**ANÁLISE DE ALGORITMOS DE ALGORITMOS DE CLASSIFICAÇÃO PARA IDENTIFICAÇÃO DE IRREGULARIDADES NO SISTEMA ELÉTRICO DE DISTRIBUIÇÃO**

JOÃO PEDRO LOBO MEDEIROS (IFPB, Campus Cajazeiras)

**E-mail:** joao.lobo@academico.ifpb.edu.br.

**Área de conhecimento:(Tabela CNPq)**: 3.04.00.00-7 Engenharia Elétrica.

**Palavras-Chave**: *Machine Learning*, Perda Não Técnica de Energia, *Random Forest* e *Multilayer Perceptron*

1. **Introdução**

As perdas de energia elétrica são um dos principais problemas das distribuidoras de energia tanto no Brasil como no mundo. Estas perdas podem ser divididas em dois tipos: a perda técnica e perda não técnica (PNT). A perda técnica tem origem em causas naturais, como o efeito Joule nos condutores e a magnetização dos transformadores. Já as PNTs ocorrem por causa de falhas no sistema de medição dos consumidores. As perdas não técnicas representam uma redução significativa no faturamento das distribuidoras. Segundo a Agência Nacional de Energia Elétrica (ANEEL), em 2019, as perdas não técnicas corresponderam a 6,44% da energia injetada no sistema elétrico do país, o que representa um custo de aproximadamente R$ 7,4 bilhões, o qual é parcialmente rateado para os consumidores regulares do sistema.

Em função do cenário descrito no parágrafo anterior, todas as distribuidoras de energia elétrica possuem departamentos específicos para o combate a esse tipo de irregularidade. A principal ação realizada para o combate as PNTs são inspeções ostensivas nas unidades consumidoras. Porém, existe um desafio para as distribuidoras nessa tarefa, o qual consiste na identificação prévia dos consumidores com PNT. Visto que cada inspeção possui um custo operacional para ser realizada, sempre que a inspeção é realizada e não há identificação de PNT, existirá um prejuízo para a concessionária.

A partir do desafio citado, é possível encontrar diferentes trabalhos na bibliografia correlata ao tema em que são aplicados diferentes algoritmos de classificação para a identificação de PNT, contudo, não há um concesso sobre qual a melhor técnica, como por exemplo, o estudo realizado por Figueroa et al. (2017), onde o algoritmo que obteve o melhor desempenho foi o MLP, com 65% na curva AUC, 13% no MCC e 14% na F1-*score*. Outro estudo que vale a pena ser citado é o realizado por Ghori et al. (2020), onde o algoritmo *Random Forest* obteve um valor de 0,982 na métrica F1-*score* e o MLP obteve um valor de 0,994 na métrica de sensibilidade. Neste contexto, o presente trabalho tem como objetivo analisar diferentes algoritmos de classificação quando aplicados na identificação de irregularidades no sistema elétrico de distribuição. A partir dos resultados do trabalho, será possível contribuir com as distribuidoras de energia para uma maior efetividade nas inspeções e em outras ações de combate às PNTs, resultando em um menor custo operacional e uma maior recuperação de energia por parte da distribuidora.

1. **Materiais e Métodos**

Os algoritmos utilizados no estudo foram: as Árvores de Decisão, o *Gradient Boosting,* o *MultiLayer Perceptron* (MLP) e o *Random Forest*. As árvores de decisão foram treinadas usando um número mínimo de 10 recordes por nó e com 200 números de tópicos. O *Gradient Boosting* foi treinado com 8 níveis de profundidade da floresta, 250 modelos e uma taxa de aprendizagem de 0,1. O MPL foi treinado com um número máximo de 200 interações, 5 camadas ocultas com 10 neurónios por camada e o *Random Forest*, foi treinado utilizando 100 árvores de decisão na execução do modelo. Estes parâmetros são valores típicos encontrados na literatura e adotados como referência no estudo realizado. A base de dados utilizada no estudo foi obtida em contato com uma distribuidora do setor elétrico do país, que pediu confidencialidade para preservar sua integridade. Para a implementação dos algoritmos de classificação, é utilizado a plataforma KNIME® *Analytics*. O KNIME® é uma solução para análise avançada de dados desenvolvida pela Universidade de Constança na Alemanha e distribuído gratuitamente na modalidade *Open Source*. A plataforma possui bibliotecas pré-instaladas com os principais algoritmos utilizados na área de ciência de dados, o que possibilita uma maior praticidade na implementação dos algoritmos na etapa de testes. Além disso, a ferramenta é do tipo *point-and-click*, ou seja, exime programação mínima, o que torna seu uso simples. Uma imagem da plataforma KNIME® é apresentada na Figura 1.



Fonte: KNIME® Analytics (2021)

Figura 1: KNIME *Analytics Platafform*

Após a aplicação dos algoritmos, os resultados são comparados a partir de métricas de desempenho pré-estabelecidas, as quais são derivadas da matriz de confusão. A matriz de confusão é uma representação útil do resultado da classificação de dados, pois, a partir dela é possível comparar os valores reais (de referência) com aqueles que foram indicados pelo algoritmo de classificação para cada estado previsto. As métricas de desempenho mais apropriadas para a avaliação um problema de classificação binária, como o caso em questão são: sensibilidade e precisão, as quais são definidas na sequência:

 Sensibilidade: É a porcentagem de irregularidades reconhecidas pelo algoritmo em comparação ao total de irregularidades existentes.

 Precisão: É a taxa de acerto dos casos identificados como irregularidade pelo algoritmo.

As métricas sensibilidade e precisão são inversamente proporcionais, por isso, o desafio de um bom classificador é obter valores elevados para ambas as métricas simultaneamente. Uma forma de se avaliar o resultado da classificação em termos das duas métricas simultaneamente é por meio da métrica F1-*score*, a qual é definhada como a média harmônica da -sensibilidade e da precisão, como expresso na Equação (1)

$F1-score =2×\frac{\left(sensibilidade ×precisão\right)}{sensibilidade+precisão}$ (1).

1. **Resultados e Discussão**

A comparação dos resultados obtidos para os dois algoritmos demonstra que o MLP apresentou o melhor desempenho de acordo com a métrica de sensibilidade, indicando que o método teve uma boa cobertura na detecção dos consumidores com perda não técnica existentes na base de dados. Adicionalmente, os valores de sensibilidade para os dois métodos são comparados no gráfico da Figura 2.



Fonte: Autoria própria

Figura 2 - Comparação da sensibilidade entre os algoritmos.

Como pode ser notado no gráfico da Figura 2, a sensibilidade do método *MultiLayer Perceptron* é superior à dos demais algoritmos. Contudo, quando a comparação dos resultados é realizada a partir da métrica precisão, pode-se observar que o melhor resultado foi obtido pelo método *Random Forest*, como apresentado no gráfico da Figura 3.

Fonte: Autoria própria

Figura 3 - Comparação da precisão entre os algoritmos.

Como pode ser observado na Figura 3, a precisão do método *Random Forest* é melhor que a dos demais algoritmos, o que demonstra que o método *Random Forest* apresentou uma quantidade menor de indicações falso-positivas.

A partir da análise apresentada para as métricas precisão e sensibilidade ainda não é possível determinar qual o método com melhor resultado geral, dado que para cada uma delas um método diferente foi superior. Assim, torna-se útil a análise a partir da métrica F1-*score*, que apresentado anteriormente consiste na média harmônica da precisão e da sensibilidade. O resultado do F1-*score* para os dois algoritmos é apresentado na Figura 4.

Fonte: Autoria própria

Figura 4 - Comparação da F1-*score* entre os algoritmos.

Como pode ser observado na Figura 4, o algoritmo MLP apresentou um valor de F1-*score* superior aos demais algoritmos. Dessa forma, é possível concluir que o algoritmo MLP apresentou o melhor resultado na identificação da perda não técnica na base de dados em questão, quando comparado ao algoritmo *Random Forest* que foi melhor na precisão.

1. **Considerações Finais**

Foram avaliados no trabalhado o desempenho dos algoritmos de classificação: árvores de decisão, *Gradient Boosting*, MLP e o *Random Forest* para a identificação da PNT em uma base de consumidores do sistema elétrico de distribuição. Como métricas de avaliação foram utilizadas a sensibilidade, a precisão e o F1-*score*.

A partir dos resultados alcançados é possível concluir que o algoritmo MLP apresentou o melhor desempenho, com um valor de F1-*score* de 0,397, o qual é 08,04% superior em relação ao resultado do *Random Forest* e 5% superior em relação a árvore de decisão, algoritmo que ficou em segundo lugar. Apesar disso os outros algoritmos apresentaram resultados satisfatórios, como o *Random Forest* que obteve uma precisão superior, o que demonstra que o método também pode ser utilizado em aplicação específicas em que se deseja priorizar a precisão da classificação. Dessa forma, conclui-se que os objetivos finais do projeto foram alcançados de forma satisfatória.

**Referências**

Figueroa, G., Chen, Y.-S., Avila, N., & Chu, C.-C. (2017). Improved practices in machine learning algorithms for NTL detection with imbalanced data. 1-5.

Ghori, K. M., Abbasi, R. A., Awais, M., Imran, M., Ullah, A., & Szathmary, L. (2020). Performance Analysis of Different Types of Machine Learning Classifiers for Non-Technical Loss Detection. 1-16