UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Engenharia e Ciências
Câmpus de Rosana

Luiz Paulo Barbosa do Nascimento Filho

DETECÇÃO DE PERDAS COMERCIAIS NA REDE DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA A PARTIR DE TÉCNICAS DE SISTEMAS INTELIGENTES

Rosana - SP
2024

Luiz Paulo Barbosa do Nascimento Filho

DETECÇÃO DE PERDAS COMERCIAIS NA REDE DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA A PARTIR DE TÉCNICAS DE SISTEMAS INTELIGENTES

Trabalho de Conclusão de Curso apresentado à Coordenadoria de Curso de Engenharia de Energia do Câmpus de Rosana, Faculdade de Engenharia e Ciências (FEC), Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia de Energia.

Orientador: Prof. Dr. Lucas Teles de Faria.

Rosana - SP
2024

F481d

Filho, Luiz Paulo Barbosa do Nascimento

Detecção de perdas comerciais na rede de distribuição de energia elétrica a partir de técnicas de sistemas inteligentes / Luiz Paulo Barbosa do Nascimento Filho. -- Rosana, 2024

60 f. : il., tabs., mapas

Trabalho de conclusão de curso (Bacharelado - Engenharia de Energia) - Universidade Estadual Paulista (UNESP), Faculdade de Engenharia e Ciências, Rosana

Orientador: Lucas Teles de Faria

1. Perdas de energia elétrica. 2. Perdas não técnicas. 3. Perdas comerciais. 4. Aprendizado de máquina. 5. Pré-processamento de dados. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Faculdade de Engenharia e Ciências
Câmpus de Rosana

LUIZ PAULO BARBOSA DO NASCIMENTO FILHO

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE "**GRADUADO EM ENGENHARIA DE ENERGIA**"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM ENGENHARIA DE ENERGIA

Prof. Dr. KLEBER ROCHA DE OLIVEIRA
Coordenador

BANCA EXAMINADORA:

Prof. Dr. LUCAS TELES DE FARIA
Orientador/UNESP-Rosana

Prof. Dr. LEONARDO HENRIQUE FARIA MACEDO POSSAGNOLO
UNESP-Rosana

Eng. GUSTAVO ESTEVO FELIX
Membro Externo

Dezembro de 2024

Dedico este trabalho aos meus pais, Luiz Paulo (Pauleta) e Márcia, que sempre me ofereceram apoio incondicional. Obrigado por acreditarem que eu era capaz.

AGRADECIMENTOS

Agradeço, em primeiro lugar, ao meu pai, pelo apoio incondicional e pela compreensão em cada etapa dessa jornada, e à minha mãe, por seu companheirismo, paciência e conselhos que foram fundamentais ao longo desse processo.

Ao meu orientador, Prof. Dr. Lucas Teles de Faria, sou profundamente grato por todos os ensinamentos, pela orientação dedicada e pela paciência durante os anos em que fui seu orientado. Agradeço também pelas oportunidades valiosas que enriqueceram minha formação acadêmica e profissional. À Andréia da Silva Santos de Faria, por todo o auxílio e colaboração nas pesquisas, e ao Prof. Dr. Leonardo Henrique Faria Macedo Possagnolo, por sua paciência, entusiasmo e genuíno interesse em esclarecer minhas dúvidas. A todos, minha gratidão por sempre estarem dispostos a contribuir para o meu aprendizado.

Aos colegas do Cursinho Alternativo Unesp Rosana (CAUR), onde tive a chance de me desenvolver não apenas tecnicamente, mas também pessoalmente. Este foi um espaço de aprendizado e crescimento que levarei comigo para sempre.

Aos amigos e colegas de graduação, que compartilharam tantas experiências, aprendizados e momentos marcantes ao longo desses anos, deixo meu sincero agradecimento. Em especial, agradeço à Maria Fernanda de Souza, pela parceria constante e pelo incentivo mútuo; à Sabrina Alves, pelo cuidado genuíno com os amigos ao redor; ao Níkolas Bernardes, pelas incontáveis risadas compartilhadas; e à Jade Zepelin, pela companhia e apoio, e também pelos momentos singelos, como passear de bike, comer churros e tomar *guarapa*, que se tornaram lembranças inesquecíveis.

A UNESP, Campus de Rosana pela estrutura física e intelectual excepcional que foi fundamental para o meu desenvolvimento técnico e para a realização deste trabalho.

À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelo apoio financeiro, concedido por meio do Processo nº 2023/03151-1 (Iniciação Científica) e 2023/14980-9 (BEPE-Iniciação Científica) que possibilitaram a realização de parte da minha pesquisa em Portugal. Durante esse período, agradeço também ao Prof. Dr. Tiago Manuel Campelos Ferreira Pinto, pela orientação atenciosa e acolhida.

Por fim, a todos que, de alguma forma, contribuíram para esta conquista, deixo meu mais sincero agradecimento.

"Ninguém nunca descobre do que realmente se trata a vida, e isso não importa. Explore o mundo. Quase tudo é realmente interessante se você se aprofundar o suficiente."

Richard P. Feynman

RESUMO

As perdas não técnicas (PNTs), ou perdas comerciais, afetam negativamente a qualidade da energia elétrica e geram custos elevados para as distribuidoras. Este trabalho analisa métodos de pré-processamento de dados aplicados à detecção de PNTs no sistema de distribuição de energia elétrica, utilizando uma rede neural artificial *perceptron* multicamadas (PMC). As técnicas avaliadas incluem remoção de *outliers*, balanceamento de classes e seleção de atributos, com foco no impacto no desempenho e na eficiência computacional do modelo. Os resultados mostraram que, sem pré-processamento, o modelo apresentou um F1-Score de 0,88 e tempo de execução de 10,93s, destacando um bom desempenho inicial. No entanto, em *datasets* maiores, a ausência de pré-processamento pode aumentar significativamente o custo computacional. Entre as técnicas aplicadas, a seleção de atributos mostrou-se eficiente para reduzir o tempo de treinamento, mantendo F1-Score satisfatórios. Por outro lado, métodos como a remoção de outliers via IQR, que reduziu 33% do conjunto de treinamento, prejudicaram o desempenho do modelo, enquanto o balanceamento com NEARMISS apresentou F1-Score de apenas 0,2±0,12. A análise também revelou que combinações específicas de pré-processamento podem equilibrar desempenho e custo computacional. A aplicação dessas técnicas permitiu identificar casos com alto F1-Score e tempo reduzido de treinamento, tornando-os ideais para implementação prática. Observou-se ainda que a simplificação dos modelos, por meio de pré-processamento, não comprometeu sua capacidade de detecção em *datasets* de menor escala, mas pode ser indispensável em problemas mais complexos. Conclui-se que o uso de aprendizado de máquina, aliado ao pré-processamento, é uma ferramenta eficaz para mitigar PNTs, especialmente em cenários onde o custo-benefício de desempenho e eficiência é crítico. Apesar dos avanços tecnológicos, o trabalho reforça a importância do conhecimento humano para interpretar resultados e explorar novas abordagens, evidenciando a ciência de dados como uma área em constante evolução.

Palavras-Chave: Aprendizado de máquina. Ciência de dados. Perdas comerciais. Perdas não técnicas. Pré-processamento de dados. Redes neurais artificiais. Sistemas de distribuição de energia elétrica.

ABSTRACT

Non-technical losses (NTL), or commercial losses, negatively impact the quality of electric power and impose significant costs on energy distribution companies. This study analyzes data preprocessing methods applied to the detection of NTL in power distribution systems using a multilayer perceptron artificial neural network (MLP). The evaluated techniques include outlier removal, class balancing, and feature selection, focusing on their impact on model performance and computational efficiency. The results showed that, without preprocessing, the model achieved an F1-Score of 0.88 and an execution time of 10.93s, demonstrating good initial performance. However, in larger datasets, the absence of preprocessing significantly increased computational costs. Among the techniques applied, feature selection proved effective in reducing training time while maintaining satisfactory F1-Score. Conversely, methods such as outlier removal via IQR, which reduced 33% of the training set, impaired the model's performance, while class balancing using NEARMISS yielded an F1-Score of only 0.2±0.12. The analysis also revealed that specific preprocessing combinations can balance performance and computational cost. Applying these techniques allowed the identification of cases with high F1-Score and reduced training times, making them ideal for practical implementation. Furthermore, simplifying models through preprocessing did not compromise their detection capabilities in smaller datasets but may become indispensable for more complex problems. In conclusion, machine learning combined with preprocessing is an effective tool to mitigate NTL, especially in scenarios where the cost-benefit of performance and efficiency is critical. Despite technological advancements, this work emphasizes the importance of human expertise in interpreting results and exploring new approaches, demonstrating that data science is a constantly evolving field.

Keywords: Artificial neural networks. Data preprocessing. Data science. Commercial losses. Machine learning. Non-technical losses. Power distribution system.

LISTA DE ILUSTRAÇÕES

Figura 1 – Composição tarifária de energia elétrica. 17
Figura 2 – Fluxograma com o procedimento para avaliação das perdas. 20
Figura 3 – Perdas não técnicas reais e regulatórias sobre baixa tensão. 20
Figura 4 – Perdas não técnicas sobre a baixa tensão na região Norte. 21
Figura 5 – Evolução dos custos das perdas no processo tarifário. 21
Figura 6 – Arquitetura da rede neural *perceptron* multicamadas. 23
Figura 7 – Representação do neurônio artificial de McCulloch-Pitts. 24
Figura 8 – Taxa de furto por setor censitário. 37
Figura 9 – Fluxograma para criação do repositório do conjunto de dados. 39
Figura 10 – Diagrama de *Venn* das UCs classificadas como *outlier.* 40
Figura 11 – Frequência de seleção de *features* com o algoritmo *random forest.* 41
Figura 12 – Frequência de seleção de *features* com o algoritmo ANOVA. 42
Figura 13 – Frequência de seleção de *features* com o algoritmo *mutual info classif.* 42
Figura 14 – Função de ativação ReLU. 44
Figura 15 – Função de ativação sigmoide e Heaviside. 45
Figura 16 – Normalização: interpretação geométrica e teorema de Tales. 45
Figura 17 – Boxplot: *outliers versus* balanceamento. 48
Figura 18 – Boxplot: *outliers versus Feature Selection.* 48
Figura 19 – F1-Score *versus* tempo de treinamento. 48

LISTA DE TABELAS

Tabela 1 – Parcelas da tarifa de energia. 15
Tabela 2 – Parcelas da tarifa de uso do sistema de distribuição. 15
Tabela 3 – Matriz de confusão. 25
Tabela 4 – Atributos estatísticos baseados em regimes de consumo. 28
Tabela 5 – Número de UCs após o balanceamento. 40
Tabela 6 – Hiperparâmetros adotados no processo de treinamento da PMC. 43
Tabela 7 – Desempenho das técnicas de pré-processamento isoladas. 47
Tabela 8 – Grupo de resultados vantajosos e seus métodos de pré-processamento. 49
Tabela 9 – Resultados para os demais métodos de pré-processamento aplicados. 50

SUMÁRIO

1 INTRODUÇÃO 14
1.1 COMPOSIÇÃO TARIFÁRIA 14
1.1.1 Tarifa de energia 14
1.1.2 Tarifa de uso do sistema de distribuição 14
1.2 PERDAS E A COMPOSIÇÃO TARIFÁRIA 16
1.3 O QUE SÃO AS PERDAS NÃO TÉCNICAS 17
1.4 PROCESSO DE INSPEÇÃO DAS UNIDADES CONSUMIDORAS 18
1.5 OBJETIVOS 18

2 PERDAS TÉCNICAS E PERDAS NÃO TÉCNICAS 19
2.1 PERDAS TÉCNICAS 19
2.2 PERDAS NÃO TÉCNICAS 19
2.3 REVISÃO BIBLIOGRÁFICA 21

3 MACHINE LEARNING 23
3.1 REDES NEURAIS ARTIFICIAIS PERCEPTRON MULTICAMADAS 23
3.2 MATRIZ DE CONFUSÃO 25
3.2.1 Acurácia 26
3.2.2 Precisão 26
3.2.3 Recall 27
3.2.4 F1-Score 27

4 TÉCNICAS DE PRÉ-PROCESSAMENTO 28
4.1 ENGENHARIA DE ATRIBUTOS 28
4.2 OUTLIER 29
4.2.1 IQR 29
4.2.2 Z-Score 29
4.2.3 Isolation forest 30
4.3 BALANCEAMENTO 30
4.3.1 SMOTE 31
4.3.2 ADASYN 32
4.3.3 Random under sampler 33
4.3.4 NEARMISS 33
4.4 FEATURE SELECTION 33
4.4.1 Random forest 34
4.4.2 ANOVA f_classif 35

4.4.3 Mutual info classif 35

5 RESULTADOS 36

5.1 PARÂMETROS PARA IMPLEMENTAÇÃO DOS MODELOS 36

5.2 ANÁLISE EXPLORATÓRIA DOS DADOS 36

5.3 CONJUNTO DE TREINAMENTO E VALIDAÇÃO 37

5.4 CRIAÇÃO DO REPOSITÓRIO DE DADOS 38

5.5 APLICAÇÃO DAS TÉCNICAS DE PRÉ-PROCESSAMENTO 39

5.5.1 Outlier 39

5.5.2 Balanceamento 40

5.5.3 Feature Selection 41

5.6 TREINAMENTO DA RNA PMC 43

5.6.1 Hiperparâmetros 43

5.6.2 Função de ativação 43

5.6.3 Normalização 44

5.7 RESULTADOS DA FASE DE TESTES 46

6 CONCLUSÃO 52

6.1 TRABALHOS FUTUROS 53

1 INTRODUÇÃO

O sistema elétrico de potência é estruturado em três principais setores: geração, transmissão e distribuição (GTD). O setor de geração é responsável por converter diferentes formas de energia em energia elétrica. Em 2022, a matriz elétrica brasileira teve sua maior produção proveniente das fontes hidráulica, eólica, biomassa e gás natural.

Devido à localização geográfica das usinas geradoras, frequentemente distantes dos centros consumidores, o setor de transmissão tem a função de transportar essa energia por longas distâncias. Para minimizar as perdas durante o transporte, a transmissão utiliza componentes como transformadores elevadores de tensão, uma medida que reduz as perdas de energia e é economicamente vantajosa.

Por fim, a energia é entregue ao setor de distribuição, onde ocorre uma nova transformação de tensão por meio de transformadores abaixadores. Essa redução de tensão ocorre por uma questão de segurança e praticidade no fornecimento de energia elétrica aos consumidores finais (EPE, 2023; Monticelli; Garcia, 2011).

1.1 COMPOSIÇÃO TARIFÁRIA

O módulo 7 dos Procedimentos de regulação tarifária (PRORET) se refere a estrutura tarifária das concessionárias de distribuição de energia elétrica, o submódulo 7.1 tem como objetivo estabelecer os procedimentos gerais a serem aplicados ao processo de definição da estrutura tarifária (ANEEL, 2023).

A tarifa de energia elétrica é o valor que o consumidor paga pelo uso da eletricidade fornecida. Sendo composta por diferentes componentes que visam cobrir os custos de geração, transmissão, distribuição e outras despesas associadas à prestação do serviço. No Brasil, a tarifa de energia elétrica pode ser dividida em duas partes principais apresentadas a seguir.

1.1.1 Tarifa de energia

A tarifa de energia (TE) é o valor pago pelo consumidor pela energia consumida, ou seja, a eletricidade efetivamente utilizada. Esse componente é diretamente influenciado pelos custos de geração de energia. A Tabela 1 apresenta as principais parcelas que compõem esta tarifa.

1.1.2 Tarifa de uso do sistema de distribuição

A tarifa de uso do sistema de distribuição (TUSD) cobre os custos de uso da rede de distribuição de energia elétrica, que são responsabilidade das distribuidoras. Esse custo é cobrado de todos os consumidores, independentemente de sua classe de consumo. A Tabela 2 apresenta as principais parcelas que compõem esta tarifa.

Tabela 1 – Parcelas da tarifa de energia.

Parcela	*Descrição*
Encargos	São taxas adicionais que visam financiar a implementação de políticas públicas e outros custos relacionados à sustentabilidade e à operação do setor elétrico. Alguns exemplos incluem a Conta de Desenvolvimento Energético (CDE), a contribuição para a energia elétrica de fontes renováveis, entre outros.
Perdas	Refere-se a remuneração básica (RB) sobre o mercado cativo onde os consumidores cativos pagam uma parte dos custos associados. O valor da RB é, então, embutido no custo total da Tarifa de Energia (TE) que os consumidores pagam. Isso inclui não só a energia consumida, mas também o custo das perdas de energia que as distribuidoras enfrentam.
Transporte	O transporte na TE está relacionado ao custo da energia gerada e sua transmissão até a distribuidora. Refere-se à transmissão de energia pela rede de alta tensão entre as usinas e as subestações, com o custo desse transporte incluído na Tarifa de Energia. O foco está no transporte da energia desde a geração até a distribuição, cobrindo os custos de geração e transmissão.
Energia	Este componente representa o custo da energia propriamente dita, ou seja, o custo da geração de eletricidade pelas usinas. O preço da energia pode variar conforme o mercado, fonte de geração e período do ano.

Fonte: Elaboração do próprio autor.

Tabela 2 – Parcelas da tarifa de uso do sistema de distribuição.

Parcela	*Descrição*
Encargos	Semelhante aos encargos na TE, os encargos na TUSD são compostos por taxas que financiam a operação do sistema de distribuição e, assim como na TE, visam cobrir custos de infraestrutura, operação e manutenção das redes de distribuição.
Perdas	As perdas na TUSD referem-se à energia que se perde durante a distribuição. As perdas podem ser tanto técnicas quanto não. O custo das perdas de energia é repassado aos consumidores por meio dessa tarifa.
Transporte	O transporte na TUSD abrange o custo de distribuição da energia da subestação até o consumidor final. Ele refere-se ao transporte de energia em baixa e média tensão pela rede de distribuição, cobrindo os custos da infraestrutura, como linhas de distribuição, transformadores e sua manutenção além dos custos de operação e manutenção.

Fonte: Elaboração do próprio autor.

1.2 PERDAS E A COMPOSIÇÃO TARIFÁRIA

No Brasil, a Agência Nacional de Energia Elétrica (ANEEL) adota, desde 1995, o modelo de regulação econômica conhecido como *price cap* no setor de distribuição de energia elétrica. Esse modelo estabelece um limite máximo para os reajustes tarifários, considerando fatores como inflação, ganhos de eficiência e investimentos necessários. Dessa forma, incentiva as concessionárias a operar com maior produtividade, recompensando aquelas que superam os padrões de desempenho estabelecidos pela regulação.

O principal objetivo do *price cap* é equilibrar a redução de custos para os consumidores com a garantia de uma remuneração justa para as concessionárias. Empresas que conseguem operar abaixo dos custos projetados pelo regulador podem reter os ganhos adicionais, enquanto aquelas que não atendem aos critérios de qualidade estão sujeitas a penalizações. Isso promove tanto a competitividade quanto a melhoria contínua nos serviços prestados.

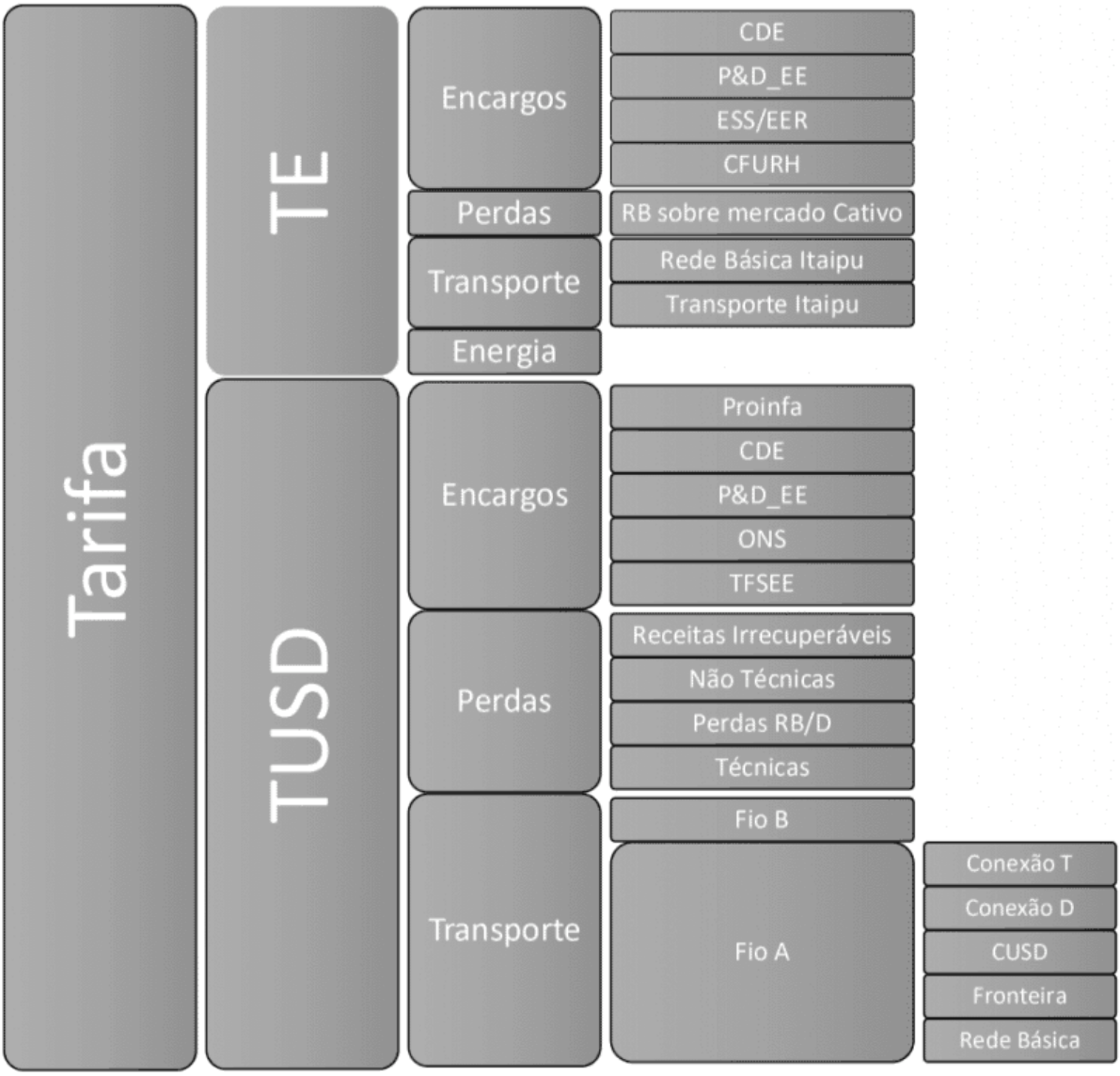
Além disso, o modelo prevê revisões tarifárias periódicas, geralmente a cada quatro ou cinco anos, durante as quais a ANEEL ajusta os parâmetros de cálculo do teto tarifário. Nesse processo, são analisados custos operacionais, investimentos realizados, qualidade do serviço e evolução da demanda. Esses ajustes garantem a sustentabilidade econômica do modelo, oferecendo proteção aos consumidores e incentivando as distribuidoras a manterem um serviço eficiente e de qualidade (CASTRO et al., 2020; INSTITUTO ACENDE BRASIL, 2007, 2011).

As perdas não técnicas (PNTs) representam uma parcela dos custos incluídos na tarifa de energia (TE) e estão associadas à tarifa de uso do sistema de distribuição (TUSD), especificamente na categoria de "perdas", como ilustrado na Figura 1.

Em síntese, definida como a energia consumida, mas não faturada, as PNTs influenciam diretamente no valor pago pelos consumidores, especialmente em regiões onde essas perdas são mais elevadas (Instituto Acende Brasil, 2017).

Considerando esses fatores, compreender a composição tarifária e o impacto das PNTs é fundamental, uma vez que essas perdas causam prejuízos financeiros significativos tanto para as concessionárias quanto para as unidades consumidoras (UCs), refletindo em acréscimos na fatura de energia. Além das perdas de receita e aumento na tarifa, as PNTs comprometem a segurança e a eficiência operacional do sistema de distribuição (Jeyaraj et al., 2020; Savian et al., 2021). Dessa forma, adotar medidas para reduzir as PNTs é essencial para melhorar a eficiência do sistema e oferecer tarifas mais justas, beneficiando tanto os consumidores quanto o setor elétrico como um todo.

Figura 1 – Composição tarifária de energia elétrica.



Fonte: (ANEEL, 2018a).

1.3 O QUE SÃO AS PERDAS NÃO TÉCNICAS

As PNTs, também chamadas de perdas comerciais, referem-se à energia que é consumida, mas não faturada. Elas ocorrem devido a diferentes fatores, tais como: (i) furto: realizado por meio de desvios na rede secundária do sistema de distribuição de energia elétrica (SDEE) ou por intervenções do tipo desvio no sistema de medição, sendo considerada um *by-pass*; (ii) fraudes: consistem na adulteração do sistema de medição de energia elétrica para reduzir ou eliminar o registro do consumo real, caracterizando-se como um furto de energia; (iii) falhas operacionais: relacionadas a erros cometidos pela distribuidora, como leituras incorretas, falhas no processo de faturamento, ausência de equipamentos de medição ou medições realizadas por equipamentos defeituosos (Faria, 2016; Instituto Acende Brasil, 2017; Ventura et al., 2023).

1.4 PROCESSO DE INSPEÇÃO DAS UNIDADES CONSUMIDORAS

O processo de inspeção realizado pelas concessionárias de energia é conduzido por equipes especializadas, que identificam possíveis irregularidades nas UCs por meio de visitas *in loco.* Esse processo depende de recursos humanos qualificados e envolve custos consideráveis, o que limita o número de inspeções que podem ser realizadas em um determinado período.

Em Faria (2016) são discutidas as principais estratégias para detectar UCs irregulares, bem como a taxa de sucesso associada a cada uma, definida em (1). As estratégias abordadas incluem campanhas de combate às PNTs, varreduras, denúncias, análise de dados e o uso de *software* especializado.

$$Taxa\ de\ sucesso\ [\%] = \frac{N°\ de\ UCs\ irregulares\ encontradas}{N°\ de\ UCs\ inspecionadas} \cdot 100 \qquad (1)$$

1.5 OBJETIVOS

As PNTs afetam negativamente a qualidade da energia elétrica e os custos das distribuidoras de energia. Adicionalmente, a detecção dessas perdas é onerosa, pois demanda recursos humanos especializados.

Nesse contexto, a aplicação de modelos de *machine learning* para classificar as UCs aumenta a eficácia do processo de inspeção. Assim, propõe-se neste estudo a avaliação de métodos de pré-processamento de dados, visando um desempenho aprimorado dos algoritmos de classificação para detecção de PNTs.

Este estudo visa realizar uma análise abrangente dos métodos para remoção de *outliers*, balanceamento de classes e seleção de *features* representativas. Desse modo, a capacidade de aprendizado e generalização dos modelos de *machine learning* é aprimorada, acrescido de uma redução do esforço computacional.

Por fim, a falta de estudos que realizam uma comparação abrangente nos métodos de pré-processamento aplicados ao contexto das PNTs motivou a realização deste trabalho.

2 PERDAS TÉCNICAS E PERDAS NÃO TÉCNICAS

2.1 PERDAS TÉCNICAS

As perdas técnicas são intrínsecas ao SDEE, resultantes de fenômenos físicos que ocorrem durante o transporte, transformação de tensão e medição de energia. Essas perdas decorrem de processos naturais, como as perdas Joule nos condutores, que ocorrem devido à resistência elétrica ao fluxo de corrente. Além disso, incluem-se as perdas por efeito de Foucault e por histerese, que estão associadas ao funcionamento dos transformadores. As perdas por efeito de Foucault surgem da formação de correntes parasitas nos núcleos ferromagnéticos, enquanto as perdas por histerese são causadas pela repetitiva magnetização e desmagnetização do núcleo durante os ciclos alternados da corrente elétrica (FITZGERALD, 2014).

No Brasil, as diretrizes e regulamentações sobre as perdas técnicas estão estabelecidas no Módulo 7 dos Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional (PRODIST), que define parâmetros e práticas para mitigar e calcular essas perdas.

Adicionalmente, é importante destacar que o aumento das PNTs também pode contribuir para o aumento das perdas técnicas no SDEE, podendo impactar negativamente nos índices de qualidade do serviço, aumentar os riscos à segurança, reduzir a eficiência do sistema e ocasionar perdas de receita para a concessionária (Messinis; Hatziargyriou, 2018).

Nesse contexto, a Figura 2 apresenta o fluxograma simplificado do procedimento de avaliação das perdas técnicas, cujo método de cálculo é descrito a seguir. A correta determinação das perdas técnicas é uma etapa essencial, pois precede a estimativa das PNTs.

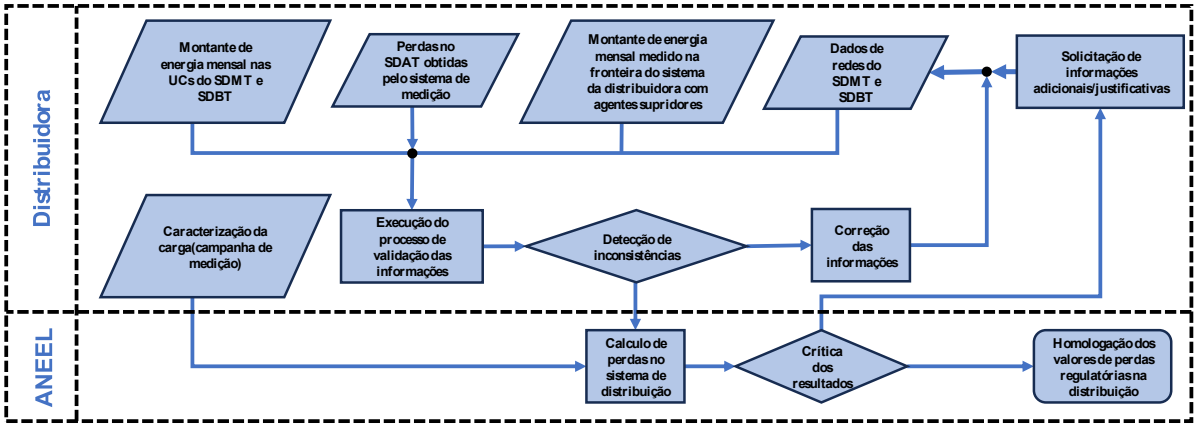
> 3 MÉTODO DE CÁLCULO
>
> 3.1 As perdas de energia nas redes e nos equipamentos associados ao sistema de distribuição de alta tensão (SDAT) são apuradas por dados obtidos do sistema de medição.
>
> 3.2 As perdas de energia nas redes e equipamentos associados ao sistema de distribuição de média tensão (SDMT) e ao sistema de distribuição de baixa tensão (SDBT) são obtidas pela aplicação do método de fluxo de potência.
>
> 3.3 Para os medidores são computadas as perdas nas bobinas de tensão localizadas nas unidades consumidoras do grupo B (ANEEL, 2018, p.10).

2.2 PERDAS NÃO TÉCNICAS

As PNTs são classificadas como reais e regulatórias. As PNTs reais correspondem à diferença entre as perdas totais e as perdas técnicas, enquanto as PNTs regulatórias refletem a parcela de custos dessas perdas que é repassada às UCs por meio da tarifa de energia elétrica. As PNTs regulatórias apresentam menor variação em comparação às PNTs reais, uma vez que

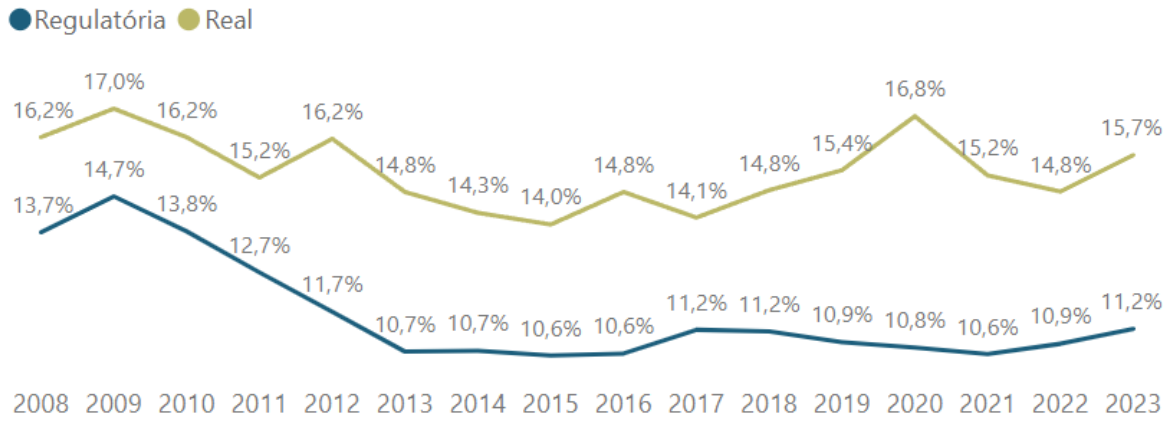
seu objetivo principal é mitigar as perdas de receita das concessionárias. A análise do portal de relatórios da ANEEL (2024) mostra, na Figura 3, a evolução das PNTs reais em relação às perdas regulatórias.

Figura 2 – Fluxograma com o procedimento para avaliação das perdas.



Fonte: Adaptado de ANEEL (2018).

Figura 3 – Perdas não técnicas reais e regulatórias sobre baixa tensão.



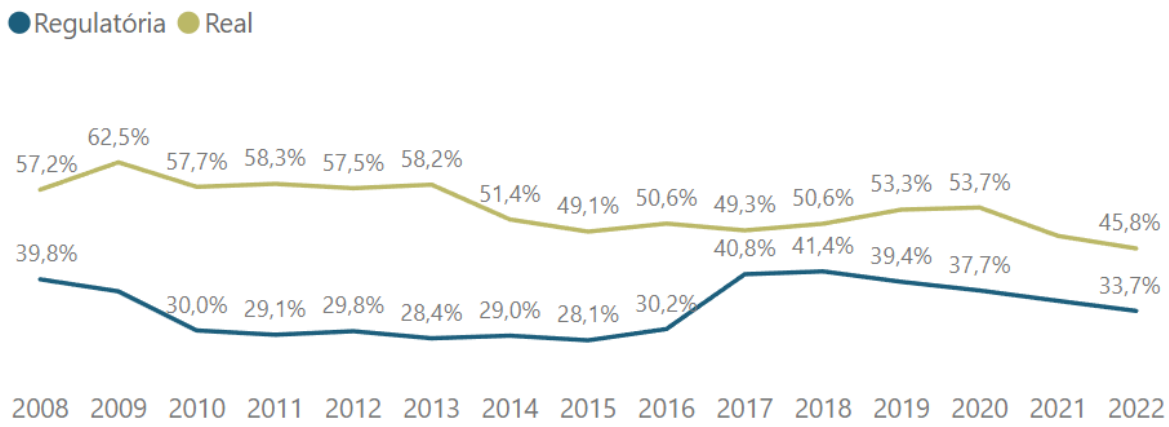
Fonte: ANEEL (2024).

A região Norte do país apresenta os maiores índices de PNTs, conforme mostrado na Figura 4. Em particular, a distribuidora de energia Amazonas Energia registrou, em 2022, um percentual de perdas não técnicas ($PPNT$) real de 120,82% em relação ao consumo do mercado de baixa tensão (Mbt), conforme apresentado em (2) (ANEEL, 2015). Desta forma a cada 1 kWh consumido no Mbt as PNTs reais são de 1,21 kWh, evidenciando a necessidade de investimentos para redução dessas perdas.

$$PPNT\ [\%] = \frac{PNT}{Mbt} \cdot 100 \qquad (2)$$

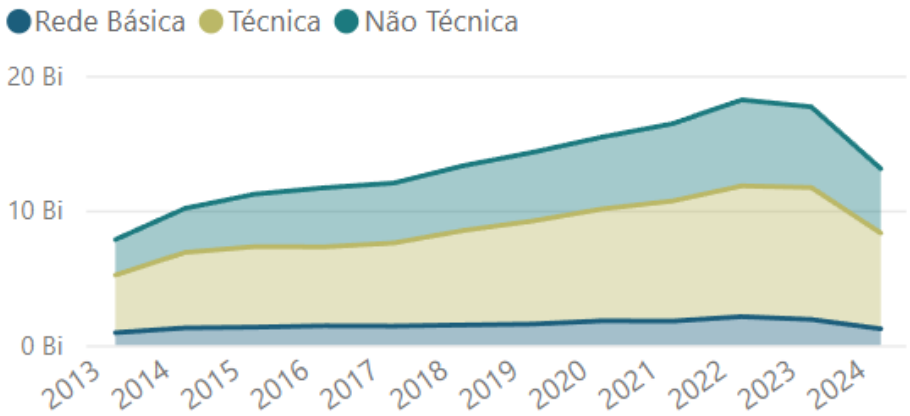
Em 2023, os custos associados às PNTs totalizaram R$ 5,981 bilhões. A Figura 5 apresenta a evolução desses custos relacionados às perdas totais ao longo dos anos.

Figura 4 – Perdas não técnicas sobre a baixa tensão na região Norte.



Fonte: ANEEL (2024).

Figura 5 – Evolução dos custos das perdas no processo tarifário.



Fonte: ANEEL (2024).

2.3 REVISÃO BIBLIOGRÁFICA

Nesta seção, apresenta-se uma revisão dos estudos recentes na literatura especializada voltados à detecção de PNTs por meio de técnicas de sistemas inteligentes. O foco está em metodologias que combinam aprendizado de máquina e análise de dados para identificar padrões de consumo irregulares, buscando soluções que tornem o processo de inspeção mais eficiente.

De acordo com Nagi et al. (2010), as PNTs são mais comuns e têm impactos mais significativos em países emergentes, embora possam afetar todas as economias. Saeed et al. (2020) sugerem que, em países como o Brasil, as estratégias para combater PNTs devem se basear em ferramentas distintas em relação àquelas empregadas em países desenvolvidos, propondo abordagens focadas em *software* em vez de *hardware*. Essa recomendação é devido ao fato de que grande parte do SDEE de países emergentes não é composta por *smart grids*.

Figueroa et al. (2018) ressaltaram a escassez de pesquisas sobre o balanceamento das classes antes da aplicação de classificadores. Dada a significativa disparidade entre o número de UCs regulares e irregulares frequentemente observadas em *datasets* relacionados a furtos de energia, o estudo propôs a utilização de sobreamostragem da classe minoritária (UCs irregulares) e subamostragem aleatória da classe majoritária (UCs regulares). O balanceamento foi testado em diferentes proporções e aplicado nos classificadores *support vector machine* (SVM) com *kernel* Linear, SVM com *kernel Radial Basis Function* (RBF) e *perceptron* multicamadas (PMC). A estratégia de balanceamento dos dados demonstrou resultados promissores, sendo avaliada por meio das métricas *Receiver Operating Characteristic-Area Under the Curve* (ROC-AUC), correlação de Matthews (MCC) e Fβ-score.

Haq et al. (2021) utilizaram uma rede neural convolucional profunda para extrair as características mais significativas de um banco de dados, enquanto um classificador SVM foi empregado para identificar UCs irregulares. A eficiência da abordagem foi validada por meio de métricas derivadas da matriz de confusão, além da análise das curvas ROC e AUC, que evidenciaram a robustez da metodologia proposta.

No estudo de Ghori et al. (2020), diversos modelos de *machine* learning foram avaliados para classificação de PNTs utilizando uma base de dados conformada por 80 mil registros de UCs, cada um com 71 características. Aproximadamente 4% dos registros referiam-se a PNTs. A base foi dividida em 80% para treinamento e 20% para validação. Para reduzir a complexidade computacional, aplicou-se o índice de Gini, que realizou a seleção das características mais significativas, resultando na seleção de 14 características. Os classificadores foram avaliados a partir de métricas extraídas da matriz de confusão, como precisão, *recall* e F1-Score. Dentre os 15 classificadores avaliados, o CatBoost e o PMC apresentaram melhor desempenho.

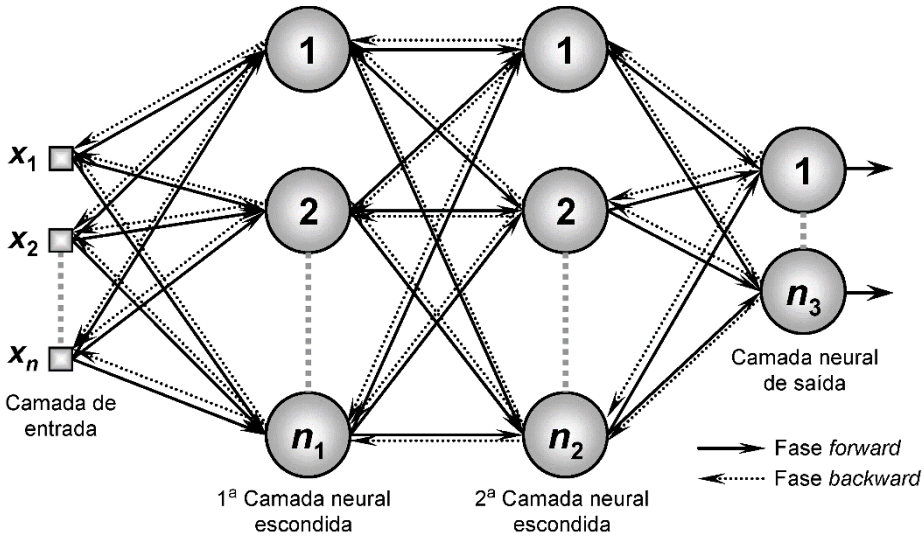
3 MACHINE LEARNING

3.1 REDES NEURAIS ARTIFICIAIS PERCEPTRON MULTICAMADAS

As redes neurais artificiais (RNAs) surgiram para representar matematicamente o processamento de informações em sistemas biológicos. Esse conceito é amplamente utilizado em modelos de reconhecimento de padrões, inspirados na maneira como os sistemas biológicos processam informações. Bishop (2006) observa que uma RNA pode ser interpretada como uma sequência de regressões logísticas aplicadas em camadas sucessivas.

A literatura apresenta diversas arquiteturas de RNAs, entre as quais destaca-se a PMC, apresentada na Figura 6, amplamente difundida para aplicação em diversos problemas, tanto de classificação quanto de predição. Neste trabalho, a PMC será aplicada em um problema de classificação supervisionada, o que envolve as etapas de treinamento e teste do modelo.

Figura 6 – Arquitetura da rede neural *perceptron* multicamadas.



Fonte: Silva et al. (2016).

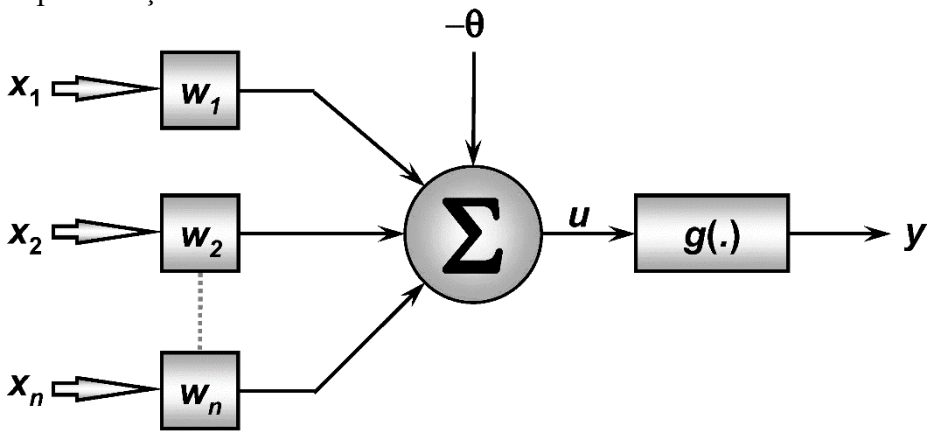
O processo de aprendizagem da RNA PMC é baseado no algoritmo de retropropagação (*backpropagation*), proposto por Rumelhart et al. (1986). Esse método, amplamente utilizado em redes neurais, permite que o modelo ajuste os pesos de cada camada de neurônios de forma a minimizar o erro entre os valores previstos e os reais.

O treinamento da RNA PMC envolve duas etapas, nas quais os fluxos de informação ocorrem em sentidos opostos. Na primeira etapa, chamada de propagação (*forward*), cada amostra com suas respectivas *features* passa pelas camadas da rede até a saída. Nesse fluxo, os neurônios processam as entradas, aplicam as funções de ativação e produzem uma saída que representa a predição do modelo. Em seguida, a predição é comparada com o valor esperado,

resultando em um erro para essa amostra específica. Na segunda etapa, conhecida como retro-propagação (*backward*), o erro calculado é transmitido de volta pelas camadas da rede. Esse processo ajusta os pesos sinápticos de cada neurônio, de maneira a reduzir progressivamente o erro (Bishop, 2006; Haykin, 2001; Silva et al., 2016).

A Figura 7 apresenta uma RNA com arquitetura *perceptron*, proposta por McCulloch e Pitts (1943) composta por um único neurônio artificial. Essa estrutura será utilizada para explicar o funcionamento de um neurônio e sua representação matemática. Para cada variável há um peso associado onde $\{x_1, x_2, \dots, x_n\}$ representam os sinais de entrada, para cada uma dessas variáveis há um conjunto variável de pesos sinápticos $\{w_1, w_2, \dots, w_n\}$, que ponderam a importância de cada entrada.

Figura 7 – Representação do neurônio artificial de McCulloch-Pitts.



Fonte: Silva et al. (2016).

O processo de combinação linear é mostrado em (3) e inclui uma variável de viés θ (comumente chamada de *bias*) que adiciona um valor constante ao resultado. Após a soma, o valor calculado (potencial de ativação) representa o domínio de uma função de ativação $g(\cdot)$, como ilustrado em (4) (Silva et al., 2016). A imagem da função de ativação representa a saída $\hat{y}$ da rede neural.

Resumidamente, o processo de aprendizado na arquitetura PMC consiste em ajustar os pesos sinápticos da rede neural para que a rede capture o padrão associado aos dados de forma que as previsões da rede se aproximem dos valores reais. Essa característica permite que o conhecimento seja distribuído nos pesos de cada neurônio, o que é fundamental para o desempenho da rede.

$$u = \sum_{i=1}^{n} w_i \cdot x_i - \theta \tag{3}$$

$$\hat{y} = g(u) \tag{4}$$

3.2 MATRIZ DE CONFUSÃO

A matriz de confusão é amplamente utilizada como uma ferramenta fundamental para avaliar o desempenho de algoritmos de classificação. Seu principal propósito é comparar a classe predita pelo modelo com a classe real de cada amostra, quantificando assim os acertos e erros do algoritmo. Na sua forma básica, a matriz de confusão é ideal para problemas de classificação binária, nos quais existem apenas duas classes, sendo, portanto, adequada para o problema abordado neste trabalho. É importante destacar que também existem variações da matriz de confusão para problemas de classificação multiclasse e multirrótulo, como descrito por Heydarian et al. (2022)

A Tabela 3 apresenta os elementos da matriz de confusão adaptados ao contexto das PNTs. Nela, Q_{VP} representa a quantidade de verdadeiros positivos, ou seja, o número de unidades consumidoras (UCs) irregulares corretamente classificadas. Q_{FN} refere-se à quantidade de falsos negativos, que corresponde ao número de UCs irregulares erroneamente classificadas como regulares. Q_{FP} indica a quantidade de falsos positivos, isto é, o número de UCs regulares incorretamente classificadas como irregulares, enquanto Q_{VN} representa a quantidade de verdadeiros negativos, ou seja, o número de UCs regulares corretamente identificadas (GEEKS FOR GEEKS, 2024c; Silveira et al., 2022).

Tabela 3 – Matriz de confusão.

Matriz de Confusão		Classe Predita	
		Irregular	Regular
Classe Real	Irregular	Q_{VP}	Q_{FN}
	Regular	Q_{FP}	Q_{VN}

Fonte: Elaboração do próprio autor.

Em problemas de classificação, especialmente em conjuntos de dados desbalanceados, como no caso das PNTs, convém utilizar métricas adequadas para avaliar o desempenho dos classificadores. Embora a acurácia seja uma métrica amplamente utilizada, ela pode ser enganosa em cenários onde as classes estão desbalanceadas. Por isso, além da acurácia, outras métricas como precisão e *recall* oferecem uma visão mais completa sobre o desempenho dos classificadores, especialmente ao lidar com falsos positivos e falsos negativos, fatores críticos na identificação de PNTs (Faria et al., 2012).

3.2.1 Acurácia

A acurácia, descrita em (5), é comumente utilizada para avaliar sistemas classificadores. Ela mede a proporção de classificações corretas em relação ao total de amostras, sendo amplamente utilizada em diversos cenários de aprendizado supervisionado. No entanto, em cenários de PNTs, essa métrica pode ser inadequada devido ao forte desbalanceamento entre as classes de UCs regulares e irregulares. Geralmente para cada nove UCs regulares, há em média apenas uma UC irregular (Faria, 2012).

Desta forma, um classificador que não seja capaz de identificar corretamente as UCs irregulares, mas que classifique todas as UCs regulares de maneira precisa, ainda pode obter uma acurácia elevada, o que pode dar a falsa impressão de que o modelo é eficaz. Isso ocorre porque a acurácia não penaliza adequadamente o erro de classificação das UCs irregulares, que são as classes de interesse no contexto das PNTs. Portanto, embora a acurácia possa ser útil em outros contextos de classificação equilibrada, em problemas como a detecção de PNTs, onde o desequilíbrio entre as classes é significativo, a acurácia pode mascarar a verdadeira performance do modelo.

$$Acurácia = \frac{Q_{VP} + Q_{VN}}{Q_{VP} + Q_{FN} + Q_{FP} + Q_{VN}} \tag{5}$$

3.2.2 Precisão

Métricas como a precisão, apresentada em (6), são fundamentais para a avaliação em problemas com classes desbalanceadas. A precisão indica a confiabilidade do classificador em identificar UCs irregulares. Em outras palavras, ela avalia o quão bem o modelo evita classificar indevidamente uma UC regular como irregular, sendo alta quando o número de falsos positivos (Q_{FP}) é próximo de zero, ou seja, quando o modelo comete poucos erros de classificar UCs regular como irregular. Em problemas de PNTs, reduzir falsos positivos é essencial, pois a classificação incorreta de UCs regulares como irregulares resulta em inspeções desnecessárias pelas equipes de campo, aumentando os custos operacionais.

Portanto, uma precisão alta no classificador não só melhora a eficácia do modelo em detectar as UCs irregulares, mas também contribui para a otimização dos processos operacionais, evitando desperdícios de recursos e tornando o sistema mais eficiente e sustentável.

$$Precisão = \frac{Q_{VP}}{Q_{VP} + Q_{FP}} \tag{6}$$

3.2.3 Recall

Por outro lado, o *recall*, descrito em (7), mensura a capacidade do classificador de identificar corretamente todas as UCs irregulares, ou seja, ele avalia a taxa de cobertura do modelo. Um *recall* alto, com falsos negativos (Q_{FN}) próximo de zero, indica que o classificador tem uma alta taxa de cobertura, ou seja, consegue detectar a maioria das UCs irregulares da área de estudo. Em outras palavras, isso significa que comete poucos erros de classificar UCs irregulares como regular. Isso é especialmente importante em PNTs, onde a falha em identificar UCs irregulares resulta em perdas financeiras e operacionais para a distribuidora de energia.

Seu comportamento é especialmente relevante no cenário de PNTs, pois contribui para a redução dos prejuízos financeiros, ao minimizar a chance de furtos não identificados e não inspecionados, que continuariam gerando custos adicionais para a distribuidora.

$$Recall = \frac{Q_{VP}}{Q_{VP}+Q_{FN}} \tag{7}$$

3.2.4 F1-Score

O F1-Score, apresentado em (8), é a média harmônica entre a precisão e o *recall*, sendo uma métrica especialmente útil para avaliar o desempenho de classificadores em cenários desbalanceados, como os encontrados em PNTs, onde a disparidade entre as classes regulares e irregulares é significativa. Ao equilibrar essas duas métricas, o F1-Score fornece uma visão mais completa da eficácia do classificador. A precisão avalia a confiabilidade do modelo na identificação de UCs irregulares, enquanto o *recall* mede sua capacidade de detectar todas as UCs irregulares, o que é crucial para evitar perdas financeiras e operacionais significativas.

Quando o F1-Score é baixo, significa que pelo menos uma das métricas (precisão ou *recall*) apresenta deficiências notáveis. Isso pode resultar em problemas como: inspeções desnecessárias de UCs regulares, o que acarreta custos extras; ou na falha em identificar UCs irregulares, o que leva a perdas de receita. Em contrapartida, um F1-Score elevado reflete um bom equilíbrio entre acurácia e eficiência operacional, indicando que o classificador é capaz de identificar corretamente a maioria das UCs irregulares enquanto minimiza os custos de inspeção. Esse equilíbrio é essencial para otimizar os processos e mitigar os impactos financeiros negativos no sistema de distribuição.

$$F1 - Score = \frac{2 \cdot Precisão \cdot Recall}{Precisão + Recall} \tag{8}$$

4 TÉCNICAS DE PRÉ-PROCESSAMENTO

4.1 ENGENHARIA DE ATRIBUTOS

A engenharia de atributos, também conhecida como criação de novas *features*, é uma técnica de ciência dos dados que visa melhorar o desempenho e a generalização de modelos de *machine learning*. Ela envolve a criação de novos atributos a partir das variáveis existentes, com o uso de transformações matemáticas. Para dados categóricos, utiliza-se o *one-hot encoding*, enquanto, para variáveis numéricas, transformações estatísticas como médias e medianas são aplicadas (GEEKS FOR GEEKS, 2024d).

No caso de séries temporais, fatores como condições meteorológicas introduzem ruídos nos dados. Avila et al. (2018) extraíram coeficientes *wavelets* e Fourier para reduzir tanto a dimensionalidade quanto os ruídos. Neste estudo, foram extraídos dezesseis atributos estatísticos a partir do histórico de consumo em kWh, conforme proposto em Ferreira (2008). Essas *features*, mais robustas e menos sensíveis a ruídos, capturam o perfil de consumo de cada UC, além de serem mais compreensíveis para os especialistas em PNTs. A Tabela 4 descreve os atributos extraídos.

Tabela 4 – Atributos estatísticos baseados em regimes de consumo.

Siglas	*Descrição das features*
REG	Número de regimes ou patamares do histórico de consumo mensal em kWh.
CV	Coeficiente de variação.
PQR	Percentual de quedas em relação ao regime vigente.
PAR	Percentual de aumentos em relação ao regime vigente.
NQR	Número de regimes de queda.
NAR	Número de regimes de aumento.
PMRI	Percentual de meses no regime inicial da série de consumo mensal.
PMRQ	Percentual de meses em regimes de queda da série de consumo mensal.
PMRA	Percentual de meses em regimes de elevação da série de consumo mensal.
NZ	Número de zeros.
NRFM	Número de regimes na faixa média.
NRABFM	Número de regimes abaixo da faixa média.
NRACFM	Números de regimes acima da faixa média.
NRFRI	Números de regimes na faixa do regime inicial.
NRABFRI	Números de regimes abaixo da faixa do regime inicial.
NRACRI	Números de regimes acima da faixa do regime inicial.

Fonte: Elaboração do próprio autor.

4.2 OUTLIER

Outliers podem ser definidos como amostras que apresentam pouca similaridade com o restante dos dados de um conjunto. Um banco de dados pode conter *outliers* devido a erros de coleta ou entrada de dados, ou até mesmo por refletirem eventos reais, como variações inesperadas no consumo de energia elétrica de uma residência, influenciadas por fatores raros ou aleatórios. Embora esses valores possam ser verdadeiros, eles podem afetar negativamente a capacidade de generalização dos modelos de aprendizado de máquina.

Nesse contexto, serão exploradas duas técnicas para detecção de *outliers* baseadas em conceitos estatísticos: o *interquartile range* (IQR) e o Z-Score. Adicionalmente, será aplicada uma outra técnica de aprendizado de máquina, o *isolation forest*, que utiliza árvores binárias para isolar anomalias.

4.2.1 IQR

O processo de exclusão de amostras consideradas *outliers* pela técnica do *inter-quartile range* (IQR) é fundamentado na análise estatística da mediana e da dispersão dos dados, utilizando os *quartis* para descrever a variabilidade e a centralidade dos conjuntos de dados. Para cada instância (*feature*), calcula-se a diferença entre o terceiro quartil (Q_3) e o primeiro quartil (Q_1), como mostrado em (9), onde há a definição do intervalo interquartil. A partir desse intervalo, determinam-se os limites inferior e superior, conforme (10) e (11), respectivamente. Após o cálculo desses limites para cada instância, são selecionadas apenas as amostras que se encontram dentro do intervalo $[L^{inf}, L^{sup}]$, e que representam o intervalo livre de *outliers* (Alabrah, 2023).

$$IQR = Q_3 - Q_1 \tag{9}$$

$$L^{inf} = Q_1 - 1{,}5 \cdot IQR \tag{10}$$

$$L^{sup} = Q_3 + 1{,}5 \cdot IQR \tag{11}$$

4.2.2 Z-Score

O Z-Score é uma medida estatística que indica quantos desvios padrão uma amostra está distante da média de uma determinada instância (*feature*) em uma base de dados. Essa métrica é amplamente utilizada tanto para padronizar diferentes distribuições quanto para detectar *outliers*.

No processo de detecção de *outliers* utilizando o Z-Score, a normalização de cada amostra original (X_i) é feita por meio da fórmula do Z-Score, resultando em (Z_i), conforme (12), onde μ representa a média e σ o desvio padrão de uma instância da base de dados.

Os *outliers* são identificados ao comparar os valores normalizados com um limiar predefinido. Comumente, utiliza-se um limiar de três desvios padrão. Apenas as amostras que atendem (13) estarão dentro do intervalo livre de *outliers* (Anusha et al., 2019).

$$Z_i = \frac{X_i - \mu}{\sigma} \tag{12}$$

$$|Z_i| < limiar \tag{13}$$

4.2.3 Isolation forest

O *isolation forest* é projetado especificamente para identificar *outliers*, ou anomalias em dados de alta dimensionalidade, utilizando um modelo que isola explicitamente os *outliers*. A premissa central do *isolation forest* é que as anomalias são escassas e significativamente diferentes tanto das amostras normais quanto entre si.

O algoritmo utiliza árvores binárias para separar as amostras. A lógica subjacente é que amostras de *outliers* geralmente exigem menos partições para serem isoladas, resultando em uma menor profundidade na árvore. Em contrapartida, amostras normais tendem a precisar de mais partições, o que implica em uma maior profundidade. Essa característica permite que o modelo calcule uma pontuação de anomalia, com amostras isoladas rapidamente recebendo pontuações mais altas e sendo detectadas como *outliers* de forma explicita.

Além de sua eficácia na detecção de *outliers*, o *isolation forest* é notável por sua simplicidade operacional, apresentando apenas dois parâmetros de ajuste: o número de árvores a serem construídas e o tamanho dos subconjuntos (Liu et al., 2008).

4.3 BALANCEAMENTO

Não é incomum encontrarmos conjunto de dados com classes mais frequentes que possuem um número de amostras significativamente maior que outras em problemas de classificação. Bases de dados com essa característica são chamadas de desbalanceadas. No contexto de PNTs, estima-se que apenas cerca de 10% das UCs pratiquem algum tipo de ação que gere PNTs (Faria et al., 2012). Por conta dessa disparidade, as bases de dados criadas pelo processo de inspeção serão um reflexo dessa estimativa, apresentando um balanceamento significativo.

> Esse fenômeno de desbalanceamento é comum em diversos outros contextos, como nos dados de diagnóstico de câncer, onde a maioria dos participantes apresenta um resultado negativo; em fraudes de pagamento, que ocorrem com menor frequência em relação às transações legítimas; e até mesmo em caixas de *spam*, onde a maior parte dos e-mails recebidos são válidos (Azank e Gurgel, 2020).

Em problemas de classificação binária, a classe com menor número de amostras é chamada de minoritária, enquanto a de maior número é denominada classe majoritária. Esse desbalanceamento pode dificultar o treinamento e a validação de modelos supervisionados, justificando a necessidade da aplicação de técnicas de pré-processamento. Em casos onde o desbalanceamento é pequeno, essa diferença pode ser insignificante e, por vezes, ignorada.

De forma simplificada, as técnicas de balanceamento podem incluir: sobreamostragem, que consiste em aumentar o número de amostras da classe minoritária, e subamostragem, que reduz o número de amostras da classe majoritária. Há também técnicas híbridas que combinam essas duas abordagens (Sharma et al., 2021).

Figueroa et al. (2018) destacam a escassez de estudos sobre o balanceamento de classes em dados de PNTs e propõem a utilização de sobreamostragem para UCs irregulares (classe minoritária) e subamostragem para UCs regulares (classe majoritária), ambas de forma aleatória. O procedimento é realizado em diferentes proporções de balanceamento e apresenta resultados promissores, mostrando que *datasets* mais balanceados podem gerar melhores resultados quando comparado a versões menos balanceadas.

4.3.1 SMOTE

O *synthetic minority over-sampling technique* (SMOTE) é uma técnica amplamente utilizada para lidar com problemas de dados desbalanceados, sendo especialmente aplicada para a sobreamostragem da classe minoritária. Seu objetivo é gerar novas amostras sintéticas da classe com menos amostras, de modo a equilibrar o conjunto de dados.

O processo de sobreamostragem do SMOTE consiste em criar amostras "sintéticas" da classe minoritária, aumentando seu tamanho para balancear a base de dados. O método funciona selecionando aleatoriamente os *k-vizinhos mais próximos* (onde *k* pode ser pré determinado) no espaço de características e, em seguida, gerando novas amostras que são combinações lineares entre a amostra original e seus vizinhos. Esse procedimento é repetido de forma iterativa até que o conjunto de dados estejam balanceado ou até que o número de amostras da classe minoritária atinja um valor pré-determinado (Chawla et al., 2002).

Uma das principais vantagens do SMOTE é que ele evita a criação de amostras duplicadas, como ocorre em técnicas mais simples, onde a classe minoritária é replicada diretamente até atingir o balanceamento. No entanto, uma das principais limitações do SMOTE é a possibilidade de *overfitting*, pois as novas amostras são geradas a partir da combinação linear de amostras existentes. Isso pode resultar em amostras sintéticas enviesadas que não capturam

a complexidade total dos dados originais, especialmente se os dados forem muito dispersos ou ruidosos.

Por fim, o SMOTE é frequentemente combinado com outras técnicas, como a subamostragem da classe majoritária, para melhorar a eficácia em cenários com desbalanceamento extremo, onde a diferença entre as classes é muito grande.

4.3.2 ADASYN

O *adaptive synthetic sampling approach for imbalanced learning* (ADASYN) é uma técnica de sobreamostragem similar ao SMOTE, mas com uma abordagem adaptativa. Sua estratégia baseia-se em gerar amostras sintéticas da classe minoritária de acordo com a dificuldade de aprendizado de cada região dos dados. Assim, mais amostras sintéticas são criadas para as áreas onde as amostras da classe minoritária são mais difíceis de classificar, enquanto menos amostras são geradas em regiões onde o aprendizado é relativamente simples.

O procedimento do ADASYN começa pela determinação do grau de desbalanceamento entre as classes. A partir disso, define-se quantas amostras sintéticas precisam ser geradas para balancear o conjunto de dados. O próximo passo é determinar o nível de equilíbrio desejado entre as classes, seguido pela identificação dos *k-vizinhos mais próximos*, usando a distância euclidiana. Com base nessa proximidade, o ADASYN decide a distribuição das novas amostras sintéticas a serem criadas.

A quantidade de amostras geradas para cada amostra minoritária é calculada de maneira adaptativa, usando a distribuição da classe minoritária. Essa distribuição mede a dificuldade de classificação de uma amostra minoritária, levando em conta o número de vizinhos da classe majoritária ao seu redor. Amostras em regiões mais complexas (com maior número de vizinhos da classe majoritária) recebem mais amostras sintéticas, enquanto as amostras em áreas mais fáceis de classificar recebem menos, tornando o processo adaptativo e automático, sendo este uma das vantagens da técnica.

Entretanto, uma possível limitação do ADASYN é sua maior complexidade computacional, especialmente ao lidar com grandes volumes de dados, em comparação ao SMOTE. Além disso, como a técnica se baseia nos *k-vizinhos mais próximos*, a escolha do valor de k pode influenciar significativamente seu desempenho, tornando-se mais um parâmetro a ser ajustado durante o processo de modelagem (He et al., 2008).

4.3.3 Random under sampler

A técnica *random under sampler* realiza a subamostragem de forma aleatória, removendo amostras da classe majoritária com o objetivo de balancear o conjunto de dados. Trata-se de uma técnica simples e eficiente, tanto em termos de implementação quanto de desempenho computacional.

Entretanto, uma de suas limitações é a possibilidade de eliminar amostras importante da classe majoritária, podendo resultar na perda de informações relevantes para o modelo. Apesar disso, o *random under sampler* é amplamente utilizado por sua rapidez e facilidade, especialmente em bases de dados com grande desbalanceamento e dimensionalidade, situação em que técnicas mais complexas podem não ser viáveis (Lemaître et. al, 2017).

4.3.4 NEARMISS

A técnica NEARMISS, proposta por Zhang e Mani (2003), realiza a subamostragem da classe majoritária com base nas distâncias Euclidianas entre as amostras das classes, sendo também uma técnica baseada no conceito de k-vizinhos mais próximos. O processo de seleção das amostras da classe majoritária para balancear o conjunto de dados pode ser feito por meio de três variantes do algoritmo:

- *NearMiss-1*: Seleciona o número necessário de amostras da classe majoritária que estão mais próximas de cada amostra da classe minoritária;
- *NearMiss-2*: Seleciona as amostras da classe majoritária que estão mais distantes de cada amostra da classe minoritária;
- *NearMiss-3*: Seleciona as amostras da classe majoritária cujos *k-vizinhos mais próximos* são exclusivamente amostras da classe minoritária. Ou seja, mantém as amostras da classe majoritária que estão em regiões predominantemente cercadas por amostras minoritárias.

4.4 FEATURE SELECTION

A etapa de *feature selection* é fundamental em projetos de aprendizado de máquina, pois envolve a seleção das características mais representativas de um conjunto de dados com base em determinados critérios. O objetivo é aprimorar o desempenho do modelo, reduzindo o *overfitting* e melhorando sua capacidade de generalização, além de diminuir o tempo de treinamento. Os métodos de seleção de *features* são geralmente classificados em três grupos: métodos de filtro, métodos *wrapper* e métodos incorporados (Geeks For Geeks, 2024b).

Os métodos de filtro utilizam técnicas estatísticas para avaliar a relevância de cada *feature* de forma independente do modelo, apresentando baixo custo computacional, sendo eficazes na remoção de *features* redundantes ou correlacionadas. No entanto, esses métodos não eliminam a multicolinearidade, pois avaliam cada *feature* isoladamente.

Os métodos *wrapper* consistem em treinar o modelo com diferentes subconjuntos de *features*, avaliando o desempenho a cada variação. A vantagem principal desses métodos consiste em encontrar o conjunto de *features* mais adequado ao modelo específico. Contudo, eles são computacionalmente custosos. Neste estudo, não foi utilizado nenhum método com essa abordagem.

Por fim, os métodos incorporados integram a seleção das *features* mais relevantes ao próprio processo de aprendizado do algoritmo. Esses métodos combinam as vantagens dos anteriores, pois levam em conta as interações entre as *features* durante a seleção (Geeks For Geeks, 2024a).

4.4.1 Random forest

O *random forest* ou floresta de decisão aleatória, é um modelo de *machine learning* baseado em um conjunto de árvores de decisão. Ele funciona construindo uma "floresta" com várias árvores de decisão que tomam decisões de forma independente, cada uma treinada com um subconjunto aleatório dos dados. No final, o modelo combina as previsões de todas as árvores (por exemplo, pela média ou voto majoritário), gerando uma resposta final mais robusta e precisa (Breiman, 2001; Louppe, 2015).

A qualidade das divisões nas árvores do *random forest* é geralmente medida pelo índice de Gini, onde valores mais baixos indicam maior separação entre as classes. A diminuição média de impurezas ou *mean decrease in impurity* (MDI) reflete a contribuição média de cada *feature* para melhorar essa qualidade ao longo de todo o modelo (Soman, 2023).

Breiman (2001) aponta que o processo de seleção de características é intrínseco aos modelos *random forest*, sendo uma medida derivada diretamente do impacto das *features* na redução de impureza dos nós. *Features* que, ao serem utilizadas em divisões, resultam em uma grande diminuição de impureza (avaliada pela MDI) são consideradas mais importantes. Assim, quanto mais frequentemente uma *feature* é utilizada e maior sua contribuição para reduzir a impureza, maior será sua importância.

Essa característica do *random forest* permite identificar as *features* mais relevantes para a separação de classes, possibilitando a simplificação do próprio modelo e o aproveitamento desse processo de seleção de *features* em outros modelos de *machine learning*.

4.4.2 ANOVA f_classif

O ANOVA f_classif é um método de seleção de *features* baseado na análise de variância (ANOVA), que calcula o valor do teste F para cada *feature* em relação à variável alvo. Esse teste é usado para avaliar se existem diferenças significativas entre as médias de classes para uma determinada *feature.*

Em bases de dados onde se assume linearidade entre as classes, o método apresenta vantagens significativas devido ao seu baixo custo computacional, sendo útil especialmente quando o objetivo é apenas reduzir o número de *features* no modelo. No entanto, em conjuntos de dados mais complexos, ele pode não ser a melhor escolha para selecionar as *features* mais representativas.

4.4.3 Mutual info classif

Por fim, o *mutual info classif* estima a informação mútua entre as variáveis, fundamentando-se na teoria das probabilidades e na teoria da informação. Ele quantifica o quanto uma variável aleatória contém informação sobre outra. No contexto da seleção de características, essa métrica permite medir a quantidade de informação compartilhada entre cada *feature* e a variável alvo na base de treinamento, identificando assim as características ou variáveis mais relevantes para a previsão (Kraskov et al., 2004).

A informação mútua é especialmente útil porque pode capturar relações não lineares entre as *features* e a variável alvo, oferecendo vantagens em conjuntos de dados mais complexos em comparação com métodos baseados em correlação, que se restringem a relações lineares.

5 RESULTADOS

Neste capítulo são apresentados os resultados deste estudo. Primeiramente, são descritas as especificidades do conjunto de dados utilizado. Os dados são tratados em uma etapa de pré-processamento e, por fim, avalia-se o desempenho do classificador rede neural PMC.

Todas as simulações deste estudo foram realizadas em um computador pessoal com sistema operacional Windows 64 bits, processador Intel i5 (2,40 GHz, 9ª Geração), 32 GB de RAM e um SSD de 520 GB.

5.1 PARÂMETROS PARA IMPLEMENTAÇÃO DOS MODELOS

A implementação deste trabalho foi realizada inteiramente em *python*, uma linguagem de programação de código aberto amplamente utilizada em ciência de dados (Van Rossum; Drake, 2009). As etapas iniciais de filtragem dos dados foram conduzidas utilizando as bibliotecas *pandas* e *numpy*, conhecidas pela sua eficiência no tratamento e manipulação de dados (Harris et al., 2020; McKinney, 2010). Para a visualização dos dados, foram empregadas as bibliotecas *matplotlib, seaborn* e *geopandas,* permitindo tanto a criação de gráficos quanto a visualização e a manipulação de dados geoespaciais (Hunter, 2007; Jordahl et al., 2020; Waskom, 2021).

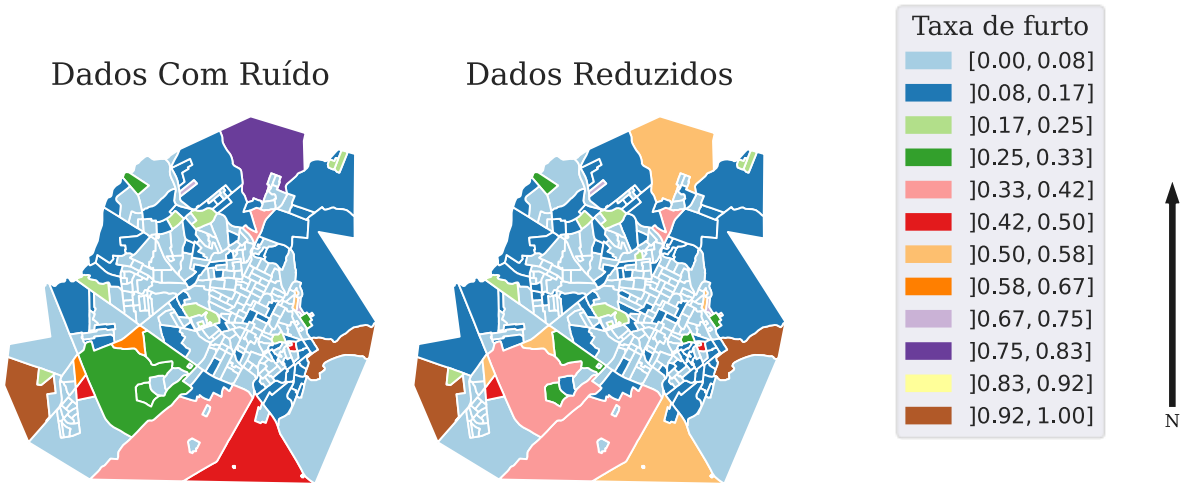
As etapas de pré-processamento, essenciais para a preparação dos dados, foram executadas com o auxílio das bibliotecas *numpy, scikit-learn* e *imbalanced-learn*, cada uma com funcionalidades específicas para manipulação e balanceamento de dados. O modelo de aprendizado de máquinas utilizado foi implementado com as ferramentas disponibilizadas pela *scikit-learn* (Harris et al., 2020; Lemaître et al., 2017; Pedregosa et al., 2011).

Todo o desenvolvimento do código foi realizado no ambiente de programação *jupyter notebook*, integrado ao *visual studio code*, proporcionando uma interface prática para a escrita e execução dos *scripts* (Kluyver et al., 2016; Visual Studio Code, 2022).

5.2 ANÁLISE EXPLORATÓRIA DOS DADOS

A base de dados original contém informações de aproximadamente 80 mil UCs de uma cidade com cerca de 200 mil habitantes, localizada no interior de São Paulo, Brasil. Os dados pertencem a uma concessionária de energia elétrica, inclui informações confidenciais e abrange o histórico de consumo mensal em kWh ao longo de três anos e a classe referente ao furto de energia, sendo elas regular e irregular. Esse intervalo foi escolhido por representar o tempo máximo permitido por lei para que a concessionária possa realizar a cobrança retroativa das UCs, conforme estabelecido por Monyelle et al., (2019, p. 49).

Figura 8 – Taxa de furto por setor censitário.



Fonte: Elaboração do próprio autor.

A taxa de furtos, definida como a razão entre o número de UCs irregulares e o total de UCs em uma determinada região, foi calculada por setor censitário da cidade. Para a avaliação proposta neste estudo, foi realizada uma redução no número de amostras mantendo a heterogeneidade da taxa de furtos próxima aos dados originais em cada área, conforme mostrado na Figura 8.

A redução no número de amostras eliminou instâncias com dados faltantes ou informações inconsistentes, uma vez que apenas as amostras com *features* de consumo completas foram selecionadas, resultando em uma base de dados reduzida com 12792 UCs. Dessa forma, garantiu-se que todas as amostras presentes estavam devidamente preenchidas, facilitando as análises subsequentes.

5.3 CONJUNTO DE TREINAMENTO E VALIDAÇÃO

O processo de separação entre os conjuntos de treinamento e teste foi realizado logo após a conclusão da análise exploratória para garantir que as mesmas combinações de técnicas de pré-processamento sejam aplicadas de forma consistente e que o conjunto de dados utilizado para treinar o modelo seja distinto daquele utilizado para validá-lo. A proporção escolhida foi de 70% dos dados para treinamento e 30% para teste, o que é uma prática comum em problemas de aprendizado supervisionado.

Essa divisão foi considerada adequada, pois garante um número suficiente de dados para o treinamento do modelo, sem comprometer a avaliação em dados novos. Em situações onde a base de dados é muito pequena, uma divisão 80:20 pode ser mais apropriada para aumentar o volume de dados de treinamento, enquanto em bases de dados muito grandes, uma proporção como 90:10 pode ser considerada adequada para avaliação (Nguyen et al., 2021).

Essa padronização na divisão dos dados nos conjuntos de treinamento e de teste garante que todas as técnicas de pré-processamento e suas respectivas combinações sejam aplicadas de forma idêntica aos mesmos dados de treinamento, assegurando uma base comparativa justa para a avaliação das técnicas de pré-processamento. Além disso, essa separação antecipada evita o problema conhecido como "*data leakage*" (vazamento de dados), que ocorre quando informações do conjunto de teste "vazam" para o conjunto de treinamento, oferecendo ao modelo informações que não estariam disponíveis em um cenário real.

Um erro comum, por exemplo, é normalizar ou preencher valores faltantes de forma sintética na base de dados antes de dividi-lo, o que faz com que os dados de teste influenciem na etapa de pré-processamento dos dados de treinamento, e vice-versa. Essa abordagem garante que o desempenho do modelo seja avaliado de forma justa, simulando melhor a capacidade de generalização de modelos em dados desconhecidos.

5.4 CRIAÇÃO DO REPOSITÓRIO DE DADOS

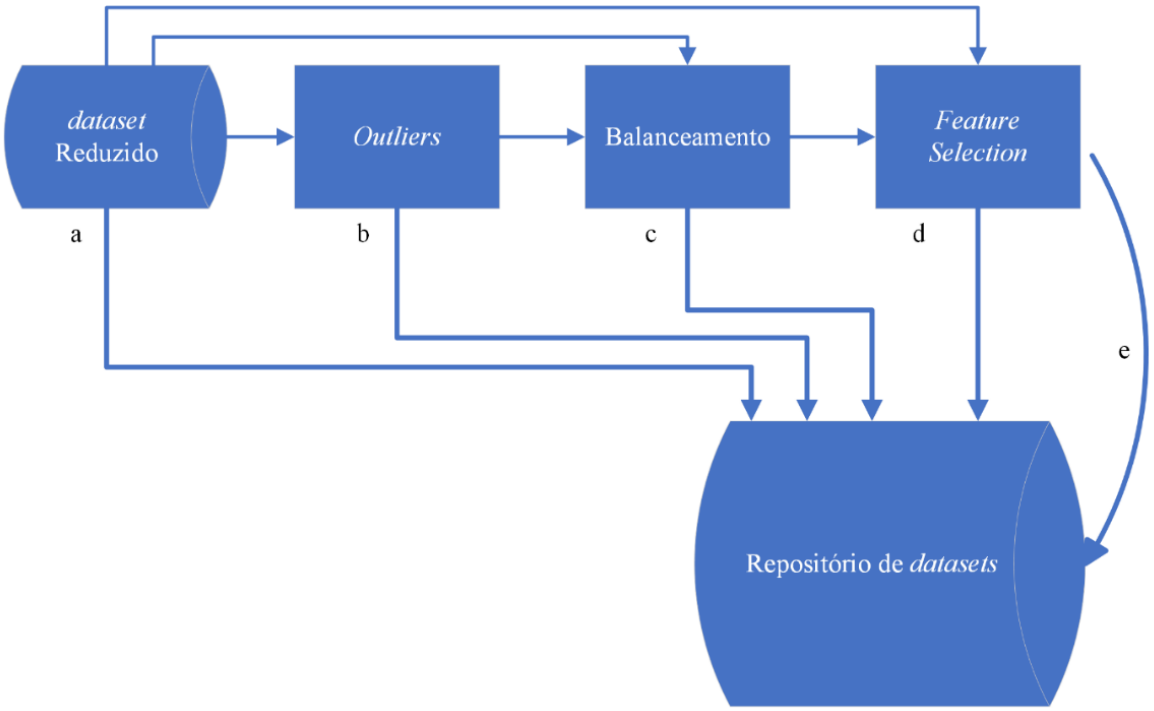
O processo de criação da base de dados foi realizado por meio da combinação das técnicas de pré-processamento, conforme ilustrado na Figura 9. Para a detecção de *outliers* e *feature selection* foram aplicadas três técnicas, enquanto no balanceamento, foram utilizadas quatro técnicas. Além disso, para cada tipo de técnica, foi considerado um cenário base no qual essas técnicas não foram aplicadas, o que adicionou uma configuração extra em cada etapa.

Assim, o total de *datasets* criados pode ser facilmente calculado pela combinação dessas técnicas, levando em conta as configurações adicionais, pelo seguinte cálculo: $(3 + 1) \cdot (4 + 1) \cdot (3 + 1) = 80$.

Dessa forma, o repositório contém 80 *datasets*, resultantes da combinação de todas as opções de pré-processamento disponíveis. Esse volume de dados possibilita testar e comparar o impacto de diferentes combinações de técnicas em cenários diversos.

Essa abordagem não só permite uma análise detalhada das técnicas de pré-processamento aplicadas, mas também oferece uma base robusta para avaliar estatisticamente como cada técnica influencia o desempenho do algoritmo de classificação. Assim, é possível identificar as melhores combinações de técnicas para lidar com os desafios apresentados pelos dados e otimizar os resultados obtidos.

Figura 9 – Fluxograma para criação do repositório do conjunto de dados.



Fonte: Elaboração do próprio autor.

5.5 APLICAÇÃO DAS TÉCNICAS DE PRÉ-PROCESSAMENTO

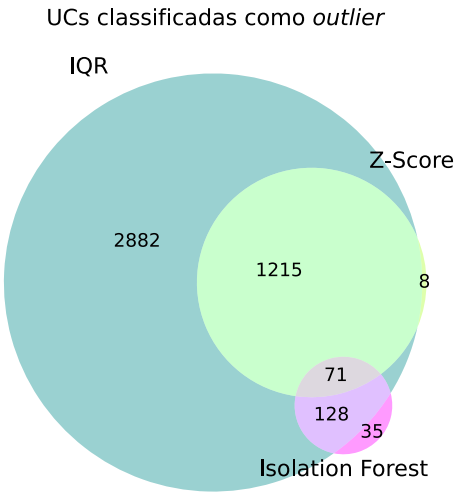
5.5.1 Outlier

Após a aplicação das técnicas de detecção de *outliers* discutidas na Seção 4.2, foi realizada uma análise para identificar quais amostras foram classificadas como *outliers* por cada método. A Figura 10 ilustra esses resultados em um diagrama de Venn, onde a técnica IQR removeu 4.296 UCs do *dataset* (sendo 2.882, 1.215, 71 e 128), enquanto o Z-Score eliminou 1.294 UCs (1.215, 71 e 8). Por fim, o *isolation forest* retirou 234 UCs (71, 128 e 35).

É importante notar que o IQR foi responsável pela remoção de aproximadamente 33,58% das amostras, o que pode impactar negativamente a capacidade de aprendizado dos modelos de *machine learning*. Essa significativa exclusão de dados poderá ser um dos fatores que contribuem para resultados inferiores nos *datasets* onde essa técnica é aplicada.

Ademais, é interessante observar que as técnicas de detecção de *outliers* podem ser utilizadas em conjunto. Por exemplo, ao selecionar as amostras identificadas como *outliers* por todos os três métodos, é possível refinar ainda mais o processo de filtragem. Nesse caso, a interseção das técnicas resulta em 71 amostras comuns ($\text{IQR} \cap \text{Z} - \text{Score} \cap \text{Isolation Forest}$), proporcionando uma abordagem mais robusta na identificação de *outliers*.

Figura 10 – Diagrama de *Venn* das UCs classificadas como *outlier*.



Fonte: Elaboração do próprio autor.

5.5.2 Balanceamento

O balanceamento dos dados foi realizado após a remoção dos *outliers*. A Tabela 5 apresenta o número de UCs restantes após essa etapa. Observa-se que os métodos de subamostragem, como o *random under sampler* e o *nearmiss*, reduziram drasticamente o número de amostras para treinamento. Isso pode ser problemático, especialmente quando combinado com técnicas de remoção de *outliers* como o IQR, a qual elimina um número significativo de dados, principalmente se essas amostras removidas pertencerem à classe minoritária de UCs irregulares.

Tabela 5 – Número de UCs após o balanceamento.

Balanceamento	*Outlier*	*UCs regulares*	*UCs irregulares*
SMOTE	Ausente	9014	9014
	IQR	5888	5888
	Z-Score	7929	7929
	Isolation Forest	8100	8100
ADASYN	Ausente	9127	9014
	IQR	5888	5879
	Z-Score	7957	7929
	Isolation Forest	8104	8100
Random Under Sampler	Ausente	935	935
	IQR	59	59
	Z-Score	527	527
	Isolation Forest	854	854
NEARMISS	Ausente	935	935
	IQR	59	59
	Z-Score	527	527
	Isolation Forest	854	854

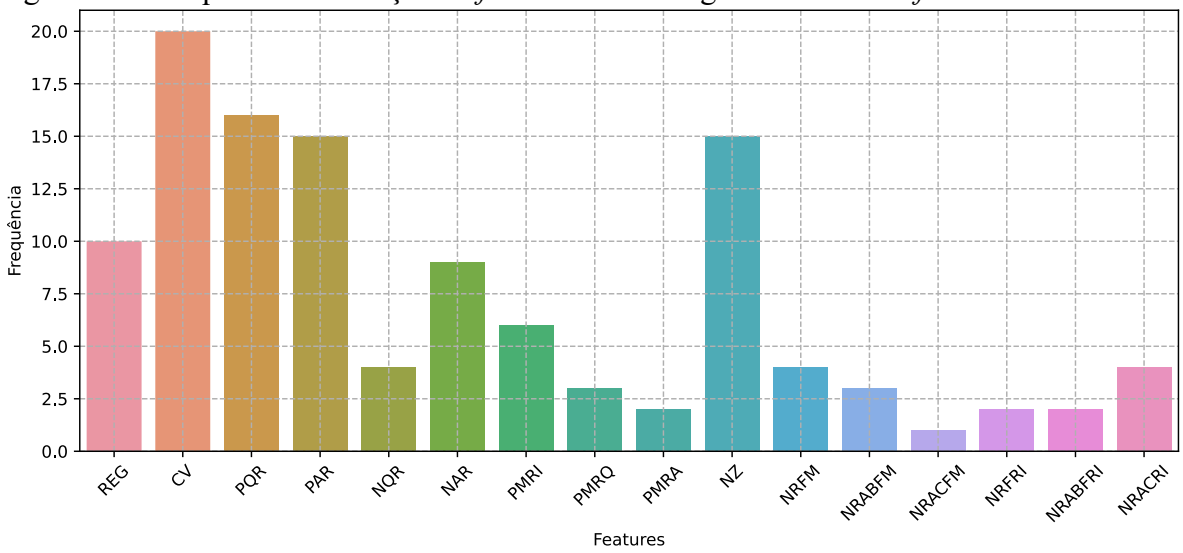
Fonte: Elaboração do próprio autor.

5.5.3 Feature Selection

A etapa de seleção das *features* mais representativas foi realizada por último, incluindo *datasets* ausentes de pré-processamento ou previamente ajustados com remoção de *outliers* e balanceamento de classes. Em seguida, cada método de seleção de *features* foi aplicado nos 20 *datasets* gerados nas etapas anteriores.

A Figura 11 apresenta a frequência de escolha de cada *feature* ao utilizar o algoritmo de *random forest*. A *feature* coeficiente de variação (CV) foi selecionada em todos os 20 *datasets* criados, sendo a mais frequente. Em segundo lugar, a *feature* percentual de quedas em relação ao regime vigente (PQR) foi escolhida 16 vezes, seguido pelo percentual de aumentos em relação ao regime vigente (PAR) e o número de zeros (NZ), cada uma presente em 15 dos 20 *datasets* criados.

Figura 11 – Frequência de seleção de *features* com o algoritmo *random forest.*

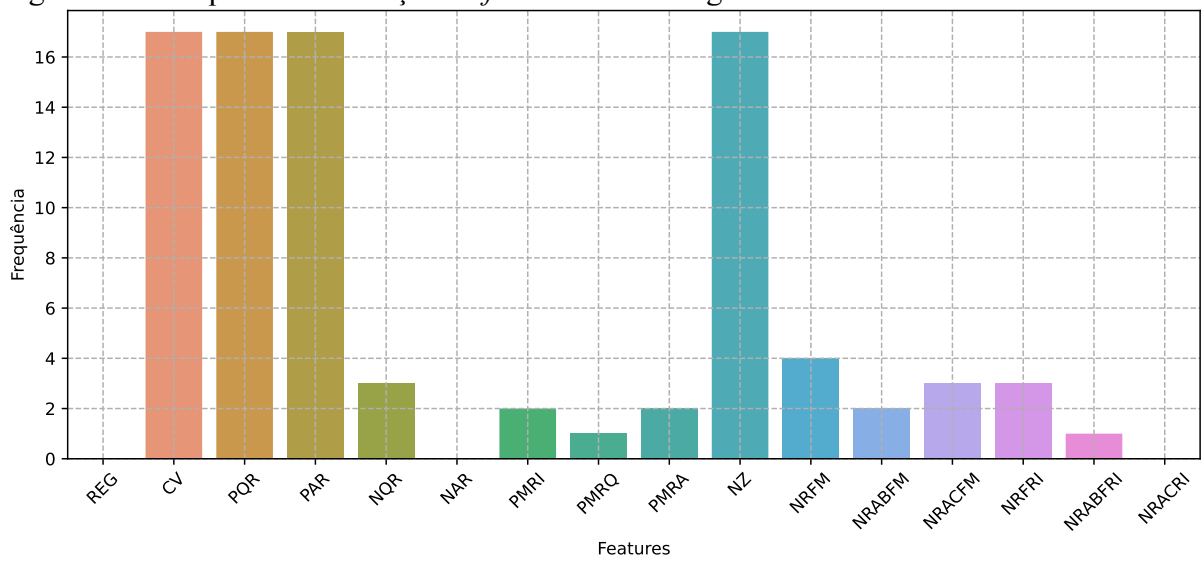


Fonte: Elaboração do próprio autor.

O algoritmo ANOVA selecionou as *features* CV, PQR, PAR e NZ em 17 dos 20 *datasets* criados, conforme ilustrado na Figura 12. Observa-se que as *features* número de regimes (REG), número de regimes de aumento (NAR) e números de regimes acima da faixa do regime inicial (NRACRI) não foram selecionadas em nenhum dos *datasets.* As demais *features* apresentaram uma baixa frequência de seleção.

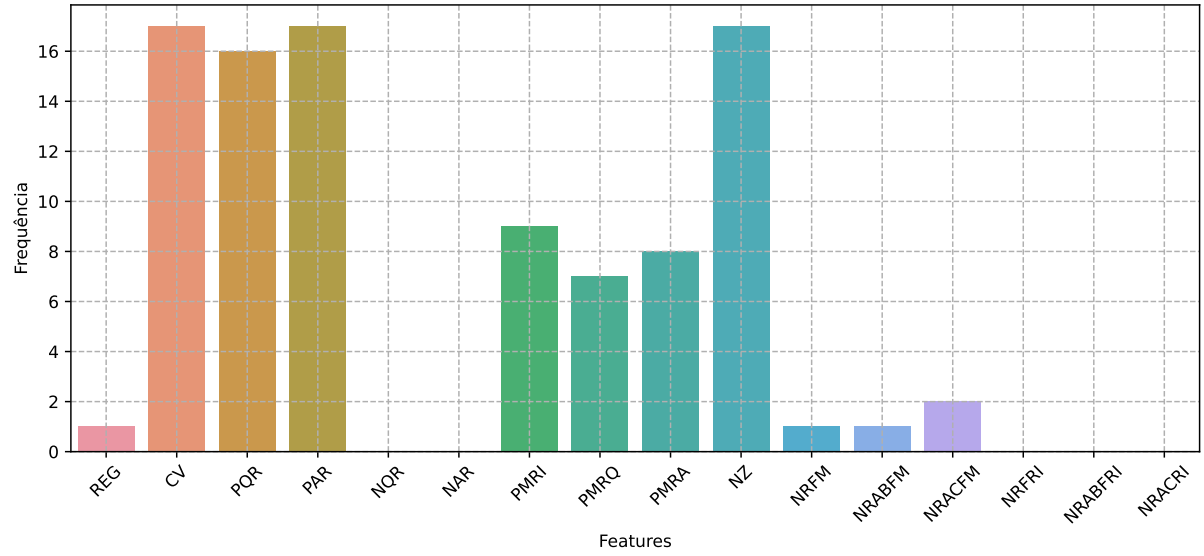
Por outro lado, o algoritmo *mutual info classif* selecionou as *features* CV, PAR e NZ em 17 dos 20 *datasets* criados. Em seguida, a *feature* PQR foi escolhida em 16 dos 20 *datasets*, conforme apresentado na Figura 13.

Figura 12 – Frequência de seleção de *features* com o algoritmo ANOVA.



Fonte: Elaboração do próprio autor.

Figura 13 – Frequência de seleção de *features* com o algoritmo *mutual info classif*.



Fonte: Elaboração do próprio autor.

Por fim, observamos que as *features* CV, PQR, PAR e NZ foram selecionadas com maior frequência ao aplicar os métodos de *feature selection* propostos. Esse resultado sugere que a escolha das *features* mais representativas pode não apenas melhorar o desempenho dos modelos de aprendizado de máquinas, bem como torná-los mais simples, de modo a reduzir o esforço computacional necessário. Portanto, um conjunto otimizado de *features* pode ser fundamental para alcançar uma boa relação custo-benefício nos modelos.

5.6 TREINAMENTO DA RNA PMC

5.6.1 Hiperparâmetros

No aprendizado supervisionado, é comum a necessidade de ajustar hiperparâmetros. A RNA PMC utilizada para realizar a classificação dos *datasets* do repositório descrito na Seção 5.4, foi implementada pela biblioteca de código aberto *scikit-learn* (Pedregosa et al., 2011).

A Tabela 6 apresenta os principais hiperparâmetros usados na etapa de treinamento, que foram os mesmos para todos os *datasets* do repositório. Essa padronização visa garantir uma base de dados para comparação justa no processo de avaliação das técnicas de pré-processamento.

Tabela 6 – Hiperparâmetros adotados no processo de treinamento da PMC.

Hiperparâmetros	*Descrição*	*Valor*
activation	Função de ativação dos neurônios	*relu*
alpha	Taxa de regularização L2, que ajuda a evitar o *overfitting* penalizando os pesos com grandes magnitudes	10^{-4}
batch_size	Tamanho do lote utilizado na atualização dos gradientes	auto
solver	Algoritmo de otimização baseado em gradiente estocástico proposto por Kingma e Ba (2015)	adam
beta_1	Influencia a suavização dos gradientes. Ele controla o quanto o otimizador vai considerar gradientes passados	0,9
beta_2	Ajusta a magnitude dos gradientes para cada parâmetro, para lidar com variações em diferentes direções de otimização	0,999
epsilon	Pequena constante que evita divisão por zero no *solver* Adam	10^{-8}
hidden_layer_sizes	Número de neurônios em cada camada oculta	100
learning_rate	Taxa de aprendizado constante	–
learning_rate_init	Valor inicial da taxa de aprendizado	10^{-3}
shuffle	Realiza o embaralhamento dos dados em cada iteração (época)	–

Fonte: Elaboração do próprio autor.

Vale destacar que seria possível realizar um ajuste fino dos hiperparâmetros (*hyperparameter tuning*) utilizando técnicas como a busca exaustiva em grade (*grid search*), que testa todas as combinações possíveis de hiperparâmetros. No entanto, essa abordagem pode gerar um espaço de busca muito grande. Para mitigar esse problema, pode-se empregar a busca randômica (*randomized search*), que seleciona aleatoriamente combinações de hiperparâmetros, sendo uma técnica amplamente utilizada (Bergstra; Bengio, 2012).

5.6.2 Função de ativação

As funções de ativação determinam como as saídas dos neurônios são calculadas nas RNAs e, consequentemente, como as informações são propagadas através da rede. No contexto deste trabalho, a função de ativação utilizada é a *rectified linear unit* (ReLU), descrita

matematicamente em (14). A ReLU, mostrada na Figura 14, permite que as saídas sejam valores no intervalo $[0, \infty)$, garantindo que os neurônios possam produzir respostas não lineares e contribuindo para a eficácia do aprendizado.

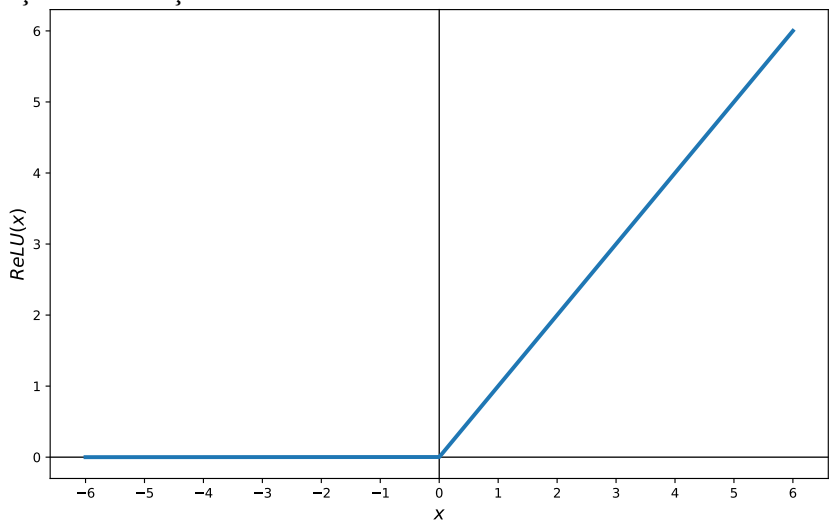
Em um problema de classificação binária, como o que estamos tratando, as classes de UCs irregulares e regulares são codificadas, respectivamente, como um e zero. Isso implica que a última camada da rede contém apenas um único neurônio, cujo objetivo é prever a probabilidade de uma amostra pertencer a cada uma das classes. Para esse neurônio, a função de ativação utilizada é a sigmoide junto com a Heaviside deslocada (ou função degrau), que é expressa em (15) e é apresentada na Figura 15, onde $\hat{y}$ representa a classe predita. A sigmoide retorna um valor no intervalo [0,1], interpretando-o como uma probabilidade (GEEKS FOR GEEKS, 2024e).

$$\mathrm{ReLU}(x) = \max(0, x) \tag{14}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{15}$$

$$\hat{y} = H = \begin{cases} 1, & \sigma(x) \geq 0{,}5 \\ 0, & \sigma(x) < 0{,}5 \end{cases}$$

Figura 14 – Função de ativação ReLU.



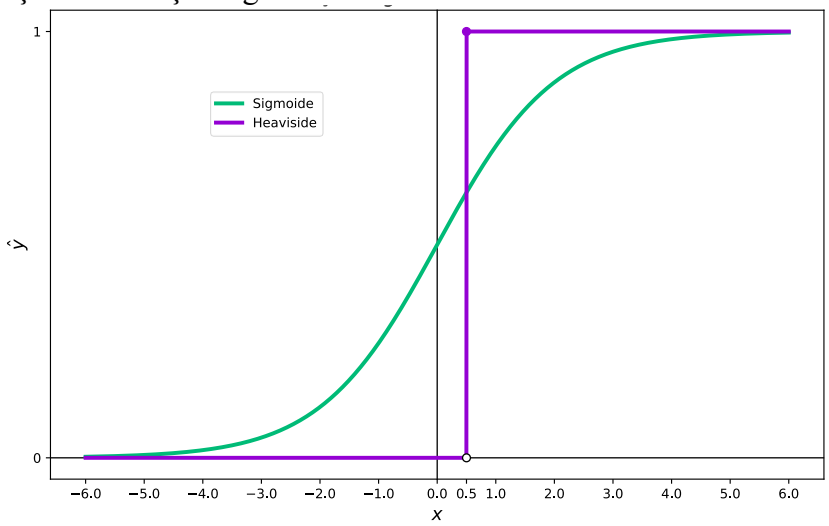
Fonte: Elaboração do próprio autor.

5.6.3 Normalização

A normalização dos dados, tanto na etapa de treinamento quanto na etapa de teste, é uma prática recomendada no desenvolvimento de modelos de aprendizado de máquina. Esse processo consiste em ajustar cada feature do *dataset* de forma individual, de modo que todas compartilhem os mesmos valores de máximo e mínimo. Essa normalização é fundamental, pois

equilibra a importância de cada *feature* durante o treinamento, permitindo que todas sejam tratadas de maneira justa e evitando que variáveis com magnitudes maiores dominem a aprendizagem do modelo.

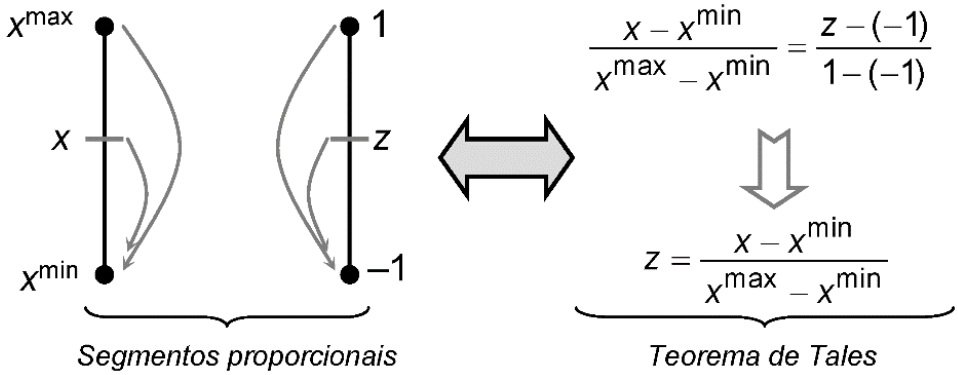
Figura 15 – Função de ativação sigmoide e Heaviside.



Fonte: Elaboração do próprio autor.

A Figura 16 ilustra uma demonstração geométrica fundamentada no Teorema de Tales, evidenciando como a proporção entre segmentos permite o ajuste para diferentes escalas. Esse princípio é diretamente aplicável ao processo de normalização, o qual pode ser adaptado para trazer variáveis de diferentes escalas a uma mesma faixa, facilitando comparações e análises.

Figura 16 – Normalização: interpretação geométrica e teorema de Tales.



Fonte: Silva et al. (2016).

A escolha da técnica de normalização deve ser guiada pelas características específicas do *dataset*, uma vez que diferentes tipos de dados podem exigir abordagens distintas. Além disso, o intervalo de saída gerado pela normalização deve ser compatível com as funções de ativação empregadas nos neurônios da RNA. Assim, a normalização utilizada é apresentada em

(16) e pode ser vista como um processo que garante uma junção adequada entre os dados e o modelo, assegurando que a informação seja utilizada de forma eficiente e eficaz ao longo do treinamento.

$$x_i' = \frac{x_i - \min(\boldsymbol{X})}{\max(\boldsymbol{X}) - \min(\boldsymbol{X})} \tag{16}$$

Por fim, é essencial que a normalização aplicada ao *dataset* de teste utilize os valores de $\min(\boldsymbol{X})$ e $\max(\boldsymbol{X})$ apresentados em (16), extraídos do *dataset* de treinamento. Isso garante consistência entre as etapas de treinamento e teste, uma vez que aplicar diferentes escalas de normalização pode introduzir distorções no modelo. Assim, o modelo será avaliado corretamente com base nas mesmas condições de treinamento.

5.7 RESULTADOS DA FASE DE TESTES

A etapa de teste avalia o aprendizado dos modelos de aprendizado de máquina, sendo diretamente influenciado pelos dados usados no treinamento. Essa etapa permite identificar quais métodos de pré-processamento proporcionam o melhor desempenho da RNA PMC.

Inicialmente, será analisado o caso base, no qual não foram aplicados os métodos de pré-processamento. A Tabela 7 apresenta os resultados das métricas descritas na Seção 3.2 nos *datasets*, onde apenas um método de pré-processamento é aplicado. O caso em que não houve a aplicação de nenhum método de pré-processamento, apresenta uma acurácia alta em comparação à literatura (Avila et al., 2018; Figueroa et al., 2018; Ghori et al., 2020).

No entanto, no contexto das PNTs, a acurácia não reflete com precisão o desempenho do classificador, uma vez que o problema é naturalmente desbalanceado, com a classe alvo sendo minoritária. Por isso, a métrica F1-Score será o foco da avaliação.

Ausente da aplicação de pré-processamento, o modelo apresentou um F1-Score de 0,88 e um tempo de execução de 10,93s, levantando dúvidas sobre a necessidade de aplicar tais técnicas para melhorar o desempenho. Entretanto, em *datasets* de maior dimensionalidade, essa abordagem pode resultar em um aumento significativo do tempo de execução, além de potenciais reduções no desempenho do modelo.

Os resultados da aplicação isolada dos métodos de pré-processamento indicam que a seleção de características reduziu o custo computacional, mantendo um F1-Score aceitável, conforme mostrado na Tabela 7.

Tabela 7 – Desempenho das técnicas de pré-processamento isoladas.

Pré-processamento	*Acurácia*	*Precisão*	*Recall*	*F1-Score*	*t_exec (s)*
Ausente	0,98	0,88	0,88	0,88	10,93
Outlier	0,98±0,01	0,90±0,09	0,55±0,44	0,57±0,44	8,48±1,95
Balanceamento	0,85±0,24	0,63±0,32	0,91±0,03	0,69±0,29	13,89±14,45
Feature Selection	0,97±0,01	0,86±0,01	0,81±0,06	0,84±0,03	6,75±2,82

Fonte: Elaboração do próprio autor.

A combinação da Figura 17 e Figura 18 permite avaliar o desempenho de cada *dataset* em relação aos métodos de pré-processamento aplicados, utilizando a distribuição do F1-Score para cada método de remoção de *outliers*, apresentada em *boxplots.*

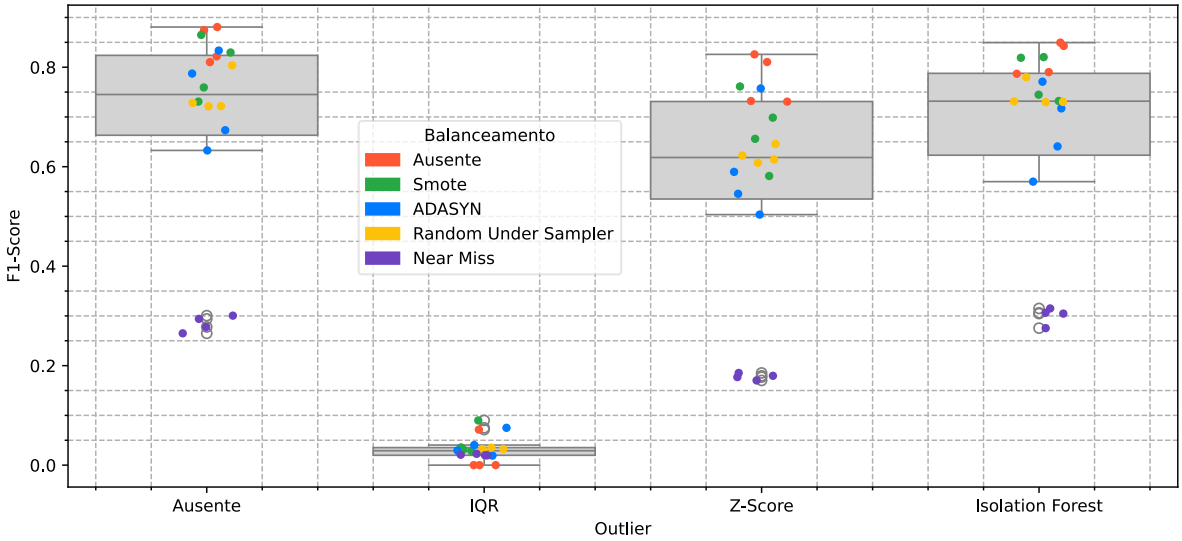
Observa-se que, quando aplicado o método de remoção de *outliers* e IQR ao *dataset*, conforme ilustrado na Figura 17, a PMC apresenta um F1-Score significativamente inferior em comparação às demais combinações de pré-processamento. Essa constatação, relevante para o problema, levanta a discussão abordada na Seção 5.5.1, pois a exclusão de cerca de 33,58% das amostras de treinamento pode ter comprometido a capacidade de aprendizagem do modelo, tornando os resultados pouco satisfatórios.

É evidente que, ao aplicar o método de balanceamento *nearmiss*, o modelo apresentou um F1-Score de $0{,}2 \pm 0{,}12$, significativamente inferior à média geral de $0{,}49 \pm 0{,}32$, impactando negativamente a eficácia das técnicas de balanceamento, como apresentado na Figura 17.

Adicionalmente, a Figura 18 permite observar o F1-Score sob a perspectiva dos métodos de *feature selection*, sem indicar uma tendência clara, seja de redução ou aumento no desempenho dos modelos. Isso sugere que a escolha das *features* mais significativas pode ser benéfica quanto a simplificação de modelos de aprendizado de máquina, sem comprometer seu desempenho.

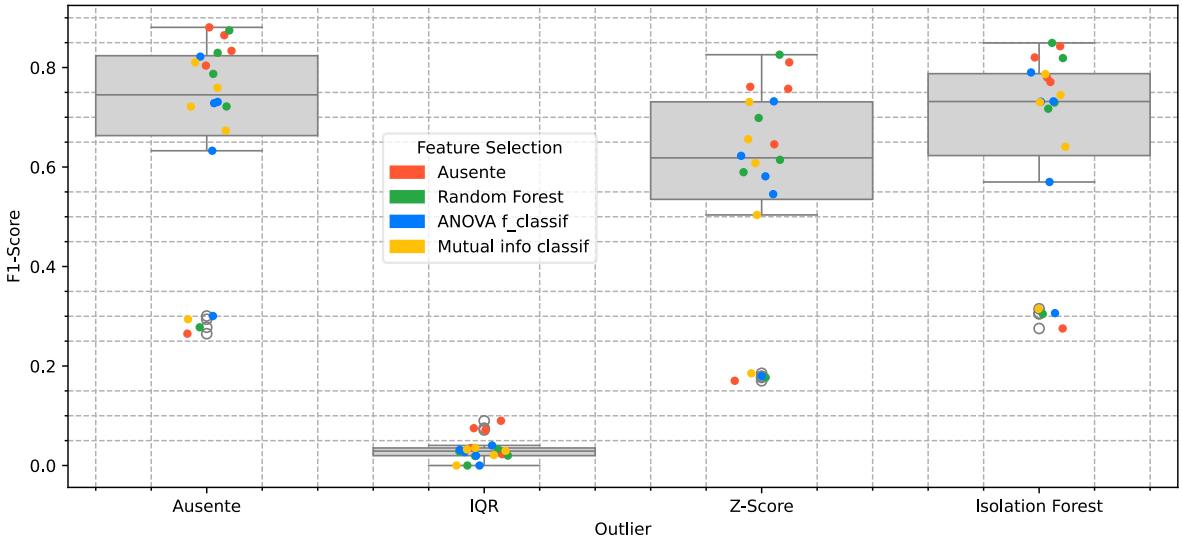
A Figura 19 mostra a relação entre a métrica F1-Score e o tempo de treinamento da PMC para cada *dataset* do repositório com os três grupos principais de resultados: Na parte inferior esquerda, estão os casos com F1-Score e tempo de execução baixos, indicando desempenho insuficiente. Na parte superior direita, os casos apresentam um bom F1-Score, mas com tempos de execução elevados, o que pode ser um fator limitante dependendo dos recursos disponíveis. Os resultados na parte superior esquerda mostram um alto F1-Score combinado com um tempo de treinamento reduzido, sendo esses os mais vantajosos pela eficiência e desempenho.

Figura 17 – Boxplot: *outliers versus* balanceamento.



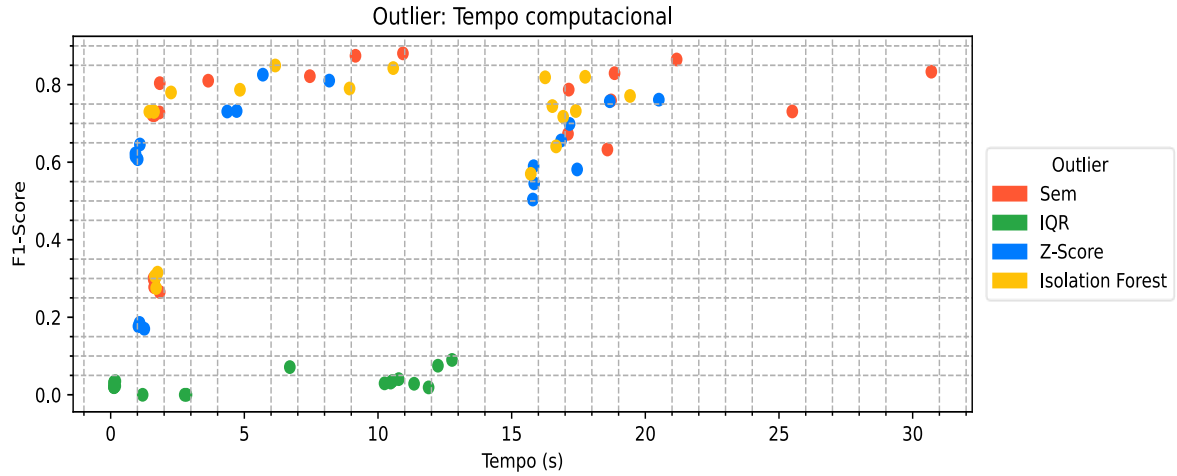
Fonte: Elaboração do próprio autor.

Figura 18 – Boxplot: *outliers versus Feature Selection.*



Fonte: Elaboração do próprio autor.

Figura 19 – F1-Score *versus* tempo de treinamento.



Fonte: Elaboração do próprio autor.

Em síntese, a aplicação dos métodos de pré-processamento gerou diversas relações entre desempenho e tempo de treinamento, evidenciando que o pré-processamento pode impactar os resultados dos modelos de aprendizado de máquina. Isso ressalta a importância de buscar empiricamente as melhores combinações, além de explorar novas metodologias. A Tabela 8 apresenta o grupo de resultados considerados mais vantajosos, com as 24 possibilidades de pré-processamento, incluindo o treinamento realizado ausente de métodos de pré-processamento.

Tabela 8 – Grupo de resultados vantajosos e seus métodos de pré-processamento.

Outlier	*Balanceamento*	*Feature Selection*	*Acurácia*	*Precisão*	*Recall*	*F1-Score*	*t_exec (s)*
Ausente	Ausente	Ausente	0,977	**0,881**	0,881	0,881	10,933
		Random Forest	0,976	0,876	0,873	**0,875**	9,159
		ANOVA f_classif	0,967	0,861	0,786	0,822	7,452
		Mutual info classif	0,965	0,856	0,769	0,810	3,649
	Random Under Sampler	Ausente	0,957	0,721	**0,908**	0,804	1,836
		Random Forest	0,934	0,606	0,893	0,722	1,567
		ANOVA f_classif	0,935	0,612	0,900	0,728	1,824
		Mutual info classif	0,933	0,603	0,898	0,721	1,633
Z-Score	Ausente	Ausente	0,976	0,823	0,798	0,810	8,176
		Random Forest	**0,978**	0,828	0,824	0,826	5,693
		ANOVA f_classif	0,966	0,743	0,721	0,732	4,713
		Mutual info classif	0,965	0,728	0,734	0,731	4,360
	Random Under Sampler	Ausente	0,938	0,510	0,880	0,646	1,094
		Random Forest	0,929	0,473	0,876	0,614	0,929
		ANOVA f_classif	0,932	0,486	0,867	0,622	**0,926**
		Mutual info classif	0,928	0,467	0,871	0,608	1,011
Isolation Forest	Ausente	Ausente	0,970	0,874	0,814	0,843	10,570
		Random Forest	0,971	0,873	0,827	0,849	6,157
		ANOVA f_classif	0,960	0,830	0,754	0,790	8,937
		Mutual info classif	0,959	0,823	0,754	0,787	4,837
	Random Under Sampler	Ausente	0,951	0,706	0,872	0,780	2,258
		Random Forest	0,937	0,634	0,859	0,730	1,621
		ANOVA f_classif	0,937	0,637	0,859	0,731	1,604
		Mutual info classif	0,937	0,634	0,861	0,730	1,445

Fonte: Elaboração do próprio autor.

Por outro lado, a Tabela 9 detalha os demais resultados obtidos, classificados como menos interessantes para utilização prática. É importante destacar que os resultados poderiam variar caso cada modelo, na etapa de treinamento, fosse submetido a uma otimização de hiperparâmetros. Além disso, a utilização de diferentes algoritmos de classificação pode impactar os resultados para cada *dataset* gerado pelas técnicas de pré-processamento. Isso sugere que combinações que apresentaram baixo desempenho para a RNA PMC podem ser mais adequadas quando aplicadas a outros algoritmos.

Tabela 9 – Resultados para os demais métodos de pré-processamento aplicados.

Outlier	*Balanceamento*	*Feature Selection*	*Acurácia*	*Precisão*	*Recall*	*F1-Score*	*t_exec (s)*
Ausente	SMOTE	Ausente	0,973	0,843	0,888	0,865	**21,172**
		Random Forest	0,964	0,769	0,900	0,830	18,849
		ANOVA f_classif	0,936	0,617	0,895	0,731	25,498
		Mutual info classif	0,946	0,662	0,891	0,759	18,718
	ADASYN	Ausente	0,966	0,789	0,883	0,834	**30,699**
		Random Forest	0,953	0,699	0,900	0,787	17,135
		ANOVA f_classif	0,898	0,486	0,908	0,633	18,582
		Mutual info classif	0,916	0,537	0,903	0,673	17,107
	NEARMISS	Ausente	0,491	0,154	0,951	0,265	1,836
		Random Forest	0,526	0,163	0,946	0,278	1,625
		ANOVA f_classif	0,573	0,178	0,951	0,301	1,602
		Mutual info classif	0,561	0,174	0,949	0,294	1,630
IQR	Ausente	Ausente	0,990	1,000	0,037	0,071	6,701
		Random Forest	0,989	0,000	0,000	0,000	2,837
		ANOVA f_classif	0,989	0,000	0,000	**0,000**	1,196
		Mutual info classif	0,989	0,000	0,000	**0,000**	2,770
	SMOTE	Ausente	0,968	0,065	0,148	0,090	12,769
		Random Forest	0,920	0,016	**0,111**	0,028	11,350
		ANOVA f_classif	0,855	0,017	0,222	0,031	10,472
		Mutual info classif	0,873	0,019	0,222	0,036	10,539
	ADASYN	Ausente	0,971	0,057	0,111	0,075	12,242
		Random Forest	0,919	0,011	0,074	0,019	11,893
		ANOVA f_classif	0,869	0,022	0,259	0,040	10,775
		Mutual info classif	0,845	0,016	0,222	0,029	10,245
		Ausente	0,569	0,018	0,741	0,035	0,176

	Random Under Sampler	Random Forest	0,533	0,017	0,741	0,033	0,134
		ANOVA f_classif	0,461	0,014	0,704	0,027	0,135
		Mutual info classif	0,508	0,017	0,778	0,032	0,129
	NEARMISS	Ausente	0,191	0,012	0,889	0,023	0,173
		Random Forest	0,187	0,010	0,778	**0,020**	0,125
		ANOVA f_classif	**0,175**	**0,010**	0,778	**0,020**	0,126
		Mutual info classif	0,228	0,011	0,778	**0,021**	0,126
Z-Score	SMOTE	Ausente	0,968	0,731	0,794	0,761	20,505
		Random Forest	0,951	0,581	0,876	0,699	17,167
		ANOVA f_classif	0,920	0,437	0,867	0,581	17,449
		Mutual info classif	0,941	0,523	0,880	0,656	16,850
	ADASYN	Ausente	0,966	0,691	0,837	0,757	18,670
		Random Forest	0,922	0,444	0,876	0,590	15,821
		ANOVA f_classif	0,912	0,408	0,824	0,545	15,843
		Mutual info classif	0,889	0,353	0,880	0,504	15,793
	NEARMISS	Ausente	0,417	0,094	0,931	0,170	1,259
		Random Forest	0,446	0,098	0,927	0,177	1,039
		ANOVA f_classif	0,448	0,099	0,940	0,180	1,091
		Mutual info classif	0,476	0,103	0,927	0,185	1,080
Isolation Forest	SMOTE	Ausente	0,964	0,816	0,825	0,820	17,750
		Random Forest	0,962	0,783	0,859	0,819	16,253
		ANOVA f_classif	0,939	0,648	0,840	0,732	17,400
		Mutual info classif	0,943	0,674	0,832	0,745	16,522
	ADASYN	Ausente	0,950	0,708	0,846	0,771	19,428
		Random Forest	0,932	0,612	0,866	0,717	16,928
		ANOVA f_classif	0,870	0,425	0,864	0,570	15,715
		Mutual info classif	0,904	0,509	0,864	0,641	16,662
	NEARMISS	Ausente	0,512	0,162	0,932	0,275	1,701
		Random Forest	0,587	0,183	0,908	0,305	1,694
		ANOVA f_classif	0,593	0,184	0,903	0,306	1,654
		Mutual info classif	0,607	0,191	0,908	0,315	1,754

Fonte: Elaboração do próprio autor.

6 CONCLUSÃO

Neste trabalho, foi realizada uma análise comparativa de métodos de pré-processamento, incluindo a remoção de *outliers*, balanceamento de classes e a seleção das *features* mais representativas, aplicados ao problema de detecção de perdas não técnicas (PNTs) por meio de uma rede neural artificial (RNA) do tipo *perceptron* multicamadas (PMC). O objetivo foi classificar as unidades consumidoras (UCs) como regulares ou irregulares em uma base de dados de uma cidade do interior de São Paulo com aproximadamente 200 mil habitantes.

O desempenho da RNA PMC foi avaliado com diferentes *datasets*, criados a partir de combinações dos métodos de pré-processamento aplicados em etapas específicas. A eficácia dos modelos foi medida principalmente pelo F1-Score, métrica relevante para o problema de detecção de PNTs, além do tempo de treinamento da PMC para cada *dataset*.

Para garantir uma avaliação justa, foi utilizado o mesmo *dataset* de teste para todos os modelos gerados. A criação do conjunto de dados inicial incluiu uma redução no número de amostras, mantendo a proporção de furtos por setor censitário próxima ao valor original. A etapa de engenharia de atributos gerou 16 *features* estatísticas, extraídas com base no consumo mensal em kWh dos consumidores ao longo de três anos.

Visando simplificar o treinamento da PMC, optou-se por não realizar o ajuste de hiperparâmetros para cada *dataset*. Embora essa decisão tenha facilitado a implementação e tornado a avaliação dos resultados mais igualitária, também limitou a possibilidade de alcançar melhores resultados.

Na etapa de teste, tanto o modelo ausente de métodos de pré-processamento quanto algumas combinações desses métodos apresentaram F1-Score satisfatórios. Observou-se que a aplicação dos métodos de pré-processamento influenciou, nesse problema, principalmente na redução do tempo de treinamento, com pouca variação na capacidade de detecção das UCs regulares e irregulares. Ressalta-se que, em *datasets* maiores, a vantagem na redução do tempo de treinamento torna-se mais expressiva. Adicionalmente, essas técnicas podem também melhorar a precisão do modelo.

O desempenho satisfatório alcançado por um subconjunto de modelos treinados evidencia a importância de utilizar múltiplas combinações de métodos de pré-processamento. Essa abordagem permite avaliar o custo-benefício de cada modelo e identificar aqueles que melhor se adequem à etapa de produção, onde o modelo será disponibilizado para a detecção de UCs irregulares e influenciará diretamente no processo de inspeção.

Por fim, a aplicação de aprendizado de máquina e ciência de dados, com o uso de métodos de pré-processamento para a detecção de PNTs, não elimina a necessidade de investimentos no desenvolvimento humano na área. Embora essas técnicas busquem automatizar a captura de padrões nos dados, a criatividade e a curiosidade dos profissionais são fundamentais para interpretar resultados, ajustar abordagens e explorar novos caminhos para a melhoria contínua, mantendo a ciência de dados como uma área sempre em evolução.

6.1 TRABALHOS FUTUROS

- **Aplicação em outros algoritmos de classificação:** Avaliar o desempenho do repositório de *datasets* desenvolvido quando aplicado a outros algoritmos de aprendizado de máquina, como árvores de decisão, *ensemble methods* (e.g., Random Forest e Gradient Boosting) e métodos baseados em *deep learning* (e.g., redes convolucionais e transformadores). Essa análise pode ampliar as comparações e identificar alternativas promissoras para a detecção de PNTs;
- **Análise em *datasets* maiores:** Aplicar as técnicas propostas em *datasets* mais amplos, que abranjam diferentes regiões e características de consumo, para avaliar a capacidade de generalização dos modelos em cenários variados;
- **Combinação de técnicas de pré-processamento:** Explorar o uso combinado de métodos de pré-processamento, como a remoção de outliers combinando os métodos disponível, tornando as etapas mais robustas e eficazes;
- **Exploração de novas técnicas de pré-processamento:** Pesquisar e avaliar novas abordagens para o pré-processamento de dados, ampliando as possibilidades de transformação e aprimoramento dos dados para a classificação.
- **Aprimoramento das características extraídas:** Investigar métodos avançados de engenharia de atributos, buscando extrair características mais representativas dos dados originais e reduzir a influência de ruídos. Exemplos incluem a aplicação de coeficientes de *wavelets*, transformada de *Fourier* e análise de séries temporais para enriquecer a representação dos dados;
- **Otimização de hiperparâmetros:** Implementar técnicas de otimização de hiperparâmetros, como busca em grade (*grid search*), busca randômica (*randomized search*) ou algoritmos evolucionários, para ajustar os modelos de classificação e alcançar melhores resultados de desempenho.

REFERÊNCIAS

ALABRAH, A. An Improved CCF detector to handle the problem of class imbalance with outlier normalization using IQR method. **Sensors**, v. 23, n. 9, p. 1–14, 30 abr. 2023.

ANEEL. **PRORET 2.6:** Perdas de energia. Disponível em: <https://www2.aneel.gov.br/cedoc/aren2015660_Proret_Submod_2_6_V3.pdf>. Acesso em: 19 nov. 2024.

ANEEL. **ANEXO 1 AIR:** Variáveis de faturamento das componentes tarifárias. Disponível em: <https://antigo.aneel.gov.br/web/guest/audiencias-publicas-antigas?p_p_id=participacaopublica_WAR_participacaopublicaportlet&p_p_lifecycle=2&p_p_state=normal&p_p_mode=view&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_pos=1&p_p_col_count=2&_participacaopublica_WAR_participacaopublicaportlet_ideDocumento=31756&_participacaopublica_WAR_participacaopublicaportlet_tipoFaseReuniao=fase&_participacaopublica_WAR_participacaopublicaportlet_jspPage=/html/pp/visualizar.jsp>. Acesso em: 21 set. 2024a.

ANEEL. **PRODIST Módulo 7:** Cálculo de perdas na distribuição. Disponível em: <https://www2.aneel.gov.br/cedoc/aren2017771_prodist_modulo_7_v5.pdf>. Acesso em: 27 set. 2024b.

ANEEL. **PRORET 7.1:** Estrutura tarifária das concessionárias de distribuição. Disponível em: <https://www2.aneel.gov.br/cedoc/aren20231060_2_1.pdf>. Acesso em: 20 nov. 2024.

ANEEL. **Relatório de perdas de energia**. Disponível em: <https://portalrelatorios.aneel.gov.br/luznatarifa/perdasenergias>. Acesso em: 12 set. 2024.

ANUSHA, P. V. et al. Detecting outliers in high dimensional data sets using Z-score methodology. **International Journal of Innovative Technology and Exploring Engineering**, v. 9, n. 1, p. 48–53, 1 nov. 2019.

AVILA, N. F.; FIGUEROA, G.; CHU, C. C. NTL detection in electric distribution systems using the maximal overlap discrete wavelet-packet transform and random undersampling boosting. **IEEE Transactions on Power Systems**, v. 33, n. 6, p. 7171–7180, 2018.

AZANK, F.; GURGEL, G. K. **Dados desbalanceados:** O que são e como lidar com eles. Disponível em: <https://medium.com/turing-talks/dados-desbalanceados-o-que-são-e-como-evitá-los-43df4f49732b>. Acesso em: 19 set. 2024.

BERGSTRA, J.; BENGIO, Y. Random Search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, p. 281–305, fev. 2012.

BISHOP, C. M. **Pattern recognition and machine learning:** Information science and statistics. 1. ed. Berlin, Heidelberg: Springer, 2006.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

CASTRO, N. DE et al. **As perdas não técnicas no setor de distribuição brasileiro: uma abordagem regulatória**. 1. ed. Rio de Janeiro, RJ: GESEL, 2020.

CHAWLA, N. V. et al. SMOTE: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, jun. 2002.

EPE - EMPRESA DE PESQUISA ENERGÉTICA. **BEN 2023: Relatório síntese**. Disponível em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-748/topico-681/BEN_Síntese_2023_PT.pdf>. Acesso em: 19 maio. 2024.

FARIA, L. T.; FELTRIN, A. P. **Estimação espaço-temporal das perdas não técnicas no sistema de distribuição de energia elétrica**. Tese de Doutorado. Ilha Solteira, SP: Universidade Estadual Paulista - Unesp, 26 fev. 2016.

FARIA, L. T.; FELTRIN, A. P.; MINUSSI, C. R. **Sistema inteligente híbrido intercomunicativo para detecção de perdas comerciais**. Dissertação de Mestrado. Ilha Solteira, SP: Universidade Estadual Paulista - Unesp, 1 mar. 2012.

FERREIRA, H. M. **Uso de ferramentas de aprendizado de máquina para prospecção de perdas comerciais em distribuição de energia elétrica**. Dissertação de Mestrado. Campinas, SP: Universidade Estadual de Campinas - Unicamp, jan. 2008.

FIGUEROA, G. et al. **Improved practices in machine learning algorithms for NTL detection with imbalanced data**. IEEE Power and Energy Society General Meeting, 2018.

FITZGERALD, A. E. **Máquinas elétricas**. 7. ed. Porto Alegre, RS: AMGH, 2014.

GEEKS FOR GEEKS. **Feature selection techniques in machine learning**. Disponível em: <https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/>. Acesso em: 4 ago. 2024a.

GEEKS FOR GEEKS. **Feature selection in Python with Scikit-Learn**. Disponível em: <https://www.geeksforgeeks.org/feature-selection-in-python-with-scikit-learn/>. Acesso em: 4 ago. 2024b.

GEEKS FOR GEEKS. **Understanding the confusion matrix in machine learning**. Disponível em: <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>. Acesso em: 18 nov. 2024c.

GEEKS FOR GEEKS. **One hot encoding in machine learning**. Disponível em: <https://www.geeksforgeeks.org/ml-one-hot-encoding/>. Acesso em: 7 nov. 2024d.

GEEKS FOR GEEKS. **Activation functions in neural networks**. Disponível em: <https://www.geeksforgeeks.org/activation-functions-neural-networks/>. Acesso em: 19 nov. 2024e.

GHORI, K. M. et al. Performance analysis of different types of machine learning classifiers for non-technical loss detection. **IEEE Access**, v. 8, p. 16033–16048, 27 jan. 2020.

HAQ, E. U. et al. A hybrid approach based on deep learning and support vector machine for the detection of electricity theft in power grids. **Energy Reports**, v. 7, p. 349–56, 5 ago. 2021.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, v. 585, n. 7825, p. 357–362, 16 set. 2020.

HAYKIN, S. **Redes neurais artificiais:** Princípios e prática. 2ª Edição ed. Porto Alegre: Tradução Paulo Martins Engel., 2001.

HE, H. et al. **ADASYN:** Adaptive synthetic sampling approach for imbalanced learning. Hong Kong: Proceedings of the International Joint Conference on Neural Networks, 26 set. 2008.

HEYDARIAN, M.; DOYLE, T. E.; SAMAVI, R. MLCM: Multi-Label Confusion Matrix. **IEEE Access**, v. 10, p. 19083–19095, 2022.

HUNTER, J. D. Matplotlib: A 2D graphics environment. **Computing in Science and Engineering**, v. 9, n. 3, 2007.

INSTITUTO ACENDE BRASIL. **Análise do processo de revisão tarifária e da regulação por incentivos**. São Paulo, SP. Disponível em: <https://acendebrasil.com.br/wp-content/uploads/2020/04/Caderno_01_Regulacao_por_Incentivos.pdf>. Acesso em: 5 jul. 2023.

INSTITUTO ACENDE BRASIL. **Tarifas de energia e os benefícios da regulação por incentivos**. São Paulo, SP. Disponível em: <https://acendebrasil.com.br/estudo/white-paper-n-3-tarifas-de-energia-e-os-beneficios-da-regulacao-por-incentivos/>. Acesso em: 17 nov. 2024.

INSTITUTO ACENDE BRASIL. **Perdas comerciais e inadimplência no setor elétrico**. São Paulo, SP. Disponível em: <https://acendebrasil.com.br/estudo/white-paper-18-perdas-comerciais-e-inadimpl-ncia-no-setor-eletrico/>. Acesso em: 15 out. 2024.

JEYARAJ, P. R. et al. Smart grid security enhancement by detection and classification of non-technical losses employing deep learning algorithm. **International Transactions on Electrical Energy Systems**, v. 30, n. 9, 2020.

JORDAHL, K. et al. **Geopandas:** v0.8.1. Disponível em: <https://geopandas.org/en/stable/index.html>. Acesso em: 6 out. 2024.

KINGMA, D. P.; BA, J. L. **Adam:** A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015.

KLUYVER, T. et al. **Jupyter Notebooks:** A publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas - Proceedings of the 20th International Conference on Electronic Publishing, ELPUB 2016, jun. 2016.

KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. **Physical Review E**, v. 69, n. 6, p. 1–16, 23 jun. 2004.

LEMAÎTRE, G.; NOGUEIRA, F.; ARIDAS, C. K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. **Journal of Machine Learning Research**, v. 18, p. 1, 2017.

LIU, F. T.; TING, K. M.; ZHOU, Z. H. **Isolation forest**. (IEEE, Ed.)Pisa, Italy: Proceedings - IEEE International Conference on Data Mining, ICDM, fev. 2008.

LOUPPE, G. **Understanding Random Forests:** From theory to practice. PhD dissertation. Liège, Bélgica: Université de Liège, jun. 2015.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The Bulletin of Mathematical Biophysics**, v. 5, n. 4, p. 115–133, 1943.

MCKINNEY, W. **Data structures for statistical computing in Python**. Proceedings of the 9th Python in Science Conference, 2010.

MESSINIS, G. M.; HATZIARGYRIOU, N. D. Review of non-technical loss detection methods. **Electric Power Systems Research**, v. 158, p. 250–266, 3 fev. 2018.

MONTICELLI, A. J.; GARCIA, A. V. **Introdução a sistemas de energia elétrica**. 2ª ed. ed. Campinas, SP: Editora da Unicamp, 2011.

MONYELLE, H.; COSTA, E. G. DA; ARAÚJO, J. F. DE. **Análise da contribuição de atributos derivados do histórico de consumo para a detecção de perdas não técnicas**. Dissertação de Mestrado. Campina Grande, PB: Universidade Federal de Campina Grande, jul. 2019.

NAGI, J. et al. Nontechnical loss detection for metered customers in power utility using support vector machines. **IEEE Transactions on Power Delivery**, v. 25, n. 2, p. 1162–1171, abr. 2010.

NGUYEN, Q. H. et al. Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. **Mathematical Problems in Engineering**, v. 2021, 8 fev. 2021.

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

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533–536, 9 out. 1986.

SAEED, M. S. et al. Detection of non-technical losses in power utilities: A comprehensive systematic review. **Energies**, v. 13, n. 18, 1 set. 2020.

SAVIAN, F. DE S. et al. Non-technical losses: A systematic contemporary article review. **Renewable and Sustainable Energy Reviews**, v. 147, 11 maio 2021.

SHARMA, A.; PUROHIT, A.; MISHRA, H. A survey on imbalanced data handling techniques for classification. **International Journal of Emerging Trends in Engineering Research**, v. 9, n. 10, 7 out. 2021.

SILVA, I. N. DA; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências Aplicadas:** Fundamentos teóricos e aspectos práticos. 2. ed. São Paulo, SP: Artliber, 2016.

SOMAN, A. B. **Gini impurity vs Gini importance vs mean decrease impurity**. Disponível em: <https://medium.com/@aneesha161994/gini-impurity-vs-gini-importance-vs-mean-decrease-impurity-51408bdd0cf1>. Acesso em: 15 out. 2024.

VAN ROSSUM, G.; DRAKE, F. L. **Python 3 reference manual**. Scotts Valley, CA: Createspace Independent Pub, 2009.

VENTURA, L. et al. Estimation of non-technical loss rates by regions. **Electric Power Systems Research**, v. 223, 1 out. 2023.

VISUAL STUDIO CODE. **Visual Studio Code documentation**. Disponível em: <https://code.visualstudio.com/docs>. Acesso em: 14 out. 2024.

VITOR G. SILVEIRA et al. **Detecção de perdas não técnicas via rede neural ARTMAP Fuzzy em sistemas de distribuição de energia elétrica**. Procedings do XXIV Congresso Brasileiro de Automática, 19 out. 2022.

WASKOM, M. Seaborn: Statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, 6 abr. 2021.

ZHANG, J.; MANI, I. **KNN approach to unbalanced data distributions: A case study involving information extraction**. Washington DC, United States of America: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Datasets, 21 ago. 2003.