



Uso de machine learning e advanced analytics no combate às perdas não-técnicas

Tema: Recuperação de energia - Perdas não-técnicas

Autores: Matheus Dantas de Lucena; Maria Eusa Alves Pinto de Oliveira; Juliano Antonio Prado da Silva

Co-Autores: Luciano Dantas Pereira Junior; Danilo Febroni Baptista

Empresa: Energisa Paraíba - Distribuidora de Energia S/A

Resumo

Historicamente, as concessionárias de energia elétrica têm como um de seus maiores desafios a problemática das perdas tanto em termos de eficiência quanto de impacto econômico. No setor de distribuição de energia elétrica, as Perdas Não Técnicas (PNT) são causadas principalmente por fraudes e furtos, assim como por falhas na medição local e no faturamento. Para o caso de Unidades Consumidoras (UC) telemedidas, estão disponíveis as informações fasoriais de tensão, corrente, potência, ângulos entre as grandezas, entre outros. No Grupo Energisa S/A, por exemplo, existem em torno de 23 mil clientes telemedidos, os quais tem suas grandezas monitoradas com o intuito de detectar irregularidades e, consequentemente a necessidade de uma intervenção. Existe, atualmente, uma ferramenta de *Advanced Analytics* (AA) que performa uma sequência de regras, as quais detectam mediante filtros, UCs que possuem indícios de irregularidades. A solução aqui proposta é composta por uma ferramenta que utiliza *Machine Learning* (ML) com o objetivo de, ao utilizar um histórico de análises, construir um modelo que detecte padrões de irregularidades, reduzindo, assim, o tempo gasto por analistas no processo de análise sem pôr em risco a qualidade dos resultados. O modelo resultou em acurácias de 100% e 90,8%, para as grandezas tensão e corrente elétrica, respectivamente, indicando o potencial do método e do modelo em auxiliar as empresas do Grupo Energisa no processo de combate às PNT.

1. Introdução

A Agência Nacional de Energia Elétrica (ANEEL) publica anualmente o Relatório de Perdas de Energia Elétrica na Distribuição, com informações sobre perdas técnicas e não técnicas na distribuição. No último relatório, referente ao ano de 2023, as Perdas Totais (PTOT) representaram 14,1% (80,2 TWh) da energia injetada, sendo 7,4% referente às Perdas Técnicas (PTEC) (42,0 TWh) e 6,7% às PNT (38,2 TWh) (ANEEL, 2024). As perdas não técnicas regulatórias, que são reconhecidas nas tarifas, foram da ordem de 27,3 TWh. As concessionárias de grande porte, cujo mercado é maior do que 700 GWh, são responsáveis por quase a totalidade dos montantes das perdas não técnicas no Brasil devido ao tamanho do mercado e à maior complexidade de se combater as perdas. Os níveis de perdas não técnicas dependem da gestão das concessionárias, das características socioeconômicas e de aspectos comportamentais existentes em cada área de concessão (ANEEL, 2024).

A perda de receita devido aos procedimentos irregulares é um dos principais focos das distribuidoras devido aos prejuízos que ela acarreta. Para identificar uma fraude ou um furto, é necessário enviar uma equipe técnica ao local para que seja feita uma inspeção. Entretanto, cada deslocamento e fiscalização gera custos a empresa que podem chegar a ter até 8 milhões de consumidores (ANEEL, 2019). Dessa forma, essas companhias buscam aprimorar as técnicas de identificação de fraude a fim de otimizar as inspeções, recuperando mais energia para cada visita realizada às unidades consumidoras (PAULO, 2020).

As técnicas mais comuns utilizadas para detecção de perda comercial em soluções sem hardware envolvem métodos de classificação como *Support Vector Machine*, Perfil de Carga, Redes Neurais Artificiais, Árvores de Decisão e Técnicas de Aprendizado de Máquinas (ML). Esses classificadores são capazes de inferir um indicador binário ou uma probabilidade da presença de perdas a partir de um conjunto de entradas. O uso dessas técnicas geralmente consiste no processamento de dados de entrada, ajuste do modelo de classificação aos dados, avaliação do desempenho e implantação do modelo. Como fonte de dados, a maior parte dos artigos utilizou informações de consumo de energia, do perfil do consumidor, da carga, tensão e correntes medidas e dos resultados de inspeção. Grande parte também fez combinações das variáveis, utilizando o consumo junto com a informação do consumidor ou com os resultados de inspeção (PAULO, 2020).

Tendo como objetivo avaliar a assertividade de um modelo ML para detecção de irregularidades em clientes telemedidos, foi desenvolvido um fluxo na ferramenta de AA, utilizando um algoritmo *XGBoost* baseado em árvores de decisão com *Gradient Boosting* (aumento de gradiente). A partir desse modelo, foram avaliados 336 UC para informações de corrente e 91 para informações de tensão. No primeiro caso, a acurácia média foi de 91,67%, para o segundo caso foi de 100%. Dessa forma, a Energia passa a ter a possibilidade de tornar seu processo de inspeção mais célere e, conseqüentemente, conseguir detectar fraudes e/ou problemas técnicos que estão incorrendo em PNT, recuperando a energia correspondente.

2. Desenvolvimento

Com o objetivo de estruturar e desenvolver o modelo, as atividades e estrutura propostas foram conduzidas conforme a seguir. Esse processo está descrito na subseção de Metodologia, e, em seguida os resultados obtidos são apresentados.

2.1. Metodologia

Para avaliar a acurácia do modelo proposto, cuja arquitetura baseou-se em ML, coletou-se resultados de análises já realizadas, os quais foram utilizados para o treinamento e validação.

2.1.1. Machine Learning

Com o aumento na complexidade da solução de problemas reais a partir de um volume cada vez maior de dados, torna-se clara a necessidade de ferramentas computacionais capazes de resolver problemas de difícil visualização direta. O processo de induzir uma hipótese ou função a partir da experiência passada por meio de algoritmos para solução de problemas denomina-se *Machine Learning* (ML). As etapas usuais do desenvolvimento de modelo de ML, bem como as classificações dos tipos de algoritmos, podem ser observadas na Figura 1.

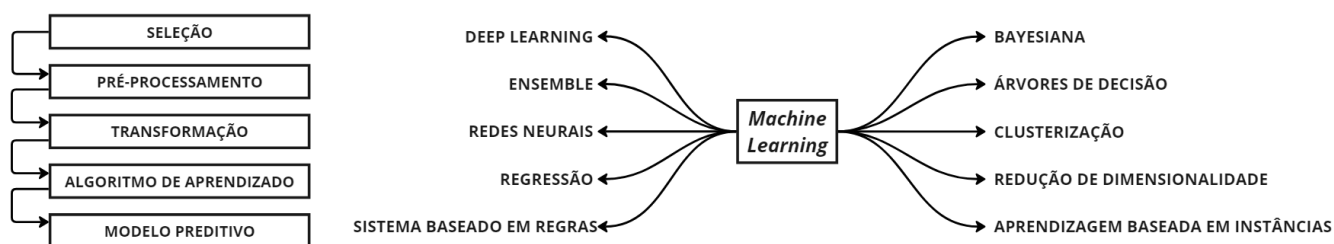


Figura 1 - Etapas de desenvolvimento de um modelo de Aprendizado de Máquina e os principais tipos de algoritmos.

No aprendizado de máquina, os algoritmos aprendem a partir de um princípio de inferência, denominado de treinamento, no qual se obtêm conclusões genéricas a partir de um subconjunto de dados. O modelo deve ser capaz de relacionar os valores dos atributos de entrada ao seu respectivo atributo de saída (alvo ou *target*), mesmo quando aplicado a novos dados nunca apresentados ao algoritmo. Essa propriedade de manter-se válido para novos objetos é conhecida por generalização de um modelo. Se o algoritmo estiver com baixa capacidade de generalização, diz-se que o modelo está superajustado aos dados de treinamento, ou em *overfitting*, e não será capaz de apresentar resultados consistentes para dados inéditos. No caso inverso, o modelo está subajustado, ou em *underfitting*, não sendo capaz de produzir uma alta taxa de acerto, mesmo no conjunto de treinamento, normalmente porque os exemplos disponíveis são pouco representativos ou o modelo utilizado não foi capaz de capturar os padrões existentes nos dados (FACELI, et al, 2011).

De acordo com a forma que se dá o sistema de aprendizado, pode-se classificar os algoritmos de ML em aprendizado supervisionado e aprendizado não supervisionado. O termo supervisionado é utilizado devido à presença de um supervisor externo, ou professor, que conhece a saída desejada para cada exemplo. Aplicações do aprendizado supervisionado incluem problemas de classificação, em que o objetivo é atribuir cada entrada a um número finito de categorias discretas, ou regressão, em que as saídas consistem em uma ou mais variáveis contínuas. Em um aprendizado não supervisionado, a figura do professor não existe e o atributo de saída não é diretamente utilizado. O objetivo é encontrar grupos semelhantes nos dados (clusterização), determinar sua distribuição (estimação de densidade) ou projetar os dados de alta dimensão em duas ou três dimensões (PAULO, 2020).

2.1.2. XGBoost

O *XGBoost*, ou *Extreme Gradient Boosting*, representa uma abordagem de ponta para ML, com desempenho de alta qualidade no combate a problemas de classificação e regressão. Ao alavancar um conjunto de árvores de decisão, o *XGBoost* constrói um modelo preditivo por meio de um processo iterativo que se concentra na minimização de erros. Esse refinamento iterativo, impulsionado pela otimização do gradiente descendente (a capacidade do modelo de encontrar a resposta com o menor erro rapidamente), permite que o *XGBoost* aprimore continuamente sua precisão preditiva, atualizando estrategicamente os parâmetros das árvores de decisão. Além disso, o *XGBoost* incorpora técnicas de regularização, ou mais simplesmente, técnicas para combater o *overfitting*, garantindo a capacidade do modelo de generalizar bem para dados não vistos. A combinação de aprendizado em conjunto, otimização de gradiente descendente e regularização torna o *XGBoost* uma ferramenta formidável para cientistas de dados e profissionais que buscam soluções altamente precisas e eficientes para desafios complexos de modelagem preditiva.

Em sua essência, o *XGBoost* emprega uma técnica chamada aumento de gradiente. É aqui que o algoritmo começa criando uma árvore de decisão simples e, em seguida, adiciona iterativamente mais árvores ao

modelo, cada uma com foco na correção dos erros cometidos pelas árvores anteriores. Esse processo continua até que um número predeterminado de árvores seja atingido ou o desempenho do modelo não melhore mais significativamente.

O *XGBoost* apresenta vários aprimoramentos importantes no algoritmo de aumento de gradiente padrão:

- Regularização: O *XGBoost* incorpora técnicas de regularização, como regularização L1 (*Lasso Regression*) e L2 (*Ridge Regression*) para evitar o *overfitting* e melhorar a generalização do modelo para ajudar a evitar que os modelos se tornem muito complexos. A regularização L1 simplifica o modelo removendo recursos menos importantes, enquanto a regularização L2 mantém o modelo estável equilibrando a influência de diferentes recursos.
- Poda de árvores: O *XGBoost* emprega uma técnica chamada "poda de árvores" para limitar a profundidade das árvores de decisão, evitando modelos excessivamente complexos e potencialmente em *overfitting*.
- Processamento paralelo: O *XGBoost* suporta processamento paralelo, permitindo tempos de treinamento mais rápidos em grandes conjuntos de dados, utilizando vários núcleos de CPU.
- Manipulação de valores ausentes: O *XGBoost* pode lidar automaticamente com valores ausentes no conjunto de dados, reduzindo a necessidade de pré-processamento extensivo de dados.

2.1.3. Avaliação de Modelos Preditivos

Para avaliar a performance de um modelo, existe uma variedade de métricas que podem ser adequadas para uma aplicação em particular. As métricas de erro mais comuns em problemas de classificação incluem a acurácia, a precisão e a sensibilidade. Essas medidas são extraídas da matriz de confusão que ilustra o número de predições corretas e incorretas em cada classe e pode ser montada conforme mostrado na Figura 2 para duas classes.

		PREVISTO	
		NEGATIVO	POSITIVO
REAL	NEGATIVO	VERDADEIRO NEGATIVO (VN)	FALSO POSITIVO (FP)
	POSITIVO	FALSO NEGATIVO (FN)	VERDADEIRO POSITIVO (VP)

Figura 2 - Matriz de Confusão

A acurácia (ac) pode ser definida conforme a equação (1) a seguir.

$$ac = \frac{VP + VN}{n}, \quad (1)$$

em que n é a quantidade de elementos da matriz de confusão.

Essa métrica também é denominada de taxa de acerto e determinará a proporção de exemplos corretamente classificados. Para aplicações em que o conjunto de dados é desbalanceado, ou seja, uma classe é bem menos representada que as demais, o desempenho de um algoritmo de ML não pode ser expresso em termos da acurácia (KUBAT e MATWIN, 1997). De fato, se o modelo ignorar a existência da classe menos representada, ainda assim é possível obter altos índices de acurácia. Para uma aplicação em que apenas 10% da população é da classe positiva, por exemplo, é possível obter uma taxa de acerto de 90% ao classificar todos os itens como negativos. Como alternativa, são utilizadas outras métricas para melhor representar a performance do modelo.

A efetividade (ef), ou precisão, é definida pela equação (2) e indica a proporção de positivos classificados corretamente dentre todos os classificados como positivos pelo modelo.

$$ef = \frac{VP}{VP + FP} \quad (2)$$

A cobertura, ou sensibilidade, é definida pela equação (3) e corresponde a taxa de acerto da classe positiva, ou seja, quantas unidades da classe positiva foram corretamente classificados dentre os que realmente eram positivos.

$$cob = \frac{VP}{VP + FN} \quad (3)$$

Para agregar as duas métricas, é possível utilizar a função *F-score*, que considera um mesmo peso de ponderação entre a precisão e a sensibilidade e é dada pela equação (4).

$$F - score = 2 \cdot \frac{ef \cdot cob}{ef + cob} = \frac{2 \cdot VP}{2 \cdot VP + FN + FP} \quad (4)$$

Outra função utilizada para agregar a efetividade e a cobertura é o *G-measure*, equação (5), que representa a média geométrica entre as duas métricas.

$$G - measure = \sqrt{ef \cdot cob} = VP \cdot \sqrt{\frac{1}{(VP + FP) \cdot (VP + FN)}} \quad (5)$$

2.1.4. Criação da Base de Dados

O objetivo principal dessa estruturação foi assegurar que o modelo tivesse acesso a dados consistentes e organizados, possibilitando uma análise eficaz e a detecção precisa das anomalias. As informações que compõem essa base de dados são descritas a seguir:

•

UC: Unidade Consumidora;

- Data & Hora: Data e horário de cada medição única;
- Classe: Classificação da medição, indicando se as tensões ou correntes apresentam alguma anomalia;
- Tensão nas Fases A, B e C: Valores de tensão (em volts) medidos nas fases A, B e C. A medição de tensão é crucial para identificar potenciais anomalias no sistema de distribuição elétrica. Desvios nos valores de tensão, como picos ou quedas abaixo dos níveis esperados, podem ser sinais de falhas técnicas, como sobrecargas, ou até mesmo indícios de manipulações fraudulentas na rede elétrica. Esses desvios afetam diretamente a classe da medição, pois valores anormais de tensão podem ser rotulados como anomalias (classe 1), enquanto medições dentro dos padrões normais são classificadas como normais (classe 0);
- Corrente nas Fases A, B e C: Valores de corrente (em ampères) medidos nas fases A, B e C. A corrente reflete o consumo de energia e desvios em seus valores podem sugerir sobrecarga, falhas nos componentes elétricos ou, ainda, intervenções fraudulentas. Essas variações na corrente são essenciais para classificar se uma medição é normal ou anômala, relacionando-se diretamente à classe (0 ou 1) da medição.

Nas Figuras Figura 3 e Figura 4 são apresentados exemplos de gráficos da grandeza tensão para duas UC diferentes, a primeira sem irregularidade e a segunda com irregularidade na tensão da fase A a partir do dia 09/12/2024. Já nas Figuras 5 e 6 são apresentados exemplos de gráficos da grandeza corrente para duas UC diferentes, a primeira sem irregularidade e a segunda com irregularidade na corrente da fase A a partir do dia 05/11/2024.

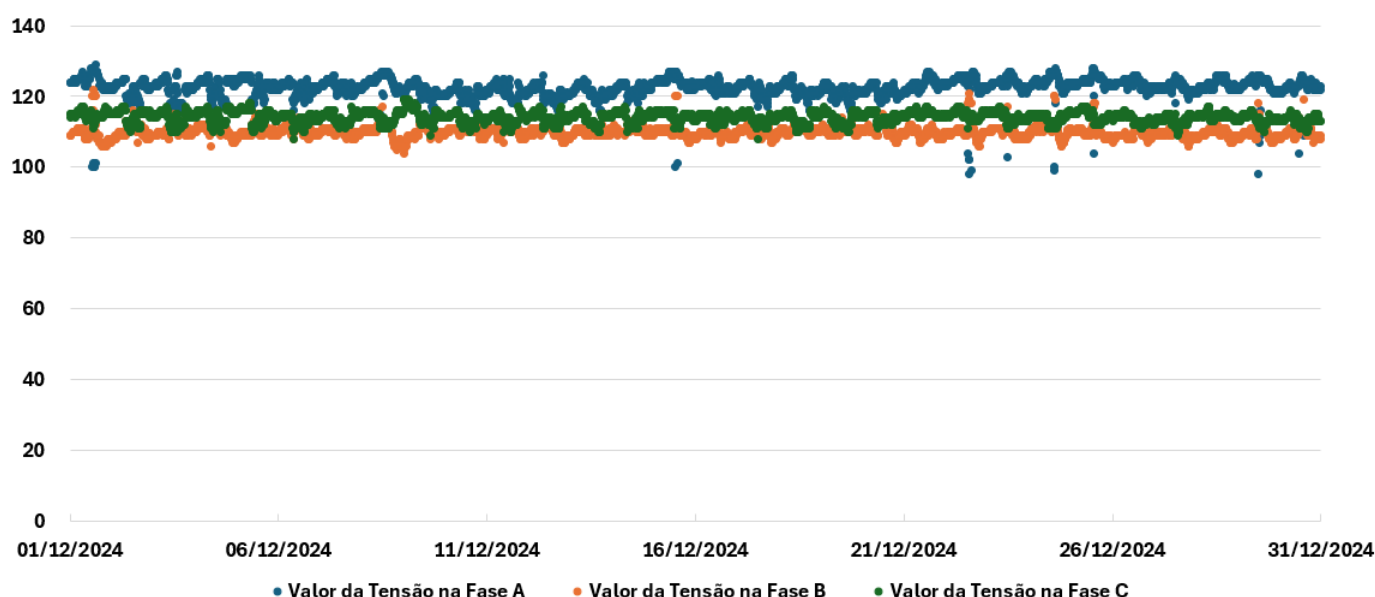


Figura 3 - Exemplo de Fasorial com Tensões "OK".

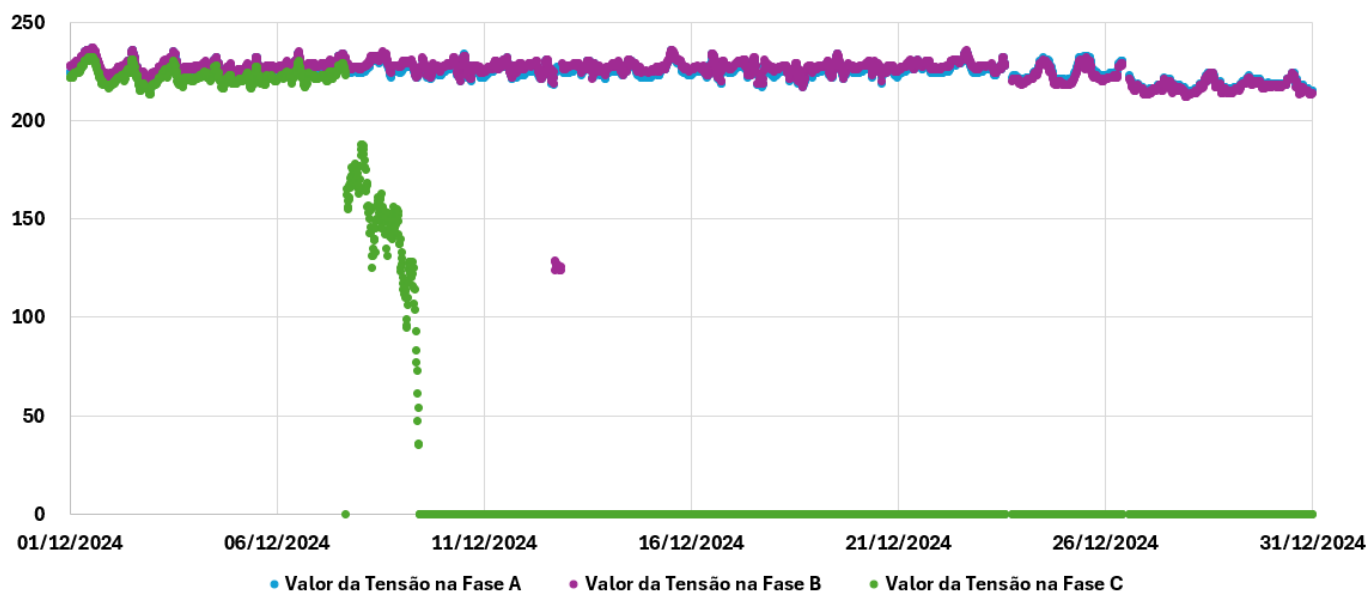


Figura 4 - Exemplo de Fasorial com Tensões "NOK".

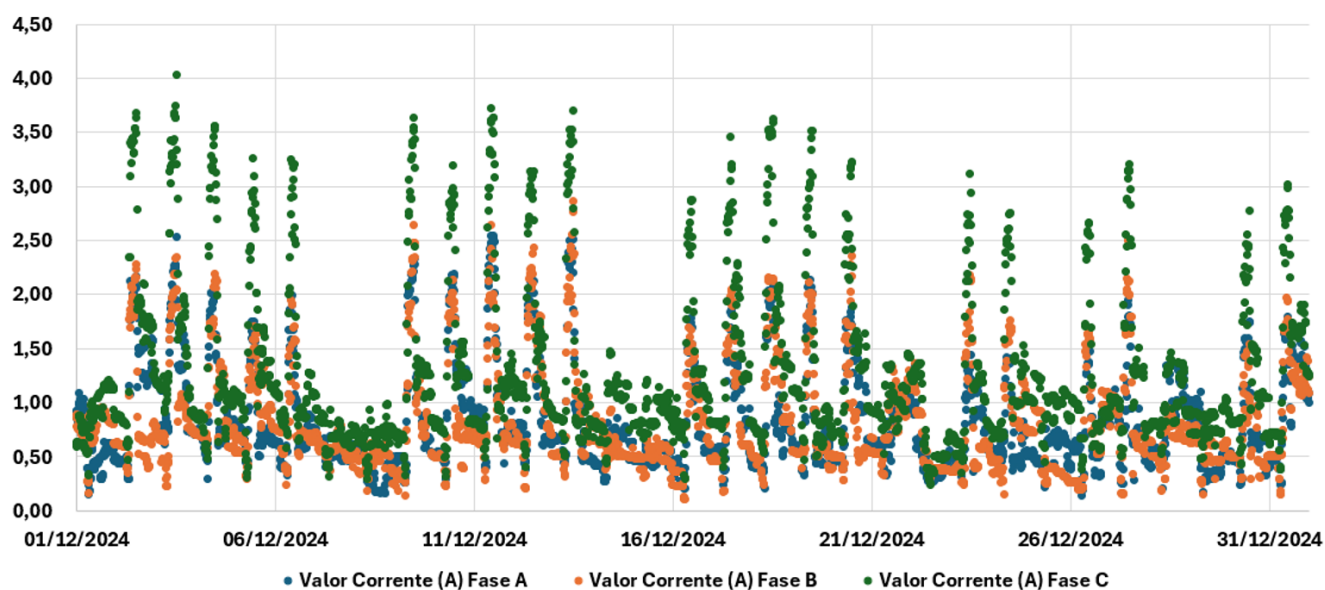


Figura 5 - Exemplo de Fasorial com Correntes "OK".

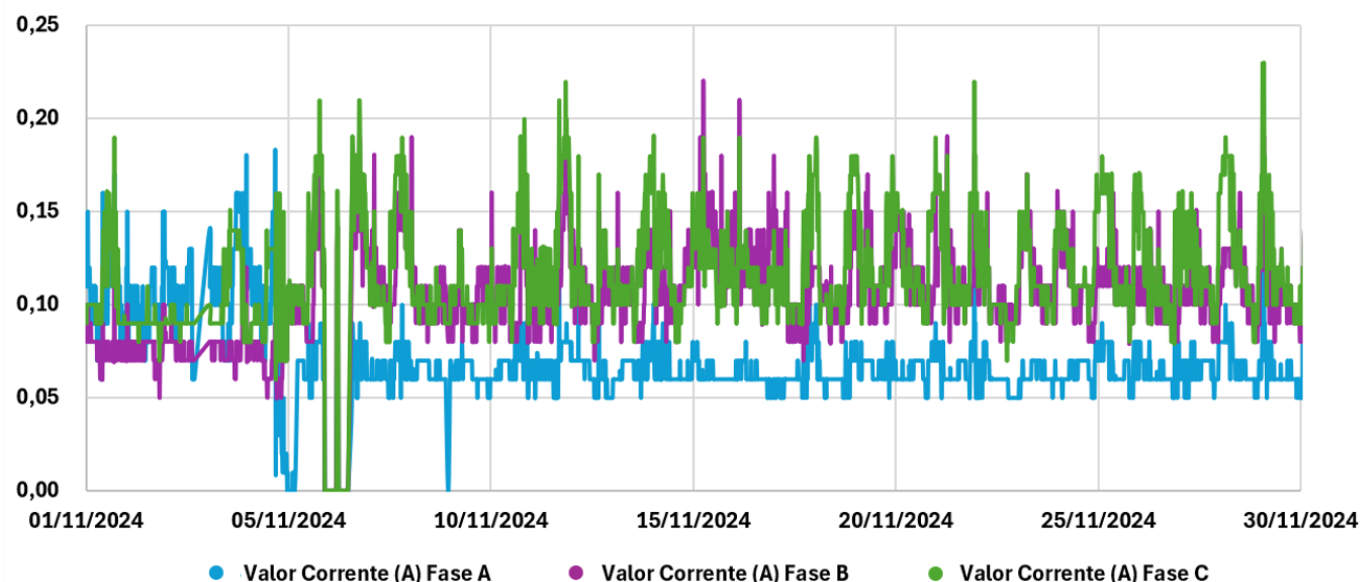


Figura 6 - Exemplo de Fasorial com Correntes "NOK".

2.2. Resultados

Os resultados obtidos foram segmentados em duas partes: (1) Modelo com dados de tensão e (2) Modelo com dados de corrente. Para os dois cenários, considerou-se um ponto de corte dos dados fasoriais como sendo de 15 dias após a data de identificação do problema.

Para ambos os modelos, as amostras foram particionadas de maneira aleatória, sendo 70% considerado para criação do modelo e 30% para validação. Além disso, as configurações apresentadas na Tabela 1, foram consideradas na construção do modelo.

Tabela 1 - Configurações do XGBooster.

Nome	Descrição	Valor
<i>Boosting rounds</i>	O número de modelos para treinar no conjunto <i>boosting</i>	100
Eta	Taxa de aprendizado	0,3
Lamba	Termo de regularização L2	1,0
Alpha	Termo de regularização L1	1,0
<i>Booster</i>	Seleção do <i>booster</i> padrão da árvore de decisão	<i>Tree</i>
Profundidade máxima	Profundidade máxima da árvore de decisão	6

2.2.1. Modelo com Dados de Tensão

Para o Modelo com Dados de Tensão, foram utilizados os dados fasoriais de 91 UC, destas 72 classificadas como "OK" (não apresentam irregularidade) e 19 como "NOK" (apresentaram alguma irregularidade). A matriz de confusão desse modelo é apresentada na Figura 7. Já os resultados completos, incluindo as métricas de avaliação do modelo preditivo são apresentados na Tabela 2.

		PREVISTO	
		NEGATIVO	POSITIVO
REAL	NEGATIVO	19 (VN)	0 (FP)
	POSITIVO	0 (FN)	72 (VP)

Figura 7 - Matriz de Confusão do Modelo de Tensão.

Tabela 2 - Resultados do Modelo de Tensão.

Regra	Quantidade	VP	FP	VN	FN	ac	Ef	cob	F-score	G-measure
Tensão	91	72	0	19	0	100%	100%	100%	100%	100%

Destaca-se, para este modelo, a qualidade dos resultados, mesmo com uma amostra de dados em volume inferior ao modelo de corrente. Isso se deve, principalmente, às características da grandeza em questão, que tendem a não variam muito com o perfil de carga do cliente.

2.2.2. Modelo com Dados de Corrente

Para o Modelo com Dados de Corrente, foram utilizados os dados fasoriais de 336 UC, destas 277 classificadas como “OK” (não apresentam irregularidade) e 59 como “NOK” (apresentaram alguma irregularidade). A matriz de confusão desse modelo é apresentada na Figura 8. Já os resultados completos, incluindo as métricas de avaliação do modelo preditivo são apresentados na Tabela 3.

		PREVISTO	
		NEGATIVO	POSITIVO
REAL	NEGATIVO	28 (VN)	31 (FP)
	POSITIVO	0 (FN)	277 (VP)

Figura 8 - Matriz de Confusão do Modelo de Corrente.

Tabela 3 - Resultados do Modelo de Corrente.

Regra	Quantidade	VP	FP	VN	FN	ac	ef	cob	F-score	G-measure
Corrente	336	277	31	28	0	90,8%	89,9%	100%	94,7%	94,8%

Destaca-se para este modelo que, apesar da grandeza avaliada (corrente) ser muito dependente do perfil de carga do cliente e ter uma elevada correlação com a sazonalidade, a qualidade dos resultados foi mantida, com uma acuraria de 90,8%.

3. Conclusão

O presente estudo focou na detecção e controle de perdas não técnicas em clientes teledados, de modo a agilizar o processo de análises das grandezas elétricas e acelerando a tomada de decisão. Por meio de uma ferramenta, na qual foi implementada um algoritmo *XGBoost*, foi possível segmentar, treinar e validar uma amostra de dados, de modo a criar um modelo que conseguisse emular uma análise realizada por um ser humano.

Foram obtidos resultados excelentes, de 100% e 90,8%, para as grandezas tensão e corrente elétrica, respectivamente, indicando o potencial do método e do modelo em auxiliar as empresas do Grupo Energisa no processo de combate às perdas não-técnicas.

Como etapas subsequentes, o método será melhorado de modo a incorporar outras informações, como por exemplo, as cadastrais, auxiliando na detecção de clientes bifásicos e com presença de geração distribuída, que podem incorrer em diagnósticos errôneos. Além disso, o modelo receberá mais amostras de dados para melhorar a acurácia em seu treinamento à medida que novas análises forem realizadas.

4. Referências bibliográficas

ANEEL. Relatórios de Consumo e Receita de Distribuição: Consumidores, Consumo, Receita e Tarifa Média – Região. In: ANEEL: Tarifas consumidores. Brasília, [2019]. Disponível em: <http://www.aneel.gov.br/relatorios-de-consumo-e-receita>. Acesso em: 01/02/2025.

ANEEL. Perdas de Energia Elétrica na Distribuição - 2024, de 18 de julho de 2024. Acesso em 02/02/2025. Disponível em: https://git.aneel.gov.br/publico/centralconteudo/-/raw/main/relatorioseindicadores/tarifaecnomico/Relatorio_Perdas_Energia.pdf

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina. Rio de Janeiro: LTC, 2011.

KUBAT, M.; MATWIN, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Fourteenth International Conference on Machine Learning, 1997, San Francisco. Proceedings [...]: Morgan Kaufmann Publishers, San Francisco, 1997.

PAULO, Fernanda Rodrigues. Detecção de fraude em unidades consumidoras não telemedidas com uso de técnicas de aprendizado de máquina. 2020.