Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Rafael Mendonça Rocha Barros

Tese de Doutorado

Advanced Analytics Aplicado à Gestão da Perda Não Técnica de Energia em Sistemas Elétricos de Distribuição

Campina Grande
2021

Rafael Mendonça Rocha Barros

Advanced Analytics Aplicado à Gestão da Perda Não Técnica de Energia em Sistemas Elétricos de Distribuição

Tese de doutorado apresentada à Coordenação do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande, como parte dos requisitos necessários para a obtenção do título de Doutor em Engenharia Elétrica.

Área de Concentração: Processamento de Energia

Orientadores:

Edson Guedes da Costa, D. Sc.

Jalberth Fernandes de Araujo, D. Sc.

Campina Grande
2021

Rafael Mendonça Rocha Barros

Advanced Analytics Aplicado à Gestão da Perda Não Técnica de Energia em Sistemas Elétricos de Distribuição

Tese aprovada em: 19/07/2021

Edson Guedes da Costa, D. Sc.
Universidade Federal de Campina Grande
Orientador

Jalberth Fernandes de Araujo, D. Sc.
Universidade Federal de Campina Grande
Orientador

Antônio Padilha Feltrin, D. Sc.
Universidade Federal do ABC
Avaliador Externo

José Roberto Sanches Mantovani, D. Sc.
Universidade Estadual Paulista
Avaliador Externo

Benemar Alencar de Sousa, D. Sc.
Universidade Federal de Campina Grande
Avaliador Interno

George Rossany Soares de Lira, D. Sc.
Universidade Federal de Campina Grande
Avaliador Interno

Campina Grande-PB

B277a Barros, Rafael Mendonça Rocha.
Advanced analytics aplicado à gestão da perda não técnica de energia em sistemas elétricos de distribuição / Rafael Mendonça Rocha Barros. – Campina Grande, 2021.
183 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2021.
"Orientação: Prof. Dr. Edson Guedes da Costa, Prof. Dr. Jalberth Fernandes de Araujo".
Referências.

1. Sistemas Elétricos de Distribuição. 2. Advanced Analytics. 3. Inferência Causal. 4. Machine Learning. 5. Maximização - Perda não Técnica - Retorno Financeiro. 6. Processamento de Energia. I. Costa, Edson Guedes da. II. Araujo Jalberth Fernandes de. III. Título.

CDU 621.3.095.2(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
POS-GRADUACAO EM ENGENHARIA ELETRICA
Rua Aprigio Veloso, 882, - Bairro Universitario, Campina Grande/PB, CEP 58429-900

REGISTRO DE PRESENÇA E ASSINATURAS

1. ATA DA DEFESA PARA CONCESSÃO DO GRAU DE DOUTOR EM CIÊNCIAS, NO DOMÍNIO DA ENGENHARIA ELÉTRICA, REALIZADA EM 19 DE JULHO DE 2021
(Nº 333)

CANDIDATO: **RAFAEL MENDONÇA ROCHA BARROS**. COMISSÃO EXAMINADORA: BENEMAR ALENCAR DE SOUZA, D.Sc., UFCG, Presidente da Comissão, EDSON GUEDES DA COSTA, D.Sc., UFCG, Orientadores, GEORGE ROSSANY SOARES DE LIRA, D.Sc., UFCG, ANTONIO PADILHA FELTRIN, Dr., UFABC, JOSÉ ROBERTO SANCHES MANTOVANI, Dr, FEIS-UNESP. TÍTULO DA TESE: Advanced Analytics Aplicado à Gestão da Perda não Técnica de Energia em Sistemas Elétricos de Distribuição. ÁREA DE CONCENTRAÇÃO: Processamento da Energia. HORA DE INÍCIO: **08h00** – LOCAL: **Sala Virtual, em virtude da suspensão de atividades na UFCG decorrente do coronavírus e de conformidade com o Art. 8º da PORTARIA PRPG/GPR Nº 003, DE 18 DE MARÇO DE 2020)**. Em sessão pública, após exposição de cerca de 45 minutos, o candidato foi arguido oralmente pelos membros da Comissão Examinadora, tendo demonstrado suficiência de conhecimento e capacidade de sistematização, no tema de sua tese, obtendo conceito APROVADO. Face à aprovação, declara o presidente da Comissão, achar-se o examinado, legalmente habilitado a receber o Grau de Doutor em Ciências, no domínio da Engenharia Elétrica, cabendo a Universidade Federal de Campina Grande, como de direito, providenciar a expedição do Diploma, a que o mesmo faz jus. Na forma regulamentar, foi lavrada a presente ata, que é assinada eletronicamente por mim, ÂNGELA DE LOURDES RIBEIRO MATIAS, e os membros da Comissão Examinadora presentes. Campina Grande, 19 de Julho de 2021.

ÂNGELA DE LOURDES RIBEIRO MATIAS
Secretária

BENEMAR ALENCAR DE SOUZA, D.Sc., UFCG
Presidente da Comissão e Examinador Interno

EDSON GUEDES DA COSTA, D.Sc., UFCG
Orientador

GEORGE ROSSANY SOARES DE LIRA, D.Sc., UFCG
Examinador Interno

ANTONIO PADILHA FELTRIN, Dr., UFABC
Examinador Externo

JOSÉ ROBERTO SANCHES MANTOVANI, Dr, FEIS-UNESP
Examinador Externo

RAFAEL MENDONÇA ROCHA BARROS
Candidato

2 - APROVAÇÃO

2.1. Segue a presente Ata de Defesa de Tese de Doutorado do candidato **RAFAEL MENDONÇA ROCHA BARROS**, assinada eletronicamente pela Comissão Examinadora acima identificada.

2.2. No caso de examinadores externos que não possuam credenciamento de usuário externo ativo no SEI, para igual assinatura eletrônica, os examinadores internos signatários **certificam** que os examinadores externos acima identificados participaram da defesa da tese e tomaram conhecimento do teor deste documento.

Documento assinado eletronicamente por **ANGELA DE LOURDES RIBEIRO MATIAS**, **SECRETÁRIO (A)**, em 10/08/2021, às 15:13, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **ANTONIO PADILHA FELTRIN**, **Usuário Externo**, em 10/08/2021, às 15:38, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **RAFAEL MENDONÇA ROCHA BARROS**, **Usuário Externo**, em 10/08/2021, às 15:47, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **EDSON GUEDES DA COSTA**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 10/08/2021, às 15:53, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **GEORGE ROSSANY SOARES DE LIRA**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 10/08/2021, às 16:10, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

Documento assinado eletronicamente por **BENEMAR ALENCAR DE SOUZA**, **PROFESSOR(A) DO MAGISTERIO SUPERIOR**, em 10/08/2021, às 21:01, conforme horário oficial de Brasília, com fundamento no art. 8º, caput, da Portaria SEI nº 002, de 25 de outubro de 2018.

A autenticidade deste documento pode ser conferida no site https://sei.ufcg.edu.br/autenticidade, informando o código verificador **1694677** e o código CRC **2FA79C71**.

Dedico este trabalho a todos os professores, por buscarem, além de compreender a realidade, proporcionar que outros também a compreendam. Pois assim é possível existir emancipação.

Agradecimentos

Agradeço aos meus pais, Eunice e Sérgio, por terem escolhido a educação como prioridade absoluta na minha criação.

Aos meus irmãos Karoline e Felipe e a todos os meus familiares, pela torcida e pelo suporte dedicados a mim durante minha jornada de crescimento.

A minha noiva Marcela, por compreender meus momentos de ausência e me apoiar durante toda minha carreira acadêmica.

Ao professor Edson Guedes, pela amizade e pelos ensinamentos transmitidos ao longo de dez anos de convívio, os quais contribuíram de forma decisiva para minha formação.

Ao professor Jalberth, pela amizade e pela ajuda durante todas as etapas deste trabalho.

Aos amigos do Laboratório de Alta Tensão da Universidade Federal de Campina Grande, pelas diversas vezes em que fui ajudado, pelo companheirismo e pelos momentos de confraternização que foram compartilhados comigo.

Aos amigos do Grupo Energisa, pelos ensinamentos fornecidos e pela confiança depositada em mim durante os anos em que trabalhei como engenheiro de distribuição, os quais complementaram de forma valiosa minha formação.

Por fim, agradeço ao contribuinte brasileiro, por financiar as políticas públicas educacionais, sem as quais eu não poderia ter chegado até aqui.

“Quando o mundo estiver unido na busca do conhecimento, e não mais lutando por dinheiro e poder, então nossa sociedade poderá enfim evoluir a um novo nível.”

Thomas Jefferson

Resumo

Uma nova metodologia para o aprimoramento da gestão da perda não técnica de energia nos sistemas elétricos de distribuição é apresentada neste trabalho. A solução consiste na utilização de técnicas baseadas em *Advanced Analytics* para construção de um processo automatizado e adaptativo capaz de identificar os fatores de risco e de proteção para ocorrência da perda não técnica; a probabilidade de existência de perda; uma estimativa da energia não medida no sistema e o potencial de retorno financeiro para inspeções *in-loco* de consumidores. No desenvolvimento da pesquisa, foi utilizada uma base de dados com informações reais de 261.489 consumidores de uma distribuidora brasileira. A Inferência Causal foi aplicada para identificar o grau de associação de diversas variáveis com a ocorrência de perda não técnica e, assim, identificar os fatores de risco e de proteção. As variáveis associadas à ocorrência de perda foram utilizadas como entrada em modelos de *Machine Learning* com o objetivo de identificar a probabilidade de ocorrência de perda, bem como fornecer uma estimativa do valor da energia não medida no sistema. No total, foram avaliados 23 modelos de classificação e sete de regressão com diferentes algoritmos e abordagens no tratamento dos dados. Os resultados dos modelos preditivos foram utilizados para calcular o potencial de retorno financeiro das inspeções de campo. Posteriormente, foi proposto um modelo para o dimensionamento ótimo da infraestrutura de combate às perdas em uma distribuidora, o qual resulta na maximização do retorno financeiro das inspeções de campo. Todas as etapas da metodologia foram validadas por meio de novas inspeções de campo realizadas em 1.417 consumidores. Os principais resultados alcançados no trabalho foram a identificação de 76 fatores de risco ou proteção para ocorrência da perda não técnica; a identificação do algoritmo *Rotation Forest* como o mais adequado para a identificação da perda, o qual apresentou uma precisão de 66,5% nas inspeções realizadas em campo; a identificação do algoritmo XGBT como o mais adequado para a predição dos valores de energia não medida; o qual apresentou um desvio de 3,02% na estimativa do valor total de energia recuperada em um grupo de consumidores; a maximização do retorno financeiro das inspeções, que foi capaz de aumentar em até 11,5 vezes o retorno das inspeções de campo no melhor caso. A partir dos resultados alcançados, é possível concluir que a nova metodologia proposta neste trabalho representa um avanço em relação aos demais trabalhos disponíveis na bibliografia, já que fornece uma solução para a caracterização da perda não técnica; identifica um novo algoritmo com desempenho superior para classificação dos

consumidores; apresenta uma abordagem para estimar a energia não medida nos sistemas; fornece uma estratégia para estimar o retorno financeiro das inspeções *in-loco* e, por fim, apresenta um método para maximização do retorno das ações de combate às perdas. De modo que todas essas contribuições preenchem lacunas existentes na bibliografia atualmente disponível. Por fim, destaca-se que os resultados deste trabalho podem ser utilizados por distribuidoras de energia elétrica para melhorar suas estratégias de gestão da perda não técnica de energia, propiciando um aumento de receita e a redução dos seus custos operacionais. O que, por sua vez, irá se refletir como redução na tarifa de energia elétrica em benefício de toda a sociedade.

Palavras-chave: *Advanced Analytics*, Inferência Causal, *Machine Learning*, maximização, perda não técnica, retorno financeiro, sistemas elétricos de distribuição.

Abstract

A new methodology for improving the management of non-technical loss in electrical distribution power systems is presented in this work. The solution consists of using techniques based on Advanced Analytics to build an automated and adaptive process capable of identifying risk and protection factors for non-technical loss' occurrence; likelihood of loss existing; an estimate of energy not measured in the system and a potential financial return for on-site inspections of consumers. To perform the research, a database with real information of 261,489 consumers from a Brazilian utility was used. Causal Inference was applied to identify the degree of association of several features with the occurrence of non-technical loss and, thus, to identify the risk and protection factors. The variables associated with the occurrence of loss were used as input into Machine Learning models in order to identify the probability of loss occurrence, as well as providing an estimate of the value of unmeasured energy in the system. In total, 23 classification models and seven regression models were evaluated with different algorithms and data approaches. The results of the predictive models were used to calculate the potential of financial return from field inspections. Subsequently, a model was proposed to determine the optimal infrastructure for loss management in a utility, which results in the maximization of the financial return of field inspections. All stages of the methodology were validated through new field inspections carried out on 1,417 consumers. The main results achieved in the work were the identification of 76 risk or protection factors for the occurrence of non-technical loss; the identification of the Rotation Forest algorithm as the most suitable for loss identification, which presented an precision of 66.5% in the field inspections carried out; the identification of the XGBT algorithm as the most suitable for prediction of unmeasured energy values; which showed a deviation of 3.02% in estimate of the total value of energy recovered in a group of consumers; maximizing the financial return of field inspections, which was able to increase the return on field inspections by up to 11.5 times in the best case. From the achieved results, it is possible to conclude that the new methodology proposed in this work represents an valuable improvement compared to other works in available bibliography, since it provides a solution for the characterization of non-technical loss; identifies a new algorithm with superior performance for classifying consumers; presents an approach to estimate unmeasured energy in systems; provides a strategy to estimate the financial return of on-site inspections and, finally, presents a method for maximizing the return of actions in non-technical loss management. So

that all these contributions fill gaps in the existing bibliography. Finally, it should be noted that this work's results can be used by utilities to improve their non-technical loss management strategies, providing an increase in revenue and a reduction in their operational costs. Which, in turn, will be reflected in a reduction in electricity tariff for the benefit of the whole society.

Keywords: Advanced Analytics, Causal Inference, Machine Learning, maximization, non-technical loss, financial return, distribution power systems.

Lista de Ilustrações

Lista de Tabelas

Lista de Abreviaturas e Siglas

AA	*Advanced Analytics*
Adaboost	*Adaptive Boosting*
ANEEL	Agência Nacional de Energia Elétrica
AUC	*Area Under Curve*
AWS	*Amazon Web Service*
BHA	*Black Hole Algorithm*
BIC	*Bayesian Information Criteria*
BT	*Bagging Tree*
CELESC	Centrais Elétricas de Santa Catarina
CNN	*Convolutional Neural Network*
CPBETD	Consumption Patter-Based Energy Theft Detector
DT	Árvore de Decisão
ELM	*Extreme Learning Machine*
FIS	*Fuzzy Inference System*
FN	Falso Negativo
FP	Falso Positivo
FR	Lógica Fuzzy
GBC	*Gradient Boosted Classifier*
GBT	*Gradient Boosted Tree*
GRI	*Generalized Rule Induction*
IA	Inteligência Artificial
ICMS	Impostos sobre a Circulação de Mercadorias e Prestação de Serviços
IPTU	Imposto Predial e Territorial Urbano
KBS	*Knowledge-Based System*
LR	Regressão Logística
LSTM	*Long Short-Term Memory*
MAP	*Maximum a Posteriori*
MAPE	*Mean Absolute Percentage Error*
ML	*Machine Learning*
MLP	*Multi-Layer Perceptron*

NB	Naïve Bayes
OPF	*Optimum-Path Forest*
OR	*Odds Ratio*
PCA	*Principal Component Analysis*
PNT	Perda Não Técnica
PTEC	Perda Técnica
RA	Risco Atribuível
RF	*Random Forest*
RMSE	*Root Mean Square Error*
RNA	Redes Neurais Artificiais
ROC	*Receiver Operating Characteristic*
RP	Razão de Prevalência
RTF	*Rotation Forest*
SMOTE	*Synthetic Minority Over-sampling Technique*
SVM	*Support Vector Machine*
TextCNN	*Text Convolutional Neural Networks*
TPE	*Tree-structured Parzen Estimators*
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
VPN	Valor Preditivo Negativo
XGBT	*eXtreme Gradient Boosted Tree*

Lista de Símbolos

L	Distância entre o 1º e o 3º quartil no gráfico de *Box-plot*
K	Fator multiplicativo no método *Z-score*
Y	Valor original da variável em um conjunto de dados
Y'	Valor normalizado de uma variável
$\hat{Y}$	Valor previsto para uma variável
$Y_{mín}$	Valor mínimo da variável em um conjunto de dados
$Y_{máx}$	Valor máximo da variável em um conjunto de dados
N	Quantidade de variáveis explicativas disponíveis
P	Função de probabilidade
k	Número de centroides em um modelo de clusterização
l	Grau de liberdade de um modelo
α	Taxa de aprendizagem do modelo
δ	Gradiente da função
S	Função de ativação de uma Rede Neural Artificial
G	Função *softmax*
$b^{(1)}$	*Bias* de uma rede neural
$w^{(1)}$	Pesos sinópticos de uma rede neural
e	Número de Euler
σ	Desvio padrão
$\bar{\sigma}$	Média do conjunto de dados
X	Variável genérica
μ	Valor médio verdadeira
$\bar{x}$	Valor médio de uma amostra
n	Número de indivíduos em uma amostra
ρ	Coeficiente de correlação de Spearman
φ	Coeficiente de correlação de Spearman para variáveis qualitativas
rg_X	Ordem de classificação da variável X
ϵ	Explicabilidade
β_0	Coeficientes da função de regressão logística

k	Parâmetro do algoritmo *Rotation Forest*
γ	Taxa para divisão de parâmetros na Otimização Bayesiano
R^2	Coeficiente de Determinação
n	Quantidade de amostras em um conjunto de dados
k	Quantidade de interação no método *k-fold*
R	Retorno financeiro de uma inspeção em campo
$\mathrm{E_{kWh}}$	Energia recuperada a partir de uma inspeção em campo
T	Tarifa de venda da energia elétrica sem impostos
C	Custo operacional médio para executar uma inspeção em campo
$\mathrm{R_P}$	Potencial de retorno financeiro de uma inspeção em campo
$\mathrm{P_{PNT}}$	Probabilidade de ocorrência de PNT em um consumidor
$\mathrm{P_{E_{kWh}}}$	Previsão da energia recuperada a partir de uma inspeção em campo

Sumário

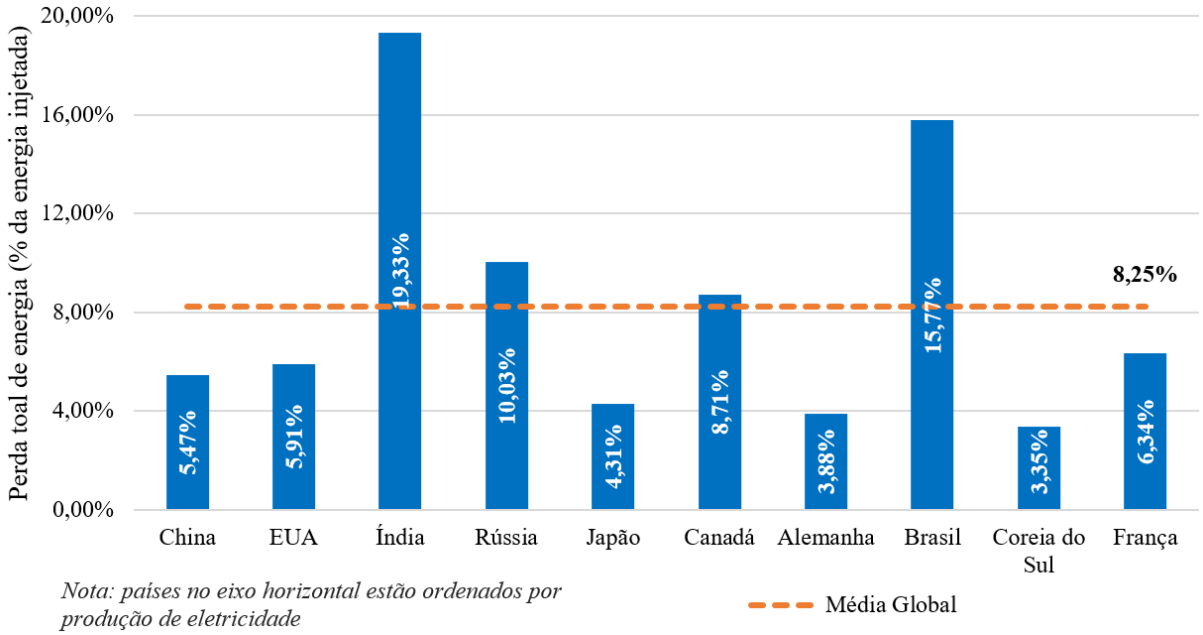
1 Introdução

1.1 Contextualização

A perda de energia no sistema elétrico de distribuição pode ser definida como a diferença entre a quantidade de energia injetada no sistema e a quantidade de energia consumida pelas cargas que é efetivamente medida pelas concessionárias (ANTMANN, 2009). A perda na distribuição de energia está diretamente relacionada à qualidade da gestão exercida pelas concessionárias, além de fatores socioeconômicos e geográficos de cada região, como o nível de escolaridade, poder aquisitivo, violência e densidade populacional. Por isso, o nível de perda nos sistemas elétricos pode variar de 3% a 40% da energia injetada em diferentes países (WORLD BANK, 2018).

No gráfico da Figura 1 é apresentado o nível de perda de energia elétrica nos 10 países com maior produção de eletricidade no mundo.

Figura 1 – Perda de energia elétrica nos 10 países com maior produção de eletricidade de mundo.



Fonte: Adaptado de World Bank (2018).

Na Figura 1, o Brasil destaca-se negativamente como o segundo país com os maiores níveis de perda de energia, perdendo apenas para a Índia, cujo índice representa aproximadamente o dobro da média global. Quando comparado a países desenvolvidos como

Estados Unidos, Japão e Alemanha, o Brasil possui um nível de perdas três vezes maior, o que consolida sua posição negativa no cenário global (WORLD BANK, 2018).

Outros países também possuem níveis elevados de perda em várias regiões do mundo, como Venezuela (36,04%), México (13,71%), Portugal (10,3%), países do sul da Ásia (18,84%), do Oriente Médio (13,49%), da África Subsaariana (11,74%) e da Europa Oriental (10,58%) (WORLD BANK, 2018).

Os dados demonstram que a perda de energia é um problema global e, por isso, o tema representa também uma área de pesquisa latente ao redor do mundo, na qual diferentes autores têm apresentado soluções para contribuir com a redução dos índices de perda nos sistemas elétricos. Tal fato é corroborado pelo crescente número de publicações neste tema como destacado em Viegas e Esteves (2017), Messinis e Hatziargyriou (2018) e Saeed *et al.* (2020).

Para o estudo da perda de energia é necessário inicialmente identificar suas origens. Com respeito a esse aspecto, a perda pode ser classificada em dois tipos distintos: perda técnica (PTEC) e perda não técnica (PNT), também conhecida como perda comercial (CIRED, 2017).

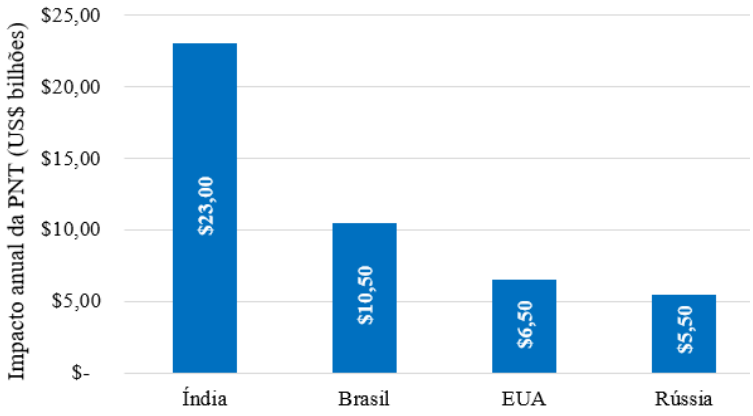
A perda técnica é inerente ao processo de condução de energia e tem origem em fenômenos físicos que ocorrem nos equipamentos elétricos, tais como, efeito Joule nos condutores, correntes parasitas e histerese no núcleo de transformadores de potência e correntes de fuga nos isoladores e para-raios. Por se tratar de um fenômeno físico relacionado às propriedades dos materiais, não é possível eliminar a perda técnica no processo de distribuição de energia. Sabendo disso, o que se busca é a operação do sistema no menor nível possível de perdas técnicas, preservando a viabilidade econômica no processo (CIRED, 2017).

Diferentemente da perda técnica, a perda não técnica não tem origem natural. Ela pode ser definida como a quantidade de energia que é consumida, mas não é registrada pelos sistemas de medição das concessionárias, ou seja, trata-se de um consumo irregular de energia com consequente perda de faturamento por parte das concessionárias. A perda não técnica pode ser provocada por falhas técnicas nos equipamentos de medição ou erros no processo de leitura e faturamento do consumo. Contudo, na maior parte dos casos, a irregularidade ocorre por má fé do consumidor, que se utiliza de diferentes métodos para burlar o sistema de medição (JAMIL e AHMAD, 2019).

1.2 Relevância

Diversos problemas estão associados à perda não técnica de energia, os quais se refletem na forma de prejuízos para toda a sociedade. Do ponto de vista econômico, segundo levantamento do Northeast Group (2017), o impacto global causado pela perda não técnica de energia é de US$ 96 bilhões por ano. O Brasil sozinho representa 11% deste impacto, valor menor apenas que o da Índia, que representa 24% do total. Na Figura 2 são apresentados os impactos econômicos causados pela perda não técnica em outros países de porte semelhante ao Brasil, em termos de produção de energia.

Figura 2 – Impacto financeiro da PNT em alguns países.



Fonte: Adaptado de Northeast Group (2017).

Os dados apresentados na Figura 2 revelam a dimensão do prejuízo causado pela perda não técnica de energia. Parte do prejuízo apresentado é assumido pelas concessionárias de distribuição e outra parte é repassada ao consumidor por meio da tarifa ou de subsídios do governo, a depender do modelo regulatório adotado no país (CIRED, 2017). Desse modo, a perda não técnica é responsável pela elevação no preço das tarifas de energia, fazendo com que consumidores regulares acabem pagando pelo consumo dos irregulares, o que é um dos impactos sociais negativos provocado pela perda não técnica. No caso do Brasil, em média 3% do custo da tarifa de energia elétrica é relativo ao custo da perda não técnica que é repassada para os consumidores, sendo que em algumas distribuidoras esse percentual chega a 21,5% do valor da tarifa (ANEEL, 2019).

Outro impacto social relevante é a redução na arrecadação de impostos. Uma vez que que a energia consumida de forma irregular não é faturada pelas concessionárias, não existe o recolhimento de impostos sobre tal consumo. Considerando o caso Brasil em que a perda não técnica anual é da ordem de 25 TWh, o estado deixa de arrecadar R$ 4,5 bilhões apenas em

Impostos sobre a Circulação de Mercadorias e Prestação de Serviços (ICMS), valor que poderia ser convertido em benefícios para a população (ANEEL, 2019).

Do ponto de vista técnico, a perda não técnica compromete a qualidade do fornecimento de energia elétrica aos consumidores. Dado que o planejamento do sistema é feito com base nas cargas regulares, é possível que equipamentos fiquem subdimensionados devido à existência das cargas irregulares que compõem a perda não técnica. Tal fato compromete todo o planejamento de expansão do sistema e causa problemas como subtensão e descontinuidade no fornecimento. Como consequência, mais uma vez, o consumidor regular é penalizado pela prática do consumidor irregular.

A perda não técnica também compromete à segurança do sistema elétrico, especialmente nos casos de consumidores clandestinos que se conectam diretamente ao sistema, intervindo na rede energizada sem nenhum equipamento de segurança. Além disso, também existe um impacto ambiental associado à perda não técnica, pois, em instalações com fraude, os consumidores tendem a aumentar o consumo de energia devido ao fato de que o aumento do consumo não representará um aumento na fatura a ser paga. Há estudos que indicam que o consumidor fraudador reduz seu consumo em até 50% após ter seu sistema de medição regularizado (ANTMANN, 2009).

Dessa forma, a eliminação da perda não técnica também aumenta a racionalidade do consumo e tende a reduzir a demanda do sistema como um todo. Consequentemente, haverá uma menor necessidade de ampliação da capacidade de geração e transmissão instalada, reduzindo os impactos ambientais causados pelo setor. Além disso, com uma carga menor, também haverá uma redução na perda técnica existente nos equipamentos, que é outro benefício da redução de perda não técnica.

A partir do exposto nos parágrafos anteriores, fica evidenciado a relevância do tema central deste trabalho, que é a gestão da perda não técnica de energia, bem como, os benefícios econômicos, técnicos e sociais que podem ser alcançados com as contribuições fornecidas pelos resultados da pesquisa.

1.3 Motivação

Diante de todo o contexto exposto, fica evidente a necessidade de concessionárias de distribuição em realizar uma gestão eficiente da perda não técnica de energia em suas áreas de concessão. Assim, praticamente todas as concessionárias estabelecem programas com o

objetivo específico de reduzir ou controlar os indicadores de perda não técnica. Dentre as ações realizadas estão campanhas educativas, ações em parceria com o poder público e instalação de equipamentos resistentes à fraude na rede elétrica.

Contudo, a ação mais efetiva executada pelas concessionárias para redução da perda não técnica é a realização de inspeções *in loco* nos sistemas de medição dos consumidores. As inspeções têm como objetivo identificar e corrigir as irregularidades na medição. Além disso, durante as inspeções são registradas evidências que possam comprovar a má fé do consumidor na manipulação do sistema de medição. A partir disso, é possível tomar medidas legais para cobrar, de forma retroativa, o consumo que deixou de ser medido, bem como denunciar a prática ilícita às autoridades competentes.

Por outro lado, a realização de inspeções *in loco* possui limitações operacionais e de custo, o qual pode variar entre US$ 30 e US$ 100 por inspeção a depender do país (MANAGEMENT SOLUTION, 2017) e (MANO, 2017). Por essa razão, não é possível inspecionar todos os consumidores em um curto período de tempo. Assim, as concessionárias costumam realizar campanhas anuais de fiscalização, nas quais grupos selecionados de consumidores são inspecionados em uma quantidade que pode variar de 1% a 10% do total, dependendo da área de concessão (MANAGEMENT SOLUTION, 2017).

A tarefa de pré-selecionar grupos de consumidores a serem fiscalizados é fundamental para definir a rentabilidade das campanhas de inspeção, uma vez que se a maioria dos consumidores fiscalizados não apresentarem irregularidades em seus sistemas de medição, as ações serão inócuas e não trarão qualquer retorno para as concessionaras. Apesar disso, as taxas de sucesso alcançadas pelas empresas no processo de seleção de consumidores são relativamente baixas. Em Management Solution (2017), Eller (2003), Mondero *et al.* (2012) e Penin (2008) são reportadas taxas de 9,0%, 8,3%, 10% e 12% respectivamente, em concessionárias de países como Brasil e Espanha.

As baixas taxa de sucesso estão associadas, principalmente, à dificuldade em se identificar quais são as causas que levam à ocorrência de perda não técnica em uma unidade consumidora. Provavelmente, as verdadeiras causas estão associadas a fatores intangíveis, como comportamento cultural, complexidade socioeconômica e nível de escolaridade do consumidor.

Apesar disso, é possível que sejam determinadas variáveis práticas que possuam alguma associação com a ocorrência de perda não técnica, ou em outras palavras, fatores de risco e fatores de proteção. Por exemplo, os consumidores que apresentam uma queda no seu consumo

médio têm maior ou menor chance de apresentar perda não técnica? O modelo do medidor de energia é um fator de risco para a ocorrência de perda?

Estas são questões práticas que ainda não foram claramente discutidas na bibliografia disponível sobre o tema. Esse entendimento é corroborado por Saeed *et al.* (2020), que analisaram um total de 85 publicações em periódicos científicos entre os anos de 2000 e 2020, que tratavam de soluções para identificação de perda não técnica no sistema de distribuição. Entre as principais conclusões dos autores, está o fato de haver uma lacuna significativa nos métodos teóricos que investigam as causas da perda não técnica.

Constata-se, portanto, a existência de uma lacuna nos trabalhos publicados quanto à determinação de fatores de risco e de proteção para ocorrência da perda não técnicas de energia em sistemas elétricos de distribuição.

Além disso, o conhecimento prévio de quais consumidores têm maior probabilidade de possuir perda não técnica é de grande relevância para a definição de estratégias de gestão por parte das concessionárias. Por isso, muitos autores têm proposto a utilização de modelos baseados em *Machine Learning* (ML) para a identificação da perda não técnica nos consumidores dos sistemas elétricos, o que é corroborado por Messinis e Hatziargyriou (2018). Os autores listam dezenas de trabalhos que utilizam diferentes algoritmos para a classificação dos consumidores.

Contudo, apesar da quantidade de publicações no tema, algumas questões importantes ainda não foram totalmente esclarecidas. Dentre essas questões, podem ser destacadas as seguintes:

- Qual é a técnica de ML mais adequada para a identificação de PNT?
- A combinação de resultados de diferentes classificadores pode melhorar a identificação de PNT?
- Em aplicações reais de campo, o desempenho dos classificadores permanece o mesmo que no conjunto de dados de treinamento/teste?
- Faz sentido realizar abordagens alternativas de classificação, como técnicas de agrupamento antes da classificação dos dados?

Alguns motivos podem ser considerados para justificar o porquê de essas questões ainda permanecerem em aberto. Como, por exemplo, o fato de que a maioria dos estudos consiste em aplicar e testar uma única técnica a um conjunto específico de consumidores. Não há padronização com relação ao tipo de variáveis utilizadas ou com relação ao tamanho do banco de dados. Até mesmo a métrica de avaliação dos resultados apresenta uma grande variabilidade,

alguns trabalhos utilizam a precisão, outros a Área sob a Curva *Receiver Operating Characteristic* (ROC), outros o *F1-score* e assim por diante (GUERRERO *et al.*, 2018), (SAEED *et al.*, 2019) e (PASSOS JR *et al.*, 2016).

A falta de padronização dificulta a comparação entre resultados de diferentes estudos. Como consequência, praticamente não existem trabalhos que se proponham a comparar os resultados de diferentes publicações. Outro problema é que existem poucos trabalhos que apresentam resultados reais de inspeções de campo, portanto, não é possível avaliar a capacidade dos algoritmos em manter o seu desempenho em aplicações reais.

Constata-se, portanto, mais uma lacuna existente na bibliografia em relação a utilização de algoritmos de classificação para a identificação da perda não técnica de energia nos sistemas de distribuição.

Outro ponto a ser destacado é que, como será discutido no Capítulo 3, na maior parte dos trabalhos disponíveis, os autores estão focados em encontrar formas de identificar a presença de perda não técnica nos consumidores sem diferenciar a quantidade de energia que deixou de ser medida em cada caso. Basicamente, a qualidade das soluções propostas é discutida em termos de métricas como precisão, acurácia, *F1-score*, dentre outras.

No entanto, a abordagem acima pode não ser a mais adequada para fins de gerenciamento da perda em concessionárias de distribuição de energia. Para ilustrar essa ideia, considere as seguintes situações:

- Situação 1: uma distribuidora seleciona 100 consumidores para inspeções em campo e encontra perda não técnica em 80 deles. No entanto, eles são pequenos consumidores ou a perda começou há apenas algumas semanas. Assim, a quantidade total de energia não medida que pode ser recuperada pela distribuidora é de apenas 1 MWh;
- Situação 2: uma distribuidora seleciona 100 consumidores para inspeções em campo e encontra perda não técnica em 30 deles. Eles são grandes consumidores e a perda começou há meses. Assim, a concessionária pôde recuperar 100 MWh de energia não medida nos consumidores.

Embora as duas situações apresentem o mesmo custo operacional para a concessionária, a Situação 2 proporciona um retorno financeiro 100 vezes maior do que a Situação 1. Logo, a Situação 2 é a mais desejada. No entanto, na Situação 1 mais consumidores com perda não técnica são identificados, ou seja, a Situação 1 apresenta uma precisão maior no processo de seleção dos consumidores.

Portanto, focar apenas em classificar consumidores entre instalações com ou sem perda não técnica, como um problema de classificação binária, não garantirá a melhor abordagem para gerenciamento da perda em concessionárias de distribuição de energia.

Em vez disso, a gestão da perda deve focar no retorno financeiro das ações. Apesar dessa conclusão, praticamente não existem trabalhos disponíveis que apresentem uma abordagem focada no retorno financeiro das inspeções em campo, o que evidencia mais uma lacuna existente na bibliografia.

Os parágrafos anteriores evidenciam uma série de lacunas existentes na bibliografia relacionada à gestão da perda não técnica de energia nos sistemas elétricos de distribuição. Fica evidente, portanto, que tais lacunas são a motivação para a realização do presente trabalho. Nesse contexto, com os resultados da pesquisa pretende-se contribuir para o esclarecimento das questões levantadas, bem como fornecer novas soluções que possam aperfeiçoar a gestão da perda não técnica nos sistemas de distribuição, ajudando a mitigar os danos causados às concessionárias e a sociedade.

A partir das motivações expostas nesta seção, foram definidos os objetivos gerais e específicos do trabalho, os quais são necessários para o alcance dos resultados desejados e são apresentados na próxima seção.

1.4 Objetivos

O objetivo geral do presente trabalho é desenvolver e validar uma nova metodologia baseada em técnicas de *Advanced Analytics* para aprimorar a gestão da perda não técnica de energia em sistemas elétricos de distribuição. Para alcançar o objetivo geral, os seguintes objetivos específicos são necessários:

- Executar uma análise de inferência causal para caracterizar a perda não técnica de energia e identificar fatores de riscos e de proteção para sua ocorrência;
- Desenvolver e validar modelos preditivos de classificação para a identificação da perda não técnica em sistemas elétricos de distribuição;
- Desenvolver e validar modelos preditivos de regressão para a previsão do valor de energia não medida e passível de recuperação em consumidores com perda não técnica;
- Desenvolver um modelo matemático para previsão do retorno financeiro decorrente das ações de inspeção em consumidores do sistema elétrico de distribuição;

- Propor um novo método de seleção de consumidores para inspeções de perda não técnica baseado na maximização do retorno financeiro das ações;
- Propor um novo método para o dimensionamento ótimo da infraestrutura de inspeção em campo de distribuidoras de energia elétrica;
- Validar os resultados obtidos a partir da realização de novas inspeções de campo em consumidores reais do sistema elétrico de distribuição.

1.5 Contribuições

A partir da motivação e dos objetivos apresentados, é possível verificar que o presente trabalho representa um avanço em relação à bibliografia disponível e que os resultados alcançados são uma importante contribuição para a gestão de perdas não técnicas em distribuidoras de energia elétrica, bem como para o estado da arte do tema. Dentre as contribuições específicas da pesquisa, destacam-se as seguintes:

- Identificação de fatores de risco e de proteção para ocorrência da perda não técnica de energia em sistemas elétricos de distribuição, os quais podem ser utilizados para definir tomadas de decisão na gestão da perda não técnica;
- Avaliação comparativa de 23 modelos de classificação distintos para a identificação da perda não técnica em sistemas elétricos, a partir da qual pode ser identificado o modelo mais adequado para a tarefa;
- Proposição de método original para definição do limiar de probabilidade em modelos de classificação capaz de maximizar de forma equilibrada a precisão e sensibilidade dos modelos;
- Criação de modelos de regressão para previsão da energia não medida em consumidores com perda não técnica;
- Definição de modelo matemático para representar o potencial de retorno financeiro de inspeções em campo em sistemas elétricos de distribuição;
- Proposição de novo método de seleção para inspeções de consumidores nos sistemas elétricos de distribuição, capaz de maximizar o retorno financeiro das ações;
- Proposição de novo método para o dimensionamento ótimo da infraestrutura de inspeções em campo em distribuidoras de energia elétrica.

Além disso, a partir da aplicação dos resultados alcançados na pesquisa por concessionárias de distribuição de energia, a pesquisa também poderá contribuir para os seguintes pontos:

- Aumento no faturamento das concessionárias de distribuição de energia elétrica;
- Redução do valor da tarifa de energia elétrica;
- Melhoria no planejamento dos sistemas elétricos de distribuição;
- Uso racional da energia nos sistemas elétricos;
- Aumento da segurança e da qualidade do fornecimento de energia nos sistemas elétricos de distribuição.

Também é uma contribuição da pesquisa a divulgação científica realizada por meio de publicações em eventos e em periódicos, os quais são listados na seção seguinte.

1.6 Publicações

A divulgação dos resultados obtidos ao longo desta pesquisa foi realizada por meio da publicação de artigos em periódicos e congressos científicos internacionais e nacionais. Os trabalhos listados a seguir foram publicados ou estão em etapa de revisão para publicação.

Periódicos Internacionais:

- **Barros, R. M. R.**; Costa, E. G.; Araujo, J. F. "Maximizing the Financial Return of Non-Technical Loss Management in Power Distribution Systems," *IEEE Transactions on Power Systems*, p.p. 1-8, 2021. [1]
- **Barros, R. M. R.**; Costa, E. G.; Araujo, J. F. "Evaluation of Classifiers for Non-Technical Loss Identification in Electric Power Systems," *International Journal of Electrical Power and Energy Systems*, vol. 132, p.p. 1-9, 2021.
- **Barros, R. M. R.**; Costa, E. G.; Araujo, J. F. "Causal Inference Analysis Applied to Non-Technical Loss Characterization in Distribution Power Systems," *IEEE Transactions on Power Delivery*, p.p. 1-8, 2021. [1]

[1] O artigo já passou pela 1ª etapa de revisão da revista e aguarda análise da ressubmissão.

- **Barros, R. M. R.**; Costa, E. G.; Araujo, J. F.; Andrade, F. L. M.; Ferreira, T. V. "The Contribution of Inrush Current to Mechanical Failure of Power Transformers Windings," *High Voltage*, vol. 4, p.p. 300-307, 2019. [2]

Congressos Internacionais:

- Alves, H. M. M.; Oliveira, I. B.; **Barros, R. M. R.**; Araujo, J. F.; Costa E. G. "Detection of Non-Technical Loss Using Data Mining," *In: 3rd International Workshop on Advanced Dielectrics and Applications*, 2019, Campina Grande.

Congressos Nacionais:

- **Barros, R. M. R.**; Costa, E. G.; Araujo, J. F.; Lira, G. R. S. "Use of ANN for Identification of Consumers with Irregular Electrical Installations," *In: 2018 Simpósio Brasileiro de Sistemas Elétricos (SBSE)*, 2018, Niterói.

1.7 Organização do Texto

Além da Introdução, o presente trabalho está organizado em mais cinco capítulos que são descritos resumidamente a seguir:

- No Capítulo 2 é realizado o embasamento teórico a respeito dos principais temas que permeiam a pesquisa, dos quais podem ser destacados os fundamentos relacionados à perda de energia no sistema elétrico de distribuição e os conceitos fundamentais do processo de *Advanced Analytics*, como o pré-processamento de dados, análise de inferência causal, clusterização de dados, além de métodos para aprendizagem de máquina e otimização;
- No Capítulo 3 são apresentadas, em ordem cronológica, as principais pesquisas realizadas nos temas correlatos a este trabalho, tais como, técnicas baseadas em aprendizado de máquina para detecção de fraudes, soluções de *smart meter* para balanço energético no sistema elétrico e soluções alternativas para redução da perda não técnica no sistema de distribuição, compondo assim o estado da arte para o tema;

[2] A publicação não está diretamente relacionada ao tema da tese.

- No Capítulo 4 são apresentados o material e os métodos empregados nas etapas executadas para alcançar os objetos da pesquisa, dentre eles as principais características da base de dados utilizada, a plataforma KNIME® e a arquitetura de computação utilizada em nuvem, bem como os detalhes das etapas de tratamento dos dados, extração e seleção de variáveis, otimização, criação de modelos preditivos e de maximização;
- No Capítulos 5 são apresentados em detalhes todos os resultados obtidos na pesquisa, bem como as discussões e possíveis conclusões que podem ser extraídas dos mesmos;
- No Capítulo 6 são apresentadas as considerações finais do trabalho, destacando-se as contribuições dos resultados alcançados e propondo trabalhos futuros para a continuação da linha de pesquisa;
- Ao final do texto, são apresentadas ainda todas as referências bibliográficas citadas ao longo do trabalho.

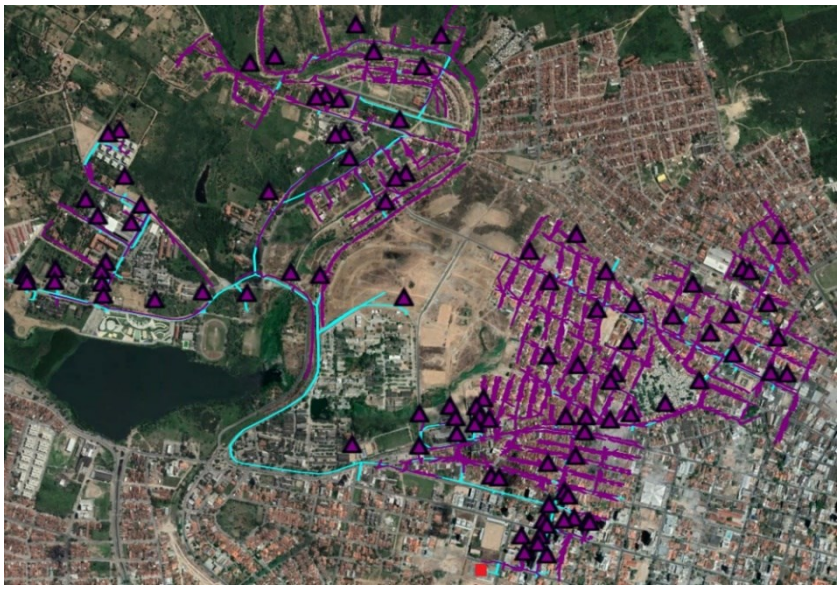
2 Fundamentação Teórica

Neste capítulo são apresentados os conceitos teóricos necessários para o entendimento da metodologia utilizada no trabalho, bem como para a análise e discussão dos resultados obtidos. Assim, são destacados os conceitos fundamentais sobre: perdas de energia no sistema elétrico de distribuição e as principais etapas do processo de *Advanced Analytics* como tratamento de dados, seleção de variáveis, inferência causa, técnicas de clusterização, técnicas de aprendizagem de máquinas e técnicas de otimização.

2.1 Perda de Energia em Sistemas Elétricos de Distribuição

O processo de distribuição de energia elétrica consiste em transportar energia a partir do sistema de transmissão, ou de unidades geradoras de médio e pequeno porte, aos consumidores finais (ABRADEE, 2017). Para realizar o transporte de energia são necessários diversos equipamentos como cabos condutores, transformadores, reguladores e equipamentos de medição, controle e proteção. Como exemplo, na Figura 3 é possível observar a representação geográfica do alimentador de distribuição que atende ao campus sede da Universidade Federal de Campina Grande.

Figura 3 – Representação geográfica do alimentador que atende ao campus sede da Universidade Federal de Campina Grande.



Fonte: Cortesia Energisa (2019).

Na Figura 3 o quadrado vermelho representa a subestação de distribuição, onde se inicia o alimentador. Os condutores de média tensão são representados pelas linhas na cor ciano e os condutores de baixa tensão são representados pelas linhas na cor magenta. Por fim, os transformadores de distribuição são representados pelos triângulos.

A partir da Figura 3 é possível constatar que o sistema de distribuição é bastante capilarizado e composto por uma grande quantidade de condutores e outros equipamentos, o que naturalmente faz com que exista uma perda significativa de energia elétrica no processo de distribuição. Esta perda pode ser classificada em dois tipos distintos de acordo com sua origem: perda técnica e perda não técnica. Ambos os tipos são discutidos nas subseções seguintes.

2.1.1 Perda Técnica

A perda técnica ocorre nos equipamentos instalados no sistema de distribuição devido a efeitos físicos que são inerentes aos materiais elétricos que constituem os equipamentos (KAGAN, OLIVEIRA e ROBBA, 2010), como por exemplo:

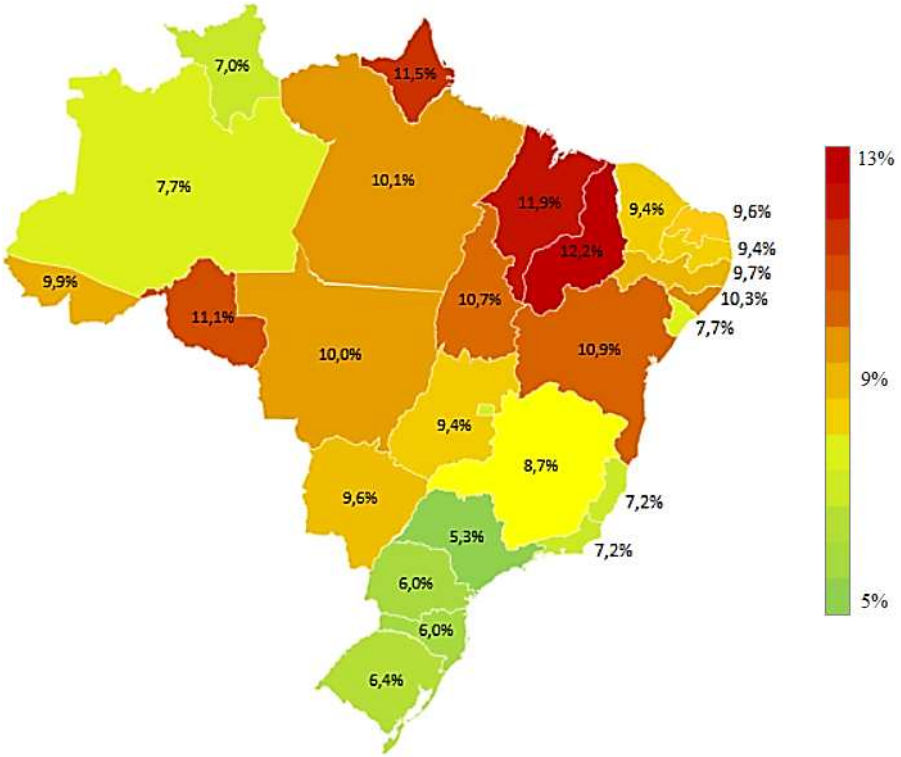
- Efeito Joule nos condutores;
- Corrente de fuga nos isoladores e para-raios;
- Correntes parasitas e histerese no núcleo de transformadores; e
- Perdas internas em equipamentos de medição, proteção e controle.

A perda técnica geralmente é contabilizada em termos percentuais em relação à energia injetada no sistema. O valor pode variar em função de alguns aspectos como densidade de carga, nível de tensão de operação, carregamento dos equipamentos e penetração de geração distribuída na rede. Na Figura 4 é possível observar o nível de perda técnica em cada estado do Brasil de acordo com estimativas da Agência Nacional de Energia Elétrica (ANEEL).

A partir da análise da Figura 4, constata-se que os estados das regiões Norte e Nordeste, em geral, possuem níveis de perda técnica maior que os estados das regiões Sul e Sudeste. Tal fato pode ser justificado pela característica da carga nestas regiões. Enquanto nas regiões Sul e Sudeste existe uma alta densidade de carga com alta participação das cargas industriais, nas regiões Norte e Nordeste prevalecem regiões com baixa densidade de carga e baixa presença industrial, o que tende a elevar os níveis de perda técnica. Tal fato ocorre porque para atender cargas geograficamente distantes é necessário um comprimento maior de linhas de distribuição, que causam maior efeito Joule, maior queda de tensão e uma maior quantidade de equipamentos

no sistema, fazendo com que o percentual de perda técnica em relação à energia injetada seja elevado.

Figura 4 – Percentual de perda técnica em relação a energia injetada no sistema de distribuição no Brasil.



Fonte: Aneel (2019).

Para estimar os níveis de perda técnica no sistema de distribuição apresentados na Figura 4, são utilizados *softwares* de fluxo carga baseados em métodos numéricos para resolver equações diferenciais relacionadas às grandezas elétricas. Dentre os *softwares* mais utilizados destacam-se o OpenDSS®, que é uma aplicação de código aberto, o SinapGrid® e o Pertec®, que são aplicações comerciais utilizadas pelas distribuidoras de energia elétrica.

Para as situações em que o nível de perda técnica for elevado, várias ações podem ser adotadas para reduzi-lo, tais como:

- Elevação do nível de tensão da rede;
- Aumento da capacidade de condução de corrente dos cabos;
- Uso de transformadores mais eficientes;
- Redução de fator de potência;
- Inserção de geração distribuída em pontos estratégicos; e
- Redução do desequilíbrio de corrente entre fases.

Contudo, para cada medida, a viabilidade de redução da perda técnica deve ser estudada em termos da comparação entre o custo dos equipamentos necessários à implantação da medida e o custo da geração de energia (ANTMANN, 2009).

Na prática, em poucos casos uma obra de melhoria da rede se viabiliza economicamente somente com o benefício da redução da perda técnica. A viabilidade é obtida devido a outros ganhos associados, como a melhoria no nível de tensão entregue ao consumidor e à redução na frequência e na duração das interrupções de fornecimento (ANTMANN, 2009).

Destaca-se por fim que a perda técnica não é a única forma de perda de energia no sistema elétrico de distribuição. A perda total do sistema é composta ainda por uma outra parcela conhecida como perda não-técnica, a qual é o foco desta pesquisa e será melhor discutida na subseção seguinte.

2.1.2 Perda Não Técnica

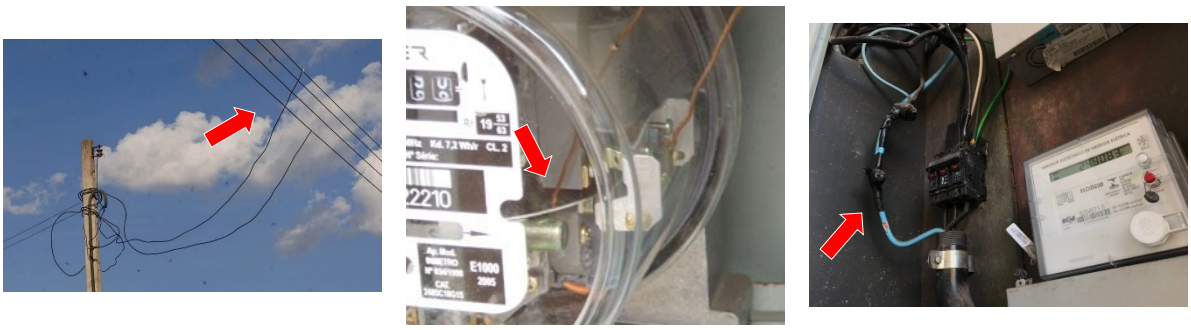
A perda não técnica pode ser definida como a energia que é consumida, mas não é registrada nos sistemas de faturamento das concessionárias (CIRED, 2017). Portanto, a perda não técnica ocorre devido a uma irregularidade no processo de medição e/ou faturamento da energia elétrica. Dentre as principais causas para esta irregularidade, podem ser destacadas as seguintes:

- Defeito técnico nos equipamentos de medição (medidores e transformadores de instrumentos);
- Erros operacionais no processo de leitura e/ou faturamento do consumo;
- Ligações clandestinas na rede de distribuição, caracterizadas pela ausência de medição de consumo;
- Adulteração nos equipamentos de medição, com objetivo de manipular os valores de energia registrados; e
- Desvios na conexão à montante do medidor (*by-pass*), fazendo com que parte da energia consumida não passe através do equipamento.

Os dois primeiros pontos destacados acima estão relacionados à gestão da distribuidora e, por isso, também são conhecidos como perda administrativa. Os outros três pontos estão relacionados à má fé do consumidor e caracterizam o crime de furto de energia, o qual está tipificado no Código Penal Brasileiro no artigo 155, que estabelece como pena a reclusão de

um a quatro anos e multa (BRASIL, 1940). Na Figura 5 são apresentados alguns registros de furto de energia na rede elétrica.

Figura 5 – Registros de furto de energia na rede elétrica: (a) ligação clandestina, (b) adulteração no medidor e (c) *by-pass* no ramal de ligação.



(a) (b) (c)

Fonte: (a) O Liberal (2018), (b) Diário de Taubaté (2019) e (c) Diário da Manhã (2019).

Na Figura 5 (a) é possível observar um caso de ligação clandestina, no qual o consumidor se conecta diretamente à rede elétrica sem nenhum tipo de medição para o seu consumo. Na Figura 5 (b) é possível observar um caso de adulteração do medidor de energia, em que foi realizada a perfuração do equipamento para inserção de um arame, com objetivo de criar resistência à rotação do disco, fazendo com que o consumo registrado seja menor que o real. Por fim, na Figura 5 (c) é possível observar um caso de desvio da corrente elétrica através de um *by-pass* no ramal de ligação. Neste caso a maior parte da energia consumida não passará através do medidor, fazendo com que o consumo faturado seja menor que o real.

A perda não técnica é um indicador difícil de ser quantificado diretamente, por isso, ela é obtida pela diferença entre a perda total e a perda técnica da distribuidora. O nível de perda não técnica pode variar significativamente, desde índices próximos a zero até 50% da energia injetada. Na Tabela 1 é apresentado um panorama global sobre os níveis de perda não técnica nas diferentes regiões do mundo.

Tabela 1 – Panorama global sobre o nível de perda não técnica nas diferentes regiões do mundo.

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%; • Bangladesh apresenta perda superiores a 20%; • Indonésia apresenta perdas de 7%; • Malásia apresenta perda de até 15%; • Tailândia apresenta perda de 11%.
América do Norte e Central	• México apresenta índices de 13%; • Estados Unidos apresentam percentuais próximos a zero, mas em algumas regiões são registradas ligações clandestinas relacionadas ao cultivo ilegal de maconha.

Região	Destaques
América do Sul	• Brasil apresenta índices de 7,3% a 25% dependendo da região; • Chile apresenta índices inferiores a 5%.
Oriente Médio e África	• Na África subsaariana o índice médio de perda é de 50%, Botsuana é uma exceção com índice de 10%; • Na África do Sul o índice médio é de 7%, mas atinge 50% entre os consumidores comerciais; • Senegal apresenta índice médio de 21%.

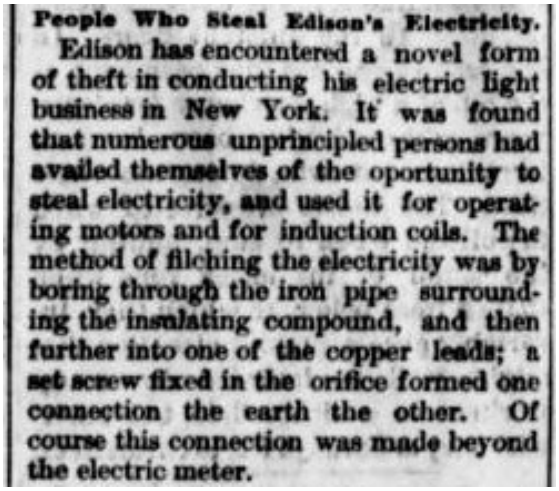
Fonte: CIRED (2017).

A partir da Tabela 1 é possível constatar que os países subdesenvolvidos e emergentes possuem índices superiores aos países desenvolvidos. Tal variação pode ser justificada pelos fatores que influenciam esse índice, dentre os quais destacam-se os seguintes:

- Poder aquisitivo da população;
- Nível de escolaridade;
- Complexidade socioeconômica da região;
- Valor da tarifa de energia elétrica em relação ao salário médio;
- Legislação vigente sobre o tema;
- Efetividade da aplicação da legislação pelo poder judiciário; e
- Gestão da concessionária.

A partir dos fatores apresentados, é possível verificar que a perda não técnica é um problema global, complexo e difícil de ser combatido. Além disso, este também não é um problema recente, como pode ser visto na Figura 6, em que é apresentada uma reportagem sobre furto de energia em 1886 em uma das primeiras redes de distribuição de eletricidade do mundo, instalada em Nova York pela *Edison General Electric Company* para a iluminação da cidade.

Figura 6 – Furto de energia noticiado em 1886 na cidade de Nova York.



People Who Steal Edison's Electricity.

Edison has encountered a novel form of theft in conducting his electric light business in New York. It was found that numerous unprincipled persons had availed themselves of the oportunity to steal electricity, and used it for operating motors and for induction coils. The method of filching the electricity was by boring through the iron pipe surrounding the insulating compound, and then further into one of the copper leads; a set screw fixed in the orifice formed one connection the earth the other. Of course this connection was made beyond the electric meter.

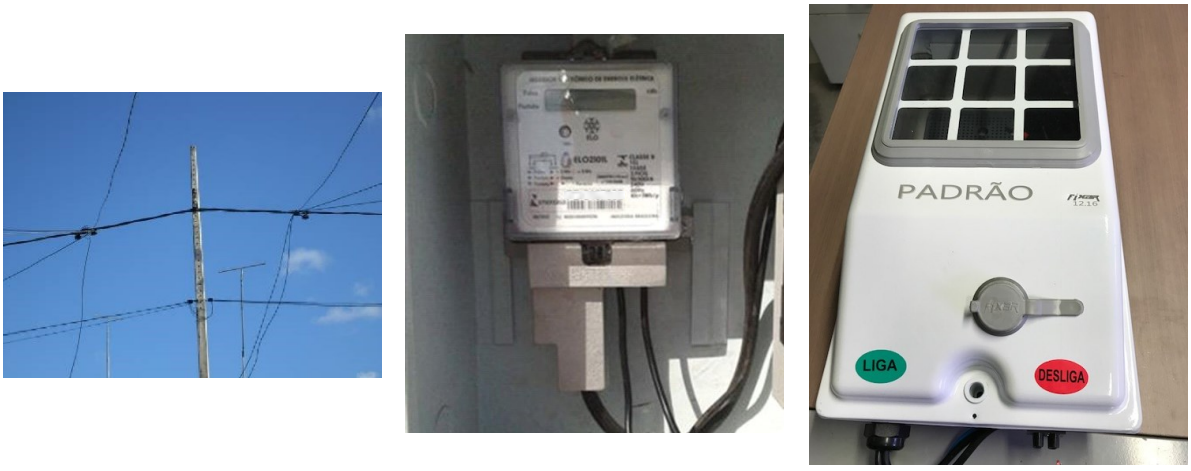
Fonte: Knight e Gordon (1886).

No caso reportado na Figura 6, o superintendente da companhia tentou resolver o problema aplicando pulsos de corrente na rede elétrica. Segundo a reportagem, os pulsos seriam suficientes para queimar as cargas irregulares conectadas, enquanto causariam apenas oscilações na iluminância das lâmpadas incandescente que compunham o sistema.

Obviamente, a solução tomada pela *Edison General Electric Company* não poderia ser aplica nos dias de hoje, pois, o sistema não é mais composto apenas por lâmpadas incandescentes. Contudo, o relato revela a necessidade histórica das concessionárias em minimizar a perda não técnica do sistema. Por isso, atualmente todas as distribuidoras possuem programas específicos de gerenciamento da perda não técnica.

O gerenciamento consiste tanto em ações de prevenção como de combate à perda não técnica. Dentre as ações de prevenção destacam-se as campanhas de conscientização, o aperfeiçoamento da legislação e as ações de blindagem da rede, como as destacadas na Figura 7.

Figura 7 – Ações de blindagem da rede para prevenção da perda não técnica: (a) blindagem de circuito de baixa tensão, (b) blindagem dos bornes do medidor e (c) blindagem da caixa de medição.



(a) (b) (c)

Fonte: (a) e (b) Cortesia Energisa (2019), (c) Fixar (2019).

Na Figura 7 (a) é possível observar uma blindagem no circuito de baixa tensão do sistema de distribuição, que consiste em substituir a rede aberta por cabos multiplexados e isolados com várias camadas de material de auto fusão. A medida tem o objetivo de dificultar o tipo de fraude apresentado na Figura 5 (a). Na Figura 7 (b) é possível observar a blindagem dos bornes do medidor por meio do encapsulamento com uma junção metálica que só pode ser aberta com uma chave de segurança. A medida tem o objetivo de dificultar o acesso aos bornes do medidor para possíveis fraudes. Por fim, na Figura 7 (c) é apresentada uma caixa de medição blindada, a qual é composta por material resistente e só pode ser aberta com uma chave de

segurança. A medida tem o objetivo de dificultar o acesso a qualquer parte do padrão de medição para possíveis manipulações.

Existem ainda outras ações de prevenção com diferentes custos e benefícios, tais ações podem ser aplicadas em conjunto e a escolha das mais apropriadas dependem das características de cada região. Contudo, as ações de prevenção têm efeito apenas paliativo e não são capazes de reduzir a perda não técnica já existente no sistema de distribuição. Por isso, as distribuidoras também investem em ações de fiscalização nas unidades consumidoras. Estas ações consistem na inspeção do sistema de medição com o objetivo de identificar possíveis fraudes ou defeitos nos equipamentos. Na Figura 8 é apresentado o procedimento de inspeção em uma unidade consumidora.

Figura 8 – Procedimento de inspeção no sistema de medição de uma unidade consumidora.

Fonte: SFn Notícias (2018).

Como pode ser observado na Figura 8, um técnico da companhia distribuidora de energia está acompanhado de um perito criminal para caracterização da irregularidade que foi detectada no sistema de medição, prática que é comum quando a fraude ocorre em consumidores de médio e grande porte. A partir das inspeções, é possível identificar as irregularidades no sistema de mediação, corrigi-las e recuperar parte do consumo de energia que deixou de ser medido.

As regras para recuperação do consumo de energia no Brasil são definidas pela Resolução Normativa nº 414/2010 da ANEEL. O capítulo XI da resolução trata dos procedimentos irregulares no sistema de distribuição e os artigos 130 e 132 definem como deve ser calculada e qual o período de recuperação da energia não medida (ANEEL, 2010). De forma resumida, o cálculo da energia recuperada deve seguir a aplicação de um dos métodos seguintes:

- Medição fiscalizadora (mínimo de 15 dias de medição);
- Fator de correção (se for detectado um erro de escala na medição);
- Medição anterior (considera-se os três maiores valores dos 12 meses anteriores a irregularidade);
- Valor da carga desviada (quando é possível fazer o levantamento de cargas);
- Medição posterior (considera-se o maior valor dos três meses seguintes a regularização).

Havendo a possibilidade de aplicação de mais de um método, deve-se optar pelo que aparece primeiro na lista apresentada. Com relação ao período de recuperação, tem-se o seguinte:

- Máximo de 36 meses, quando o início da irregularidade pode ser identificado;
- Máximo de seis meses, quando o início da irregularidade não pode ser identificado;
- Máximo de três meses, quando se trata de falha nos equipamentos de medição, sem que haja má fé do consumidor na irregularidade.

O período de recuperação é restrito ainda à data em que foi realizada a última inspeção de perda não técnica na unidade consumidora.

Assim, as inspeções devem realizadas de maneira a priorizar os consumidores que podem apresentar os maiores valores de energia recuperada, segundo os critérios regulatórios. Além disso, diferentemente das ações de prevenção que são consideradas investimentos e, portanto, são remuneradas na tarifa de energia elétrica, as ações de inspeção são consideradas custos operacionais e, portanto, devem ser realizadas da maneira mais eficiente possível.

Para aumentar a eficiência, é necessário direcionar as ações de inspeção para as unidades consumidoras que efetivamente possuem irregularidades no sistema de medição, pois, caso uma inspeção não identifique nenhum procedimento irregular, ela representará um desperdício de mão de obra e materiais para a distribuidora.

Cada inspeção tem um custo médio que pode variar de U$$ 30 a U$ 100 dependendo do país (MANAGEMENT SOLUTION, 2017) e (MANO, 2017). O custo operacional não é o único associado à inspeção, pois, existe um custo relacionado à imagem da companhia de distribuição, uma vez que um procedimento de inspeção pode constranger um consumidor regular. Por isso, é necessário atuar com a maior taxa de acerto possível nos procedimentos de inspeção.

Além disso, também é desejável que, em um cenário onde a quantidade de inspeções que podem ser realizadas é limitada, as inspeções sejam direcionadas prioritariamente para os consumidores com maior quantidade de energia não medida e que pode ser recuperada pelas distribuidoras. Pois, assim será garantido o maior retorno financeiro das ações.

Apesar dos aspectos destacados, a taxa de acerto praticada pelas concessionárias ao redor do mundo é baixa, ficando abaixo dos 15% na média dos registro disponíveis na bibliografia (ELLER, 2003), (COMETTI e VAREJÃO, 2007), (PENIN, 2008), (MONDERO *et al.*, 2012) e (MANAGEMENT SOLUTION, 2017). Tal fato representa um problema ainda não resolvido para as concessionárias de energia e está associado à complexidade do processo de identificação de fraudes, dado que em muitos casos o perfil de consumo de um consumidor fraudulento é igual ao de um consumidor regular. Além disso, a quantidade de consumidores irregulares é pequena quando comparada ao total de consumidores conectados à rede, fato que também dificulta a identificação dos consumidores irregulares.

O fato de ser um problema ainda não resolvido configura uma oportunidade de pesquisa para o desenvolvimento de novos métodos que possam contribuir com o aperfeiçoamento do processo de gestão da perda não técnica nas distribuidoras de energia elétrica. Nesse contexto, uma possibilidade é a aplicação de técnicas baseadas em *Advanced Analytics*, as quais são discutidas na próxima seção.

2.2 O Processo de *Advanced Analytics*

O termo *Advanced Analytics* (AA) tem sido utilizado para se referir a uma ampla variedade de técnicas de análises de dados, as quais têm como objetivo extrair informações que possam ser utilizadas para uma tomada de decisão de forma antecipada. Algumas dessas técnicas incluem mineração de dados, aprendizado de máquina, análise preditiva, análise de *clusters* e otimização. A utilização do termo se tornou comum nos últimos anos devido ao avanço no uso combinado de técnicas estatísticas e algoritmos computacionais para solução de problemas complexos envolvendo grandes volumes de dados. Portanto, o termo está mais relacionado a um processo do que a uma técnica específica (TECHOPEDIA, 2017).

Geralmente o processo de *Advanced Analytics* possui aplicação em áreas de negócio com disponibilidade de grandes volumes de dados históricos, como por exemplo, setor financeiro, seguradoras e empresas de vendas. Dentre os problemas que podem ser resolvidos com a aplicação do *Advanced Analytics*, podem ser citados:

- Definição do risco de uma operação financeira;
- Definição do perfil de consumo de um nicho social;
- Definição do preço ótimo de um produto; e
- Previsão da demanda em um setor de mercado.

Com a redução dos custos computacionais, cada vez mais empresas possuem bancos de dados com o histórico de suas operações, o que aumenta a possibilidade de aplicação do processo de *Advanced Analytics* para tomadas de decisões. Tal fato é corroborado pelo número crescente de publicações com a aplicação do conceito em diversas áreas, como Santesteban (2019), Zinsli (2020) e Heiney *et al.* (2021).

A principal vantagem de utilizar o *Advanced Analytics* é a possibilidade de combinar um grande conjunto de técnicas analíticas para realizar um monitoramento preditivo de alguma variável de interesse. Nesse contexto, o termo "*Advanced*" é utilizado no sentido de estar à frente (LIN, TABERNA e SIMON, 2018).

O *Advanced Analytics* pode ser implementado utilizando pacotes computacionais especialmente desenvolvidos para este fim. Dentre as ferramentas gratuitas disponíveis podem ser destacadas as seguintes:

- Anaconda®: distribuição das linguagens de programação Python e R;
- Weka®: plataforma de código aberto desenvolvida em Java;
- H2O.ai®: plataforma de código aberto com foco em aprendizagem de máquina;
- KNIME®: plataforma gratuita que contempla todas as etapas do *Advanced Analytics* e também possui integração com as demais plataformas citadas.

A partir do uso dos pacotes computacionais citados é possível aplicar diferentes técnicas para cada etapa do processo de forma estruturada. Algumas das principais técnicas utilizadas são descritas nas subseções seguintes.

2.2.1 Pré-processamento dos Dados

O pré-processamento dos dados é a primeira etapa do processo de *Advanced Analytics*. Esta etapa consiste em organizar os dados disponíveis em um formato que seja adequado à aplicação das técnicas de aprendizagem de máquina, além de garantir a qualidade e a consistência das informações (OLIVERI *et al.*, 2019).

O pré-processamento se faz necessário porque, no mundo real, existem vários problemas no armazenamento de grandes volumes de dados, como combinações impossíveis (ex.: sexo = masculino e grávida = sim), valores inconsistentes (ex.: peso = − 100 kg) e dados ausentes. Além disso, a forma como os dados são apresentados podem influenciar de maneira significativa o desempenho dos algoritmos computacionais. Por exemplo, é mais eficiente trabalhar com um conjunto de números entre zero e um, do que com um conjunto entre um e 1.000.000. Pois, no segundo caso os números de valor elevado podem ganhar uma importância maior no modelo, sem que tenham necessariamente uma importância estatística.

Os principais pontos da etapa de pré-processamento dos dados são descritos brevemente nas subseções seguintes.

2.2.1.1 Balanceamento

O balanceamento consiste em igualar ou aproximar o número de amostras em categorias distintas de um conjunto de dados. Por exemplo, em um problema em que se deseja classificar um e-mail como *spam* ou não-*spam* é desejável que a amostra de dados disponível para treinamento contenha um número próximo de e-mails do tipo *spam* e do tipo não-*spam*. Caso contrário, o algoritmo de classificação poderá ficar enviesado e classificar todas as amostras como sendo igual ao tipo predominante no conjunto de treinamento.

Caso a amostra de dados disponível seja naturalmente desbalanceada, é possível recorrer aos métodos conhecidos como *undersampling* e *oversampling*. O primeiro consiste em remover amostras da classe majoritária até que a quantidade fique próxima à da classe minoritária. O segundo consiste em duplicar ou criar novas amostras da classe minoritária até que a quantidade fique próxima à da classe majoritária.

A vantagem do *undersampling* é o menor custo computacional, já que resultará em um conjunto de dados menor que o original, e sua desvantagem é a possível perda de informação com o descarte de amostras. Por outro lado, a vantagem do *oversampling* é o fato de que nenhuma informação será descartada, e sua desvantagem é o maior esforço computacional, já que o conjunto de dados poderá ser muito maior que o original. De modo geral, o *oversampling* é mais utilizado que o *undersampling*, pois, a menos que o custo computacional seja crítico, a informação contida nos dados tem maior valor (LEMAÎTRE, NOGUEIRA e ARIDAS, 2017).

Um dos principais métodos de *oversampling* é o *Synthetic Minority Over-sampling Technique* (SMOTE). O método consiste no balanceamento da distribuição de dados entre as diferentes classes, a partir da adição de novas linhas de dados sintéticos (artificiais) à classe minoritária. De modo simplificado, são criadas novas linhas de dados sintéticos a partir da

extrapolação entre um objeto real de uma determinada classe e um de seus vizinhos mais próximos da mesma classe. Em seguida, é selecionado um ponto ao longo da linha entre esses dois objetos e determina-se os atributos do novo objeto com base neste ponto escolhido aleatoriamente. Maiores informações sobre o método SMOTE podem ser encontradas em Chawla *et al.* (2002).

2.2.1.2 *Missing Values*

Conforme comentado, é comum que em bancos de dados reais algumas informações estejam ausentes, problema conhecido como *missing values*. Para tratar os *missing values* existem basicamente duas opções: excluir os registros associados à ausência de dados ou preencher os valores ausentes com dados sintéticos obtidos a partir de uma regra pré-definida.

A desvantagem da exclusão de registros é a perda de informações possivelmente úteis. Por outro lado, o preenchimento de dados com valores fictícios também pode distorcer informações naturalmente associadas ao conjunto de dados. É o caso de quando a ausência de dados em si indica um comportamento útil para a resolução de um problema. Por exemplo, um cliente insatisfeito com um atendimento virtual tem o hábito de não preencher os campos de *feedback*, nesse casso a ausência de valores no campo indica uma característica sobre a satisfação do cliente.

De todo modo, o procedimento mais usual é a substituição dos *missing values* por algum valor pré-definido, o qual pode ser a média dos vizinhos, a moda de todo o conjunto ou ainda um valor fixo (MIRKES *et al.*, 2016).

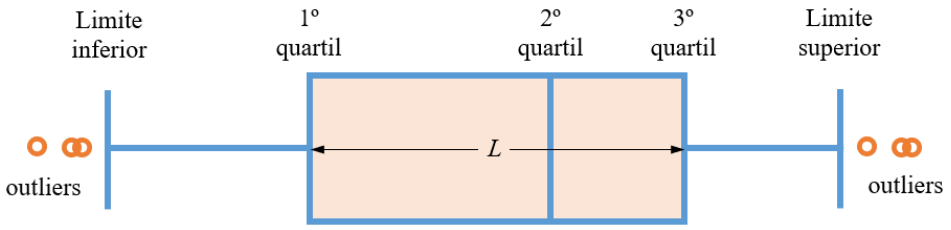
2.2.1.3 *Outliers*

O *outlier* é um elemento de um conjunto de dados que possui um grande afastamento dos demais elementos do grupo. Este elemento atípico geralmente prejudica a interpretação de testes estatísticos aplicados ao grupo. Um único *outlier*, por exemplo, pode distorcer completamente o valor da média de uma série de dados (SARABANDO, 2009).

A definição de um elemento como um *outlier* é subjetiva, ou seja, não segue uma regra fixa. Contudo, existem alguns métodos comumente aplicados para identificação dos *outliers*, como, por exemplo, o gráfico *Box-plot* e o *Z-scores*. Uma vez identificado como *outlier*, o elemento pode ser excluído ou substituído por outro valor a partir de uma regra pré-definida, assim como no caso dos *missing values*.

O gráfico de *Box-plot* consiste na representação dos elementos de uma amostra por meio de quartis, como ilustrado na Figura 9. Para definição dos *outliers*, calcula-se a distância L entre o 1º quartil e o 3º quartil do conjunto de dados. São considerados *outliers* os elementos maiores que o 3º quartil mais 1,5L ou menores que o 1º quartil menos 1,5L. O fator de 1,5 pode sofrer variações dependendo da aplicação (MOROCO, 2003).

Figura 9 – Ilustração dos elementos de um gráfico *Box-plot*.



Fonte: Adaptado de NeuroMat (2017).

Já no método *Z-scores*, é calculado o desvio padrão (σ) da média do conjunto de dados ($\bar{\sigma}$) e a condição para ser considerado um *outlier* é apresentada na Equação (1).

$$Se\ \ X > (\bar{\sigma} + K\sigma)\ \ ou\ \ X < (\bar{\sigma} - K\sigma)\ \ \rightarrow X\ \text{é}\ outilier. \qquad (1)$$

De acordo com a Equação (1), é considerado *outlier* o elemento que for maior que a média mais K vezes o desvio padrão ou menor que a média menos K vezes o desvio padrão. Em que K é um fator que pode variar conforme o tamanho da amostra (SARABANDO, 2009).

2.2.1.4 Normalização

Em um conjunto de dados diferentes variáveis numéricas podem possuir ordens de grandeza muito distintas, por exemplo, a variável “idade” variando entre 0 e 90, e a variável “saldo bancário” variando entre 1.000 e 100.000. Nesse caso, a diferença de escala entre as diferentes variáveis pode enviesar o comportamento de alguns métodos de aprendizagem de máquina. Por isso, é recomendável que as variáveis numéricas em um conjunto de dados sejam normalizadas (KUMAR e AZAD, 2017).

Normalizar um conjunto de valores significa representá-los em uma escala padronizada, como de 0 a 1, por exemplo. Existem diferentes formas de se normalizar um conjunto de dados e uma das mais usuais é a normalização de base unitária que pode ser obtida com a aplicação da Equação (2).

$$Y' = \frac{Y - Y_{mín}}{Y_{máx} - Y_{mín}}. \quad (2)$$

Em que Y' é o valor normalizado da variável, Y é o valor original, $Y_{mín}$ e $Y_{máx}$ são os valores mínimos e máximos das variáveis no conjunto de dados. Na normalização de base unitária todos os valore Y' ficam situados no intervalo entre 0 e 1.

2.2.1.5 Categorização de Variáveis

Variável categórica é uma variável que se apresenta em escala nominal, cujas categorias identificam a classe à qual um elemento pertence, como, por exemplo, o gênero que é uma variável categórica com duas classes (masculino e feminino).

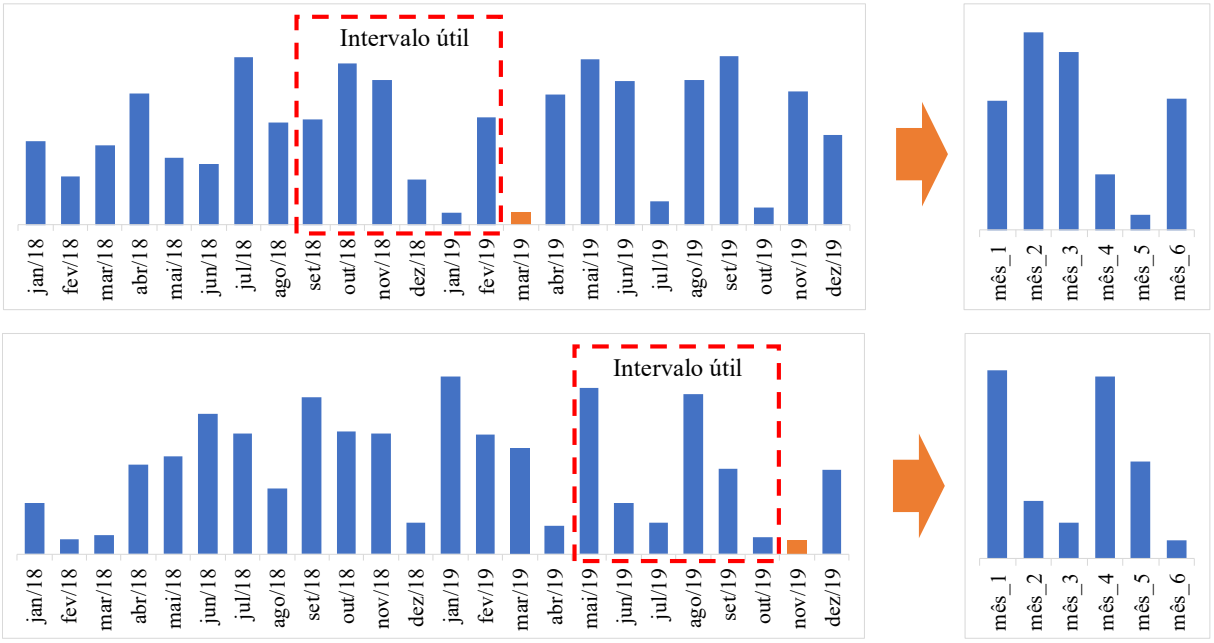
Existem situações em que é útil representar de forma categórica uma variável originalmente numérica. Como por exemplo, ao invés de representar a idade de uma pessoa em anos de vida, representá-la por categorias como jovem, adulto e idoso. Nesse caso, cada categoria agrupa um determinado intervalo numérico referente à quantidade de anos de vida. O processo é conhecido como categorização de variáveis e pode tornar um conjunto de dados mais explicativo, nos casos em que pequenas variações no valor numérico não representam mudanças relevantes do ponto de vista prático (NIST, 2013).

2.2.1.6 Janelamento

O janelamento de uma série de dados é o processo de seleção de uma janela de tempo móvel, para a qual o conjunto de dados será observado. O processo pode ser útil em situações em que se deseja analisar dados anteriores a um determinado evento, como o ilustrado na Figura 10, em que o evento de interesse é a ocorrência de um valor mínimo.

No exemplo da Figura 10 o objetivo é identificar as causas da ocorrência do valor mínimo na série de dados, o qual está destacado na cor laranja. Nesse caso não faz sentido observar dados posteriores a ocorrência do valor mínimo, apenas os dados anteriores ao evento. Assim, aplica-se o processo de janelamento para criar uma nova série de dados contendo apenas os valores anteriores ao evento. Como pode ser observado na figura, é selecionado o intervalo de seis meses anteriores a ocorrência do valor mínimo e, como resultado, é possível obter novas series de dados contendo apenas os dados anteriores à ocorrência do evento de interesse, as quais são ilustradas no lado direito da Figura 10.

Figura 10 – Ilustração do processo de janelamento em uma série de dados cujo objetivo é capturar os dados referentes ao período de seis meses anteriores ao mínimo da série.



Fonte: Autoria própria.

A partir dos tratamentos descritos nos itens anteriores é possível obter um conjunto de dados consistente e adequado à utilização nas demais etapas do *Advanced Analytics*. Por esta razão, o pré-processamento é considerado uma das etapas mais importantes de todo o processo e, além disso, é responsável por aproximadamente 60% de todo o tempo gasto na sua execução.

2.2.2 *Feature Engineering*

Feature Engineering é o processo de criação de novas variáveis a partir de variáveis originalmente disponíveis em um banco de dados. O objetivo é que as variáveis derivadas sejam mais explicativas que as originais do ponto de vista estatístico. Por exemplo, supondo um problema em que se deseja prever a população de um local na próxima década e a única variável disponível é a quantidade de habitantes registrada a cada ano nos últimos 50 anos. Neste caso, pode ser mais eficiente utilizar a derivada da série de dados populacionais do que a série original, pois, a derivada de uma série indica sua taxa de crescimento (ZABOKRTSKY, 2016).

A ideia no processo de criação de *Feature Engineering* é criar o máximo de informação possível a partir dos dados disponíveis. Para isso podem ser aplicadas transformações de domínio como a transformada de Fourier e a transformada de Wavelet, ou ainda calculadas métricas estatísticas como média, moda, mediana, curtose, assimetria, desvio padrão, intervalo interquartil, dentre outras possibilidades.

Como pode ser notado no parágrafo anterior, o processo de *Feature Engineering* não possui uma metodologia bem definida, por isso, ele exige criatividade e conhecimento da área na qual o problema está inserido. Como se trata de um processo livre, é possível que sejam criadas variáveis com algum nível de redundância entre si, contudo, é possível tratar este ponto com o processo de seleção de variáveis que será discutido a seguir.

2.2.3 Seleção de Variáveis

Para garantir o melhor desempenho de um modelo preditivo é necessário determinar, dentre todas as variáveis disponíveis, o subconjunto de variáveis independentes que melhor explique a variável resposta. Para tanto, é necessário realizar um processo de seleção de variáveis, cujo objetivo principal é diminuir a variância da estimativa da variável resposta e o custo da coleta de informações. O objetivo é alcançado por meio de um conjunto com menor número de variáveis e maior explicabilidade que o original. Assim, por produzir uma representação mais compacta do conceito a ser aprendido, o processo de seleção de variáveis melhora o desempenho de modelos preditivos (URBANOWICZ *et al.*, 2017).

Diferentes métodos podem ser utilizados no processo de seleção de variáveis, dentre os quais destaca-se a Inferência Causal. Pois, com ela é possível inferir a presença e a magnitude das relações de causa e efeito, ao invés de apenas definir a correlação como faz a maioria dos métodos tradicionais de análise estatística. Na subseção seguinte são fornecidos maiores detalhes sobre o conceito de Inferência Causal.

2.2.3.1 Inferência Causal

A Inferência Causal pode ser definida como o processo de inferir a presença e a magnitude das relações de causa e efeito a partir dos dados (REIS, 2020). Contudo, não é trivial afirmar que uma variável é a causa de um evento, uma vez que um conjunto de condições, incluindo condições subjetivas, deve ser atendido. Assim, existem critérios que são sugeridos para tratar a Inferência Causal. Um exemplo amplamente utilizado na bibliografia são os critérios de causalidade de Hill, que propõe nove condições para estabelecer uma relação causal entre um fator e um efeito (HILL, 1965). A lista dos critérios é apresentada na Tabela 2.

Tabela 2 – Critérios de Hill para causalidade.

Critério	Descrição
1 – Força da associação	Uma associação forte tem maior probabilidade de ser causal do que uma associação fraca. Já que associações fracas podem ser resultantes de viés de seleção, confusão ou acaso.
2 – Consistência	Se a associação se observa repetidamente em diferentes populações e circunstâncias, tem maior probabilidade de ser causal.
3 – Especificidade	A causa apenas conduzirá a um efeito e não a múltiplos efeitos. Esse é um critério questionável, já que algumas exposições conferem risco para vários desfechos.
4 – Temporalidade	A causa precede o efeito.
5 – Relação dose-resposta	O aumento da exposição à causa aumenta o efeito.
6 – Plausabilidade	A associação tem uma explicação plausível.
7 – Coerência	A assunção de causalidade deverá estar ligada a outras observações.
8 – Evidência experimental	A relação causa e efeito é observável experimentalmente.
9 – Analogia	O observado é análogo ao que se sabe sobre outras situações semelhantes.

Fonte: Hill (1965).

Na prática, raramente é possível provar todos os nove critérios de Hill, bem como determinar as causas necessárias ou suficientes de um evento, como a ocorrência de PNT em um consumidor. Além disso, há autores que afirmam que as proposições causais não podem ser totalmente provadas, pois, encontram falhas ou limitações práticas em todas as abordagens (ROTHMAN e GREENLAND, 2004).

No entanto, a análise de Inferência Causal pode ser utilizada para sugerir quais variáveis possuem maior probabilidade de serem causas de um evento. Em outras palavras, mesmo que não seja possível determinar a causa da perda não técnica, é possível inferir quais são as condições em que é mais provável que ela ocorra. Menezes (2001) afirma que os critérios da força da associação e temporalidade são indispensáveis à causalidade, e que os demais critérios podem contribuir para sua inferência, mas não necessariamente determinam-na.

Nesse contexto, a análise de Inferência Causal consiste em medir a força da associação entre as variáveis conhecidas (possíveis causas) e a presença de PNT nos consumidores (efeito), além de garantir a condição da temporalidade e avaliar as demais condições subjetivas na Tabela 2, a fim de descartar a possibilidade da associação ser uma mera coincidência.

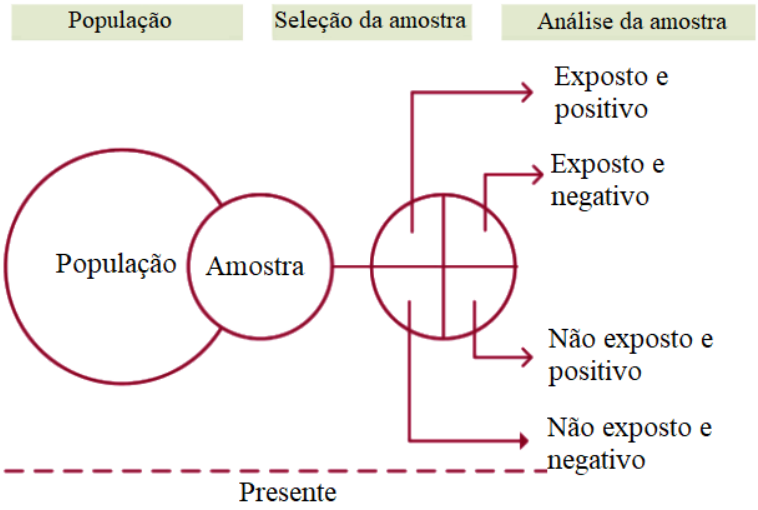
Para definir quais medidas de associação devem ser utilizadas, é necessário definir inicialmente em que tipo de estudo a Inferência Causal será aplicada. Existem basicamente dois tipos de estudos de Inferência Causal, o experimental e o observacional. No estudo experimental uma amostra de indivíduos é obtida aleatoriamente. Em seguida, divide-se aleatoriamente a amostra em dois grupos, um dos grupos é exposto à possível causa (grupo-

caso) e outro não (grupo-controle). Por fim, compara-se a ocorrência do evento de interesse nos dois grupos (desfecho) (GUILLOUX *et al.*, 2019).

O estudo experimental é o que mais se aproxima do ideal para a análise de Inferência Causal, pois, a aleatoriedade na seleção das amostras garante que os possíveis fatores de confusão estão balanceados nos dois grupos. Assim, para qualquer intervenção realizada, o desfecho será devido à intervenção e não a outros fatores. Contudo, esse tipo de estudo costuma ter um custo elevado e, em alguns casos, sua execução é inviável, como no caso do problema da perda não técnica de energia (GUILLOUX *et al.*, 2019)

Por isso, também podem ser realizados estudos observacionais, em que uma amostra de indivíduos é observada de forma passiva, sem a interferência do observador. A amostra é obtida por oportunidade e não necessariamente de maneira aleatória. Posteriormente, separam-se os indivíduos que estão expostos à possível causa dos que não estão e observa-se a prevalência do evento de interesse em cada grupo. O tipo mais simples de estudo observacional é o transversal, em que as informações relacionadas aos indivíduos da amostra são extraídas no mesmo momento da seleção. Uma ilustração do estudo transversal é apresentada na Figura 11.

Figura 11 – Ilustração do estudo observacional transversal.

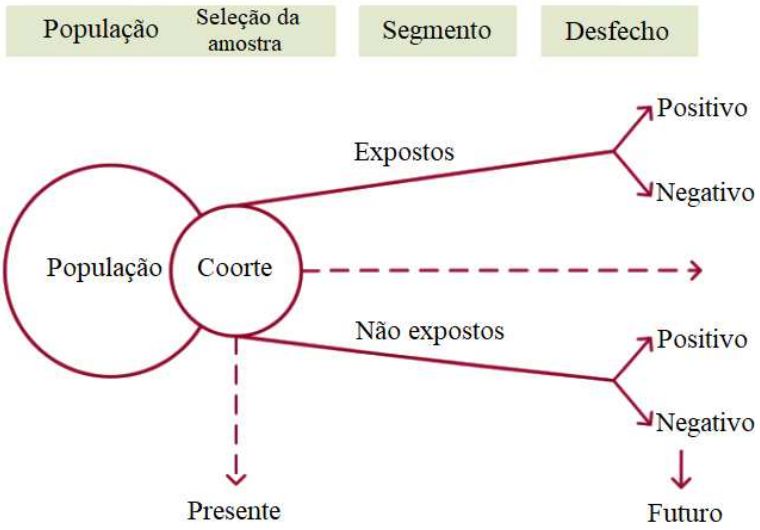


Fonte: Adaptado de Guilloux *et al.* (2019).

Como observado na Figura 11, no estudo transversal é possível determinar apenas a prevalência de determinado evento em cada grupo da amostra. Sendo necessário ainda a inferência de outros critérios da Tabela 2 para demonstração de causalidade.

Outro tipo de estudo observacional é o de coorte, em que se realiza o acompanhamento de uma amostra por determinado período tempo para verificar o desfecho no futuro, conforme ilustração da Figura 12.

Figura 12 – Ilustração do estudo observacional de coorte.

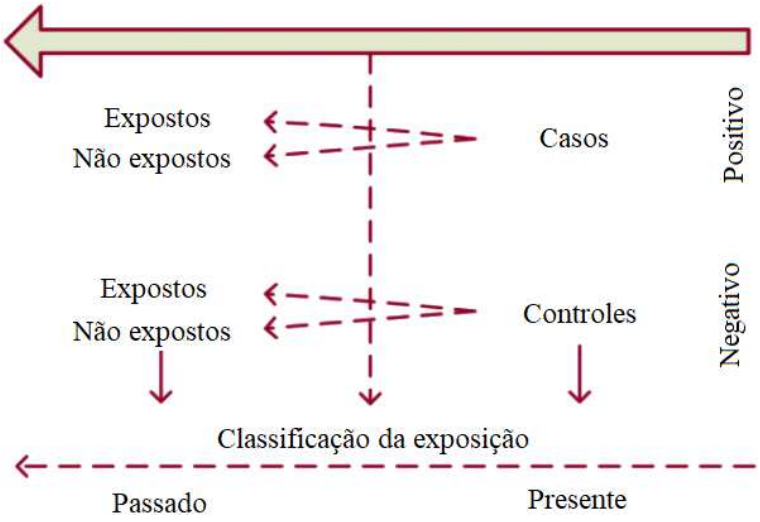


Fonte: Adaptado de Fletcher e Fletcher (2006).

Como pode ser visto na Figura 12, no estudo de coorte a amostra inicial não possui desfecho conhecido, pois ele ocorrerá com o passar do tempo. Neste caso, somente ao final do período de acompanhamento é calculada a incidência do evento de interesse em cada grupo.

Por fim, o estudo observacional também pode ser do tipo caso-controle. Nesse caso, selecionam-se indivíduos que desenvolveram e que não desenvolveram um desfecho de interesse, posteriormente, procura-se avaliar a exposição passada dos indivíduos a fatores que se acredita serem causas do desfecho, conforme ilustrado na Figura 13.

Figura 13 – Ilustração do estudo observacional de caso-controle.



Fonte: Adaptado de Fletcher e Fletcher (2006).

No caso do problema da perda não técnica, a abordagem utilizada para Inferência Causal é do tipo observacional caso-controle. A partir do histórico de inspeções realizadas por uma

concessionária é possível identificar consumidores em que a presença ou a ausência da PNT é conhecida em um dado momento. Esses consumidores são selecionados como amostra da população total, aqueles que apresentaram registro de PNT no momento da inspeção são os casos, e os que não apresentaram PNT são os controles. Em seguida, informações retroativas a data da inspeção são obtidas para serem utilizadas na inferência causal.

Como afirmado anteriormente, para cada tipo de estudo, diferentes medidas de associação devem ser utilizadas. Na Tabela 3 é apresentado um resumo sobre as medidas de associação para os estudos observacionais.

Tabela 3 – Medidas de associação para estudos observacionais.

Tipo de Estudo	**Medida de associação**		
	Tipo	**Medida**	**Equação**
Transversal	Prevalência	Razão de Prevalência (RP)	$RP = \frac{a}{a+b} / \frac{c}{c+d}$ (3)
Coorte	Incidência	Risco Atribuível (RA)	$RA = \frac{a}{a+b} - \frac{c}{c+d}$ (4)
Caso-controle	Chance	*Odds Ratio* (OR)	$OR = \frac{a}{c} / \frac{b}{d}$ (5)

Fonte: Guilloux *et al.* (2019).

As equações expostas na Tabela 3 se referem aos valores da tabela de distribuição conjunta apresentada na Figura 14.

Figura 14 – Exemplo de tabela de distribuição conjunta de um estudo observacional.

		Efeito	
		Sim	Não
Exposição	Sim	**a**	**b**
	Não	**c**	**d**

Fonte: Autoria própria.

Na Figura 14, os indivíduos da amostra observada são distribuídos de acordo com a sua exposição a possível causa e o seu desfecho em relação ao efeito de interesse. Assim, a representa o total de indivíduos que foram expostos e desenvolveram o efeito; b o total de indivíduos que foram expostos e não desenvolveram o efeito; c o total de indivíduos que não foram expostos e desenvolveram o efeito, e por fim, d representa o total de indivíduos que não foram expostos e não desenvolveram o efeito.

A tabela de distribuição conjunta da Figura 14 só pode ser utilizada para avaliar a exposição a variáveis qualitativas. Portanto, as medidas apresentadas na Tabela 3 também se

referem a variáveis qualitativas. Quando se tratam de variáveis quantitativas, é necessário utilizar medidas de tendência central, como a média ou a mediana, ou medidas de dispersão, como a variância e o desvio padrão.

Contudo, é necessário se atentar para o fato de que apenas a informação da média de uma amostra, por exemplo, é uma informação pobre do ponto de vista estatístico, uma vez que o valor é obtido a partir de uma amostra e, portanto, não reflete necessariamente o valor verdadeiro da população total. Portanto, além de informar o valor de uma medida de associação ou de tendência central para uma determinada amostra, também é necessário informar qual a probabilidade do valor obtido ser o valor verdadeiro para a população total.

O procedimento para isso é a utilização de um coeficiente de confiança. Ao invés de calcular um único valor, deve-se calcular uma faixa de valores associada ao coeficiente de confiança, o qual informará a probabilidade do valor verdadeiro estar contido na faixa de valores obtida.

A confiança estatística de uma amostra está relacionada à variância dos dados nessa amostra e na população total, o que por sua vez está relacionada à distribuição de probabilidade dos dados. A distribuição de probabilidade real dificilmente é conhecida, contudo, quando o conjunto amostral possui um número suficientemente grande de amostras, a distribuição de probabilidade se aproxima da distribuição normal (ROSS, 2009). Considerando a distribuição normal, o intervalo para o valor da média de uma amostra com um coeficiente de confiança de 95% pode ser calculado como apresentado na Equação (6).

$$\bar{x} - 1{,}96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96\frac{\sigma}{\sqrt{n}}, \tag{6}$$

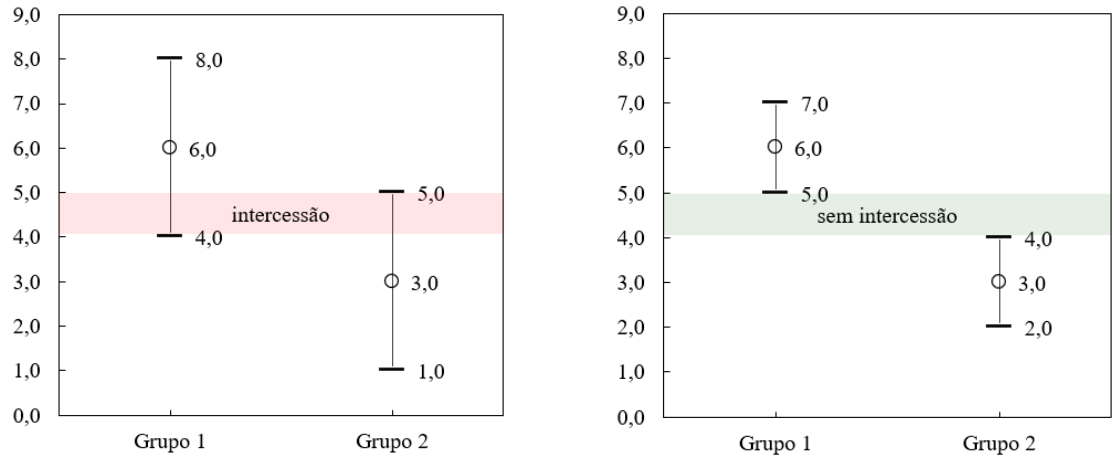
em que $\bar{x}$ é a média da amostra, σ é o desvio padrão da média da amostra, n é o número de indivíduos na amostra e μ é o valor verdadeiro da média para a população total.

O intervalo definido na Equação (6) é, portanto, uma média confiável para a amostra. No caso prático do problema da perda não técnica, se houver diferença entre a média confiável de uma variável no grupo de consumidores com PNT e no grupo sem PNT, então, isso significa que a variável tem alguma associação com a ocorrência ou com a ausência de PNT, e pode ser utilizada para caracterizar a perda não técnica no conjunto.

Contudo, se a média confiável tem algum intervalo de intercessão entre os dois grupos, isso significa que não há associação significativa entre a variável e a presença de PNT e,

portanto, ela não pode ser utilizada para caracterizar a perda não técnica no conjunto. Uma ilustração dessa ideia é fornecida na Figura 15.

Figura 15 – Ilustração das diferenças de médias confiáveis. Em (a) não existe diferença significativa entre a média do Grupo 1 e do Grupo 2, em (b) existe diferença significativa entre as médias do Grupo 1 e do Grupo 2.



Fonte: Autoria própria.

Nos gráficos da Figura 24 são apresentados os intervalos de confiança da média de uma variável em dois grupos diferentes. São indicados os limites inferior e superior, bem como o valor médio central do intervalo. No gráfico (a), apesar do valor central ser distinto, há uma intercessão entre o intervalo da média para o Grupo 1 e para o Grupo 2, portanto, não há diferença significativa para essa variável em ambos os grupos. Por outro lado, no gráfico (b) não há interseção, o que indica que há diferença significativa para a variável entre os dois grupos, ou seja, a variável pode ser utilizada para caracterizar os grupos.

O conceito de intervalo de confiança também deve ser aplicado às medidas de associação apresentadas na Tabela 3, uma vez que elas são obtidas a partir de amostras da população total. No caso do *Odds Ratio* (OR), o seu valor pode ser expresso como na Equação (7), para um coeficiente de confiança de 95% e considerando-se uma distribuição normal de probabilidade.

$$e^{\ln(OR)-1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \leq OR^{*} \leq e^{\ln(OR)+1{,}96\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}, \tag{7}$$

em que OR é o *Odds Ratio* da amostra de dados, OR^{*} é o valor real do *Odds Ratio* para a população total, e a, b, c, d são as quantidades da tabela de distribuição conjunta da Figura 14. O intervalo de valores do *Odds Ratio* deve ser interpretado como da seguinte maneira:

- Se o intervalo de $OR^{*} > 1 \rightarrow$ A exposição é um fator de risco para o efeito;

- Se o intervalo de $OR^* < 1 \rightarrow$ A exposição é um fator de proteção para o efeito;
- Se o intervalo de $OR^* \supset 1 \rightarrow$ Não existe associação entre a exposição e o efeito.

As medidas de associação discutidas até o momento são úteis para indicar quais variáveis podem ser utilizadas para caracterizar a ocorrência de perda não técnica no sistema de distribuição. Contudo, diferentes variáveis podem conter algum grau de redundância na informação que representam. Por isso, no processo de seleção de variáveis faz-se necessário a eliminação de variáveis redundantes, com o objetivo de diminuir o custo computacional e os possíveis fatores de confusão em uma amostra de dados.

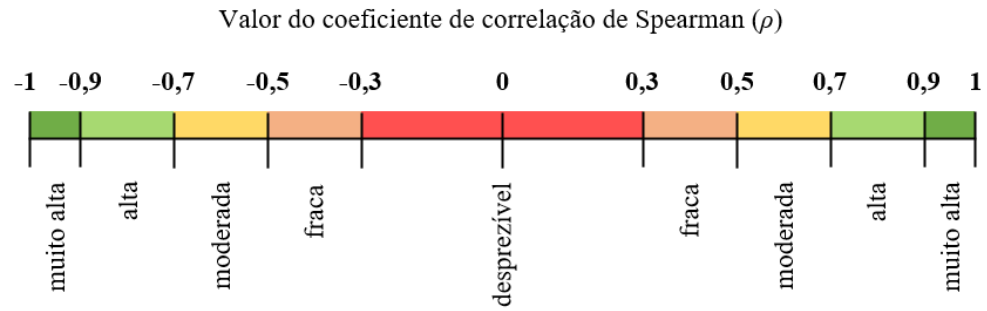
Para determinar a redundância de informações entre duas variáveis, pode ser utilizado o coeficiente de correlação de Spearman. Este coeficiente é uma medida não paramétrica de correlação que avalia quão bem a relação entre duas variáveis pode ser descrita por uma função monotônica (CHEN e POPOVICH, 2002). O seu valor pode ser determinado por ρ conforme a Equação (8).

$$\rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rg_X}\sigma_{rg_Y}}, \tag{8}$$

em que rg_X e rg_Y são as ordens de classificações das variáveis X e Y; σ_{rg_X} e σ_{rg_Y} são os desvios padrão das variáveis de rg_X e rg_Y.

Quando as variáveis são do tipo qualitativas, o coeficiente de correlação de Spearman é baseado no número de pares concordantes e discordantes e, neste caso, o coeficiente é conhecido como φ. O valor do coeficiente sempre estará entre −1 e +1 e pode ser interpretado como na Figura 27.

Figura 16 – Nível de redundância de informação entre variáveis de acordo com o valor do coeficiente de Spearman.



Fonte: Adapitado de Chen e Popovich (2002).

Como pode ser observado na Figura 27, $\rho = 0$ indica nenhuma redundância entre as variáveis, e $\rho = -1$ ou $\rho = 1$ indica uma redundância completa entre as variáveis, ou seja, uma variável explica completamente a outra, ou ainda, tem 100% de explicabilidade para outra.

O termo explicabilidade, pode ser empregado para se referir ao quanto da variância de um evento pode ser explicada pelos dados de variáveis conhecidas. Segundo Fieller, Hartley e Pearson (1957), a explicabilidade ϵ pode ser expressa em termos do coeficiente de correlação de Spearman ρ, como na Equação (9).

$$\epsilon = \rho^2. \tag{9}$$

Quando $\epsilon = 1$ (ou 100%) para duas variáveis X e Y, significa que o comportamento de X pode ser perfeitamente predito por Y. Além disso, somar os valores de explicabilidade de um conjunto de variáveis independentes em relação a uma variável dependente, como a ocorrência de PNT, indicará o quanto da variância do evento pode ser explicada pelas variáveis independentes conhecidas.

Diante do exposto na presente subseção, pode-se concluir que a realização da análise de Inferência Causal na etapa de seleção de variáveis é muito útil, pois, isso garante que as variáveis que serão utilizadas como entrada nos modelos preditivos possuirão algum grau de interferência na ocorrência da perda não técnica de energia. Tal consideração não poderia ser feita somente com os modelos de aprendizagem de máquina, que apesar de conseguirem atribuir pesos maiores as variáveis mais relevantes, não conseguem diferenciar correlação de causalidade.

2.2.4 Clusterização

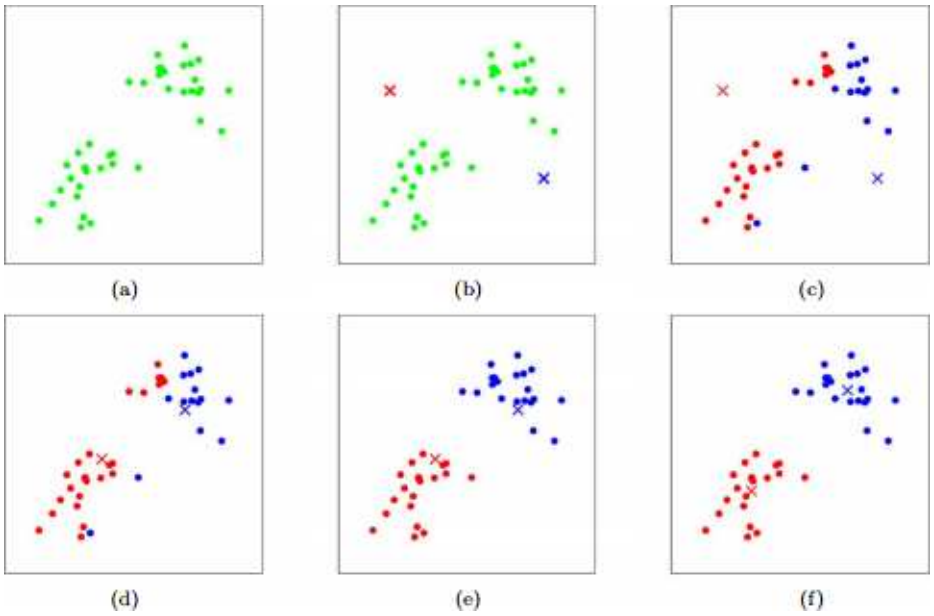
Clusterizar um conjunto de dados consiste em separar as amostras em grupos distintos, de modo que os indivíduos em um mesmo grupo compartilhem semelhanças entre si mais do que compartilham com indivíduos de outros grupos. O tipo de semelhança compartilhada e a quantidade de grupos dependem de cada conjunto de dados e da natureza do problema em questão (SCHUBERT, 2017).

A clusterização também pode ser definida como um método de aprendizagem de máquina não supervisionado, pois não depende de nenhuma amostra da variável resposta para treinamento. Na verdade, nos problemas de clusterização não existe uma variável resposta, o

que se busca é identificar os diferentes padrões existentes no conjunto de dados por meio de algoritmos de agrupamento, dentre os quais destaca-se o *k-means*.

O algoritmo *k-means* é o mais popular algoritmo de clusterização e já foi considerado um dos 10 mais utilizados na área de ciência de dados (WU *et al.*, 2008). Foi originalmente proposto em 1982 por Stuart Lloyd como uma técnica para modulação por código de pulso (LLOYD, 1982) e tem como objetivo dividir um conjunto de dados em k grupos distintos, de modo que os indivíduos de um mesmo grupo possuam características semelhantes entre si (PIECH e NG, 2013). O funcionamento do algoritmo *k-means* é baseado na distância euclidiana entre cada indivíduo do conjunto e um determinado centroide, cujo conceito está ilustrado na Figura 17.

Figura 17 – Ilustração do funcionamento do algoritmo *k-means*: (a) conjunto de dados original, (b) escolha aleatória dos centroides iniciais, (c-f) ilustração das duas primeiras iterações do algoritmo.



Fonte: Piech e Ng (2013).

Na Figura 17 (a) observa-se o conjunto de dados original que se deseja agrupar. Na Figura 17 (b) são propostos, de forma aleatória, dois centroides iniciais que definirão os dois grupos. Na Figura 17 (c) cada indivíduo é atribuído ao grupo cuja distância euclidiana para o centroide seja a menor. Na Figura 17 (d) a média de cada grupo formado se torna o novo centroide. Na Figura 17 (e) é feita a redistribuição dos indivíduos entre os dois grupos com base na menor distância euclidiana para os novos centroides. Na Figura 17 (f) novamente a média de cada grupo formado se torna o novo centroide. O processo iterativo descrito se repete até que não haja mudança de grupo entre os indivíduos ou até que um determinado critério de parada seja alcançado.

Apesar da sua popularidade, o algoritmo *k-means* possui alguns inconvenientes, como o fato de que o usuário deve informar previamente o número k de centroides. Nesse contexto, é útil a utilização do algoritmo *x-means*, que é uma versão modificada do *k-means*.

O *x-means* determina automaticamente o melhor número de centroides para um determinado conjunto de dados. O algoritmo começa com um conjunto mínimo de centroides e, em seguida, iterativamente aumenta essa quantidade verificando se o uso de mais centroides faz sentido de acordo com os dados (PELLEG e MOORE, 2000).

A determinação do número de centroides é baseada no *Bayesian Information Criteria* (BIC). O BIC tem como pressuposto a existência de um "modelo verdadeiro" que descreve a relação entre a variável dependente e as diversas variáveis explicativas nos diversos modelos sob seleção. Assim o critério é definido como a estatística que maximiza a probabilidade de se identificar o modelo verdadeiro dentre os avaliados (SCHWARZ, 1978). O valor do critério BIC para um determinado modelo é obtido pela Equação (10).

$$BIC = -\log(P) + \frac{l}{2}\log(k), \tag{10}$$

em que P é a função de probabilidade, k o número de centroides e l o número de parâmetros livres do modelo. O modelo que apresentar o menor BIC é considerado o modelo verdadeiro.

Dessa forma, o algoritmo *x-means* pode ser utilizado para separação dos consumidores em uma base de dados com o objetivo de agrupar aqueles com características semelhantes, e verificar se alguma dessas características está associada à ocorrência de perda não técnica, por exemplo.

2.2.5 Aprendizagem de Máquina

Os termos Aprendizagem de Máquina ou *Machine Learning* (ML) são utilizados para se referir a um campo da Inteligência Artificial (IA) dedicado ao desenvolvimento de algoritmos para análise automática de dados. Os algoritmos se baseiam na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana. Em outras palavras, a partir do aprendizado com exemplos históricos um modelo computacional pode executar uma tarefa sem que haja instruções explicitamente programadas para isso (SIMON, 2013).

O conceito não é recente, ele foi introduzido em 1956 na Conferência de Dartmouth (*Dartmouth Summer Research Project on Artificial Intelligence*). Contudo, foi nos últimos 10 anos que a aplicação da Aprendizagem de Máquina cresceu de forma exponencial para solução dos mais diversos tipos de problema. O fato pode ser atribuído ao aumento da capacidade de processamento computacional e ao seu barateamento, além do aumento da disponibilidade de grandes volumes de dados e da evolução dos algoritmos, que são cada vez mais robustos (MAHRI, ROSTAMIZADEH e TALWALKAR, 2018).

Dentre as aplicações da Aprendizagem de Máquina destacam-se os problemas de classificação e regressão. No problema de classificação, o objetivo é atribuir um rótulo a um indivíduo no conjunto de dados, de modo a classificá-lo em uma das possíveis classes existente. O problema de detecção de fraudes é um caso típico de classificação. O setor financeiro foi o primeiro a explorar os algoritmos de classificação com objetivo de detecção de fraudes, sendo possível encontrar propostas de soluções desde o final da década de 90 (GHOSH e REILLY, 1994).

Já nos problemas de regressão, o objetivo é prever o valor de uma variável de interesse a partir do conhecimento de outras variáveis. A variável de interesse pode ser contínua ou discreta. Quando a variável é do tipo discreta, o problema de regressão torna-se um problema de classificação. Portanto, a classificação pode ser vista como um caso particular da regressão.

Diferentes algoritmos podem ser utilizados em problemas de classificação e regressão. Alguns podem ser utilizados para ambos os tipos, enquanto outros só se aplicam aos casos de classificação. Na Tabela 4 são listados alguns dos principais algoritmos de Aprendizagem de Máquina disponíveis na bibliografia.

Tabela 4 – Principais algoritmos de Aprendizagem de Máquina.

Tipo	**Algoritmo**	**Sigla**	**Aplicação**
Linear generalizado	Regressão Logística	LR	Classificação
Baseado em árvores	Árvore de Decisão	DT	Classificação e Regressão
Bayesiano	Naïve Bayes	NB	Classificação
Rede neural	*Multi-Layer Perceptron*	MLP	Classificação e Regressão
Espacial	*Support Vector Machine*	SVM	Classificação
Fuzzy	Classificador *Fuzzy*	FR	Classificação
Ensemble	*Gradient Boosted Tree*	GBT	Classificação e Regressão
	eXtreme Gradient Boosted Tree	XGBT	Classificação e Regressão
	Bagging Tree	BT	Classificação e Regressão
	Random Forest	RF	Classificação e Regressão
	Rotation Forest	RTF	Classificação e Regressão

Fonte: Autoria Própria.

Como pode ser visto na Tabela 4, os algoritmos listados contemplam uma ampla variedade de tipos. Os algoritmos do tipo *ensemble* baseiam-se na execução em cadeia de modelos de algoritmos mais simples. Os vários modelos são combinados para produzir resultados aprimorados. Dessa forma, os algoritmos do tipo *ensemble* tendem a produzir soluções mais precisas do que um único modelo produziria individualmente (DEMIR, 2016).

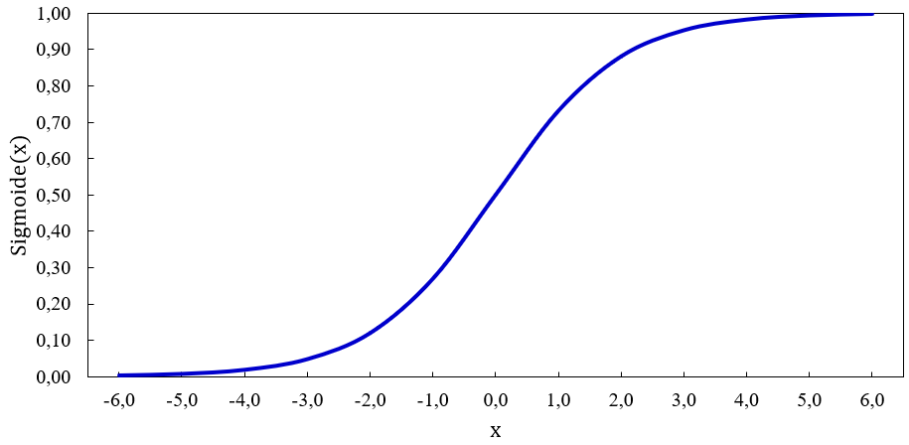
Nas subseções seguintes serão apresentadas as principais características dos algoritmos listados na Tabela 4, bem como, serão fornecidas referências clássicas nas quais podem ser encontrados maiores detalhes dos algoritmos.

2.2.5.1 Regressão Logística

A Regressão Logística (LR) é um dos modelos mais básicos para classificação binária de dados. É baseado em um método estatístico cujo objetivo é encontrar uma equação que prevê o resultado para uma variável dependente binária, a partir de uma ou mais variáveis independentes. O algoritmo foi nomeado devido à função utilizada no núcleo do método, a função logística, também chamada de função sigmoide (BROENLEE, 2016). A função logística é dada na Equação (11) e representada graficamente na Figura 18.

$$sigmoide(x) = \frac{1}{(1 + e^{-x})}. \tag{11}$$

Figura 18 – Representação gráfica da função sigmoide.



Fonte: Autoria própria.

Como observado na Figura 18, a função sigmoide recebe como argumento qualquer número real e o mapeia entre zero e um. No algoritmo de Regressão Logística os valores das

variáveis independentes de entrada são combinados linearmente por meio da função sigmoide ponderada por coeficientes (β) para se determinar a probabilidade do valor da variável dependente de saída y pertencer a classe padrão. A Equação (12) representa a Regressão Logística para o caso de uma única variável independente x.

$$y = \frac{e^{(\beta_0+\beta_1 \cdot x)}}{(1 + e^{(\beta_0+\beta_1 \cdot x)})}. \tag{12}$$

Uma vez que a variável dependente y é binária, ela só pode assumir os valores zero ou um. Assim, o resultado da Equação (12) pode ser interpretado como a probabilidade da variável dependente pertencer à classe padrão. Portanto, a solução do algoritmo de Regressão Linear consiste na determinação dos coeficientes β, o que é realizado de forma iterativa com métodos de minimização do erro como o Gradiente Descendente ou métodos Quase-Newton (BROENLEE, 2016).

O algoritmo de Regressão Logística destaca-se por sua simplicidade, facilidade de interpretação dos resultados e baixo custo computacional. Contudo, apresenta como principal desvantagem o desempenho limitado em problemas de domínio complexo. Informações mais detalhadas do algoritmo LR podem ser encontradas em Broenlee (2016).

2.2.5.2 Árvore de Decisão

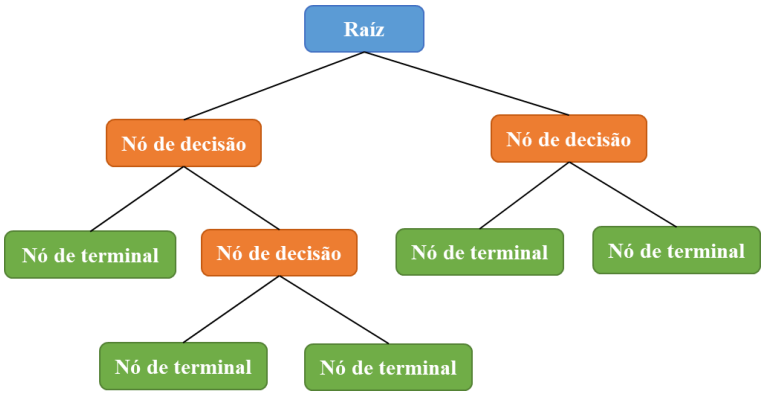
O algoritmo da Árvore de Decisão (DT) baseia-se em uma estrutura semelhante a um fluxograma, como apresentado na Figura 19. Cada nó de decisão (ou nó interno) representa um teste de um atributo, cada ramificação representa o resultado do teste e cada nó terminal (ou folha) representa uma classe. Os caminhos da raiz à folha representam as regras de classificação (ROKACH e MAIMON, 2008).

Na Figura 19, para que sejam possíveis as divisões nos nós de decisão, é necessário que as variáveis do problema sejam categóricas. No caso de serem números reais, elas devem ser discretizadas previamente. A construção de uma árvore de decisão para problemas reais envolve duas tarefas básicas: decidir quais variáveis escolher como raiz e quais condições usar para dividir os nós.

A raiz da árvore é definida como a variável que melhor divide a classes existente, essa mesma lógica é aplicada para os nós de decisão subsequentes. Portanto, o critério de divisão dos ramos da árvore constitui a etapa mais importante do algoritmo. Os critérios de qualidade

mais utilizados para divisão das árvores são o índice Gini, *Chi-Square* e o *Information Gain* (SANTANA, 2017).

Figura 19 – Estrutura básica de uma árvore de decisão.



Fonte: Adaptado de Chauhan (2020).

As Árvores de Decisão constituem um dos algoritmos mais populares de Aprendizagem de Máquina e suas principais vantagens são a simplicidade, facilidade de interpretação dos resultados e o baixo custo computacional. Por outro lado, as principais desvantagens são a instabilidade e a possibilidade de *overfitting*. A instabilidade pode ocorrer quando pequenas variações nos dados de entrada geram resultados significativamente diferentes na saída. O *overfitting* ocorre quando a árvore se torna muito especifica para os dados do treinamento, apresentando baixa capacidade de generalização.

Informações mais detalhadas do algoritmo DT podem ser encontradas em Rokach e Maimon (2008).

2.2.5.3 Naïve Bayes

Naïve Bayes (NB) é um classificador probabilístico simples baseado na aplicação do teorema de Bayes com a suposição de independência entre as variáveis de entrada. No início do algoritmo também é assumido que todas as variáveis possuem a mesma importância, ou seja, recebem o mesmo peso para o cálculo da saída (KHURANA, 2020).

O teorema de Bayes descreve a probabilidade de um evento ocorrer dado a probabilidade de ocorrência de um outro evento que já acorreu. O teorema é descrito pela Equação (13).

$$P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)}, \tag{13}$$

em que Y e X são os eventos; $P(Y)$ e $P(X)$ são as probabilidades marginais ou priori, eles representam as probabilidades de os eventos ocorrerem sem o conhecimento de nenhuma condição prévia; $P(Y|X)$ e $P(Y|X)$ são as probabilidades condicionais ou posteriori, eles representam a probabilidade do primeiro evento ocorrer dado que o segundo evento é verdadeiro. Para a verificação do teorema, $P(X)$ deve ser diferente de zero.

A suposição de Naïve pode ser expressa matematicamente pela Equação (14).

$$P(Y|X) = P(Y) \cdot P(X). \tag{14}$$

A partir da suposição na Equação (14) e assumindo que Y é a variável dependente em um problema de classificação, enquanto X é o vetor de variáveis independentes $\{x_1, \dots, x_n\}$, a Equação (13) pode ser reescrita como a Equação (15).

$$P(Y|x_1, \dots, x_n) = \frac{P(Y) \cdot \prod_{i=1}^{n} P(x_i|Y)}{P(x_1, \dots, x_n)}. \tag{15}$$

Como o denominador na Equação (15) permanece constante para uma dada entrada, a probabilidade condicional é proporcional ao termo no numerador, como escrito na Equação (16).

$$P(Y|x_1, \dots, x_n) \propto P(Y) \cdot \prod\prod_{i=1}^{n} P(x_i|Y). \tag{16}$$

A regra de classificação do algoritmo Naïve Bayes é baseada na Equação (16) e consiste em encontrar a probabilidade do conjunto de entradas $\{x_1, \dots, x_n\}$ para todos as classes possíveis da variável de saída Y e, posteriormente escolhe-se a saída associada com probabilidade máxima. Isso pode ser expresso matematicamente como na Equação (17).

$$\hat{Y} = \arg\max_{Y} P(Y) \cdot \prod\prod_{i=1}^{n} P(x_i|Y). \tag{17}$$

Portanto, a execução do algoritmo consiste no cálculo das probabilidades $P(Y)$ e $P(x_i|Y)$, o que pode ser feito com o uso da técnica *Maximum a Posteriori* (MAP). Nota-se a partir da demonstração exposta que a execução do algoritmo Naïve Bayes é relativamente

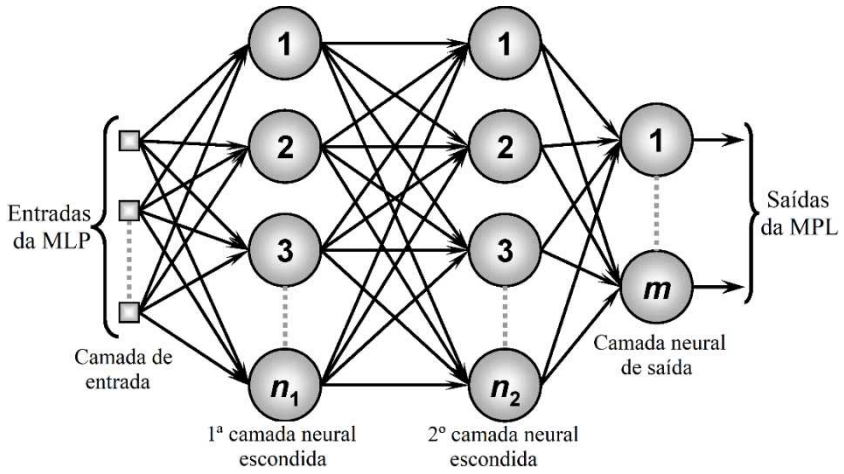
simples. Isso só é possível devido à suposição de Naïve de que as variáveis de entrada são independentes entre si.

Tal suposição dificilmente se verifica em aplicações reais, contudo, é útil para a solução prática do algoritmo. Por isso, o algoritmo NB pode ser muito rápido em comparação com métodos mais sofisticados, uma vez que cada distribuição de probabilidade pode ser estimada independentemente como uma distribuição unidimensional. A velocidade de convergência e a simplicidade são as principais vantagens do algoritmo Naïve Bayes, enquanto que sua principal desvantagem é o desempenho limitado em domínios complexos. Informações mais detalhadas do algoritmo NB podem ser encontradas em Rish (2001).

2.2.5.4 *Multi-Layer Perceptron*

Redes Neurais Artificiais (RNA) são algoritmos de aprendizagem inspirados nas redes neurais biológicas que constituem os cérebros humanos. Esses algoritmos aprendem por meio de exemplos e melhoram progressivamente o desempenho para fazer tarefas como predição, aproximação e classificação. O tipo de RNA mais utilizado é o *Multi-Layer Perceptron* (MPL), o qual é composto por uma camada de entrada, uma ou mais camadas intermediárias e uma camada de saída. As camadas intermediárias também são chamadas de camadas escondidas ou camadas ocultas, tal denominação ocorre porque não é possível prever a saída desejada nas camadas intermediárias (SILVA, SPATTI e FLAUZINO, 2010). A estrutura básica de uma MLP é apresentada na Figura 20.

Figura 20 – Estrutura de uma Rede Neural *Multi-Layer Perceptron*.



Fonte: Moreira (2018).

Na Figura 20 cada nó representa um neurônio artificial, exceto para a camada de entrada. Cada neurônio possui um valor associado conhecido como *bias*, utilizado para controlar o grau

de liberdade da rede, e cada ramo que conecta dois neurônios possui um peso, o qual é conhecido como peso sináptico em referência ao neurônio biológico. Os pesos são utilizados para ponderar as entradas em cada camada, ou seja, quanto maior o peso para uma entrada mais influência ela exercerá na rede.

Matematicamente, uma MLP simples com apenas uma entrada, uma camada oculta e um neurônio, pode ser definida como na Equação (18).

$$f(x) = G\left(b^{(2)} + w^{(2)}S\left(b^{(1)} + w^{(1)}x\right)\right), \tag{18}$$

em que x é o vetor de entrada e $f(x)$ é a saída da MLP, $b^{(1)}$ e $b^{(2)}$ são os *bias* dos neurônios na camada intermediária e na camada de saída respectivamente, $w^{(1)}$ e $w^{(2)}$ são os pesos sinápticos na camada intermediária e na camada de saída respectivamente, e G e S são funções de ativação.

A função de ativação S é responsável por definir a ativação de saída do neurônio em termos do seu nível de ativação interna, em outras palavras é uma função que mapeia as entradas ponderadas para a saída de cada neurônio. Escolhas típicas para S são a tangente hiperbólica apresentada na Equação (19), e a função sigmoide na Equação (11).

$$tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}. \tag{19}$$

A função de ativação G é utilizada para forçar a saída da MLP a representar a probabilidade dos dados pertencerem a classe definida, para isso, G é definida como a função *softmax* apresentada na Equação (20).

$$softmax(\boldsymbol{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}}. \tag{20}$$

Para $i = 1, \dots, K$ e $\boldsymbol{x} = (x_1, \dots, x_K)$.

A função *softmax* recebe como entrada um vetor de K números reais e o normaliza em uma distribuição de probabilidade que consiste em K probabilidades proporcionais às exponenciais dos números de entrada. Ou seja, antes da aplicação da *softmax*, alguns componentes do vetor podem ser negativos ou maiores que um; e a soma dos mesmos pode não

ser igual a 1, mas depois de aplicar a *softmax*, cada componente estará no intervalo (0,1) e a soma dos componentes será 1, por isso, podem ser interpretados como probabilidades.

Dessa forma, a construção de uma MLP para modelagem de um dado problema consiste em determinar os valores de $\{w^{(1)}, b^{(1)}, w^{(2)}, b^{(2)}\}$, o que envolve três etapas básicas:

- Treinamento: Etapa de aprendizagem. Um conjunto de dados é apresentado sucessivamente na entrada da rede e sua saída é comparada a um conjunto de referência (variável resposta). A cada iteração os parâmetros da rede são ajustados com o objetivo de minimizar a diferença entre a saída e a referência.
- Validação: Etapa de supervisão da aprendizagem. É executada paralelamente ao treinamento para garantir a generalidade da rede, ou seja, evitar que a rede fique enviesada para o conjunto de treinamento (*overfitting*).
- Teste: Etapa de verificação da aprendizagem. Consiste em utilizar um conjunto de dados distinto daqueles utilizados no treinamento e validação para averiguar o nível de conhecimento adquirido pela rede ao final do processo.

Na etapa de aprendizagem uma entrada é aplicada à MLP e seu efeito é propagado pela rede, camada a camada, com os pesos da rede iguais a um palpite inicial. Ao final, o valor da saída é comparado ao valor da entrada e uma função de erro é calculado. Para execução da etapa de aprendizagem é utilizado o algoritmo Levenberg–Marquardt, o qual consiste na combinação dos métodos Gradiente Descendente e Gauss-Newton. O algoritmo inicia com o método do Gradiente Descendente para reduzir o erro da estimativa inicial, na sequência o algoritmo converge para o método de Gauss-Newton, o qual utiliza informações da derivada segunda da função erro, o que torna a convergência do problema muito rápida (BISHOP, 1995).

Dado o exposto, é possível afirmar que a MLP é uma das técnicas de AI mais populares por sua capacidade de resolver problemas estocástico, o que geralmente permite soluções aproximadas para problemas complexos. Dentre as principais aplicações da MLP estão problemas que envolvem classificação e previsão de dados (SILVA, SPATTI e FLAUZINO, 2010).

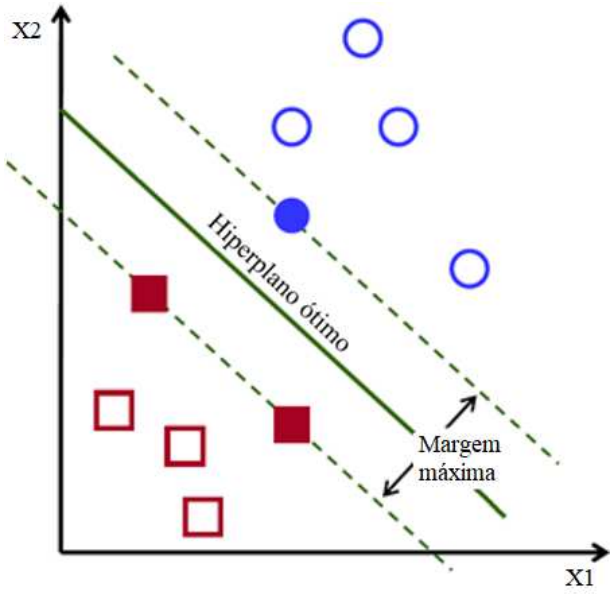
Dentre as desvantagens do uso da MLP está a falta de justificativa para a resposta, pois, o algoritmo pode ser visto como uma "caixa preta", no qual não se sabe exatamente por que o modelo chegou a um determinado resultado. Além disso, a etapa de treinamento pode apresentar esforço computacional elevado a medida em que se aumenta o número de camadas ocultas e a quantidade de neurônios em cada camada (ALPAYDIN, 2010). Informações mais detalhadas do algoritmo MLP podem ser encontradas em Bishop (1995).

2.2.5.5 *Support Vector Machine*

O *Support Vector Machine* (SVM) é um algoritmo de aprendizagem de máquina focado na classificação binária de dados que foi proposto por Boser, Guyon e Vapnik (1992). O funcionamento do SVM é baseado na representação espacial dos dados de entrada, a partir da qual é estabelecido um limite de separação entre duas classificações distintas. O limite que define a separação de classes é denominado de hiperplano. A dimensão do hiperplano depende da quantidade de variáveis de entrada e da natureza do problema. Para um problema linear, por exemplo, se houver apenas duas variáveis de entrada, o hiperplano será uma reta, se houver três variáveis de entrada, será um plano e assim sucessivamente (SARADHI, KAMIK e MITRA, 2005).

Portanto, pode-se afirmar que o objetivo do SVM é encontrar um hiperplano em um espaço N-dimensional que classifique distintamente os pontos de dados, em que N é o número de variáveis no problema. Porém, existem muitos hiperplanos possíveis que podem ser escolhidos, nesse caso, deve-se encontrar o hiperplano que maximize a distância entre os pontos de dados das duas classes distintas, esta distância recebe o nome de margem. Na Figura 21 está ilustrado o conceito de classificação do SVM.

Figura 21 – Ilustração de uma classificação linear realizada pelo *Support Vector Machine*.



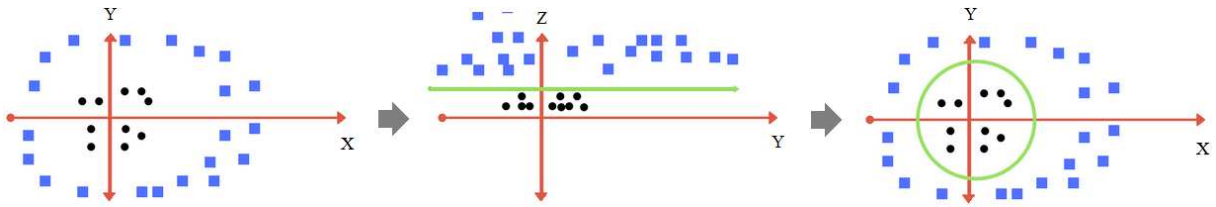
Fonte: Gandhi (2018).

O exemplo na Figura 21 refere-se a um conjunto de 10 indivíduos que possuem duas variáveis características associadas, X1 e X2. A variável resposta é a classe dos indivíduos, a qual é representada pelas cores azul e vermelho. A linha verde representa o hiperplano ótimo que separa as duas classes. Os indivíduos destacados, que ficam nas extremidades internas dos

grupos, são chamados de vetores de suporte, pois, a partir deles são definidas as linhas de margem da separação. Se um vetor de suporte for excluído da base de dados o hiperplano ótimo irá mudar, por isso, esses indivíduos têm importância central no SVM.

O exemplo da Figura 21 é didático, contudo, os problemas reais têm natureza não linear na sua maioria. Por isso, o hiperplano terá uma forma mais complexa do que uma reta. Para lidar com esse tipo de situação, o SVM utiliza um conjunto de funções para transformar o domínio dos dados, de modo a diminuir a complexidade na determinação do hiperplano. As funções são conhecidas como *kernel*, e um exemplo de transformação de domínio é apresentada na Figura 22.

Figura 22 – Transformação de domínio de variáveis por meio de uma função *kernel*: domínio de dados original no eixo X-Y → domínio transformado no eixo X-Y-Z → transformação inversa para o domínio X-Y.



Fonte: Patel (2017).

No exemplo da Figura 22, o domínio original das variáveis é representado no plano X-Y (gráfico à esquerda), neste domínio não é possível definir o hiperplano de forma linear, por isso, a função *kernel* dada na Equação (21) é utilizada para realizar uma transformação e adicionar mais uma dimensão por meio do eixo Z.

$$Z = X^2 + Y^2. \tag{21}$$

No novo domínio, ao plotar o plano Y-Z (gráfico ao centro da Figura 22) uma separação clara dos dados é visível e pode ser representada por meio da reta verde. Ao aplicar a transformação inversa, obtém-se novamente o plano original e a reta de separação é mapeada como um limite circular (gráfico à direita da Figura 22).

O princípio de funcionamento apresentado faz com que o SVM seja considerado um dos melhores métodos para problemas de classificação binária, pois, o algoritmo funciona bem mesmo em domínios complexos. Por isso, o SVM possui grande abrangência de aplicações em diversas áreas, como finanças, biologia, medicina, logística, entre outras. As principais vantagens da técnica são: bom desempenho de generalização, simplicidade matemática e possibilidade de interpretação geométrica dos resultados. Como desvantagem está o fato de que pode apresentar esforço computacional muito elevado em algumas situações, pois o esforço

varia com o cubo da quantidade de dados analisados (PORTELA *et al.*, 2017). Informações mais detalhadas do algoritmo SVM podem ser encontradas em Boser, Guyon e Vapnik (1992).

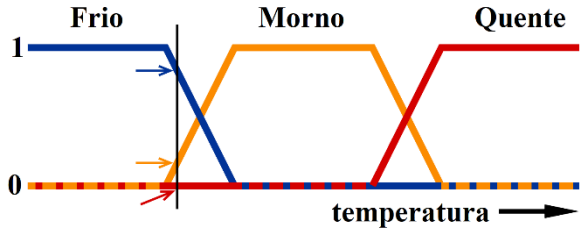
2.2.5.6 Classificador *Fuzzy*

O algoritmo Classificador *Fuzzy* (FR) é baseado na lógica *fuzzy*, também conhecida como lógica difusa, que é um tipo de operação de rotulagem suave em que uma variável desconhecida é atribuída a um ou mais estados com graus variados, usando conjuntos e regras difusas (*fuzzy*). O principal diferencial do método é que ele tem por objetivo modelar modos de raciocínio aproximados ao invés de precisos. Na lógica clássica as proposições só podem ser verdadeiras ou falsas, contudo, na lógica *fuzzy* podem existir valores intermediários entre o verdadeiro e o falso. Nesse caso, as proposições passam a ter um grau de pertinência (ZADEH, 2008).

O grau de pertinência é um valor que varia no intervalo de zero a um e é utilizado para classificar os elementos do conjunto entre as classes possíveis. Um grau de pertinência igual a um indica que o elemento pertence completamente a uma classe, e um grau de pertinência igual a zero indica que ele não pertence completamente a classe. Qualquer valor intermediário indica que o elemento pertence simultaneamente a mais de uma classe, com graus de pertinência que somam um.

A classificação *Fuzzy* pode ser melhor compreendida a partir da análise do exemplo ilustrado na Figura 23, em que o valor da temperatura de um sistema é mapeado a partir de três funções de pertinência, “Frio”, “Morno” e “Quente”.

Figura 23 – Ilustração de uma classificação *Fuzzy* para a temperatura em três estados.



Fonte: Adaptado de Shaw (2004).

Na Figura 23 cada estado pode assumir o valor de pertinência entre zero e um. No ponto indicado pela linha vertical na figura, o valor de pertinência é zero para a classe “Quente”, 0,2 para a classe “Morno” e 0,8 para a classe “Frio”. Tais valores indicam que a temperatura registrada com certeza não é quente, mas, pode ser classificada como relativamente fria ou

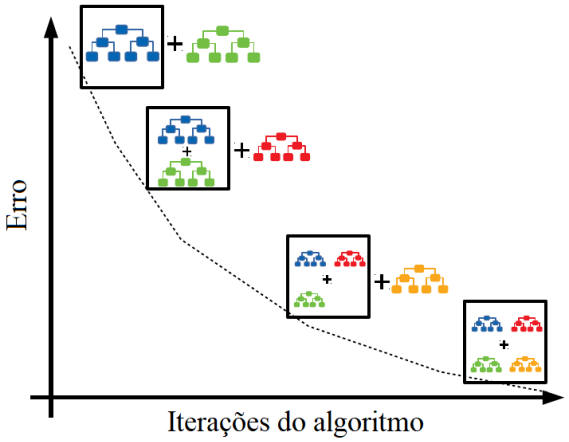
ainda levemente morno. A classificação final no algoritmo é obtida por um conjunto de regras do tipo "*if-then*" definidas pelo usuário.

O Classificador *Fuzzy* é especialmente vantajoso em problemas nos quais existe alguma incerteza sobre o estado dos indivíduos do conjunto. Outra vantagem do método é a interpretabilidade dos resultados, já que as etapas e as instruções lógicas que levam à previsão da classe são rastreáveis e compreensíveis. Por outro lado, a principal desvantagens do método é que o desempenho do modelo geralmente é limitado pelo conhecimento do especialista na configuração dos parâmetros, como o número de funções de pertinência de cada variável e o número de regras. Informações mais detalhadas do algoritmo FR podem ser encontradas em Berthold (2003).

2.2.5.7 *Gradient Boosted Tree*

O *Gradient Boosted Trees* (GBT) é um algoritmo proposto por Friedman (2001) que possui aplicação em problemas de classificação. O algoritmo é baseado na montagem sucessiva de árvores de decisão simples. A ideia é construir árvores de decisão sucessivas, de tal modo que exemplos classificados incorretamente por árvores anteriores sejam melhores classificados nas árvores seguintes. Ou seja, a árvore atual depende das anteriores tendo maior foco no erro das últimas, dessa forma, exemplos classificados incorretamente anteriormente são ponderados com maior peso nas iterações seguintes (HASTIE, TIBSHIRANI e FRIEDMAN, 2009). A ideia geral do algoritmo GBT é ilustrada na Figura 24.

Figura 24 – Ilustração do funcionamento algoritmo *Gradient Boosted Trees*.



Fonte: Adapitado de Pal (2020).

Na Figura 24 é ilustrado o funcionamento do algoritmo GBT por meio da montagem sucessiva de árvores de decisão. O procedimento tem o objetivo de minimizar uma função de erro, mais especificamente o erro médio quadrático, dado na Equação (22).

$$Erro = \sum \left(y_i - y_i^p\right)^2. \tag{22}$$

Em que y_i é o i-ésimo exemplo na base de treinamento e y_i^p é a previsão do modelo para o i-ésimo exemplo.

Para minimizar a função de erro o modelo subsequente deve possuir um gradiente decrescente, o que é representado matematicamente na Equação (23), a qual resulta na Equação (24).

$$y_i^{p+1} = y_i^p - \alpha \times \frac{\delta \sum \left(y_i - y_i^p\right)^2}{\delta y_i^p}, \tag{23}$$

$$y_i^{p+1} = y_i^p - 2\alpha \sum \left(y_i - y_i^p\right), \tag{24}$$

em que α representa o deslocamento de cada passo do processo iterativo e é conhecido como a taxa de aprendizagem do algoritmo. Portanto, o algoritmo GBT basicamente atualiza as previsões de forma que a soma dos erros seja próxima de zero e os valores previstos estejam suficientemente próximos dos valores reais.

O GBT tem se destacado como um método de aprendizagem robusto e é indicado para detecção de anomalias, na qual os dados costumam ser altamente desequilibrados, como transações com cartão de crédito e segurança cibernética (HASTIE, TIBSHIRANI e FRIEDMAN, 2009). Estas aplicações se assemelham a tarefa de detecção de perda não-técnica de energia no sistema de distribuição, que é o foco deste trabalho.

Dentre as desvantagens do método, podem ser citados o custo computacional, o risco de *overfitting* e a baixa explicabilidade do resultado final. O *overfitting* ocorre quando o modelo fica enviesado para uma dada amostra de dados utilizada na etapa de treinamento. A desvantagem pode ser minimizada pela escolha adequada da taxa de aprendizagem e o número de nós das árvores de decisão (HASTIE, TIBSHIRANI e FRIEDMAN, 2009). Informações mais detalhadas do algoritmo GBT podem ser encontradas em Friedman (2001).

2.2.5.8 eXtreme Gradient Boosted Tree

O *eXtreme Gradient Boosted Tree* (XGBT) implementa algoritmos de aprendizado de máquina sob a estrutura *Gradient Boosting* descrita na seção anterior, com base no pacote computacional XGBoost. O XGBoost é uma biblioteca otimizada, projetada para ser eficiente, flexível e portátil. Trata-se de uma solução de código aberto distribuída em vários ambientes de desenvolvimento distintos como C++, Java, Python e R.

Do ponto de vista conceitual o funcionamento do algoritmo XGBT é similar ao algoritmo GBT, contudo ele foi desenvolvido para ser computacionalmente mais eficiente em aplicações com grandes volumes de dados. Algumas das diferenças em relação ao algoritmo GBT são relacionadas a seguir.

- O algoritmo suporta computação distribuída;
- Possui um parâmetro de randomização para o treinamento, o que o torna mais resistente ao *overfitting*;
- Explora a existência de matrizes esparsas para reduzir o custo computacional.

A desvantagem do algoritmo XGBT em relação ao GBT fica por conta do maior número de parâmetros que precisam configurados pelo usuário na inicialização do algoritmo. Informações mais detalhadas do algoritmo XGBT podem ser encontradas em XGBoost Documentation (2020).

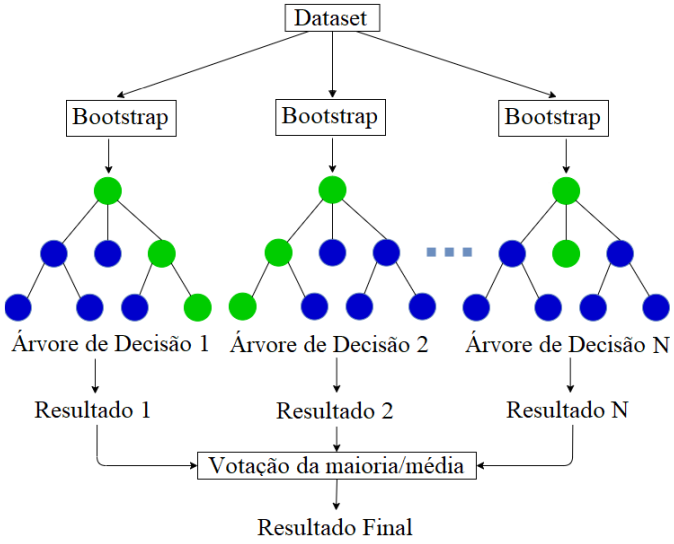
2.2.5.9 Bagging Tree

O algoritmo *Bagging Tree* (BT) baseia-se em um princípio simples de funcionamento, o qual consiste em combinar a saída de várias Árvores de Decisão criadas aleatoriamente para gerar a saída final (BREIMAN, 1996). Como será visto adiante, o termo *Bagging* é um acrônimo para *bootstrap aggregation*, que se refere aos procedimentos que estão na base do funcionamento do algoritmo, o qual está ilustrado na Figura 25.

Como ilustrado na Figura 25, inicialmente o algoritmo separa o conjunto de dados disponível para treinamento em vários subconjuntos aleatoriamente, cada subconjunto possui uma quantidade de amostras que pode ser menor ou igual a quantidade total, portanto, podem existir amostras repetidas nos subconjuntos, esse processo de amostragem é conhecido como *bootstrap*. Posteriormente, os subconjuntos são utilizados para treinar Árvores de Decisão simples, como descrito na subseção 2.2.5.2. Por fim, os resultados das classificações individuais de cada árvore são agrupados (*aggregation*) para se verificar qual foi a classe que se repetiu mais vezes, esta classe é escolhida como resultado final do método *Bagging Tree*. Caso o

problema seja uma regressão, então o resultado do algoritmo será a média dos resultados individuais de cada Árvore de Decisão.

Figura 25 – Ilustração do funcionamento do algoritmo *Bagging Tree*.



Fonte: Adaptado de Sharma (2020).

A ideia do algoritmo BT é que um grande número de modelos não correlacionados operando como um comitê de votação, tende a apresentar um desempenho melhor do que qualquer um dos modelos individuais constituintes. Em outras palavras, as árvores protegem umas às outras de seus erros individuais (ZOGHBI, 2020).

Dentre as principais vantagens do algoritmo *Bagging Tree* estão a quantidade reduzida de parâmetros que precisam ser configuradas e a maior resistência ao *overfitting*. Por outro lado, a principal desvantagem é que um grande número de árvores pode tornar o algoritmo muito lento e ineficaz para previsões em tempo real. Em geral, o algoritmo é rápido de treinar, mas lento para criar previsões depois de treinado. Informações mais detalhadas do algoritmo BT podem ser encontradas em Breiman (1996).

2.2.5.10 Random Forest

O algoritmo *Random Forest* (RF) é uma extensão do *Bagging Tree*, com uma etapa extra no procedimento de *bootstrap*. No caso do RF, além de obter subconjuntos aleatórios de dados, também é feita uma seleção aleatória de variáveis para cada subconjunto, ou seja, em vez de utilizar todas variáveis para o treinamento de todas as árvores, o algoritmo utiliza combinações diferentes de variáveis em cada árvore (NAGPAL, 2017).

A vantagem do algoritmo *Random Forest* é que ele aumenta a independência das árvores constituintes, já que elas são treinadas com combinações diferentes de variáveis. Isso tende a

melhorar a acurácia geral do método e a torná-lo mais resistente ao *overftting* (DONGES, 2020). Como desvantagem pode-se citar o fato de que o algoritmo é menos performático em problemas de regressão, dado que o valor final é baseado na média das previsões das árvores constituintes, não há garantia de que o erro da previsão será reduzido. Essa desvantagem é compartilhada pelos algoritmos baseados no método *Bagging*, nesse caso os algoritmos baseados no método *Boosting*, como os descritos nas subseções 2.2.5.7 e 2.2.5.8, tendem a se sair melhor. Informações mais detalhadas do algoritmo RF podem ser encontradas em Breiman (2001).

2.2.5.11 Rotation Forest

O algoritmo *Rotation Forest* (RTF) foi proposto por Rodriguez, Kuncheva e Alonso (2006) e é baseado na extração de características do conjunto de dados por meio do método *Principal Component Analisys* (PCA). O PCA é um procedimento matemático que utiliza uma transformação ortogonal para converter um conjunto de variáveis correlacionadas em um conjunto de variáveis linearmente independentes, as quais são chamadas de componentes principais. O número de componentes principais é sempre menor ou igual ao número de variáveis originais, e os primeiros componentes possuem a maior parte da informação dos dados, por isso, o PCA também é utilizado para a redução de variáveis (ABDI e WILLIAMS, 2010).

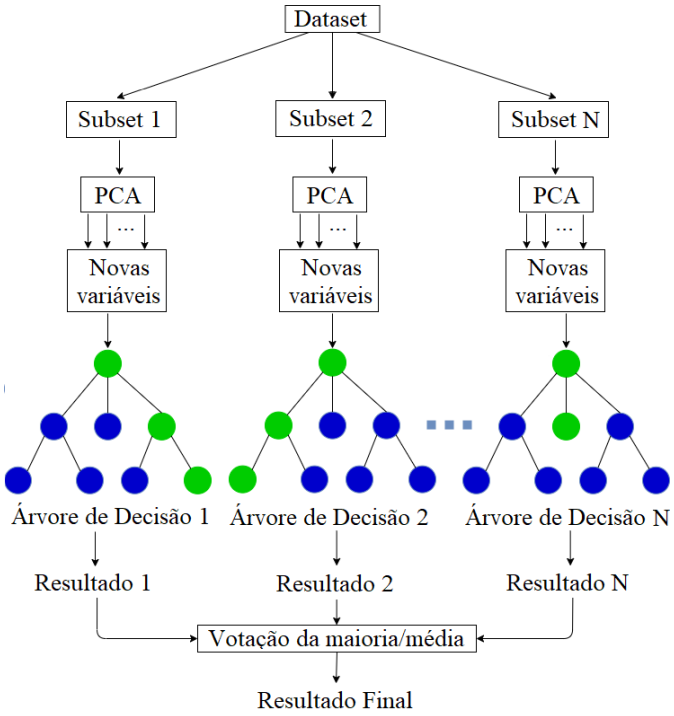
O funcionamento do *Rotation Forest* é semelhante ao do algoritmo *Random Forest* apresentado na subseção 2.2.5.10, porém, com algumas diferenças importantes, as quais podem ser melhor compreendidas a partir da ilustração apresentada na Figura 26.

Na Figura 26 inicialmente o conjunto de variáveis é separado em k subconjuntos aleatoriamente, em que cada subjconjunto possui N/k variáveis, sendo N a quantidade total de variáveis no *dataset*. Posteriormente o PCA é aplicado em cada subconjunto e os componentes principais resultantes são organizados em uma matriz chamada Matriz de Rotação (por isso o termo *Rotation*). Em seguida, a Matriz de Rotação é utilizada para o treinamento de Árvores de Decisão simples. Por fim, os resultados individuais de cada árvore são agrupados e utilizados para definir o resultado final do algoritmo, assim como ocorre no *Random Forest*.

O algoritmo *Rotation Forest* tem se destacado por oferecer um desempenho superior ao de algoritmos mais populares, especialmente em domínios complexos (BAGNALL, FLYNN, *et al.*, 2018). Uma das razões para o bom desempenho do RTF é a diversidade promovida pela etapa de extração de variáveis com o PCA para cada árvore de decisão. A desvantagem do algoritmo fica por conta do maior esforço computacional e a maior complexidade para

implementação. Informações mais detalhadas do algoritmo RTF podem ser encontradas em Rodriguez, Kuncheva e Alonso (2006).

Figura 26 – Ilustração do funcionamento do algoritmo *Rotation Forest*.



Fonte: Adaptado de Sharma (2020).

2.2.6 Otimização de Hiperparâmetros

Um hiperparâmetro é uma variável cujo valor é utilizado para controlar o processo de treinamento em modelos de Aprendizagem de Máquina. Em qualquer algoritmo, esses parâmetros necessitam ser inicializados antes da etapa de treinamento do modelo. São exemplos de hiperparâmetros: taxa de aprendizagem, número de épocas, camadas ocultas e funções de ativação (PRABHU, 2018).

Os hiperparâmetros são importantes porque controlam diretamente o comportamento do algoritmo de treinamento e têm um impacto significativo no desempenho do modelo que está sendo treinado. Uma boa escolha de hiperparâmetros pode aumentar significativamente o desempenho do modelo, por isso, existem técnicas de otimização desenvolvidas com o foco em determinar os melhores valores para os hiperparâmetros de um modelo em uma dada aplicação (CLAESEN e MOOR, 2015). Dentre as técnicas disponíveis, destaca-se a Otimização Bayesiana por utilizar uma estratégia de busca dividida em duas etapas, o que reduz a quantidade de iterações necessárias para atingir o resultado. Os principais aspectos da técnica são destacados a seguir.

2.2.6.1 Otimização Bayesiana

A otimização Bayesiana é uma técnica de otimização baseada em uma estratégia de duas fases. Na primeira, chamada de "aquecimento" é executada uma busca na qual as combinações de parâmetros são escolhidas e avaliadas aleatoriamente. Com base no resultado da fase de aquecimento, a segunda fase tenta encontrar combinações de parâmetros promissoras, que também são avaliadas.

Na segunda fase, a técnica utiliza o *Tree-structured Parzen Estimators* (TPE) para encontrar boas combinações de parâmetros e o algoritmo termina após um número especificado de iterações (BERGSTRA *et al.*, 2011). No TPE, em cada rodada um grupo de parâmetros candidatos são avaliados e divididos entre bons e ruins de acordo com uma taxa, a qual é representada pela letra γ.

O valor de γ é usado pelo TPE para dividir as combinações de parâmetros já avaliadas em boas e ruins com base em sua pontuação. O valor geralmente é escolhido para estar entre 0,15 e 0,30. Um valor γ de 0,25 significa que as combinações de 25% dos melhores parâmetros pertencerão à boa distribuição e o restante à ruim. Para ambos os grupos, uma função de densidade de probabilidade é construída.

O TPE tenta encontrar a próxima combinação de parâmetros, maximizando a melhoria esperada. Isso é feito escolhendo aleatoriamente candidatos do espaço de parâmetros. O número de candidatos que serão sorteados por rodada deve ser definido como um parâmetro de entrada do algoritmo. Para maximizar a melhoria esperada, deve-se maximizar a probabilidade da boa distribuição dividida pela probabilidade da má distribuição de cada candidato. O candidato com a maior melhoria esperada é o próximo parâmetro a ser avaliado. Quanto maior o número de candidatos escolhido, mais o algoritmo explorará bons espaços de parâmetro.

Ao avaliar iterativamente uma configuração promissora de hiperparâmetro com base em um resultado preliminar, a Otimização Bayesiana visa reunir observações que revelem o máximo de informações possível sobre o espaço de hiperparâmetros e, em particular, a localização do ponto ótimo. Na prática, a Otimização Bayesiana obtém melhores resultados em um número menor de buscas em comparação a outras técnicas também populares, como a *Grid Search* ou a *Random Search*, o que pode ser atribuído ao fato da técnica utilizar resultados de passos anteriores para orientar as buscas dos passos seguintes (SNOEK, LAROCHELLE e ADAMS, 2012).

2.2.7 Medidas de Desempenho para Modelos Preditivos

A avaliação do resultado obtido a partir da aplicação de um modelo preditivo, ou a comparação entre os resultados de diferentes modelos, exige o uso de medidas robustas que possam contemplar todos os aspectos sensíveis a classificação de dados. Além disso, deve-se levar em conta a natureza do problema e as características do conjunto de dados em cada caso, pois, a escolha de uma métrica inadequada pode levar a falsas conclusões de sucesso para os resultados de um modelo.

Felizmente existem medidas amplamente utilizadas na área de ciência de dados, as quais atendem tal finalidade. No caso de problemas de classificação, destaca-se o uso da Matriz de Confusão, por combinar simplicidade e completude de análise. Para os problemas de regressão, as principais métricas de avaliação são: o *Root Mean Square Error* (RMSE) e coeficiente de determinação (R^2) para o caso de ajuste de curvas. A seguir são apresentas as principais características de cada uma das métricas.

2.2.7.1 Matriz de Confusão

A Matriz de Confusão é uma ferramenta padrão de análise de desempenho para classificadores. Ela consiste na representação em forma de matriz indicando a frequência de classificação para cada classe do modelo e suas possíveis combinações. Cada linha da matriz representa os indivíduos em uma classe prevista, enquanto cada coluna representa os indivíduos em uma classe real (ou vice-versa) (KELLEHER, NAMEE e D'ARCY, 2015).

Na Figura 27 é apresentado o exemplo de uma Matriz de Confusão referente a um problema de classificação entre as classes positivo e negativo em um grupo de 10 indivíduos.

Figura 27 – Exemplo de Matriz de Confusão.

		Classe real		
		positivo	**negativo**	
Classe predita	**positivo**	**3 (VP)**	**1 (FP)**	75% (precisão)
	negativo	**2 (FN)**	**4 (VN)**	67% (valor preditivo negativo)
		60% (sensibilidade)	80% (especificidade)	70% (acurácia)

Fonte: Autoria própria.

Na Figura 27 é possível constatar que quatro indivíduos foram classificados como positivos, sendo que três realmente eram, os quais são denominados Verdadeiros Positivos (VP) e um foi classificado erroneamente, o qual é denominado de Falso Positivo (FP). De modo análogo, seis indivíduos foram classificados como negativos, dos quais quatro realmente eram, os quais são denominados Verdadeiros Negativos (VN) e dois foram classificados erroneamente, os quais são denominados Falsos Negativos (FN).

A partir das quantidades de VP, FP, VN e FN é possível determinar várias métricas estatísticas associadas ao desempenho do classificador, sendo que as cinco principais estão destacadas na Figura 27 e são descritas a seguir:

- Sensibilidade: Indica o percentual de indivíduos da classe positiva que foram identificados corretamente pelo classificador. A sensibilidade é definida conforme a Equação (25). No exemplo da Figura 27, dentre os cinco positivos existentes no grupo, apenas três foram identificados pelo classificador (sensibilidade = 60%).

$$Sensibilidade = \frac{VP}{VP + FN}. \tag{25}$$

- Especificidade: É o equivalente da sensibilidade para a classe dos negativos e é definida conforme Equação (26). No exemplo da Figura 27, dentre os cinco negativos existentes no grupo, quatro foram identificados pelo classificador (especificidade = 80%).

$$Especificidade = \frac{VN}{VN + FP}. \tag{26}$$

- Precisão: Indica o nível de acerto da classificação positiva e é definida conforme Equação (27). No exemplo da Figura 27, dentre os quatro indivíduos classificadas como positivos, apenas três realmente eram (precisão = 75%).

$$Precisão = \frac{VP}{VP + FP}. \tag{27}$$

- Valor Preditivo Negativo (VPN): É o equivalente da precisão para a classe dos negativos e é definido conforme Equação (28). No exemplo da Figura 27, dentre

os seis indivíduos classificados como negativos, apenas quatro realmente eram (valor preditivo negativo = 67%).

$$Valor\ preditivo\ negativo = \frac{VN}{VN + FN}. \tag{28}$$

- Acurácia: Indica o nível de acerto geral do classificador, incluindo as classes positivas e negativas. A acurácia é definida conforme Equação (29). No exemplo da Figura 27, do total de 10 indivíduos, sete foram classificados corretamente em suas respectivas classes (acurácia = 70%).

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN}. \tag{29}$$

Dentre as métricas destacadas as mais importantes para uma classificação binário, como no caso do problema da perda não técnica, são a sensibilidade e precisão. Não por acaso, as duas métricas guardam uma relação de compromisso entre si. No limite, é simples se obter uma alta precisão com baixa sensibilidade ou uma alta sensibilidade com uma baixa precisão. Contudo, o desafio para um bom classificador é obter a combinação de alta sensibilidade com alta precisão. Esta relação de compromisso pode ser mensurada por meio de uma métrica conhecida como F1-*score*, a qual consiste na média harmônica entre a sensibilidade e a precisão, como expresso na Equação (30).

$$\text{F1-}score = 2 \cdot \frac{\text{sensibilidade} \times \text{precisão}}{\text{sensibilidade} + \text{precisão}}. \tag{30}$$

O valor de F1-*score* varia entre zero e um, em que zero representa o máximo desequilíbrio, ou seja, uma das medidas na equação é 100% e a outra 0%. Por outro lado, F1-*score* é igual a 1 somente quando as duas medidas na equação forem iguais a 100%.

2.2.7.2 Root Mean Square Error

O *Root Mean Square Error* (RMSE) é uma das métricas mais populares para avaliar a qualidade de modelos em problemas de regressão. O RMSE é definido pelo desvio padrão dos erros das previsões realizadas para cada ponto do conjunto de dados, os quais são conhecidos como resíduos. Os resíduos representam a distância entre os pontos previstos e os pontos

verdadeiros no problema de regressão (TIWARI, 2019). O valor do RMSE é calculado como expresso na Equação (31).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}, \tag{31}$$

em que n é a quantidade de amostras no conjunto de dados; Y é o valor real do ponto; e $\hat{Y}$ é o valor previsto para o ponto.

O valor do RMSE informa quão boa é a concentração dos pontos de dados reais ao redor da curva de regressão. Uma curva de regressão perfeita, ou seja, que sobrepõe todos os pontos da amostra, resultaria em um RMSE igual a zero. Por outro lado, não há um valor máximo para o RMSE, já que ele tem a mesma escala da variável original. Portanto, o valor do RMSE só pode ser comparado em um mesmo conjunto de dados (SRIVASTAVA, 2019).

Um valor de referência para comparação com o RMSE de um modelo preditivo pode ser obtido a partir de uma previsão ingênua dos dados, conhecida como *Naïve Predictive Model*, o qual consiste em utilizar o valor médio do conjunto de dados como previsão para todos os pontos da amostra. Se o RMSE de um modelo preditivo for menor que o RMSE do *Naïve Predictive Model*, então, o modelo é considerado relevante (BROWNLEE, 2021).

2.2.7.3 Coeficiente de Determinação

O coeficiente de determinação, também conhecido como R^2, é uma medida estatística que representa a proporção da variância de uma variável dependente que é explicada por uma variável independente em um modelo de regressão (FERNANDO, 2020). O valor de R^2 é calculado como expresso na Equação (32).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \tag{32}$$

em que n é a quantidade de amostras no conjunto de dados; Y é o valor real do ponto; $\hat{Y}$ é o valor previsto para o ponto; e $\bar{Y}$ é o valor médio do conjunto de valores reais.

O valor de R^2 informa em que medida o valor previsto explica a variância do valor verdadeiro no problema de regressão, portanto, o R^2 pode ser considerado uma medida da qualidade do ajuste de curvas. O valor do coeficiente varia entre zero e um, em que um indica

o ajuste perfeito da variável prevista com a variável verdadeira, enquanto zero indica que a previsão não possui nenhuma capacidade de explicar a variância do valor verdadeiro.

Contudo, é importante que destacar que o R^2 não indica necessariamente a qualidade de um modelo preditivo e, portanto, não deve ser utilizado como parâmetro de qualidade para seleção de modelos, para esta finalidade o RMSE é a melhor escolha. O coeficiente R^2 apenas indica o quão bem o valor previsto explica as variações no valor real (SRIVASTAVA, 2019).

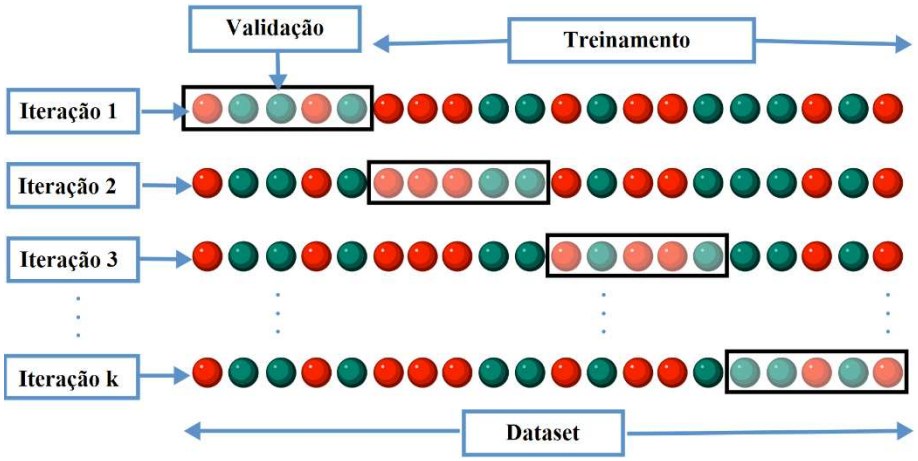
2.2.8 Validação Cruzada

Para garantir que as medidas de desempenho discutidas na subseção anterior representem uma medida generalizada da qualidade dos métodos, faz-se necessário que elas sejam obtidas por meio de um processo de validação robusto, o qual é conhecido como validação cruzada.

Na validação cruzada o objetivo é avaliar como será o desempenho dos modelos preditivos mediante entradas de dados genéricas, que sejam diversas entre si e distintas do conjunto utilizado para a etapa de treinamento. Assim, é possível estimar o desempenho dos modelos em aplicações reais (STONE, 1974).

O método mais popular de validação cruzada é *k-fold*, o qual consiste em particionar o conjunto de dados disponível em vários subconjuntos e, posteriormente, realizar várias iterações de treinamento e validação do modelo preditivo com os diferentes subconjuntos (BROWNLEE, 2018). Na Figura 28 é ilustrado o funcionamento da validação cruzada *k-fold*.

Figura 28 – Ilustração do processo de validação cruzada *k-fold*.



Fonte: Adaptado de (GUFOSOWA, 2019).

Como pode ser visto na Figura 28, o conjunto total de dados é dividido em k subconjuntos. Em cada iteração, um subconjunto é selecionado como o conjunto de validação, enquanto os demais k-1 subconjuntos são combinados como um conjunto de treinamento. O procedimento é repetido k vezes para cada um dos subconjuntos. Ao final, o modelo terá classificado todos os dados disponíveis a partir de conjuntos de treinamento independentes. Por fim, os resultados de cada iteração são coletados para compor o desempenho final do modelo. Quanto maior o valor de k, melhor será a generalização dos resultados, porém, maior será também o custo computacional. Um valor tipicamente utilizado para equilibrar a relação de compromisso é k = 10 (BROWNLEE, 2018).

Diante do exposto na presente seção, é possível concluir que as diferentes técnicas que compõem o processo de *Advanced Analytics* podem ser utilizadas de maneira combinada para construir uma solução robusta e adaptativa para o problema de detecção de perda não técnica no sistema de distribuição. Tal solução pode garantir que o máximo de informação seja extraído dos dados disponíveis e que os modelos criados sejam os mais adequados para a natureza do problema. A abrangência e a flexibilidade da abordagem *Advanced Analytics* representam um avanço importante em relação às soluções que foram propostas na bibliografia, o que poderá ser constatado na seção seguinte, na qual é apresentado o estado da arte para o tema.

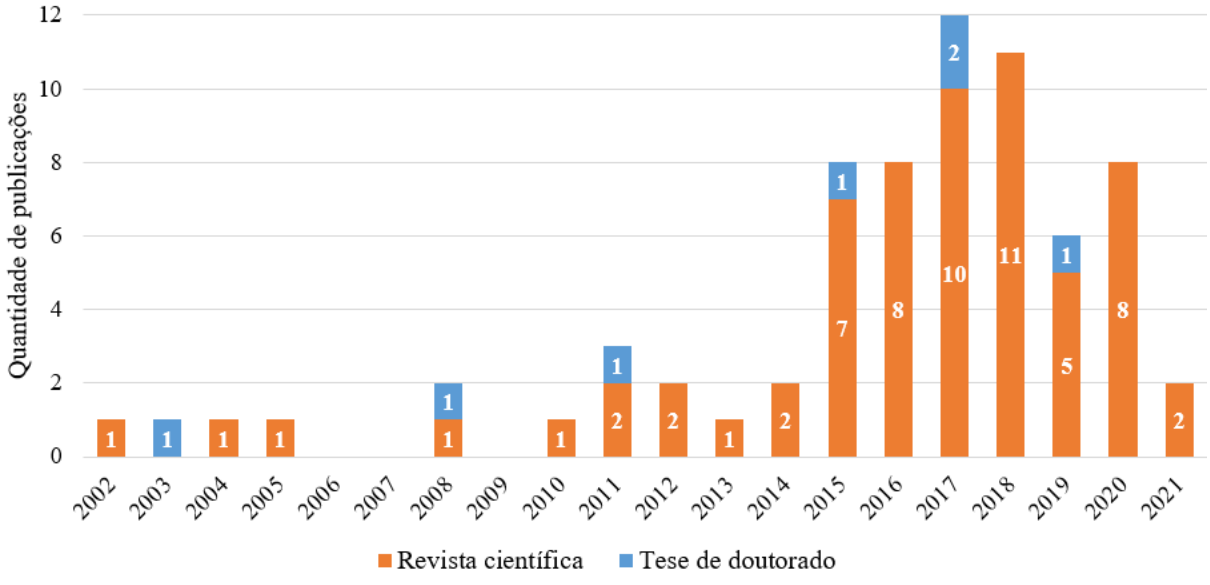
3 O Estado da Arte

Neste capítulo são apresentados os principais trabalhos correlatos ao tema em estudo disponíveis na bibliografia científica. O objetivo do capítulo é compor o estado da arte para o tema, destacando as principais contribuições de cada autor, bem como identificar as limitações das soluções apresentadas e as oportunidades de melhorias existentes.

3.1 Análise Quantitativa

A partir de uma extensa pesquisa, foram obtidas as publicações mais relevantes nos principais meios de divulgação cientifica, notadamente as bases do *IEEExplore*®, *ScienceDirect*®, *IET Digital Library*® e bancos de teses e dissertações de vários países. No gráfico da Figura 29 é possível visualizar a quantidade de trabalhos identificados em revistas científicas e em teses de doutorado por ano de publicação.

Figura 29 – Quantidade de publicações relacionadas ao tema em revistas científicas e em teses de doutorado.

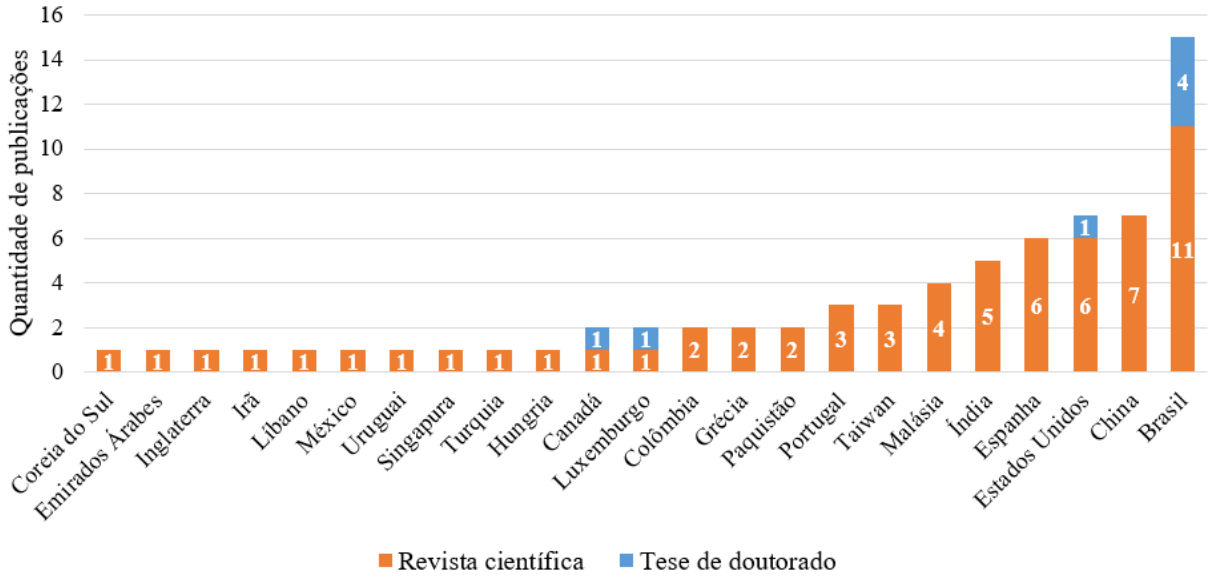


Fonte: Autoria própria.

A partir da Figura 29 é possível constatar que a quantidade de publicações tem crescido significativamente nos últimos cinco anos, o que demonstra a relevância e atualidade do tema em estudo e o interesse tanto da comunidade científica quanto do setor elétrico em soluções que

possam ser aplicadas para a redução da perda não técnica de energia. Adicionalmente, na Figura 30 o número de publicações é agrupado por país de origem.

Figura 30 – Quantidade de publicações relacionadas ao tema por país de origem.



Fonte: Autoria própria.

Na Figura 30 nota-se uma concentração de publicações nos países que apresentam níveis de perda não técnica de energia elevado, como os destacados anteriormente na Tabela 1, com exceção dos Estados Unidos que mesmo possuindo baixos índices de perda é o terceiro país com mais publicações. Podem ser destacados os países emergentes como o Brasil, que possui o maior número de publicações, seguido de China e Índia. Destacam-se ainda Espanha e Portugal na Europa e Malásia e Taiwan na Ásia. Na África, onde estão os países com os maiores índices de perda não técnica de energia do mundo, não foram encontradas publicações. O fato pode ser atribuído aos baixos índices de publicação científica no continente.

No que diz respeito às linhas de pesquisa desenvolvidas, é possível observar que a maior parte das soluções utilizadas até o momento para o problema da perda não técnica no sistema elétrico baseiam-se na aplicação de Inteligência Artificial para identificação de consumidores irregulares ou de infraestrutura avançada de medição para realização de balanço energético no sistema, como pode ser constatado na análise descritiva apresentada na seção seguinte.

3.2 Análise Descritiva

O primeiro trabalho de destaque que trata da problemática da perda não técnica de energia foi publicado no final do século XX. Dick (1995) fez um relato sobre o furto de energia elétrica no Reino Unido e as principais ações executadas pelas companhias de distribuição para combater a prática. Segundo o autor, na época o furto de energia representava uma perda de faturamento de £50 milhões por ano para companhias de distribuição. A principal forma de furto era a interferência indevida no medidor de energia, que na época era um equipamento facilmente manipulado, o simples fato de incliná-lo poderia alterar o valor de energia medido.

Ainda segundo Dick (1995), as ações das companhias eram baseadas nas denúncias da própria população e dos funcionários da empresa e consistiam na troca do medidor por modelos mais robustos e na instalação do equipamento em local externo à residência. No trabalho, o autor menciona ainda o início do desenvolvimento dos medidores eletrônicos baseados em componentes de estado sólidos, o quais eram mais resistentes a interferências dos consumidores.

Na sequência, Ghajar, Khalife e Richani (2002) propuseram uma modernização do sistema de medição de energia elétrica, com o objetivo de reduzir a possibilidade de erros e fraudes no sistema de medição. Foi proposto um sistema chamado *Automatic Meter Reading* (AMR), que incluía o uso dos recém-criados medidores eletrônicos e a comunicação por rádio frequência para coleta remota da leitura de consumo e envio dos dados para uma central de computação, na qual as faturas seriam processadas automaticamente via *software*. Apesar de visionário, o trabalho é completamente propositivo e não apresenta nenhuma implementação prática. O adjetivo visionário deve-se ao fato de que a ideia do autor é a base da infraestrutura avançada de medição proposta como solução em trabalhos posteriores.

No ano seguinte, Eller (2003) utilizou Redes Neurais Artificiais (RNA) para classificar consumidores como possíveis fraudadores no sistema de distribuição da Centrais Elétricas de Santa Catarina (CELESC). O autor tomou como base informações do histórico de consumo de energia elétrica e dados do Imposto Predial e Territorial Urbano (IPTU) das residências. Segundo simulações apresentadas pelo autor, seria possível atingir uma taxa de acerto de até 50% na classificação realizada. Contudo, não foram realizadas verificações em dados reais. Eller (2003) é o primeiro autor a aplicar técnicas de mineração de dados para identificação de perda não técnica no sistema elétrico de distribuição, dentre os trabalhos analisados. Até então as técnicas haviam sido aplicadas para outros setores, como transações bancárias, telefonia e cartão de crédito.

Smith (2004) apresentou uma perspectiva histórica do furto de energia em vários países, destacando o crescimento exponencial que houve entre as décadas de 1980 e 2000. Como possível solução, o autor propôs o uso de caixas de medição blindadas, nas quais haveria maior dificuldade de manipulação dos medidores. O autor também destaca que mudanças estruturantes na legislação do setor elétrico são necessárias para a redução do furto de energia. Foram destacados casos de sucesso em alguns países onde houve privatização de concessionárias e implantação de regras de incentivo econômico para o combate às perdas. No trabalho de Smith (2004) é apresentada uma boa perspectiva global da perda não técnica de energia e seus impactos. Contudo, o autor não explora a aplicação direta de soluções técnicas para o problema.

Tomando proveito da evolução da área de inteligência artificial, Nizar, Dong e Wang (2008) utilizaram uma técnica conhecida como *Extreme Learning Machine* (ELM) para classificar consumidores em dois grupos: regulares e irregulares. O trabalho baseou-se na curva de consumo de energia diária em kWh para identificar perfis relacionados à ocorrência de perda não técnica. Os autores compararam os resultados obtidos com o emprego do ELM com os obtidos com a técnica *Support Vector Machine* (SVM), considerada uma das principais técnicas de classificação disponível na época. Com o uso do ELM foi possível obter menor custo computacional e maior acurácia na classificação das curvas de consumo, em comparação ao SVM. Dentre as limitações do trabalho, está o fato de que o método só pode ser aplicado nos casos em que há medição de consumo de hora em hora. Além disso, os autores classificam como irregular os consumidores cujo perfil da curva de consumo é diferente de um perfil típico, porém, tal fato não é suficiente para determinar que o consumidor possui perda não técnica.

Uma abordagem multidisciplinar para o tratamento da perda não técnica de energia foi apresentada por Penin (2008). O autor propôs soluções baseadas em melhores práticas de várias concessionárias, tais como, medição remota, medidores pré-pagos, gestão de selos e auditorias em campo. Em uma segunda etapa, o autor utilizou RNA para identificar o perfil de consumo de clientes fraudadores com base no histórico de inspeções realizadas em uma concessionária. Segundo o autor, os resultados alcançados foram superiores aos praticados pela concessionária com as regras convencionais. Pode-se concluir que o trabalho de Penin (2008) é abrangente e ao mesmo tempo superficial, pois, trata de diversos temas simultaneamente sem se aprofundar em nenhum deles. A abordagem com RNA foi bastante limitada, por se basear apenas no histórico de consumo dos clientes, de modo que unidades com queda de consumo são classificadas como suspeitos, e unidades sem queda de consumo são classificados como regulares. O que é muito trivial para identificação de casos reais de perda não técnica.

Um trabalho utilizando a técnica SVM para identificar consumidores com suspeita de perda não técnica em uma concessionária de energia da Malásia foi desenvolvido por Nagi *et al.* (2010). Foram utilizados como dados de entrada as informações de histórico de consumo mensal, tipo de medidor, histórico de detecção de fraudes, dívidas de faturamento e relatório de irregularidades. Contudo, o foco da técnica era a detecção de quedas abruptas no histórico de consumo. Adicionalmente, os autores criaram uma etapa de tomada de decisão manual para o modelo, na qual um analista decide qual consumidor será encaminhado para inspeção em campo, tomando como base a probabilidade indicada pelo SVM e seu o próprio conhecimento. Com a etapa de tomada de decisão manual, a taxa de acerto foi de 64% para as inspeções realizadas em campo. Sem a etapa manual, a taxa de acerto foi de 26%. Tal fato demonstra que a maior parte da inteligência estava no conhecimento do analista e não no modelo baseado em SVM. No trabalho existe ainda outra limitação, que é se basear apenas na queda de consumo para indicar a ocorrência de perda não técnica em unidades consumidoras.

O trabalho foi aprimorado no ano seguinte por Nagi *et al.* (2011) substituindo a etapa de tomada de decisão manual pelo algoritmo *Fuzzy Inference System* (FIS). Trata-se de um algoritmo do tipo *if-then* em que as regras são implementadas com base no conhecimento tácito do usuário. No trabalho, o FIS atua como um esquema de pós-processamento para o refinamento da seleção de consumidores suspeitos com maior probabilidade de fraude. Segundo os autores, foi possível melhorar em 17% o desempenho do modelo com a implementação do FIS. O trabalho resolve um problema da primeira proposta dos autores, pois transfere para a máquina o conhecimento do especialista, que antes era um pré-requisito para uso do modelo.

Angelos *et al.* (2011) propuseram uma metodologia para a classificação de consumidores com suspeita de perda não técnica baseada na mudança do perfil de consumo de energia elétrica. A metodologia é composta de duas etapas, na primeira o algoritmo *Fuzzy c-Means* é utilizado para agrupar consumidores com perfis de consumo semelhantes em diferentes *clusters*. O agrupamento é baseado em cinco métricas derivadas de uma série de seis meses de consumo. Na segunda etapa, uma classificação baseada na lógica *Fuzzy* é executada utilizando informações de uma série de 12 meses de consumo, os quais são posteriores aos seis meses utilizados na primeira etapa. Por fim, a distância euclidiana da classificação aos centros dos *clusters* originais é calculada. Em seguida, as medidas de distância euclidiana da classificação são normalizadas e ordenadas, gerando uma pontuação de zero a um. Ao final, os potenciais fraudadores apresentam as maiores pontuações.

A metodologia proposta por Angelos *et al.* (2011) destaca-se por ser não supervisionada e necessitar de poucas informações de entrada, o que facilita sua aplicação. Contudo, o método

limita-se a identificar fraudes em que há uma mudança no perfil de consumo do cliente necessariamente entre os períodos utilizados na primeira e na segunda etapa.

Na Espanha, León *et al.* (2011) propuseram uma nova metodologia para identificação de consumidores com potencial de fraude no sistema de distribuição. O trabalho é baseado no método de classificação *Generalized Rule Induction* (GRI) e utiliza métricas de variabilidade e tendência extraídas da série de consumo mensal dos consumidores. O método consiste em avaliar um conjunto de cinco regras do tipo *if-then* definidas pelos autores. A depender da combinação de resultados o consumidor é classificado como normal ou como suspeito. Segundo os autores, utilizando um conjunto de dados da distribuidora espanhola Endesa, foi possível obter uma taxa de acerto de 20%, enquanto a média histórica da companhia era de 7%. A metodologia proposta por León *et al.* (2011) possui como vantagem a simplicidade de implementação do método GRI, contudo, não fica claro no trabalho como as regras utilizadas foram obtidas, deixando a entender que elas são derivadas do conhecimento tácito dos autores, o que dificultaria a escalabilidade do método para outras distribuidoras.

Até o momento se destacavam como principais técnicas preditivas para a identificação de perda não técnica a RNA e o SVM. Neste contexto, Ramos *et al.* (2011) propuseram o uso de um novo algoritmo como alternativa aos já tradicionais. Trata-se do *Optimum-Path Forest* (OPF), que é um tipo de classificador baseado em uma computação combinatória de caminhos ideais. Dentre as vantagens do OPF, os autores destacam a ausência de parâmetros e o baixo custo computacional, o que torna a técnica apropriada para aplicações de tempo real. Para aplicação do modelo foram utilizados dados de medição de demanda de 15 em 15 minutos e algumas métricas adicionais como fator de carga e demanda contratada de um conjunto de consumidores de uma concessionária brasileira. Os autores compararam os resultados obtidos com o OPF com os de outras técnicas como RNA e SVM. Em todos os casos apresentados o uso do OPF proporcionou ganhos tanto em acurácia quanto em tempo de processamento. A abordagem com o OPF proposta por Ramos *et al* (2011) é uma contribuição importante, pois, é uma alternativa de classificador rápido que pode ser utilizado para detecção de fraude em tempo real em grandes clientes. Contudo, o modelo baseia-se em poucas métricas e, por isso, a quantidade de tipos de fraude detectáveis são limitadas.

Bastos (2011) propôs uma metodologia para diagnóstico das perdas não técnicas utilizando Redes Bayesianas e investigação de conformidade. O procedimento consiste em inspecionar uma amostra de consumidores e a partir dos resultados utilizar o Teorema de Bayes para confirmar hipóteses iniciais. A partir da metodologia e utilizando um estudo de caso, o autor estimou as perdas não técnicas de uma distribuidora segundo as suas causas e regiões de

incidência. O trabalho de Bastos (2011) permite que seja analisado o custo-benefício de diferentes ações para minimização da perda não técnica em uma distribuidora e pode ser útil na definição de um plano estratégico. Contudo, a metodologia limita-se a oferecer uma visão macroscópica da perda, não sendo possível identificar individualmente quem são os consumidores irregulares.

Com o patrocínio da concessionária espanhola Endesa, Mondero *et al.*, (2012) aplicaram o coeficiente de *Pearson* e Redes Bayesianas para identificar consumidores com perda não técnica na concessionária. Cada técnica foi utilizada para identificar um determinado perfil de irregularidade. O coeficiente de *Pearson* foi utilizado para casos de quedas abruptas de consumo e as Redes Bayesianas foram utilizadas para casos sem queda de consumo. Neste caso, foram analisadas informações relacionadas ao histórico de inspeções disponível na empresa, tais como, máximo e mínimo consumo no período, fator de potência, variabilidade do consumo, entre outros. O trabalho de Mondero *et al.* (2012) representa uma contribuição importante, pois, foi o primeiro a diferenciar perfis distintos de fraude, utilizando técnicas específicas para identificar cada um. Por outro lado, a abordagem com o coeficiente de *Pearson* pode ser considerada enviesada, pois, basicamente se propõe a identificar clientes com queda abrupta no consumo de energia, o que por si só não significa uma irregularidade.

Huang, Lo e Lu (2013) apresentaram uma abordagem baseada em Estimação de Estados para determinar a carga aproximada de transformadores de distribuição em concessionárias Taiwanesas. Os valores estimados foram utilizados para detectar possíveis violações dos medidores de energia e fornecer evidências quantitativas da perda não técnica no sistema de distribuição. O método proposto consiste em utilizar medições de *smart meters*, disponíveis em diferentes pontos do sistema, para obter uma curva de carga em cada transformador a partir de um método de Estimação de Estados. Na sequência a curva de carga estimada do transformador é comparada com a curva de carga formada pela agregação das medições de faturamento dos clientes. Por fim, a variância entre as duas curvas de carga é avaliada, os transformadores que apresentaram maior variância são classificados como portadores de perda não técnica. Os autores avaliaram o método em uma rede simplificada de 39 barras, em que foi possível demonstrar o aumento da variância entre as curvas no momento em que um erro de medição foi introduzido no sistema. Diferentemente dos estudos discutidos até o momento, o trabalho de Huang, Lo e Lu (2013) não se baseia em *software*, mas em *hardware*. Tal fato pode agregar maior confiabilidade a solução, porém, também impõe maiores custos de implementação.

Novamente com patrocínio da Endesa, Guerrero *et al.* (2014) utilizaram a técnica *Knowledge-Based System* (KBS) para implementar uma solução de detecção de perda não

técnica baseada na expertise dos inspetores da companhia. A solução consiste em um sistema baseado em regras de validação de dados preestabelecidas, as quais foram obtidas a partir de entrevistas com inspetores da companhia com anos de experiencia na detecção de perda não técnica. Foram implementadas regras para validação de informações contratuais, histórico de consumo e histórico de inspeções. Também foi implementa uma técnica de reconhecimento de texto para identificar palavras-chave nas observações realizadas em ordens de serviço. Segundo os autores, o uso de regras baseadas no conhecimento de especialistas apresentou um desempenho superior ao de métodos supervisionados como RNA. O trabalho de Guerrero *et al.* (2014) é uma contribuição importante, pois apresenta um método para transferir para máquina o conhecimento humano, contudo, a solução limita-se a ser no máximo tão boa quanto um ser humano.

A Teoria dos Jogos foi utilizada por Amin *et al.* (2015) para determinar a quantidade, o tipo e a localização ideal das medições de um sistema de *smart meter* em uma rede distribuição, do ponto de vista de detecção da perda não técnica. Foram considerados vários fatores na análise como custo da infraestrutura, probabilidade de cometimento de fraude, custo da fraude, importância de cada consumidor, dentre outros. O principal objetivo da Teoria dos Jogos é obter uma solução de equilíbrio para todas as variáveis, levando em conta as relações de dependência existente. O trabalho de Amin *et al.* (2015) é uma contribuição relevante pois oferece um método para dimensionamento de um sistema de *smart meter* com viés de redução de perda não técnica, contudo, trata-se de um trabalho teórico e de difícil validação prática.

Na sequência, vários trabalhos propuseram o uso de *smart meter* para localização de perda não técnica, como Xu (2015) que analisou separadamente dados de corrente, tensão e energia em cada fase da instalação para identificar condições anormais de medição. Jokar, Arianpoo e Leung (2015) propuseram um novo algoritmo chamado *Consumption Patter-Based Energy Theft Detector* (CPBETD) para detecção de perda não técnica a partir da característica da curva de carga diária dos consumidores. Os autores conseguiram identificar anomalias relacionadas ao furto de energia, como por exemplo, uma queda de demanda em um indivíduo no horário em que todos os demais consumidores próximos elevam seu consumo. Os trabalhos de Xu (2015) e Jokar, Arianpoo e Leung (2015) apresentam métodos confiáveis de localização de perda não técnica, pois, baseiam-se em medições, contudo, apenas uma pequena parte do sistema de distribuição possui a infraestrutura avançada de medição utilizada nos trabalhos, o que diminui seu potencial de aplicação.

Jindal *et al.* (2016) propuseram uma solução para detecção em tempo real da perda não técnica de energia em vários níveis do sistema de distribuição. A solução é baseada em um

sistema de *smart grid* com infraestrutura avançada de medição em todas as barras, incluindo nos consumidores. No esquema proposto um alimentador é dividido em segmentos e, para determinar a perda de cada segmento, é realizada a comparação entre as potências medidas na entrada e na saída do segmento. Para os casos em que a perda é considerada elevada, é feito uma segunda análise agora a nível de consumidor. Para cada consumidor são obtidas informações de carga instalada, número de habitantes, horário de funcionamento e temperatura atual. Com base nas informações e no método das Árvores de Decisão é obtido o consumo esperado para a unidade. Na sequência, o consumo real é comparado com o consumo esperado, os casos em que há divergência acima de um limite são indicados como possíveis fraudadores.

A metodologia proposta por Jindal *et al.* (2016) é inovadora por combinar dados de medição de *smart meter* com técnicas de aprendizagem de máquina para obter uma classificação em tempo real dos consumidores. Contudo, a metodologia dificilmente poderá ser implementada em uma situação prática devido à complexidade das informações necessárias, como infraestrutura de medição avançada em todas as barras do sistema e informações de número de moradores de todas residência.

Outros trabalhos seguiram na mesma linha de tentar segregar a perda do sistema por segmentos de rede, utilizando para isso medições de *smart meter* localizadas em vários pontos da rede, como os trabalhos de Leite e Mantovani (2016) e Zhan *et al.* (2016). No último, a infraestrutura avançada de medição também foi utilizada para localização de faltas no sistema. Os trabalhos também encontraram aplicação limitada devido à necessidade de um grande número de medições ao longo da rede, o que ainda não é uma realidade viável para as distribuidoras. Com objetivo de contornar esta dificuldade, Madrigal, Rico e Uzcaregui (2017) apresentaram um método no qual a perda não técnica no sistema pode ser estimada com base em poucas medições instaladas em pontos aleatórios da rede, para tanto, um modelo baseado em Regressão Linear Múltipla foi desenvolvido. Apesar de utilizar poucas medições o método de Madrigal, Rico e Uzcaregui (2017) serve apenas para estimar o nível de PNT global do sistema, não sendo possível identificar a perda no nível de consumidores.

Tomando proveito de novos algoritmos de otimização que foram desenvolvidos, Ramos *et al.* (2016) aplicaram o *Black Hole Algorithm* (BHA) para selecionar um conjunto de métricas adequado à caracterização de perda não técnica nos consumidores. A ideia é selecionar, a partir de um conjunto com várias métricas, o subconjunto mínimo que permita a melhor taxa de reconhecimento de uma variável alvo, que no caso é a presença de perda não técnica nos consumidores. Para validar a eficácia do método, os autores utilizaram o algoritmo OPF para classificar um conjunto de consumidores em regulares ou irregulares. Inicialmente foram

utilizadas todas as oito métricas disponíveis no conjunto de dados, como demanda máxima, demanda mínima, fator de carga, dentre outras. Na sequência, o algoritmo BHA foi aplicado para selecionar o conjunto ótimo de métricas e, por fim, o algoritmo foi OPF novamente aplicado com o novo conjunto de métricas para classificar os consumidores em regular ou irregular. O algoritmo BHA selecionou apenas quatro das oito métricas disponíveis e o resultado da nova classificação foi 12% superior ao inicial.

No trabalho de Ramos *et al.* (2016) é abordada uma questão relevante, porém pouco enfatizada nos trabalhos apresentados até então, que é o processo de seleção das variáveis mais adequadas para a correta caracterização da perda não técnica nos consumidores. O processo é determinante para o sucesso da identificação de perda não técnica, podendo ter mais influência que a própria técnica de classificação utilizada, por isso, mereceria ser melhor explorado em outros trabalhos.

Na linha de pesquisa de infraestrutura avançada de medição, Zanetti (2017) apresentou uma solução para detecção de perda não técnica na qual todos os transformadores de distribuição possuem medição no lado secundário. A medição dos transformadores é comparada então com o somatório das medições dos consumidores conectados ao equipamento. O autor também apresenta uma sugestão para estimar a perda técnica na rede secundária do sistema, a qual se baseia no comprimento da rede. A busca por consumidores fraudulentos é iniciada quando ocorre uma inconsistência entre a energia fornecida pelo transformador e o somatório dos consumos dos clientes e da perda técnica estimada. A partir da análise dos instantes anteriores e posteriores ao início da inconsistência, o autor consegue detectar qual o consumidor estava furtando energia, geralmente é aquele que apresentou uma queda de consumo no mesmo intervalo de tempo.

A metodologia proposta por Zanetti (2017) apresenta vantagens como detecção em tempo real de fraudadores e a possibilidade de determinar o montante de energia que está sendo desviado e, portanto, pode ser utilizada para priorizar as ações nos maiores casos perda. Porém, assim como comentado para os demais trabalhos baseados em *smart meters*, a limitação da solução está no fato de que a infraestrutura avançada de medição ainda não é uma realidade viável para as distribuidoras de energia. Além disso, a metodologia proposta só é capaz de detectar fraudes que tenham iniciado após o começo da análise das medições, o que também limita a quantidade de casos detectáveis.

Viegas, Esteves e Vieira (2018) propuseram um novo método para detecção de perda não técnica baseado no algoritmo de *Gustafson-Kessel fuzzy* e em medições de um sistema de *smart meter* de uma distribuidora portuguesa. O método consiste em agrupar os consumidores,

que são conhecidamente regulares, de acordo com o seu perfil de consumo diário, o agrupamento é realizado com o uso do algoritmo *Gustafson-Kessel fuzzy*. Assim, é possível obter um banco de dados com vários perfis de consumo regulares para a distribuidora. Para cada nova medição obtida do sistema de *smart meter*, o algoritmo *Gustafson-Kessel fuzzy* é novamente aplicado para tentar associá-la a algum dos grupos de consumidores regulares existente. Por fim, é calculada a distância do novo indivíduo para o centroide do grupo ao qual ele foi associado. Segundo os autores, quanto maior a distância para o centroide, maior será a probabilidade de existir perda não técnica na medição em questão. Os autores afirmam que é possível obter uma taxa de acerto superior a 60% com o uso de método. Contudo, o método só pode ser aplicado a sistemas com infraestrutura avançada de medição.

Baseado na utilização de medições de *smart meters* Singh, Bose e Joshi (2018) monitoraram a dinâmica de consumo de energia elétrica de algumas unidades consumidoras ao longo do tempo, mudanças significativas foram apontadas como possíveis fraudes. O método é baseado na distância de *Kullback-Leibler*, também chamada de entropia relativa, a qual é utilizada para calcular a distância entre as distribuições de probabilidade das variações no histórico de consumo. Segundo os autores, é possível obter índices de acerto de até 70%, contudo, os testes só consideraram casos em que há variações bruscas no perfil de consumo o que limita a abrangência dos resultados apresentados para casos práticos, em que pode haver fraudes sem variações abruptas de consumo.

Um método baseado em *hardware* foi proposto por Aryanezhad (2019) para eliminação da perda não técnica do sistema de distribuição. O método consiste em instalar um regulador de tensão controlado por *software* no início da rede secundária capaz de executar rotinas para reduzir ou aumentar a tensão da rede em determinados intervalos. O objetivo é obter valores de tensão que sejam impróprios para o funcionamento de equipamentos elétricos. Clientes regulares não seriam afetados devido à instalação de um dispositivo de supervisão da tensão instalado junto ao medidor que sempre ajustaria o nível de tensão para o valor nominal da rede. Desse modo, somente os clientes irregulares teriam o funcionamento de seus equipamentos prejudicados, eliminando assim a perda não técnica do sistema. O autor construiu um protótipo para demonstrar o funcionamento do sistema proposto.

Apesar de ser eficaz na eliminação de clientes irregulares, o método proposto por Aryanezhad (2019) pode ser considerado agressivo, pois, levaria a queima de equipamentos elétricos e a possíveis acidentes envolvendo pessoas nas instalações irregulares. Além disso, o método tem um custo de implementação elevado e só eliminaria os consumidores conectados diretamente à rede elétrica, sem a presença de medidores.

Punmiya e Choe (2019) criaram métricas derivadas das medições de *smart meters* em um processo conhecido como *Feature Engineering*. O método consiste em calcular diversas métricas derivadas, tais como desvio padrão, máximo, mínimo, média da curva de consumo diário dos consumidores, dentre outros. As métricas criadas são utilizadas como entrada para um modelo de predição baseado no *Gradient Boosted Classifier* (GBC). Segundo o autor, as métricas derivadas podem ser mais explicativas para a ocorrência de perda não técnica que a informação original da curva de consumo diária, contudo, não foram apresentadas validações com testes reais para o método proposto.

Seguindo a mesma linha de pesquisa, Alves (2019) analisou a contribuição de métricas derivadas do histórico de consumo de energia elétrica para identificação de fraudes em unidades consumidoras. Foram criadas métricas a partir das séries de consumo mensal, considerando características como sazonalidade, amplitude e análise no domínio da frequência. Após a criação das métricas, o autor aplicou os métodos *Correlation Based Feature Selection* e *Relief* para seleção das variáveis mais explicativas dentre as disponíveis. Por fim, uma RNA foi utilizada para medir a taxa de acerto em um modelo considerando todas as métricas criadas, um considerando apenas as métricas selecionadas pelos métodos citados e outro considerando o próprio histórico de consumo como entrada. Segundo o autor, o modelo considerando apenas as métricas selecionadas resultou em uma melhoria de 10 pontos percentuais na taxa de acerto, em relação ao caso considerando apenas o histórico de consumo como entrada. O trabalho de Alves (2019) reforça a conclusão de que a criação de novas variáveis em complemento as variáveis nativas de um banco de dados, associada às técnicas de seleção de variáveis, tem o potencial de tornar os modelos preditivos mais precisos.

Glauner (2019) desenvolveu um sistema de realidade virtual para seleção de consumidores com suspeita de perda não técnica. O sistema consiste na visualização espacial em 3D das unidades consumidores por meio do HoloLens, óculos de realidade virtual da *Microsoft*®. No holograma é possível visualizar a probabilidade de cometimento de fraude de cada unidade consumidora, a qual é obtida a partir de diferentes métodos de aprendizagem de máquina, como SVM e *Decision Tree* (DT). Além disso, também é possível visualizar outras informações dos consumidores, como histórico de consumo e informações contratuais. A decisão final sobre quais unidades serão direcionadas para inspeção em campo fica por conta de um especialista que opera o sistema. O sistema desenvolvido por Glauner (2019) é uma forma avançada de selecionar consumidores para inspeção, porém, o custo para sua implementação em larga escala tende a ser elevado, pois, cada consumidor teria quer modelado individualmente em um ambiente de realidade 3D.

Uma nova técnica para o balanceamento de amostras em problemas de PNT que utilizam séries de consumo obtidas em sistemas de *smat meter* foi apresentada por Aslam *et al.* (2020). A técnica foi nomeada de IQMOT e é baseada no *oversampling* da classe minoritária a partir dos valores dos quartis calculados para as séries de consumo. Os quartis funcionam como limites para os valores sintéticos a serem gerados. Calcula-se o percentual de consumidores que estão em cada quartil e posteriormente os valores são utilizados como pesos para gerar uma nova curva de consumo sintética que é similar as curvas reais.

Os autores argumentam que a nova técnica supera os problemas de *overfiting* observados em técnicas tradicionais como o SMOTE. Para validar os resultados, foram realizadas simulações com dados obtidos de um sistema de *smart meter* da *State Grid Corporation of China* com 38.757 consumidores. Os autores utilizaram o algoritmo de *Long Short-Term Memory* (LSTM) e o *Adaptive Boosting* (Adaboost) para identificar os consumidores com PNT. Os resultados foram comparados com os de uma base de dados que utilizou o SMOTE para balanceamento das amostras, e uma base de dados não balanceada. A conclusão dos autores é que o IQMOT é superior, pois, apresentou *F1-score* de 0,954, enquanto o SMOTE foi de 0,901, e base não balanceada foi de 0,753.

Calvo *et al.* (2020) propuseram a segmentação prévia dos consumidores de acordo com o tipo de PNT cometida, para posteriormente realizar a classificação dos dados com modelos personalizados para cada grupo. Três tipos de PNT foram consideradas: manipulação de medidores, desvio de energia e falha técnica. Os autores realizaram inspeções em grupos de consumidores selecionados com a abordagem de segmentação e sem a segmentação. O algoritmo *Gradient Boosting Tree* (GBT) foi utilizado como classificador. Os resultados mostraram que a etapa de segmentação prévia promoveu uma melhora de aproximadamente 70% na precisão da classificação. Os autores também argumentam que a segmentação aumenta a interpretabilidade dos resultados dos modelos. Apesar disso, os resultados apresentados podem ser considerados frágeis, pois não indicam a quantidade de consumidores que foram inspecionados, e a precisão isoladamente não é uma boa métrica de avaliação, pois, pode mascarar resultados enviesados.

Uma abordagem de detecção de PNT baseada no algoritmo *Text Convolutional Neural Networks* (TextCNN) foi apresentada por Feng *et al.* (2020). A principal vantagem do método apresentado é que se torna possível a utilização de métricas de consumo intradiárias, enquanto nos trabalhos anteriores as métricas eram mensais, semanais ou diárias. Com a metodologia apresentada por Feng *et al.* (2020) é possível, por exemplo, identificar uma anomalia de comportamento em uma determinada hora do dia, o que é útil em algumas modalidades de

fraude. Segundo os resultados apresentados pelos autores, a metodologia apresentou uma melhoria de 35% na classificação dos consumidores, quando comparada com o uso de métricas diárias de consumo.

Ghori *et al.* (2020) realizaram uma comparação entre diferentes algoritmos de classificação aplicados a uma mesma base de dados de consumidores do Paquistão. No total, 14 classificadores diferente foram utilizados para identificar a presença de PNT em um grupo de mais de 80 mil consumidores. A principal conclusão dos autores foi que os algoritmos do tipo *ensemble* apresentam melhor desempenho em comparação com os algoritmos clássicos, como Redes Neurais, Regressão Logística e SVM, com destaque para o *CatBoost* e o *RadomForest*. Os resultados apresentados por Ghori *et al.* (2020) são interessantes e podem ser utilizados para direcionar a escolha de algoritmos em problema de PNT. Contudo, ainda carecem de aperfeiçoamento, pois não foram realizados testes reais e, assim, não é possível saber se os resultados são fidedignos.

Uma abordagem focada no retorno financeiro das ações de inspeção em campo foi apresentada por Massaferro, Martino e Fernandez (2020). Os autores utilizaram uma análise de risco Bayesiana para considerar o retorno financeiro com a energia recuperada nas ações, versos o custo das inspeções. Posteriormente, os consumidores foram selecionados de forma a maximizar o retorno financeiro total. A abordagem foi testada em uma base de dados de consumidores uruguaios. Segundo os resultados apresentados, a abordagem proposta seria capaz de aumentar o retorno financeiro das ações em até 90%. O trabalho de Massaferro, Martino e Fernandez (2020) é muito útil em aplicações reais, pois, propõe a maximização do retorno financeiro das ações. Porém, carece de robustez em sua metodologia, pois, não foram realizados testes de campo, e além disso, para prever a energia recuperada nos consumidores, os autores utilizaram a informação de demanda contratada, o que está disponível apenas para um número pequeno de consumidores e não poderia ser tomada como único parâmetro.

Raggi *et al.* (2020) apresentaram uma técnica de análise de dados para detecção e localização de perdas não técnicas em sistemas de distribuição que possuem *smart meters*. O método proposto é baseado na estimação de estados e propõe-se a lidar com dados não estruturados ou com baixa qualidade (*bad data*). Os autores utilizaram redes reais de 34 e 1.682 barras para simular a presença de PNT na rede. De acordo com os resultados apresentados, o método seria capaz de identificar consumidores com perdas a partir de 2 kW na baixa tensão ou 23 kW na média tensão. A proposta de Raggi *et al.* (2020) pode ser considerada uma contribuição relevante, pois, apresenta um método de estimação de estados específico para a localização de PNT. Contudo, a premissa para utilização do método é que todos consumidores

possuam sistema de *smart meter*, o que ainda não é uma realidade provável no médio prazo para a maioria dos sistemas de distribuição.

Saeed *et al.* (2020) apresentaram uma revisão sistemática das principais soluções propostas para a detecção da perda não técnica de energia entre os anos de 2010 e 2020. O trabalho de Saeed *et al.* (2020) pode ser considerado uma atualização dos trabalhos de Viegas *et al.* (2017) e Messinis e Hatziargyriou (2018), mas se destaca por oferecer listar de forma quantitativa informações que podem ser úteis para condução de novas pesquisas, tais como, os algoritmos utilizados, as métricas mais comuns, os tipos de validação, dentre outras. Saeed *et al.* (2020) também aplicou um novo algoritmo chamado *Boosted* C5.0, o qual é baseado em árvores de decisão, para a identificação de PNT. Os resultados foram apresentados em outra publicação no mesmo ano e, a partir da comparação com algoritmos clássicos, os autores concluem que o algoritmo apresenta um desempenho superior.

Recentemente, Esmael *et al.* (2021) propuseram uma nova abordagem para a tarefa de identificar PNT nos sistemas de energia que ao invés de treinar classificadores, utiliza o conceito de recuperação de informação. No trabalho, são utilizadas Redes Neurais Convolucionais (CNN) para extrair variáveis relevantes de séries de consumo de energia na forma de representações visuais. Posteriormente, essas variáveis são codificadas em assinaturas textuais. Dessa forma, torna-se possível indexar o conjunto de assinaturas e, posteriormente, realizar uma busca por uma assinatura que está associada a um determinado perfil de PNT. Os autores testaram a nova proposta em uma base de consumidores da distribuidora CPFL no Brasil. Os resultados indicaram uma taxa de acerto de 25,4%, que é relativamente baixa quando comparada com a taxa de 19% obtida aleatoriamente na amostra. Ainda assim, a proposta de Esmael *et al.* (2021) pode ser considerada relevante, pois indica uma nova abordagem para identificação da PNT sem a necessidade do treinamento de classificadores, e pode apresentar resultados mais robustos com a continuação da pesquisa.

Até aqui é possível concluir que a maior parte dos trabalhos é baseada em métodos de aprendizagem de máquina supervisionados. Contudo, Sharma e Majumdar (2021) propuseram um método não supervisionado para a identificação de PNT chamado de *Recursive Transform Learning*. O método proposto é conceitualmente simples, a partir de uma série histórica de consumo, é realizada a previsão de consumo para o próximo intervalo de medição. Se o valor medido for muito diferente da previsão, o consumidor é indicado como suspeito de PNT. O método opera online com medições provenientes de um sistema de *smart meter*. O método foi testado em 240 consumidores em Delhi na Índia, e o resultado demonstrou que o desempenho alcançado é similar à de métodos supervisionados. Os autores argumentam que, com o mesmo

desempenho, torna-se vantajoso a utilização de métodos não supervisionados, pois, nesse caso não há necessidade de se possuir uma base de consumidores com a informação prévia de presença de PNT.

Uma análise descritiva para os principais trabalhos relacionados ao tema de perda não técnica foi apresentada nos parágrafos anteriores. A partir desta análise, é possível tomar conhecimento das principais soluções e abordagens que são propostas na bibliografia mundial. De forma complementar ao que foi exposto e com o objetivo de consolidar o estado da arte apresentado para o tema, os trabalhos discutidos estão sumarizados na Tabela 5 com as suas principais contribuições.

Tabela 5 – Resumo com as principais contribuições de cada pesquisa sobre o tema em estudo.

Trabalho	**Principal Contribuição**
Dick (1995)	Relatou o furto de energia no Reino Unido e as principais ações tomadas pelas companhias de distribuição na época.
Ghajar, Khalife e Richani (2002)	Propôs uso de medidores eletrônicos e leitura remota com RF
Eller (2003)	Utilizou RNA com informações do histórico de consumo e do IPTU
Smith (2004)	Propôs o uso de caixas de medição blindadas, privatização das concessionárias e mudanças na legislação
Nizar, Dong e Wang (2008)	Utilizou o ELM em comparação ao SVM para identificar degrau de consumo
Penin (2008)	Fez uma revisão das melhores práticas para a gestão da perda não técnica e utilizou RNA para identificar degrau de consumo
Nagi *et al.* (2010)	Utilizou o SMV para identificar degrau de consumo e adicionou uma etapa de análise manual para tomada de decisão final
Angelos *et al.* (2011)	Propôs a classificação dos consumidores com base na lógica *Fuzzy* e na mudança do perfil de consumo histórico
León *et al.* (2011)	Realizou a classificação dos consumidores com base no método GRI e métricas de variabilidade e tendência da série de consumo
Nagi *et al.* (2011)	Substituiu a etapa manual pelo algoritmo *Fuzzy Inference System* (FIS), em que as regras são implementadas com base no conhecimento tácito dos usuários.
Ramos *et al.* (2011)	Utilizou o *Optimum-Path Forest* (OPF) como alternativa às técnicas tradicionais de classificação
Bastos (2011)	Utilizou Redes Bayesianas e investigação de conformidade para fazer o diagnóstico da PNT conforme causas e região de incidência
Mondero *et al.* (2012)	Aplicou o coeficiente de Pearson e Redes Bayesianas para identificar tipos distintos de fraude
Huang, Lo e Lu (2013)	Utilizou dados de *smart meter* e estimação de estados para comparar a curva de carga de transformadores com as medições dos consumidores associados ao equipamento
Guerrero *et al.* (2014)	Aplicou o KBS para implementar um sistema de detecção de perda não técnica baseada no conhecimento de especialistas
Amin *et al.* (2015)	Propôs a aplicação da Teoria dos Jogos para dimensionamento de um sistema de *smart meter* com viés de redução de perda não técnica

Trabalho	Principal Contribuição
Xu (2015)	Analisou os sinais de corrente, tensão e potência em cada fase para identificar condições anormais na medição relacionadas a PNT
Jokar, Arianpoo e Leung (2015)	Detectou a perda não técnica com base em comportamentos anômalos da curva de carga diária medida nos consumidores
Jindal *et al.* (2016)	Propôs um esquema baseado em *smart meter* e Árvores de Decisão para identificar a PNT em tempo real em diferentes níveis
Ramos *et al.* (2016)	Aplicou o *Black Hole Algorithm* (BHA) para selecionar o melhor conjunto de variáveis para caracterização da PNT nos consumidores
Leite e Mantovani (2016)	Utilizou o algoritmo A-Star para localizar regiões com maior ocorrência de perda não técnica, baseado em um gráfico de controle multivariado.
Zhan *et al.* (2016)	Utilizou o balanço energético de medições de *smart meter* para localizar simultaneamente PNT e faltas na rede, com o objetivo de aumentar a viabilidade financeira das soluções.
Madrigal, Rico e Uzcaregui (2017)	Apresentou um método para estimar a PNT no sistema elétrico com a necessidade de poucas medições em pontos aleatórios do sistema
Zanetti (2017)	Utilizou medições no secundário dos transformadores de distribuição para comparar com o consumo dos clientes associados ao equipamento
Viegas, Esteves e Vieira (2018)	Utilizou o método *Gustafson-Kessel fuzzy* para definir perfis de consumo regulares e definiu a probabilidade de cometimento de fraude com base na distância do indivíduo para o centroide do grupo
Singh, Bose e Joshi (2018)	Monitorou a dinâmica do consumo por meio da entropia relativa das distribuições de probabilidade do histórico de consumo
Aryanezhad (2019)	Propôs um sistema para alterar a tensão da rede e prejudicar o funcionamento de equipamentos de clientes irregulares
Punmiya e Choe (2019)	Calculou métricas derivadas da curva de consumo diário em um processo de *Feature Engineering* para melhorar a caracterização da PNT
Alves (2019)	Avaliou a contribuição de métricas derivadas do histórico de consumo mensal para melhorar o desempenho de modelos preditivos
Glauner (2019)	Criou um sistema de realidade virtual para seleção de consumidores com suspeita de fraude
Aslam *et al.* (2020)	Apresentou o método IQMOT para o balanceamento de amostras baseado nos valores interquartis das métricas de consumo.
Calvo *et al.* (2020)	Propôs a segmentação prévia dos consumidores de acordo com o tipo de fraude cometida como forma de melhorar o resultado dos classificadores.
Feng *et al.* (2020)	Apresentou uma abordagem que utiliza métricas de consumo intradiárias, permitindo a identificação de anomalias em determinadas horas do dia.
Ghori *et al.* (2020)	Realizou uma comparação entre diferentes algoritmos para identificação da PNT e conclui que os métodos baseados em *ensemble* são os mais apropriados para o problema.
Massaferro, Martino e Fernandez (2020)	Apresentou uma abordagem para maximizar o retorno financeiro das ações de inspeção baseada em análise Bayesiana.
Raggi *et al.* (2020)	Apresentou um método para identificação de PNT baseada em medições de *smart meter* e estimação de estados.
Saeed *et al.* (2020)	Realizou uma revisão sistemática de várias soluções para identificação de PNT e aplicou um novo algoritmo ao problema chamado Boosted C5.0, o qual é baseado em árvores de decisão.
Esmael *et al.* (2021)	Utilizou CNNs para extrair métricas visuais das séries de consumo, que posteriormente são codificadas na forma de texto e utilizadas para criar assinaturas associadas a perfis de PNT.

Trabalho	Principal Contribuição
Sharma e Majumdar (2021)	Propôs um método não supervisionado para a identificação de PNT, o qual consiste em comparar o valor de energia medido com um *forecast* obtido previamente.

Fonte: Autoria própria.

3.3 Análise Qualitativa

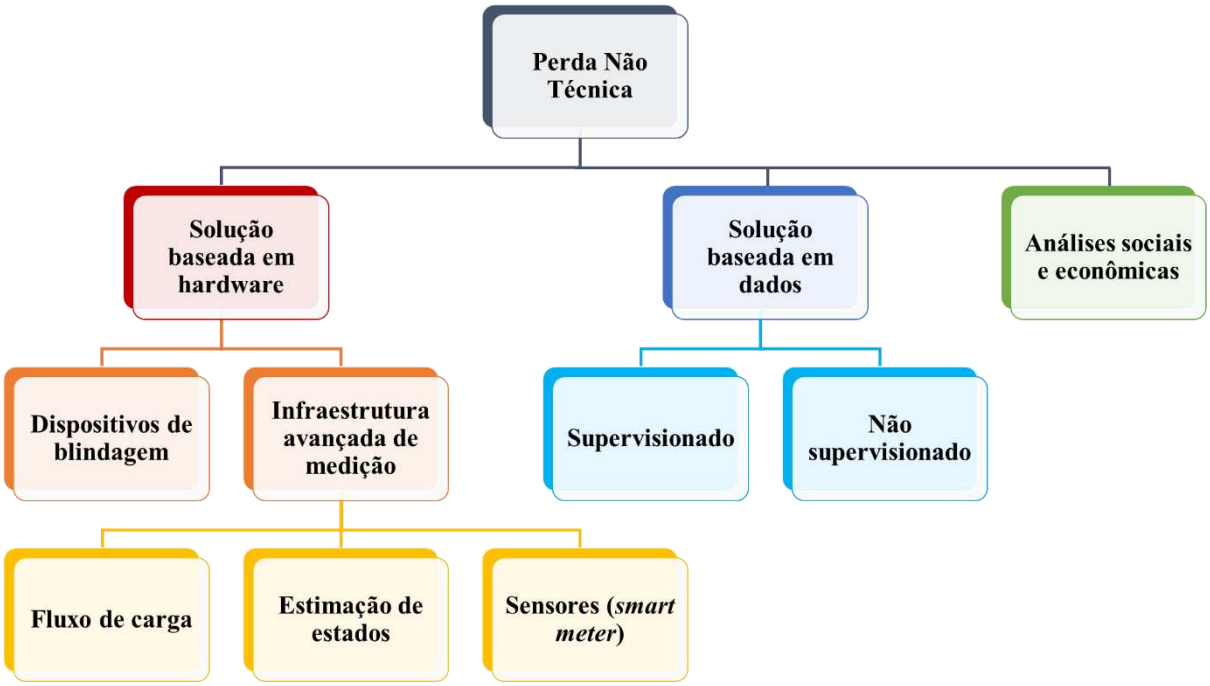
Diante do exposto nas subseções anteriores, é possível constatar que o problema da perda não técnica de energia no sistema de distribuição tem motivado pesquisas por todo o mundo de forma crescente nos últimos anos. Inicialmente as pesquisas consistiam em utilizar informações relacionadas ao histórico de consumo para aplicar técnicas de aprendizagem de máquina com objetivo de encontrar padrões relacionados à ocorrência de perda não técnica. Diferentes técnicas foram aplicadas, destacando-se como principais o *Support Vector Machine*, Árvores de Decisão, Redes Neurais Artificiais, Regressão Logística e Redes Bayesianas. Em geral a proposta dos trabalhos é utilizar uma única técnica como solução para classificação dos consumidores. Alguns trabalhos propuseram novos algoritmos e utilizavam a comparação com os clássicos para validarem os resultados.

Na sequência, com o surgimento da infraestrutura avançada de medição, os trabalhos começaram a se basear em dados de sistemas de *smart meters*. Em geral, a proposta é utilizar medições em vários pontos da rede para segregar a perda em segmentos, por meio do balanço energético entre as medições. Alguns autores propuseram o monitoramento em tempo real das medições, para detectar a fraude no instante em que ela se inicia. Nos trabalhos mais recentes, têm se buscado soluções que possam diminuir o custo da infraestrutura avança de medição, como a estimação de estados, além de técnicas mais avançadas de análise de dados.

Como visto nos parágrafos anteriores, diferentes abordagens podem ser identificadas em cada pesquisa, cada qual com suas vantagens e desvantagens. Com o objetivo de resumir as diferentes abordagens, uma categorização de métodos é apresentada na Figura 31, na qual a maior parte dos trabalhos pode ser classificado em uma ou mais categoria.

A partir da análise crítica realizada nos trabalhos que compõe o estado da arte no tema, é possível destacar algumas das principais vantagens e desvantagens para os métodos apresentados na Figura 31. Com relação às soluções baseadas em *hardware*, as principais vantagens são a possibilidade de detectar a ocorrência de PNT em tempo real e a elevada taxa de acerto das indicações.

Figura 31 – Métodos apresentados na revisão bibliográfica para abordagem da perda não técnica.



Fonte: Autoria própria.

Por outro lado, a principal desvantagem das soluções baseadas em *hardware* é o custo de aquisição e manutenção da infraestrutura necessária para implementação das soluções. Na maior parte dos casos, o custo inviabiliza a adoção das soluções em larga escala. Por essa razão a maior parte dos trabalhos apresentam apenas simulações ou resultados de provas de conceitos em um pequeno grupo de consumidores. Pode-se afirmar que, até o momento, a infraestrutura avançada de medição não se viabiliza economicamente apenas aplicações de controle da PNT, sendo necessário que outras funcionalidades sejam agregadas à tecnologia para que ela se torne atrativa no futuro.

Com relação as soluções baseadas em dados, as principais vantagens são o baixo custo de aquisição e a facilidade de implementação em larga escala, dado que a maior parte dos trabalhos utilizam informações que já são acessíveis para as concessionárias. Por outro lado, a principal desvantagem está associada às taxas de acerto relativamente baixas das indicações que são alcançadas pelos algoritmos de classificação. Parte do problema está relacionada à baixa qualidade dos dados disponíveis, a característica de alto desbalanceamento entre o número de amostras de consumidores com PNT e sem PNT, e ao viés de seleção que é inerente ao problema, uma vez que no histórico de informações disponível, os consumidores não foram inspecionados de maneira aleatória, mas sim direcionados a partir de algum critério adotado pelas concessionárias.

Com relação aos trabalhos que realizam análises sociais e econômicas, a principal vantagem dos estudos é a oferta de uma percepção mais abrangente dos impactos da perda não técnica de energia para as sociedades, que vai desde o aspecto econômico até o cultural. Como principal desvantagens desses trabalhos, destaca-se o fato de que na maioria das vezes não oferecem soluções concretas para o problema.

Com relação à localização do presente trabalho no diagrama da Figura 31, o mesmo pode ser classificado como uma solução baseada em dados, que utiliza predominantemente métodos supervisionados de aprendizagem de máquina. De forma complementar e com o objetivo de posicionar as contribuições do presente trabalho entre aquelas disponíveis na bibliografia, é apresentado na Tabela 6 um quadro sinóptico com as técnicas aplicadas por cada autor para identificação da perda não técnica de energia elétrica.

Tabela 6 – Quadro sinóptico com as principais abordagens utilizadas em cada pesquisa.

Trabalho	Inferência Causal	Superv.	Não Superv.	Multi-algoritmo	Otimização financeira	Validação em campo	Outros
Dick (1995)							X
Ghajar, Khalife e Richani (2002)							X
Eller (2003)		X					
Smith (2004)							X
Nizar, Dong e Wang (2008)		X					
Penin (2008)		X					
Nagi *et al.* (2010)		X				X	X
Angelos *et al.* (2011)			X				
León *et al.* (2011)			X				X
Nagi *et al.* (2011)		X	X			X	
Ramos *et al.* (2011)		X					
Bastos (2011)		X					
Mondero *et al.* (2012)		X				X	
Huang, Lo e Lu (2013)							X
Guerrero *et al.* (2014)			X			X	X
Amin *et al.* (2015)		X					X
Xu (2015)							X
Jokar, Arianpoo e Leung (2015)							X
Jindal *et al.* (2016)		X					X
Ramos *et al.* (2016)			X				
Leite e Mantovani (2016)							X
Zhan *et al.* (2016)							X
Madrigal, Rico e Uzcaregui (2017)		X					X

Trabalho	Inferência Causal	Superv.	Não Superv.	Multi-algoritmo	Otimização financeira	Validação em campo	Outros
Zanetti (2017)							X
Viegas, Esteves e Vieira (2018)			X				
Singh, Bose e Joshi (2018)			X				X
Aryanezhad (2019)							X
Punmiya e Choe (2019)			X				
Alves (2019)			X				
Glauner (2019)							X
Aslam *et al*. (2020)		X					
Calvo *et al.* (2020)			X			X	
Feng *et al.* (2020)		X					X
Ghori *et al.* (2020)				X			
Massaferro, Martino e Fernandez (2020)		X			X		
Raggi *et al.* (2020)							X
Saeed *et al.* (2020)		X					
Esmael *et al*. (2021)		X					X
Sharma e Majumdar (2021)			X				
Presente pesquisa	X	X	X	X	X	X	

Fonte: Autoria própria.

A partir da análise da Tabela 6, nota-se que nenhuma pesquisa abordou o problema da perda não técnica com as mesmas técnicas utilizadas na presente pesquisa simultaneamente. Destaca-se em especial o uso da Inferência Causal, que não foi utilizada em nenhuma pesquisa, a otimização financeira e a abordagem multi-algoritmo que foram utilizadas em apenas uma pesquisa cada, e a validação em campo que foi realizada em quantidade mínima de trabalhos. Logo, torna-se evidente que a abordagem oferecida pela presente pesquisa representa uma evolução em relação os trabalhos desenvolvidos até o momento.

3.4 Principais Lacunas

A partir do estado da arte exposto neste capítulo, é possível extrair contribuições relevantes para a gestão da perda não técnica no sistema elétrico de distribuição. Contudo, ainda não está disponível uma solução definitiva para o problema, pois as soluções propostas ainda apresentam limitações e pontos de melhoria que motivam a realização de novos trabalhos. Dentre as principais limitações, podem ser destacadas as seguintes:

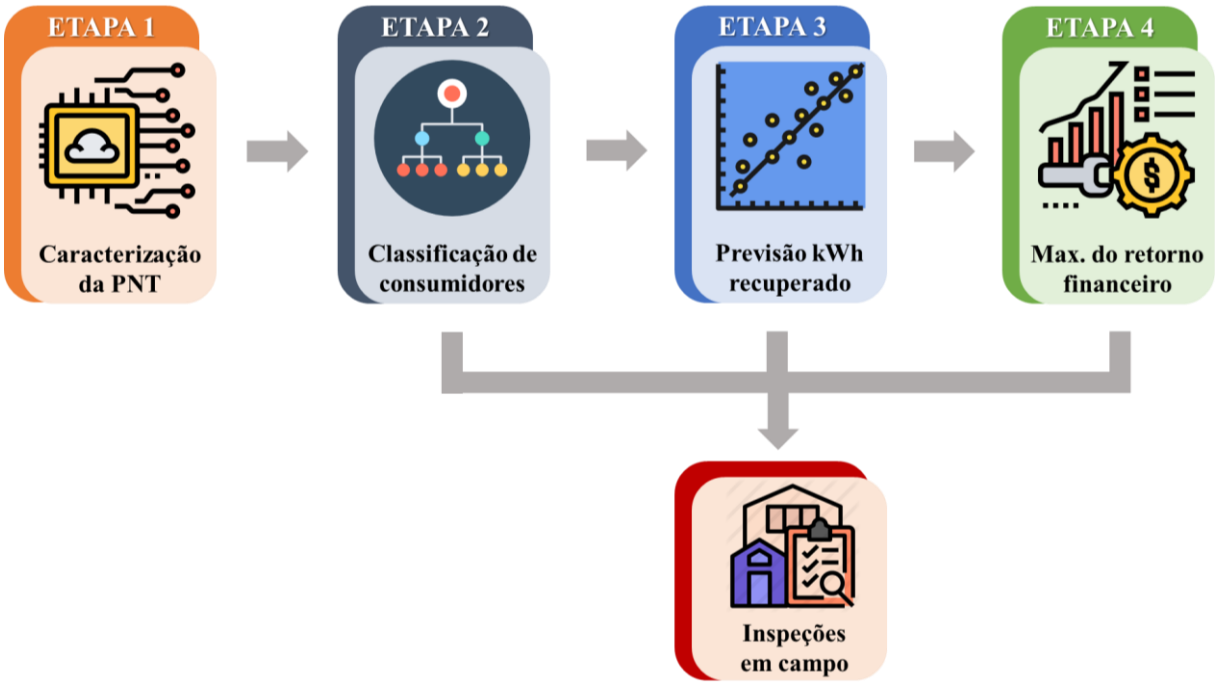
- Número elevado de falsos positivos resultantes dos métodos de aprendizagem de máquina;
- Custo elevado de implementação para as soluções baseadas em *smart meter*;
- Falta de utilização de métricas não relacionadas ao histórico de consumo;
- Falta de generalização das soluções, pois, a maioria tem o objetivo de identificar um tipo específico de fraude, geralmente o que apresenta queda no histórico de consumo;
- Falta de padronização na avaliação do desempenho dos trabalhos, o que dificuldade a comparação dos resultados de diferentes técnicas;
- Utilização de métricas inadequadas para a avaliação dos modelos, o que pode levar a uma conclusão de sucesso errônea;
- Ausência de soluções multi-algoritmos, pois, em boa parte dos trabalhos o objetivo é validar a aplicação de um algoritmo específico frente aos demais;
- Ausência de testes em campo com quantidade relevante de consumidores para validação das soluções propostas em aplicações reais;
- Falta de foco no retorno financeiro das ações de gestão da perda não técnica.

Nota-se, portanto, que apesar das várias soluções propostas ainda existem oportunidades de melhorias que podem ser exploradas em novas pesquisas para se obter soluções mais avançadas para a gestão da perda não técnica de energia. Neste contexto, o presente trabalho se propõe a contribuir para o aprimoramento da gestão da perda não técnica por meio da metodologia que é apresentada na seção seguinte.

4 Metodologia

Neste capítulo são apresentados todos os detalhes da metodologia utilizada para o desenvolvimento e a obtenção dos resultados da pesquisa. Para facilitar o entendimento do leitor, a metodologia desenvolvida no trabalho foi dividida em 4 etapas principais, as quais estão esquematizadas no fluxograma da Figura 32.

Figura 32 – Principais etapas para o desenvolvimento da pesquisa.



Fonte: Autoria própria.

Como pode ser visto na Figura 32, a primeira etapa da pesquisa consiste na caracterização da perda não técnica, o que é feito a partir da extração de dados reais de uma concessionária de energia, seguido do pré-processamento dos dados, do processo de *feature engineering* e, por fim, da identificação dos fatores de risco e de proteção para ocorrência da PNT por meio da análise de Inferência Causal. Os procedimentos da Etapa 1 são descritos em detalhes na seção 4.1.

Na Etapa 2 os fatores de risco e proteção para a PNT são utilizados como variáveis de entrada em 23 modelos distintos de classificação. Os diferentes modelos são avaliados quanto ao seu desempenho e ao seu custo computacional para identificação da PNT entre os consumidores da base de dados coletada. Na avaliação dos modelos são comparados os

resultados do processo de validação cruzada com inspeções em campo realizadas a partir das indicações dos modelos. Os procedimentos da Etapa 2 estão descritos em detalhes na seção 4.2.

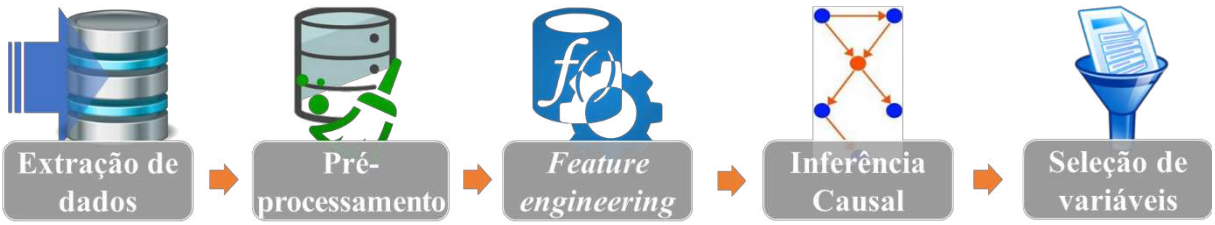
Na Etapa 3 as variáveis resultantes da Etapa 1 são avaliadas quanto o seu grau de associação com a energia recuperada nas inspeções em campo. As variáveis mais associadas são selecionadas como entradas de sete modelos distintos de regressão. Os diferentes modelos são avaliados quanto à qualidade da previsão dos valores de energia recuperada em inspeções em campo. Assim como na Etapa 2, são comparados os resultados do processo de validação cruzada com inspeções em campo realizadas em consumidores reais. Os procedimentos da Etapa 3 são apresentados de forma detalhada na seção 4.3.

Por fim, na Etapa 4 é proposto um modelo para a maximização do retorno financeiro das inspeções. Os resultados dos melhores modelos preditivos identificados nas Etapas 2 e 3 são utilizados para construção de uma equação matemática, cujo objetivo é prever o potencial de retorno financeiro de uma inspeção em campo para cada consumidor da base dados. A partir do potencial de retorno financeiro, é identificado o conjunto ótimo de consumidores que devem ser inspecionados em campo, para garantir o máximo retorno financeiro à concessionária. O modelo proposto na Etapa 4 também é validado a partir de novas inspeções em campo.

4.1 Etapa 1: Caracterização da PNT

A caracterização da PNT tem como objetivo identificar, dentre as informações disponíveis para uma concessionária de energia, quais são os fatores de risco e os fatores de proteção para ocorrência de perda não técnica em um consumidor. Para tanto, são executados os procedimentos ilustrados na Figura 33.

Figura 33 – Etapas do processo de caracterização da perda não técnica de energia.

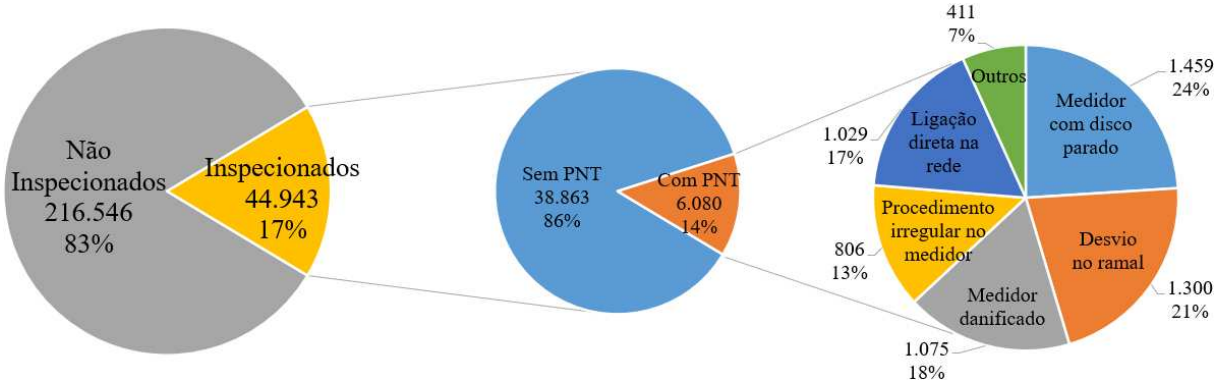


Fonte: Autoria própria.

Como apresentado na Figura 33, inicialmente é realizada a extração de dados dos sistemas da concessionária. Para realização da pesquisa, foram obtidos dados referentes a um município da região centro-oeste brasileira que possui 261.489 unidades consumidoras. Dentre

os consumidores presentes na base de dados, 17% já haviam passado por um processo de inspeção nos últimos dois anos, como apresentado na Figura 34. O período de dois anos para verificação das inspeções foi considerado por representar um período em que há maior confiabilidade nas informações disponíveis na base dados, registros antigos apresentam uma quantidade significativa de inconsistências, que a concessionária atribui a uma migração de sistemas ocorrida no passado.

Figura 34 – Características da base de dados coletada.



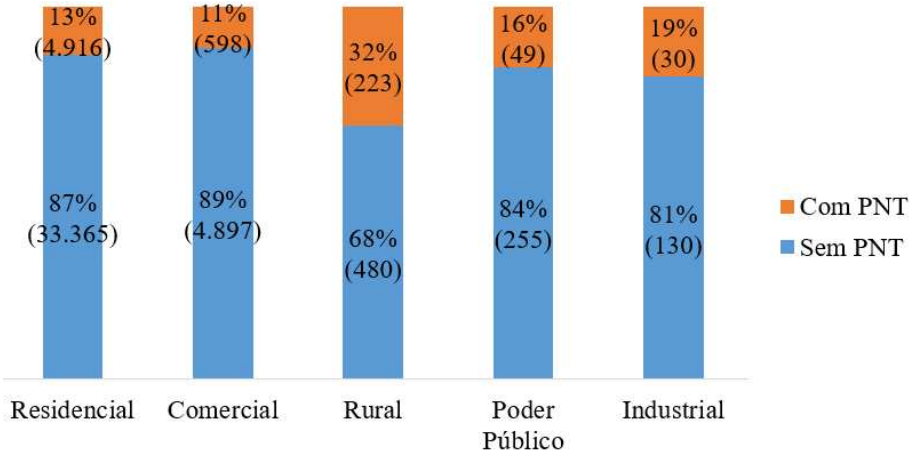
Fonte: Autoria própria.

Com pode ser visto na Figura 34, dentre os 44.943 consumidores que já foram inspecionados, 86% não apresentavam irregularidades no sistema de medição, enquanto em 14% foi identificado a presença de PNT. O índice de 14% representa, portanto, a taxa de sucesso praticada pela concessionária nos seus procedimentos de inspeção em campo.

Ainda na Figura 34, é possível observar os principais tipos irregularidades encontradas nos consumidores com perda não técnica. Os casos de medidores com o disco parado, que representam 24% do total, podem ser causados tanto por falha técnica do medidor quanto por intervenções indevidas do consumidor. Os desvios no ramal de ligação representam 21% do total e são muito utilizados pela sua simplicidade de execução, como exemplificado na Figura 5 (c). Os danos e procedimentos irregulares nos medidores representam 18% e 13% do total respectivamente. Esses procedimentos são executados pelos consumidores com objetivo de adulterar o funcionamento do medidor, um exemplo desse procedimento é apresentado na Figura 5 (b). As ligações diretas na rede, em que o consumo não é registrado por um medidor, representam 17% do total, um exemplo dessa irregularidade é apresentado na Figura 5 (a). Os demais 7% dos casos de perda não técnica são referentes a outros tipos de irregularidade, como o isolamento do condutor neutro, a inversão de fases na ligação com o medidor e problemas nos transformadores de instrumentos.

Ainda com relação às características da base de dados obtida, na Figura 35 é apresentada a ocorrência de PNT por classe de consumo.

Figura 35 – Ocorrência de perda não técnica por classe de consumo na base de dados coletada.



Fonte: Autoria própria.

Como observado na Figura 35, a classe Rural se destaca por apresentar quase um terço dos consumidores inspecionados com perda não técnica. A distribuidora atribui o índice mais elevado devido ao pouco número de ações que são realizadas nas áreas rurais em função da dificuldade de acesso e baixa densidade populacional, o que criaria a sensação de liberdade nos consumidores para prática das irregularidades. Apesar do maior percentual de perda não técnica, a classe Rural representa apenas 7,8% do faturamento de energia da distribuidora, o que pode contribuir para o pouco número de ações da distribuidora nas áreas rurais.

Além das informações referentes às inspeções em campo realizada nos consumidores, foram coletados todos os dados disponíveis nos sistemas comerciais, técnicos e operacionais da concessionária. No total, 12 fontes diferentes de dados foram utilizadas. Cada fonte contém um tipo diferente de informação, conforme descrito na Tabela 7.

Tabela 7 – Extração de dados nos sistemas da distribuidora de energia.

Fonte	Qtd. de registros	Descrição
1	$> 50 \cdot 10^6$	Histórico de interações realizadas nos canais de atendimento ao cliente, como sítio, telefone, aplicativo e redes sociais.
2	$> 21 \cdot 10^6$	Histórico de consumo mensal de energia.
3	$> 17 \cdot 10^6$	Histórico do processo de leitura mensal do medidor, incluindo observações dos leituristas.
4	$> 49 \cdot 10^6$	Histórico do processo de faturamento mensal, incluindo as observações feita pela concessionária para cada fatura.
5	$> 26 \cdot 10^6$	Histórico das datas de pagamento para cada fatura de consumo mensal.
6	$> 48 \cdot 10^6$	Histórico das datas de vencimento para cada fatura de consumo mensal.
7	$> 14 \cdot 10^6$	Histórico de ordens de serviço executadas em cada consumidor.

Fonte	Qtd. de registros	Descrição
8	$> 595 \cdot 10^3$	Histórico de inspeções *in loco* executadas em consumidores. Este dado indica a presença ou não de perda não técnica nos consumidores no momento das inspeções.
9	$> 128 \cdot 10^3$	Histórico de pré-inspeções executadas em cada consumidor. Uma pré-inspeção consiste em uma inspeção visual realizada no sistema de medição por um técnico, que posteriormente pode indicar a necessidade de realização de uma inspeção *in loco*.
10	$> 106 \cdot 10^3$	Histórico de ocorrências de perdas não técnicas. Fornece informações como a quantidade de energia recuperada pela concessionária após a detecção de uma fraude.
11	$> 232 \cdot 10^3$	Histórico das obras de regularização. Uma regularização representa algum tipo de equipamento que foi instalado no sistema de distribuição para evitar fraudes.
12	$> 284 \cdot 10^6$	Informações do cadastro geral do consumidor. Inclui informações como endereço, tipo de conexão, nível de tensão e modelo do medidor. Inclui também o histórico de mudanças cadastrais dos consumidores.

Fonte: Autoria própia.

Uma vez que os dados apresentados na Tabela 7 são obtidos, faz-se necessário executar a etapa de pré-processamento. O objetivo do pré-processamento é corrigir erros que são comumente encontrados em bancos de dados que armazenam informações transacionais, como no caso das distribuidoras de energia. Também objetiva-se obter um formato de dados que seja mais adequado para utilização em modelos de Aprendizagem de Máquina. Todos os procedimentos executados no pré-processamento dos dados são descritos na subseção 2.2.1.

Seguindo o fluxo indicado na Figura 33, o próximo passo é a execução do procedimento de *feature engineering*, o qual consiste em obter novas variáveis a partir das variáveis originalmente existentes no conjunto de dados. A ideia é que as novas variáveis obtidas sejam mais explicativas em relação à ocorrência da perda não técnica. Maiores informações sobre o processo de *feature engineering* são fornecidas na subseção 2.2.2.

Após a etapa de *feature engineering*, é realizada a análise de Inferência Causal, cujo objetivo principal é identificar variáveis que estejam associadas de maneira possivelmente causal com a ocorrência de perda não técnica, indicando quais são fatores de risco e quais são fatores de proteção. O tipo de amostra de dados obtida no problema o caracteriza como um estudo de caso-controle, conforme descrito na subseção 2.2.3.1. Os consumidores considerados no estudo caso-controle são aqueles que já passaram por um procedimento de inspeção *in-loco* registrado no banco de dados coletado. Dentre estes consumidores, os "casos" são aqueles em que foi identificado a presença de perda não técnica, enquanto os "controles" são os consumidores sem a presença de perda não técnica.

A partir da separação dos "casos" e dos "controles", é analisado o histórico de informações dos últimos dois anos para determinar a associação das variáveis existentes com os "casos". O grau de associação de cada variável é determinado pelas medidas *Odds Ratio* para as variáveis categóricas e Média Confiável, com coeficiente de confiança de 95%, para as

variáveis numéricas. Aquelas variáveis que não demonstram possuir associação com a ocorrência dos “casos” são eliminadas do conjunto de dados.

Posteriormente, o coeficiente de correlação de Spearman (ρ) é calculado entre as variáveis restantes com o objetivo de eliminar a redundância do conjunto. Os pares de variáveis que possuem $|\rho| \geq 0{,}7$ são considerados redundantes. Nesse caso, é feita a eliminação da variável que possui menor associação com a ocorrência da perda não técnica. Por fim, é calculada a explicabilidade (ϵ) do conjunto de dados resultante, conforme Equação (9). O valor de ϵ indica o quanto da ocorrência da perda não técnica de energia pode ser descrito pelo conjunto de variáveis obtidas na análise de Inferência Causal.

As variáveis resultantes dos procedimentos descritos anteriormente são avaliadas como fatores de risco ou de proteção para ocorrência da perda não técnica, a depender dos valores das medidas *Odds Ratio* e Média Confiável obtidos para cada uma. São considerados fatores de risco para ocorrência da perda não técnica as variáveis cujo *Odds Ratio* é maior que 1,0 ou a Média Confiável é maior no grupo “casos” do que no grupo “controles”. São considerados fatores de proteção para ocorrência de perda não técnica as variáveis cujo *Odds Ratio* é menor que 1,0, ou a Média Confiável é maior no grupo “controles” do que no grupo “casos”.

O conjunto de dados obtidos a partir dos procedimentos descritos nesta seção é utilizado para o treinamento e a validação de modelos preditivos de classificação e regressão, como detalhado nas seções seguintes.

4.2 Etapa 2: Classificação de Consumidores

A etapa de classificação de consumidores tem como objetivo identificar, a partir das variáveis obtidas na Etapa 1, quais são os consumidores que possuem perda não técnica de energia. Para tanto, são utilizados diferentes modelos preditivos de classificação, bem como, diferentes abordagens para o tratamento das informações disponíveis.

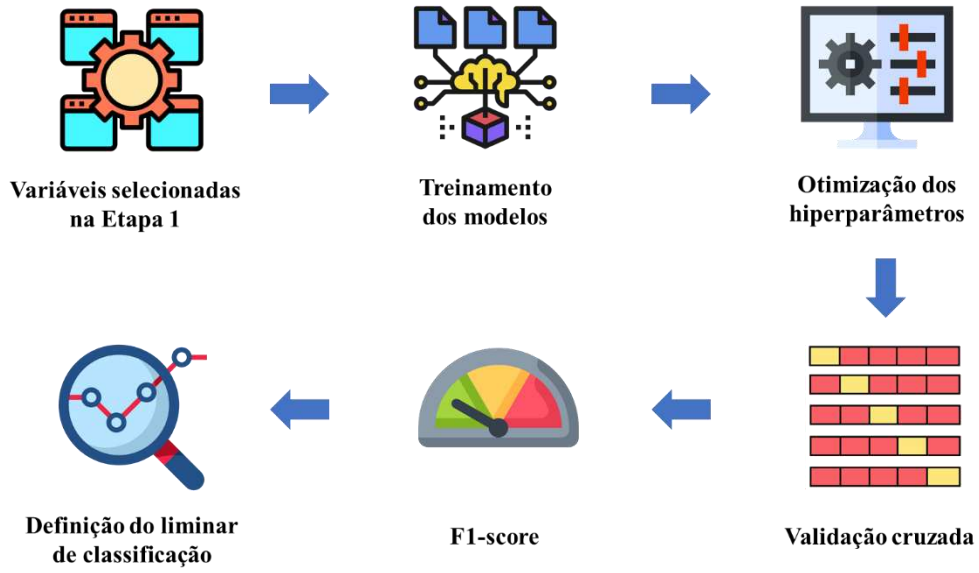
Inicialmente, os algoritmos descritos na subseção 2.2.5 são aplicados de forma individual. Posteriormente, o método *x-means* descrito na subseção 2.2.4 é utilizado para realizar uma segregação preliminar na base de dados. Por fim, é realizado a classificação dos consumidores a partir de um critério de votação entre todos os modelos construídos. Os resultados dos diferentes modelos e abordagens são avaliados de acordo com seu desempenho e seu custo computacional. Para tanto são realizadas validações cruzadas com o método *k-fold* descrito na subseção 2.2.8, bem como, inspeções reais em campo para novos consumidores do

sistema de distribuição. Uma descrição detalhada de cada abordagem é apresentada nas subseções seguintes.

4.2.1 Classificadores Individuais

Uma visão geral das etapas executadas no processo de classificação com algoritmos individuais é apresentada na Figura 36 e uma descrição detalhada é fornecida na sequência.

Figura 36 – Procedimentos executados para a classificação dos consumidores com classificadores individuais.



Fonte: Autoria própria.

Conforme a Figura 36, inicialmente o conjunto de variáveis obtidos na Etapa 1 é utilizado para treinar o conjunto de 11 classificadores apresentados na Tabela 4. Os classificadores foram selecionados por representarem o conjunto de algoritmos mais utilizados nos trabalhos relacionados na bibliografia correlata ao tema, conforme discussões na seção 3.2. A exceção são os algoritmos *Rotation Forest* (RTF) e *Bagging Tree* (BT), os quais ainda não foram aplicados para problemas de perda não técnica de energia e representam, portanto, uma contribuição desta pesquisa. Os dois algoritmos foram adicionados devido às suas características serem semelhantes ao *Random Forest* (RT), porém, com relatos de desempenho superior em algumas aplicações de classificação de dados, conforme discussão da subseção 2.2.5.

Seguindo a sequência da Figura 36, o algoritmo de otimização Bayesiana é utilizado para determinar os hiperparâmetros mais adequados em cada classificador. Os detalhes da otimização Bayesiana são discutidos na subseção 2.2.6.1 e os parâmetros utilizados no processo são apresentados na Tabela 8.

Tabela 8 – Parâmetros utilizados no algoritmo de otimização Bayesiana.

Parâmetro	Valor
Quantidade de iterações de aquecimento	30
Quantidade máxima de iterações	100
Taxa de divisão (γ)	0,2
Quantidade de candidatos por rodada	100

Fonte: Autoria própria.

Os valores na Tabela 8 foram escolhidos de acordo as recomendações de Bergstra *et al.* (2011). Além dos parâmetros descritos na Tabela 8, para cada classificador foram definidos os espaços de hiperparâmetros no qual é realizada a busca Bayesiana para definição do valor ótimo, conforme apresentado na Tabela 9.

Tabela 9 – Espaços de hiperparâmetros utilizados na otimização Bayesiana.

Modelo	Hiperparâmetro	Limite inferior	Limite superior
LR	Número máximo de iterações	1.000	
	Tolerância (*epsilon*)	10^{-5}	
DT	Número mínimo de derivações por nó	2	
	Critério para derivação	*Gini index*	
NB	Limiar de classificação	0,0	1,0
	Desvio padrão mínimo	0,0	1,0
	Desvio padrão mínimo do limiar de classificação	0,0	1,0
MLP	Número de camadas escondidas	1	10
	Número de neurônios por camada	2	35
SVM	Potência da função *kernel*	1	2
	Coeficiente da função *kernel*	-1,0	1,0
FR	Norma triangular (*T-norm*)	mín/máx; produto; Lukasiewicz; ou Yager	
GBT	Número de árvores	20	200
	Número de níveis das árvores	2	20
XGBT	Número de árvores	50	500
	Número de níveis das árvores	2	10
	Número máximo de *bins*	100	500
BT	*Bag size*	10%	100%
RF	Número de árvores	100	500
RTF	Tamanho do grupo de variáveis	2	10
	Taxa de remoção	20%	80%

Fonte: Autoria própria.

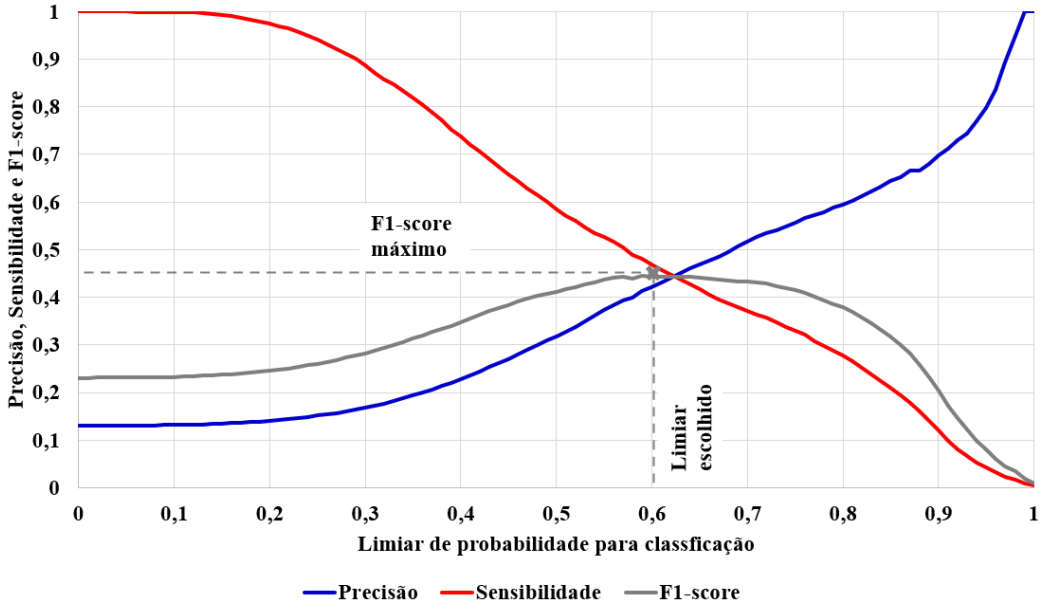
Os limites inferiores e superiores apresentados na Tabela 9 foram definidos manualmente de acordo com a experiência do autor. Além disso, os classificadores LR e DT não requerem ou são pouco influenciados pela otimização de hiperparâmetros, por isso, os hiperparâmetros destes classificadores foram definidos manualmente de acordo com a experiência do autor.

Na sequência, utilizando os hiperparâmetros ótimos definidos na otimização Bayesiana, os classificadores são testados a partir do processo de validação cruzada com o método *k-fold*, o qual foi discutido em detalhes na subseção 2.2.8. No método *k-fold* o valor de *k* é definido como igual a 10, com objetivo de equilibrar a relação de compromisso entre custo computacional e a generalização dos resultados.

Os resultados da validação cruzada de cada modelo são utilizados para construir as respectivas matrizes de confusão, conforme descrição na subseção 2.2.7.1. Além disso, também é registrado o tempo de processamento de cada modelo, com objetivo de estabelecer uma comparação do custo computacional de cada um.

A partir da matriz de confusão, é obtido o valor da métrica F1-*score* para cada modelo, a qual é descrita na subseção 2.2.7.1. Destaca-se que o valor de F1-*score* depende do limiar de probabilidade escolhido para classificação no modelo. Desse modo, é possível avaliar o resultado para diferentes limiares de classificação e, posteriormente, selecionar o valor que corresponda ao F1-*score* máximo. Esta análise de sensibilidade é ilustrada na Figura 37.

Figura 37 – Análise de sensibilidade do limiar de probabilidade para classificação de um modelo preditivo em função da métrica F1-*score*.



Fonte: Autoria própria.

Como pode ser verificado no exemplo da Figura 37, o valor do limiar de probabilidade que corresponde ao maior F1-*score* é 0,6. Caso o valor de 0,5 fosse selecionado, o que ocorre na maioria dos estudos que não realizam a análise de sensibilidade, o valor do F1-*score* seria 8% menor. Além disso, o valor da precisão do modelo seria 25% menor, enquanto a sensibilidade seria 25% maior.

No exemplo descrito na Figura 37 fica evidente como a maximização da métrica F1-*score* resulta em um equilíbrio entre a precisão e a sensibilidade do modelo preditivo, o que é fundamental para o problema de identificação da perda não técnica em campo. Fica evidente também que a métrica precisão sozinha não é capaz de informar sobre a qualidade do classificador. Pois, a precisão do modelo pode ser facilmente elevada escolhendo-se um limiar de classificação mais alto. Contudo, nesse caso a sensibilidade do modelo seria muito baixa, o que significa que ele não teria uma cobertura satisfatória dos casos de PNT existentes. Em outras palavras, apenas um pequeno grupo de consumidores com alta probabilidade de PNT seria selecionado, deixando a maioria dos casos de PNT existentes fora da classificação.

O destaque do parágrafo anterior se faz necessário, pois em muitos trabalhos na bibliografia são relatados estudos em que se obteve altos valores de precisão na identificação da perda não técnica, porém, não são informados os valores de sensibilidade dos modelos. O que pode indicar que os modelos possuem uma baixa cobertura e não seriam adequados para utilização em casos reais por concessionárias de energia.

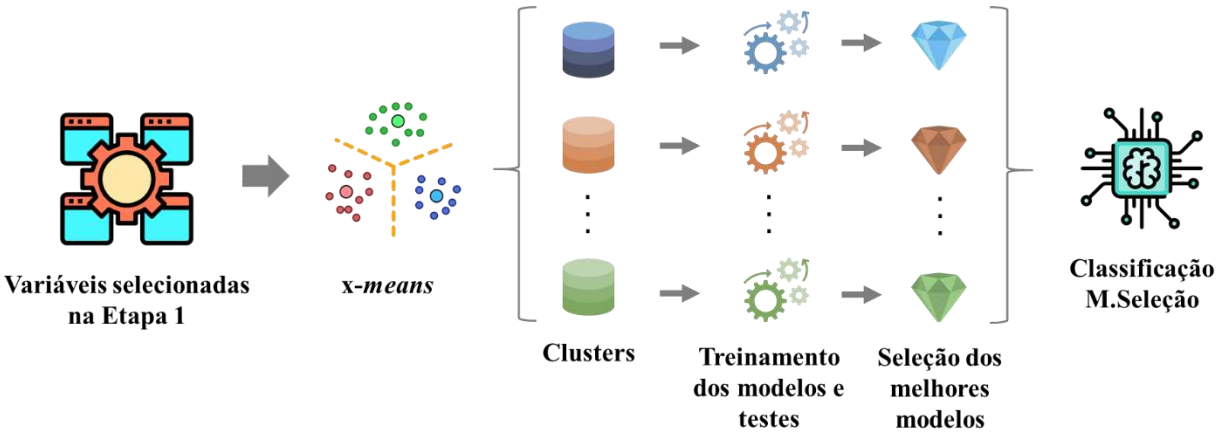
Por isso, no presente trabalho optou-se por se realizar a análise de sensibilidade do limiar de classificação ilustrada na Figura 37, com o objetivo de maximizar a métrica F1-*score* e, assim, garantir um equilíbrio entre precisão e sensibilidade nos modelos preditivos desenvolvidos.

Portanto, os resultados dos modelos apresentados no Capítulo 5 se referem sempre ao limiar de classificação que maximiza o valor do F1-*score* no modelo em questão. Assim, o valor do limiar de classificação pode ser diferente para cada modelo.

4.2.2 Classificação Segmentada

A principal diferença entre o procedimento de classificação segmentada, o qual será denominado de modelo M.Seleção, e a classificação com classificadores individuais descrita na subseção anterior é que, no primeiro caso, é executada uma etapa preliminar de segregação da base de dados com o algoritmo x-*means*, conforme ilustração apresentada na Figura 38.

Figura 38 – Procedimentos executados para classificação dos consumidores com a segregação prévia da base de dados.



Fonte: Autoria própria.

Como pode ser visto na Figura 38, uma segmentação é realiza no banco de dados antes da etapa de classificação. O algoritmo x-*means* é utilizado como método de agrupamento, enquanto a segregação dos consumidores é feita a partir do consumo médio mensal de energia e no tipo de atividade de cada consumidor. As duas informações foram selecionadas, pois, a partir delas os consumidores são segregados de acordo com seu porte e perfil de consumo. Os parâmetros utilizados na execução do algoritmo *x-means* são apresentados na Tabela 10.

Tabela 10 – Parâmetros utilizados no método *x-means*.

Parâmetro	Valor
Qtd. mínima de *clusters*	2
Qtd. máxima de *clusters*	10
Quantidade máxima de iterações	1.000

Fonte: Autoria própria.

Os valores apresentados na Tabela 10 foram definidos com base na expectativa dos agrupamentos que podem ser realizados a partir do perfil de consumo e do porte dos consumidores. Os *clusters* criados a partir da aplicação do algoritmo x-*means* são tratados como bases de dados independentes e, para cada um deles, é executada a etapa de classificação com classificadores individuais descrita na subseção 4.2.1. Ou seja, para cada *cluster* serão avaliados 11 algoritmos distintos de classificação.

Em seguida, são selecionados os melhores classificadores em cada *cluster* distintamente, com base na métrica F1-*score*. Por fim, os resultados das classificações parciais dos *clusters* são agrupados para compor a classificação final do modelo que será denominado "M.Seleção". Dessa forma, é possível que na classificação obtida no modelo M.Seleção hajam contribuições de diferentes algoritmos de classificação.

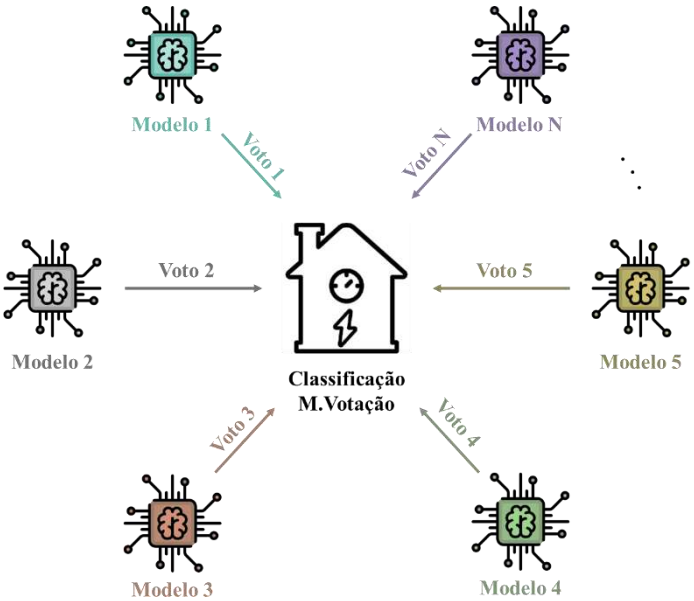
A ideia por trás da estratégia de classificação descrita nos parágrafos anteriores é reduzir a heterogeneidade dos dados, e avaliar se o procedimento facilita a tarefa de aprendizado de padrões relacionados a perda não técnica de energia. Em outras palavras, a estratégia proposta permite avaliar a hipótese de que um algoritmo pode ser mais adequado para identificar a PNT em pequenos consumidores residenciais, do que em grandes consumidores industriais, por exemplo. Se a hipótese for verdadeira, o resultado da classificação obtida no modelo M.Seleção será melhor do que aquele obtido com classificadores individuais descrito na subseção 4.2.1.

4.2.3 Critério de Votação

Cada algoritmo e cada estratégia de classificação apresentados nas subseções 4.2.1 e 4.2.2 podem ser considerados como um modelo de classificação distinto. Assim, dezenas de modelos de classificação diferentes estarão disponíveis ao final das etapas descritas nas subseções anteriores. Uma avaliação sobre o desempenho dos diferentes modelos pode ser executada para definir o mais adequado na tarefa de identificação da perda não técnica no sistema de distribuição.

Por outro lado, uma avaliação alternativa também é considerada no estudo, a qual consiste na criação de um novo modelo que será denominado de "M.Votação". O modelo M.Votação consiste em um critério de votos para classificar um consumidor quanto à presença ou não de PNT. Os votos são atribuídos por cada um dos modelos desenvolvidos anteriormente, conforme ilustrado na Figura 39.

Figura 39 – Ilustração do funcionamento do modelo M.Votação.



Fonte: Autoria própria.

Como visto na Figura 39, a classificação do consumidor no modelo M.Votação será baseada na classificação que o consumidor recebeu nos demais modelos desenvolvidos. A quantidade de votos necessários para definir que um determinador consumidor possui PNT também será objeto de avaliação no modelo M.Votação.

A ideia no modelo M.Votação é verificar se existe alguma melhoria na identificação da perda não técnica quando se considera a coincidência na classificação oferecida por diferentes modelos preditivos, com diferentes estratégias de classificação. No caso da hipótese ser verdadeira, o resultado obtido no modelo M.Votação será melhor do que aqueles obtidos nos demais modelos descritos nas subseções 4.2.1 e 4.2.2.

4.3 Etapa 3: Previsão da Energia Recuperada

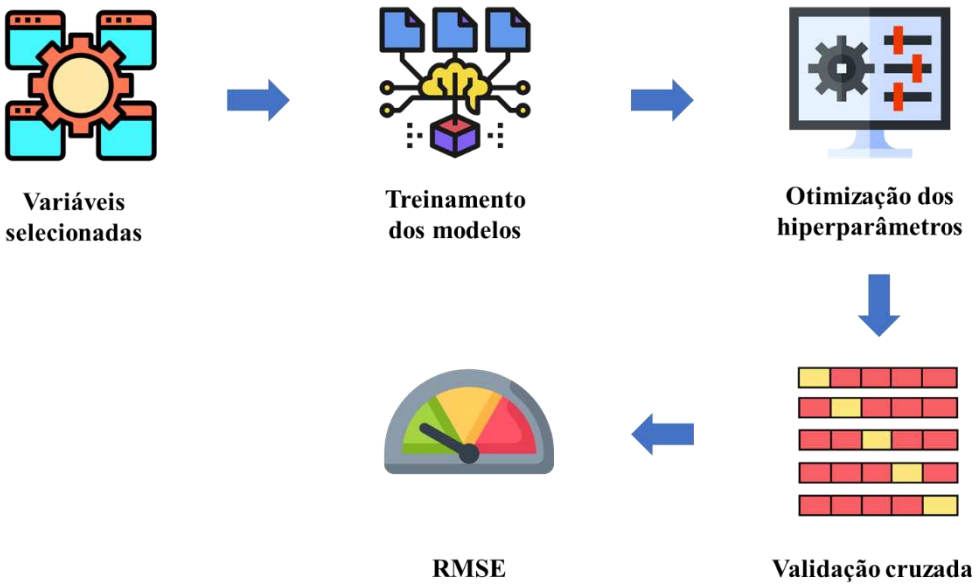
Quando um consumidor é identificado com perda não técnica, é possível que a concessionária recupere total ou parcialmente a energia que deixou de ser medida no período de até 36 meses anteriores a data da inspeção, conforme regras estabelecidas pela ANEEL e discutas na subseção 2.1.2. A quantidade de energia recuperada a partir de uma inspeção em campo é uma informação fundamental para determinar o retorno obtido pela concessionária com aquela ação. Sob esse ponto de vista, é mais interessante para uma concessionária identificar a PNT em um consumidor que possibilite a recuperação de 1.000 kWh de energia, do que em dois consumidores que possibilitem a recuperação de 100 kWh cada, por exemplo.

Por isso, o objetivo da Etapa 3 da metodologia desenvolvida nesta pesquisa é estimar um potencial de energia recuperada para cada consumidor da base dados, no caso de uma inspeção ser realizada. Para tanto, os procedimentos descritos na Figura 40 são executados.

Como pode ser observado, os procedimentos descritos na Figura 40 são semelhantes aos descritos na Figura 36 da seção 4.2. Contudo, existe uma diferença fundamental para o caso em questão, pois se trata de um problema de regressão, enquanto que na seção 4.2 é abordado um problema de classificação. Por essa razão, os algoritmos utilizados na presente etapa restringem-se aos sete indicados na Tabela 4 como apropriados para problemas de regressão.

Existem ainda outras diferenças entre os procedimentos descritos nas Figura 40 e Figura 36, uma delas diz respeito a seleção das variáveis que são utilizadas como entrada nos modelos preditivos. No caso da previsão do valor da energia recuperada, não é possível aplicar a análise de Inferência Causal, como feito na Etapa 1 da metodologia. Pois, a energia recuperada é uma grandeza numérica, e não um evento como é a ocorrência de perda não técnica.

Figura 40 – Procedimentos executados para a previsão do potencial de energia recuperada nas inspeções em campo.



Fonte: Autoria própria.

Assim, para seleção das variáveis na presente etapa, é calculado o valor do coeficiente de correlação de Spearman (ρ) entre todas as variáveis disponíveis e a variável energia recuperada. As variáveis que apresentam algum grau de correlação com a energia recuperada com uma confiança igual ou maior que 95% são selecionadas. Além disso, o coeficiente de correlação de Spearman também é calculado mutuamente entre as variáveis selecionadas, com objetivo de identificar informações redundantes. Os pares de variáveis que apresentam $|\rho| \geq 0{,}7$ são considerados redundantes e, nesses casos, a variável com menor associação à energia recuperada é eliminada da base de dados.

Com relação à otimização dos hiperparâmetros, ela é executada com as mesmas configurações da Tabela 8 e com os mesmos intervalos de busca definidos na Tabela 9. Por fim, destaca-se que, enquanto na Etapa 2 a avaliação dos modelos preditivos é realizada a partir da métrica F1-*score*, na presente etapa utiliza-se a métrica de erro *Root Mean Square Error* (RMSE), por ser a mais recomendada para comparação de modelos de regressão, conforme discutido na subseção 2.2.7.2.

4.4 Etapa 4: Maximização do Retorno Financeiro

O objetivo da Etapa 4 é determinar um modelo de seleção de consumidores para inspeções em campo, focado na maximização do retorno financeiro proporcionado às

concessionárias de energia. Para tanto, inicialmente é determinada uma equação que modela o retorno financeiro gerado por uma inspeção em campo e, posteriormente, os resultados das Etapa 2 e Etapa 3 são utilizados para definir o grupo de consumidores que maximiza a equação.

A equação pode ser definida a partir da informação de que o retorno financeiro de uma inspeção é dado pela diferença entre a receita gerada pela energia recuperada e o custo operacional da inspeção. A relação é apresentada matematicamente na Equação (33).

$$R = (E_{kWh} \times T) - C, \tag{33}$$

em que R é o retorno financeiro da inspeção; E_{kWh} é a energia recuperada a partir da inspeção; T é a tarifa de venda de energia sem impostos praticada pela concessionária; e C é o custo operacional médio para realizar uma inspeção. Para a concessionária considerada neste estudo, T = R$ 592,43/MWh e C = R$ 156,12/inspeção.

O valor do retorno financeiro R na Equação (33) só pode ser conhecido após a realização da inspeção em campo e posterior cálculo da energia recuperada E_{kWh}. Devido a tal limitação, propõe-se como alternativa para realização da presente análise a utilização de um potencial de retorno financeiro, o qual pode ser obtido antes da realização da inspeção em campo e é expresso na Equação (34).

$$R_P = \left(P_{PNT} \times P_{E_{kWh}} \times T\right) - C, \tag{34}$$

em que R_P é o potencial de retorno financeiro de uma inspeção em campo; P_{PNT} é a probabilidade de ocorrência de PNT no consumidor, a qual é fornecida por um dos modelos preditivos descritos na Etapa 2; e $P_{E_{kWh}}$ é a previsão de energia recuperada a partir da inspeção em campo, a qual é fornecida por um dos modelos preditivos descritos na Etapa 3.

A ideia na estimativa apresentada na Equação (34) é ponderar o valor previsto de energia recuperada ($P_{E_{kWh}}$) pela probabilidade de ocorrência de PNT (P_{PNT}). Desse modo, os modelos preditivos de classificação e regressão atuam como corretores um do outro. Por exemplo, quando um consumidor não possui PNT em campo, espera-se que o modelo de classificação forneça um baixo valor de P_{PNT}. Portanto, se o valor de energia recuperada previsto pelo modelo de regressão para este mesmo consumidor for muito elevado, ele será reduzido pela multiplicação com o valor P_{PNT}. O raciocínio oposto também pode ser aplicado, com um baixo valor de energia recuperada prevista corrigindo um alto valor de probabilidade de ocorrência

da PNT. Portanto, o erro global do valor do potencial de retorno financeiro (R_P) na Equação (34) tende a ser reduzido.

Uma vez que o potencial de retorno financeiro tenha sido definido na Equação (34), é possível realizar as análises necessárias para a maximização do seu valor. A análise de maximização pode ser aplicada em dois cenários distintos:

- Cenário 1: a infraestrutura para gestão da PNT é fixa;
- Cenário 2: a infraestrutura para gestão da PNT é passível de dimensionamento.

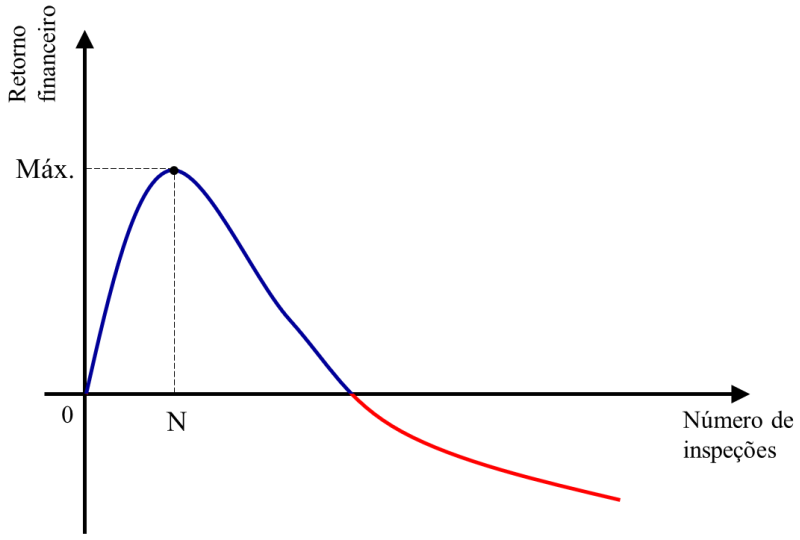
Entenda-se por infraestrutura para gestão da PNT as equipes de campo, os veículos, os equipamentos utilizados nas inspeções e as equipes de *back office*.

No Cenário 1, a quantidade de consumidores que podem ser inspecionados já está definida pelo tamanho da infraestrutura, resta, portanto, definir quais são os consumidores. Para tanto, o valor de R_P é calculado para todos os consumidores da base de dados de acordo com a Equação (34). Posteriormente, os consumidores são ordenados de forma decrescente de acordo como o valor R_P. A seleção de consumidores para inspeção em campo deve ser realizada de acordo com a ordem de prioridades obtida. Como os maiores retornos estão concentrados nas inspeções iniciais, é garantido que qualquer número de inspeções gerará o maior retorno possível.

No Cenário 2, a infraestrutura pode ser vista como uma variável a ser ajustada até o ponto de máximo retorno financeiro. Devido ao fato de existir um número limitado de consumidores com PNT na base de dados, é natural que após um determinado número de inspeções o retorno das ações seja negativo, ou seja, o custo operacional das inspeções será superior à receita obtida com a energia recuperada. Por outro lado, realizar poucas inspeções trará um retorno financeiro abaixo do possível. Portanto, existe uma relação entre o número de inspeções em campo que são realizadas em um conjunto de consumidores e o retorno financeiro obtido com as ações, esta relação está ilustrada na Figura 41.

Conforme ilustrado na Figura 41, existe um número específico de inspeções, o qual é indicado pela letra N no gráfico, que maximiza o retorno financeiro obtido com as ações em campo. A abordagem proposta nesta seção pode ser utilizada para determinar esse ponto. Uma vez que os valores de R_P tenham sido calculados para todos os consumidores da base de dados, e eles estejam ordenados de forma decrescente, o gráfico da Figura 41 pode ser traçado em termos do valor acumulado de R_P, e a quantidade ótima de inspeções pode ser determinada para o conjunto de dados em questão.

Figura 41 – Relação teórica entre número de fiscalizações e retorno financeiro.



Fonte: Autoria própria.

A partir do conhecimento do número ótimo de inspeções, a concessionária pode dimensionar a infraestrutura necessária para executá-las em um período desejado. Portanto, a presente metodologia representa também um recurso para o dimensionamento ótimo da infraestrutura de gestão da perda não técnica nas concessionárias de energia. O que pode ser muito útil para o planejamento das concessionárias, dado que não existe um critério definido para essa tarefa disponível na bibliografia correlata ao tema.

4.5 Inspeções em Campo

Com objetivo de validar os resultados obtidos nas Etapas 2, 3 e 4 em aplicações reais, foram realizadas novas inspeções em campo com o apoio da concessionária que disponibilizou a base de dados para realização dos estudos.

Os modelos preditivos desenvolvidos na Etapa 2 são utilizados para selecionar novos consumidores na base dados, para os quais ainda não há informação da presença de perda não técnica, ou seja, não há registro de inspeções realizadas nos últimos dois anos. Em seguida, os consumidores selecionados são enviados às equipes de campo para as inspeções *in loco*.

Porém, como as inspeções em consumidores regulares não são desejadas pela concessionária, apenas os consumidores classificados com PNT pelos modelos preditivos são selecionados. Além disso, a concessionária também limitou a quantidade de inspeções que poderiam ser realizadas, em função da limitação de trabalho das equipes de campo. Por isso, o número de testes para cada modelo preditivo foi limitado a 200 consumidores. Assim, foram

enviados para inspeção os 200 consumidores com maior probabilidade de possuírem PNT de acordo com cada um dos modelos descritos na Etapa 2.

Em seguida, os resultados das inspeções de campo são comparados com os resultados obtidos no processo de validação cruzada. Desta forma, é possível medir a confiabilidade dos modelos, ou seja, verificar se o desempenho simulado permanece o mesmo em inspeções reais de campo. No entanto, como apenas os consumidores indicados com a presença de PNT são inspecionados, não é possível calcular a métrica F1-*score* para os resultados de campo. Portanto, a comparação é feita pela métrica de precisão.

A partir do cálculo da energia recuperada com as novas inspeções em campo, também é possível realizar a validação dos valores previsto com os modelos de regressão descritos na Etapa 3. Por fim, com o objetivo de validar os resultados de maximização do retorno financeiro descritos na Etapa 4, foram encaminhados para inspeções em campo mais dois grupos de consumidores:

- Grupo 1: os 200 consumidores com maior probabilidade de possuírem PNT, de acordo com o melhor classificador obtido na Etapa 2; e
- Grupo 2: os 200 consumidores com maior potencial de retorno financeiro, de acordo com o cálculo da Equação (34).

O Grupo 1 representa a abordagem clássica de seleções de consumidores para inspeções de PNT, que foca em identificar o maior número de consumidores com PNT possível, enquanto o Grupo 2 representa a nova abordagem proposta na Etapa 4, focada na maximização do retorno financeiro das inspeções. Portanto, a comparação dos resultados de ambos os grupos permitirá medir o ganho obtido com as diferentes abordagens de seleção e compará-lo com os resultados simulados previamente na base de dados.

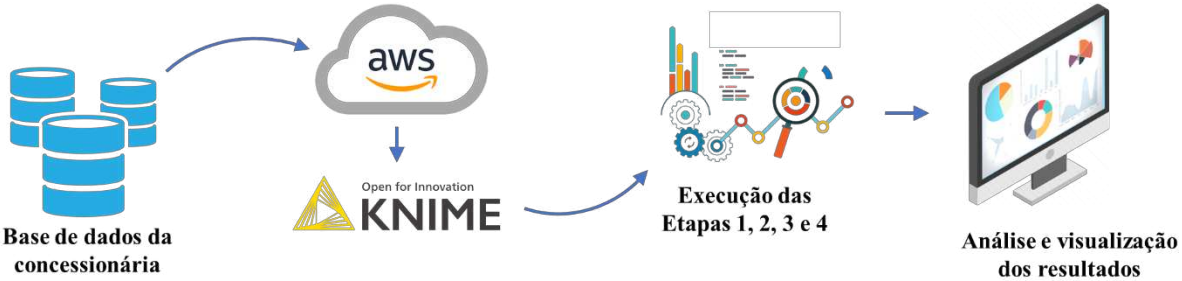
A validação a partir de inspeções reais em campo garante robustez aos resultados obtidos na pesquisa, pois indicam em que medida eles podem ser reproduzidos em aplicações reais. Além disso, poucos trabalhos na bibliografia realizaram a validação dos seus resultados em campo, configurando, assim, um diferencial da presente pesquisa.

4.6 Ambiente Computacional

A execução das etapas descritas anteriormente exige um esforço computacional significativo. Assim, para que o desenvolvimento da pesquisa seja exequível, faz-se necessário a escolha de soluções adequadas de *software* e *hardware*. Na Figura 42 é apresentado uma visão

geral da execução das etapas da metodologia, destacando-se as soluções computacionais utilizadas.

Figura 42 – Visão geral da execução das etapas da metodologia, destacando-se as soluções de *hardware* e *software*.



Fonte: Autoria própria.

Como observado na Figura 42, as informações coletadas da base de dados da concessionária são transferidas para a plataforma de computação em nuvem *Amazon Web Service* (AWS), a qual foi utilizada como solução de *hardware* na pesquisa. O serviço AWS foi contratado sob demanda para um servidor do tipo r5a.4xlarge que possui as configurações apresentadas na Tabela 11.

Tabela 11 – Configurações do servidor contratado para computação em nuvem.

Característica	Descrição
Processador	• Intel Xeon Platinum Série 8000; • *Clock* de até 3,1 GHz; • 16 núcleos de processamento.
Memória	• 128 GB DDR4; • 2,1 GHZ.
Armazenamento	• 30 GB tipo SSD.

Fonte: Autoria própria.

A opção pela contratação de um servidor em nuvem ocorreu em função do elevado custo computacional dos processamentos executados na pesquisa, os quais demandam a disponibilidade de uma estrutura de *hardware* robusta, como as descritas na Tabela 11. No caso da AWS é possível contratar apenas o tempo de uso do servidor a um custo de US$ 1,64/hora, enquanto que a aquisição de um servidor físico com configurações semelhantes teria um custo aproximado de R$ 70.000,00. Por isso, tornou-se mais viável a utilização do serviço AWS sob demanda.

Retornando a análise da Figura 42, é possível notar que a plataforma de análise de dados KNIME® é utilizada como solução de *software*. O KNIME® ou *Konstanz Information Miner* consiste em um ambiente computacional gratuito de código aberto, desenvolvido em 2006 por

pesquisadores da Universidade de Constança na Alemanha, o qual é dedicado ao processo de *Advanced Analytics*. No KNIME® estão disponíveis vários pacotes computacionais com técnicas para as diferentes etapas do processo de *Advance Analytics* integradas em um conceito de construção modular. Além disso, a ferramenta também possui integração com várias outras soluções de análise de dados como as linguagens *Python*, R e Java® e os pacotes computacionais Weka® e H2O®. No KNIME® também é possível desenvolver rotinas personalizadas com o uso das linguagens Java *script*, R ou *Python*, as rotinas podem ser compartilhadas posteriormente com outros usuários em um ambiente de colaboração *online* chamado NodePit.

Os principais motivos para escolha do KNIME® no desenvolvimento desta pesquisa foram: a gratuidade; a universalidade, devido à integração com outras soluções; a robustez para trabalhar com grande volume de dados; a agilidade, por permitir o desenvolvimento de forma estruturada; e a completude, por permitir o desenvolvimento de todas as etapas do processo em um único ambiente.

A partir do exposto no presente capítulo, pretende-se garantir o entendimento de como os resultados da pesquisa foram obtidos, bem como, possibilitar a sua reprodutibilidade. No capítulo seguinte todos os resultados obtidos na pesquisa são apresentados e discutidos de forma detalhada.

5 Resultados e Discussões

Neste capítulo são apresentados e discutidos todos os resultados da pesquisa, os quais foram obtidos a partir do material e dos métodos descritos no capítulo anterior. Os resultados são apresentados de acordo com a ordem de descrição das etapas no Capítulo 4.

5.1 Caracterização da PNT

A primeira etapa da metodologia consiste em caracterizar a perda não técnica de energia a partir de variáveis que foram criadas nos processos de extração de dados, pré-processamento e *feature engineering*, conforme detalhado na seção 4.1. Na Tabela 12 é apresentado um resumo das variáveis obtidas ao final do processo.

Tabela 12 – Resumo das variáveis obtidas nos processos de extração, pré-processamento e *feature engineering*.

Categoria	Quantidade	Descrição
Ações de combate às perdas	22	Variáveis relacionadas às ações para controle da PNT nos consumidores. Ex.: tempo desde a última inspeção e indicação de blindagem no sistema de medição.
Ordens de Serviços	33	Variáveis relacionadas à execução de ordens de serviço nos consumidores. Ex.: quantidades de cortes de energia por falta de pagamento e troca de medidores.
Atendimentos	8	Variáveis relacionadas às interações nos canais de atendimento ao consumidor. Ex.: quantidades de atendimentos por *chatbot*, site ou ligação telefônica.
Faturamento	17	Variáveis relacionadas ao processo de faturamento do consumo. Ex.: tempo médio para pagamento e quantidade de faturas vencidas sem pagamento.
Leitura do consumo	9	Variáveis relacionadas ao processo de leitura do consumo. Ex.: suspeitas de fraude indicadas pelo leiturista e quantidade de meses sem a realização da leitura do medidor.
Cadastro comercial	20	Variáveis relacionadas às informações comerciais dos consumidores. Ex.: tipo de atividade, classe de consumo e tempo desde a primeira ligação na rede.
Cadastro Técnico	26	Variáveis relacionadas às informações técnicas dos consumidores. Ex.: modelo do medidor, nível de tensão e quantidade de fases.
Consumo de energia	47	Variáveis relacionadas às séries de consumo mensal de energia. Ex.: desvio padrão, derivada e distância interquartil.

Fonte: Autoria própria.

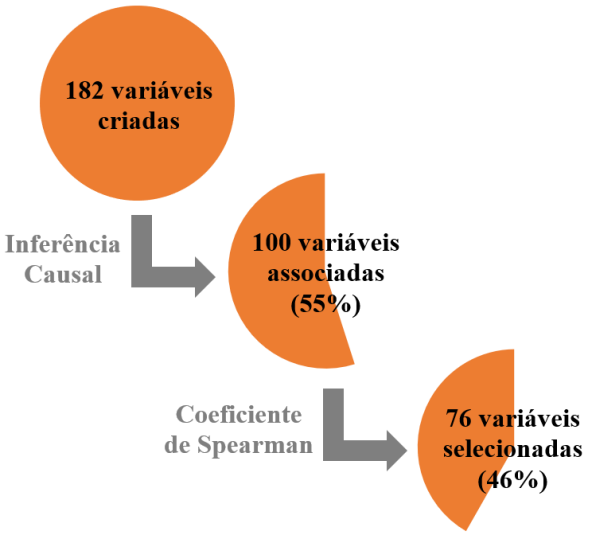
Conforme apresentado na Tabela 12, no total foram obtidas 182 variáveis distintas a partir das informações disponíveis nos bancos de dados da concessionária. Devido à limitação

de espaço, optou-se por apresentar na Tabela 12 apenas um resumo baseado nas características comuns das variáveis. Contudo, uma descrição detalhada de todas a 182 variáveis é fornecida na Tabela 23, a qual é apresentada no Apêndice A ao final do texto.

A maioria das variáveis na Tabela 12 são relacionadas ao consumo de energia, ou seja, foram obtidas a partir da série histórica de consumo de energia elétrica. O resultado deve-se ao fato de que o processo de criação das variáveis é focado na descoberta de informações relacionadas à ocorrência de PNT, e um dos principais efeitos da PNT em um consumidor é a redução ou alteração no seu perfil de consumo.

Apesar do grande número de variáveis obtidas, ainda não é possível afirmar que tais variáveis são adequadas para caracterizar a PNT. Para tanto, foi realizada a análise de Inferência Causal, como descrito na seção 4.1. Como resultado, 100 variáveis apresentaram alguma associação significativa com a presença ou ausência de PNT nos consumidores. No entanto, 24 pares de variáveis apresentaram coeficiente de correlação de Spearman superior a 0,7 e foram considerados redundantes. Assim, após a eliminação da redundância, 76 variáveis independentes restaram na base de dados, conforme ilustrado na Figura 43.

Figura 43 – Resultado do processo de seleção de variáveis para caracterização da perda não técnica de energia.



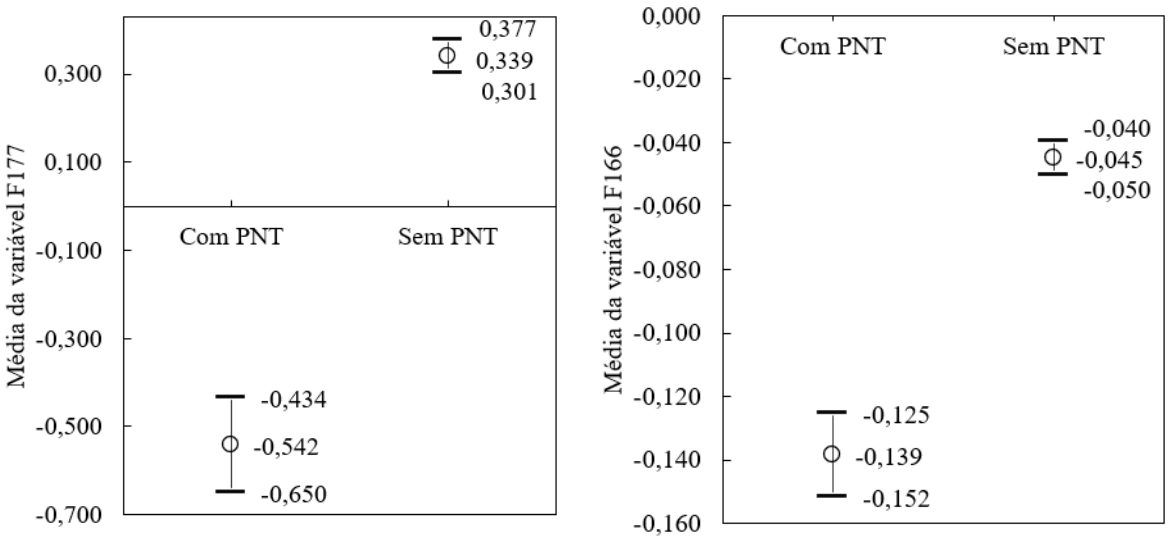
Fonte: Autoria própria.

Como apresentado na Figura 43, 46% das variáveis originalmente criadas foram selecionadas para caracterização da perda não técnica de energia e serão utilizadas como entrada nos modelos de classificação. Na Tabela 23 do Apêndice A estão indicadas quais foram as variáveis selecionadas, bem como, as variáveis associadas e não selecionadas.

Na Figura 44 são apresentados gráficos com as médias confiáveis para as variáveis denominadas F177 e F166. A partir dos gráficos e das análises realizadas na sequência é

possível verificar como as variáveis selecionadas podem ser utilizadas para caracterização da PNT no conjunto de consumidores.

Figura 44 – Média confiável das variáveis (a) F177 e (b) F166 para os grupos de consumidores com e sem perda não técnica de energia.



(a) (b)

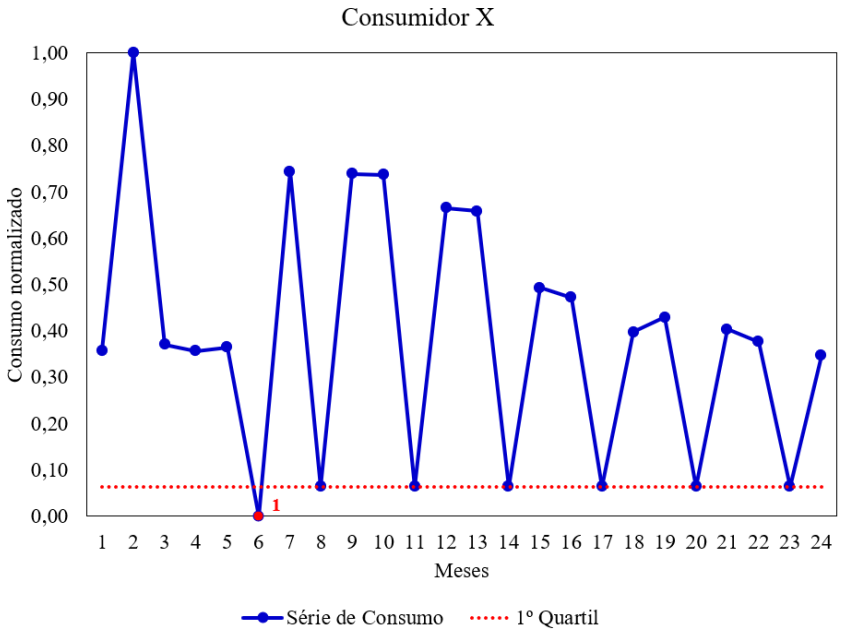
Fonte: Autoria própria.

A variável F177 apresentada na Figura 44 (a) representa a diferença relativa entre a quantidade de vezes que o consumo de um mês foi inferior ao 1º quartil da série completa de consumos em um indivíduo e em um grupo de indivíduos com características semelhantes em termos de consumo, localização e informações cadastrais. Na Figura 45 são apresentados gráficos que facilitam o entendimento de como a variável F177 é obtida.

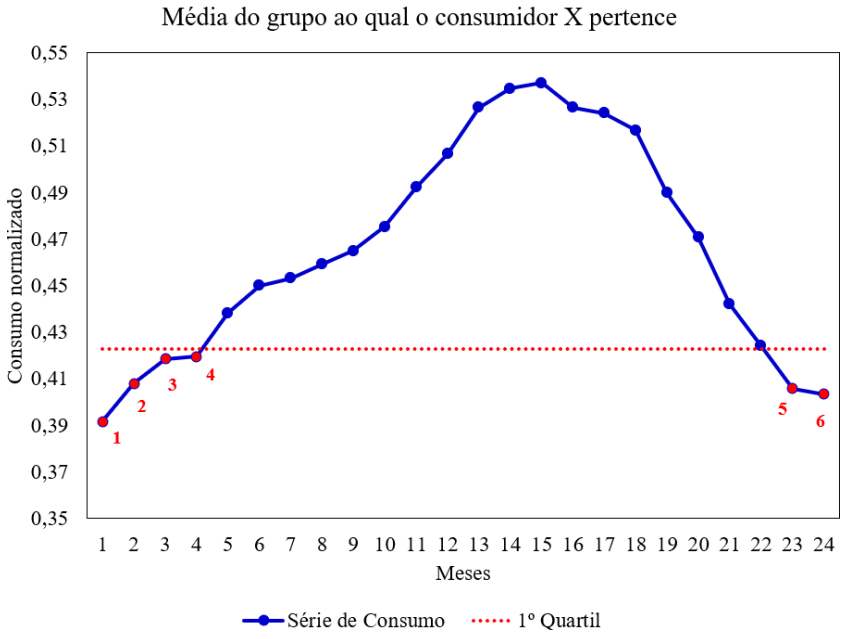
Na Figura 45 (a) é apresentada a série de consumo mensal normalizado para um determinado consumidor X, e na Figura 45 (b) é apresentada a média de consumo mensal para o grupo de consumidores ao qual o consumidor X pertence. Como pode ser visto na Figura 45 (a), existe apenas um único mês para o qual o valor do consumo está abaixo do 1º quartil, enquanto que na Figura 45 (b) existem seis meses para os quais o valor do consumo está abaixo do 1º quartil da série. Portanto, para o caso do consumidor X representado na Figura 45, o valor da variável F177 será igual a (1-6)/6 = – 0,833. Ou seja, o consumidor X possui uma quantidade de meses com consumo abaixo do 1º quartil 83,3% menor que a média do seu grupo.

Na prática, a variável F177 é uma medida que contabiliza períodos não continuados de baixo consumo de energia. Períodos de menor consumo são esperados devido às características sazonais de uma determinada região. No entanto, estas características afetam todos os consumidores de um grupo da mesma maneira. Portanto, se um indivíduo difere de seus semelhantes quanto à medida em questão, ele se configura como uma exceção ao grupo.

Figura 45 – Séries de consumo mensal com indicação do 1º quartil da série para (a) um consumidor e (b) um grupo de consumidores.



(a)



(b)

Fonte: Autoria própria.

Após a discussão da Figura 45, fica claro na Figura 44 (a) que para um consumidor com PNT a quantidade de períodos de menor consumo de energia é cerca de 54,2% menor que a média do seu grupo. Enquanto que para consumidores sem PNT, a quantidade é cerca de 33,9% maior que a média do seu grupo. Portanto, a análise da Figura 44 (a) sugere que consumidores com PNT possuem uma quantidade menor de períodos não continuados de baixo consumo de

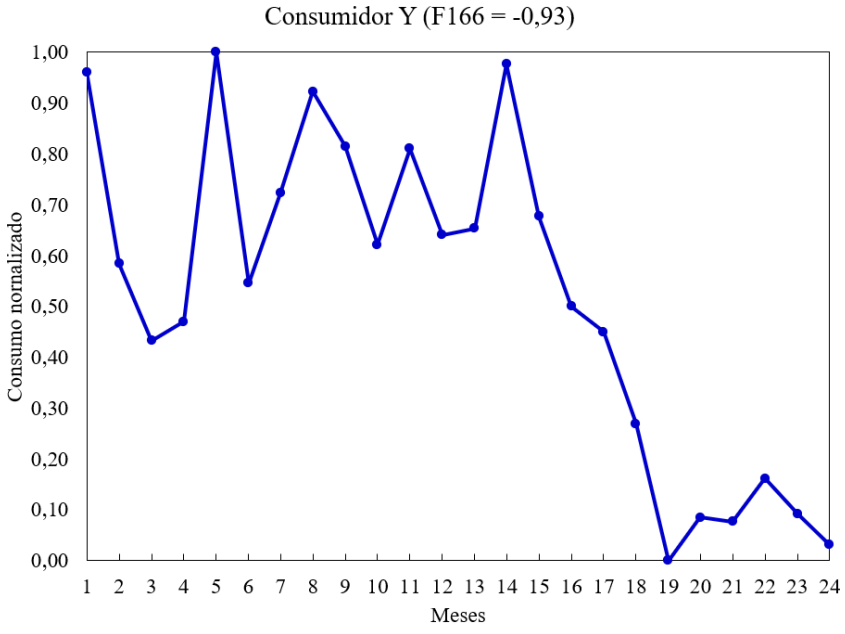
energia do que consumidores sem PNT. Uma explicação possível é que consumidores com PNT tendem a não modular seu consumo em função das características sazonais, como o clima ou as bandeiras tarifárias.

Voltando a análise para a Figura 44 (b), a variável F166 representa a derivada acumulada da série normalizada de consumo de energia de um consumidor, ou seja, a variável é uma medida da tendência da série. Os gráficos apresentados na Figura 46 auxiliam no entendimento do cálculo da variável F166.

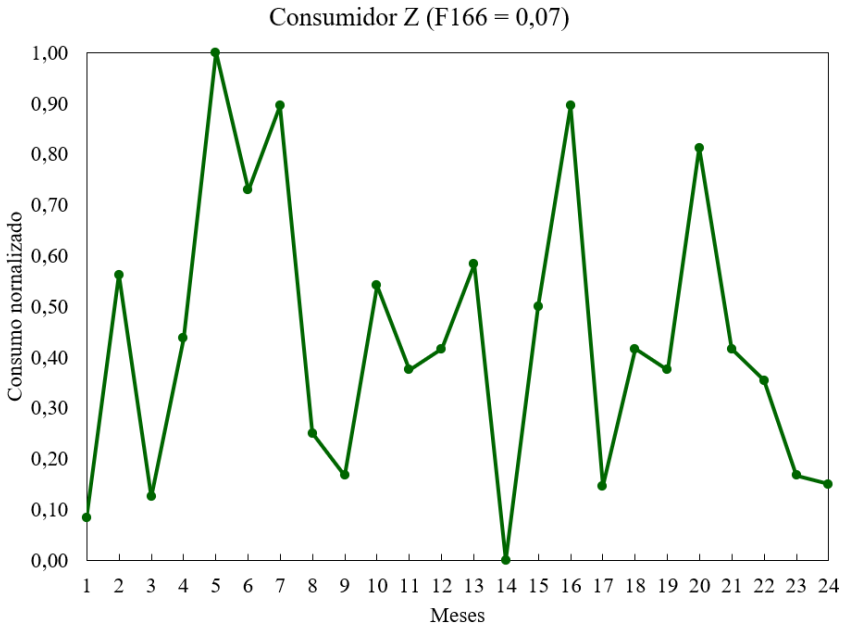
Na Figura 46 (a) é apresentada a série de consumo mensal de um determinado consumidor Y, cuja derivada acumulada da série é – 0,93. O valor negativo indica uma tendência de decrescimento na série, o que pode ser observado visualmente no gráfico a partir do mês 15. Na Figura 46 (b) é apresentada a série de consumo mensal de um determinado consumidor Z, cuja derivada acumulada da série é 0,07. O valor próximo a zero indica que não há uma tendência nem decrescimento nem de crescimento na série, o que também pode ser observado visualmente no gráfico.

Portanto, a variável F166 indica a tendência de decrescimento ou crescimento na série de consumo mensal dos consumidores. Um valor negativo indica um decrescimento, um valor positivo indica um crescimento e um valor próximo a zero indica estabilidade. Como a série de consumos é normalizada entre zero e um, o valor de F166 sempre estará entre – 1 e 1, com os valores extremos representando uma série monotônica decrescente e uma série monotônica crescente, respectivamente.

Figura 46 – Séries de consumo mensal normalizado com derivada acumulada igual a (a) -0,93 e (b) 0,07.



(a)



(b)

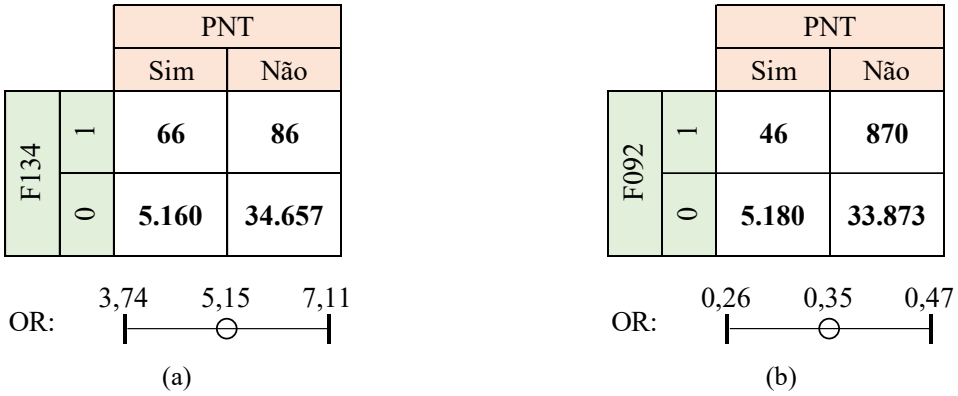
Fonte: Autoria própria.

O resultado das médias confiáveis da variável F166 apresentado na Figura 44 (b) indica que o valor de F166 para os consumidores com PNT é de cerca de – 0 ,139, enquanto que para os consumidores sem PNT é de cerca de – 0,045. Portanto, a análise da Figura 44 (b) sugere que os consumidores com PNT possuem uma tendência mais decrescente para a série de consumo mensal do que os consumidores sem PNT. Uma explicação possível é que consumidores com PNT tendem a apresentar uma queda no consumo após a instalação da irregularidade no sistema de medição.

As análises realizadas para a Figura 44 dizem respeito a variáveis numéricas. Porém, variáveis categóricas também fazem parte do grupo selecionado para a caracterização da PNT. No caso das variáveis categóricas ao invés da média confiável, utiliza-se o *Odds Ratio* como medida de associação, conforme descrito na subseção 2.2.3.1. Com o objetivo de ilustrar o uso das variáveis categóricas, são apresentados na Figura 47 os valores de *Odds Ratio* calculados para as variáveis denominadas F134 e F092.

Na Figura 47 (a) é apresentada a tabela de distribuição conjunta para a variável F134, a qual indica a presença de um modelo de específico medidor. Se F134 for igual a um, significa que o consumidor possui o modelo instalado, se F134 for igual a zero, significa que o consumidor não possui o modelo. O *Odds Ratio* calculado para a tabela de distribuição conjunta da Figura 47 (a) ficou entre 3,74 e 7,11, com valor central de 5,15. Isso indica que um consumidor com o modelo de medidor em questão tem uma chance cerca de 5,15 vezes maior de possuir PNT do que um consumidor que não possui o modelo.

Figura 47 – Valores de *Odds Ratio* calculados para as variáveis (a) F134 e (b) F092.



Fonte: Autoria própria.

Portanto, a análise da Figura 47 (a) sugere que o modelo de medidor em questão é um fator de risco para ocorrência de PNT em um consumidor. Uma possível explicação é que o modelo possui fragilidades construtivas, o que torna sua adulteração mais fácil, ou ainda que existem pessoas especializadas na adulteração do modelo na região.

O parágrafo anterior representa uma conclusão importante para a gestão de perdas na concessionária que, a partir desta conclusão, pode suspender a aquisição de novas unidades do respectivo modelo de medidor, já que o mesmo representa um fator de risco para ocorrência de PNT.

Voltando a análise para a Figura 47 (b), na qual é apresentada a tabela de distribuição conjunta para a variável F092, a qual indica que o consumidor desenvolve a atividade de comércio atacadista. Se F092 for igual a um, significa que o consumidor é um comércio atacadista, se F092 for igual a zero, significa que o consumidor não é um comércio atacadista. O *Odds Ratio* calculado para a tabela de distribuição conjunta da Figura 47 (b) ficou entre 0,26 e 0,47, com valor central de 0,35. Como o valor do *Odds Ratio* é menor que um, conclui-se que a atividade de comércio atacadista é um fator de proteção para a ocorrência de PNT. Além disso, é possível afirmar que um consumidor que desenvolve a atividade de comércio atacadista tem uma chance cerca de 2,9 vezes menor de possuir PNT, do que um consumidor que desenvolve outro tipo de atividade, em que o valor 2,9 é obtido pela divisão da unidade pelo *Odds Ratio* (2,9 = 1/0,35). Uma possível explicação é que comércios atacadistas geralmente fazem parte de uma rede franquias, de modo que a fatura de energia elétrica não é paga localmente, o que pode diminuir a motivação para os administradores locais cometerem fraudes.

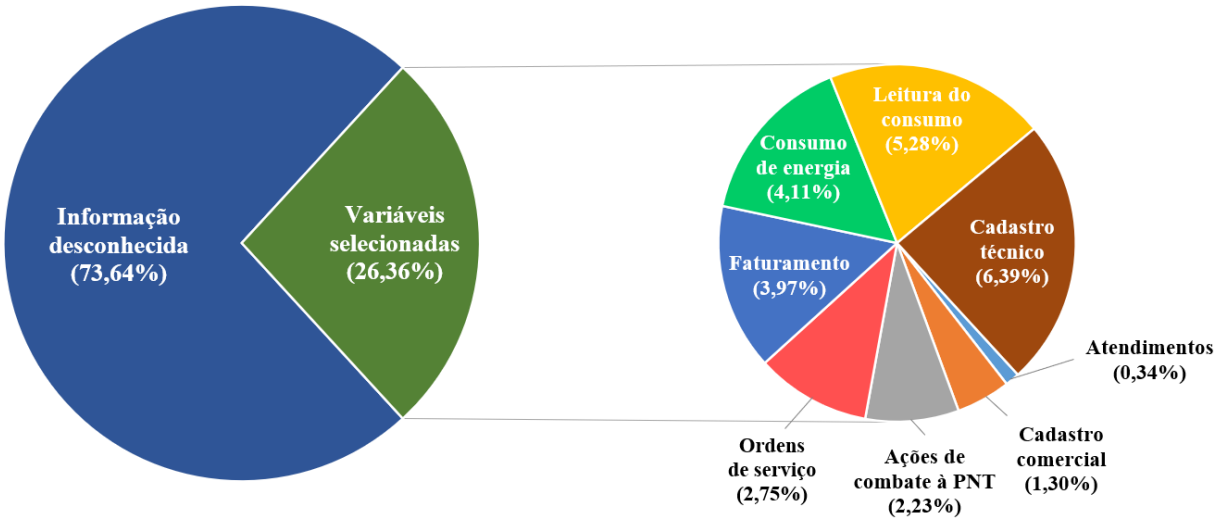
A partir da análise realizada para as quatro variáveis nos parágrafos anteriores, fica evidente como os resultados obtidos podem ser utilizados para caracterizar a presença de PNT nos consumidores da base dados, bem como, para tomadas de decisão relacionadas à gestão de

perda nas concessionárias. O mesmo tipo de análise pode ser realizado para todas as 76 variáveis selecionadas, o que permite que a concessionária conheça quantitativa e qualitativamente os fatores de risco e proteção para a ocorrência de PNT em seus consumidores.

Adicionalmente, é interessante conhecer em que medida o conjunto de 76 variáveis selecionadas pode descrever o comportamento da PNT no conjunto de consumidores. Uma maneira adequada de se fazer isso é por meio do cálculo da explicabilidade ϵ do conjunto de variáveis, o qual foi definido na Equação (9) da subseção 2.2.3.1.

Na Figura 48 é apresentado o resultado do cálculo da explicabilidade para o conjunto de 76 variáveis. O resultado também é apresentado de forma segregada pelas categorias listadas na Tabela 12.

Figura 48 – Explicabilidade obtida a partir do conjunto de 76 variáveis independentes selecionadas em relação à variável dependente que representa a ocorrência de PNT nos consumidores.



Fonte: Autoria própria.

A partir da análise da Figura 48, constata-se que o conjunto de variáveis selecionadas apresentam uma explicabilidade de 26,36% em relação à ocorrência de PNT na base de consumidores. Isto significa que 26,36% da variância da perda não técnica pode ser explicada pelo conjunto de variáveis selecionas. O valor pode ser considerado um bom resultado, dada a complexidade do problema e o fato de que em problemas de análise de Inferência Causal não se espera alcançar valores superiores a 50% (ROTHMAN e GREENLAND, 2004).

A análise da Figura 48 também demonstra que o conjunto de variáveis relacionadas ao cadastro técnico é o que apresenta maior explicabilidade, seguido do conjunto relacionado ao processo de leitura do consumo e ao conjunto relacionado ao consumo de energia. Comparando esse resultado com a quantidade de variáveis em cada conjunto que é apresentada na Tabela 12, destaca-se o conjunto relacionado a leitura do consumo, o qual possui apenas 9 variáveis e ainda

assim figura entre os mais explicativos. O resultado indica que as informações coletadas no processo de leitura do consumo, como as indicações de suspeita de fraude que são feitas pelos leituristas são valiosas e, portanto, devem ser monitoradas pela concessionária para auxiliar na gestão da PNT.

Por caracterizarem a presença da perda não técnica entre os consumidores, o conjunto de variáveis foi utilizado como entrada em modelos de classificação com o objetivo de classificar os consumidores em relação à presenta da PNT. Os resultados obtidos são apresentados e discutidos na seção seguinte.

5.2 Classificação de Consumidores

Para a etapa de classificação de consumidores, inicialmente, foram definidos os hiperparâmetros ótimos de cada algoritmo listado na Tabela 9. Os valores foram definidos a partir da aplicação do método de otimização Bayesiana, como descrito na subseção 4.2.1. Os resultados da otimização são apresentados na Tabela 13.

Tabela 13 – Hiperparâmetros ótimos definidos na otimização Bayesiana para a classificação.

Modelo	**Hiperparâmetro**	**Valor ótimo**
LR	Número máximo de iterações	1.000
	Tolerância (*epsilon*)	10^{-5}
DT	Número mínimo de derivações por nó	2
	Critério para derivação	*Gini Index*
NB	Limiar de classificação	0,0014
	Desvio padrão mínimo	0,814
	Desvio padrão mínimo do limiar de classificação	0,511
MLP	Número de camadas escondidas	6
	Número de neurônios por camada	26
SVM	Potência da função *kernel*	2
	Coeficiente da função *kernel*	0
FR	Norma triangular (*T-norm*)	Produto
GBT	Número de árvores	157
	Número de níveis das árvores	4
XGBT	Número de árvores	396
	Número de níveis das árvores	3
	Número máximo de *bins*	276
BT	*Bag size*	100%

Modelo	Hiperparâmetro	Valor ótimo
RF	Número de árvores	250
RTF	Tamanho do grupo de variáveis	40
	Taxa de remoção	40%

Fonte: Autoria própria.

Os valores apresentados na Tabela 13 para os classificadores LR e DT foram definidos manualmente de acordo com a experiência do autor, uma vez que os classificadores não requerem ou são pouco influenciados pela otimização de hiperparâmetros.

Utilizando os hiperparâmetros ótimos e o conjunto de variáveis selecionadas na etapa anterior, foi realizado o treinamento dos modelos. Posteriormente, foram realizados testes para avaliação de desempenho a partir do processo de validação cruzada *k-fold*, conforme descrito na subseção 4.2.1. Os resultados da validação cruzada são apresentados na Tabela 14.

Tabela 14 – Resultado da validação cruzada para os classificados individuais.

Modelo	Limiar de classificação	Precisão	Sensibilidade	F1-*score*
GBT	0,21	46,12%	43,42%	0,45
XGBT	0,59	41,26%	48,12%	0,44
BT	0,24	46,96%	41,81%	0,44
RTF	0,20	41,90%	46,59%	0,44
RF	0,26	47,20%	41,39%	0,44
SVM	0,30	42,43%	39,88%	0,41
LR	0,61	37,01%	45,12%	0,41
DT	0,17	40,19%	40,53%	0,40
MLP	0,84	36,46%	43,90%	0,40
NB	0,20	30,16%	42,50%	0,35
FR	0,83	23,35%	43,15%	0,30

Fonte: Autoria Própria.

Os resultados na Tabela 14 são apresentados em ordem decrescente do valor da métrica F1-*score*. Como pode ser observado, o modelo GBT apresentou o valor de 0,45 para o F1-*score*, sendo o maior valor para o conjunto testado e, portanto, pode ser considerado o melhor classificador de acordo com a métrica.

A análise da Tabela 14 também revela que os cinco melhores classificadores (GBT, XGBT, BT, RTF e RF) possuem valores semelhantes de F1-*score*, e são todos algoritmos do tipo *ensemble*. O resultado sugere que os algoritmos do tipo *ensemble* são mais adequados para a identificação de PNT no sistema de distribuição do que os demais tipos. Uma possível razão

para o bom desempenho é o fato de que algoritmos baseados em *ensemble* tendem a reduzir o viés da classificação, pois, incorporam vários classificadores distintos com diferentes padrões de erro para compor seus resultados. Assim, há uma diminuição no impacto causado pelo viés existente nos dados de treinamento.

Como os dados de treinamento são obtidos a partir do resultado de inspeções de campo anteriores, há um viés natural de seleção, pois, as inspeções não são realizadas aleatoriamente, mas sob algum critério estabelecido pela concessionária. Portanto, a característica dos métodos do tipo *ensemble* é especialmente vantajosa neste cenário.

Voltando a análise para o final da Tabela 14, o modelo FR aparece como o pior classificador do conjunto, apresentando um F1-*score* igual a 0,30, valor que é 33,33% menor que os melhores resultados da tabela. O baixo desempenho pode ser justificado pelo fato de que os classificadores *fuzzy* são mais adequados em cenários de incerteza na classificação, como por exemplo, a previsão do tempo de vida restante de equipamentos. Contudo, para o caso da identificação de PNT, a técnica *fuzzy* não se mostrou adequada.

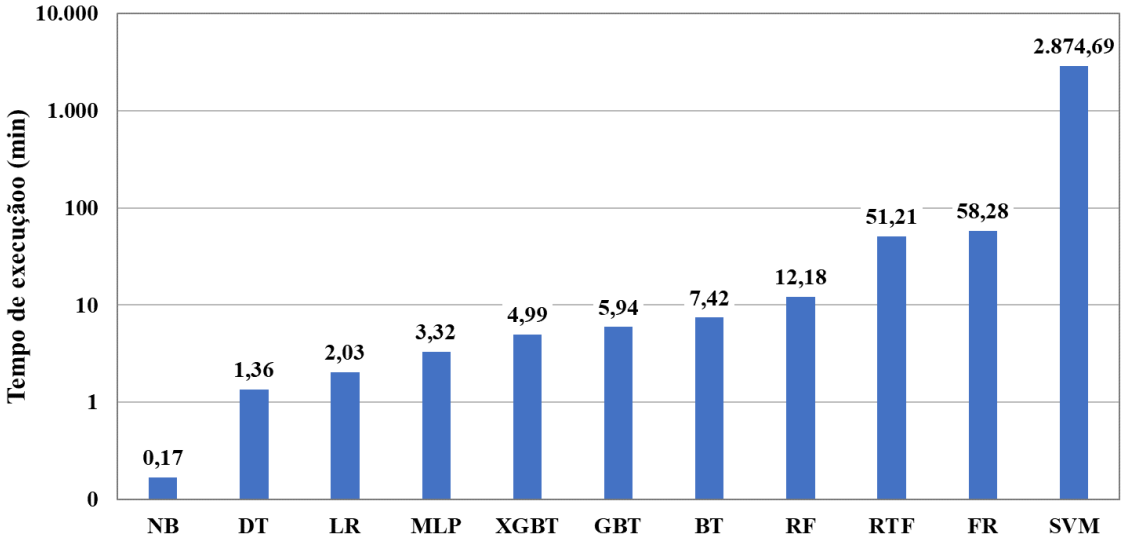
Os demais algoritmos na Tabela 14 (SVM, LR, DT e MLP) apresentaram desempenho intermediário, com valores de F1-*score* entre 0,4 e 0,41. A exceção foi o algoritmo NB, que apresentou F1-*score* igual a 0,35. Uma possível razão para o desempenho do NB é que ele emprega uma função de hipótese muito simples (linear) e ignora as relações de dependência entre as variáveis.

Por fim, é importante ressaltar que os valores do limiar de classificação apresentados na Tabela 14 representam o ponto para o qual o valor da métrica F1-*score* é máximo em cada classificador, conforme discutido na Figura 37 da subseção 4.2.1. O resultado demonstra a importância da análise dos limiares mais adequados para cada situação, uma vez que os valores mudam significativamente entre os diferentes classificadores. Em outras palavras, se um valor fixo fosse assumido como limiar para todos os modelos, os valores de F1-*score* na Tabela 14 seriam menores.

A maximização da métrica F1-*score* também é a razão pela qual os valores de precisão e sensibilidade apresentados na Tabela 14 são equilibrados na maioria dos modelos, pois, a métrica F1-*score* é definida pela média harmônica dos valores de precisão e sensibilidade, como apresentado na Equação (30). Assim, ao maximizar a média harmônica busca-se os maiores valores de precisão e sensibilidade simultaneamente, os quais só podem ser alcançados de forma equilibrada. Caso a relação entre o limiar de classificação, a precisão e a sensibilidade de um modelo fossem lineares, o valor máximo de F1-*score* ocorreria para valores idênticos de precisão e sensibilidade. A ideia foi já discutida na subseção 4.2.1 e ilustrada na Figura 37.

De maneira adicional, o desempenho dos modelos de classificação também foi avaliado a partir do tempo de execução do processo de validação cruzada. Os tempos de execução para cada modelo são apresentados na Figura 49.

Figura 49 – Tempo de execução do processo de validação cruzada em minutos.



Fonte: Autoria própria.

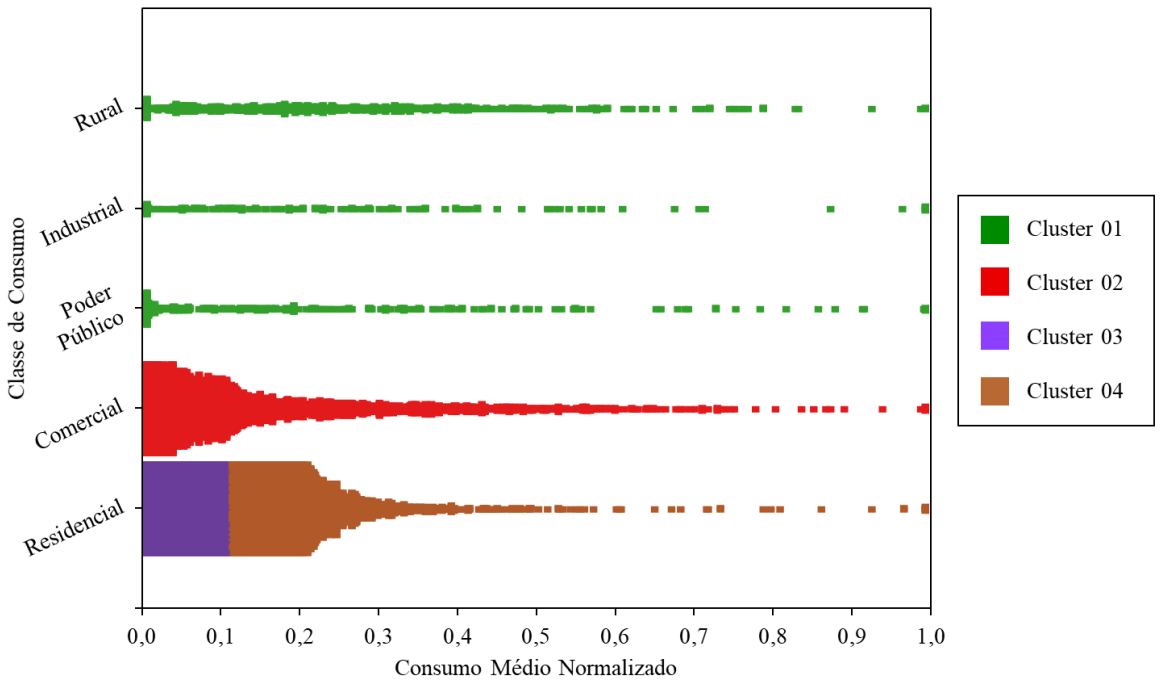
Note-se inicialmente que a escala vertical do gráfico na Figura 49 é logarítmica. A partir da análise das informações no gráfico, constata-se que o modelo NB é aquele com o menor custo computacional, exigindo apenas 0,17 minutos para executar o processo de validação cruzada. O resultado pode ser justificado pelo fato do algoritmo NB utilizar uma função de hipótese linear. Na outra extremidade do gráfico, a modelo SVM aparece com o maior custo computacional, com um tempo de execução de 2.874,69 minutos, o que representa praticamente dois dias completos de processamento computacional. O tempo de processamento do SVM é 49 vezes maior que o segundo modelo com maior custo computacional.

O elevado custo computacional do SVM pode ser atribuído ao fato de que o processamento do algoritmo é proporcional ao cubo da quantidade de variáveis utilizadas para o treinamento, enquanto o conjunto de treinamento possui 76 variáveis distintas. De fato, alguns autores não recomendam a utilização do SVM em problemas com um grande conjunto de dados (BANERJEE, 2020). Além disso, ao comparar o tempo de processamento do SVM com o desempenho do modelo apresentado na Tabela 14, é possível concluir que a relação custo-benefício oferecida pelo algoritmo não é razoável, já que o modelo obteve apenas o 6º melhor desempenho em termos de F1-*score*. Portanto, com base nos resultados, é possível afirmar que o SVM não é um algoritmo indicado para a tarefa de identificação da PNT no sistema de distribuição.

Com relação aos demais classificadores apresentados na Figura 49, pode-se afirmar que eles apresentaram um custo computacional razoável, com o tempo de processamento inferior a uma hora. Os classificadores XGBT, GBT, BT e RF merecem ser destacados por aliarem um elevado desempenho na Tabela 14 e um baixo custo computacional na Figura 49.

Na sequência da metodologia, a base de dados foi segmentada por meio do algoritmo *x-means* e das informações de consumo médio mensal de energia e tipo de atividade de cada consumidor, como descrito na subseção 4.2.2. O resultado da segmentação dos dados é apresentado na Figura 50 na forma de um gráfico de dispersão, no qual as coordenadas cartesianas representam as duas variáveis consideradas na clusterização, a coleção de pontos representa os indivíduos do conjunto e as cores representam o *cluster* ao qual os indivíduos pertencem.

Figura 50 – Resultado da clusterização com o algoritmo *x-means*.



Fonte: Autoria própria.

A partir da análise da Figura 50, constata-se que foram criados quatro *clusters* para o conjunto de dados, os quais apresentam boa compacidade e boa separabilidade, sendo que 3% dos dados foram rotulados com o Cluster 01, 12% com o Cluster 02, 28% com o Cluster 03 e 57% com o Cluster 04. Nota-se ainda que o Cluster 01 agrupou consumidores das atividades rural, industrial e serviço público com qualquer nível de consumo de energia. O Cluster 02 agrupou consumidores da atividade comercial com qualquer nível de consumo de energia. O

Cluster 03 agrupou consumidores da atividade residencial com baixo consumo de energia e o Cluster 04 agrupou consumidores da atividade residencial com alto consumo de energia.

A classe residencial foi a única dividida entre dois *clusters*, o que pode ser justificado pelo fato de 85% dos consumidores serem da classe residencial, o que faz com a classe apresente a maior heterogeneidade de indivíduos.

Cada um dos *clusters* obtidos foi considerado um base de dados independente e foi utilizado para treinar novamente os classificadores listados na Tabela 14, com exceção do SVM, que não foi incluído devido ao seu custo computacional proibitivo. Portanto, cada classificador passa por quatro processos de treinamento distintos, e ao final os resultados são agregados para compor a classificação do conjunto total, como descrito na subseção 4.2.2. Em seguida, a validação cruzada *k-fold* é aplicada para avaliar o desempenho dos novos modelos obtidos. Os resultados da validação cruzada para a classificação segmentada são apresentados na Tabela 15.

Tabela 15 – Resultado da validação cruzada para a classificação segmentada.

Modelo	**Limiar de Classificação**	**Precisão**	**Sensibilidade**	**F1-*score***
RF_segmentado	0,24 – 0,34	46,05%	41,26%	0,44
GBT_segmentado	0,18 – 0,22	43,14%	43,70%	0,43
RTF_segmentado	0,16 – 0,33	42,84%	43,46%	0,43
BT_segmentado	0,17 – 0,27	41,73%	44,34%	0,43
XGBT_segmentado	0,46 – 0,61	42,23%	43,49%	0,43
LR_segmentado	0,59 – 0,68	37,38%	43,88%	0,40
DT_segmentado	0,11 – 0,21	42,67%	36,93%	0,40
MLP_segmentado	0,70 – 0,85	28,44%	51,45%	0,37
NB_segmentado	0,16 – 0,37	29,86%	41,10%	0,35
FR_segmentado	0,01 – 0,50	16,85%	52,33%	0,25

Fonte: Autoria Própria.

A primeira observação que pode ser feita a partir da análise da Tabela 15 é com relação ao limiar de classificação dos modelos, neste caso, é apresentado um intervalo de valores ao invés de um valor único. Isto deve-se ao fato de que os modelos representam a composição de quatro classificações distintas e independentes, correspondentes aos quatros *clusters*. Portanto, a classificação de cada *cluster* apresenta um valor diferente de limiar, e a classificação final composta terá quatro valores diferentes de limiar, os quais estão delimitados pelo intervalo apresentado na tabela.

Os resultados na Tabela 15 estão listados em ordem decrescente da métrica F1-*score* e, como pode ser notado a partir da análise da tabela, apesar de haver mudanças na ordem de apresentação dos modelos, os resultados podem ser considerados semelhantes aos da Tabela 14. Pois, os métodos de *ensemble* continuam superando os demais algoritmos e os valores de F1-*score* são apenas um pouco menores do que os da Tabela 14. O resultado demonstra, portanto, que a classificação segmentada não foi capaz de melhorar o desempenho dos modelos de classificação para identificação da perda não técnica.

Contudo, ainda é possível construir um novo classificador a partir da seleção dos melhores resultados obtidos em cada *cluster*. O procedimento foi descrito na subseção 4.2.2 a partir da ilustração da Figura 38. O novo modelo será denominado M.Seleção e o seu resultado de validação cruzada é apresentado na Tabela 16.

Tabela 16 – Resultado e composição do modelo M.Seleção.

Cluster	**Modelo**	**Limiar de classificação**	**Precisão**	**Sensibilidade**	**F1-*score***
01	RF	0,34	55,30%	58,60%	0,57
02	RF	0,24	39,54%	35,33%	0,37
03	RTF	0,23	44,93%	35,47%	0,40
04	XGBT	0,61	44,58%	47,03%	0,46
-	M.Seleção	0,23 - 0,61	45,20%	42,77%	0,44

Fonte: Autoria própria.

A partir da análise da Tabela 16, constata-se que o modelo RF foi o melhor classificador para os Clusters 01 e 02, o modelo RTF foi o melhor classificador para o Cluster 03 e, por fim, o modelo XGBT foi o melhor classificador para o Cluster 04. Mais uma vez, os melhores classificadores são todos do tipo *ensemble*, o que corrobora a conclusão já discutida de que os métodos *ensemble* são os mais adequados para a identificação da PNT.

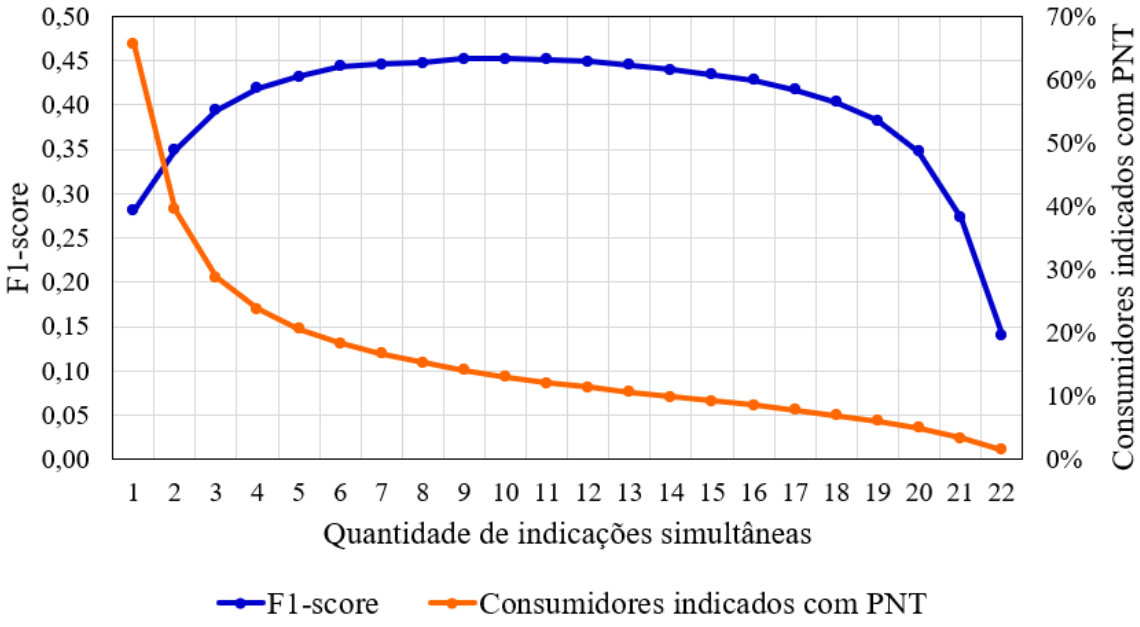
O valor da métrica F1-score nos *clusters* individuais variou entre 0,57 no Cluster 01 e 0,37 no Cluster 02. A composição dos resultados dos *clusters*, representada pelo modelo M.Seleção apresentou um F1-*score* de 0,44, o qual é menor que o valor para o modelo GBT na Tabela 14. Portanto, conclui-se que mesmo com a construção do modelo M.Seleção, a segmentação do banco de dados não foi capaz melhorar o desempenho da classificação dos consumidores.

No entanto, a segmentação dos dados ainda pode ser útil em cenários específicos. Como pode ser notado na Tabela 16, o desempenho do Cluster 01 é significativamente melhor do que

o caso geral. Assim, em um cenário em que haja interesse de se identificar a PNT nesse grupo específico de consumidores, a classificação segmentada proporcionará um melhor resultado.

Até o momento, 22 classificadores diferentes foram avaliados (11 na Tabela 14, 10 na Tabela 15 e 1 na Tabela 16), contudo, ainda é possível a construção do 23º classificador, a partir de um critério de votação que considera a coincidência de resultados em todos os demais 22 modelos, o qual foi descrito na subseção 4.2.3. O novo modelo será denominado M.Votação e o seu resultado é apresentado na Figura 51.

Figura 51 – Resultado do modelo M.Votação.



Fonte: Autoria própria.

O eixo horizontal do gráfico na Figura 51 representa o número de indicações simultâneas dos 22 classificadores distintos, que são consideradas para assumir que um consumidor possui PNT no modelo M.Votação. Por exemplo, o valor 22 no eixo horizontal indica que um consumidor só é considerado com PNT no modelo M.Votação, caso ele tenha recebido a indicação de PNT em todos os demais 22 classificadores. Da mesma maneira, o valor 1 no eixo horizontal indica que um consumidor é considerado com PNT no modelo M.Votação, caso ele tenha recebido a indicação de PNT em pelo menos um dos demais 22 modelos.

O eixo vertical esquerdo do gráfico na Figura 51 indica o valor da métrica F1-*score* para o modelo M.Votação, o valor da métrica em cada ponto do modelo é representado pela curva azul no gráfico. Por fim, o eixo vertical direito indica a porcentagem do total de consumidores que foram indicados com PTN pelo modelo M.Votação, a percentagem para cada ponto do modelo é representada pela curva laranja no gráfico.

A partir da análise do gráfico na Figura 51, constata-se que quanto maior o número de indicações simultâneas consideradas, mais restritivo é o modelo M.Votação. Ou seja, ao se considerar que um consumidor possui PNT quando foi indicado por pelo menos um classificador, 66% de todos os consumidores da base de dados são indicados com PNT. Por outro lado, ao se considerar que um consumidor possui PNT quando foi indicado por todos os 22 classificadores, apenas 2% de todos os consumidores da base de dados são indicados com PNT.

Contudo, mais restritivo não significa melhor desempenho, pois, como pode ser observado na curva azul, o maior desempenho é obtido no meio do gráfico, quando se considera uma quantidade de nove classificadores simultâneos para indicação da PNT. Nesse ponto, o valor da métrica F1-*score* é de 0,45, o que representa o mesmo valor do modelo GBT na Tabela 14. Portanto, a partir dos resultados apresentados, pode-se concluir que o critério de votação utilizado no modelo M.Votação não proporcionou melhoria significativa no desempenho obtido para identificação da PNT no conjunto de consumidores.

Os resultados dos 23 modelos analisados até o momento foram obtidos a partir do processo de validação cruzada, o qual representa um resultado simulado a partir da base de dados disponível. Contudo, para garantir robustez às conclusões da pesquisa, torna-se necessário avaliar o quão fidedigno são os resultados em aplicações reais. Para tanto, os modelos desenvolvidos foram utilizados para selecionar novos consumidores em inspeções reais de campo. As inspeções foram executadas por equipes especializadas da concessionária que disponibilizou a base de dados para a pesquisa, conforme descrito na seção 4.5.

Como discutido na seção 4.5, a concessionária apresentou algumas restrições para realização das inspeções. Primeiramente, não é desejável fiscalizar consumidores regulares, portanto, apenas consumidores indicados com PNT pelos classificadores foram enviados para as equipes de campo. Em segundo lugar, as equipes de campo têm uma capacidade limitada de trabalho. Portanto, apenas 200 consumidores foram inspecionados para cada classificador. Como um consumidor pode ter sido indicado por mais de 1 classificador, o número total de consumidores inspecionados foi de 1.349. Os resultados das inspeções são apresentados na Tabela 17.

Os modelos na Tabela 17 estão listados em ordem decrescente de acordo com o valor da precisão nas inspeções de campo. Como discutido na seção 4.5, devido ao fato de apenas os consumidores indicados com PNT terem sido inspecionados, não é possível determinar o valor da métrica F1-*score* para as inspeções em campo. Por isso, a métrica utilizada para avaliação dos resultados é a precisão. Além disso, a métrica *Mean Absolute Percentage Error* (MAPE) é

utilizada para comparar o desvio entre os valores da precisão na validação cruzada e nas inspeções em campo.

Tabela 17 – Resultado da classificação para as inspeções em campo.

Modelo	Limiar de classificação	Precisão		Desvio
		Validação Cruzada	Inspeção em campo	
RTF	0,56	71,06%	66,50%	6,86%
RTF_segmentado	0,56	69,44%	66,00%	5,21%
RF	0,54	71,87%	65,00%	10,56%
RF_segmentado	0,53	70,94%	65,00%	9,13%
M.Votação	-	70,83%	63,53%	11,49%
BT	0,53	68,44%	57,00%	20,06%
XGBT_segmentado	0,87	66,70%	56,00%	19,10%
BT_segmentado	0,54	67,66%	53,50%	26,46%
XGBT	0,87	66,82%	53,00%	26,08%
GBT	0,54	68,44%	52,00%	31,61%
M.Seleção	0,86	65,18%	51,00%	27,81%
DT	0,78	62,61%	49,50%	26,48%
SVM	0,60	64,73%	46,50%	39,20%
NB	0,93	53,51%	46,00%	16,34%
LR	0,91	56,13%	45,50%	23,37%
NB_segmentado	0,93	54,36%	44,50%	22,15%
DT_segmentado	0,72	60,86%	43,00%	41,53%
GBT_segmentado	0,62	68,50%	41,50%	65,05%
MLP	0,93	44,42%	40,50%	9,67%
LR_segmentado	0,91	52,94%	39,50%	34,03%
FR_segmentado	0,99	18,53%	18,00%	2,96%
MLP_segmentado	0,99	30,34%	14,50%	109,27%
FR	0,99	25,19%	10,50%	139,92%

Fonte: Autoria Própria.

A partir da análise da Tabela 17, constata-se inicialmente que os valores do limiar de classificação para os modelos são maiores que os apresentados nas Tabela 14, Tabela 15 e Tabela 16. A alteração foi necessária devido à limitação no número de consumidores selecionados. Como deseja-se selecionar apenas 200 consumidores em cada modelo, o valor do limiar de classificação foi elevado para tornar os modelos mais restritivos, até o ponto em que apenas 200 consumidores fossem indicados com PNT, os quais correspondem aos 200 consumidores com maior probabilidade de ocorrência de PNT no conjunto.

Para que a comparação dos resultados entre a validação cruzada e as inspeções em campo seja coerente, é necessário utilizar o mesmo limiar de classificação para os modelos nos dois casos. Por isso, o processo de validação cruzada foi executado novamente com a alteração dos limiares de classificação dos modelos. Conforme visto na Figura 37, quanto maior o limiar de classificação de um modelo, maior será sua precisão, por isso, os valores de precisão para a validação cruzada apresentados na Tabela 17 são maiores que os valores nas Tabela 14, Tabela 15 e Tabela 16.

A análise da Tabela 17 também revela que, nas inspeções em campo, os métodos do tipo *ensemble* mais uma vez apresentaram os melhores resultados de classificação. Destaca-se o modelo RTF que apresentou o melhor resultado de campo, com 66,50% de precisão, ou seja, 133 dos 200 consumidores inspecionados apresentaram PNT. O modelo RTF também apresentou uma boa fidedignidade, ou seja, o valor da precisão das inspeções em campo é próximo do valor simulado na validação cruzada, com um desvio entre os dois valores de apenas 6,86%.

O modelo RTF_segmentado apresentou um resultado de campo próximo ao RTF, com uma precisão de 66% e um desvio de 5,21% em relação ao valor da precisão na validação cruzada. O mesmo comportamento pode ser observado para os modelos RF e RF_segmentado, os quais apresentaram precisão de campo iguais a 65%, mas, o modelo segmentado apresentou um desvio um pouco menor em relação ao resultado da validação cruzada. Os resultados sugerem que, embora o processo de segmentação não melhore o desempenho da classificação, ele pode melhorar a fidedignidade dos modelos em alguns casos.

Seguindo a ordem na Tabela 17, o modelo M.Votação aparece com uma precisão de 63,53% para as inspeções em campo e também possui uma boa fidedignidade, com um desvio de 11,49% em relação à validação cruzada. O resultado demonstra que, embora o critério de votação não supere os melhores modelos, ele ainda é melhor do que alguns dos classificadores mais populares na bibliografia, como o SVM, MLP, NB e DT.

Por fim, constata-se com a análise da Tabela 17 que alguns classificadores que apresentaram bom desempenho nos resultados de validação cruzada, não mantiveram o desempenho nas inspeções em campo, é o caso dos modelos XGBT e GBT. A razão é que esses classificadores apresentaram baixa fidedignidade, com desvios de 26,08% e 31,61% entre resultados de campo e validação cruzada, respectivamente. O resultado demonstra a importância da realização das inspeções em campo para a avaliação da fidedignidade dos modelos. A fidedignidade é uma característica muito importante, pois, demonstra em que

medida os resultados obtidos a partir de simulações de conjuntos de dados podem ser extrapolados para aplicações reais de campo.

Considerando todas as análises realizadas nesta seção de resultados, conclui-se que o algoritmo RTF pode ser considerado como o classificador mais adequado para fins de identificação de PNT no sistema de distribuição. Pois, o modelo foi capaz de combinar bom desempenho, alta fidedignidade e razoável custo computacional. Destaca-se ainda que o RTF é seguido de perto pelos outros modelos do tipo *ensemble*, que também podem ser considerados soluções adequadas para a identificação da PNT.

Na seção seguinte são apresentados os resultados obtidos na Etapa 3 da metodologia, os quais se referem aos modelos de regressão, construídos com objetivo de prever o valor da energia recuperada em inspeções de campo.

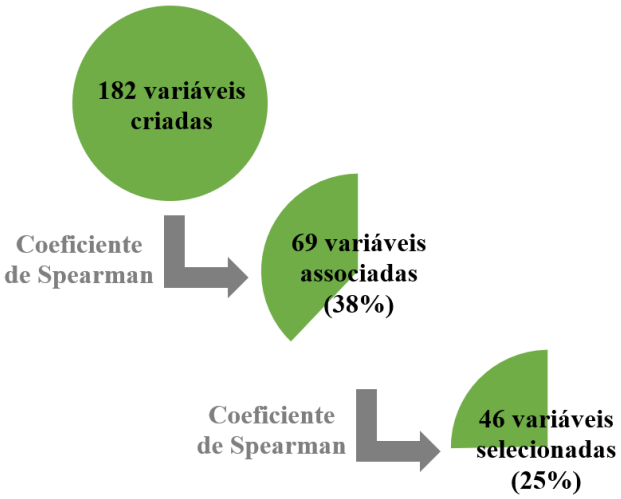
5.3 Previsão da Energia Recuperada

Os procedimentos realizados na etapa de previsão da energia recuperada são semelhantes aos realizados na etapa de classificação dos consumidores discutidos na seção anterior. As principais diferenças foram discutidas na seção 4.3 e se referem ao fato de que os modelos no caso em questão são do tipo regressão, as variáveis selecionadas são diferentes e a métrica de avalição é o RMSE ao invés do F1-*score*.

Inicialmente, foi determinada a associação entre todas as 182 variáveis criadas na Etapa 1 e a variável energia recuperada, por meio do coeficiente de Spearman. Posteriormente, para identificar a redundância de informações, o coeficiente de Spearman foi calculado também entre as variáveis que possuem associação. Ao final, foram selecionadas apenas as variáveis independentes e associadas à energia recuperada. A relação de todas as variáveis selecionadas é apresentada na Tabela 23 do Apêndice A ao final do texto. Adicionalmente, o resultado da seleção de variáveis para os modelos de regressão é ilustrado na Figura 52.

A partir da análise da Figura 52 constata-se que foram identificadas 69 variáveis com algum nível de associação à energia recuperada, o que representa 38% do total de variáveis disponíveis. Contudo, 23 pares de variáveis apresentaram redundância de informações, por isso, foram eliminadas da base de dados as variáveis que tinham menor associação com a energia recuperada. Assim, restaram 46 variáveis independentes na base dados, o que representa 25% do total de variáveis disponíveis.

Figura 52 – Resultado do processo de seleção de variáveis para previsão da energia recuperada.



Fonte: Autoria própria.

O conjunto de 46 variáveis foi utilizado para o treinamento dos sete modelos listados na Tabela 4 da subseção 2.2.5 que podem ser utilizados em problemas de regressão. Porém, inicialmente foram determinados os hiperparâmetros ótimos dos modelos com o método de otimização Bayesiana. Os valores dos hiperparâmetros ótimos obtidos são apresentados na Tabela 18.

Tabela 18 – Hiperparâmetros ótimos definidos na otimização Bayesiana para a regressão.

Modelo	Hiperparâmetro	Valor ótimo
DT	Número mínimo de derivações por nó	2
	Critério para derivação	*Gini Index*
MLP	Número de camadas escondidas	10
	Número de neurônios por camada	48
GBT	Número de árvores	89
	Número de níveis das árvores	3
XGBT	Número de árvores	100
	Número de níveis das árvores	6
	Número máximo de *bins*	256
BT	*Bag size*	87%
RF	Número de árvores	100
RTF	Tamanho do grupo de variáveis	2
	Taxa de remoção	27%

Fonte: Autoria própria.

Utilizando-se os hiperparâmetros ótimos apresentados na Tabela 18 e o conjunto de 46 variáveis selecionadas, foram realizados os treinamentos dos modelos de regressão.

Posteriormente, foi executada a validação cruzada dos modelos com o método *k-fold*. Os resultados do processo de validação cruzada são apresentados na Tabela 19.

Tabela 19 – Resultado da validação cruzada para os modelos de regressão.

Modelo	RMSE	kWh Total (Real)	kWh Total (Previsto)	Desvio
XGBT	$4{,}94\times10^{-2}$	5.423.413	5.178.763	-4,51%
BT	$4{,}95\times10^{-2}$	5.423.413	5.496.140	1,34%
RTF	$5{,}03\times10^{-2}$	5.423.413	5.371.502	-0,96%
RF	$5{,}04\times10^{-2}$	5.423.413	5.753.474	6,09%
GBT	$5{,}50\times10^{-2}$	5.423.413	4.348.145	-19,83%
MLP	$5{,}62\times10^{-2}$	5.423.413	5.413.491	-0,18%
DT	$7{,}21\times10^{-2}$	5.423.413	6.321.571	16,56%

Fonte: Autoria própria.

Os modelos na Tabela 19 estão apresentados na ordem crescente dos valores do erro RMSE. Como pode ser visto, o menor valor de erro foi obtido pelo modelo XGBT, o qual é seguido de perto pelo modelo BT, com erros de $4{,}94\times10^{-2}$ e $4{,}95\times10^{-2}$ respectivamente. Na outra extremidade da Tabela 19, o modelo DT aparece com o maior valor de erro, o qual é 46% maior que no modelo XGBT. Os demais modelos (RTF, RF, GBT e MPL) aparecem com valores intermediários de erro entre $5{,}03\times10^{-2}$ e $5{,}62\times10^{-2}$.

A partir dos resultados discutidos no parágrafo anterior, nota-se que os modelos do tipo *ensemble* também se sobressaem em problemas de regressão. Contudo, é necessário fazer a ressalva de que neste caso apenas os modelos MLP e DT não são do tipo *ensemble*, o que torna a comparação menos robusta do que a realizada no problema de classificação da seção anterior, em que havia uma diversidade maior de modelos.

Com relação à ordem de grandeza do erro RMSE, é importante destacar que os valores de energia recuperada utilizados para o treinamento dos modelos foram normalizados entre 0,0 e 1,0, pois, como discutido na subseção 2.2.1.4, a normalização é um procedimento que pode melhorar o desempenho de modelos preditivos. Diante desta informação, os valores do erro RMSE na ordem de 10^{-2} que são apresentados na Tabela 19 podem ser considerados significativos em uma escala de 0,0 a 1,0.

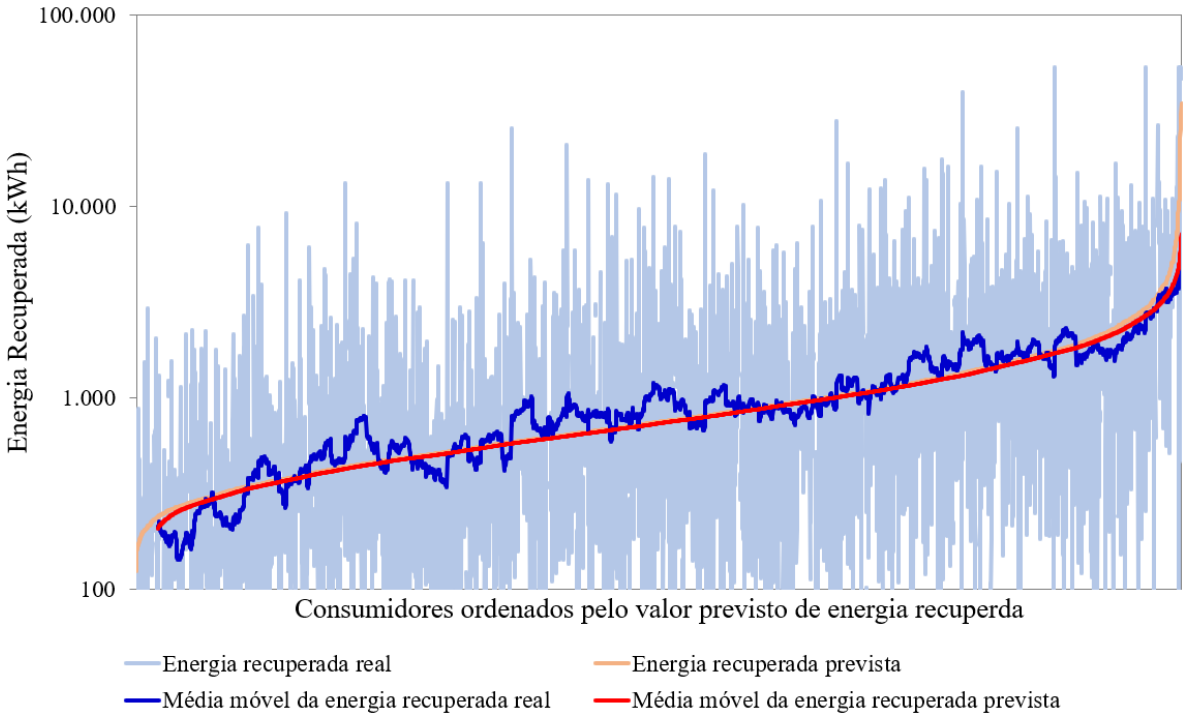
Isto significa que para um consumidor individual, o desvio entre os valores previsto e real de energia recuperada podem ser elevados. Entretanto, a análise das últimas colunas da Tabela 19 demonstra que quando se trata do valor total de energia recuperada, os desvios são pequenos. O valor total é obtido a partir da soma da energia recuperada em cada consumidor da base de dados e conforme apresentado, com exceção dos modelos GBT e DT, os resultados

previstos apresentaram um desvio relativamente pequeno, variando entre – 4,51% e 6,09% em relação ao valor real.

Portanto, a partir dos resultados da Tabela 19 pode-se concluir que os modelos de regressão obtidos são melhores em prever o valor total da energia recuperada em um conjunto de consumidores, do que em prever o valor individual em cada consumidor.

Com objetivo de aprofundar a análise dos resultados da regressão, o modelo com melhor desempenho na Tabela 19 (XGBT) foi escolhido para elaboração do gráfico que está apresentado na Figura 53, no qual os valores real e previsto da energia recuperada são apresentados para todos os consumidores utilizados no processo de validação cruzada. Para facilitar a visualização do gráfico, os consumidores foram ordenados na ordem crescente do valor previsto de energia recuperada.

Figura 53 – Valores real e previsto de energia recuperada para todos os consumidores no processo de validação cruzada do modelo XGBT.



Fonte: Autoria própria.

Como pode ser observado, no gráfico na Figura 53 são apresentados os valores de energia recuperada real (curva azul claro), energia recuperada prevista (curva vermelho claro), média móvel da energia recuperada real (curva azul escuro) e média móvel da energia recuperada prevista (curva vermelho escuro).

A partir da comparação das curvas de energia recuperada real e energia recuperada prevista, verifica-se que na maioria dos casos existe um desvio significativo entre os dois

valores. Isto deve-se ao fato de que, como já discutido anteriormente, o valor do erro RMSE entre as duas séries é significativo.

Entretanto, a análise do gráfico também demonstra que a previsão foi capaz de diferenciar grupos de consumidores com baixos valores de energia recuperada daqueles com altos valores de energia recuperada. A observação fica evidente na comparação dos valores nas extremidades esquerda e direita do gráfico, como a escala vertical do gráfico é logarítmica, os valores à direita são 100 vezes maiores do que os valores à esquerda.

O aspecto destacado no parágrafo anterior é corroborado pela análise das curvas de média móvel no gráfico, a partir da qual consta-se que a média móvel dos valores reais possui um bom ajuste com a média móvel dos valores previstos. De fato, o coeficiente de determinação (R^2) das duas curvas de média móvel é 0,91, enquanto o máximo valor possível para R^2 é 1,0, o qual indicaria um ajuste perfeito das curvas, como discutido na subseção 2.2.7.3.

Portanto, a partir dos resultados discutidos na Figura 53, conclui-se que apesar do modelo de regressão desenvolvido não ser capaz de prever de modo satisfatório o valor de energia recuperada para consumidores individuais, ele é capaz de prever de modo satisfatório a recuperação de energia em nível de grupo, o que é suficiente para os propósitos da pesquisa.

Assim como foi feito para os modelos de classificação, os resultados dos modelos de regressão também foram testados em aplicações reais de campo. Para tanto, foram utilizadas as mesmas inspeções realizadas nos 1.349 consumidores e discutidas na seção anterior. Assim, os modelos de regressão listados na Tabela 19 foram aplicados aos novos consumidores com objetivo de prever o valor de energia recuperada com as inspeções de campo. Os resultados obtidos são apresentados na Tabela 20.

Tabela 20 – Resultado dos modelos de regressão para as inspeções de campo.

Modelo	RMSE	kWh Total (Real)	kWh Total (Previsto)	Desvio
XGBT	$4{,}76\times10^{-2}$	386.286	397.956	3,02%
BT	$4{,}78\times10^{-2}$	386.286	421.304	9,07%
GBT	$4{,}83\times10^{-2}$	386.286	327.124	- 15,32%
RTF	$4{,}85\times10^{-2}$	386.286	376.027	- 2,66%
RF	$4{,}89\times10^{-2}$	386.286	437.496	13,26%
MLP	$4{,}96\times10^{-2}$	386.286	293.695	- 23,97%
DT	$8{,}00\times10^{-2}$	386.286	667.840	72,89%

Fonte: Autoria própria.

A partir das irregularidades identificadas com as inspeções, a concessionária calculou o valor de energia a ser recuperada dos consumidores, o qual corresponde a 386.286 kWh e é

apresentado na terceira coluna da Tabela 20. O modelo que mais se aproximou do valor calculado pela concessionária foi o RTF, que apresentou um desvio de apenas – 2,66%, seguido de perto pelo modelo XGBT, com um desvio de 3,02%. Os modelos BT, GBT, RF e MLP apresentaram valores intermediários de desvio, variando entre – 23,97% e 9,07%. Já o modelo DT apresentou um desvio significativo de 72,89%, o que o torna um método não recomendável para previsão do valor de energia recuperada.

Na segunda coluna da Tabela 20 são apresentados os valores do erro RMSE em ordem crescente. Como pode ser observado, a ordem é semelhante à da Tabela 19, com o modelo XGBT apresentando o menor valor de erro. A única alteração é o modelo GBT que subiu duas posições. Além disso, os valores do erro RMSE na Tabela 20 também são semelhantes aos apresentados na Tabela 19, o que demonstra que os resultados simulados tem boa fidedignidade, com exceção do modelo DT que apresentou um desvio em relação ao valor total da energia recuperada muito maior que o da Tabela 19.

A partir dos resultados discutidos nesta seção, é possível concluir que dentre os modelos de regressão avaliados, o XGBT se apresentou como a melhor opção para a previsão da energia recuperada, por combinar o menor valor de erro RMSE, um baixo valor de desvio em relação ao valor total da energia recupera e uma boa fidedignidade entre o desempenho medido nas simulações e nos testes reais em campo.

Com os resultados dos modelos de classificação e regressão obtidos até o momento, é possível avançar para a última etapa da metodologia, que corresponde a Etapa 4 discutida na seção 4.4. O principal objetivo da Etapa 4 é a maximização do retorno financeiro obtido com as inspeções de campo e os resultados alcançados são apresentados na seção seguinte.

5.4 Maximização do Retorno Financeiro

Na etapa de maximização do retorno financeiro os melhores modelos de classificação e de regressão identificados nas etapas anteriores são utilizados em conjunto com a Equação (34), a qual foi apresentada na seção 4.4 e modela o potencial de retorno financeiro das inspeções. A partir dos resultados dos modelos e da equação é possível determinar o conjunto ótimo de consumidores que devem ser inspecionados para maximizar o retorno financeiro da concessionária.

Para a análise de maximização, inicialmente, foi determinado para todos os consumidores da base dados um valor de probabilidade para a presença de PNT (P_{PNT}), e uma

previsão para a energia recuperada em uma inspeção em campo ($P_{E_{kWh}}$). Para tanto, foram utilizados os modelos RTF para classificação e XGBT para regressão, por terem sido considerados os melhores modelos nos resultados das seções anteriores.

Na sequência, os valores de P_{PNT} e $P_{E_{kWh}}$ foram aplicados na Equação (34) para determinar o valor do potencial de retorno financeiro de cada consumidor (R_P). A partir do valor de R_P é possível realizar a análise de maximização, a qual pode ser aplicada em dois cenários distintos, conforme discutido na seção 4.4:

- Cenário 1: a infraestrutura para gestão da PNT é fixa;
- Cenário 2: a infraestrutura para gestão da PNT é passível de dimensionamento.

No Cenário 1 a quantidade de consumidores que podem ser inspecionados é pré-determinada pela infraestrutura disponível. Assim, a maximização é realizada pela ordenação dos consumidores em ordem decrescente do valor R_P.

Uma maneira de analisar como a maximização é aplicada ao Cenário 1, é comparando os resultados de duas seleções distintas de consumidores. Na primeira, denominada "Seleção 1" os consumidores são ordenados em ordem decrescente do valor P_{PNT}. Na segunda, denominada "Seleção 2" os consumidores são ordenados em ordem decrescente do valor R_P. A Seleção 1 representa a abordagem clássica de seleção para inspeções de campo, em que a probabilidade de ocorrência de PNT é o viés utilizado. Já a Seleção 2 representa a nova abordagem proposta, em que o foco é o retorno financeiro das inspeções.

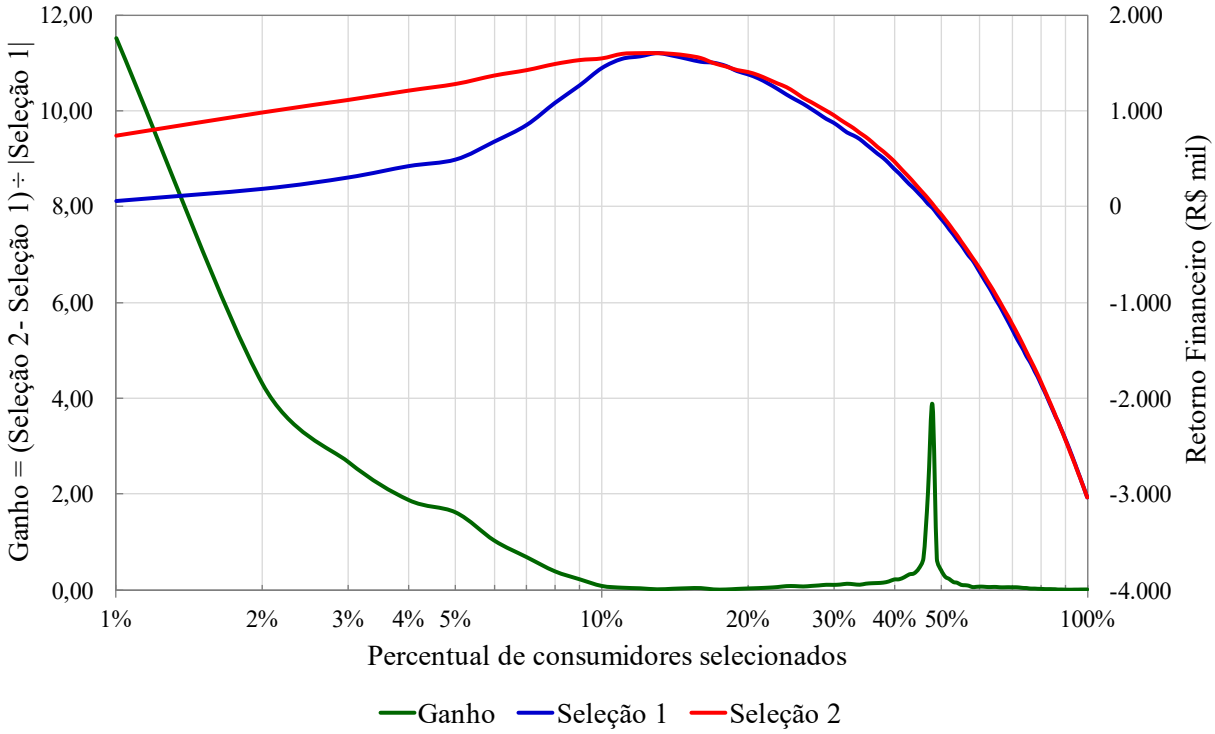
O resultado da análise do retorno financeiro para o Cenário 1 é exibido no gráfico da Figura 54, no qual estão apresentados os valores do retorno financeiro para a Seleção 1 (curva azul) e para a Seleção 2 (curva vermelha), além do ganho obtido na Seleção 2 em relação a Seleção 1 (curva verde).

Conforme afirmado anteriormente, no Cenário 1 a porcentagem de consumidores selecionados para inspeção é pré-estabelecida pela infraestrutura disponível na concessionária. Contudo, como a infraestrutura varia de acordo com a concessionária, optou-se por apresentar no gráfico da Figura 54 o intervalo de seleção de 1% a 100% do total de consumidores, assim, é possível verificar todas as situações possíveis.

Como pode ser observado na Figura 54, quando o percentual de seleção é de 1% dos consumidores, o retorno financeiro na Seleção 1 é de R$ 59.492, enquanto na Seleção 2 é de R$ 745.350. Neste caso, o ganho obtido na Seleção 2 em relação a Seleção 1 é de 11,53 vezes. Se o percentual de seleção for definido em 2%, o ganho é de 4,30 e continua significativo até o percentual de 10%, depois do qual o retorno financeiro em ambos os casos é semelhante,

embora a Seleção 2 permaneça com um retorno maior para todos os percentuais. Na porcentagem de 48%, o retorno financeiro da Seleção 1 torna-se negativo, enquanto o retorno da Seleção 2 permanece positivo, por isso, a curva de ganho (curva verde) apresenta uma breve elevação nesta porcentagem.

Figura 54 – Retorno financeiro no Cenário 1: comparação entre a Seleção 1 e Seleção 2 de acordo com o percentual de consumidores selecionados.

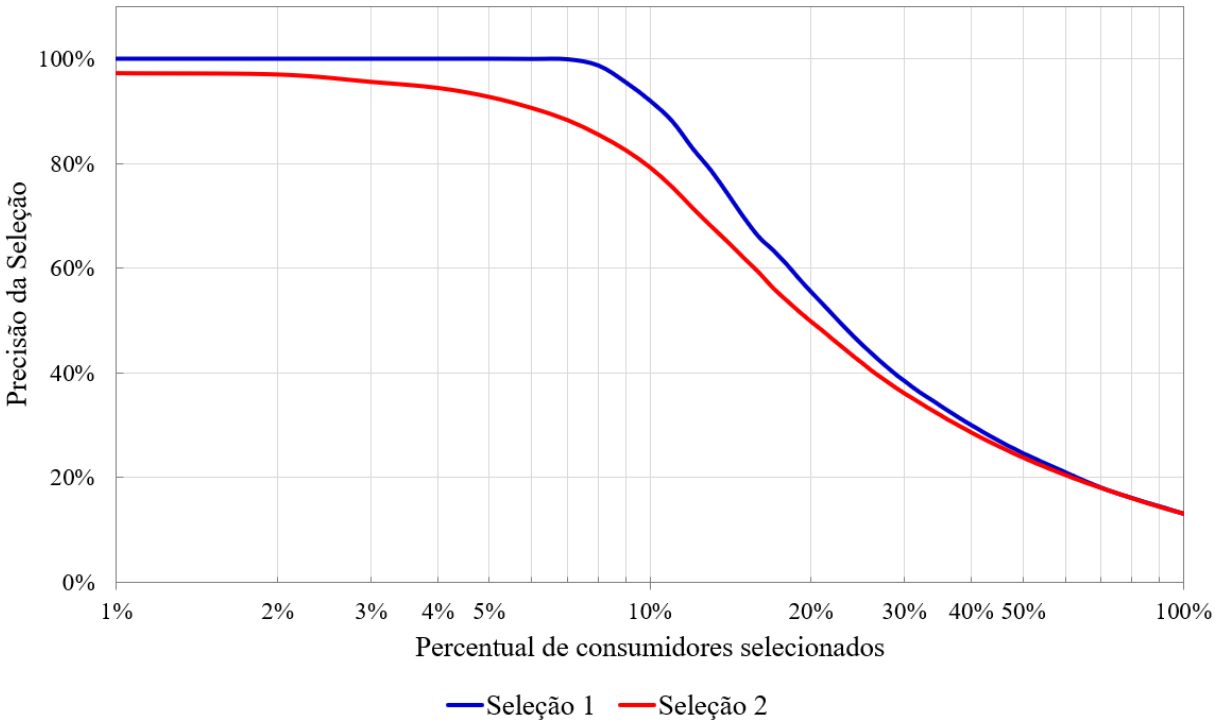


Fonte: Autoria própria.

A partir da análise no parágrafo anterior é possível concluir que a abordagem proposta melhora significativamente o retorno financeiro das fiscalizações, especialmente em situações em que o percentual de consumidores fiscalizados é inferior a 10%. Pode-se concluir ainda que quanto menor for o número de consumidores inspecionados, mais relevante torna-se o ganho proporcionado pela estratégia de maximização. Desse modo, a aplicação da metodologia proposta torna-se especialmente interessante em concessionárias que possuem infraestrutura de gestão de perdas pequena.

Como comentado anteriormente, o foco da Seleção 2 é o retorno financeiro, o que ocorre em detrimento da precisão da seleção. Por isso, embora o retorno financeiro seja maior na Seleção 2, espera-se que este caso tenha uma precisão menor em relação a Seleção 1, uma vez que a probabilidade de ocorrência de PNT foi o viés de seleção no segundo caso. A expectativa é confirmada na Figura 55, na qual a precisão da seleção é apresentada para ambos os casos de acordo com a porcentagem de consumidores selecionados.

Figura 55 – Comparação entre a precisão da Seleção 1 e Seleção 2 de acordo com o percentual de consumidores selecionados.



Fonte: Autoria própria.

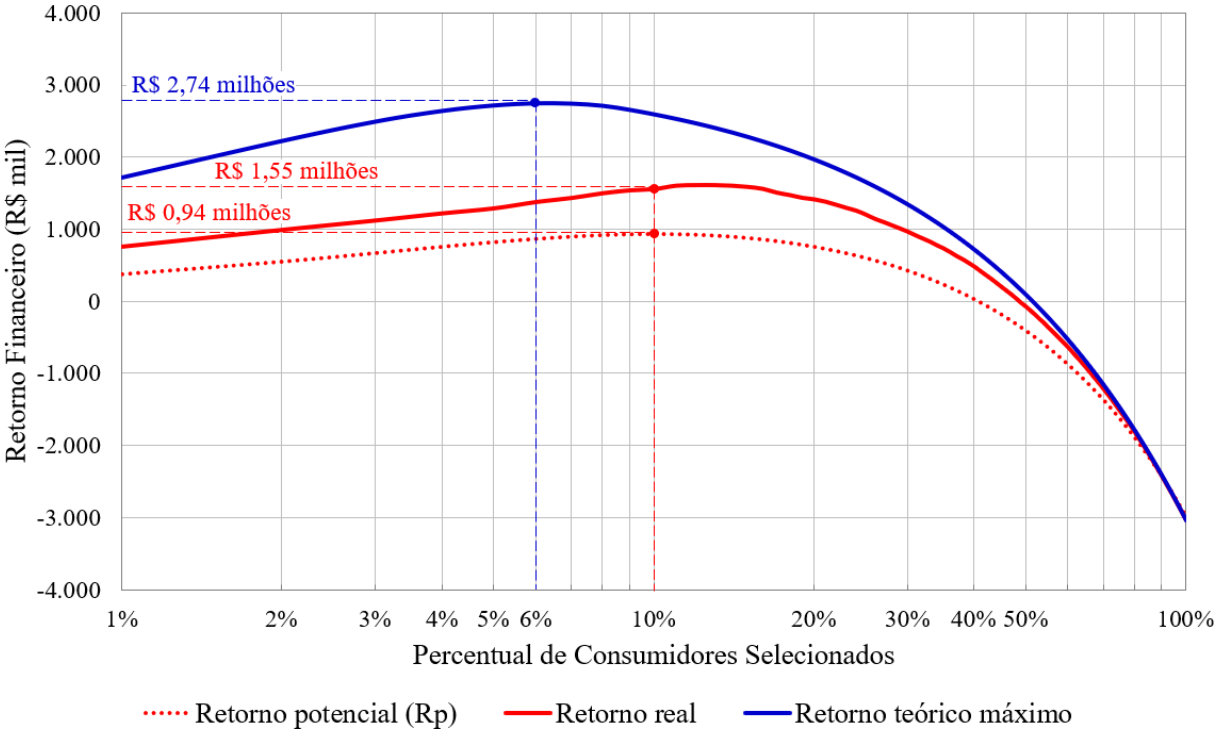
Com a análise na Figura 55, observa-se que a Seleção 1 tem uma precisão maior em comparação a Seleção 2 para todas as porcentagens de seleção. Além disso, na Seleção 1 é possível alcançar uma precisão próxima de 100% ao selecionar menos de 7% dos consumidores para inspeções, ou seja, todos os consumidores selecionados apresentarão PNT caso o grupo seja menor que 7% do total. Este resultado indica o bom desempenho do classificador *Rotation Forest* para a indicação dos consumidores com alta probabilidade de possuir PNT.

Com relação ao Cenário 2, tem-se que a infraestrutura de gestão da perda é passível de dimensionamento e, portanto, a quantidade de inspeções realizadas pode ser considerada uma variável adicional a ser definida na análise de maximização. Assim, no Cenário 2 o objetivo da análise é determinar o percentual de consumidores que, quando inspecionados, fornecem o máximo retorno financeiro à concessionária. O resultado da análise de maximização para o Cenário 2 é apresentado na Figura 56.

No gráfico da Figura 56 o percentual de seleção no eixo horizontal considera os consumidores ordenados em ordem decrescente do valor do retorno potencial R_P (curva vermelha pontilhada), o qual foi determinado com base na Equação (34), assim como feito para a análise do Cenário 1. O valor do retorno real (curva vermelha contínua) corresponde ao valor efetivamente obtido com as inspeções em campo e, por fim, o valor do retorno teórico máximo

(curva azul) corresponde ao valor que seria obtido caso as inspeções fossem executadas exatamente na ordem decrescente do retorno financeiro real.

Figura 56 – Retorno financeiro no Cenário 2 de acordo com o percentual de consumidores selecionados.



Fonte: Autoria própria.

O retorno teórico máximo representa a seleção perfeita de consumidores, em que a energia que pode ser recuperada em cada consumidor é conhecida antes da realização das inspeções. Portanto, corresponde a uma situação teórica, impossível de ser alcançada em aplicações reais. O valor é apresentado no gráfico apenas como uma referência para medir o quão próximo a abordagem proposta na pesquisa está do máximo teórico que poderia ser alcançado.

A partir da análise da Figura 56, verifica-se que na curva do retorno potencial o valor máximo ocorre para um percentual de seleção de 10% dos consumidores. Neste ponto, o potencial de retorno financeiro calculado é R$ 0,94 milhões, enquanto que o valor real do retorno financeiro no mesmo ponto é R$ 1,55 milhões. O fato do valor real ser 65% superior ao valor potencial previsto indica que o modelo está subestimando o retorno financeiro das inspeções, o que não é necessariamente um problema para aplicações reais, pois, o retorno será maior do que o planejado. Além disso, nota-se que as curvas de retorno potencial e retorno real apresentam comportamentos semelhantes, o que é suficiente para identificação do ponto máximo a partir da curva de retorno potencial.

A análise da Figura 56 também revela que o retorno teórico máximo para a base de consumidores em questão é de R$ 2,74 milhões, e poderia ser obtido com a inspeção de 6% dos consumidores. O resultado demonstra, portanto, que o ponto ótimo de inspeção obtido a partir da metodologia proposta foi capaz de atingir 57% do máximo valor teórico do retorno financeiro, o que pode ser considerado um bom resultado, visto que a situação teórica é uma utopia.

A análise do gráfico revela ainda que o retorno financeiro das inspeções torna-se negativo quando o número de consumidores selecionados é maior que 49% do total, podendo atingir um saldo negativo de – R$ 3,03 milhões caso 100% dos consumidores fossem inspecionados. O resultado demonstra que é inviável a realização de inspeções de forma ostensiva e indiscriminada nos consumidores, por isso, a importância de que as concessionárias disponham de uma metodologia robusta e efetiva para a seleção dos consumidores, como a apresentada nesta pesquisa.

Por fim, assim como nas etapas anteriores da metodologia, os resultados de maximização do retorno financeiro foram validados em uma aplicação real de campo. Para tanto, os modelos discutidos anteriormente foram aplicados em novos consumidores e, em seguida, dois grupos de 200 consumidores cada foram selecionados para novas inspeções em campo. O Grupo 1 representa a Seleção 1 e o Grupo 2 representa a Seleção 2 que foram discutidas durante a análise de maximização do Cenário 1.

Apenas o Cenário 1 pode ser validado em campo, pois, para a validação do Cenário 2 seria necessário inspecionar todos os consumidores da área de concessão, o que não é exequível. Além disso, dos novos consumidores selecionados, apenas 68 não estavam inclusos no grupo de 1.349 consumidores inspecionados nas etapas anteriores. Portanto, apenas 68 novas inspeções foram realizadas. Assim, no total 1.417 inspeções de campo foram realizadas para validar as etapas da metodologia da presente pesquisa.

Os resultados das inspeções de campo para validação da etapa atual da metodologia são apresentados na Tabela 21.

Tabela 21 – Resultado das inspeções de campo para validação da maximização do retorno financeiro.

Item	Grupo 1	Grupo 2
Consumidores com PNT	133	80
Consumidores sem PNT	67	120
Energia recuperada total	110,554 MWh	175,029 MWh
Custo operacional total	R$ 31.224	R$ 31.224

Fonte: Autoria própria.

Come pode ser observado na Tabela 21, as inspeções no Grupo 1 tiveram um número maior de consumidores identificados com PNT, o que era esperado, uma vez que a probabilidade de PNT foi o viés de seleção neste grupo. Por outro lado, o total de energia recuperada no Grupo 2 é 58% maior do que no Grupo 1, o que demonstra que este grupo proporcionou um maior retorno financeiro, uma vez que o custo das fiscalizações é o mesmo em ambos os grupos.

Os valores de precisão da seleção e retorno financeiro em ambos os grupos podem ser calculados a partir da Tabela 21 e são apresentados na Tabela 22.

Tabela 22 – Precisão e retorno financeiro das inspeções em campo.

Grupo	**Precisão da Seleção**	**Retorno Financeiro**
Grupo 1	66,5%	R$ 34.266
Grupo 2	40,0%	R$ 72.469
Variação	− 26,5 p.p.	+ 111,5%

Fonte: Autoria própria.

A partir da análise da Tabela 22 constata-se que a precisão no Grupo 1 é 26,5 p.p. maior do que no Grupo 2, porém, o retorno financeiro no Grupo 2 é R$ 38.203 maior, o que representa um aumento de 111,5% em relação ao Grupo 1. Portanto, os resultados das inspeções de campo demonstram que a metodologia de maximização proposta na pesquisa é capaz de aumentar significativamente o retorno financeiro das inspeções de campo em aplicações reais, o que corrobora os resultados simulados discutidos anteriormente e garante robustez as conclusões que foram expostas.

A partir dos resultados apresentados e discutidos nesta seção, é possível constatar que todos os objetivos estabelecidos no início do trabalho foram alcançados de forma satisfatória. Além disso, fica claro também a contribuição dos resultados para a gestão da perda não técnica no setor elétrico de distribuição. Por fim, na seção seguinte são apresentadas as conclusões finais e as sugestões de trabalhos futuros.

6 Conclusões

Uma nova metodologia baseada em técnicas de *Advanced Analytics* foi desenvolvida e validada na pesquisa, a qual contribui para o aprimoramento da gestão da perda não técnica de energia em sistemas elétricos de distribuição. Com a análise de Inferência Causal foi possível caracterizar a presença da perda não técnica em um grupo de consumidores e identificar 76 fatores de risco e de proteção para a sua ocorrência. Uma análise detalhada de algumas variáveis demonstrou como elas podem levar a conclusões valiosas para tomada de decisão na gestão da perda na concessionária. Além disso, com base no coeficiente de correlação de Spearman, verificou-se que o conjunto de variáveis foi capaz de explicar 26,36% da variância da perda não técnica no grupo.

Um total de 23 modelos distintos de classificação foram avaliados a partir do custo computacional e do desempenho na identificação da perda não técnica em um conjunto de consumidores. Os resultados alcançados foram validados com novas inspeções de campo e demonstraram que os métodos *ensemble* são os mais adequados para fins de identificação da perda não técnica, pois combinam alto desempenho, alta fidedignidade e razoável custo computacional. Dentre os algoritmos analisados destacou-se o *Random Forest*, por ter apresentado o melhor desempenho nos resultados de campo, com uma precisão de 66,50% e um desvio de apenas 6,86% em relação aos resultados simulados.

Diferentes abordagens também foram avaliadas na identificação da perda não técnica, como a segregação preliminar de dados e o uso de um critério de votação combinando muitos classificadores. Os resultados demonstraram que estas abordagens não promoveram melhora no desempenho da classificação. Apesar disso, as abordagens podem ser úteis em alguns cenários específicos, conforme discutido na apresentação dos resultados.

Além disso, sete modelos distintos de regressão foram utilizados para prever a energia que pode ser recuperada em um consumidor caso seja realizado uma inspeção no mesmo. Os resultados demonstraram que os modelos não são bons para prever a energia recuperada de consumidores individualmente, mas são bons para prever a energia recuperada em nível de grupo, o que foi suficiente para os objetivos da pesquisa. Dentre os modelos avaliados o *eXtreme Gradient Boosted Tree* se destacou por apresentar os menores erros. Nas inspeções de campo o desvio entre o valor real e o valor previsto pelo modelo para a energia recuperada total foi de apenas 3,02%.

Um modelo matemático para previsão do retorno financeiro decorrente das ações de inspeção foi proposto e, a partir dele, um método para maximização do retorno financeiro das inspeções de campo foi desenvolvido. Os resultados demonstraram que o método proposto foi capaz de aumentar o retorno financeiro das inspeções em até 11,53 vezes em relação à abordagem clássica utilizada na bibliografia. Novas inspeções de campo foram realizadas e demonstraram que o método de maximização permanece robusto em aplicações reais, uma vez que proporcionou um aumento de 111,5% no retorno financeiro em relação à abordagem clássica. Assim, quando aplicado em grandes concessionárias, este método pode fornecer um aumento de milhões de reais na receita das companhias.

O método também pode ser utilizado para dimensionar a infraestrutura ótima para gestão das perdas em uma concessionária, o que pode ser muito útil já que não há critérios definidos na bibliografia para este fim. A partir deste método foi definido o número ótimo de inspeções que deveriam ser realizados na concessionária em estudo com objetivo de alcançar o maior retorno financeiro possível, com isso foi possível alcançar 57% do valor teórico máximo para o retorno financeiro.

Por fim, é possível afirmar que os resultados alcançados nesta pesquisa representam uma contribuição relevante para as concessionárias de distribuição de energia elétrica no gerenciamento da perda não técnica, bem como, para o estado da arte no tema. Pois, fornecem uma nova metodologia capaz de caracterizar e identificar a perda não técnica nos sistemas elétricos, bem como, maximizar o retorno financeiro obtido nas inspeções de campo, de modo a preencher lacunas existentes na bibliografia disponível até então.

6.1 Trabalhos Futuros

Diante da relevância dos resultados alcançados no trabalho, torna-se natural que existam possibilidades de desdobramentos da pesquisa, ou memo de complementação da metodologia utilizada. Por isso, são apresentados a seguir sugestões de trabalhos futuros que podem ser realizados na continuação desta linha de pesquisa.

- Aperfeiçoar os modelos de regressão para melhorar a previsão de energia recuperada em nível de consumidores individuais. Apesar de ser uma tarefa relativamente complexa, dado a forma como o valor da energia recuperada é definido pela concessionária, previsões mais precisas em nível individual podem

promover melhorias no modelo de maximização do retorno financeiro, além de auxiliarem no planejamento das concessionárias;

- Incorporar nas análises o aumento incremental do consumo de energia que surge após a eliminação das irregularidades no sistema de medição que provocam a perda não técnica. O valor é conhecido como energia agregada e é incorporado no faturamento mensal das concessionárias após a realização das inspeções. Este valor pode influenciar no cálculo do retorno financeiro que é obtido com as inspeções, por isso, a sua incorporação nas análises tornaria os resultados alcançados ainda mais realistas;
- Considerar na análise do retorno financeiro a possibilidade de inadimplência dos consumidores nos pagamentos referentes a energia recuperada nas inspeções. Para tanto, é necessário desenvolver novos modelos preditivos para prever a probabilidade de um consumidor identificado com perda não técnica não pagar a cobrança de energia recuperada que será realizada;
- Incorporar dados de *smart meter* disponíveis na concessionária para criação de novas variáveis. Alguns consumidores possuem medição do tipo *smart meter*, com a disponibilização das grandezas elétricas tensão, corrente, fase e potência em intervalos periódicos de tempo. Além disso, há uma tendência de que este tipo de medição seja expandido para cada vez mais consumidores. Assim, as informações provenientes das medições podem ser utilizadas para criação de novas variáveis com alta associação à presença de perda não técnica, o que colaboraria para o aumento do desempenho dos modelos preditivos;
- Avaliar o uso da abordagem *self-taught learning* (*deep learning*) como alternativa às Etapas 1 e 2 do trabalho;
- Desenvolver um modelo de otimização para definir a sequência de execução mais eficiência das inspeções de campo do ponto de vista operacional. O algoritmo da colônia de formigas é um ponto de partida promissor para esta tarefa.

Referências

ABDI, H.; WILLIAMS, L. J. Principal component analysis. **Wiley Interdisciplinary Reviews: Computational Statistics**, v. 2, p. 433-459, 2010. ISSN doi: 10.1002/wics.101.

ABRADEE. A Distribuição de Energia. **Associação Brasileira de Distribuidores de Energia Elétrica**, 2017. Disponivel em: <http://www.abradee.com.br/setor-de-distribuicao/a-distribuicao-de-energia/>. Acesso em: 19 Setembro 2019.

ALPAYDIN, E. **Introduction to machine learning**. 2nd. ed. Cambridge: MIT Press, 2010.

ALVES, H. M. D. M. **Análise da Contribuição de Atributos Derivados do Histórico de Consumo para a Detecção de Perdas Não Técnicas**. Univ. Fed. de Campinha Grande. Dep. de Eng. Elétrica. Dissertação de Mestrado. Campina Grande, p. 66. 2019.

AMIN, S. et al. Game-Theoretic Models of Electricity Theft Detection in Smart Utility networks. **IEEE Control Systems Magazine**, p. 66-81, Febuary 2015. ISSN 1066-033X.

ANEEL. **Resolução Normativa nº 414**. Agência Nacional de Energia Elétrica. Brasília, p. 201. 2010.

ANEEL. **Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional - PRODIST: Módulo 7 - Cálculo de Perdas na Distribuição**. Agência Nacional de Energia Elétrica. Brasília, p. 27. 2018.

ANEEL. **Perdas de Energia Elétrica na Distribuição**. Agência Nacional de Energia Elétrica. Brasília, p. 21. 2019.

ANGELOS, E. W. S. D. et al. Detection and Identification of Abnormalities in Custumer Consumptions in Power Distribution Systems. **IEEE Transaction on Power Delivery**, v. 26, n. 4, p. 2436-2442, october 2011. ISSN 0885-8977.

ANTMANN, P. **Reducing Technical and Non-Technical Losses in the Power Sector**. World Bank Group Energy Sector Strategy. Washington, D.C., p. 35. 2009.

ARYANEZHAD, M. A novel approac to detection and prevention of electricity pilferage over power distribution network. **Electrical Power and Energy Systems**, v. 111, p. 191-200, April 2019. ISSN 0142-0615.

ASLAM, Z. et al. A Combined Deep Learning and Ensemble Learning Methodology to Avoid Electricity Theft in Smart Grids. **Energies**, v. 13, p. 2-24, October 2020. ISSN doi:10.3390/en13215599.

BAGNALL, A. et al. Is rotation forest the best classifier for problems with continuous features? **ArXiv** , 2018.

BANERJEE, A. Computational Complexity of SVM. **Medium**, 2020. Disponivel em: <https://alekhyo.medium.com/computational-complexity-of-svm-4d3cacf2f952>. Acesso em: 10 junho 2021.

BASTOS, P. R. F. D. M. **Diagnóstico de perdas comerciais de energia elétrica na distribuição usando redes Bayesianas**. Univ. Fed. de Campina Grande. Dep. de Eng. Elétrica. Tese de doutorado. Campina Grande, p. 140. 2011.

BERGSTRA, J. S. et al. **Algorithms for Hyper-Parameter Optimization**. Proceeding of the 24th Neural Information Processing Systems Conference. Granada: [s.n.]. 2011. p. 1-9.

BERTHOLD, M. R. Mixed fuzzy rule formation. **International Journal of Approximate Reasoning**, v. 32, p. 67-84, 2003. ISSN doi: 10.1016/S0888-613X(02)00077-4.

BISHOP, C. M. **Neural Networks for Pattern Recognition**. Oxford: Clarendon Press, 1995.

BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. **A Training Algorithm for Optimal Margin Classifiers**. Proceedings of The Annual Workshp on Computacional Learning. Pttsburgh: ACM Press. 1992. p. 144-152.

BRASIL. **DECRETO-LEI Nº 2.848**. Presidência da República. Brasília. 1940.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123-140, 1996. ISSN doi: 10.1007/BF00058655.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, p. 5-32, 2001. ISSN doi: 10.1023/A:1010933404324.

BROENLEE, J. Logistic Regression for Machine Learning, Apr. 1, 2016. [Online]. Available:. **Machine Learning Mastery**, 2016. Disponivel em: <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>. Acesso em: 11 Janeiro 2021.

BROWNLEE, J. A Gentle Introduction to k-fold Cross-Validation. **Machine Learning Mastery**, 2018. Disponivel em: <https://machinelearningmastery.com/k-fold-cross-validation/>. Acesso em: 27 maio 2021.

BROWNLEE, J. Regression Metrics for Machine Learning. **Machine Learning Mastery**, 2021. Disponivel em: <https://machinelearningmastery.com/regression-metrics-for-machine-learning/>. Acesso em: 27 maio 2021.

CALVO, A. et al. Knowledge-Based Segmetationn to Improve accuracy and Explainability in Non-Technical Losses Detection. **Energies**, v. 13, p. 2-15, October 2020. ISSN doi:10.3390/en13215674.

CHAUHAN, N. S. Decision Tree Algorithm Explained. **KDnuggets**, 2020. Disponivel em: <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>. Acesso em: 25 maio 2021.

CHAWLA, N. V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, June 2002. 321-357.

CHEN, P. Y.; POPOVICH, P. M. **Correlation:** Parametric and nonparametric measures. Thousand Oaks: Sage Publications, 2002.

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

CLAESEN, M.; MOOR, B. D. **Hyperparameter Search in Machine Learning**. Proceedings of The XI Metaheuristics International Conference. Agadir: [s.n.]. 2015. p. 1-5.

COMETTI, E. S.; VAREJÃO, F. M. **Melhoramento da Identificação de Perdas Comerciais Através da Análise Computacional Inteligente do Perfil de Consumo e dos Dados Cadastrais de Consumidores**. Anais do II Congresso Brasileiro de Eficiência Energética. Vitória: CBEE. 2007. p. 1-11.

DEMIR, N. Ensemble Methods: Elegant Techniques to Produce Improved Machine Learning Results. **Developers**, 2016. Disponivel em: <https://www.toptal.com/machine-learning/ensemble-methods-machine-learning#:~:text=Ensemble%20methods%20are%20techniques%20that,winning%20solutions%20used%20ensemble%20methods.>. Acesso em: 24 Maio 2021.

DIÁRIO DA MANHÃ. Mais de 700 "gatos" na rede elétrica de Passo Fundo no primeiro semestre. **Diário da Manhã**, 2019. Disponivel em: <https://diariodamanha.com/noticias/mais-de-700-gatos-na-rede-eletrica-de-passo-fundo-no-primeiro-semestre/>. Acesso em: 23 Setembro 2019.

DIÁRIO DE TAUBATÉ. EDP registra 2.119 ocorrências de fraudes de energia no Vale do Paraíba no 1º semestre de 2019. **Diário de Taubaté e Região**, 2019. Disponivel em: <https://www.diariodetaubateregiao.com.br/dt/edp-registra-2-119-ocorrencias-de-fraudes-de-energia-no-vale-do-paraiba-no-1o-semestre-de-2019/>. Acesso em: 23 Setembro 2019.

DICK, A. J. **Theft of electricity-how UK electricity companies detect and deter**. Proceedings of the European Convention on Security and Detection. Brighton: IET. 1995. p. 90-95.

DONGES, N. A Complete Guide to the Random Forest Algorithm. **Built In**, 2020. Disponivel em: <https://builtin.com/data-science/random-forest-algorithm>. Acesso em: 26 maio 2021.

ELLER, A. N. **Arquitetura de Informações para o Gerenciamento de Perdas Comerciais de Energia Elétrica**. Univ. Fed. de Santa Catarina, Dep. de Eng. de Produção, Tese de Doutorado. Florianópolis, p. 128. 2003.

ESMAEL, A. et al. Non-Technical Loss Detection in Power Grid Using Information Retrieval Approaches: A Comparative Study. **IEEE Access**, v. 9, p. 40635-40648, March 2021. ISSN doi: 10.1109/ACCESS.2021.3064858.

FENG, X. et al. A Novel Electricity Theft Detection Scheme based on Text Convolutional Neural Networks. **Energies**, v. 13, p. 1-17, November 2020. ISSN doi:10.3390/en13215758.

FERNANDO, J. R-Squared Definition. **Investopedia**, 2020. Disponivel em: <https://www.investopedia.com/terms/r/r-squared.asp#:~:text=R%2Dsquared%20(R2),variables%20in%20a%20regression%20model.>. Acesso em: 27 maio 2021.

FIELLER, E. C.; HARTLEY, H. O.; PEARSON, E. S. Tests for rank correlation coefficients. **Biometrika**, v. 44, p. 470-481, 1957. ISSN doi: 10.1093/biomet/44.3-4.470.

FIXAR. Caixa de Medição Individual Monofásica. **Fixar Indutrial**, 2019. Disponivel em: <http://www.fixarba.com.br/produtos-lista.php?categoria=18>. Acesso em: 23 Setembro 2019.

FLETCHER, R. H.; FLETCHER, S. W. **Epidemiologia Clínica**. 4ª. ed. Massacchusetts: Jones & Bartlett, 2006.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **The Anals of Statistics**, v. 29, n. 5, p. 1189-1232, April 2001.

GANDHI, R. Support Vector Machine — Introduction to Machine Learning Algorithms. **Toawards Data Science**, 2018. Disponivel em: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Acesso em: 10 Outubro 2019.

GARTNER. **Magic Quadrant for Data Science anda Machine Learning Platforms**. Gartner, Inc. [S.l.], p. 79. 2019. (ID G00354456).

GHAJAR, R.; KHALIFE, J.; RICHANI, B. Desing and cost analysis of an automatic meter reading system for Eletricité du Liban. **Utilities Policy**, v. 9, p. 193-205, January 2002. ISSN 0957-1787.

GHORI, K. et al. Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection. **IEEE Access**, v. 8, p. 16033-16048, January 2020. ISSN doi: 10.1109/ACCESS.2019.2962510.

GHOSH, S.; REILLY, D. L. **Credit Card Fraud Detection with a Neural-Network**. Proceedingsof the 27th Annual Hawaii International Conference on System Science. Maui: [s.n.]. 1994. p. 621-630.

GLAUNER, P. **Artificial Intelligence for the Detection of Electricity Theft and Irregular Power Usage in Emerging Markets**. Univ. of Luxembourg. Faculty of Sciences, Technology and Comunication. Doctoral Thesis. Luxembourg, p. 152. 2019.

GLAUNER, P. et al. **Non-Technical Losses in the 21st Century: Causes, Economic Effects, Detection and Perspectives**. University of Luxembourg. Luxembourg, p. 10. 2018.

GUERRERO, J. I. et al. Improving Knowledge--Based System whit statistical techniques, text mining, and neural networks for non-thecnical loss detection. **Knowledge-Based Systems**, p. 1-13, August 2014. ISSN 0950-7051.

GUERRERO, J. I. et al. Non-Technical Losses Reduction by Improving the Inspections Accuracy in a Power Utility. **IEEE Transactions on Power Systems**, 2018. pp. 1209-1219.

GUFOSOWA. K-fold cross validation. **Criative Common License**, 2019. Disponivel em: <https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.svg>. Acesso em: 27 maio 2021.

GUILLOUX, A. G. A. et al. **Estudos Epidemiológios**. [S.l.]: [s.n.], 2019.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. **The Elements of Statistical Learning**. New York: Springer, 2009.

HEINEY, J. et al. Intel Realizes $25 Billion by Applying Advanced Analytics from Product Architecture Design Through Supply Chain Planning. **INFORMS Journal on Applied Analytics**, 2021. 9-25.

HILL, A. B. The Environment and Disease: Association or Causation? **" in Proceedings of the Royal Society of Medicine**, London, 1965. 295-300.

HUANG, S.-C.; LO, Y.-L.; LU, C.-N. Non-technical Loss Detection using State Estimation and Analysis of Variance. **IEEE Transaction on Power Systems**, v. 28, n. 3, p. 2959-2966, August 2013. ISSN 0885-8950.

HUSSAIN, Z. et al. Methods and Techniques of Electricity Thieving in Pakistan. **Journal of Power Energy Engineering**, v. 4, p. 1-10, September 2016. ISSN 2327-5901.

JAMIL, F.; AHMAD, E. Policy considerations for limiting electricity theft in the developing countries. **Energy Policy**, v. 129, p. 452-458, February 2019. ISSN 0301-4215.

JINDAL, A. et al. Decision Tree and SVM-based Data Analytics for Theft Detection in Smart Grid. **IEEE Transactions on Industrial Informatics**, v. 12, n. 3, p. 1005-1016, March 2016. ISSN 1551-3203.

JOKAR, P.; ARIANPOO, N.; LEUNG, V. C. M. Eletricity Theft Detection in AMI Using Customers' Consumption Patterns. **IEEE Transactions on Smart Grid**, p. 1-11, 2015. ISSN 1949-3053.

KAGAN, N.; OLIVEIRA, C. C. B. D.; ROBBA, E. J. **Introdução aos Sistemas de Distribuição de Energia Elétrica**. 2ª. ed. São Paulo: Blucher, 2010.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics**. 1st. ed. Cambridge: The MIT Press, 2015.

KHURANA, S. Naive Bayes Classifiers. **GeeksforGeeks**, 2020. Disponivel em: <https://www.geeksforgeeks.org/naive-bayes-classifiers/>. Acesso em: 25 maio 2021.

KNIGHT, W.; GORDON, S. People who steal Edison's eletricity. **The Daily Yellowstone Journal**, Miles City, 27 March 1886.

KUMAR, K.; AZAD, S. K. **Database normalization desing pattern**. Proceedings of 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). Mathura: IEEE. 2017. p. 318-322.

LEITE, J. B.; MANTOVANI, J. R. S. Detecting and Locating Non-Technical Losses in Modern Distribution Networks. **IEEE Transaction on Smart Grid**, v. 9, n. 2, p. 1023-1032, June 2016. ISSN 1949-3053.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curso of Imbalanced Datasets in Machine Learning. **Journal of machine Learnig Research**, v. 18, p. 1-5, January 2017. ISSN 1533-7928.

LEÓN, C. et al. Variability and Trend-Based Generalized Rule Induction Model to NTL Detection in Power Companies. **IEEE Transactions on Power Systems**, v. 26, n. 4, p. 1798-1807, November 2011. ISSN 0885-8950.

LIN, Z.; TABERNA, P.-L.; SIMON, P. Advanced analytical techniques to characterize materials for electrochemical capacitors. **Current Opinion in Electrochemistry**, v. 9, p. 18-25, June 2018. ISSN 2451-9103.

LLOYD, S. P. Least Squares Quantization in PCM. **IEEE Transaction on Information Theory**, v. 28, n. 2, p. 129-137, 1982.

MADRIGAL, M.; RICO, J. J.; UZCAREGUI, L. Estimation of Non-Thecnical Energy Losses in Electrical Distribution Systems. **IEEE Latin America Transaction**, v. 15, n. 8, p. 1447-1452, August 2017. ISSN 1548-0992.

MAHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of Machine Learning**. 2nd. ed. Cambridge: MIT Press, 2018.

MANAGEMENT SOLUTION. **Fraud management in the energy industry**. Management Solution. Madrid, p. 42. 2017.

MANO, R. Non-thecnical Losses in Utility Business - What It Is and Why It Matters to All of Us. **Metering and Smart Energy International**, v. 4, p. 110-112, 2017.

MASSAFERRO, P.; MARTINO, J. M.; FERNANDEZ, A. Fraud Detection in Electric Power Distribuiton: An Approach that Maximizes the Economic Return. **IEEE Transactions on Power Systems**, v. 35, p. 1-9, January 2020. ISSN doi: 10.1109/TPWRS.2019.2928276.

MENEZES, A. M. B. **Epidemiologia das Doenças Respiratórias**. Rio de Janeiro: Revinter, v. 1, 2001.

MESSINIS, G. M.; HATZIARGYRIOU, N. D. Review of non-technical loss detection methods. **Electric Power Systems Research**, v. 158, p. 250-266, February 2018. ISSN 0378-7796.

MIBORROW, S. Decision tree learning. **Creative Commons license**, 2011. Disponivel em: <https://upload.wikimedia.org/wikipedia/commons/f/f3/CART_tree_titanic_survivors.png>. Acesso em: 24 Maio 2021.

MIRKES, E. M. et al. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. **Computers in Biolgy and Medicine**, v. 75, p. 203-2016, August 2016. ISSN 0010-4825.

MONDERO, I. et al. Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees. **Electrical Power and Energy Systems**, v. 34, p. 90-98, November 2012. ISSN 0142-0615.

MOREIRA, S. Rede Neural Perceptron Multicamadas. **Medium**, 2018. Disponivel em: <https://medium.com/ensina-ai/rede-neural-perceptron-multicamadas-f9de8471f1a9>. Acesso em: 08 Outubro 2019.

MOROCO, J. **Análise Estatíistica de dados - com utilização do SPSS**. Lisboa: Sílabo, 2003.

NAGI, J. et al. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector machines. **IEEE Transaction on Power Delivery**, v. 25, n. 2, p. 1162-1171, April 2010. ISSN 0885-8977.

NAGI, J. et al. Improving SVM-Based Nontechnical Loss Detection in Power Utility Using the Fuzzy Inference System. **IEEE Transaction on Power Delivery**, v. 26, n. 2, p. 1284-1285, April 2011. ISSN 0885-8977.

NAGPAL, A. Decision Tree Ensembles- Bagging and Boosting. **Towards Data Science**, 2017. Disponivel em: <https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9>. Acesso em: 26 maio 2021.

NEUROMAT. Elements of a boxplot. **Wikimedia Commons**, 2017. Disponivel em: <https://commons.wikimedia.org/wiki/File:Elements_of_a_boxplot_pt.svg>. Acesso em: 1 Outubro 2019.

NIST. **Engineering Statistics Handbook**. Whasington, D.C.: U.S. Department of Commerce, 2013.

NIZAR, A. H.; DONG, Z. Y.; WANG, Y. Power Utility Nonthecnical Loss Analysis With Extreme Learning Machine Method. **IEEE Transaction on Power Systems**, v. 23, n. 3, p. 946-955, August 2008. ISSN 0885-8950.

NORTHEAST GROUP. **Electricity Theft and Non-Technical Losses: Global Markets, Solutions and Vendors**. Northeast Group, IIc. Whashington, D.C., p. 49. 2017.

O LIBERAL. STJ confirma corte de luz para quem faz gato na rede elétrica. **O Liberal**, 2018. Disponivel em: <https://liberal.com.br/arquivo-de-noticias/brasil-e-mundo/brasil/stj-confirma-corte-de-luz-para-quem-faz-gato-na-rede-eletrica-793131/>. Acesso em: 23 Setembro 2019.

OLIVERI, P. et al. The impact of signal pre-processing on the final interpretation of analytical outcomes – A tutorial. **Analytica Chimica Acta**, v. 1058, p. 9-17, June 2019. ISSN 0003-2670.

PAL, A. Gradient Boosting Trees for Classification: A Beginner's Guide. **Madium**, 2020. Disponivel em: <https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea>. Acesso em: 26 maio 2021.

PASSOS JR., L. A. et al. Unsupervised Non-technical Losses Identification Through Optimum-path Forest. **Electric Power Systems Research**, 2016. pp. 413-423.

PATEL, S. Chapter 2 : SVM (Support Vector Machine) — Theory. **Medium**, 2017. Disponivel em: <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>. Acesso em: 10 Outubro 2019.

PELLEG, D.; MOORE, A. **X-means:** Extending k-means with Efficient Estimation of the Number of Clusters. Proceedings of the Seventeenh International Conference on Machine Learning. San Francisco: [s.n.]. 2000. p. 1-8.

PENIN, C. A. D. S. **Combate, Prevenção e Otimização das Perdas Comerciais de Energia Elétrica**. Univ. de São Paulo, Escola Polotécnica, Tese de Doutorado. São Paulo, p. 227. 2008.

PIECH, C.; NG, A. K-means. **Stanford**, 2013. Disponivel em: <http://stanford.edu/~cpiech/cs221/handouts/kmeans.html>. Acesso em: 06 Outubro 2019.

PORTELA, C. et al. Tutorial de SVM. **Machine Learning Laboratory in Finance and Organizations**, 2017. Disponivel em: <https://lamfo-unb.github.io/2017/07/13/svm/>. Acesso em: 10 Outubro 2019.

PRABHU, R. Understanding Hyperparameters and its Optimisation techniques. **Towards Data Science**, 2018. Disponivel em: <https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568>. Acesso em: 7 Outubro 2019.

PUNMIYA, R.; CHOE, S. Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing. **IEEE Transaction on Smart Grid**, v. 10, n. 2, p. 2326-2329, January 2019. ISSN 1949-3053.

RAGGI, L. M. R. et al. Non-techincal Loss Identification by Using Data Analytics and Customer Smart Meters. **IEEE Transactions on Power Delivery**, v. 35, p. 2700-2710, Febuary 2020. ISSN doi: 10.1109/TPWRD.2020.2974132.

RAMOS, C. C. O. et al. A New Approach for nntechnical Losses Detection Based on Optimum-Path Forest. **IEEE Transaction on Power Systems**, v. 26, n. 1, p. 181-188, Febuary 2011. ISSN 0885-8950.

RAMOS, C. C. O. et al. On the Study of Commercial Losses in Brazil: A Binary Black Hole Algorithm for Theft Characterization. **IEEE Transaction on Smart Grid**, v. 9, n. 2, p. 676-683, April 2016. ISSN 1949-3053.

REIS, R. C. P. D. Estatística e Inferência Causal. **Univ. Fed. do Rio Grande do Sul. Instituto de Matemática e Estatística**, 2020. Disponivel em: <http://www.leg.ufpr.br/lib/exe/fetch.php/2020_cafe_do_dest_ufpr.pdf>. Acesso em: 18 maio 2021.

RISH, I. **An empirical study of the naive Bayes classifier**. Workshop on Empirical Methods in Artificial Intelligence. Seattle: [s.n.]. 2001.

RODRIGUEZ, J. J.; KUNCHEVA, L. I.; ALONSO, C. J. Rotation Forest: A new classifier ensemble method. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 28, p. 1619-1630, 2006. ISSN doi: 10.1109/TPAMI.2006.211.

ROKACH, L.; MAIMON, O. **Data mining with decission trees:** theory and applications. Hackensack: World Scientific, 2008.

ROSS, S. **Introduction to Probability and Statistics for Engineers and Scientists**. 4th. ed. Cambridge: Academic Press, 2009.

ROTHMAN, K. J.; GREENLAND, S. Causation and Causal Inference in Epidemiology. **American Journal of Public Health**, v. 95, p. 144-150, Novembro 2004. ISSN doi: 10.2105/AJPH.2004.059204.

SAEED, M. et al. An Efficient Boosted C5.0 Decision-Tree-Based Classification Approach for Detecting Non-Technical Losses in Power Utilities. **Energies**, v. 13, p. 1-19, June 2020. ISSN doi: 10.3390/en13123242.

SAEED, M. S. et al. Ensemble Bagged Tree Based Classification for Reducing Non-Technical Losses in Multan Electric Power Company of Pakistan. **Electronics**, 2019.

SAEED, M. S. et al. Detection of Non-Technical Losses in Power Utilities—A Comprehensive Systematic Review. **Energies**, Setembro 2020. 1-25.

SANTANA, F. Árvores de Decisão (Projeto passo a passo). **Minerando Dados**, 2017. Disponivel em: <https://minerandodados.com.br/arvores-de-decisao-conceitos-e-aplicacoes/>. Acesso em: 25 maio 2021.

SANTESTEBAN, L. G. Precision viticulture and advanced analytics. A short review. **Food Chemistry**, v. 279, p. 58-62, May 2019. ISSN 0308-8146.

SARABANDO, P. **Outliers: Conceitos Básicos**. Escola Superior de Viseu. Dep. de Matemática. Viseu, p. 28. 2009.

SARADHI, V.; KAMIK, H.; MITRA, P. **Decomposition Method for Support Vector Clustering**. Proceeding of the 2nd International Conference on Intelligent Sensing and Information. Hammamet: [s.n.]. 2005. p. 268-271.

SCHUBERT, E. **Knowledge Discovery in databases - Part III - Clustering**. Heidelberg University. Heidelberg, p. 433. 2017.

SCHWARZ, G. Estimating the Dimension of a Model. **The Annals of Statistics**, v. 6, n. 2, p. 461-464, 1978.

SFN NOTÍCIAS. Enel pagará R$ 1 mil para quem denunciar 'gato' na rede elétrica em municíos da região. **SFn Notícias**, 2018. Disponivel em: <http://www.sfnoticias.com.br/enel-pagara-r-1-mil-para-quem-denunciar-gato-na-rede-eletrica-em-municipios-da-regiao>. Acesso em: 22 Setembro 2019.

SHARMA, A. Decision Tree vs. Random Forest – Which Algorithm Should you Use? **Analytics Vidhya**, 2020. Disponivel em: <https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/>. Acesso em: 26 maio 2021.

SHARMA, S.; MAJUMDAR, A. Unsupervised Detection of Non-Technical Losses via Recursive Transform Learning. **IEEE Transaction on Power Delivery**, v. 36, p. 1241-1244, April 2021. ISSN doi: 10.1109/TPWRD.2020.3029439.

SHAW, R. S. Fuzzy logic temperature. **Creative Commons license**, 2004. Disponivel em: <https://commons.wikimedia.org/wiki/File:Warm_fuzzy_logic_member_function.gif>. Acesso em: 26 maio 2021.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas**. São Paulo: Artliber, 2010.

SIMON, P. **Too Big to Ignore:** The Business Case for Big Data. Hoboken: John Wiley & Sons, 2013.

SINGH, S. K.; BOSE, R.; JOSHI, A. Entropy-based electricity theft detection in AMI network. **IET Cyber-Physical Systems: Theory & Applications**, v. 3, n. 2, p. 99-105, October 2018. ISSN 2398-3396.

SMITH, T. B. Eletricity theft: a comparative analysis. **Energy Policy**, n. 32, p. 2067-2076, 2004. ISSN 0301-4215.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. **Practical Bayesian Optimization of Machine Learning Algorithms**. Neural Information Processing Systems 2012. Lake Tahoe: [s.n.]. 2012. p. 1-9.

SRIVASTAVA, T. 11 Important Model Evaluation Metrics for Machine Learning Everyone should know. **Analytics Vidhya**, 2019. Disponivel em: <https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/>. Acesso em: 27 maio 2021.

STONE, M. Cross-Validatory Choice and Assessment of Statistical Predictions. **Journal of the Royal Statistical Society**, v. 36, p. 111-147, 1974. ISSN doi: 10.1111/j.2517-6161.1974.tb00994.x.

TECHOPEDIA. Advanced Analytics. **Techopedia**, 2017. Disponivel em: <https://www.techopedia.com/definition/32370/advanced-analytics>. Acesso em: 29 Setembro 2019.

TIWARI, S. Complete Guide to Machine Learning Evaluation Metrics. **Medium**, 2019. Disponivel em: <https://medium.com/analytics-vidhya/complete-guide-to-machine-learning-evaluation-metrics-615c2864d916>. Acesso em: 27 maio 2021.

URBANOWICZ, R. J. et al. **Relief-Based Feature Selection: Introduction and Review**. University of Pennsylvania. Institute for Biomedical Informatics. Philadelphia, p. 43. 2017.

VIEGAS, J. L. et al. Solutions for detection of non-technical losses in the electricity grid: A review. **Renewable and Sustainable Energy Reviews**, v. 80, p. 1256-1268, May 2017. ISSN 1364-0321.

VIEGAS, J. L.; ESTEVES, P. R.; VIEIRA, S. M. Clustering-based novelty detection for identification of non-technical losses. **Electrical Power and Energu Systems**, v. 101, p. 301-310, March 2018. ISSN 0142-0615.

WORLD BANK. Electric power transmission and distribution losses (% of output). **The World Bank**, 2018. Disponivel em: <https://data.worldbank.org/indicator/EG.ELC.LOSS.ZS>. Acesso em: 27 junho 2019.

WU, X. et al. Top 10 algorithms in data minig. **Knowledge and Information System**, v. 14, n. 1, p. 1-37, 2008.

XGBOOST DOCUMENTATION. XGBoost Documentation. **XGBoost**, 2020. Disponivel em: <https://xgboost.readthedocs.io/en/latest/>. Acesso em: 11 Janeiro 2021.

XU, V. Z. **A Design of Theft Detection Framework for Smart Grid Network**. University of Waterloo. Dep. of Computer Science. Thesis. Waterloo, p. 91. 2015.

ZABOKRTSKY, Z. **Feature Engineering in Machine Learning**. Charles University in Prague. Institute of Formal and Applied Linguistics. Prague, p. 20. 2016.

ZADEH, L. A. Fuzzy logic. **Scholarpedia**, v. 3, p. 1766, 2008. ISSN doi: 10.4249/scholarpedia.1766.

ZANETTI, M. **Sistema de Identificação de Consumidores Fraudulentos em Redes Elétricas Inteligentes**. Pontifícia Univ. Católica do Paraná. Dep. de Informática. Tese de doutorado. Curitiba, p. 147. 2017.

ZHAN, T.-S. et al. Non-technical loss and power blackout detection under advanced metering infrastructure using a cooperative game based inference mechanism. **IET Generation, Transmission & Distribution**, v. 10, n. 4, p. 873-882, 2016. ISSN 1751-8687.

ZINSLI, J. Advanced analytics drive IoT success. **Hydrocarbon Processing**, September 2020. pp.85-86.

ZOGHBI, R. Bagging (Bootstrap Aggregating), Overview. **Medium**, 2020. Disponivel em: <https://medium.com/swlh/bagging-bootstrap-aggregating-overview-b73ca019e0e9>. Acesso em: 26 maio 2021.

Apêndice A – Lista de Variáveis

As variáveis obtidas no processo de caracterização da perda não técnica de energia são descritas na Tabela 23. Na tabela também é indicado quais variáveis apresentação associação com a presença de perda não técnica e com o valor da energia recuperada, bem como, quais foram selecionadas como entradas nos modelos preditivos.

Tabela 23 – Variáveis criadas no processo de caracterização da perda não técnica.

Grupo	Tipo	Nome	Descrição	Assoc. com PNT?	Selec. para PNT?	Assoc. com kWh?	Selec. para kWh?
Ações de combate às perdas	Categórica	F001	Inspeção anterior com resultado normal	N	N	S	S
Ações de combate às perdas	Categórica	F002	Inspeção anterior com detecção de fraudes	S	S	S	N
Ações de combate às perdas	Categórica	F003	Inspeção anterior com detecção de defeitos	N	N	N	N
Ações de combate às perdas	Categórica	F004	Inspeção anterior com detecção de auto religamento	S	S	N	N
Ações de combate às perdas	Numérica	F005	Quantidade de dias desde a última inspeção	N	N	N	N
Ações de combate às perdas	Numérica	F006	Diferença entre as médias de consumo antes e após um degrau, vezes TEMPO_DEGRAU	N	N	S	N
Ações de combate às perdas	Numérica	F007	Diferença entre o KWH_RECUPERA_ESTIMADO do consumidor e do grupo ao qual ele pertence	N	N	S	S
Ações de combate às perdas	Categórica	F008	Inspeção anterior em outas unidades do mesmo titular com detecção de fraudes	S	N	S	N
Ações de combate às perdas	Categórica	F009	Inspeção anterior em outas unidades do mesmo titular com detecção de defeitos	N	N	N	N
Ações de combate às perdas	Categórica	F010	Inspeção anterior em outas unidades do mesmo titular com detecção de auto religamento	S	N	N	N
Ações de combate às perdas	Categórica	F011	Pré-inspeção anterior com resultado normal	S	S	N	N
Ações de combate às perdas	Categórica	F012	Pré-inspeção anterior com indícios de fraude	N	N	N	N
Ações de combate às perdas	Categórica	F013	Pré-inspeção anterior com indícios de defeito	S	S	S	S
Ações de combate às perdas	Categórica	F014	Pré-inspeção anterior com indícios de auto religamento	S	S	N	N
Ações de combate às perdas	Numérica	F015	Quantidade de dias desde a última pré-inspeção	N	N	N	N
Ações de combate às perdas	Categórica	F016	Regularização padrão do sistema de medição	S	S	N	N
Ações de combate às perdas	Categórica	F017	Substituição de medidor obsoleto	S	S	N	N
Ações de combate às perdas	Categórica	F018	Instalação de telemetria no sistema de medição	N	N	N	N
Ações de combate às perdas	Categórica	F019	Instalação de caixa de medição blindada	S	S	N	N
Ações de combate às perdas	Categórica	F020	Externalização do sistema de medição	N	N	N	N
Ações de combate às perdas	Categórica	F021	Regularização de consumidor clandestino	S	S	N	N
Ações de combate às perdas	Numérica	F022	Quantidade de dias desde a última regularização	N	N	N	N
Ordens de Serviços	Categórica	F023	Inspeção no sistema de medição	S	S	N	N
Ordens de Serviços	Categórica	F024	Detecção de irregularidade no sistema de medição	S	N	S	S
Ordens de Serviços	Categórica	F025	Instalação de medidor	S	S	S	S

Grupo	Tipo	Nome	Descrição	Assoc. com PNT?	Selec. para PNT?	Assoc. com kWh?	Selec. para kWh?
Ordens de Serviços	Numérica	F026	Quantidade de cortes de fornecimento realizados	S	S	S	S
Ordens de Serviços	Numérica	F027	Quantidade de religações realizadas	N	N	N	N
Ordens de Serviços	Categórica	F028	Mudança de titularidade	N	N	S	S
Ordens de Serviços	Numérica	F029	Quantidade de vistorias realizadas	S	S	N	N
Ordens de Serviços	Categórica	F030	Negociação para religamento após corte de fornecimento	S	S	S	S
Ordens de Serviços	Categórica	F031	Pré-inspeção no sistema de medição	S	N	N	N
Ordens de Serviços	Categórica	F032	Desligamento sem a retirada do medidor	S	S	S	S
Ordens de Serviços	Numérica	F033	Quantidade de solicitações para benefício de baixa-renda	N	N	N	N
Ordens de Serviços	Categórica	F034	Troca do disjuntor geral	N	N	S	S
Ordens de Serviços	Categórica	F035	Retirada de medidor	S	S	S	S
Ordens de Serviços	Categórica	F036	Religação com a instalações do medidor	S	S	S	S
Ordens de Serviços	Numérica	F037	Quantidade de solicitações para redução da carga declarada	S	N	N	N
Ordens de Serviços	Categórica	F038	Regularização do sistema de medição (outras)	S	S	N	N
Ordens de Serviços	Numérica	F039	Quantidade de ocorrências de dano elétrico	N	N	N	N
Ordens de Serviços	Numérica	F040	Quantidade de solicitações para aumento da carga declarada	N	N	S	S
Ordens de Serviços	Numérica	F041	Quantidade de solicitações para receber fatura por e-mail	N	N	N	N
Ordens de Serviços	Categórica	F042	Aviso de dano no sistema de distribuição	N	N	N	N
Ordens de Serviços	Numérica	F043	Quantidade de decisões judiciais relacionadas ao consumidor	N	N	N	N
Ordens de Serviços	Categórica	F044	Troca do ramal de serviço	N	N	N	N
Ordens de Serviços	Categórica	F045	Retirada do ramal de serviço	S	S	N	N
Ordens de Serviços	Numérica	F046	Quantidade de instalação/manutenção da telemetria da medição	N	N	N	N
Ordens de Serviços	Categórica	F047	Instalação de blindagem nos bornes do medidor	S	S	N	N
Ordens de Serviços	Numérica	F048	Quantidade de solicitações para cancelar fatura por e-mail	N	N	N	N
Ordens de Serviços	Categórica	F049	Desligamento com a retirado do medidor	N	N	S	S
Ordens de Serviços	Categórica	F050	Blindagem da rede BT de distribuição	N	N	S	S
Ordens de Serviços	Categórica	F051	Solicitação de débito automático	N	N	N	N
Ordens de Serviços	Categórica	F052	Interação realizada em canal de atendimento	N	N	N	N
Ordens de Serviços	Categórica	F053	Cancelamento de débito automático	N	N	N	N
Ordens de Serviços	Categórica	F054	Ligação de nova unidade consumidora	N	N	S	S
Ordens de Serviços	Numérica	F055	Adesão ao sistema de mini ou micro geração	N	N	N	N
Atendimentos	Numérica	F056	Quantidade de atendimentos realizados em agências físicas	S	S	N	N
Atendimentos	Numérica	F057	Quantidade de atendimentos realizados no website	N	N	S	S
Atendimentos	Numérica	F058	Quantidade de atendimentos realizados por telefone	S	S	N	N
Atendimentos	Categórica	F059	Atendimentos realizados por *chatbot*	N	N	N	N
Atendimentos	Numérica	F060	Quantidade de atendimentos realizados por mobile app	N	N	N	N

Grupo	Tipo	Nome	Descrição	Assoc. com PNT?	Selec. para PNT?	Assoc. com kWh?	Selec. para kWh?
Atendimentos	Numérica	F061	Quantidade de atendimentos em totens de autoatendimento	N	N	N	N
Atendimentos	Categórica	F062	Atendimentos realizados pela web	N	N	N	N
Atendimentos	Numérica	F063	Quantidade de dias desde o último atendimento	S	S	S	S
Faturamento	Categórica	F064	Conta refaturada	N	N	N	N
Faturamento	Numérica	F065	Quantidade de meses faturados pela média de consumo	S	S	S	S
Faturamento	Numérica	F066	Quantidade de meses faturados com consumidor desligado	S	S	S	S
Faturamento	Numérica	F067	Quantidade de meses com consumidor desligado sem fatura	S	S	S	N
Faturamento	Numérica	F068	Quantidade de meses com consumo menor que 50% da média	S	S	S	S
Faturamento	Numérica	F069	Quantidade de meses com consumo maior que 150% da média	N	N	S	S
Faturamento	Categórica	F070	Ocorreu a virada do registrador do medidor	S	S	S	S
Faturamento	Numérica	F071	Quantidade de faturas com consumo negativo	S	S	S	S
Faturamento	Numérica	F072	Quantidade de meses com faturamento do custo de disponibilidade	S	S	N	N
Faturamento	Numérica	F073	Quantidade média de dias para pagamento da fatura após o recebimento	S	S	N	N
Faturamento	Numérica	F074	Quantidade de faturas com pagamento pendente	S	S	S	S
Faturamento	Categórica	F075	Classe de consumo residencial	N	N	N	N
Faturamento	Categórica	F076	Classe de consumo comercial	S	N	N	N
Faturamento	Categórica	F077	Classe de consumo poder público	N	N	N	N
Faturamento	Categórica	F078	Classe de consumo industrial	S	S	N	N
Faturamento	Categórica	F079	Classe de consumo rural	S	N	S	S
Faturamento	Categórica	F080	Indica se a UC é classificada como baixa renda	N	N	N	N
Leitura do consumo	Numérica	F081	Quantidade de indicações de suspeita de fraude pelo leiturista	N	N	N	N
Leitura do consumo	Numérica	F082	Quantidade de indicações de suspeita de defeito pelo leiturista	S	S	S	S
Leitura do consumo	Numérica	F083	Quantidade de indicações de unidade desligada pelo leiturista	S	S	S	S
Leitura do consumo	Numérica	F084	Quantidade de leituras com consumo negativo	S	S	S	N
Leitura do consumo	Numérica	F085	Quantidade de indicações de unidade desocupada pelo leiturista	S	S	S	S
Leitura do consumo	Numérica	F086	Quantidade de meses sem realização de leitura	S	S	N	N
Leitura do consumo	Numérica	F087	Quantidade de indicações de visor do medidor sujo pelo leiturista	N	N	S	S
Leitura do consumo	Categórica	F088	Quantidade de indicações de consumo clandestino pelo leiturista	S	S	N	N
Leitura do consumo	Categórica	F089	Quantidade de indicações de unidade auto religada pelo leiturista	N	N	S	S
Cadastro comercial	Categórica	F090	Atividade residencial	N	N	N	N
Cadastro comercial	Categórica	F091	Atividade relacionada a comércio varejista	N	N	N	N
Cadastro comercial	Categórica	F092	Atividade relacionada a comércio atacadista	S	S	N	N
Cadastro comercial	Categórica	F093	Atividade relacionada a serviços	N	N	N	N
Cadastro comercial	Categórica	F094	Atividade relacionada a restaurante	N	N	N	N
Cadastro comercial	Categórica	F095	Atividade relacionada a órgão público	N	N	N	N
Cadastro comercial	Categórica	F096	Atividade relacionada a escritório	N	N	N	N

Grupo	Tipo	Nome	Descrição	Assoc. com PNT?	Selec. para PNT?	Assoc. com kWh?	Selec. para kWh?
Cadastro comercial	Categórica	F097	Atividade relacionada a fabricação	N	N	N	N
Cadastro comercial	Categórica	F098	Atividade relacionada a hotel	N	N	N	N
Cadastro comercial	Categórica	F099	Atividade relacionada a centro médico	N	N	N	N
Cadastro comercial	Categórica	F100	Atividade relacionada a obra civil	S	S	N	N
Cadastro comercial	Categórica	F101	Atividade relacionada a escola	N	N	N	N
Cadastro comercial	Categórica	F102	Atividade relacionada a agronegócio	S	S	N	N
Cadastro comercial	Categórica	F103	Atividade relacionada a comunicação	N	N	N	N
Cadastro comercial	Categórica	F104	Atividade relacionada a eventos	N	N	N	N
Cadastro comercial	Categórica	F105	Consumidor ligado	S	S	S	S
Cadastro comercial	Categórica	F106	Consumidor desligado	S	N	N	N
Cadastro comercial	Numérica	F107	Quantidade de dias que a UC está ligada	N	N	S	S
Cadastro comercial	Numérica	F108	Quantidade de dias que a UC está desligada	N	N	S	N
Cadastro comercial	Numérica	F109	Quantidade de dias que a UC foi religada	S	S	S	S
Cadastro Técnico	Categórica	F110	Ligação monofásica	S	S	N	N
Cadastro Técnico	Categórica	F111	Ligação trifásica	S	N	N	N
Cadastro Técnico	Categórica	F112	Medidor fabricado antes de 2001	S	N	N	N
Cadastro Técnico	Categórica	F113	Medidor fabricado entre 2001 e 2006	N	N	N	N
Cadastro Técnico	Categórica	F114	Medidor fabricado entre 2007 e 2013	S	N	S	S
Cadastro Técnico	Categórica	F115	Medidor fabricado a partir de 2014	S	S	S	S
Cadastro Técnico	Categórica	F116	Medidor tipo A	S	S	N	N
Cadastro Técnico	Categórica	F117	Medidor tipo B	S	S	S	S
Cadastro Técnico	Categórica	F118	Medidor tipo C	S	S	S	S
Cadastro Técnico	Categórica	F119	Medidor tipo D	S	S	N	N
Cadastro Técnico	Categórica	F120	Medidor tipo E	N	N	N	N
Cadastro Técnico	Categórica	F121	Medidor tipo F	S	S	S	S
Cadastro Técnico	Categórica	F122	Medidor tipo G	S	S	N	N
Cadastro Técnico	Categórica	F123	Medidor tipo H	N	N	N	N
Cadastro Técnico	Categórica	F124	Medidor tipo I	N	N	N	N
Cadastro Técnico	Categórica	F125	Unidade sem medidor cadastrado	S	S	N	N
Cadastro Técnico	Categórica	F126	Medidor tipo J	S	S	N	N
Cadastro Técnico	Categórica	F127	Medidor tipo L	N	N	N	N
Cadastro Técnico	Categórica	F128	Medidor tipo M	S	S	N	N
Cadastro Técnico	Categórica	F129	Unidade com outros tipos de medidor	S	S	N	N
Cadastro Técnico	Categórica	F130	Medidor tipo N	N	N	N	N
Cadastro Técnico	Categórica	F131	Medidor tipo O	N	N	N	N
Cadastro Técnico	Categórica	F132	Medidor tipo P	S	S	N	N
Cadastro Técnico	Categórica	F133	Medidor tipo Q	N	N	N	N

Grupo	Tipo	Nome	Descrição	Assoc. com PNT?	Selec. para PNT?	Assoc. com kWh?	Selec. para kWh?
Cadastro Técnico	Categórica	F134	Medidor tipo R	S	S	N	N
Cadastro Técnico	Categórica	F135	Medidor tipo S	S	S	N	N
Consumo de energia	Categórica	F136	Unidade residencial com média de consumo entre 100-220 kWh	S	S	N	N
Consumo de energia	Categórica	F137	Unidade residencial com média de consumo entre 220-500 kWh	S	S	N	N
Consumo de energia	Categórica	F138	Unidade residencial com média de consumo entre 0-100 kWh	S	S	N	N
Consumo de energia	Categórica	F139	Unidade residencial com média de consumo entre 500-1000 kWh	N	N	N	N
Consumo de energia	Categórica	F140	Unidade residencial com média de consumo maior que 1000 kWh	N	N	N	N
Consumo de energia	Categórica	F141	Unidade comercial com média de consumo entre 0-500 kWh	S	S	N	N
Consumo de energia	Categórica	F142	Unidade comercial com média de consumo entre 500-2000 kWh	N	N	N	N
Consumo de energia	Categórica	F143	Unidade comercial com média de consumo entre 2000-5000 kWh	N	N	N	N
Consumo de energia	Categórica	F144	Unidade comercial com média de consumo maior que 5000 kWh	N	N	N	N
Consumo de energia	Categórica	F145	Unidade rural com média de consumo entre 300-1000 kWh	S	S	N	N
Consumo de energia	Categórica	F146	Unidade rural com média de consumo entre 1000-5000 kWh	N	N	N	N
Consumo de energia	Categórica	F147	Unidade rural com média de consumo entre 0-300 kWh	S	S	N	N
Consumo de energia	Categórica	F148	Unidade rural com média de consumo maior que 5000 kWh	N	N	N	N
Consumo de energia	Categórica	F149	Unidade pública com média de consumo entre 0-2000 kWh	N	N	N	N
Consumo de energia	Categórica	F150	Unidade pública com média de consumo entre 2000-5000 kWh	N	N	N	N
Consumo de energia	Categórica	F151	Unidade pública com média de consumo entre 5000-10000 kWh	N	N	N	N
Consumo de energia	Categórica	F152	Unidade pública com média de consumo maior que 10000 kWh	N	N	N	N
Consumo de energia	Categórica	F153	Unidade industrial com média de consumo entre 0-1000 kWh	N	N	N	N
Consumo de energia	Categórica	F154	Unidade industrial com média de consumo entre 1000-3000 kWh	N	N	N	N
Consumo de energia	Categórica	F155	Unidade industrial com média de consumo entre 3000-7000 kWh	N	N	N	N
Consumo de energia	Categórica	F156	Unidade industrial com média de consumo maior que 7000 kWh	S	S	N	N
Consumo de energia	Numérica	F157	Consumo médio mensal	N	N	S	N
Consumo de energia	Numérica	F158	Variação absoluta entre a média de consumo do último semestre e do semestre anterior	S	N	S	N
Consumo de energia	Numérica	F159	Variação relativa entre a média de consumo do último semestre e do semestre anterior	S	S	S	N
Consumo de energia	Numérica	F160	Variação absoluta entre a média de consumo do último ano e do ano anterior	S	N	S	N
Consumo de energia	Numérica	F161	Variação relativa entre a média de consumo do último ano e do ano anterior	S	S	S	N
Consumo de energia	Numérica	F162	Quantidade de meses desde que houve um degrau na série de consumo mensal	S	N	S	N
Consumo de energia	Numérica	F163	Percentual de queda da média de consumo após um degrau na série mensal	S	N	S	N
Consumo de energia	Numérica	F164	Coeficiente de inclinação da reta que aproxima a série de consumo mensal	S	N	S	N
Consumo de energia	Numérica	F165	Desvio padrão da série de consumo mensal	S	N	S	S
Consumo de energia	Numérica	F166	Primeira derivada da série de consumo mensal	S	S	S	S
Consumo de energia	Numérica	F167	Quantidade de meses com consumo abaixo do 1º quartil da série de consumo mensal	S	S	S	S
Consumo de energia	Numérica	F168	Quantidade de meses com consumo acima do 3º quartil da série de consumo mensal	S	S	S	S

Grupo	Tipo	Nome	Descrição	Assoc. com PNT?	Selec. para PNT?	Assoc. com kWh?	Selec. para kWh?
Consumo de energia	Numérica	F169	Distância entre o 1º e o 3º quartil da série de consumo mensal	S	N	N	N
Consumo de energia	Numérica	F170	Diferença entre a MEDIA_CONS do consumidor e do grupo ao qual ele pertence	S	S	S	N
Consumo de energia	Numérica	F171	Diferença entre o DEGRAU6_kWh do consumidor e do grupo ao qual ele pertence	S	N	S	S
Consumo de energia	Numérica	F172	Diferença entre o DEGRAU6_% do consumidor e do grupo ao qual ele pertence	S	N	S	N
Consumo de energia	Numérica	F173	Diferença entre o DEGRAU12_kWh do consumidor e do grupo ao qual ele pertence	N	N	S	S
Consumo de energia	Numérica	F174	Diferença entre o DEGRAU12_% do consumidor e do grupo ao qual ele pertence	N	N	S	N
Consumo de energia	Numérica	F175	Diferença entre o DESVIOP_CONS do consumidor e do grupo ao qual ele pertence	S	N	S	N
Consumo de energia	Numérica	F176	Diferença entre a DERIV_CONS do consumidor e do grupo ao qual ele pertence	S	S	S	N
Consumo de energia	Numérica	F177	Diferença entre o MENOR_1QUARTIL_CONS do consumidor e do grupo ao qual ele pertence	S	S	S	N
Consumo de energia	Numérica	F178	Diferença entre o MAIOR_3QUARTIL_CONS do consumidor e do grupo ao qual ele pertence	S	S	S	N
Consumo de energia	Numérica	F179	Diferença entre o DIST_INTERQUARTIL_CONS do consumidor e do grupo ao qual ele pertence	S	S	N	N
Consumo de energia	Numérica	F180	Diferença entre o DEGRAU_% do consumidor e do grupo ao qual ele pertence	S	N	S	S
Consumo de energia	Numérica	F181	Diferença entre o TEMPO_DEGRAU do consumidor e do grupo ao qual ele pertence	S	N	S	N
Consumo de energia	Numérica	F182	Diferença entre o INCLINA_CONS do consumidor e do grupo ao qual ele pertence	S	N	S	N

Fonte: Autoria própria.