Comparative analysis of machine learning algorithms for power factor prediction in smart grid applications

1st Graciélli D. L. Faccin
PPGEEC
UTFPR
Pato Branco, Brazil
graciellifaccin@alunos.utfpr.edu.br

2nd Elioenai M. F. Diniz
PPGEEC
UTFPR
Pato Branco, Brazil
elioenai@alunos.utfpr.edu.br

3rd Felipe R. Teixeira
DAELE
UTFPR
Pato Branco, Brazil
feliperoberto@alunos.utfpr.edu.br

4th Gustavo W. Denardin
PPGEEC
UTFPR
Pato Branco, Brazil
gustavo@utfpr.edu.br

5th Jean P. da Costa
PPGEEC
UTFPR
Pato Branco, Brazil
jpcosta@utfpr.edu.br

***Abstract*—The escalating complexity of electrical energy systems demands innovative approaches to power management and predictive analysis. This groundbreaking study introduces a novel machine learning methodology for power factor prediction within a university campus setting, leveraging a sophisticated data normalization technique to address asymmetries in data collection. By systematically comparing multiple machine learning algorithms—including Logistic Regression, Random Forests, Support Vector Machines (SVM), k-Nearest Neighbors, and Multi-Layer Perceptron Neural Networks—the research provides insights into predictive performance. Notably, the research also developed an innovative data standardization mechanism using arithmetic functions, which effectively mitigates data asymmetry challenges. These results not only advance our understanding of power factor dynamics but also offer a robust framework for energy management in complex electrical systems, particularly in regions with adversarial operational characteristics.**

***Index Terms*—Power Factor, comparative, machine learning, decision tree**

I. Introdução

O fator de potência (FP) é uma medida comum em ambientes com equipamentos que consomem eletricidade, como indústrias, fábricas, empresas comerciais e grandes instalações residenciais que utilizam motores, transformadores e outros aparelhos de alta potência. Ele é a razão entre a potência ativa (útil) e a potência aparente (total) de um sistema elétrico, indicando a eficiência no uso da energia. Valores ideais de FP variam entre 0,92 e 1, representando menor desperdício de energia reativa [14].

Enfrentamos um problema na central energética da universidade relacionado a penalidades por variação excessiva do FP. O FP precisa manter-se estável e dentro do intervalo específico. Quando ele se desvia desse intervalo, especialmente ao ficar abaixo de 0,92, ocorrem perdas de eficiência energética. Isso não apenas compromete a estabilidade do sistema elétrico, como também gera sanções financeiras pela concessionária. Este estudo tem como objetivo prever o FP em um bloco universitário, utilizando técnicas de aprendizado de máquina.

Alguns estudos utilizam diferentes modelos de aprendizado de máquina, com os mesmos demostrando eficácia para a tarefa de analise preditiva para esse contexto em sistemas elétricos. Esses estudos vão desde previsões de carga, otimização de sistemas fotovoltaicos (PV) e predição de propriedades de materiais. Além disso, o método SHAP [13], é empregado para analisar as contribuições das variáveis preditoras.

Este estudo preenche lacunas importantes no campo de pesquisa sobre previsão do FP em sistemas elétricos. Um diferencial deste trabalho em relação a outros trabalhos anteriores relacionados, é o fato de em nossa pesquisa se utilizar dados obtidos no hemisfério sul, oferecendo uma perspectiva única e pouco explorada na literatura. Além disso, realizamos uma validação abrangente de diferentes modelos de aprendizado de máquina, contribuindo para uma compreensão mais profunda da eficácia de várias técnicas neste contexto. Um aspecto inovador do nosso trabalho é a implementação de um mecanismo de resumo baseado em funções aritméticas com agrupamentos, que permite uma análise mais refinada dos padrões de consumo energético. Estas contribuições visam expandir o conhecimento no campo e fornecer bases sólidas para futuras pesquisas na área de previsão e gestão de sistemas elétricos complexos.

Com isso, é proposto uma abordagem para reduzir na medida que se padroniza as dimensões com o uso de funções

CAPES (Finance Code 001), and FINEP

aritméticas básicas. Além disso foi feito um estudo sobre diferentes modelos de ML com o uso de validação cruzada, visando determinar qual dos modelos possui maior capacidade para criar aderência de predição ao fenômeno analisado com o uso desses dados. Em conjunto a isso foi realizado um corte temporal e aplicado numa arvore de decisão, para propiciar entendimento sobre o processo de tomada de decisão na predição feita.

O objetivo deste estudo é fazer uma analise comparativa de diferentes algoritmos de ML com dados de entrada representadas por meio de um mecanismo proposto que reduz ao mesmo que padroniza as dimensões para evitar problemas de assimetria na quantidade de dados coletados. A abordagem metodológica inclui a obtenção dos dados para esses experimentos, coletados dentre os dias 1 à 7 de agosto de 2024, com intervalos de 15 minutos e o treinamento do modelo foi estruturado com uma separação temporal dos dados, onde $\frac{2}{3}$ dos dados iniciais foram utilizados para treinar o modelo, e o $\frac{1}{3}$ restante foi reservado para testar sua capacidade de estimar a classe para os próximos momentos. Todos os dados são medições reais de um prédio em um campus universitário localizado dentro do espaço correspondente ao fuso GMT-3. Os testes feitos para validar o uso dos modelos foi feito com métricas, além de um modelo de arvore de decisão treinado para emular explicabilidade para entender o processo de estimação com uso de poucos atributos preditores.

II. Trabalhos Relacionados

É apresentado uma exploração da aplicação de máquinas de vetores de suporte (SVM) na previsão de carga elétrica de curto prazo, incorporando fatores meteorológicos para melhorar a precisão das predições [1]. O estudo utilizou dados históricos de carga e variáveis meteorológicas como temperatura e umidade relativa do ar para treinar o modelo SVM. Os autores compararam o desempenho do SVM com métodos tradicionais de regressão linear múltipla, demonstrando que o SVM alcançou maior precisão na previsão de carga. Este trabalho exemplifica o potencial do aprendizado de máquina na predição de grandezas elétricas, destacando a capacidade de lidar com relações não-lineares complexas entre múltiplas variáveis de entrada e a grandeza elétrica a ser prevista.

A investigação sobre a otimização do posicionamento, dimensionamento e FP operacional de PV em redes de distribuição [2]. Utilizando índices de estabilidade de tensão, os autores determinaram a localização ideal para a instalação de PV e otimizaram seu tamanho para minimizar perdas do sistema. Além disso, o estudo explorou a otimização do FP do PV, demonstrando que ajustar o FP entre 0,85 atrasado e 1,0 (unitário) pode levar a reduções significativas nas perdas do sistema e melhorias no perfil de tensão. Para o sistema IEEE 33 barras, por exemplo, a otimização do FP para 0,88 resultou em uma redução adicional de 20,2% nas perdas do sistema em comparação com a operação em FP unitário. Esses resultados destacam a importância de considerar o FP na operação de sistemas com alta penetração de PV.

Foi desenvolvido um modelo de aprendizado de máquina interpretável para prever o FP de compostos termoelétricos do tipo diamante. Utilizando uma técnica de "stacking" para combinar múltiplos modelos e incorporando descritores tanto elementares quanto estruturais, eles alcançaram um coeficiente de determinação (R^2) superior a 0,95 no conjunto de teste [3]. O estudo empregou o método SHAP (SHapley Additive exPlanations) [13] para interpretar os resultados, revelando correlações negativas entre o FP e a eletronegatividade dos ânions, e positivas com o volume por átomo. Este trabalho demonstra a viabilidade de combinar alta precisão preditiva com interpretabilidade em modelos de aprendizado de máquina para materiais termoelétricos, permitindo extrair insights físicos relevantes dos resultados.

No trabalho é apresentado um modelo de predição do FP em sistemas elétricos trifásicos, utilizando técnicas de aprendizado de máquina. [4] O estudo se concentra na aplicação da regressão linear para prever o FP em instalações de média tensão. A pesquisa propõe um modelo que analisa dados, monitorados durante o período de 7 dias, com intervalos de 5 minutos, de variáveis elétricas como tensão, corrente e potência ativa, para estimar o FP futuro. O objetivo é desenvolver uma abordagem que permita a compensação reativa antecipada, sem a necessidade de monitoramento contínuo. Os resultados demonstram que o modelo proposto pode prever o FP com alta precisão, mesmo em cenários onde há a presença de fontes de energia renovável, contribuindo para uma gestão energética mais eficiente e econômica.

III. Metodologia

Os dados foram coletados por meio de sensores distribuídos estrategicamente pelo campus da universidade. Esses dados, juntamente com a infraestrutura de coleta e armazenamento, são gerenciados em um Sistema de Gerenciamento de Banco de Dados (SGBD) Postgres. Devido ao volume e à complexidade dos dados, foi implementada a particionamento das tabelas, garantindo uma melhor performance e escalabilidade no gerenciamento da informação.

A. Abordagem

Devido às assimetrias observadas na quantidade de dados amostrados dentro de um mesmo intervalo de tempo, ou seja, para diferentes momentos amostrados com o mesmo intervalo, podem ocorrer variações na cardinalidade dos dados.

Na figura 1 existe um esquemático para o processo de padronização, que utiliza formas matemáticas junto de um processo de agrupamento, para padronizar as dimensões dos dados.

Os dados são originalmente coletados em intervalos de uma hora. No entanto, para atender aos requisitos específicos do projeto, foi necessário replicar os dados coletados em uma determinada hora para intervalos de 15 minutos. Essa abordagem permitiu a obtenção dos dados necessários com a granularidade exigida para a realização dos experimentos, mantendo a consistência e a integridade dos resultados.

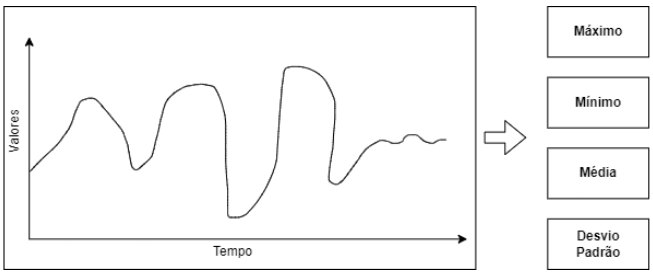

Fig. 1. Representação do processo de padronização dos dados

B. Experimentos

Os experimentos decorreram no ambiente Google Colab, utilizando a linguagem Python na sua versão 3.10, com o uso das bibliotecas Meteostat, Sklearn, Numpy, Pandas, e Matplotlib.

É realizada uma análise comparativa entre diferentes algoritmos de Machine Learning, com a finalidade de entender se algum, e qual entre eles. Possui a maior capacidade de aderência aos dados do fenômeno que está sendo mapeado. No caso, a média para o Fator de Potência na próxima janela de tempo. Para isso, foram utilizados alguns métodos que diferem entre si, na forma como geram suas estimativas. Todos eles foram aplicados sem nenhum ajuste de parâmetro, sendo aplicados com seus parâmetros padrões dispostos na biblioteca. Os métodos foram:

- Regressão Logística: A regressão logística é um método estatístico usado para modelar a probabilidade de ocorrência de um evento binário, em função de uma ou mais variáveis independentes [6],
- Floresta Aleatória: A Floresta Aleatória é um algoritmo de aprendizado de máquina que combina múltiplas árvores de decisão para melhorar a precisão da classificação ou regressão. Isso funciona ao criar várias árvores com diferentes subconjuntos dos dados e então combinar suas previsões [7].
- Máquina de Vetores de Suporte: SVM são algoritmos de aprendizado de máquina usados para classificação e regressão, que funcionam encontrando o hiperplano que melhor separa as classes no espaço de características, maximizando a margem entre as classes [8].
- K-ésimo Vizinho mais Próximo: É um método de classificação que atribui uma classe a um ponto de dados com base nas classes dos K pontos mais próximos no espaço de características, utilizando uma métrica de distância, geralmente a Euclidiana [9].
- Rede Neural Multi Camadas Perceptron: É uma arquitetura de rede neural composta por camadas de nós (neurônios) organizadas em uma entrada, uma ou mais camadas ocultas e uma camada de saída. É amplamente utilizada em tarefas de classificação e regressão, aprendendo representações complexas por meio do algoritmo de retropropagação com o uso do algoritmo de gradiente descendente [10].

O experimento envolveu a aplicação de validação cruzada para treinar e avaliar diferentes modelos de machine learning. Para cada fold, os dados foram divididos em conjuntos de treino e teste, seguidos de um processo de normalização. Em seguida, os modelos foram treinados e suas previsões avaliadas, permitindo o cálculo de métricas de desempenho como acurácia, precisão, revocação e kappa, além dos tempos de execução tanto para o treinamento quanto para a predição. As médias e os desvios padrão dessas métricas foram calculados para comparar a eficácia dos modelos. A validação cruzada foi realizada em dez divisões aleatórias não estratificadas, garantindo uma análise robusta das métricas de desempenho e tempos de execução para cada modelo avaliado.

A figura 4 ilustra o uso de uma única árvore de decisão, ao invés de uma floresta de árvores, para facilitar a construção de um mapa de decisão claro e interpretável. Essa abordagem permite visualizar diretamente as regras de classificação, proporcionando uma interpretação simples e direta das decisões tomadas pelo modelo.

C. Resultados

O FP representa a relação entre a potência ativa e a potência aparente em um circuito elétrico, sendo predominantemente influenciado pela reatância do sistema. Na imagem 2 observado, que simplesmente nenhum fator meteorológico teve correlação forte o suficiente para suportar relação de causalidade, assim nenhum teve um impacto tão significativo. [11].

Com exceção da irradiância em suas três formas, que apresenta um coeficiente positivo considerável em relação às medidas do fator de potência, especialmente em relação ao valor máximo.

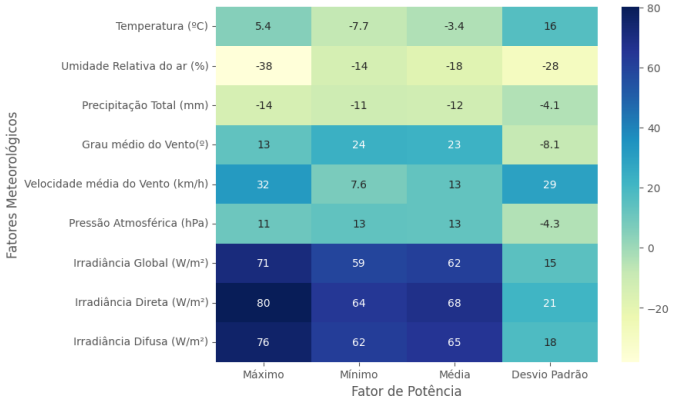

Fig. 2. Correlação de Pearson para Grandezas Meteorológicas em relação com medidas do Fator de Potência

Na imagem 3 é apresentado o Coeficiente de Variação para o FP. Como em alguns instantes alguns valores estavam tendo valores com módulo enormes, foi aplicada uma atenuação em 10.000%, independentemente do sinal. Observa-se que, embora geralmente próximo de zero, o coeficiente exibe picos extremos, tanto positivos quanto negativos. No entanto os valores negativos são mais proeminentes, indicando que o FP se torna mais instável em condições negativas.

TABELA I
MÉTRICAS DE DESEMPENHO DENTRO DO CONTEXTO DE VALIDAÇÃO CRUZADA

Métricas	Random Forest	Logistic Regression	SVM	k-NN	MLP
Acurácia (%)	94.04 ± 3.20	93.74 ± 3.18	**94.78** ± 2.77	93.89 ± **2.25**	94.34 ± 3.19
Precisão (%)	93.59 ± 3.89	93.32 ± 3.63	**94.50** ± 3.24	93.86 ± **2.14**	94.25 ± 3.24
Revocação (%)	92.64 ± 4.25	92.44 ± 3.98	**93.46** ± 3.52	92.34 ± **2.83**	93.10 ± 3.39
Kappa (%)	86.10 ± 8.06	85.62 ± 7.51	**87.80** ± 6.43	85.89 ± **4.67**	87.15 ± 6.32
Perda Logarítmica	2.15 ± 1.10	2.26 ± 1.15	**1.88 ± 1.00**	2.20 ± 0.81	1.93 ± 1.10
Tempo de Treinamento (ms)	1015.01 ± 300.65	71.63 ± 11.25	26.45 ± 10.63	**1.43 ± 1.53**	759.89 ± 568.73
Tempo de Predição (ms)	17.51 ± 9.38	**0.36 ± 0.04**	4.67 ± 2.87	2.17 ± 0.88	0.69 ± 0.37

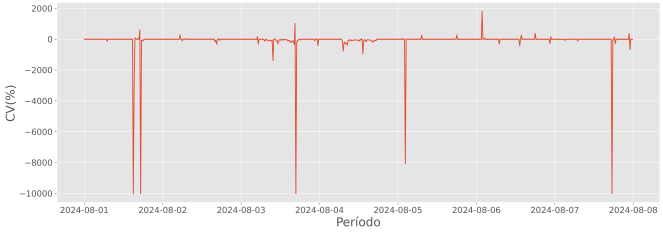

Fig. 3. Coeficiente de Variação para o Fator de Potência

As métricas de desempenho, como acurácia, revocação, precisão, coeficiente kappa de Cohen, e perda logarítmica foram calculadas para os conjuntos empregados no treinamento e teste. Na tabela I, pode ser vista uma relação para esses valores. Como pode ser observado os melhores valores são atingidos pela modelo de SVM [8], o que de certa forma é esperado devido ao fato de se lidar com uma alta dimensionalidade e uma cardinalidade relativamente baixa, cenários onde o SVM performa bem [12].

Ao analisar os tempos de treinamento, salta aos olhos a discrepância do modelo K-NN frente aos outros, isso na verdade é esperado e esconde uma armadilha, pois esse modelo não tem parâmetros aprendíveis ele apenas armazena as instancias dadas em treinamento para então realizar os cálculos de proximidade na etapa de inferência. Portanto o melhor tempo de tempo é na verdade o da Regressão Logística, já que a mesma apenas aplica uma multiplicação de coeficientes e aplicação sobre a função sigmóide.

Realizou-se o desenvolvimento e avaliação de um modelo de árvore de decisão para classificar dados em duas classes: ”Negativo” e ”Adequado”. Para o treinamento deste modelo foi utilizado uma estrutura com uma separação temporal dos dados, onde $\frac{2}{3}$ dos dados iniciais foram utilizados para treinar o modelo, e o $\frac{1}{3}$ restante foi reservado para testar sua capacidade de estimar a classe para os próximos momentos. Com isso na figura 5 é apresentada a matriz de confusão para as predições deste modelo.

A estrutura da árvore de decisão gerada é visualizada na imagem, onde cada nó representa uma decisão baseada em um valor específico de uma característica do conjunto de dados. A coloração da árvore indica as classes predominantes em cada nó, com tons de laranja representando a classe ”Negativo” e tons de azul a classe ”Adequado”. A métrica gini, presente em cada nó, indica a pureza da divisão entre as classes, com valores mais próximos de zero sugerindo maior homogeneidade dentro do nó.

IV. CONCLUSÃO

Este estudo demonstrou a eficácia das técnicas de aprendizado de máquina na previsão do FP. A SVM se destacou como o método mais preciso, alcançando uma acurácia média de 94,78% com um menor tempo de treinamento. Sendo uma descoberta significativa para a otimização da gestão energética, permitindo ajustes antecipados nos sistemas elétricos. A análise de árvore de decisão também forneceu uma percepção valiosa sobre as variáveis preditivas mais relevantes.

Apesar dos resultados inicialmente promissores, ainda existem trabalhos a serem feitos para sanar potenciais problemas na abordagem como a replicação dos experimentos com dados advindos de outras épocas do ano, com um conjunto maior. Para trabalhos futuros, o aprimoramento dessas tarefas, e a incorporação de variáveis adicionais que são endógenas ao problema, como a informação da quantidade de pessoas utilizando a rede.

Em conclusão, este estudo representa um avanço significativo na aplicação de técnicas de aprendizado de máquina para a previsão do FP. Os resultados obtidos demonstram a viabilidade e o caminho para futuras pesquisas que podem refinar e expandir essas técnicas. A capacidade de prever o FP tem implicações importantes para a eficiência energética, a estabilidade da rede e a integração de fontes de energia renovável, contribuindo para o desenvolvimento de redes elétricas mais inteligentes, eficientes e sustentáveis.

AGRADECIMENTOS

Agradecemos sinceramente à CAPES (Código Financeiro 001), e FINEP pelo apoio financeiro.

REFERÊNCIAS

[1] J. Liu, Z. Xu, W. Fan, Y. Wang, and W. Mo, ”Application of SVM Method with Meteorological Factors in Power Load Forecasting,” in Proc. 2023 3rd Int. Conf. Energy Eng. Power Syst. (EEPS), 2023, pp. 339-342, doi: 10.1109/EEPS58791.2023.10257017.

[2] M. A. Rasheed, R. Verayiah, and B. Saleh, ”Optimal Placement, Sizing and Operating Power Factor of PV for Loss Minimization and Voltage Improvement in Distribution Network via DigSilent,” in Proc. 2020 2nd Int. Conf. Smart Power Internet Energy Syst. (SPIES), 2020, pp. 126-131, doi: 10.1109/SPIES48661.2020.9243100.

[3] Z. Yang, Y. Sheng, C. Zhu, J. Ni, Z. Zhu, J. Xi, W. Zhang, and J. Yang, ”Accurate and explainable machine learning for the power factors of diamond-like thermoelectric materials,” J. Materiomics, vol. 8, no. 3, pp. 633-639, 2022, doi: 10.1016/j.jmat.2021.11.010.

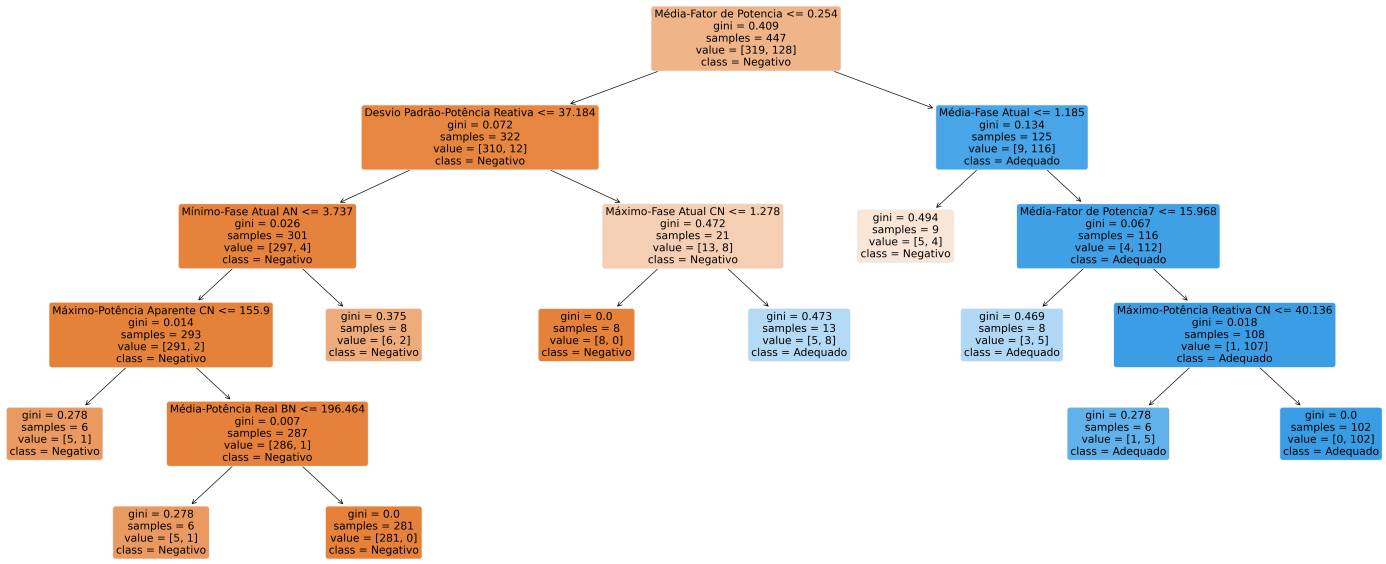

Fig. 4. Processo de classificação feito pela Árvore de Decisão

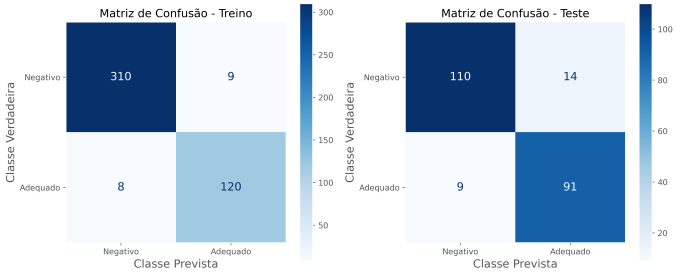

Fig. 5. Matriz de Confusão para as predições

[4] J. M. Gámez Medina, J. de la Torre y Ramos, F. E. López Monteagudo, L. C. Ríos Rodríguez, D. Esparza, J. M. Rivas, L. Ruvalcaba Arredondo, and A. A. Romero Moyano, "Power Factor Prediction in Three Phase Electrical Power Systems Using Machine Learning," Sustainability, vol. 14, no. 15, Art. no. 9113, 2022, doi: 10.3390/su14159113.

[5] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artificial Intelligence - Volume 2*, San Francisco, CA, USA, 1995, pp. 1137-1143.

[6] J. Tolles and W. J. Meurer, "Logistic Regression: Relating Patient Characteristics to Outcomes," JAMA, vol. 316, no. 5, pp. 533-534, Aug. 2016, doi: 10.1001/jama.2016.7653.

[7] T. K. Ho, "Random decision forests," in Proc. 3rd Int. Conf. Document Anal. Recognit., vol. 1, IEEE Comput. Soc. Press, 1995, pp. 278-282, doi: 10.1109/ICDAR.1995.598994.

[8] C. Cortes and V. Vapnik, "Support-Vector Networks," Mach. Learn., vol. 20, no. 3, pp. 273-297, Sep. 1995, doi: 10.1007/BF00994018.

[9] T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21-27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[10] D. E. Rumelhart et al., "Learning Representations by Back-Propagating Errors," Nature, vol. 323, no. 6088, pp. 533-536, Oct. 1986, doi: 10.1038/323533a0.

[11] V. dos R. Alves et al., "Análise da Influência do Fator de Potência no Cálculo de Perdas Técnicas em Redes de Distribuição," in Anais do Simpósio Brasileiro de Sistemas Elétricos 2020, sbabra, 2020, doi: 10.48011/sbse.v1i1.2393.

[12] R. Hafiz and S. Saeed, "Hybrid whale algorithm with evolutionary strategies and filtering for high-dimensional optimization: application to microarray cancer data", Plos One, vol. 19, no. 3, p. e0295643, 2024. https://doi.org/10.1371/journal.pone.0295643

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765-4774.

[14] COPEL, "Fator de potência - Copel - Pura Energia." [Online]. Available: https://web.archive.org/web/20240505153843/https://www.copel.com/site/copel-distribuicao/para-sua-empresa/fator-de-potencia/. Accessed: May 5, 2024.