. Agência Brasil. Taxa de desemprego passa de 13,3% para 14,6% no terceiro trimestre.

[7] Cabral, J. E., Gontijo, E. M., Pinto, J. O. P., and Filho, J. R. (2004). Fraud detection in electrical energy consumers using rough sets. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, 4: 3625–3629.

[8] Cyro, M., Karla, F., Marley, V., Marco, P. e Gustavo, C. (2008). Indicações de Suspeitos de Irregularidade em Instalações Elétricas de Baixa Tensão. Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais (SBRN), 6(1): 16-28.

[9] Dutra, B (2016). Furto de energia eleva conta de luz de quem paga em 17%. Revista Extra: Globo Comunicações. Acesso em 21 abr. 2022.

[10] Faria, R. A. D.; Fonseca, K. V. O.; Schneider, B.; Nguang, S. K. Collusion and Fraud Detection on Electronic Energy Meters - A Use Case of Forensics Investigation Procedures, 2014 IEEE Security and Privacy Workshops, San Jose, CA, pp.65-68, 2014. doi: 10.1109/SPW.2014.19.

[11] Filho, J. R.; Gontijo, E. M.; Delaiba, A. C.; Mazina, E.; Cabral, J. E.; Pinto, J. O. P.Fraud identification in electricity company customers using decision tree, 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), pp.3730-3734, v.4, 2004. doi: 10.1109/ICSMC.2004.1400924.

[12] Gunturi, S. K., Sarcar, D., (2021). Ensemble machine learning models for the detection of energy theft.Electric Power Systems Research. Volume 192, March 2021, 106904. Sciencedirect.

[13] Jang, H.S., Bae, K.Y., Park, H.S., Sung, D.K., 2016. Solar power prediction based on satellite images and support vector machine. IEEE Trans. Sustain. Energy 7, 1255–1263. http://dx.doi.org/10.1109/TSTE.2016.2535466.

[14] Jokar, P.; Arianpoo, N.; Leung, V.C.M. A survey on security issues in smart grids, Secur. Commun. Netw., v.9, pp.262-273, 2012. doi: 10.1002/sec.559.

[15] Light, Histórico e Perfil corporativo, Home. A companhia. Perfil corporativo. 2021. Disponível em http://ri.light.com.br/a-companhia/historico-e-perfil-corporativo/.

[16] Muniz, C., M. Vellasco, M., Tanscheit, R., Figueiredo, K. A Neuro-fuzzy System for Fraud Detection in Electricity Distribution. IFSAEUSFLAT 2009, 6(3): 1096-1101.

[17] Megahed, T.F., Abdelkader, S.M., Zakaria, A., 2019. Energy management in zero-energy building using neural network predictive control. IEEE Internet of Things J. 6, 5336–5344. http://dx.doi.org/10.1109/JIOT.2019.2900558.

[18] Nizar, A. H., Dong, Z. H., Zhao, J. H., e Zhang, P. (2007). A Data Mining Based NTL Analysis Method. IEEE Power Engineering Society (PES) General Meeting, 1(4): 1-8.

[19] Pollock, D. S. G. (1999). A Handbook of TimeSeries Analysis, Signal Processing and Dynamics. Academic Press, New York, San Diego Edition.

[20] Pinto, T., Faia, R., Navarro-Caceres, M., Santos, G., Corchado, J.M., Vale, Z., 2019. Multi-agent-based CBR recommender system for intelligent energy management in buildings. IEEE Syst. J. 13, 1084–1095. http://dx.doi.org/10.1109/JSYST.2018.2876933.

[21] Rauber, T.; Drago, I., Varejão, F. e Queiroga, R. (2005) Extração e Seleção de Características na Identificação de Perdas Comerciais na Distribuição de Energia Elétrica. XXV Cong. Soc. Bras. Comp.

[22] Rong, J., Tagaris, H., Lachsz, A., and Jeffrey, M. (2002). Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In 2002 Trans. and Distribution Conf. and Exhibition 2002: Asia Pacific. IEEE/PES, 3: 2251–2256.

[23] TAVARES, P. de Campos; Algoritmo, in "Enciclopédia Verbo Luso-Brasileira da Cultura, Edição Século XXI", Volume II, Editorial Verbo, Braga, Janeiro de 1998 ISBN 972-22-1864-6.

[24] Tree Boosting With XGBoost – Why Does XGBoost Win "Every" Machine Learning Competition?. Synced (em inglês). 22 de outubro de 2017. Consultado em 4 de janeiro de 2020.

[25] Sagi, Omer; Rokach, Lior (2021). "Approximating XGBoost with an interpretable decision tree". Information Sciences. 572 (2021): 522-542. doi:10.1016/j.ins.2021.05.055.

[26] Igor Kononenko, Matjaž Kukar, in Machine Learning and Data Mining, 2007. https://www.sciencedirect.com/topics/computer-science/reinforcement-learning.

[27] L. Kaufman e P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, 1o ed. Wiley-Blackwell, 2005.

[28] L. Damaceno. Junho 10, 2020. Regressão Linear ? https://medium.com/@lauradamaceno/regress%C3%A3o-linear-6a7f247c3e29

[29] J. Castro, N.; Miranda, M. e Vardiero, P. (2019) Perdas não técnicas na distribuição de energia elétrica: o caso da Light. Rio de Janeiro: 348 p. :il, ; 25cm.

[30] Relatório da Agência Nacional de Energia Elétrica, ANEEL. Perdas de Energia Elétrica na Distribuição, Edição|01/2021.

[31] SILVA, L. G. W. Desenvolvimento de uma metodologia integrada para alocação otimizada de dispositivos de controle e proteção, previsão de carga em sistemas de energia elétrica em sistemas de distribuição de energia elétrica. 2002. 84f. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia, Universidade Estadual Paulista, Ilha Solteira, 2002.

[32] FREITAS, B. M., & HOLLANDA, L. Micro e Minigeração no Brasil: Viabilidade Econômica e Entraves do Setor. FGV Energia, mai. 2015.

[33] LOPO, A. B. Análise do desempenho térmico de um sistema de aquecimento solar de baixo custo. 2010. 82 f. Dissertação (Mestrado) - Curso de Engenharia Mecânica, Universidade Federal do Rio Grande do Norte, Natal, 2010.

[34] CHAPMAN, S J. **Fundamentos de máquinas elétricas**. Porto Alegre: AMGH, 2013.

[35] TIMMONS, D. et al. Decarbonizing residential building energy: a cost-effective approach. Elsevier, Amsterdam, v. 92, p. 382-392, 2016.

[36] ABRADEE. Associação Brasileira de Distribuidores de Energia Elétrica. Como é calculado o índice de perdas da distribuidora?. Disponível em: https://www.abradee.org.br/portal/index.php/faq/296-como-e-calculado-o-indice-de-perdas-da-distribuidora. Acesso em: 8 mar. 2023.

[37] ANEEL. (2015). Manual de contabilização e classificação das perdas técnicas na distribuição de energia elétrica. Brasília: Agência Nacional de Energia Elétrica.

[38] FRAGOAS, A.G. Estudo de caso do uso de bancos de capacitores em uma rede de distribuição primaria - indicativos da sua viabilidade econômica. 2008. 72 f. TCC (Graduação) - Curso de Engenharia Elétrica, Universidade de São Paulo, São Carlos, 2008.

[39] ESPOSITO, A. S & FUCHS, P. G. Desenvolvimento tecnológico e inserção da energia solar no Brasil. Revista do BNDES, Rio de Janeiro, n. 40, p. 85-113, dez. 2013.

[40] ANICETO, D.M. Importância da correção do fator de potência nas instalações elétricas industriais. IPOG: Revista Especialize On-line, Goiânia, v. 1, n. 11, p.1-16, jul. 2016.

[41] LAVIERI, Arthur. “Isoladores Elétricos - componentes básicos para um sistema elétrico”. Canal Energia. São Paulo, 23 de fevereiro de 2010.

[42] NAKAGOMI, R. M. Proposição de um sistema para simulação de faltas de alta impedância em redes de distribuição. Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia e Automação Elétricas. São Paulo, 2006. 90p.

[43] PAIVA, I., CASTRO, N., & LIMA, A. P. Aspectos Teóricos e Analíticos da Segurança Energética e os Desafios do Setor Elétrico Brasileiro. Texto De Discussão do Setor Elétrico n.º 71. GESEL UFRJ. Rio de Janeiro, RJ, mai. 2017.

[44] NASCIMENTO, P. A. M. M. Considerações sobre as indústrias de equipamentos para produção de energias eólica e solar fotovoltaica e suas dimensões científicas no Brasil. Radar, junho de 2015.

[45] SOUZA, L. E., & CAVALCANTE, A. M. Concentrated Solar Power deployment in emerging economies: The cases of China and Brazil. 2016.

[46] CIRED. **Reduction of Technical and Non=Technical Losses in Distribution Networks**. CIRED Working Group on Losses Reduction. [S.I], p. 114. 2017.

[47] NORTHEAST GROUP. Electricity Theft and Non-Technical Losses: Global Markets, Solutions and Vendors. Northcast Group, lle. Washintogn, D.C., p. 49. 2017.

[48] ANEEL. (2019). Perdas de Energia Elétrica na Distribuição. AgÊsncia Nacional de Energia Elétrica. Brasília: p 21. 2019.

[49] CHAPMAN, S J. **Fundamentos de máquinas elétricas**. Porto Alegre: AMGH, 2013.

[50] MENDONÇA, I.M. et al. Geração de energia eólica: estudos de viabilidade via análise estatística da velocidade dos ventos. Multiverso, Juiz de Fora, v. 2, n. 1, p.80- 87, 2017.

[51] URBANETZ JUNIOR, J. Sistemas fotovoltaicos conectados a redes de distribuição urbanas: sua influência na qualidade da energia elétrica e análise dos parâmetros que possam afetar a conectividade. 2010. 192 f. Tese (Doutorado) - Curso de Engenharia Civil, Universidade Federal de Santa Catarina, Florianópolis, 2010.

[52] BONFIM, H.V.; Santos, S.F.; Silva, L.I.S.; Primo, A.; Mendonça, M.A. Utilização de conceitos de cálculo para verificação da eficiência de uma placa solar. Ciências exatas e tecnológicas. v.4, nº. 1, p.29-34, Aracaju, março de 2017. ISSNS impresso 1980-1777. ISSN eletrônico 2316-3135.

[53] CAMINHA, Amadeu C., Introdução à Proteção dos Sistemas Elétricos. Edgard Blücher Ltda, 2010.

[54] PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011.

[55] DOMINGOS, P. A few useful things to know about machine learning. Communications of the ACM, ACM, v. 55, n. 10, p. 78–87, 2012.

[56] SMOLA, A.; VISHWANATHAN, S. Introduction to machine learning. Cambridge University, UK, v. 32, p. 34, 2008.

[57] ARANHA, Christian Nunes. Uma Abordagem de Pré-processamento Automático para Mineração: Sob o Enfoque da Inteligência Computacional. Rio de Janeiro, 2007 Tese (Engenharia Elétrica) - PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO, 2007.

[58] EL NAQA, Issam; MURPHY, Martin J. What is machine learning?. In: machine learning in radiation oncology. Springer, Cham, 2015. p. 3-11.

[59] MAYER-SCHÖNBERGER, V.; CUKIER, K. Big Data. Edição traduzida. Rio de Janeiro: Elsevier, 2013.

[60] 2022 *Data Science Academy*, Capítulo 6 – O perceptron – parte 1. Deep Learning Book 2007. https://www.deeplearningbook.com.br/o-perceptron-parte-2/.

[61] P Lusztin. Decision Trees: from 0 to XGBoost & LighGBM, Capítulo 6 – Machine Learning. May 22, 2022. https://medium.com/mlearning-ai/decision-trees-from-0-to-xgboost-lightgbm-a5f6827dfa23.

[62] KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

[63] Bernhard E. Bose. Isabelle M. Guvon, Vladimir N. Vapnik KOHAVI, R. A training algorithm for optimal margin classifiers. . July 1992Pages 144–152.

[64] Christopher Bishop. Pattern Recognition and Machine Learning. 17 deagosto de 2006.

[65] S. Haykin Neural networks and learning machines. 3rd ed. Rev. ed of: Neural networks. 2nd ed., 1999.

[66] G. V. Lima dos Anjos Sistema Inteligente para Identificação de Suspeitos de Fraude no Consumo de Água. Pontífica Universidade Católica do Rio de Janeiro – PUC. dissertação de mestrado. abril de 2022.

[67] BS de Araujo, HLS de Almeid. Métodps de Inteligência Computacional Aplicados à Detecção de Fraude de Consumidores de Energia Elétrica. IEEE Latin America …, 2019.

[68] Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, [2020]. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/lei/l14020.htm.18 de set. de 2021

[69] N. T. AlvesI; B. F. CalvoII; J. A. A. Casanova, A. A. C. Neto IV; N. A. Santos. The use of programming languages and computer software in psychological science. Temas psicol. vol.22 no.3 Ribeirão Preto dez. 2014.

[70] P. Burman, "A Comparative Study of Ordinary Cross-Validation, v-Fold CrossValidation and the Repeated Learning-Testing Methods", Biometrika, vol. 76, no 3, p. 503, set. 1989, doi: 10.2307/2336116.

[71] J. Taylor and S. Taylor, Introduction To Error Analysis: The Study of Uncertainties in Physical Measurements, ser. ASMSU/Spartans.4.Spartans Textbook. University Science Books, 1997. [Online]. Available: https://books.google.com.br/books?id=giFQcZub80oC.

[72] V. L. Santos. Uso de Machine Learning para identificação de solicitação de teste de confirmação em projeto de teste de software. Universidade Federal Rural de Pernambuco – UFRPE. Junho de 2022.

[73] D. C. R. Souza. Aplicação de Redes Neurais Artificiais para Estimação de Indicadores de Segurança Estática e Dinâmica de Sistemas Elétricos de Potência. Fevereiro de 2022.

[74] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020.

[75] ANEEL. (2023). Revisão Tarifária Periódica – RTP. Resolução Homologatória N.3.177, 14 de março de 2023: Agência Nacional de Energia Elétrica.

[76] Scikit Learning 2023. Supervised Learning. User guide. 2007 - 2023, scikit-learn developers (BSD License). https://scikit-learn.org/stable/user_guide.html.

[77] Pandas Introduction 2023. https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/.

[78] R. Johansson. Numpy. Numerical Python: Scientific Computing and Data Science Applications with Numpy, SciPy and Matplotlib 2nd ed. Edition.