v7.n11.2026

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

ANÁLISE DO POTENCIAL EÓLICO EM PERNAMBUCO: UMA APLICAÇÃO DE ALGORITMOS DE *MACHINE LEARNING* E *ENSEMBLE LEARNING* PARA PREVISÃO DA VELOCIDADE DO VENTO

ANALYSIS OF WIND ENERGY POTENTIAL IN PERNAMBUCO: AN APPLICATION OF MACHINE LEARNING AND ENSEMBLE LEARNING ALGORITHMS FOR WIND SPEED FORECASTING

ANÁLISIS DEL POTENCIAL EÓLICO EN PERNAMBUCO: UNA APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING Y ENSEMBLE LEARNING PARA LA PREDICCIÓN DE LA VELOCIDAD DEL VIENTO

Viviane Farias Felipe[1], Jader da Silva Jale[2], Silvio Fernando Alves Xavier Júnior[3]

e7119605
https://doi.org/10.47820/recima21.v7i11.9605

PUBLICADO: 11/2026

RESUMO

Este estudo investiga o uso de modelos de aprendizado de máquina para a previsão da velocidade do vento e análise do potencial eólico no estado de Pernambuco, considerando as estações de Recife, Palmares, Garanhuns, Serra Talhada e Petrolina. Foram avaliados algoritmos como LightGBM, XGBoost, CatBoost e *Random Forest*, com base em dados meteorológicos contendo variáveis como temperatura do ar, umidade relativa, pressão atmosférica, radiação solar e velocidade do vento. O pré-processamento incluiu tratamento de valores ausentes, padronização dos dados e incorporação de variáveis temporais. Os modelos foram avaliados com base nas métricas RMSE, MAE e R^2. Os resultados indicaram desempenho consistente dos métodos baseados em *boosting*, com destaque para Recife, que apresentou os menores erros (RMSE = 0,3683, MAE = 0,2836, R^2 = 0,7582), e para Palmares, que apresentou a maior capacidade explicativa (R^2 = 0,8185). A análise de importância das variáveis evidenciou a predominância da componente temporal da velocidade do vento como principal fator preditivo. O teste de Diebold-Mariano indicou, em geral, ausência de diferenças estatisticamente significativas entre os modelos.

PALAVRAS-CHAVE: Modelagem Preditiva. Séries Temporais. Recursos Eólicos. Métodos de *Boosting*.

ABSTRACT

This study investigates the use of machine learning models for wind speed prediction and wind energy potential analysis in the state of Pernambuco, considering the stations of Recife, Palmares, Garanhuns, Serra Talhada, and Petrolina. Algorithms such as LightGBM, XGBoost, CatBoost, and Random Forest were evaluated using meteorological data including air temperature, relative humidity, atmospheric pressure, solar radiation, and wind speed. The preprocessing stage included missing data treatment, data standardization, and the incorporation of temporal variables. The models were evaluated based on RMSE, MAE, and R^2 metrics. The results indicated consistent performance of boosting-based methods, with Recife presenting the lowest errors (RMSE = 0.3683, MAE = 0.2836, R^2 = 0.7582), while Palmares showed the highest explanatory power (R^2 = 0.8185).

[1] Universidade Federal Rural de Pernambuco (UFRPE), Mestre.
[2] Universidade Federal Rural de Pernambuco (UFRPE), Doutor.
[3] Universidade Estadual da Paraíba (UEPB), Doutor.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Variable importance analysis highlighted the predominance of the temporal component of wind speed as the main predictive factor. The Diebold-Mariano test indicated, in general, no statistically significant differences between the models.

KEYWORDS: *Predictive Modeling. Time Series. Wind Resources. Boosting Methods.*

RESUMEN
Este estudio investiga el uso de modelos de aprendizaje automático para la predicción de la velocidad del viento y el análisis del potencial eólico en el estado de Pernambuco, considerando las estaciones de Recife, Palmares, Garanhuns, Serra Talhada y Petrolina. Se evaluaron algoritmos como LightGBM, XGBoost, CatBoost y Random Forest, utilizando datos meteorológicos que incluyen variables como temperatura del aire, humedad relativa, presión atmosférica, radiación solar y velocidad del viento. El preprocesamiento incluyó el tratamiento de valores faltantes, la estandarización de los datos y la incorporación de variables temporales. Los modelos fueron evaluados mediante las métricas RMSE, MAE y R². Los resultados indicaron un desempeño consistente de los métodos basados en boosting, destacándose Recife por presentar los menores errores (RMSE = 0.3683, MAE = 0.2836, R² = 0.7582), y Palmares por mostrar la mayor capacidad explicativa (R² = 0.8185). El análisis de importancia de variables evidenció la predominancia de la componente temporal como principal factor predictivo. La prueba de Diebold-Mariano indicó, en general, ausencia de diferencias estadísticamente significativas entre los modelos.

PALABRAS CLAVE*: Modelado Predictivo. Series Temporales. Recursos Eólicos. Métodos de Boosting.*

INTRODUÇÃO

O crescimento da demanda mundial por energia, aliado às preocupações relacionadas às mudanças climáticas e aos impactos ambientais decorrentes da utilização de combustíveis fósseis, tem impulsionado a busca por fontes energéticas renováveis capazes de promover o desenvolvimento sustentável e reduzir as emissões de gases de efeito estufa. Nesse contexto, a diversificação da matriz energética tornou-se uma estratégia fundamental para garantir segurança energética, sustentabilidade ambiental e competitividade econômica. Entre as diversas fontes renováveis disponíveis, a energia eólica destaca-se por apresentar elevada capacidade de geração de eletricidade a partir de um recurso natural abundante, renovável e de baixo impacto ambiental durante sua operação (Santos *et al.*, 2013).

No Brasil, esse crescimento tornou-se mais evidente a partir dos anos 2000, especialmente na região Nordeste, que apresenta condições atmosféricas favoráveis à geração de energia devido à ocorrência de ventos persistentes, velocidades médias elevadas e reduzida turbulência atmosférica. Como resultado, o Nordeste consolidou-se como a principal região produtora de energia eólica do país, desempenhando papel estratégico na expansão da geração renovável brasileira (Santos *et al.*, 2013; ANEEL, 2024).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

A avaliação do potencial eólico constitui uma etapa essencial para o planejamento, implantação e operação de empreendimentos voltados à geração de energia por meio dos ventos. Essa avaliação depende diretamente do conhecimento detalhado do comportamento da velocidade do vento, variável meteorológica caracterizada por elevada variabilidade temporal e espacial, influenciada por fatores atmosféricos, geográficos e climáticos. Além disso, a variável apresenta padrões sazonais, ciclos diários e dependências temporais que tornam sua modelagem um problema complexo. Dessa forma, a obtenção de previsões confiáveis torna-se fundamental para a estimativa do potencial energético disponível e para a tomada de decisões no setor (Miao *et al.*, 2019; Bagci *et al.*, 2021).

Tradicionalmente, estudos voltados à caracterização do potencial eólico utilizam abordagens estatísticas baseadas no ajuste de distribuições probabilísticas, destacando-se a distribuição de Weibull como uma das mais empregadas na literatura. Trabalhos recentes também têm explorado distribuições mais flexíveis e modelos mistos para representar adequadamente o comportamento da velocidade do vento em diferentes regiões (Wang *et al.*, 2021; Santos *et al.*, 2024). Entretanto, o aumento da disponibilidade de dados meteorológicos e o avanço das técnicas computacionais têm favorecido a utilização de métodos de aprendizado de máquina como alternativas capazes de capturar relações complexas e não lineares presentes nos dados ambientais.

O aprendizado de máquina (*Machine Learning*) tem sido amplamente aplicado em problemas de previsão meteorológica devido à sua capacidade de identificar padrões ocultos em grandes volumes de dados sem a necessidade de pressupostos rígidos sobre a estrutura das relações entre as variáveis. Em aplicações relacionadas à velocidade do vento, esses métodos permitem incorporar simultaneamente informações meteorológicas, variáveis temporais e atributos derivados, proporcionando previsões mais precisas e robustas (Pedregosa *et al.*, 2011). Além disso, técnicas de *Ensemble Learning*, baseadas na combinação de múltiplos modelos preditivos, têm demonstrado desempenho superior em diversos problemas ambientais, contribuindo para a redução dos erros de previsão e para o aumento da capacidade de generalização dos modelos (Amini *et al.*, 2022; Nourani *et al.*, 2020).

No estado de Pernambuco, apesar da relevância estratégica da energia eólica para a matriz energética nordestina, ainda existe a necessidade de estudos que integrem diferentes técnicas de aprendizado de máquina para análise do regime de ventos e avaliação do potencial eólico em distintas regiões climáticas do estado. Pernambuco apresenta significativa diversidade ambiental e atmosférica, abrangendo áreas litorâneas, agrestes e semiáridas, o que resulta em diferentes regimes de vento e potencialidades energéticas. Assim, a análise regionalizada

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

dessas características torna-se importante para subsidiar o planejamento energético e identificar áreas com maior potencial para aproveitamento dos recursos eólicos.

Diante desse cenário, este estudo tem como objetivo avaliar o desempenho de algoritmos de aprendizado de máquina na previsão da velocidade do vento em cinco localidades do estado de Pernambuco, considerando suas distintas condições meteorológicas e regionais. Especificamente, busca-se comparar diferentes algoritmos de regressão, avaliar a contribuição das variáveis meteorológicas e temporais para o processo preditivo e identificar os modelos com melhor desempenho em cada localidade. Ressalta-se que a análise se concentra na previsão da velocidade do vento, variável relevante para os estudos de recursos eólicos, não abrangendo o cálculo direto da potência ou da energia elétrica efetivamente gerada pelos aerogeradores.

1. REFERENCIAL TEÓRICO

A previsão da velocidade do vento tem sido amplamente investigada na literatura, especialmente em função de sua importância para o planejamento e a otimização de sistemas de geração de energia eólica. Devido à natureza estocástica e altamente variável do vento, sua modelagem requer o uso de técnicas capazes de lidar com incertezas e com a complexidade das interações entre variáveis meteorológicas. Nesse contexto, diferentes abordagens têm sido propostas, incluindo métodos estatísticos tradicionais e técnicas mais recentes baseadas em aprendizado de máquina.

Os métodos estatísticos clássicos, como regressão linear, regressões polinomiais e modelos lineares regularizados, são frequentemente utilizados como ponto de partida na modelagem da velocidade do vento, principalmente devido à sua simplicidade e interpretabilidade. Esses modelos permitem estabelecer relações diretas entre a variável resposta e as variáveis explicativas, sendo úteis como referência inicial. No entanto, apresentam limitações significativas quando aplicados a dados meteorológicos, uma vez que tais dados frequentemente exibem comportamento não linear e interações complexas, dificultando a capacidade desses modelos em representar adequadamente os padrões presentes nos dados (Wang *et al.*, 2021).

Diante dessas limitações, técnicas de aprendizado de máquina (*Machine Learning*) têm se destacado como alternativas mais flexíveis na modelagem de fenômenos ambientais. Essas técnicas são capazes de aprender padrões diretamente a partir dos dados, sem a necessidade de especificação prévia de uma forma funcional, o que as torna particularmente adequadas para problemas envolvendo alta complexidade e não linearidade. No aprendizado supervisionado,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

utilizado neste estudo, o modelo é treinado a partir de exemplos previamente conhecidos, nos quais existe uma variável resposta associada a um conjunto de variáveis explicativas, permitindo a construção de modelos capazes de realizar previsões para novas observações (Hastie *et al.*, 2009; Pedregosa *et al.*, 2011).

A representação adequada da componente temporal também constitui um aspecto importante na previsão da velocidade do vento. A incorporação de informações temporais permite representar padrões associados aos ciclos diários e sazonais, bem como a dependência entre observações consecutivas. A utilização de variáveis temporais e de valores defasados da velocidade do vento tem sido empregada em estudos de previsão meteorológica e energética, contribuindo para uma representação mais adequada da dinâmica dos ventos (Amini *et al.*, 2022; Pedregosa *et al.*, 2011).

Embora modelos individuais possam apresentar bom desempenho, sua capacidade de generalização pode ser limitada em cenários caracterizados por elevada variabilidade e incerteza. Nesse sentido, o *Ensemble learning* surge como uma abordagem baseada na combinação de múltiplos modelos com o objetivo de melhorar o desempenho preditivo e reduzir o erro de generalização (Nourani *et al.*, 2020). Entre as principais estratégias destacam-se o *bagging* (bootstrap aggregating) e o *boosting*. O *bagging* consiste na geração de múltiplas amostras dos dados por meio de reamostragem com reposição, permitindo o treinamento de diversos modelos independentes, cujas previsões são posteriormente combinadas. O *boosting*, por sua vez, baseia-se em um processo sequencial de aprendizado, no qual cada novo modelo é ajustado com foco na correção dos erros cometidos pelos modelos anteriores (Bagci *et al.*, 2021).

Entre os algoritmos utilizados em problemas de regressão destacam-se a Ridge Regression, os métodos baseados em Máquinas de Vetores de Suporte (Support Vector Regression) e as técnicas de ensemble fundamentadas em *bagging* e *boosting*, como *Random Forest*, XGBoost, LightGBM e CatBoost (Hoerl; Kennard, 1970; Smola; Schölkopf, 2004; Breiman, 2001; Chen; Guestrin, 2016; Ke *et al.*, 2017; Prokhorenkova *et al.*, 2018). Essas abordagens apresentam potencial para modelar relações não lineares e complexas presentes em fenômenos atmosféricos e têm sido empregadas em estudos voltados à previsão de variáveis meteorológicas e à avaliação de recursos energéticos renováveis (Amini *et al.*, 2022).

No contexto de Pernambuco, a aplicação dessas técnicas mostra-se relevante diante da diversidade ambiental e atmosférica do estado, que apresenta diferentes condições climáticas e regimes de vento. Dessa forma, a utilização de modelos de aprendizado de máquina e *Ensemble learning* possibilita investigar o comportamento da velocidade do vento sob diferentes condições

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

regionais e identificar abordagens com maior capacidade preditiva para cada localidade analisada.

2. METODOLOGIA

2.1. Área de estudo e base de dados

O objeto de análise deste estudo é o estado de Pernambuco, localizado na região Nordeste do Brasil, com área territorial de aproximadamente 98.067,877 km². O estado apresenta diversidade climática significativa, com predominância de clima tropical, caracterizado por temperaturas elevadas ao longo do ano e regimes de precipitação que variam conforme a região, influenciando diretamente o comportamento da velocidade do vento.

Para a realização deste estudo, foram utilizados dados meteorológicos provenientes do Instituto Nacional de Meteorologia (INMET), uma das principais fontes oficiais de dados climáticos no Brasil. As bases de dados utilizadas consistem em séries temporais contendo variáveis relevantes para a modelagem da velocidade do vento, incluindo temperatura do ar, pressão atmosférica, umidade relativa, radiação solar e velocidade do vento.

A análise foi conduzida considerando cinco estações meteorológicas localizadas no estado de Pernambuco, selecionadas de modo a contemplar diferentes contextos regionais: Recife (Região Metropolitana do Recife), Palmares (Mata Pernambucana), Garanhuns (Agreste Pernambucano), Serra Talhada (Sertão do Pajeú) e Petrolina (São Francisco Pernambucano). Essa seleção permite investigar diferenças no comportamento da velocidade do vento entre localidades com características climáticas e geográficas distintas, sem pressupor que uma única estação represente integralmente toda a região correspondente.

Inicialmente, os dados foram organizados cronologicamente e submetidos a procedimentos de verificação de consistência, remoção de registros incompletos e adequação das variáveis necessárias para a modelagem. A variável resposta considerada neste estudo corresponde à velocidade do vento (vel), enquanto as demais variáveis meteorológicas foram utilizadas como potenciais preditoras nos modelos de aprendizado de máquina. Dessa forma, seja o conjunto de observações representado por

$$\{(x_i, y_i)\}_{i=1}^{n} \quad (1)$$

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

em que xix_i = (x_{i1}, x_{i2},...,x_{ip}) representa o vetor de variáveis explicativas e yi corresponde à velocidade do vento observada na i-ésima observação. O objetivo consiste em estimar uma função de regressão

$$y_i = f(x_i) + \varepsilon_i \tag{2}$$

Em que ε_i representa o erro aleatório associado à observação, assumido com média zero e variância constante. Essa estrutura fundamenta a modelagem preditiva adotada neste trabalho, na qual diferentes algoritmos de aprendizado de máquina são utilizados para estimar a função f (·) a partir dos dados observados.

2.2. Pré-processamento e construção dos atributos

Inicialmente, os dados foram organizados em ordem cronológica e filtrados para o período de interesse compreendido entre os anos de 2020 e 2023. Em seguida, foram removidas variáveis que não contribuíam diretamente para o processo de modelagem preditiva, tais como identificadores geográficos e administrativos das estações. Considerando a natureza temporal da velocidade do vento, foram construídos atributos capazes de representar padrões sazonais e dependências temporais presentes nos dados. Para representar a sazonalidade anual, foi calculado o dia do ano associado a cada observação e aplicadas transformações trigonométricas seno e cosseno:

$$DOY_{sin} = sin\left(\frac{2\pi.d}{365{,}25}\right), \tag{3}$$

$$DOY_{cos} = cos\left(\frac{2\pi.d}{365{,}25}\right), \tag{4}$$

em que d representa o dia do ano correspondente à observação

De forma análoga, para representar os ciclos diários associados à dinâmica atmosférica, foram construídas as variáveis:

$$HORA_{sin} = sin\left(\frac{2\pi.h}{24}\right), \tag{5}$$

$$HORA_{cos} = cos\left(\frac{2\pi.h}{24}\right), \tag{6}$$

em que h corresponde à hora da observação.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Além disso, foi criada uma variável defasada da velocidade do vento (lag), definida como:

$$Vel_{lag_1} = Vel_{t-1} \quad (7)$$

permitindo incorporar a dependência temporal existente entre observações consecutivas da série.

A utilização de variáveis defasadas é amplamente empregada em problemas de previsão de séries temporais, pois permite explorar informações históricas relevantes para o processo preditivo (Hastie *et al.*, 2009). Foram realizados testes preliminares com diferentes estruturas de defasagem, sendo selecionada a variável Vel_{lag_1} por apresentar desempenho satisfatório e menor complexidade computacional quando comparada a estruturas mais extensas. Após a construção dos atributos, observações contendo valores ausentes foram removidas da análise.

2.3. Modelagem e validação temporal

Para cada estação meteorológica foi ajustado um conjunto de modelos de regressão pertencentes a diferentes classes de aprendizado de máquina, permitindo comparar abordagens lineares, não lineares e baseadas em *ensemble learning* na previsão da velocidade do vento. A seleção desses algoritmos buscou contemplar diferentes estratégias de modelagem, desde métodos lineares tradicionais até técnicas avançadas de aprendizado de máquina capazes de representar relações complexas entre as variáveis explicativas e a variável resposta. Foram avaliados os modelos *Ridge Regression*, Regressão Polinomial, Support Vector Regression (SVR), Support Vector Regression com *Kerne*l Polinomial (SVR Poly), *Random Forest*, XGBoost, LightGBM e CatBoost.

A *Ridge Regression* foi empregada como modelo linear regularizado (Hastie *et al.*, 2009), enquanto a Regressão Polinomial foi utilizada para representar possíveis relações não lineares entre as variáveis explicativas e a velocidade do vento. Os modelos SVR e SVR Poly foram considerados devido à sua capacidade de modelar padrões complexos por meio de funções kernel (Drucker *et al.*, 1997). Por sua vez, os algoritmos *Random Forest* (Breiman, 2001), XGBoost (Chen; Guestrin, 2016), LightGBM (Ke *et al.*, 2017) e CatBoost (Prokhorenkova *et al.*, 2018) representam métodos baseados em conjuntos de árvores de decisão, amplamente utilizados em problemas de regressão devido à sua capacidade de capturar relações não lineares e interações complexas entre variáveis.

As análises foram realizadas na linguagem Python. O pré-processamento dos dados e a implementação dos modelos foram conduzidos utilizando as bibliotecas Pandas e NumPy,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

enquanto os algoritmos de aprendizado de máquina foram implementados por meio da biblioteca Scikit-Learn (Pedregosa *et al.*, 2011). Os modelos XGBoost, LightGBM e CatBoost foram ajustados utilizando suas respectivas bibliotecas especializadas (Chen; Guestrin, 2016; Ke *et al.*, 2017; Prokhorenkova *et al.*, 2018).

Os algoritmos de aprendizado de máquina utilizados neste estudo possuem hiperparâmetros responsáveis por controlar aspectos relacionados à complexidade dos modelos, regularização e capacidade de generalização. A definição adequada desses parâmetros influencia diretamente o desempenho preditivo dos algoritmos e o risco de sobreajuste (*overfitting*).

Neste trabalho, os hiperparâmetros foram definidos com base em recomendações da literatura especializada em aprendizado estatístico e aprendizado de máquina (Hastie *et al.*, 2009), bem como em experimentações preliminares conduzidas durante o desenvolvimento das análises. Essa estratégia buscou equilibrar capacidade preditiva, estabilidade numérica e capacidade de generalização dos modelos. Os principais hiperparâmetros utilizados em cada algoritmo são apresentados na Tabela 1.

Tabela 1. Hiperparâmetros utilizados nos modelos de aprendizado de máquina

Modelo	**Configuração utilizada**
Ridge	`Ridge(alpha=0.1)`
Random Forest	`RandomForestRegressor(n_estimators=500, max_depth=10)`
XGBoost	`XGBRegressor(n_estimators=5000, learning_rate=0.05, max_depth=5)`
LightGBM	`LGBMRegressor(n_estimators=5000, learning_rate=0.01, max_depth=7)`
CatBoost	`CatBoostRegressor(iterations=1000,learning_rate=0.01, depth=6)`
SVR	`SVR(kernel='rbf', C=10, epsilon=0.01)`
SVR Poly	`SVR(kernel='poly', C=1.0, epsilon=0.01)`
Regressão Polinomial	`PolynomialFeatures(degree=2) + Ridge(alpha=0.1)`

Fonte: Elaborado pelos autores.

A avaliação dos modelos foi realizada por meio de uma estratégia de validação temporal do tipo *hold-out*. Nesse procedimento, os registros mais recentes foram reservados para teste, enquanto as observações anteriores foram utilizadas para treinamento dos modelos.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Considerando a data máxima disponível em cada estação meteorológica, os quatro meses finais da série temporal foram destinados ao conjunto de teste, enquanto os demais registros compuseram o conjunto de treinamento. A escolha dos quatro meses finais como conjunto de teste buscou garantir uma quantidade suficiente de observações para avaliação dos modelos, preservando simultaneamente um volume adequado de dados para treinamento. Além disso, essa definição buscou reproduzir um cenário real de previsão, no qual os modelos são ajustados utilizando informações históricas e posteriormente empregados para estimar observações futuras ainda não observadas.

Essa abordagem preserva a ordem cronológica das observações e evita o uso de informações futuras durante o treinamento dos modelos (*data leakage*), característica essencial em problemas de previsão de séries temporais. Além disso, a divisão temporal adotada permite avaliar a capacidade de generalização dos algoritmos em condições semelhantes às encontradas em aplicações reais de previsão da velocidade do vento.

2.4. Avaliação e comparação dos modelos

A avaliação dos modelos foi realizada com base nas métricas Raiz do Erro Quadrático Médio *(Root Mean Squared Error* – RMSE), Erro Absoluto Médio (*Mean Absolute Error* – MAE) e coeficiente de determinação (R^2). Essas métricas foram calculadas utilizando o conjunto de teste obtido pela estratégia de validação temporal *hold-out*, permitindo avaliar a capacidade de generalização dos modelos para observações não utilizadas durante o treinamento (Pedregosa *et al.*, 2011; Hastie *et al.*, 2009).

As métricas utilizadas são apresentadas a seguir:

Raiz do Erro Quadrático Médio (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

em que y_i representa o valor observado $\hat{y}_i$ o valor previsto pelo modelo e n o número de observações do conjunto de teste. Essa métrica penaliza erros de maior magnitude, sendo útil para avaliar modelos nos quais grandes desvios de previsão devem ser minimizados.

Erro Absoluto Médio (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

em que y_i representa o valor observado $\hat{y}_i$ e n possuem as mesmas definições apresentadas anteriormente. O MAE representa a magnitude média dos erros de previsão, atribuindo o mesmo peso a todos os desvios em termos absolutos.

Coeficiente de Determinação (R²)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (10)$$

em que $\bar{y}$ representa a média dos valores observados. O (R^2) indica a proporção da variabilidade da variável resposta explicada pelo modelo, sendo que valores mais próximos de 1 indicam maior capacidade explicativa.

Além das métricas de desempenho, a interpretação dos modelos foi realizada por meio da *Permutation Importance*, que avalia a relevância das variáveis explicativas a partir do impacto do seu embaralhamento sobre o desempenho preditivo. Maiores alterações nas previsões indicam maior importância da variável, enquanto pequenas alterações sugerem menor contribuição para o modelo (Molnar, 2022). Para verificar se as diferenças de desempenho entre os modelos eram estatisticamente significativas, foi aplicado o teste de Diebold-Mariano, cuja hipótese nula estabelece igualdade na acurácia preditiva entre dois modelos (Diebold; Mariano, 1995). Dessa forma, a análise conjunta das métricas RMSE, MAE e (R^2), da *Permutation Importance* e do teste de Diebold-Mariano permitiu avaliar, respectivamente, o desempenho preditivo, a relevância das variáveis e a significância das diferenças entre os modelos.

3. RESULTADOS E DISCUSSÃO

A Tabela 2 apresenta medidas estatísticas e informações geográficas das cinco estações meteorológicas analisadas, permitindo caracterizar o comportamento da velocidade do vento nas diferentes regiões do estado de Pernambuco.

Tabela 2. Estatísticas descritivas da velocidade do vento para as estações meteorológicas analisadas

Estação	Latitude	Longitude	x̄	sd	p25	p75	Máx	CV	AI
Recife	-8.05928	-34.95924	1.4332	0.8383	0.7	2.1	3.8	0.5849	1.4
Palmares	-8.66667	-35.56806	1.9565	1.3434	0.9	2.9	8.5	0.6867	2.0
Garanhuns	-8.91095	-36.49338	2.3716	1.2236	1.6	3.1	9.2	0.5159	1.5
Serra Talhada	-7.95428	-38.29508	2.2055	1.0737	1.4	2.9	7.8	0.4868	1.5
Petrolina	-9.38832	-40.52326	3.0285	1.4620	2.2	4.0	9.3	0.4827	1.8

Fonte: Elaborado pelos autores.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Observa-se que as velocidades médias do vento variam entre as diferentes regiões de Pernambuco, refletindo características climáticas e geográficas específicas de cada localidade. Petrolina apresentou a maior velocidade média (3,0285 m/s), seguida por Garanhuns (2,3716 m/s) e Serra Talhada (2,2055 m/s), enquanto Recife apresentou a menor média (1,4332 m/s). De modo geral, as estações do interior apresentaram velocidades médias superiores às observadas em Recife e Palmares.

Quanto à variabilidade, Petrolina apresentou o maior desvio-padrão (1,4620 m/s), enquanto Recife apresentou o menor (0,8383 m/s). Em termos relativos, Palmares apresentou o maior coeficiente de variação (0,6867), indicando maior heterogeneidade dos valores em relação à sua média, enquanto Petrolina apresentou o menor (0,4827).

Os maiores valores máximos foram registrados em Petrolina (9,3 m/s), Garanhuns (9,2 m/s) e Palmares (8,5 m/s), enquanto Recife apresentou o menor máximo (3,8 m/s). Esses resultados evidenciam diferenças na intensidade e na variabilidade do vento entre as localidades, fornecendo uma caracterização inicial das séries utilizadas na modelagem preditiva.

A Tabela 3 apresenta o desempenho dos modelos preditivos nas diferentes estações meteorológicas analisadas. De modo geral, observa-se que menores valores de RMSE e MAE indicam melhor desempenho preditivo, enquanto valores mais elevados de R^2 refletem maior capacidade explicativa dos modelos.

Tabela 3. Desempenho dos principais modelos preditivos nas estações meteorológicas analisadas

Estação	Modelo	$RMSE$	MAE	R^2
Recife	CatBoost	0.368281	0.283573	0.758204
	LightGBM	0.370517	0.286221	0.755259
Petrolina	LightGBM	0.526804	0.407339	0.791677
	XGBoost	0.534354	0.415297	0.785663
Garanhuns	CatBoost	0.549297	0.423468	0.764000
	LightGBM	0.549878	0.423666	0.763501
Palmares	CatBoost	0.552453	0.407629	0.820037
	Random Forest	0.554772	0.407184	0.818523
Serra Talhada	LightGBM	0.550638	0.417091	0.641219
	XGBoost	0.567860	0.434305	0.618425

Fonte: Elaborado pelos autores.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Observa-se que a estação Recife apresentou o melhor desempenho entre todas as localidades, com os menores valores de RMSE e MAE, indicando maior previsibilidade da velocidade do vento nessa região. De forma semelhante, Petrolina também apresentou resultados satisfatórios, destacando-se pelo maior R^2 (0,791677), o que sugere boa capacidade dos modelos em explicar a variabilidade dos dados.

Por outro lado, a estação de Serra Talhada apresentou o pior desempenho geral, com os maiores valores de RMSE e MAE e os menores valores de R^2, evidenciando maior dificuldade dos modelos em capturar os padrões da velocidade do vento nessa localidade. Esse comportamento pode estar associado a uma maior variabilidade dos dados ou à influência de fatores locais mais complexos.

As estações de Garanhuns e Palmares apresentaram desempenho intermediário, com métricas relativamente próximas entre si. Em particular, Palmares destacou-se por apresentar o maior valor de R^2 entre todas as estações (0,818523), embora seus erros médios não tenham sido os menores, indicando que, apesar de boa capacidade explicativa, ainda há presença de variações nos erros de predição.

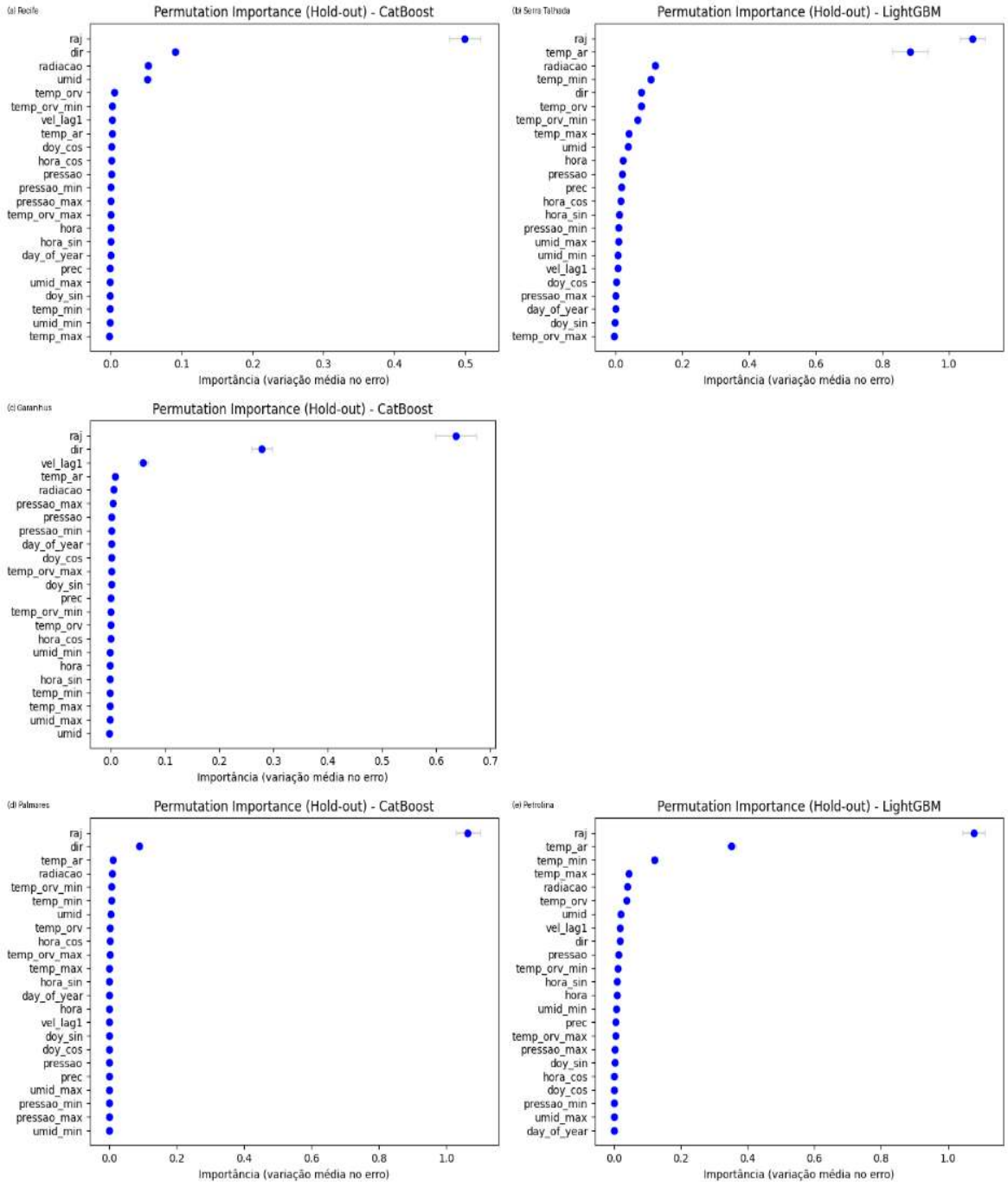
De forma geral, observa-se um comportamento relativamente estável dos modelos nas estações de Recife, Petrolina e Garanhuns, onde as diferenças entre os algoritmos são pequenas, sugerindo consistência nos resultados. Em contrapartida, maior variabilidade no desempenho é observada em Serra Talhada, indicando maior sensibilidade dos modelos às características dos dados nessa região.

Além disso, os modelos baseados em técnicas de *boosting* (LightGBM, XGBoost e CatBoost) apresentaram desempenho consistente na maioria das estações, reforçando sua eficácia na modelagem de variáveis meteorológicas complexas.

A Figura 1 apresenta a análise de importância das variáveis por permutação (*Permutation Importance)* para a modelagem da velocidade do vento nas estações de Recife, Serra Talhada, Garanhuns, Palmares e Petrolina, considerando os modelos de melhor desempenho em cada localidade.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figura 1. Permutation Importance (*hold-out*) para as estações meteorológicas analisadas: (a) Recife, (b) Serra Talhada, (c) Garanhuns, (d) Palmares e (e) Petrolina.



Fonte: Elaborado pelos autores.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Observa-se que a variável associada ao histórico da velocidade do vento (raj) se destaca como a mais influente em todas as estações analisadas, evidenciando a forte dependência temporal do processo eólico e indicando que os valores passados exercem papel determinante na capacidade preditiva dos modelos.

Nas diferentes localidades, embora a variável raj concentre a maior parcela da importância, outras variáveis meteorológicas, como temperatura do ar, radiação solar e umidade relativa, também contribuem para a modelagem, ainda que em menor magnitude. Em Recife, Garanhuns e Palmares, onde o modelo CatBoost apresentou melhor desempenho, nota-se uma distribuição mais equilibrada da importância entre as variáveis secundárias, sugerindo maior capacidade do modelo em capturar interações não lineares entre os preditores. Por outro lado, em Petrolina e Serra Talhada, modeladas com LightGBM, observa-se maior concentração da importância na variável temporal, indicando maior dependência do comportamento histórico da série.

As variáveis associadas à sazonalidade, representadas por componentes harmônicas do tempo (como seno e cosseno do dia do ano e da hora), apresentam baixa contribuição relativa na maioria das estações. Esse comportamento sugere que os padrões cíclicos exercem influência limitada na escala temporal considerada ou que tais efeitos já estão implicitamente incorporados na variável autorregressiva utilizada nos modelos.

De forma complementar, os resultados do teste de Diebold-Mariano indicam que, na maior parte das estações, não há diferença estatisticamente significativa entre os modelos avaliados, como observado em Recife (estatística DM = -0,9212; p-valor = 0,3569), Garanhuns (estatística DM = -0,0812; p-valor = 0,9353) e Palmares (estatística DM = -0,5284; p-valor = 0,5972), reforçando a similaridade no desempenho preditivo. Em Petrolina, foi identificada evidência marginal de diferença entre os modelos (estatística DM = -1,7786; p-valor = 0,0753), enquanto em Serra Talhada observou-se diferença estatisticamente significativa (estatística DM = -3,9942; p-valor = 0,0001), indicando maior sensibilidade da modelagem às escolhas algorítmicas nessa localidade.

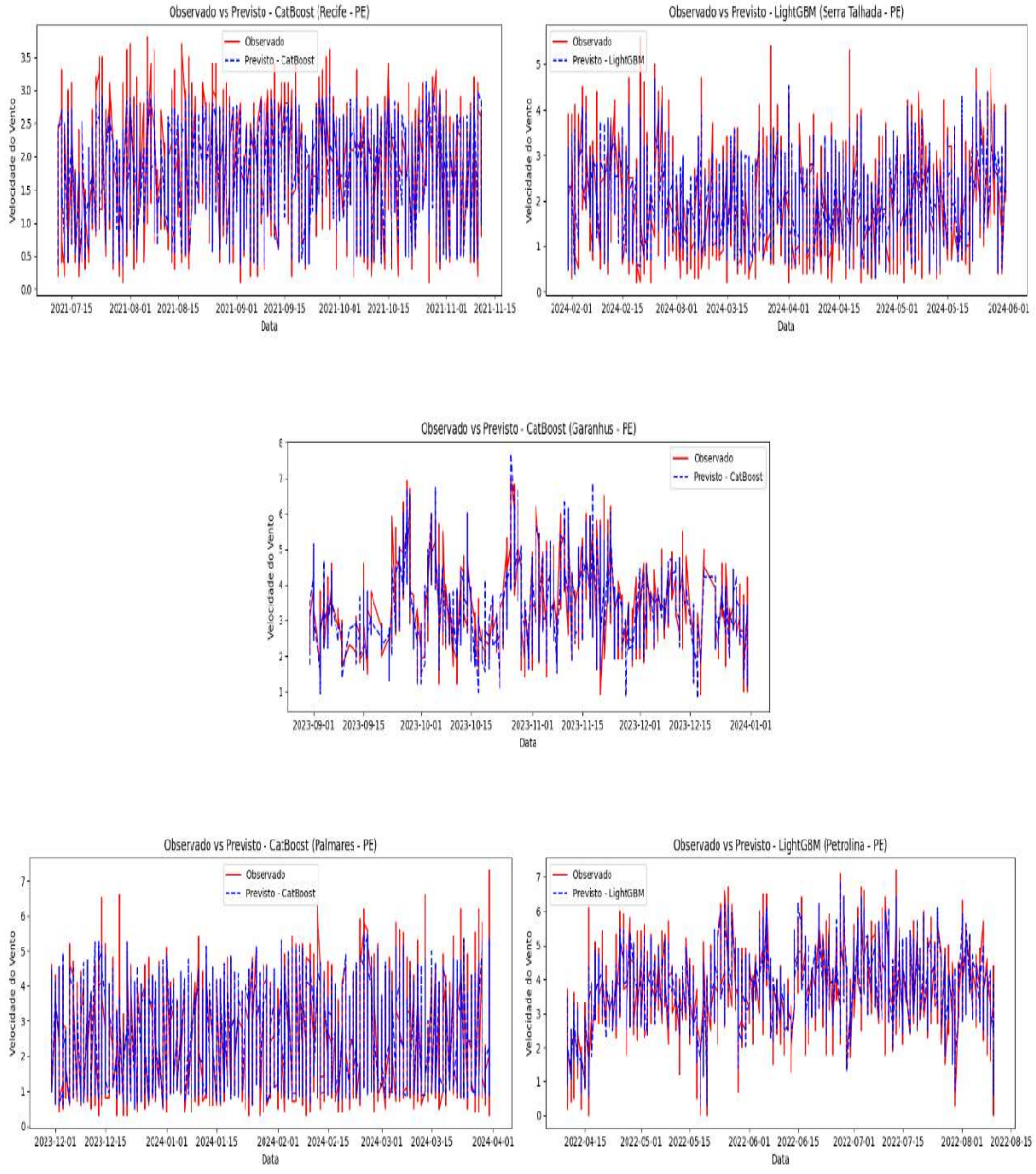
Os resultados sugerem que a predominância da variável temporal contribui para a convergência no desempenho dos modelos, reduzindo as diferenças entre as abordagens de machine learning. Assim, embora existam variações pontuais entre os algoritmos, especialmente em cenários com maior complexidade da série, os modelos analisados demonstram comportamento consistente na representação da dinâmica do vento, reforçando sua aplicabilidade na análise do potencial eólico no estado de Pernambuco.

A Figura 2 apresenta a série temporal com os valores observados (linha vermelha contínua) e os valores previstos (linha azul tracejada) da velocidade do vento para diferentes localidades do estado

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

de Pernambuco — Petrolina, Recife, Serra Talhada, Garanhuns e Palmares — utilizando os modelos LightGBM e CatBoost, conforme indicado em cada subfigura. Observa-se que os modelos conseguem acompanhar satisfatoriamente a dinâmica temporal da variável, reproduzindo a variabilidade característica da velocidade do vento nas diferentes regiões analisadas.

Figura 2. Valores observados versus previstos da velocidade do vento para cada estação



Fonte: Elaborado pelos autores.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Em Petrolina e Serra Talhada, onde foi empregado o LightGBM, nota-se que o modelo apresenta boa capacidade de ajuste às oscilações da série, especialmente na representação dos picos e vales. Em Petrolina, essa aderência é mais evidente, com forte correspondência entre os valores observados e previstos ao longo do período. Já em Serra Talhada, embora o modelo acompanhe a tendência geral, percebe-se uma maior dispersão entre as curvas em alguns trechos, indicando maior dificuldade em capturar variações mais abruptas, comportamento que é coerente com o menor coeficiente de determinação observado para essa localidade na Tabela 3.

Nas estações de Recife, Garanhuns e Palmares, modeladas com CatBoost, também se observa uma boa correspondência entre as séries. Em Recife, as curvas praticamente se sobrepõem na maior parte do período, evidenciando um ajuste consistente do modelo. Em Garanhuns, o modelo mantém boa capacidade de acompanhar a variabilidade temporal, ainda que com pequenas diferenças pontuais. Em Palmares, destaca-se a forte aderência entre os valores observados e previstos, mesmo diante de flutuações mais intensas, o que está em consonância com o melhor desempenho estatístico apresentado na Tabela 3 para essa estação.

As diferenças entre os valores observados e previstos, caracterizadas por episódios de subestimação e superestimação, são pontuais e esperadas em séries meteorológicas, dada a natureza complexa e não linear dos processos atmosféricos. Assim, a análise visual das figuras, aliada aos resultados quantitativos, confirma que os modelos empregados apresentam desempenho robusto, sendo capazes de capturar tanto padrões de baixa quanto de alta frequência, o que reforça sua aplicabilidade em contextos como previsão operacional e planejamento energético.

4. CONSIDERAÇÕES FINAIS

Este trabalho evidenciou o potencial das técnicas de aprendizado de máquina na modelagem da velocidade do vento em diferentes regiões do estado de Pernambuco, destacando a eficiência dos modelos baseados em *boosting* na captura de padrões complexos e não lineares presentes em séries meteorológicas.

De maneira geral, observou-se que a estação de Recife apresentou o melhor desempenho preditivo, com menores valores de erro e boa capacidade de ajuste, indicando um comportamento mais estável e previsível da velocidade do vento. Em contrapartida, embora Palmares tenha apresentado o maior coeficiente de determinação (R^2), sugerindo elevada capacidade explicativa dos modelos, seus erros médios indicam a presença de maior variabilidade nas previsões. Já a estação de Serra Talhada destacou-se negativamente,

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

evidenciando maior dificuldade dos modelos em representar a dinâmica do vento, possivelmente devido à maior complexidade dos fatores locais.

Os resultados também mostraram que, apesar da aplicação de diferentes algoritmos, há uma relativa convergência no desempenho entre os modelos em algumas estações, fato reforçado pelos testes estatísticos realizados. Isso sugere que a forte influência da componente temporal da velocidade do vento reduz as diferenças entre abordagens, tornando o histórico da variável um fator dominante na modelagem.

A análise de importância das variáveis confirmou que informações passadas da velocidade do vento desempenham papel central na previsão, enquanto variáveis meteorológicas complementares contribuem de forma secundária, mas relevante para o refinamento dos modelos.

Do ponto de vista aplicado, os achados deste estudo reforçam a viabilidade do uso de modelos de *machine learning* como ferramentas de apoio ao planejamento energético, especialmente em regiões com elevado potencial eólico. Além disso, contribuem para o avanço de metodologias preditivas voltadas à área ambiental, oferecendo subsídios para estudos futuros e aplicações operacionais.

Como perspectivas futuras, recomenda-se a ampliação do conjunto de dados, incluindo séries mais longas e maior número de estações meteorológicas, bem como a exploração de modelos híbridos e técnicas mais avançadas, como redes neurais profundas. A incorporação de dados de reanálise atmosférica e variáveis adicionais também pode contribuir para melhorar o desempenho preditivo, especialmente em regiões com maior variabilidade climática.

REFERÊNCIAS

AMINI, A.; DOLATSHAHI, M.; KERACHIAN, R. Adaptive precipitation nowcasting using deep learning and ensemble modeling. Journal of Hydrology, v. 612, p. 128197, 2022.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (ANEEL). Site oficial da ANEEL. Brasília: ANEEL, 2024. Disponível em: https://www.gov.br/aneel/pt-br.

BAGCI, K.; ARSLAN, T.; CELIK, H. E. Inverted kumarswamy distribution for modeling wind speed data: Lake Van, Turkey. Renewable and Sustainable Energy Reviews, v. 135, p. 110110, 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S1364032120304019.

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

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM
SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 785–794, 2016.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

DIEBOLD, F. X.; MARIANO, R. S. Comparing predictive accuracy. Journal of Business and Economic Statistics, v. 13, n. 3, p. 253–263, 1995.

DRUCKER, H.; BURGES, C. J. C.; KAUFMAN, L.; SMOLA, A. J.; VAPNIK, V. Support vector regression machines. Advances in Neural Information Processing Systems, v. 9, 1997.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. ed. [S.l.]: Springer, 2009.

HOERL, A. E.; KENNARD, R. W. Ridge regression: Biased estimation for nonorthogonal problems.
Technometrics, v. 12, n. 1, p. 55–67, 1970.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, v. 30, 2017.

MIAO, S.; GU, Y.; LI, D.; LI, H. Determining suitable region wind speed probability distribution using optimal score-radar map. Energy Conversion and Management, v. 183, p. 590–603, 2019.

MOLNAR, C. Interpretable machine learning. Lulu.com, 2022.

NOURANI, V.; PARHIZKARI, A.; BAGHETANI, S. Ensemble modeling in hydrological prediction: Concepts, methods, and applications. Journal of Hydrology, v. 590, p. 125155, 2020.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: *Machine learning* in Python. Journal of *Machine Learning* Research, v. 12, p. 2825–2830, 2011.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. CatBoost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems,
v. 31, 2018.

SANTOS, F. S. dos; NASCIMENTO, K. K. F. do; JALE, J. da S.; XAVIER JÚNIOR, S. F. A.; FERREIRA, T. A. Brazilian wind energy generation potential using mixtures of Weibull distributions. Renewable and Sustainable Energy Reviews, v. 189, p. 113990, 2024.

SANTOS, J. d.; SANTOS, R.; SILVA, T. Energia renovável e impactos ambientais no Brasil. Revista Brasileira de Energia Sustentável, v. 2, p. 45–58, 2013.

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. Statistics and Computing, v. 14, p. 199–222, 2004.

WANG, Y.; LI, Y.; ZOU, R.; SONG, D. Bayesian infinite mixture models for wind speed distribution estimation. Energy Conversion and Management, v. 236, p. 113946, 2021.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.