MODELAGEM PREDITIVA E INTELIGÊNCIA ARTIFICIAL PARA ANÁLISE E PREVISÃO DE DADOS DE ENERGIA ELÉTRICA

Lincon Vinicius Mendes Pereira (PIBIC/CNPq/FA/UEM). E-mail: ra132857@uem.br.
Luiz Fernando Cótica (Orientador). E-mail: lfcotica@uem.br.

Universidade Estadual de Maringá, Centro de Ciências Exatas, Maringá, PR.

Área e subárea do conhecimento: Ciência da Computação / Metodologia e Técnicas da Computação.

Palavras-chave: Séries temporais; aprendizado de máquina; vazamento de dados.

RESUMO

Dando continuidade ao projeto de 2025, que construiu uma base consolidada de dados públicos do setor elétrico brasileiro, este trabalho aplica e compara quatro técnicas de previsão de séries temporais com os modelos SARIMA, Random Forest, XGBoost e Regressão Linear múltipla à carga média diária do Sistema Interligado Nacional e à potência instalada no país, em registros diários de 1999 a 2025. A ordem cronológica foi respeitada na divisão treino/teste e na busca de hiperparâmetros, e todos os modelos foram comparados a referências ingênuas. Na carga, identificou-se vazamento por identidade do balanço energético, pois a soma da geração por fonte correlaciona 0,9992 com o alvo; após defasar as exógenas em um dia, a Regressão Linear obteve RMSE de 2.389 MW e R^2 de 0,882, 40,1% abaixo do erro da melhor referência. Na potência instalada, a variação diária tem autocorrelação de primeira ordem de −0,5025, o valor esperado sob ruído independente, que impõe um piso de erro de 5.424 MW; na escala mensal, o SARIMA com deriva atingiu RMSE de 1.645 MW e R^2 de 0,705. A formulação do problema e a verificação de vazamento pesam, portanto, mais que a sofisticação do algoritmo. Os modelos foram integrados ao Atlas da Energia do NAPI EZC, plataforma em Streamlit criada na etapa anterior.

INTRODUÇÃO

O planejamento do Sistema Interligado Nacional (SIN) depende de duas informações complementares: a demanda de energia elétrica a atender a cada instante, uma grandeza operacional, publicada diariamente pelo Operador Nacional do Sistema Elétrico (ONS), com forte variação por dia da semana, feriado e estação, e a capacidade instalada de geração disponível para atendê-la, grandeza estrutural de evolução muito mais lenta. Prever ambas é essencial para a programação do despacho e para as políticas de expansão do setor.

Este trabalho dá continuidade ao projeto de 2025, no qual dados públicos da ANEEL e do ONS foram coletados por webscraping, padronizados e armazenados em um banco NoSQL que deu origem ao Atlas da Energia do NAPI EZC, plataforma interativa em Streamlit. Consolidada a base, o objetivo agora é aplicar e comparar técnicas de modelagem preditiva e de aprendizado de máquina — SARIMA, Random Forest, XGBoost e Regressão Linear múltipla — à carga elétrica diária do SIN e à potência instalada, comparando todo modelo a referências ingênuas, auditando os atributos em busca de vazamento e publicando os resultados na própria plataforma.

MATERIAIS E MÉTODOS

A base, herdada da etapa anterior, reúne registros diários do sistema elétrico brasileiro de 1999 a 2025 (9.675 dias) em 44 colunas: carga por subsistema, geração por fonte, grandezas hidrológicas, energia armazenada e potência instalada, entre outras. Definiram-se dois alvos, ambos a partir de 01/01/2000: a carga média diária do subsistema brasileiro, em MWmédio, até 24/06/2025, e a potência instalada em operação, em MW, até 01/01/2024. Quatro dias de carga com valor zero, fisicamente impossíveis, foram tratados como ausentes e interpolados.

A engenharia de atributos seguiu uma regra única: para prever o dia D, só entra informação disponível até D−1. Cada alvo recebeu 23 atributos: defasagens da própria série, médias móveis sobre a série já defasada, variáveis de calendário e exógenas defasadas em um dia, e cada coluna candidata foi antes auditada quanto ao risco de ser uma decomposição do próprio alvo.

Em ambos os problemas, os últimos 365 dias foram reservados para teste e o restante para treino, sem embaralhamento, seguindo as práticas de validação temporal de Hyndman e Athanasopoulos (2021). Os hiperparâmetros do Random Forest (BREIMAN, 2001) e do XGBoost (CHEN; GUESTRIN, 2016) saíram de busca aleatória com TimeSeriesSplit de três dobras; a ordem do SARIMA veio do AIC na escala diária e do erro fora da amostra na mensal. O desempenho foi medido por MAE, RMSE, MAPE e R^2 e comparado às referências ingênuas persistência (D−1),

sazonal (D−7), médias móveis e tendência linear recente. A implementação usou Python 3.11.4 (pandas, statsmodels, scikit-learn, xgboost e holidays), com sementes fixadas em 42; os artefatos de treino alimentam as páginas de previsão do Atlas da Energia, pelo mesmo MongoDB construído em 2025.

RESULTADOS E DISCUSSÃO

A primeira versão da análise da carga usava as exógenas do mesmo dia do alvo e alcançava R^2 próximo de 0,97 no teste. A revisão revelou a causa: por identidade do balanço energético, a soma da geração por fonte e a carga são a mesma grandeza (correlação de 0,9992), e o modelo apenas recuperava a carga de hoje a partir de uma decomposição dela mesma. A correção não foi descartar as colunas, mas sim defasar todas as exógenas numéricas em um dia.

Tabela 1 – Desempenho na previsão da carga elétrica diária no conjunto de teste (25/06/2024 a 24/06/2025), após a correção do vazamento.

Modelo	RMSE (MW)	MAPE (%)	R^2	Tipo
Regressão Linear	2.389,4	2,40	0,882	Modelo
Random Forest	2.676,6	2,44	0,851	Modelo
XGBoost	2.809,3	2,70	0,836	Modelo
Sazonal (D−7)	3.991,1	3,70	0,669	Referência
Persistência (D−1)	5.635,4	5,36	0,341	Referência
SARIMA	6.873,7	6,43	0,019	Modelo

A Regressão Linear obteve o menor RMSE e o maior R^2, reduzindo em 40,1% o erro da melhor referência ingênua. A explicação é estrutural: árvores de decisão nunca produzem valores fora da faixa vista no treino. No teste, 21 dias superaram o máximo do treino (91.360 MWmédio), teto que Random Forest e XGBoost não ultrapassam e a Regressão Linear supera (até 96.452 MWmédio): em série com crescimento sustentado, extrapolar é decisivo. Já o SARIMA, univariado, reproduz o ciclo semanal, mas converge para um patamar em 365 passos.

Para a potência instalada, o diagnóstico precedeu a modelagem. Em 24 anos a média anual nunca recuou, mas a série diária cai em 50,4% dos dias e oscila milhares de MW, o que nenhum mecanismo físico explica. A Figura 1 mostra a origem da contradição: a variação diária não tem estrutura e sua autocorrelação de primeira ordem vale −0,5025, o valor esperado sob ruído independente. Daí um piso

de erro de 5.424 MW, do qual o melhor modelo diário dista apenas 5,77% — já opera no ótimo.

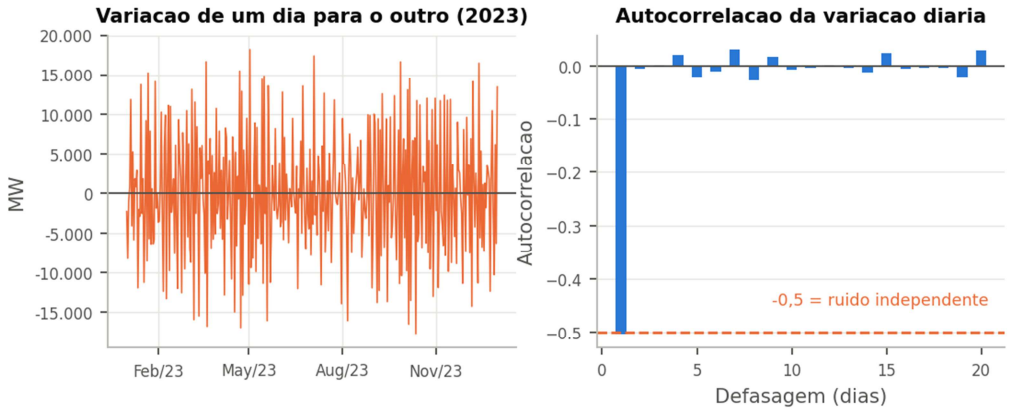


Figura 1 – Diagnóstico do ruído na potência instalada: à esquerda, a variação diária em 2023, sem estrutura; à direita, sua autocorrelação.

Como a média de n observações reduz o desvio do ruído por √n, a previsão foi reformulada na escala mensal, em que o desvio da variação cai de 5.534,5 MW para 997,3 MW. Nessa escala o SARIMA com deriva atingiu RMSE de 1.645 MW, MAPE de 0,73% e R^2 de 0,705, 35,0% melhor que a referência de tendência linear. Os mesmos dados produzem R^2 de 0,181 na escala diária e 0,705 na mensal: a diferença está na granularidade, não no algoritmo. Os resultados alimentam as novas análises Previsão da Carga Elétrica Diária e Expansão da Potência Instalada, no Atlas da Energia do NAPI EZC.

CONCLUSÕES

O desempenho não é atributo do algoritmo, e sim do encontro entre a estrutura da série e a formulação do problema. Para a carga diária, a Regressão Linear múltipla foi o melhor modelo (RMSE de 2.389 MW; R^2 de 0,882), 40,1% à frente da melhor referência ingênua, por ser a única capaz de extrapolar; para a potência instalada, a previsão significativa é a mensal, com RMSE de 1.645 MW e R^2 de 0,705. Dois diagnósticos foram decisivos: o vazamento por identidade contábil e o piso de ruído, que redirecionou o trabalho para a escala em que o sinal existe. Ambos mostram que referências ingênuas e auditoria do significado das variáveis são condições para interpretar qualquer métrica. Integrados ao Atlas da Energia do NAPI EZC, os modelos entregam à plataforma sua camada preditiva; como continuidade, preveem-se validação walk-forward e modelos probabilísticos.

AGRADECIMENTOS

Agradeço ao NAPI EZC pelo fomento, à Fundação Araucária e ao CNPq pelo apoio financeiro e ao orientador Prof. Luiz Fernando Cótica pelas contribuições essenciais a esta pesquisa.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA – ANEEL. **Dados abertos ANEEL**. Disponível em: https://dadosabertos.aneel.gov.br/. Acesso em: 21 ago. 2026.

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

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting**: principles and practice. 3. ed. Melbourne: OTexts, 2021. Disponível em: https://otexts.com/fpp3/. Acesso em: 21 ago. 2026.

NAPI EZC. **Atlas da Energia**. Disponível em: https://napiezc-science.streamlit.app/. Acesso em: 21 ago. 2026.

OPERADOR NACIONAL DO SISTEMA ELÉTRICO – ONS. **Portal de dados abertos**. Disponível em: https://dados.ons.org.br/. Acesso em: 21 ago. 2026.