UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA MECÂNICA
CURSO DE GRADUAÇÃO EM ENGENHARIA DE ENERGIAS RENOVÁVEIS

ARTHUR CARVALHO FERNANDES

APLICAÇÃO DE *MACHINE LEARNING* PARA A PREVISÃO DA VELOCIDADE DO VENTO EM DIFERENTES LOCALIDADES PARA UM HORIZONTE DE PREVISÃO DE 24 HORAS

FORTALEZA
2025

ARTHUR CARVALHO FERNANDES

APLICAÇÃO DE *MACHINE LEARNING* PARA A PREVISÃO DA VELOCIDADE DO VENTO EM DIFERENTES LOCALIDADES PARA UM HORIZONTE DE PREVISÃO DE 24 HORAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Energias Renováveis do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Engenheiro de Energias Renováveis.

Orientadora: Prof.ª Dra. Carla Freitas de Andrade.

Co-orientador: Prof. Dr. Paulo Alexandre Costa Rocha.

FORTALEZA

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F398a Fernandes, Arthur Carvalho.
Aplicação de Machine Learning para a previsão da velocidade do vento em diferentes localidades para um horizonte de previsão de 24 horas / Arthur Carvalho Fernandes. – 2025.
91 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia de Energias Renováveis, Fortaleza, 2025.
Orientação: Profa. Dra. Carla Freitas de Andrade.
Coorientação: Prof. Dr. Paulo Alexandre Costa Rocha.

1. Previsão da velocidade do vento. 2. Aprendizado de máquina. 3. Previsão numérica do tempo. I. Título.

CDD 621.042

ARTHUR CARVALHO FERNANDES

APLICAÇÃO DE *MACHINE LEARNING* PARA A PREVISÃO DA VELOCIDADE DO VENTO EM DIFERENTES LOCALIDADES PARA UM HORIZONTE DE PREVISÃO DE 24 HORAS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Energias Renováveis do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia de Energias Renováveis.

Aprovada em: 31/07/2025.

BANCA EXAMINADORA

_____

Prof.ª Dra. Carla Freitas de Andrade (Orientadora)
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Paulo Alexandre Costa Rocha (Co-orientador)
Universidade Federal do Ceará (UFC)

_____

Prof. Dr. Francisco Ilson da Silva Junior
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

À minha mãe, Renata Cardoso, e ao meu pai, Giovanni Fernandes, cujo amor incondicional e apoio constante são minhas fontes de inspiração.

Às minhas avós, Maria Luíza e Francisca Cardoso, e aos meus avôs Ayrton e Lourival, cuja dedicação e valores transmitidos desde os primeiros anos de vida formaram a base do meu caráter.

Aos meus padrinhos, Ana Luiza e Bráulio Fernandes, pelos sábios conselhos, pela amizade sincera e pelo apoio constante em cada etapa da minha jornada acadêmica.

À minha tia, Herlane Fernandes, que esteve ao meu lado em todas as circunstâncias da graduação, oferecendo seu tempo e auxílio.

Aos meus irmãos, Ana Rita, Giovanna e João Pedro, que, mesmo com o passar dos anos serão sempre minha fonte de proteção.

À minha sobrinha e afilhada, Ayla Isabella, que chegou de forma inesperada e inundou minha vida de amor e afeto.

À minhas bisavó Maria Francisca e ao meu bisavô Mariano, cujo legado de força e sabedoria atravessa gerações e me inspira a honrar nossas raízes.

Aos meus amigos Gilberto, Marcos Davi, Lucas e Wilson por todos os momentos alegres vividos de forma única desde o momento em que nos conhecemos.

Aos meus amigos do intercâmbio, Mariana, Giordano, Sayuri, Leandro, Lucas, Vinicius, Ingrid, Anna Flávia, Clémence, Emilie e Charlotte. Graças a vocês pude construir um segundo lar fora do Brasil e tive forças para superar todos os desafios no caminho.

Aos meus colegas de trabalho que me fizeram evoluir como pessoa e profissional.

À Prof.ª Dra. Carla Andrade, minha orientadora, pela confiança depositada em meu potencial e pelas valiosas orientações que guiaram cada etapa deste trabalho.

Ao Prof. Dr. Francisco Ilson da Silva Júnior, pelos ensinamentos transmitidos ao longo da graduação e por todo auxílio durante meu período de estudos na França.

À Universidade Federal do Ceará, pela excelência de ensino e pelas inúmeras oportunidades de aprendizado e crescimento profissional que me foram proporcionadas.

À CAPES, pelo investimento em minha formação e pela bolsa de estudos que tornou possível a realização de um grande sonho pessoal.

Aos membros da banca examinadora – Prof.ª Dra. Carla Freitas de Andrade, Prof. Dr. Francisco Ilson da Silva Júnior e Prof. Dr. Paulo Alexandre Costa Rocha – agradeço pelo tempo dedicado e pelas contribuições que certamente puderem enriquecer este trabalho.

A todos que, de forma direta ou indireta, fizeram parte desta trajetória, o meu mais sincero e profundo obrigado.

“Impossível é uma palavra muito grande que gente pequena usa para tentar nos oprimir.”

(Charlie Brown Jr).

RESUMO

Em consonância com o crescimento acelerado das fontes renováveis na matriz energética brasileira e mundial, especialmente da energia eólica, a previsão precisa da velocidade do vento tornou-se essencial para o planejamento e a operação eficiente do sistema elétrico nacional. Nessa tangente, a aplicação de algoritmos de aprendizado de máquina surge como uma alternativa para complementar e corrigir as limitações existentes nos modelos tradicionais de previsão numérica do tempo. Diante desse contexto, o objetivo geral do trabalho é de alcançar uma melhora na previsão da velocidade do vento com o uso de um algoritmo de aprendizado de máquina para quatro localidades distintas, são elas: Senhor do Bonfim (BA), Conde (BA), Mossoró (RN) e Rio Grande (RS). Os dados de previsão utilizados são dos modelos globais de Previsão Numérica do Tempo *ARPEGE* e *GFS*. Em um período aproximado de seis meses de dados, foi feita uma análise comparativa preliminar dos dados medidos e dados previstos, uma análise de variáveis usando métodos de *feature engineering*, uma otimização dos hiperparâmetros para encontrar a melhor combinação, e o treinamento do modelo de aprendizado de máquina para dados de *ARPEGE* e *GFS* a fim de compará-los com as suas previsões originais e com o método da Persistência. O algoritmo de aprendizado utilizado foi o *XGBoost*, e com a metodologia utilizada foi alcançado resultados significativos utilizando métricas como o *RMSE* e o R^2 para avaliação. Em Mossoró foram obtidas as melhores performances do modelo, alcançando uma melhora de 73,13% para *XGBoost ARPEGE* e de 45,45% para *XGBoost GFS* em relação às suas previsões originais. Registrando um coeficiente de determinação (R^2) acima de 0,55, os dois algoritmos registraram uma redução de mais de 40% do *RMSE* médio se comparados ao método da Persistência no mesmo período.

Palavras-chave: Previsão da velocidade do vento; Aprendizado de máquina; Previsão numérica do tempo, Energia eólica, XGBoost.

ABSTRACT

In line with the accelerated growth of renewable sources in the Brazilian and global energy matrix, especially wind energy, the precise forecasting of wind speed has become essential for the planning and efficient operation of the national electrical system. In this context, the application of machine learning algorithms emerges as an alternative to complement and correct the existing limitations in traditional numerical weather prediction models. Given this context, the general objective of this work is to achieve an improvement in wind speed forecasting by using a machine learning algorithm for four distinct locations: Senhor do Bonfim (BA), Conde (BA), Mossoró (RN), and Rio Grande (RS). The forecast data used are from the global Numerical Weather Prediction models ARPEGE and GFS. With approximately six months of data, a preliminary comparative analysis of measured and predicted data was conducted, along with a variable analysis using feature engineering methods, hyperparameter optimization to find the best combination, and the training of the machine learning model for ARPEGE and GFS data in order to compare them with their original forecasts and with the Persistence method. The machine learning algorithm used was XGBoost, and with the applied methodology, significant results were achieved using metrics such as RMSE and R^2 for evaluation. The best model performances were obtained in Mossoró, achieving an improvement of 73.13% for XGBoost ARPEGE and 45.45% for XGBoost GFS in relation to their original forecasts. Reaching a coefficient of determination (R^2) above 0.55, both algorithms registered a reduction of more than 40% in the average RMSE when compared to the Persistence method in the same period.

Keywords: Wind speed forecasting; Machine learning; Numerical weather prediction; Wind energy; XGBoost.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ONS	Operador Nacional do Sistema Elétrico
NWP	Numerical Weather Prediction
MAPE	Mean Absolute Percentage Error
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
NRMSE	Normalized Root Mean Squared Error
R^2	Coeficiente de Determinação
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
ANN	Artificial Neural Network
GFS	Global Forecast System
ARPEGE	Action de Recherche Petite Echelle Grand Echelle
ECMWF	European Centre for Medium-Range Weather Forecasts
INMET	Instituto Nacional de Meteorologia
EMA	Estação Meteorológica Automática

SUMÁRIO

1 INTRODUÇÃO

Desde os tempos antigos a previsão do tempo é usada pela sociedade, como por exemplo para a previsão da chuva e se o dia será de sol, para a previsão do inverno e do verão, com o intuito de saber o tempo certo para a colheita, entre outros. (PAROLINI, 2022)

Evoluindo para o contexto atual, a previsão do tempo se tornou crucial para anteciparmos fenômenos potencialmente catastróficos como furacões e grandes tempestades, assim como fenômenos do cotidiano como a variação de temperatura, a velocidade do vento, a umidade do ar, entre outros (SHEN et al., 2022). Variáveis essas que também influenciam em setores econômicos como o de agricultura e o de energia, esse último que precisa de uma previsibilidade para o planejamento energético do país. (SHEN et al., 2023)

A matriz energética mundial está cada vez mais renovável e, no Brasil, entre 2018 e 2023, a participação combinada de energia solar e eólica na matriz elétrica brasileira cresceu de 9 % para 22 %, com a capacidade de geração dessas fontes crescendo em torno de 180% entre 2018 e 2022 (REUTERS, 2024). Por isso, é preciso prever com o máximo de acurácia possível qual será o impacto dos fenômenos e das variáveis meteorológicas, principalmente para o setor elétrico, em que a previsão meteorológica é de suma importância para a dosagem da distribuição de energia.

Esses fatores influenciam na geração de energia, podendo ser maior ou menor por causa deles. O Operador Nacional do Sistema Elético (ONS) deve dar o aval corretamente para quando as energias devem ser distribuídas para as casas com o menor custo e no momento em que é necessário. Sabe-se que os maiores momentos de consumo de eletricidade no Brasil é no período da noite, entre 17h e 22h. (ROBERTS et al., 2014).

Em 1950 houve a implementação do primeiro modelo de Previsão Numérica do Tempo (*NWP*). Desde então, esses modelos são otimizados de maneira que a acurácia seja maior. De acordo com Lynch (2006), os modelos de *NWP* possuem uma limitação em resolver fenômenos de microescala, pois tais eventos ocorrem em uma escala espacial inferior à resolução de suas grades de cálculo.

Para contornar essas e outras limitações, uma das técnicas mais populares nos dias de hoje é o uso de aprendizado de máquina (*machine learning*) para o pós-processamento dos resultados, melhorando o desempenho do modelo e corrigindo os vieses existentes frente a processos físicos complexos (RASMUSSEN et al., 2021).

Nos últimos anos, a Inteligência Artifical (IA) vem sendo aplicada em diversos contextos, dentro deles existem os algoritmos de aprendizado de máquinas, criados para aprenderem através de variáveis de entrada comportamentos complexos e padrões que seriam difícéis de prever com equações (HAM; KIM; LUO, 2023) .O Brasil possui alguns modelos que usam algoritmo de *machine learning* e visam melhorar a acurácia da Previsão Numérica do

Tempo, tais como o *Brazilian Developments on the Regional Atmospheric Modeling System* (*BRAMS*).

Este trabalho possui o propósito de estudar e avaliar um algortimo de aprendizado de máquina para a correção de modelos de Previsão Numérica do Tempo de forma que seja alcançada uma aprimoração da precisão da previsão da velocidade do vento para quatro diferentes localidades no Brasil.

2 OBJETIVOS

O presente trabalho possui como objetivo principal:

- Alcançar uma melhora na previsão da velocidade do vento com o uso de algortimo de aprendizado de máquina para as quatro localidades em estudo.

O objetivo geral se desdobra dentre os seguintes objetivos específicos:

- Investigar diferentes modelos de Previsão Numérica do Tempo e seus resultados para as quatro localidades;
- Analisar o uso de diferentes variáveis de entrada para a construção do modelo *machine learning*;
- Investigar os melhores hiperparâmetros necessários para a melhora do resultado do modelo;
- Comparar a previsão da velocidade do vento no horizonte de 24 horas originada pelo modelo de *machine learning* com os modelos de Previsão Numérica do Tempo e o método da Persistência.

3 REVISÃO BIBLIOMÉTRICA

3.1 INTRODUÇÃO

Para a revisão bibliométrica foi utilizada a linguagem de programação R e a biblioteca *Bibliometrix*, a qual permite uma análise mais generalizada do trabalho no quesito produção científica ao longo do tempo.

A base de dados utilizada para coleta das publicações foi o *Scopus*, uma referência para a busca de produção científica no mundo. A extração dos dados dos artigos data do dia 5 de dezembro de 2024, gerando um documento em .csv que é lido pelo *Bibliometrix*.

A análise bibliométrica é fundamental para mapear o assunto do trabalho no mundo da produção científica, ou seja, analisar o crescimento do interesse no assunto através das produções científicas publicadas através de palavras-chaves específicas.

3.2 SITUAÇÃO DO ASSUNTO NA LITERATURA

Para situar o assunto nos artigos publicados, foram utilizadas as palavras chaves *Wind Power Forecasting* ou *Wind Speed Forecasting* e *Gradient Boosting* ou *SVR* ou *Neural Networks* ou *Machine Learning*.

Os artigos encontrados com a palavra chave *Wind Speed Forecasting* se concentravam muito em aplicações para energia eólica, por isso foi acrescentado *Wind Power Forecasting* para mapear uma maior quantidade de artigos, visto que esses trabalhos que fazem previsão de geração eólica possuem insights valiosos relacionando previsões numéricas do tempo e algoritmos de *machine learning*, e que também usam a variável de velocidade do vento e outras variáveis meteorológicas para munir seus modelos de previsão.

Foi utilizado a ferramenta de filtragem do *Scopus* para obter-se artigos:

- a partir de todas as combinações possíveis das palavras-chaves;
- de todos os anos disponíveis;
- somente na língua inglês;
- somente publicada em jornais, sendo artigos ou revisões;
- limitados às áreas de engenharia, energia, ciência da computação, matemática, ciência do meio-ambiente e economia e finanças.

A partir desses critérios, foi feito o *download* do arquivo e a importação no

Bibliometrix para que pudesse ser feita a análise. Na Tabela 1 encontra-se o resumo dos resultados encontrados.

Tabela 1. Síntese de resultados da busca de produções científicas na base *Scopus*.

Intervalo de tempo	1996 - 2024
Total de fontes	259
Total de documentos	1111
Quantidade de autores	2456
Taxa de crescimento anual de publicações	5,71%

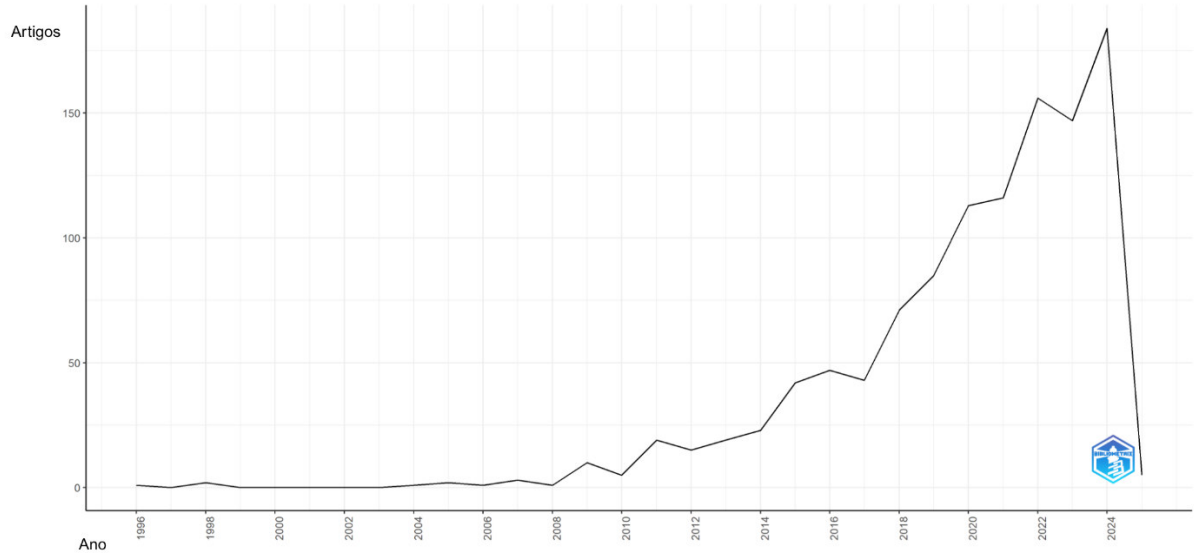
Fonte: Elaborado pelo autor (2025).

Logo, a partir desses dados foram feitas as análises de uma forma mais detalhada em cima das pesquisas realizadas, das novas tecnologias aplicadas, das perspectivas futuras e do desenvolvimento da área até o presente momento.

3.3 ANÁLISE QUANTITATIVA DAS PRODUÇÕES CIENTÍFICAS SOBRE O TEMA

Tendo as informações base da pesquisa até agora, é possível adentrar mais detalhadamente nos resultados gerados pelo *Bibliometrix*.

Figura 1. Produção científica anual.



Fonte: Adaptado de Base *Scopus*, ferramenta *Bibliometrix* (2025).

Primeiramente, vê-se a taxa crescente de publicações através da Figura 1, atingindo o pico no ano de 2024. Ao total são 184 documentos publicados.

Em relação a produção científica por país, observa-se com a Tabela 2 que o Brasil é o 5° país que mais publicou sobre o tema, com 104 documentos. A China lidera o ranking com 2439 publicações.

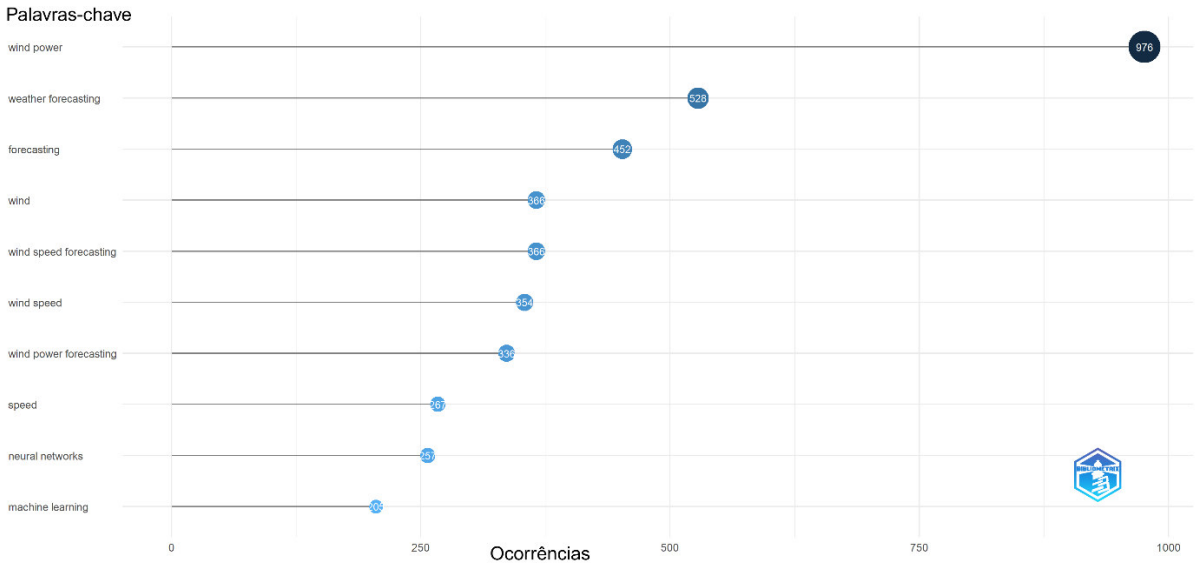
Tabela 2. Número de pulicações científicas por país.

País	**Publicações**
China	2439
Índia	301
Eua	150
Irã	107
Brasil	104
Austrália	83
Espanha	83
Turquia	83
Canadá	69
Malásia	69

Fonte: Base *Scopus*, ferramenta *Bibliometrix* (2025).

No que tange às palavras-chaves, o gráfico "*Keyword Plus*" feito pelo Bibliometrix (Figura 2), mostra as 10 palavras-chaves mais utilizadas em todo o banco de dados. Há a aparição de 4 das 6 palavras-chaves utilizadas para o filtro da busca: *Wind Speed Forecasting*, *Wind Power Forecasting*, *Neural Networks* e *Machine Learning*.

Figura 2. Palavras-chaves mais frequentes.

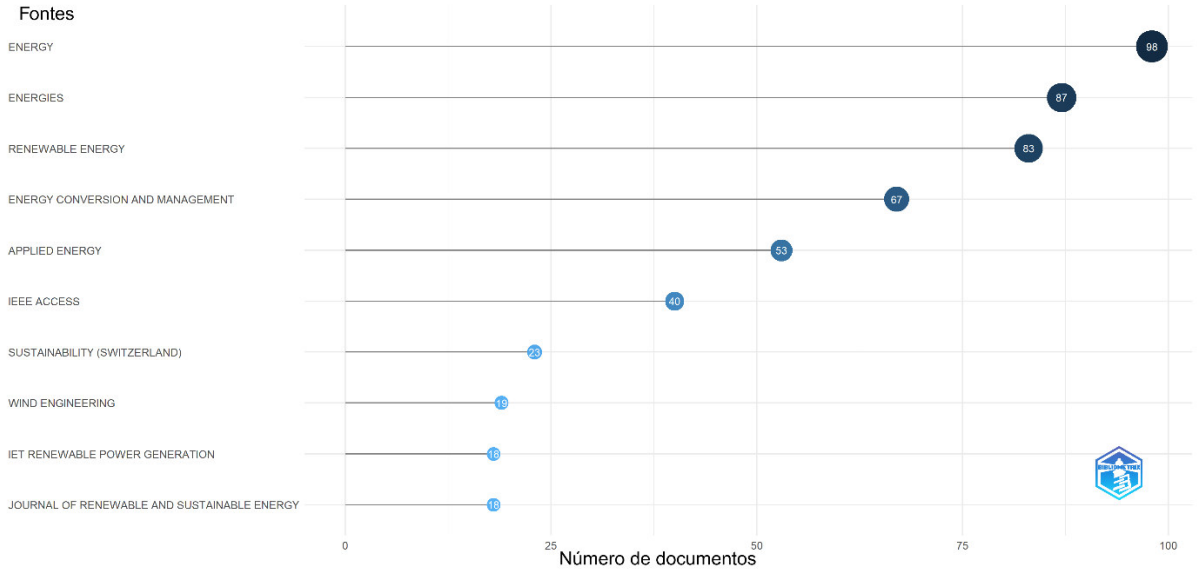


Fonte: Adaptado de Base *Scopus*, ferramenta *Bibliometrix* (2025).

Seguindo essa linha de raciocínio, as fontes mais relevantes sobre o tema são expostas na Figura 3. A fonte "*Energy*" foi a que teve mais artigos publicados, 98. Em

seguida tem-se a "*Energies*" com 87 e a "*Renewable Energy*" com 83, sendo essas as 3 mais relevantes.

Figura 3. Fontes mais relevantes.



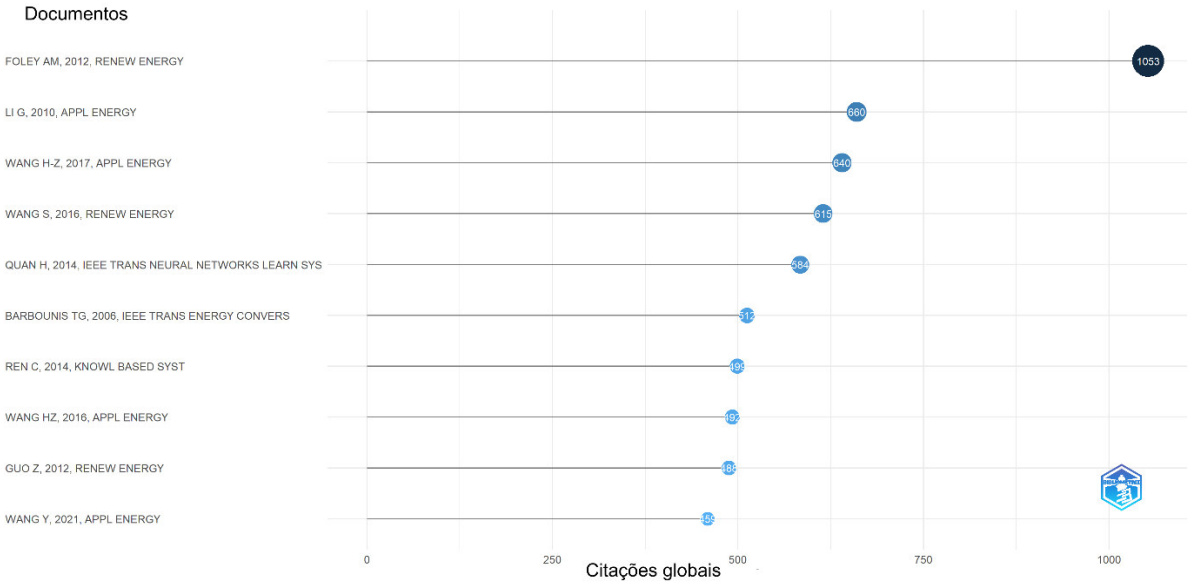
Fonte: Adaptado de Base *Scopus*, ferramenta *Bibliometrix* (2025).

3.4 ANÁLISE QUALITATIVA DOS DOCUMENTOS MAIS CITADOS

A partir do relatório gerado pelo *Bibliometrix* tem-se, ainda, as publicações mais citadas, para as quais é imprescindível ter um olhar mais atento para identificar o porquê e observar suas metodologias e conclusões.

Como pode-se observar na Figura 4, o artigo mais citado é de Foley AM (2012), publicado na *Renewable Energy* em 2012. Ele foi citado 1053 vezes, e tem como título "Métodos atuais e avanços na previsão da produção de energia eólica".

Figura 4. As 10 publicações mais citadas.



Fonte: Adaptado de Base *Scopus*, ferramenta *Bibliometrix* (2025).

A partir da Figura 4, a tabela 3 foi criada, nela contêm o primeiro autor do artigo, o seu título, o ano de publicação, se foi publicado em jornal ou revista e o total de citações.

Tabela 3. Resumo das principais informações das 10 publicações mais citadas.

Autor	Título	Ano	Jornal	Citações
Aoife M. Foley	Current methods and advances in forecasting of wind power generation	2012	Renewable Energy	1053
Gong Li	On comparing three artificial neural networks for wind speed forecasting	2010	Applied Energy	660
Huai-zhi Wang	Deep learning based ensemble approach for probabilistic wind power forecasting	2017	Applied Energy	640
Shouxiang Wang	Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method	2016	Renewable Energy	615
Hao Quan	Short-Term Load and Wind Power Forecasting Using Neural Network-Based Prediction Intervals	2014	IEEE	584
T.G. Barbounis	Long-term wind speed and power forecasting using local recurrent neural network models	2006	IEEE	512
Chao Ren	Optimal parameters selection for BP neural network based on particle swarm optimization: A case	2014	Knowledge-Based Systems	499

	study of wind speed forecasting			
H. Z. Wang	Deep belief network based deterministic and probabilistic wind speed forecasting approach	2016	Applied Energy	492
Zhenhai Guo	Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model	2012	Renewable Energy	488
Yun Wang	A review of wind speed and wind power forecasting with deep neural networks	2021	Applied Energy	459

Fonte: Adaptado de Base *Scopus*, ferramenta Bibliometrix (2025).

Em Foley et al. (2012), é feita uma análise comparativa dos modelos de previsão, categorizados em análise de séries temporais históricas e variáveis de entrada utilizando modelos de Previsão Numérica do Tempo, trazendo uma abordagem de modelos físicos, estatísticos e de aprendizado de máquina. Tem-se como conclusão que o uso dos modelos híbridos, que integram mais de uma abordagem de previsão, entregam um resultado melhor em termos de acurácia e confiabilidade, reduzindo o erro associado.

Em Li Gong et al. (2010), foi empregado três tipos de Redes Neurais Artificiais – *Feedforward Back-Propagation* (FFBP), *Radial Basis Function* (RBF) e *Adaptative Linear Element* (ADALINE) – para a previsão do vento no horizonte de previsão de 1 hora a frente usando dados históricos de dois locais de observação em Dakota do Norte, nos Estados Unidos. A análise focou-se no impacto dos diferentes parâmetros de entrada e taxas de aprendizado dos algoritmos, avaliando-se as métricas de erro médio absoluto (*MAE*), raiz do erro quadrático médio (*RMSE*) e erro percentual médio absoluto (*MAPE*). Os resultados indicaram que a escolha do modelo de Rede Neural afeta significativamente o desempenho da previsão, sem que um único modelo performe consistentemente melhor que os outros em diferentes conjuntos de dados e métricas, destacando, assim, a necessidade de um método robusto que combine previsões de vários modelos diferentes para poder tratar as inconsistências na sua seleção.

Em Wang et al. (2017) é feita uma abordagem para previsão probabilística de energia eólica (WPF) que reúne Redes Neurais Convolucionais (*CNN*) com Regressão Quantílica (QR). Utilizando dados históricos do parque eólico Milky Way em Shandong, na China, as séries de entrada são decompostas em várias frequências para aumentar a precisão da previsão. É concluído que o modelo proposto supera os métodos tradicionais de previsão, obtendo melhorias de até 62,67% na pontuação de intervalo e 50,84% no erro médio de cobertura, destacando a eficácia da aprendizagem profunda (*Deep Learning*) na captura dos comportamentos complexos dos dados de energia eólica para melhorar a previsão.

Em Wang Shouxiang et al. (2016), a previsão da velocidade do vento é composta

por uma Decomposição de Modo Empírico de Conjunto (EEMD) com uma rede neural de Algoritmo Genético-Propagação de Retorno (GA-BP). Os estudos foram realizados baseados em um parque eólico na Mongólia e concluiu-se que houve melhora na precisão da previsão devido ao método utilizado, alcançando um erro percentual absoluto médio (*MAPE*) de 6,82% para a previsão de curtíssimo prazo e 8,08% para a previsão de curto prazo, superando os métodos de previsão tradicionais.

Em Quan Hao et al. (2014), é implementado um método baseado em Rede Neural para a construção de intervalos de previsão com o intuito de quantificar as possíveis incertezas associadas às previsões de previsão de carga eólica. As demandas elétricas de Cingapura e Nova Gales do Sul, na Austrália, assim como a geração de energia eólica do parque eólico Capital, são usadas para validar o método proposto. Os resultados mostram que o método pôde construir intervalos de previsão de maior qualidade para previsões de carga e geração de energia eólica em um curto espaço de tempo.

Em Barbouis T.G. et al. (2006), foram desenvolvidos modelos avançados de previsão da produção de energia eólica utilizando modelos determinísticos de micro e mesoescala juntamente com técnicas de *Model Output Statistics* e inteligência artificial. A metodologia envolveu o treinamento de várias arquiteturas de redes neurais, avaliando o *MAE* e o *RMSE*. Foram observadas melhorias significativas na precisão da previsão, demonstrando a eficácia dos modelos propostos na tradução de dados meteorológicos em previsões confiáveis de energia eólica.

Em Ren Chao et al. (2014), utiliza-se um método chamado IS-PSO-BP (*Improved Particle Swarm Optimization - Back Propagation*) para prever a velocidade do vento usando modelos de seleção de dados longitudinais e laterais. Esse método, IS-PSO-BP, é comparado com as redes neurais *Back-Propagation* tradicionais e os modelos *Autoregressive Integrated Moving Average* (*ARIMA*). Os resultados indicaram que, mesmo que o modelo BP tenha um desempenho melhor em média, o modelo proposto pode superá-lo em condições de parâmetros ideais, demonstrando impactos significativos da dimensão de entrada na precisão da previsão.

Em Wang H.Z. et al. (2016), é feita uma abordagem com modelos híbridos para a previsão da velocidade do vento, integrando as redes neurais artificais com modelos de mesoescala de quinta geração para aumentar a precisão da previsão. A metodologia emprega uma combinação de técnicas de mineração de dados e extração de recursos para lidar com a natureza volátil das séries de velocidade do vento, utilizando simulações repetidas 20 vezes para incrementar a confiabilidade do modelo. Os resultados mostraram que o método proposto superou significativamente os modelos tradicionais, como o *Autoregressive Moving Average* (*ARMA*), demonstrando uma nova abordagem eficaz para previsões determinísticas e probabilísticas de energia eólica.

Em Guo Zhenhai et al. (2012), é avaliado um modelo modificado de rede neural *Feed-Forward* baseado em uma Decomposição de Modo Empírico (M-EMDFNN) com o fito de melhorar a precisão da previsão da velocidade do vento. Os dados são decompostos em várias funções, usando a técnica de decomposição de modo empírico, seguida pela previsão de cada função com uma rede neural *feed-forward* (*FNN*). A comparação foi feita entre os modelos M-EMDFNN com os modelos tradicionais *FNN* e EMD-FNN e concluiu-se que o modelo M-EMDFNN performa melhor no que tange a previsão de velocidade do vento.

Em Wang Yun et al. (2021), é investigada a previsão de potência e velocidade do vento utilizando uma combinação de tradicionais e avançados tipos de modelos de aprendizagem profunda (*Deep Learning*). É feito um pré-processamento de dados por meio de técnicas de processamento de sinais, extração de recursos usando vários algoritmos e otimização de modelos com abordagens híbridas, incluindo *LSTM* e *CNNs*. A análise se concentra no desempenho desses modelos e conclui que os modelos híbridos aumentam significativamente a precisão da previsão se comparados com modelos tradicionais.

Portanto, através da leitura e análise dos modelos utilizados pelos 10 artigos mais citados gerados pelos filtros de pesquisa no *Scopus*, foi visto que há uma grande concentração nos estudos de Redes Neurais dos mais diferentes tipos para a previsão de energia eólica, seja velocidade do vento ou potência. Os dados variam de procedência meteorológica como o *Global Forecast System* (*GFS*) e o *European Centre for Medium-Range Weather Forecasts* (*ECMWF*), ou de dados de turbinas reais, ou uma combinação de dados de turbinas e dados de previsão. Em todos, ainda, é comum o uso de erros como o *MAE* ou o *RMSE* para a análise comparativa.

4 REFERENCIAL TEÓRICO

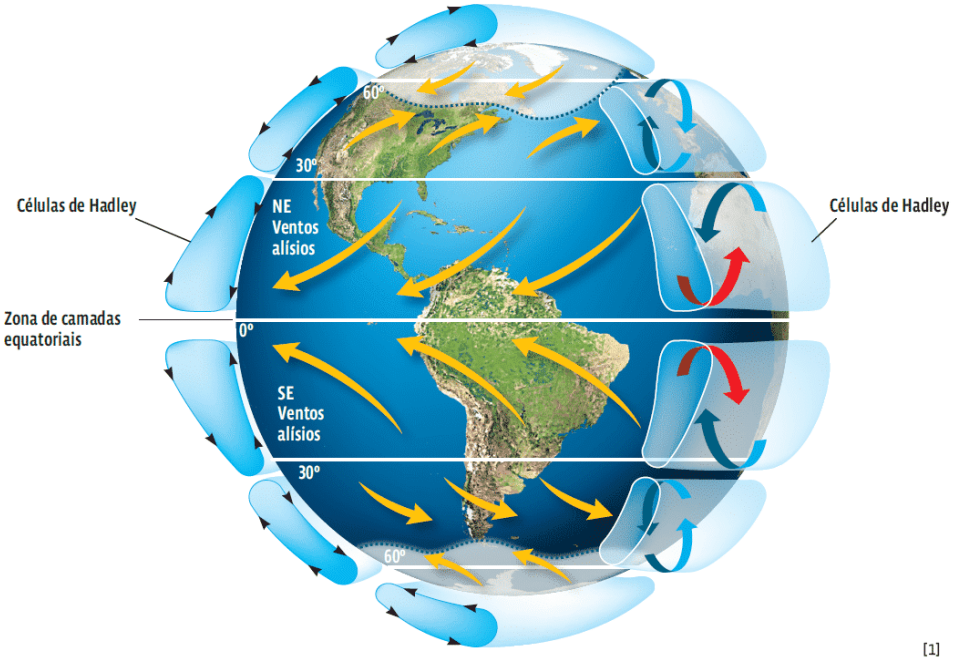
4.1 O VENTO E SUAS CARACTERÍSTICAS

O vento é conhecido por ser uma variável de difícil previsão devido ao seu comportamento estocástico (LEI, M et al., 2009). A definição de algo estocástico segundo a estatística é algo que depende ou resulta de uma variável aleatória (PAPOULIS, A.; PILLAI, S. U, 2002). O vento combina fenômenos determinísticos com variabilidades aleatórias e turbulentas em diferentes escalas. Essa natureza híbrida e comportamento caótico tornam a sua previsão extremamente desafiadora, exigindo cada vez mais melhorias dos modelos probabilísticos existentes e sendo um desafio até mesmo para modelos com inteligência artificial e de séries temporais sofisticadas (WANG, H. et al., 2019).

Os ventos são originados devido aos gradientes de pressão atmosférica. O ar mais denso tende a descer e o ar menos denso tende a subir, criando assim movimento de áreas de alta para áreas de baixa pressão que é ocasionado devido ao aquecimento desigual da Terra. Além disso, o vento também sofre influência da rotação da Terra devido ao efeito Coriolis, afetando esse gradiente de pressão e fazendo com que o vento não sopre diretamente das zonas de alta para as de baixa pressão (AHRENS, 2018).

Para o Brasil, há destaque para os ventos alísios, que são ventos que sopram de sudeste no Hemisfério Sul e são influenciados pelo Efeito Coriolis e pelo gradiente de temperatura das zonas da Linha do Equador em relação aos polos terrestres. Além deles, há também a influência dos ventos do oeste que afetam o Sul do país e a massa equatorial atlântica que está presente com mais força no Norte e no Nordeste. A Figura 5 ilustra esse comportamento.

Figura 5. Movimento das massas de ar na Terra.



Fonte: Braga (2022).

O vento também sofre grandes influências do relevo do local. Cidades que são cercadas por relevos montanhosos possuem uma circulação de ar diferente devido aos obstáculos presentes e a topografia do terreno.

Reduzindo a escala de análise, dentro do país possuímos diversos padrões regionais de circulação de vento, como as relações entre vento terral e maral. As brisas marítmas e terrestres são originadas devido ao aquecimento desigual entre a superfície terrestre e o oceano.

Durante o dia, quando a superfície terrestre se aquece mais rapidamente que o oceano, o ar sobre a terra fica menos denso e sobe, enquanto que o ar mais denso do oceano tende a ficar por baixo e se direcionar a terra devido ao gradiente de pressão, a isso apelidados de brisa marítima. Já durante a noite podemos observar o efeito contrário, e apelidamos de brisa terrestre (AHRENS, 2018).

A influência de regiões montanhosas pode ser observada em montanhas e vales, onde as encostas das montanhas se aquecem ou se resfriam mais rapidamente que o vale e geram ventos ascendentes e descendentes que podem ser montanha-vale ou vale-montanha a depender se é dia ou noite (AHRENS, 2018).

Ademais, além da influência da diferença de pressão e temperatura e do relevo, a intensidade do vento pode ser intensificada pela topografia do terreno. As condições de circulação do vento mudam drasticamente de acordo com a complexidade da topografia local. Isso remete à classificar os terrenos em simples ou complexos, de tal forma que simples seria um terreno com baixas irregularidades e complexo seria um terreno extremamente irregular, tendo assim efeito maior no fluxo de ar do local (MANWELL; MCGOWAN; ROGERS,

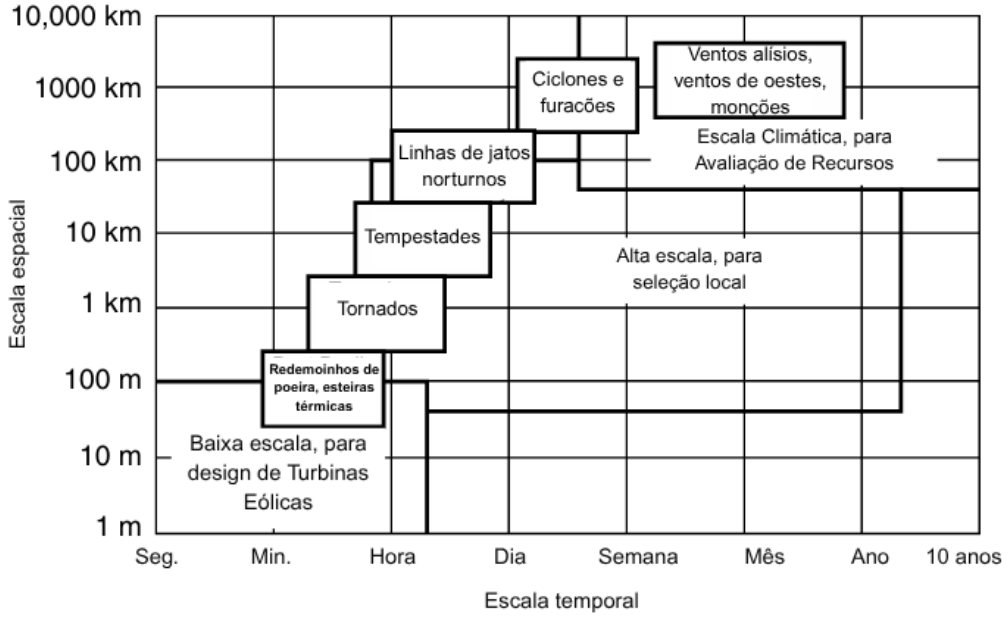
2009).

A velocidade do vento também varia fortemente conforme a altura em que a mesma é analisada. A superfície terreste exerce uma força de atrito sobre o vento de tal forma que o fluxo seja retardado e, a medida em que a altura em relação ao solo aumenta, esse efeito torna-se mais desprezível (MANWELL; MCGOWAN; ROGERS, 2009).

A viscosidade do ar sofre influências consideráveis na camada limite atmosférica e depende da rugosidade do terreno. A tensão de cisalhamento na camada limite é o produto da viscosidade pelo gradiente de velocidade na direção perpendicular à superfície, isso influencia na energia total do fluxo de ar e diminui a sua velocidade. Se a altitude aumentar, esse efeito de superfície que influencia no gradiente de velocidade dos fluxos de ar tende a ser desprezível e a velocidade do vento tende a tornar-se uniforme (MANWELL; MCGOWAN; ROGERS, 2009).

Como ilustrado na Figura 6, os movimentos atmosféricos podem variar no tempo (de segundos a meses) e no espaço (de centímetros a milhares de quilômetros). As variações no tempo podem ser anuais, interanuais, diurnas ou de curto prazo.

Figura 6. Movimentos atmosféricos variando no tempo e espaço.



Fonte: Adaptado de MANWELL; MCGOWAN; ROGERS (2009).

As variações interanuais são em escalas maiores que um ano e são importantes principalmente no contexto de produção de energia eólica, pois é necessário estimar como é o comportamento histórico do vento no local para poder traçar as metas de geração dos parques. De acordo com a meteorologia é necessário 30 anos de dados para determinar com fidelidade os valores de longo prazo de um clima local (MANWELL; MCGOWAN; ROGERS, 2009).

Ao diminuirmos a escala teremos as variações anuais, estas estão relacionadas à sazonalidade eólica. No Brasil, a temporada de altos ventos começa em torno de junho, com o

começo do inverno. Na europa os altos ventos também são no inverno, mas são no período oposto ao que ocorre no Brasil, devido à diferença de hemisfério. Estudos dizem que o verão é mais difícil de prever devido a maior incidência de radiação e uma maior quantidade de fenômenos atmosféricos (STENSRUD, 2007).

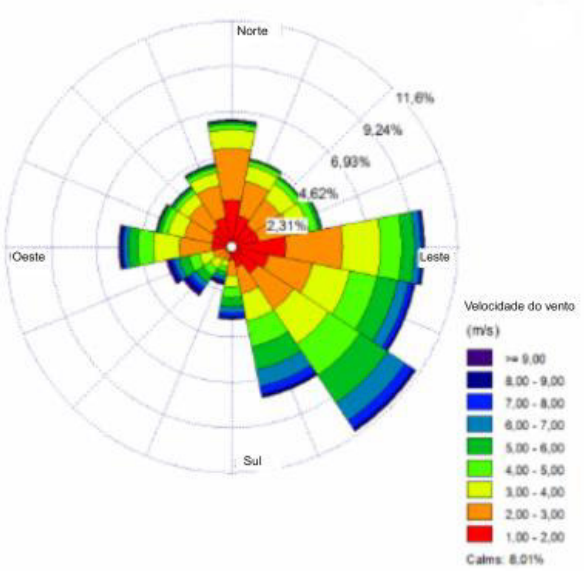
As variações diárias estão relacionadas à mudança da intensidade do vento ao longo das 24 horas. Como o aquecimento depende do sol, é normal a velocidade do vento aumentar ao longo do dia e diminuir durante a noite. Essa variação também depende da elevação do local (MANWELL; MCGOWAN; ROGERS, 2009).

As variações de curto prazo incluem turbulências e rajadas de vento e normalmente essas medições são feitas em dados de 1 segundo ou mais, normalmente analisada por dados de anemômetros que são agregados em dados 10 minutais. Uma variabilidade acima do normal em relação a velocidade média histórica da localidade pode ser definida como turbulência (MANWELL; MCGOWAN; ROGERS, 2009).

Além das variações temporais, a intensidade do vento também varia conforme sua direção, sendo utilizada para isso a rosa dos ventos. A rosa dos ventos é uma representação gráfica que combina a frequência e a intensidade do vento em várias direções para um local específico, dividido em 16 setores (N, NE, L, SE, S, SW, O, NW, etc), normalmente associando as velocidades de acordo com a escala de Beaufort. A escala de Beaufort é uma medida de calmaria ou alta turbulência do vento, classificada de acordo com a sua velocidade (CHAMMI REDDY et al., 2016).

Para exemplificação, na Figura 7 é ilustrada a rosa dos ventos para a estação SBGL na Baía da Guanabara, Rio de Janeiro, feita no estudo de de Oliveira-Júnior (2017).

Figura 7. Rosa dos ventos da estação meteorológica SBGL, localizada na Baía de Guanaraba (RJ), para o período de 2003-2013.



Fonte: Adaptado de de Oliveira-Júnior (2017).

Diferentes estudos na região Nordeste indicam que a direção predominante dos ventos é leste, com variações para sudeste e nordeste, principalmente ao longo do litoral.

Para exemplificar, um estudo realizado em Cruz das Almas (BA) revelou que a direção predominante foi de SE (entre 90° e 180°) em mais de 88% das ocorrências, com velocidades médias em torno de 2,7 m/s (brisa leve) (CONCEIÇÃO; SOUZA; BARROS, 2011).

Em Petrolina (PE), a direção predominante foi sudeste (SE) (aproximadamente 105°–135°), com uma velocidade média de 8,4 m/s. Essa classificação é considerada excelente para a energia eólica, de acordo com o NREL (SANTOS et al., 2019).

O estudo da direção do vento não só ajuda a identificar padrões locais, mas também é fundamental para prever sua velocidade. Em Souza et al., 2016, modelos de séries temporais, como *ARIMA*, *Holt-Winters* e redes neurais, foram aplicados em cidades como Fortaleza, Parnaíba e São Luís. Utilizando a direção dominante e variáveis meteorológicas correlacionadas como variáveis, esses modelos mostraram-se eficazes para estimar médias mensais, com erros percentuais variando entre 8,7% e 10,5%. A rede neural apresentou um erro de aproximadamente 8,7% em Fortaleza, enquanto o modelo *ARIMA* registrou um erro de cerca de 9,7% em Parnaíba.

A utilização de técnicas e análises fundamentadas na representação da rosa dos ventos são essenciais para o planejamento de parques eólicos, previsão da produção de energia e otimização das operações das turbinas em tempo real (RAMACHANDRAN et al., 2020).

4.2 ESTAÇÕES METEOROLÓGICAS

Uma estação meteorológica é uma instalação munida de instrumentos de medição e sensores, localizada em uma área estratégica de uma região, para medição de diversas variáveis ambientais, tais como a temperatura, a pressão atmosférica, a velocidade do vento, etc (RENKEER, 2025). O seu objetivo principal é coletar dados meteorológicos e climáticos para monitoramento do clima e previsão das condições futuras do tempo de uma região a curto e médio prazo. O estudo climático histórico também permite caracterizar o clima de uma região, identificando padrões.

As aplicações das estações meteorológicas reiteram a sua importância. No setor agrícola, antecipar o conhecimento das condições climáticas significa tomada de decisões mais assertivas com relação a fases do cultivo como a programação de plantio ou colheita. No setor de segurança, pode-se prever enchentes e prevenir acidentes, enquanto que para a

navegação e aviação os dados sobre o vento, visibilidade, tempestades e chuvas das estações são fundamentais para orientar os voos e as navegações marítimas.

Existem dois tipos de estações meteorológicas, as convencionais e as automáticas. As estações convencionais são mais manuais e necessitam de um observador para a coleta dos dados. Nela são utilizados equipamentos analógicos como termômetros. Já as estações meteorológicas automáticas (EMAs) possuem essa coleta de dados de uma maneira automatizada, sem a necessidade de um observador. Isso é possível através de sensores que operam através de sinais elétricos que são captados por um sistema de aquisição de dados (*datalogger*) que é responsável pelo armazenamento e o processamento das informações. Normalmente esses dados são coletados a cada minuto, e são integralizados na agregação horária para serem transmitidos (INMET, 2011).

No Brasil existem várias instituições que operam redes de estações meteorológicas para monitoramento do clima, a principal delas é o Institudo Nacional de Meteorologia (INMET). O INMET foi criado em 18 de novembro de 1909, inicialmente como a Diretoria de Meteorologia e Astronomia, vinculada ao então Ministério da Agricultura, Indústria e Comércio. Esse órgão adotou o nome INMET em 1992 e atualmente opera diretamente sob o Ministério da Agricultura e Pecuária (MAPA) (BRASIL, 2024).

Hoje, o INMET possui mais de 700 estações automáticas e convencionais em todo o país. As estações convencionais fazem medições três vezes ao dia (0 UTC, 12 UTC e 18 UTC) enquanto que as automáticas coletam os dados a cada minuto e estes são agregados de forma horária para sua disponibilização (INMET, 2011).

4.3 PREVISÃO NUMÉRICA DO TEMPO (NWP)

4.3.1 Contexto histórico

Do inglês *Numerical Weather Prediction* (*NWP*), a Previsão Numérica do Tempo é um método de previsão que emprega equações que descrevem os fluidos e os fluxos atmosféricos (PU; KALNAY, 2018). Essas equações são não-lineares, não permitindo assim uma solução analítica e sendo necessário o uso de técnicas para solução numérica em função do tempo.

Com o avanço computacional, as Previsões Numéricas do Tempo se tornaram cada vez mais potentes em modelar e fazer previsões do clima. Os precursores dos modelos numéricos para a previsão diária do clima foram o supercomputador da *National Oceanic And*

Atmospheric Administration (*NOAA*), nos Estados Unidos, e o *European Centre for Medium-Range Weather Forecasts* (*ECMWF*), na Europa (NOAA, 2007; LYNCH, 2010).

O objetivo de uma *NWP* é resolver as equações diferenciais originadas dos modelos que explicam como a atmosfera local é governada através dos seus fluxos e complexidades. Antes dessas equações serem resolvidas numericamente por supercomputadores, Lewis Fry Richardson em 1922 tentou resolvê-las a mão. Ele dividiu a região de interesse em quadrados, como se fosse uma malha, leu quais eram as condições atmosférica usando uma interpolação manual e mesmo com uma metodologia minuciosa a sua previsão foi uma queda de 145 mb em apenas 6 horas na pressão atmosférica, algo fisicamente impossível (PU; KALNAY, 2018).

Isso gerou uma frustração entre os apoiadores do método e o assunto só voltou a tona décadas depois em 1950 com Jule Charney. Utilizando equações de vorticidade potencial barotrópica e simplificando a atmosfera para uma camada única e sem variação vertical, Charney conseguiu bons resultados e esse feito marcou o início da previsão numérica operacional moderna (PU; KALNAY, 2018).

Desde essa primeira tentativa bem sucedida, a Previsão Numérica do Tempo pôde sentir uma evolução exponencial juntamente com o avanço computacional e avanços na ciência atmosférica. Na década de 1970, dados observacionais e modelos globais multi-níveis começaram a ser inseridos nas simulações e a melhorar os resultados obtidos.

A partir de 1990, dados de satélites revolucionaram a qualidade das previsões. No processo de evolução dos modelos *NWP*, veio a assimilação de dados (*data assimilation*). A assimilação de dados é a combinação de observações atmosféricas, como medições de estações meteorológicas e satélites, com o estado atual previsto por um modelo. Isso é feito para que a condição inicial seja mais precisa (PU; KALNAY, 2018).

No Brasil, o Instituto Nacional de Meteorologia (INMET) e o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC/INPE) são a referência para a Previsão Numérica do Tempo operacional do país. O CPTEC, fundado em 1994, foi pioneiro na América do Sul com a implementação de modelos de mesoescala como o *ETA*, o *BRAMS* e o *WRF*.

Atualmente o Brasil integra redes internacionais de observação, mas ainda sofre com as problemáticas relacionadas a infraestrutura computacional e qualidade de observações que se dificultam em um país de tamanho continental. Ainda assim, a *NWP* é uma ferramenta crucial para os setores como da agricultura, energia e da defesa civil, sendo usado cada vez mais inteligência artificial e assimilação de dados para aprimoração das previsões temporais (MARQUES, 2024).

4.3.2 Características dos modelos de *Numerical Weather Prediction*

Existem muitos modelos de *NWP*, cada um com diferentes escalas temporais e espaciais simuladas de formas diferentes. Os objetivos podem ser diversos, de identificar eventos meteorológicos especícifos como frentes frias que podem resfriar uma região e se estender por semanas, como a previsão de fenômenos meteorológicos sensíveis para a manutenção da vida humana como furacões, tempestades, etc (PU; KALNAY, 2018).

As escalas espaciais resolvidas pelas equações numéricas diminuem cada vez mais de tal forma que alguns modelos podem ter uma resolução horizontal suficiente para simular processos de mesoescala (PU; KALNAY, 2018).

Apesar da redução das escalas espaciais, existem fenômenos que possuem uma escala espacial menor que a resolução do modelo numérico, logo não podem ser representados. Um exemplo de fenômeno que não pode ser representado tão facilmente é a formação de nuvens, que pode influenciar a irradiação na terra, a precipitação no local, a umidade, entre outras mais variáveis atmosféricas que se comportariam diferente com um céu limpo, sem nuvens. Existem parametrizações que são necessárias para que esses fenômenos possam ser capturados pelos modelos (STENSRUD, 2007).

As equações que governam a evolução da atmosfera são baseadas na segunda lei de Newton ou conservação do momento (três equações para x, y e z), na equação da continuidade ou conservação da massa, a equação de estado para gases ideais, a primeira lei da termodinâmica ou da conservação de energia e a equação de conservação para a massa de água, totalizando 7 equações (LYNCH, 2017).

Como foi explicitado anteriormente, essas equações formam um sistema de equações diferenciais parciais, não lineares e acopladas, cuja solução analítica é impraticável devido à complexidade e à variabilidade espacial e temporal dos processos atmosféricos e para tornar viável a resolução dessas equações, utilizam-se métodos numéricos (LYNCH, 2017).

O método das diferenças finitas é um dos mais utilizados. Esse método consiste em discretizar o domínio contínuo da atmosfera em uma grade de pontos espaciais tridimensionais e temporais. A ideia é que as derivadas parciais existentes sejam aproximadas para valores dos pontos existentes na malha, tornando as equações diferenciais em um sistema de equações algébricas que podem ser resolvidas iterativamente por computadores.

Além da discretização, é necessário especificar as condições de contorno que definem o comportamento das variáveis meteorológicas nas bordas do modelo. Em modelos globais como o *ECMWF*, essas condições são naturais visto que o modelo abrange todo o globo. Já para modelos regionais, as condições laterais devem ser especificadas ao longo do

tempo, normalmente vindo de um modelo global (MESINGER, 2005).

Além das condições de contorno laterais, existem também as condições de contorno inferiores e superiores. Essas consideram a interação com a superfície, como o relevo e a topografia do terreno, o solo, a temperatura da superfície, entre outras variáveis (RAMS, 2003).

4.3.3 Horizontes de previsão

É conhecido que o sistema atmosférico é caótico, e isso dificulta a sua previsão, visto que esta é sensível às condições iniciais e variáveis de entrada.

Nesse contexto, é necessário introduzir o conceito de horizonte de previsão. O horizonte de previsão é qual o tempo futuro que é previsto, ou seja, 24 horas se é desejado prever o dia de amanhã. Esse intervalo costuma ser horário, pois os modelos numéricos fornecem saídas nesse formato (MONTEIRO et al., 2009).

A distinção entre os horizontes de previsão não é universal e varia conforme as fontes. Segundo CHANG (2014), os horizontes podem ser divididos em: curtíssimo prazo (alguns minutos até 1 hora), essencial para a gestão operacional dos parques eólicos; curto prazo (1 hora até algumas horas), importantes para planejamento energético de parques e ajustes em tempo real (ALKESAIBERI; HARROU; SUN, 2022); médio prazo (algumas horas a 1 semana), que indicam tendências meteorológicas (HANIFI et al., 2020); e longo prazo (1 semana a 1 ano, ou mais), voltadas para planejamento de manuntenções e planejamento de geração a longo prazo e previsões sazonais.

Na maioria dos casos os erros das previsões tendem a aumentar em consonância com o aumento do horizonte de previsão (FOLEY et al., 2011).

4.3.4 Tipos de modelos: físicos, estatísticos e híbridos ou globais e regionais

4.3.4.1 Modelos físicos, estatísticos e híbridos

As previsões podem ser classificadas em três tipos: físicas, estatísticas e híbridas. Os modelos físicos utilizam equações diferenciais para simular a atmosfera e fenômenos meteorológicos. Os modelos estatísticos, como *ARMA* (*Auto-Regressive Moving Average*) e

ARIMA (*AutoRegressive Integrated Moving Average*), baseiam-se em dados históricos para prever tendências futuras. A comparação costuma ser com o modelo de Persistência, que serve de referência para previsões de curto prazo, assumindo que o evento futuro será igual ao observado no passado (DUPRÉ et al., 2020).

Modelos híbridos combinam os modelos *NWP* e os estatísticos para aprimorar a precisão. Uma abordagem híbrida clássica é o *MOS* (*Model Output Statistics*), que ajusta as previsões físicas com correções estatísticas derivadas de erros anteriores. Tais modelos vêm ganhando popularidade, pois permitem refinar previsões com base em condições locais (HANIFI et al., 2020; DUPRÉ et al., 2020).

4.3.4.2 Modelos globais e regionais

Os modelos de previsão podem ser globais ou modelos de mesoescala, que são os modelos regionais. Esses tipos de modelos abragem uma área limitada, pois eles visam representar fenômenos específicos daquele local e para isso é feito um *downscalling* – redução da escala temporal e espacial do modelo. Modelos de mesoescala podem levar em consideração até mesmo características do terreno para poderem capturar fenômenos específicos (PIELKE, 2002).

Como já foi exposto anteriormente, é importante lembrar que as condições de contorno laterais para as equações dos modelos regionais é originada dos resultados de modelos globais. Em outras palavras, *AROME*, que é um modelo de mesoescala com um foco maior em previsões para a França e países próximos, usa como condições de contorno os resultados das simulações rodadas previamentes pelo modelo global *ARPEGE*. Isso provoca uma dependência entre um modelo e outro, então mesmo que seja feito *downscalling*, no modelo de mesoescala ainda existe a dependência de um modelo global. (PIELKE, 2002)

Dentre os mais famosos e utilizados modelos *NWP* globais estão o *Global Forecast System* (*GFS*), desenvolvido pelos Estados Unidos e operado pela *NOAA* desde 1980; o *ECMWF*, modelo global europeu em operação desde 1979 e reconhecido por sua alta precisão; e o francês *ARPEGE* (*Action de Recherche pour la Petite Echelle et la Grande Echelle*), desenvolvido pela *Météo-France* e introduzido em 1995. Outros modelos globais importantes incluem o *UK Met Office Unified Model*, do Reino Unido (desde 1991), e o *ICON* (*Icosahedral Nonhydrostatic Model*), da Alemanha, lançado em 2015.

Em relação aos modelos regionais, destacam-se o *WRF* (*Weather Research and Forecasting Model*), criado por uma colaboração de instituições norte-americanas como *NCAR* e *NOAA* em 2000, amplamente usado para pesquisa e previsão operacional de alta

resolução; o *HIRLAM*, desenvolvido por um consórcio de países nórdicos em 1985; e o *BRAMS* (*Brazilian developments on the Regional Atmospheric Modeling System*), criado no Brasil pelo INPE e CPTEC no final dos anos 1990, voltado para previsão regional e estudos de qualidade do ar.

4.4 MODELOS DE APRENDIZADO DE MÁQUINA

4.4.1 Origem dos algoritmos de *machine learning*

Os algoritmos de *machine learning* remontam às primeiras décadas do início da computação, sendo enraizado com base em estatística, teoria da informação e inteligência artificial. Criado por Frank Rosenblatt em 1958, um dos primeiros e mais conhecidos algortimos foi o *Perceptron*, considerado como o primeiro modelo de rede neural artificial (RUSSELL; NORVIG, 2013).

Na década seguinte os avanços foram vistos em novos algoritmos como o *k-Nearest Neighbors* e o uso de regressão linear começou a ser aplicado em classificações e predições (RUSSELL; NORVIG, 2013).

Nas últimas duas décadas do século XX, a ascensão das redes neurais foi a criação do algoritmo de retropropagação (*backpropagation*). Nesse período também surgiram outros métodos como o *Support Vector Machines* (*SVM*) e as Árvores de Decisão. A partir desses modelos é marcado o início da era do machine learning moderno, sendo tratado atualmente com grandes volumes de dados e supercomputadores (RUSSELL; NORVIG, 2013).

Já no começo do século XXI novos algorítimos como *Random Forest* e a técnica de *Ensemble Methods* expandiram as áreas de aplicabilidade exponencialmente devido ao poderio de previsão melhorado (RUSSELL; NORVIG, 2013).

4.4.2 Aprendizado supervisionado e não supervisionado

Dentre as abordagens de *machine learning*, destacam-se dois grandes grupos: o aprendizado supervisionado e o não supervisionado.

O aprendizado supervisionado se baseia em dados rotulados, ou seja, pares de entrada e saída conhecidos, com o objetivo de treinar modelos capazes de generalizar para

novos dados. Dentro dessa categoria, a regressão é usada para prever variáveis contínuas, como temperatura, velocidade do vento ou radiação solar, enquanto a classificação é aplicada em problemas com saídas discretas, como prever se haverá ou não chuva em determinada localidade. Modelos supervisionados incluem algoritmos como regressão linear, árvores de decisão, *XGBoost* e redes neurais profundas (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Por outro lado, o aprendizado não supervisionado lida com dados não rotulados, buscando encontrar estrutura ou padrões ocultos. Técnicas como *clustering* (ex: *K-means*, *DBSCAN*) e análise de componentes principais (PCA) são amplamente utilizadas para reduzir a dimensionalidade, detectar anomalias ou agrupar eventos estatisticamente semelhantes — como regimes sinóticos ou padrões de bloqueios atmosféricos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

4.4.3 Problemas de regressão e problemas de classificação

Dentro do aprendizado supervisionado, existe a necessidade de escolha entre um modelo de regressão ou classificação. Isso depende diretamente da natureza do problema em questão.

Problemas de regressão envolvem a previsão de valores contínuos, como estimar a temperatura máxima de amanhã com base em variáveis como pressão, umidade e radiação solar. Já a classificação é apropriada para situações em que se deseja prever categorias, como determinar se o dia será “ensolarado”, “chuvoso” ou “nublado”.

Em meteorologia, essa distinção é crítica: usar regressão em um problema de classes, ou vice-versa, pode comprometer a validade da previsão. Por exemplo, prever a velocidade do vento exige um modelo de regressão, enquanto prever a ocorrência de eventos extremos (como “vento forte” ou “vento fraco”) pode ser tratado como classificação.

Ademais, muitos sistemas híbridos modernos combinam ambas as abordagens para enriquecer as previsões — por exemplo, classificando padrões atmosféricos e, dentro de cada classe, aplicando regressões específicas para variáveis físicas.

4.4.4 Overfitting e underfitting

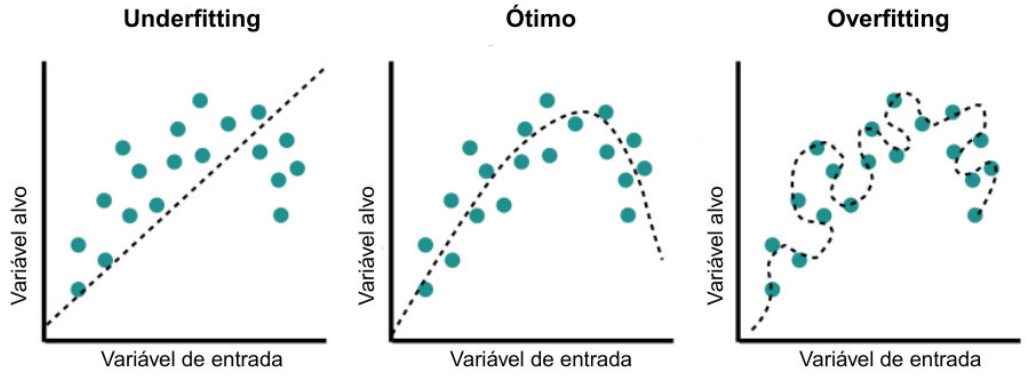
Outros dois fatores importantíssimos na hora de analisar um modelo de aprendizado de máquina é o *overfitting* e *underfitting*.

O *overfitting* ocorre quando um modelo aprende detalhes e ruído do conjunto de treino em excesso, perdendo a capacidade de generalizar para novos dados — típica situação de baixo enviesamento e alta variância (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).Isso é comum em modelos demasiado complexos ou quando os dados são escassos e ruidosos , resultando em performance ruim em ambientes reais.

Já o *underfitting* reflete um modelo simples demais para capturar o padrão dos dados, tendo um alto enviesamento; ele falha tanto nos dados de treino quanto nos de teste. Para mitigar esses problemas, usam-se técnicas que penalizam a complexidade do modelo e evitam o ajuste exagerado (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A Figura 8 ilustra a diferença entre um modelo com *underfitting*, com *overfitting* e em estado ideal.

Figura 8. Representação de modelos com *underfitting*, com valores otimizados e com *overfitting*.



Fonte: Adaptado de Rathod (2021).

Outra maneira importante de melhorar o resultado de um modelo é a validação cruzada. Essa técnica permite estimar a capacidade real de generalização do modelo e prevenir vazamento de dados, garantindo robustez na seleção de hiperparâmetros. Além disso, técnicas como aumento de dados (como rotação, ruído, variação de escala, síntese de dados) expandem a variedade do treinamento, reduzindo *overfitting* e, em certos casos, até amenizando *underfitting* por introdução de novos padrões (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

4.4.5 Séries temporais

Uma das aplicabilidades principais e que é objeto do presente trabalho é a análise de séries temporais com *machine learning*. Para isso, deve ser considerado várias particularidades que não são tratadas por modelos tradicionais.

Em primeiro plano, há a dependência temporal: cada observação pode depender fortemente das anteriores, exigindo métodos que preservem essa ordem para capturar padrões de autocorrelação e sazonalidade (HYNDMAN; ATHANASOPOULOS, 2021). Também há a estacionariedade, condição na qual propriedades estatísticas da série como média e variância são constantes ao longo do tempo — é frequentemente um pré-requisito para modelos clássicos (ex: *ARIMA*), sendo necessária a aplicação de transformações como diferenciação ou testes como o ADF para garantir viabilidade estatística (HYNDMAN; ATHANASOPOULOS, 2021).

Como descrito por Lazzeri (2020), para problemas de série temporal é usual tranformá-la em um problema supervisionado por meio de janelas deslizantes (*slidding window* ou *rolling window*), onde os últimos *n* valores (*lags*) viram variáveis de entrada para predizer os próximos *m* passos à frente. As janelas podem ser deslizantes, exemplificado na Figura 9, em que as variáveis de entrada são valores passados que preveem um valor futuro, que para a próxima predição irá ser usada como valor passado. Essas janelas deslizam seguindo a previsão, diferindo das janelas fixas. Em relação as janelas fixas, elas se baseiam no uso de *lagged values* ou *lag features* (valores ou *features* defasadas) igualmente às deslizantes. A diferença é que nesse caso, a janela não é deslizante e não acompanha a previsão do momento t+1, as variáveis passadas usadas na previsão sempre permanecerão as mesmas independente do horizonte de previsão.

Figura 9. Representação esquemática de janelas deslizantes de tamanho 4 para um horizonte de previsão fixo de 3.

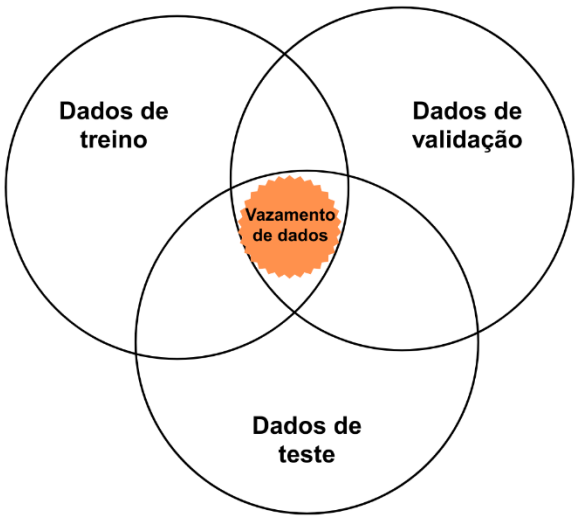
Fonte: Adaptado de BENSON et al. (2020).

Um ponto crítico na modelagem de séries temporais é a divisão dos dados em

treino, validação e teste, que deve respeitar a ordem cronológica para evitar vazamento de dados (*data leakage*). O vazamento de dados é um erro sutil na divisão do conjunto de dados de um modelo de série temporal. Ao fazer a divisão dos dados de treino, validação e teste, é comum usar estratégias de randomização para que haja o fator aleatório e o modelo generalize melhor. No entanto, em dados temporais, isso pode ocasionar dados do futuro estarem em dados do passado, ou seja, datas que fazem parte da sazonalidade do período de teste estarem no conjunto de dados do período de treino, isso é um vazamento de dados futuros dentro do conjunto de treino e pode gerar um *overfitting* mascarado (LAZZERI, 2020).

A Figura 10 foi criada para ilustrar uma divisão de conjunto em que ocorre vazamento de dados, evidenciando que os dados de testes são vistos pelo modelo em outras etapas do seu treinamento.

Figura 10. Divisão de dados de treino, validação e teste com vazamento de dados.



Fonte: Elaborado pelo autor (2025).

4.4.6 Modelos mais usados

Em relação aos tipos de modelos mais utilizados, existem os modelos de Redes Neurais e as Árvores de Decisão. Os tópicos seguintes irão discutir com maior profundidade sobre cada um deles.

4.4.6.1 Redes neurais artificiais (*ANNs*)

Os modelos de redes neurais artificais (*ANNs*), incluindo perceptrons multicamadas (*MLPs*), redes recorrentes (*RNNs*) e variantes modernas como *LSTMs*, são modelos computacionais inspirados na estrutura do sistema nervoso humano. A literatura clássica sobre redes neurais artificiais descreve com profundidade como esses modelos se inspiram no funcionamento do cérebro humano, em especial nas conexões sinápticas entre neurônios.

Segundo Haykin (2001), as redes neurais artificiais buscam reproduzir, de forma abstrata, o processo de aprendizagem biológica, no qual os "pesos sinápticos" são ajustados a partir de estímulos, analogamente às forças das conexões entre neurônios no cérebro. Esses pesos determinam a influência de cada entrada na saída do neurônio, e seu ajuste ocorre durante o treinamento com base em uma função de custo que mede o erro entre a previsão da rede e o valor real. O objetivo é minimizar essa função por meio de algoritmos como o gradiente descendente, utilizando o método de retropropagação (*backpropagation*) para calcular os gradientes de forma eficiente.

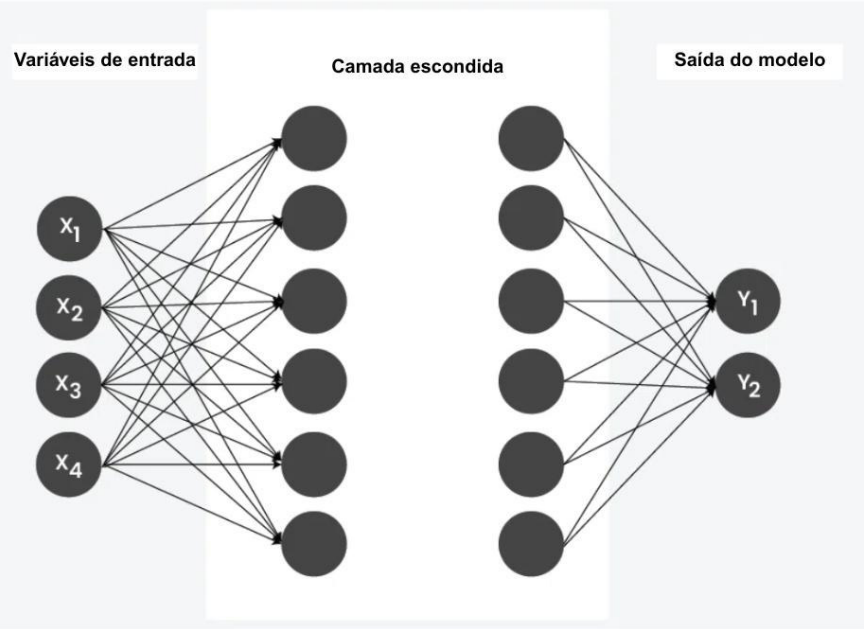
Grossberg (1988) também destaca que esses sistemas adaptativos são capazes de aprender padrões complexos com base em experiências passadas, reforçando a analogia com o cérebro. No entanto, os modelos enfrentam desafios importantes, como o *overfitting* — quando a rede se ajusta excessivamente aos dados de treinamento e perde capacidade de generalização. Para mitigar esse problema, técnicas como regularização (L1, L2) e normalização de lotes (*batch normalization*) são empregadas para estabilizar o aprendizado e melhorar o desempenho em dados não vistos (HAYKIN, 2001).

Segundo Géron (2022), em uma arquitetura de rede neural cada neurônio artificial integra entradas por meio de somas ponderadas, adiciona um termo de viés e aplica uma função de ativação não linear — como sigmoide, tangente hiperbólica ou ReLU — para gerar sua saída. A escolha da função de ativação é fundamental: sigmoides oferecem suavidade e limites, tanh centra os dados em torno de zero, e a ReLU economiza potência computacional e ajuda a mitigar os problemas de gradientes, embora possa causar unidades “mortas”.

Em termos estruturais, as *ANNs* modernas geralmente estruturam-se em camadas: uma camada de entrada, várias camadas ocultas e uma camada de saída, podendo ser totalmente conectadas (*MLP*) ou com realimentações, como nas *RNNs* . Nas *RNNs*, ciclos de realimentação permitem que o modelo retenha informações de entradas anteriores, criando uma memória temporal (GÉRON, 2022).

A Figura 11 ilustra como funciona a arquitetura de uma rede neural artificial.

Figura 11. Arquitetura de uma ANN simples com uma camada escondida.



Fonte: Adaptado de GeeksforGeeks (2025).

Para tarefas de previsão — como previsão do tempo — as arquiteturas recorrentes especializadas se destacam. Modelos simples de *RNN* já capturam dependências curtas, mas sofrem com o desvanecimento ou explosão de gradiente. Por isso nasceram arquiteturas como *LSTM* (*Long Short-Term Memory*), apresentando mecanismos internos de portas que decidem quando armazenar, liberar ou esquecer informações, permitindo modelar dependências de longo prazo em séries temporais.

Em aplicações meteorológicas, essas redes são capazes de integrar dados sequenciais (temperatura, pressão, umidade ao longo do tempo) e capturar padrões sazonais e tendências, gerando previsões mais precisas.

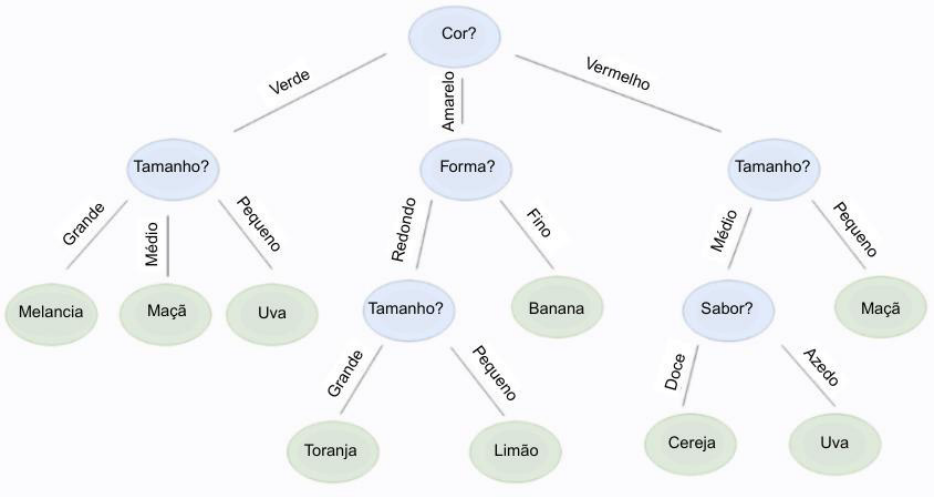
4.4.6.2 Árvores de decisão

Segundo Géron (2022), uma árvore de decisão consiste em um modelo de aprendizado supervisionado que é estruturado em folhas e ramos, onde as decisões são tomadas a partir dos dados iniciais, seguindo diferentes caminhos de ramificação. O algoritmo otimiza a divisão dos dados através de um critério para atingir o resultado esperado, que pode ser minimizar erros (*RMSE*, *MAE*). Essa característica faz com que uma árvore de decisão seja fácil de visualizar e de entender por não especialistas. O fluxograma de uma árvore de decisão é ilustrado na Figura 12.

Além disso, para o algoritmo de árvore de decisão não é necessário uma normalização dos dados, pois ele lida bem com variações numéricas e categórias nos dados de

entrada, sendo robusto para o tratamento de outliers. Apesar dessas características positivas, as árvores tendem a sofrer de alta variância e podem sofrer *overfitting* muito fácil se não forem adequadamente podadas. Existem alguns métodos de *pruning* (poda) que suavizam esse problema, mas não garantem uma boa generalização (GÉRON, 2022).

Figura 12. Arquitetura de um algoritimo de árvore de decisão.



Fonte: Adaptado de Gunji (2020).

Para mitigar essas limitações, existem modelos de árvore mais robustos como o *Random Forest* e o *XGBoost*. De acordo com o seu inventor, Breiman (2001), a Floresta Aleatória (*Random Forest*) combina múltiplas árvores de decisão treinadas por amostragem e em cada nó é realizada uma divisão aleatória dos atributos (*features*). As árvores são treinadas independentemente e o modelo cria diferentes subconjuntos, o que reduz a variância e o *overfitting* que uma única árvore de decisão poderia gerar.

Um dos benefícios da utilização do *Random Forest* é que ele é capaz de lidar com grandes volumes de dados, mantendo uma boa performance mesmo com variáveis irrelavantes como features do modelo. É importante destacar que, apesar de deixar o modelo mais robusto, ao combinar centenas de árvores o modelo perde interpretabilidade e pode ser computacionalmente mais custoso (GÉRON, 2022).

Embora tanto o *Random Forest* quanto o *XGBoost* sejam algoritmos baseados em árvores de decisão, eles diferem fundamentalmente na construção de suas florestas. Enquanto o *Random Forest* adota o método de *bagging* (*bootstrap aggregating*), onde várias árvores são construídas de forma paralela e independente, com suas previsões combinadas por votação no caso de problemas categóricos ou por uma média em problemas de regressão, o *XGBoost* (*Extreme Gradient Boosting*) utiliza *boosting*, onde as árvores são construídas de forma sequencial. Assim, não é gerado um conjunto de *x* árvores construídas aleatoriamente, e sim um árvore por vez. Cada nova árvore recebe o erro da árvore anterior e tenta mitigá-lo.

Por ter essa abordagem, o *XGBoost* é geralmente mais preciso, principalmente

para conjunto de dados complexos ou com ruídos. Além da construção das árvores, o *XGBoost* é otimizado para performance, pois ele possui uma regularização para evitar *overfitting* e possui um controle mais fino dos hiperparâmetros, o que o torna uma das escolhas preferidas em competições de ciência de dados (CHEN; GUESTRIN, 2016).

Além de sua performance no treinamento de modelos, o *XGBoost* também é otimizado para consumir menos da CPU do computador, tornando treinos mais complexos e com alto volume de dados mais rápidos e menos custosos.

4.4.7 Contextualização do uso de *machine learning* em previsões meteorológicas

O uso de algoritmos de *machine learning* em previsão meteorológica tem raízes pioneiras na década de 1990, com modelos de rede neural profunda aplicados à previsão de precipitação. Desde então, métodos clássicos como *Support Vector Machines (SVM)*, *Random Forest* e regressão Gausiana foram aplicados à previsão de variáveis meteorológicas, enquanto a partir de 2010 emergiram abordagens profundas como *Convolutional Neural Networks* (*CNNs*), *Long Short-Term Memory* (*LSTMs*) e arquiteturas híbridas como *ConvLSTM* (REICHSTEIN et al., 2019).

Os avanços em modelos de previsão numérica, potenciados por abordagens híbridas que combinam modelos físicos e estatísticos, conduzem naturalmente à adoção de *machine learning* para diversas variáveis meteorológicas — como vento, chuva ou temperatura — com resultados promissores. Diversos estudos têm utilizado *LSTM* e suas variantes para previsão de velocidade do vento em diferentes escalas temporais.

Um estudo em Cuba feito por Barrios, Lorenzo e Rodriguez (2022), utilizou *LSTM* com previsões do modelo *WRF* para produzir *nowcasting* até 2 h, medido em *MAE*, *RMSE* e correlação, com resultados robustos em quatro eventos representativos.

Em Shin, Min e Kim (2022), *Random Forests* ganharam destaque em previsões diárias de velocidade do vento em escala nacional (Coreia do Sul), com $RMSE < 0{,}8$ m/s, superando métodos logarítmicos e *SVR*.

Em Hanifi et al. (2020) é feito uma revisão bibliográfica das métricas para previsões de vento e previsões de geração eólica e concluiu-se que os erros usados mais comuns são o Erro Quadrático Médio (*RMSE*), o Erro Quadrático Médio Normalizado (*NRMSE*) e o Erro Absoluto Médio (*MAE*), expressos na unidade da variável, ou em porcentagem.

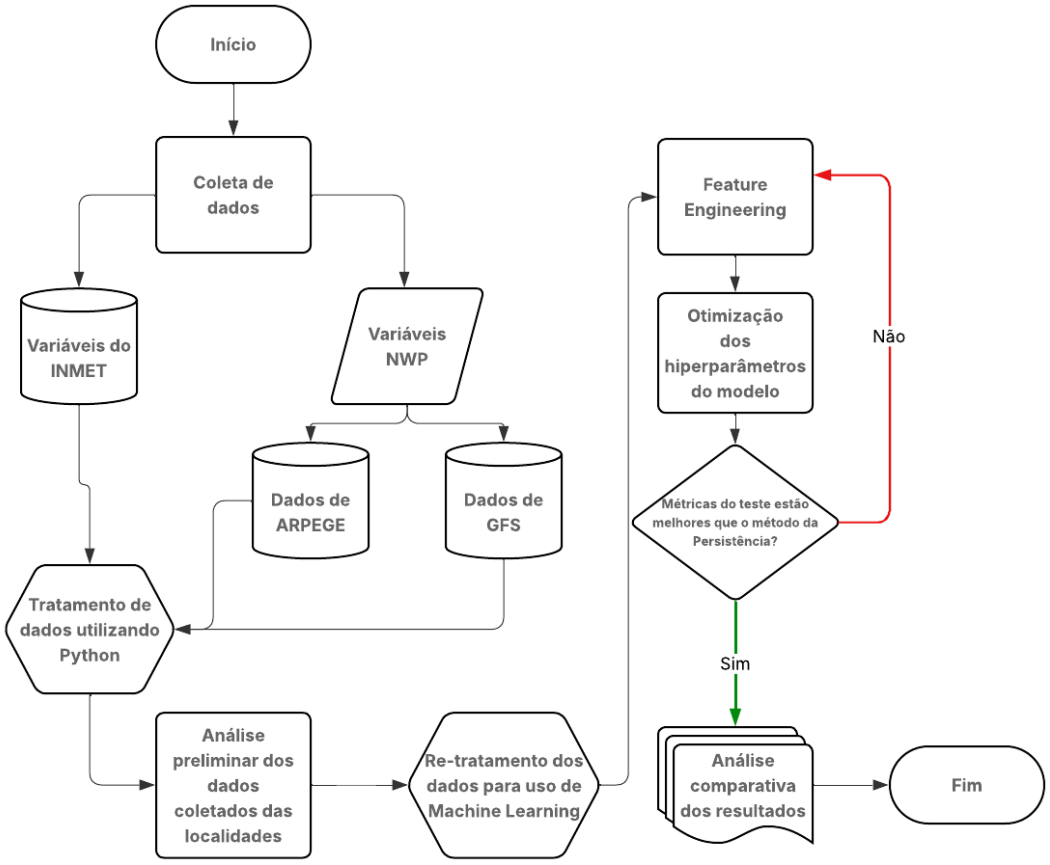
De acordo com Giebel e Kariniotakis (2017), o *NRMSE* para previsões de geração eólica em um horizonte de previsão de 24 horas está entre 9 e 14%, enquanto Monteiro et al.

(2009) indica que esse erro pode atingir valores de 14 a 17% para horizontes maiores que 48 horas.

5 METODOLOGIA

O presente trabalho foi estruturado em diversas etapas ilustradas resumidamente no fluxograma da Figura 13. Para cada tópico do fluxograma, será discutido o porquê e a justificativa para as metodologias usadas.

Figura 13. Fluxograma representativo das etapas da metodologia do trabalho.



Fonte: Elaborado pelo autor (2025).

O próximo tópico expõe, primeiramente, a coleta de dados, seguindo as etapas representativas no fluxograma.

5.1 COLETA DE DADOS

A coleta de dados é de suma importância para a confiabilidade dos resultados do trabalho, por isso é importante utilizar plataformas validadas pelo mercado e pela literatura, a fim de obter bons dados e respaldo nos resultados.

Para a coleta dos dados medidos foram utilizados dados fornecidos pelo INMET através de suas estações automáticas disponíveis no Mapa das Estações. Os dados são de

quatro localidades distintas, sendo elas: Senhor do Bonfim (BA), Conde (BA), Mossoró (RN) e Rio Grande (RS).

Para a escolha das localidades a premissa era de selecionar uma cidade do Nordeste, uma do Sul, uma do Norte e uma do Centro-Oeste, para analisar a mudança do resultado de acordo com a região. Contudo, na prática, foi difícil de encontrar estações para cada região com dados válidos para o trabalho. Sendo assim, foram escolhidas as quatro localidades citadas acima pela qualidade dos dados meteorológicos, essas apresentavam menos dados faltantes.

A Tabela 4 mostra a cidade e estado das localidades escolhidas, assim como o código do INMET das Estações Meteorológicas Automáticas e suas as coordenadas.

Tabela 4. Síntese das informações relacionadas às EMAs utilizadas.

Cidade	**Estado**	**Código da estação**	**Latitude**	**Longitude**
Senhor do Bonfim	Bahia	A428	-10.44	-40.15
Conde	Bahia	A431	-12.04	-37.68
Mossoró	Rio Grande do Norte	A318	-4.90	-37.37
Rio Grande	Rio Grande do Sul	A802	-32.08	-52.17

Fonte: Elaborado pelo autor (2025).

No que se tange aos dados de previsão, foi utilizado um site *open-source* de dados climáticos e previsão meterológica, o *Open-Meteo*. Neste site é disponibilizado uma API gratuita com inúmeros modelos *NWP* de todo o mundo, na Tabela 5 há uma lista dos principais modelos de previsão numérica do tempo disponíveis.

Tabela 5. Principais modelos de NWP disponíveis na API do Open-Meteo.

Nome	**Tipo**	**Origem**
GFS Global 0,11°/0,25°	Global	Estados Unidos
ARPEGE World	Global	França
Uk Met Office UK 2km	Mesoescala	Reino Unido
ECMWF IFS 0,25°	Global	Europa
DWD ICON Global	Global	Alemanha

Fonte: Elaborado pelo autor (2025).

5.1.1 Variáveis do INMET

Através do Mapa das Estações do INMET, foram exportados dados CSV mensais para cada uma das quatro localidades. Os dados começam no dia primeiro de janeiro e terminam no dia primeiro de julho, totalizando 6 meses de dados. No documento é coletado valores de máximo, mínimo e instântaneo das variáveis meteorológicas disponíveis. As variáveis estão listadas na Tabela 6.

Tabela 6. Variáveis meteorológicas coletadas das EMAs do INMET.

Variável	**Unidade**	**Estatísticas descritivas**
Temperatura	°C	Máximo, Mínimo, Instântaneo
Umidade	%	Máximo, Mínmo, Instântaneo
Ponto de orvalho	°C	Máximo, Mínmo, Instântaneo
Pressão	hPa	Máximo, Mínmo, Instântaneo
Velocidade do vento	m/s	Instântaneo
Direção do vento	°	Instântaneo
Rajadas	m/s	Instântaneo
Radiação	kJ/m²	Instântaneo
Chuva	mm	Instântaneo

Fonte: Elaborado pelo autor (2025).

Na Figura 14 pode-se observar como é uma EMA do INMET. Ela contém sensores, um mastro com caixa data-logger, painel solar para funcionamento mesmo sem energia da rede elétrica pública, pára-raios e um cercado. Segundo INMET (2011), todos os equipamentos incluindo os sensores são fixados no mastro, que possui 10 metros de altura. O pluviômetro e os aparelhos para medição da radiação solar ficam situados fora do mastro, mas dentro do cercado.

Figura 14. Estação Meteorológica Automática do INMET em funcionamento.

Fonte: INMET (2011).

O cálculo do valor instântaneo é feito da seguinte maneira: a estação coleta um valor a cada 5 segundos, durante 60 segundos teremos 12 valores. O valor instântaneo é a média das 12 amostras coletadas próximos a hora medida, ou seja, para 12:00, o valor instântaneo será a média dos 12 valores registrados de 11h:59m:00s a 11h:59m:59s (INMET, 2011).

Esse cálculo para a velocidade e direção do vento sofre uma pequena modificação. O vento é medido a cada 0,25 segundo, sendo 4 valores por segundo. A cada 3 segundos é calculado a média móvel da sua velocidade e direção. Então, é feita a média dos últimos 10 minutos de dados para cada hora (INMET, 2011). Para melhor entendimento, será explicitado um exemplo para as duas variáveis.

Para um valor de velocidade do vento às 14:00:00, nos últimos 10 minutos – ou seja, desde 13:50:00 – obterão-se 200 valores para tirar uma média, sendo que cada valor é uma média de 3 segundos, esse resultado final será o valor instântaneo da velocidade do vento. Para a direção, o fato dela ser uma variável circular, varia de 0° a 360°, faz-se necessário realizar a transformação para coordenadas polares e a média é realizada em cima dos vetores resultantes, para então calcular o arcotangente e obtermos o resultado final da direção do vento instântanea para a hora (INMET, 2011).

5.1.2 Variáveis das Previsões Numéricas do Tempo

Como já mostrado na Tabela 5, há inúmeros modelos *NWP* disponíveis no site *open-source Open-Meteo*. A escolha de quais modelos utilizar foi feita baseado em quais os modelos mais citados na literatura, principalmente para artigos relacionados a previsão de velocidade do vento e produção de energia eólica.

Em de Farias (2020), foram utilizados *GFS* e *WRF* como inputs para um modelo de *machine learning* que prevê a geração eólica de dois parques eólicos localizados no Nordeste e no Sul do Brasil. Já no estudo de Baggio et al. (2025) são utilizados os modelos *ARPEGE* e *AROME* para a previsão da velocidade do vento. O modelo *AROME* é um modelo de mesoescala francês que utiliza como condições de contorno laterais dados resultantes do modelo global *ARPEGE*.

Sendo assim, como as localidades se encontram no Brasil, é necessário escolher um modelo global para que as previsões tenham maior acurácia. Foram escolhidos então, os modelos *Global Forecast Sytem* e *ARPEGE*.

5.1.2.1 Modelo *GFS* 0,11°

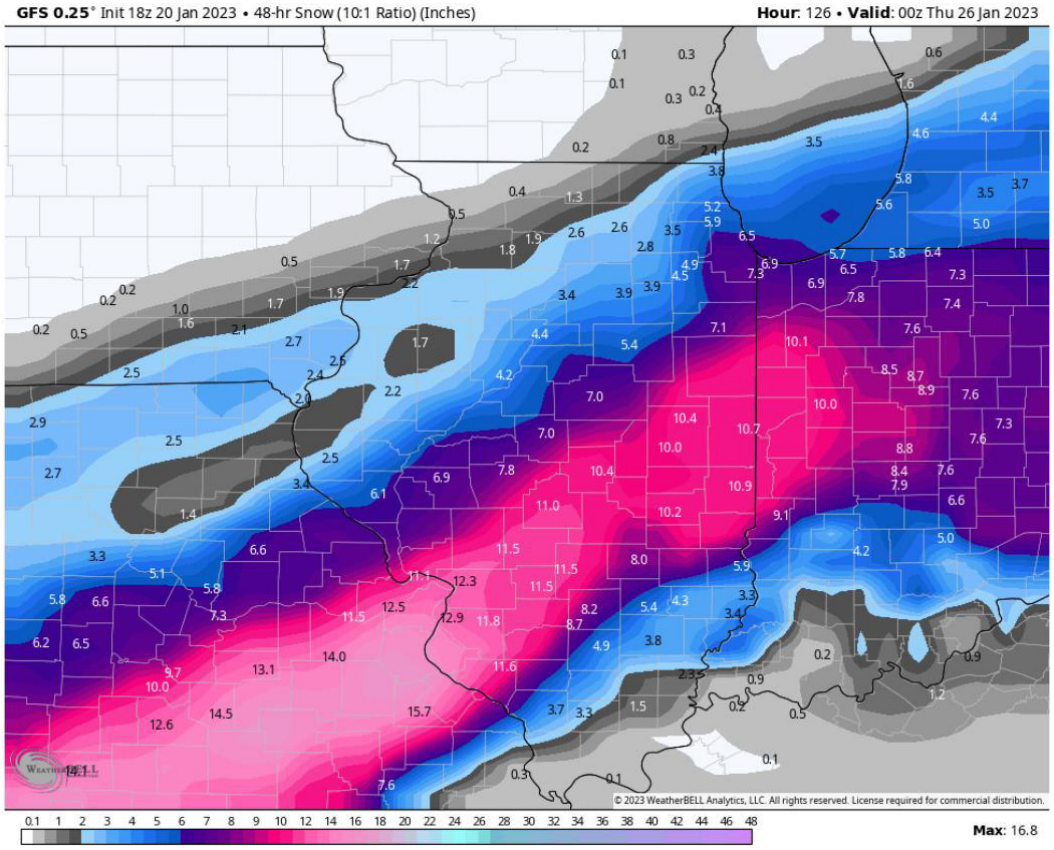
Neste trabalho foi utilizado o modelo *GFS* 0,11° de alta resolução (aproxidamente 13 km). O modelo é determinístico e prevê um dado valor para um instante t, e não uma probabilidade. O cálculo do *GFS* é feito quatro vezes ao dia em 00, 06, 12 e 18 do Tempo Universal Coordenado (UTC), fornecendo previsões de até 16 dias a frente. O modelo produz previsões com resolução de 1 hora para as primeiras 120 horas, depois com intervalos de 3 horas entre 120 e 240 horas e por último de 12 horas entre 240 e 384 horas. (NOAA, 2025)

Em relação a sua grade, sua resolução horizontal é de 13 km e ele conta com 127 camadas verticais, do solo até a mesopausa que está localizada a cerca de 80 km acima da superfície (GFS, 2022). A mesopausa é a camada limite superior entre a mesosfera e a termosfera.

Existem outros modelos com resoluções espaciais horizontais de 0,25, 0,50 e 1 graus, foi escolhido para o trabalho o modelo com menor resolução, com 0,11°.

Na Figura 15 pode-se observar como exemplo uma malha de previsão do modelo *GFS*, que no caso é o modelo com resolução horizontal de 0,25°, onde está sendo simulada uma previsão para uma tempestade de neve em janeiro de 2023.

Figura 15. Simulação de cálculo do modelo GFS 0,25° para previsão de uma tempestade de neve em janeiro de 2023. A localidade não foi informada.



Fonte: National Weather Service (2023).

As variáveis coletadas do modelo foram as previsões da velocidade e direção do vento a 10 metros, da temperatura na superfície e a 2 metros de altitude, e da pressão atmosférica na superfície para as quatro localidades. Os dados são armazenados na forma de JSON.

5.1.2.2 Modelo *ARPEGE* 0,11°

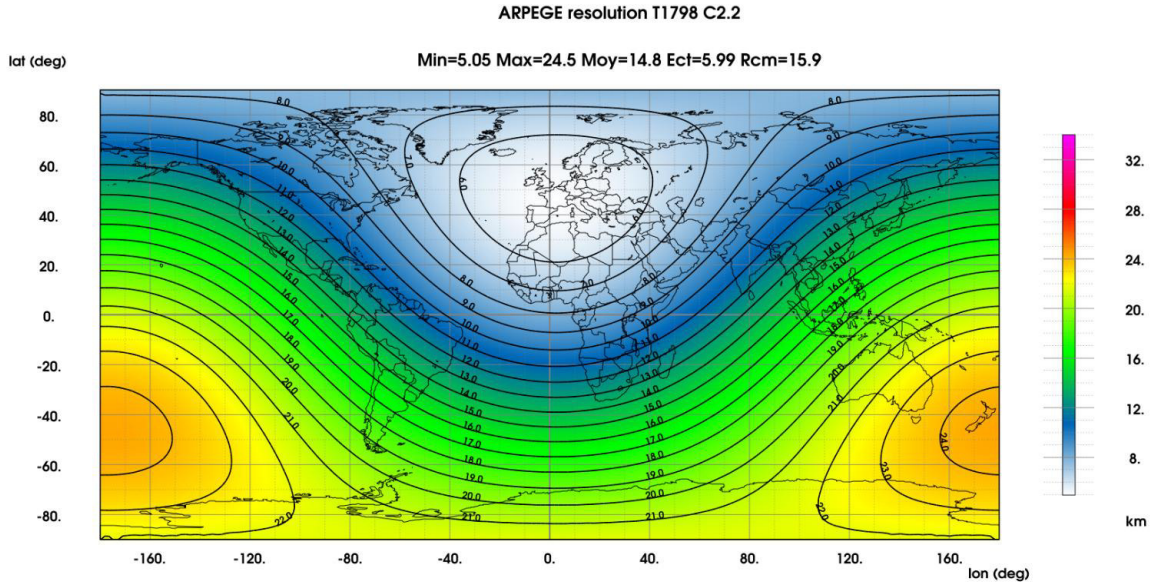
A *ARPEGE* (Action de Recherce Petite Échelle Grande Échelle) é um modelo *NWP* determinístico global desenvolvido pela *Météo-France* em colaboração com o *ECMWF*. O modelo utiliza uma análise por assimilação 4D-var onde utiliza observações de radiossondas, aviões, estações terrestres, sinais de GPS em solo e satélite, entre outros (CNRM, 2022).

A *ARPEGE* adota uma resolução horizontal variável a depender da região e utiliza a representação em elementos finitos na vertical. A resolução horizontal é de 5 km em território francês, e pode variar de 11 km até 24km nos países vizinhos e para o resto do mundo. O modelo possui 105 níveis verticais atingindo até 70 km de altitude. A grade usada é

de 0,05° x 0,05° para a Europa e de 0,1° x 0,1° para o resto do mundo no ARPEGE Global, tendo este último aproximadamente 11 km de resolução horizontal, com as previsões atingindo até 102 horas. O modelo é executado quatro vezes ao dia (00, 06, 12 e 18 UTC). Para a integração temporal, é empregado um esquema semi-implícito semi-Lagrangiano com um passo temporal de 240 segundos (CNRM, 2022).

A Figura 16 ilustra a variação da resolução horizontal em km do modelo ARPEGE no globo.

Figura 16. Ilustração da resolução horizontal em km do modelo numérico ARPEGE.



Fonte: CNRM (2022).

Na Figura 16 é possível ver como a resolução horizontal do modelo numérico ARPEGE varia com a região, sendo menos preciso para regiões distantes da Europa.

As variáveis coletadas do modelo foram as previsões da velocidade e direção do vento a 10 metros, da temperatura a 20 metros e a 2 metros de altitude, e da pressão atmosférica na superfície para as quatro localidades. Os dados são armazenados na forma de JSON.

5.2 TRATAMENTO DOS DADOS

A primeira limpeza e tratamento de dados foi feita com o exclusivo objetivo de analisar os dados medidos pelas estações meteorológicas para as quatro localidades e comparar com os dados previstos pelos modelos *GFS* e *ARPEGE*.

Para isso foi utilizado a linguagem de programação *Python* e suas bibliotecas (*pandas*, *numpy*, etc) para a:

- remoção de valores ausentes;
- filtragem de valores negativos para a velocidade do vento caso existam;
- remoção de dados duplicados que são comuns ao coletar dados mensais – dados do fim e começo do mês anterior podem ser duplicados ao coletar dados do mês seguinte;
- alinhamento temporal entre os dados medidos e dados previstos;
- interpolação e extrapolação de dados.

É importante salientar que foi feita a interpolação e extrapolação de dados previstos de temperatura. Isso foi feito pois as medições são a 10 metros, enquanto que para *GFS* as previsões de temperatura disponíveis eram a 2 metros de altitude e na superfície e para *ARPEGE* as previsões de temperatura disponíveis eram a 20 metros e a 2 metros de altitude. Apesar do objetivo ser a previsão da velocidade do vento, a ideia era de deixar as variáveis de entrada ao máximo possível na mesma altitude que a variável alvo para que os erros fossem suavizados, mesmo que o uso de interpolações tragam um pequeno erro associado.

Em relação a variável de pressão atmosférica, apesar de os modelos a fornecerem na altitude de superfície, não foi feita a sua extrapolação. Além de não ter sido possível coletar dados de variáveis de pressão para uma outra altitude para extrapolar, foi observado que a variabilidade dessa variável era muito baixa e a sua diferença entre os valores previstos a superfície e medidos a 10 metros eram bem próximos. Foi entendido, portanto, que o seu comportamento seria útil para ajudar o modelo a prever a velocidade do vento mesmo a utilizando de forma íntegra.

5.3 ANÁLISE PRELIMINAR DOS DADOS

Antes de começar a aplicação do algoritmo de *machine learning* foi realizada uma análise preliminar dos dados das quatro localidades.

A análise dos dados medidos foi feita através de gráficos que ilustram a média horária da velocidade e direção do vento, temperatura e pressão atmosférica, a variabilidade dessas variáveis através de box plots, e a sua frequência através de histogramas. Além disso, foi feita uma rosa dos ventos para cada localidade com os dados medidos e dados previstos de velocidade e direção do vento a fim de comparar a sua similaridade.

Essa análise teve o fito de analisar quais as características dos dados medidos pelas estações meteorológicas e comparar com os dados previstos.

5.4 RE-TRATAMENTO DOS DADOS PARA IMPLEMENTAÇÃO DE MACHINE LEARNING

Esse re-tratamento foi feito da mesma maneira que o tratamento inicial, limpando valores faltantes, removendo duplicatas, etc. A atenção maior veio para a ordenação da série temporal e a concatenação dos dados medidos e previstos. Foi realizado o mesmo tratamento após a análise preliminar dos dados devido aos requisitos para o uso dos dados em algoritmos de série temporal.

Quando adentramos ao mundo do *machine learning*, mais especificamente no caso de séries temporais, é crucial que todos os dados estejam perfeitamente ordenados para que seja feita a divisão entre dados de treino, validação e teste corretamente.

5.5 ENGENHARIA DE FEATURES

Segundo Reneau et al. (2023), *feature engineering* (engenharia de *features*) é o processo de modelar e selecionar as características mais relevantes dos dados que podem ser usados para treinar um modelo, impactando significativamente no desempenho do mesmo. Ao utilizá-la, é possível capturar os padrões complexos que existem em dados de série temporais, como a sazonalidade (diária, mensal).

Dentre vários métodos existentes, foram empregados os métodos de *lag features*, e *time-based features* (features baseadas no tempo).

Além da otimização das variáveis de entrada, é crucial uma divisão correta dos dados de treino, validação e teste, principalmente pelo escopo do problema ser uma série temporal, como já foi explicado anteriormente. Esta divisão será explicada no tópico seguinte, seguida dos métodos de *feature engineering*.

5.5.1 Treino, validação e teste

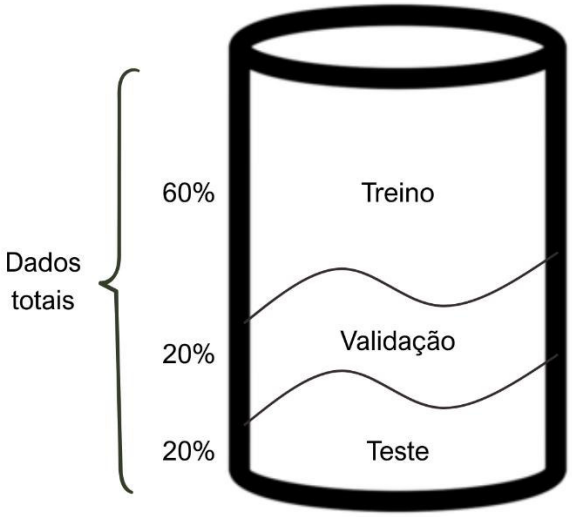
Para este trabalho foi feito uma separação entre os dados da *ARPEGE* e os dados da *GFS*, ou seja, um modelo será treinado com as previsões de um e outro será treinado com as previsões do outro, a fim de comparar os resultados.

No conjunto de treinamento tem-se os dados que o modelo vai usar para entender

os padrões e a relação entre as variáveis de entrada e a variável alvo, é onde há o aprendizado da máquina. No conjunto de validação, tem-se dados que não foram vistos durante o treinamento e são usados para otimizar os hiperparamêtros do modelo. Já o conjunto de teste é composto com o resto dos dados que devem ser previstos.

Para evitar vazamento de dados na série temporal, foi feita a divisão cronologicamente ordenada, sendo os primeiros 60% de dados para treinamento do modelo, 20% após para a validação e os últimos 20% para o teste. Essa divisão é ilustrada na Figura 17.

Figura 17. Divisão dos dados de treino, validação e teste utilizada.



Fonte: Elaborado pelo autor (2025).

5.5.2 Lag features

Este método se caracteriza por incorporar valores passados na predição atual, fornecento uma espécie de contexto que pode aumentar a acurácia do modelo. Como o vento é estocástico e as variáveis meteorológicas são conhecidamente caóticas e de difícil previsão, a inclusão de dados medidos passados para capturar tendências de curto prazo tendem a melhorar a precisão da previsão.

No estudo de Reneau et al. 2023 a utilização de *lag features* melhorou o desempenho dos modelos usados (MLP, CNN e LSTM) para o coeficiente R^2 em diferentes base de dados como o Traffic Dataset – conjunto de dados que contém a taxa de ocupação de rodovias na Baía de São Francisco.

Para a escolha do tamanho da janela do passado aplicada às variáveis foram feitos testes com os valores de 1, 3, 6, 12 e 24 para a localidade de Mossoró (RN) e os dados de

previsão da *GFS*. Cada número representa a quantidade de dados passados, então para uma janela com 24 dados, sendo estes horários, isso representa incorporar às *features* dados das 24 horas anteriores. Essa análise é importante pois em um dado momento o custo computacional para aumentar a quantidade de dados passados coletados começa a não ser benéfico em relação ao desempenho do modelo, podendo aumentar a complexidade desnecessariamente.

5.5.3 Time-based features

As *time-based features* (*features* baseadas no tempo) são as especifidades temporais que podem ser informadas ao modelo, tais como qual o dia da semana, mês do ano, se é final de semana ou não, etc. Elas podem melhorar o modelo, pois algumas variáveis alvos possuem influência no tempo, as vendas por exemplo são maiores nos fins de semana (GORDON, 2023).

O estudo de Bansal, Balaji e Lalani (2025) dissertou acerca da transformação de *features* cíclicas como horas, dias ou meses em senoides e como isso preserva a natureza contínua dos padrões temporais e capturam melhor sua periodicidade, enquanto que utilizar as representações numéricas podem criar descontinuidades artificais. Nesse mesmo trabalho foi demonstrado que a aplicação de codificação sinusoidal nas variáveis temporais superou consistentemente a codificação numérica em todos os modelos avaliados.

Esse entendimento é crucial para o presente trabalho pois uma das variáveis de entrada é a direção do vento, uma variável circular que varia de 0° a 360° e deve ser corretamente transformada para coordenadas polares para que o modelo não entenda a diferença de grandeza dos valores como um padrão descontínuo. A parametrização realizada para essa variável está ilustrada nas equações 1 e 2.

Seja θ a direção do vento em graus, então:

$$wind\ direction_{sin} = \sin(\frac{\pi}{180} \cdot \theta) \quad (1)$$

$$wind\ direction_{cos} = \cos(\frac{\pi}{180} \cdot \theta) \quad (2)$$

Para além da transformação da direção do vento em seno e cosseno, foram feitas algumas tentativas de diferentes variáveis cíclicas como entrada do modelo para melhora de sua performance, tais como a hora do dia, o dia do ano, o mês do ano, a interação entre valor de temperatura e a hora correspondente, entre outros. Mais detalhes serão discutidos nos resultados e discussões.

5.6 OTIMIZAÇÃO DOS HIPERPARÂMETROS

Os hiperparâmetros são parâmetros que são definidos antes do treinamento do algortimo, diferindo dos parâmetros internos que são ajustados automaticamente, e são responsáveis por controlar como é feito o aprendizado do modelo.

Pelos benefícios citados na metodologia, o *XGBoost* foi escolhido para ser o algortimo usado no trabalho. Para ele, existem os parâmetros *n_estimators* que se refere à quantidade de árvores que serão criadas durante o treinamento, a taxa de aprendizado (*learning_rate*) que é uma taxa da rapidez do modelo de aprender o padrão dos dados e o *max_depth* que é a profundidade máxima das árvores (FILHO, 2024).

A otimização de hiperparâmetros é um processo crucial em *machine learning* e visa determinar a combinação ideal das configurações de um algortimo, a fim de maximizar a sua performance, isso é fundamental para evitar *overfitting* ou *underfitting* do modelo. Essa otimização pode ser feita por meio de tentativa e erro, mas isso pode ser muito demorado e pouco eficaz. Para isso existem algumas técnicas como a busca aleatória (*random search*) e a busca em grade (*grid search*). A utilização do método de *random search* é feita informando as configurações e realizando a otimização através de combinações aleatórias dos hiperparâmetros. Já para o método de *grid search*, as configurações que são determinadas são testadas com todas as combinações possíveis.

Para a otimização dos hiperparâmetros e *features* foi utilizado como parâmetro os dados de Mossoró (RN) e as previsões de *GFS*. O par foi escolhido aleatoriamente e essa simplificação foi realizada devido a complexidade e uso computacional necessário para essa análise, onde caso contrário seria necessário realizar diversas iterações custosas para cada par localidade – previsão. É importante lembrar que o conjunto de dados usados para o ajuste desses hiperparâmetros é o conjunto de validação.

Para o algoritmo *XGBoost*, foi utilizado o método *GridSearchCV* para combinação e validação dos hiperparâmetros que variavam o número de arvores criadas (*n_estimators*), a profundidade de cada árvore (*max_depth*) e a taxa de aprendizado (*learning rate*). A métrica utilizada para validação foi o *RMSE*, de tal forma que o objetivo seja a minimização do mesmo.

A Tabela 7 mostra os valores testados para cada hiperparâmetro.

Tabela 7. GridSearchCV para otimização de hiperparâmetros.

Hiperparâmetro	**Valores de teste**
n_estimators	[100, 200, 500, 1000]
max_depth	[1, 2, 3, 4, 5]

learning_rate	[0,001, 0,01, 0,1]

Fonte: Elaborado pelo autor (2025).

Dessa forma, existem 60 combinações possíveis no gridseacrh, sendo que elas foram calculadas uma a uma com o armazenamento do *RMSE* da combinação. Além da análise do *RMSE* da combinação, será análisado o impacto de cada hiperparâmetro no *RMSE* para mapear quais são mais influentes no aprendizado do modelo.

5.7 TREINAMENTO E TESTE DO ALGORITMO

Após a otimização dos hiperparâmetros com os conjunto de dados de validação, foi obtido a configuração ideal para o treinamento do modelo com dados que ainda não foram vistos por ele nenhuma vez, os dados do conjunto de teste que representam 20% dos últimos dados totais após o tratamento, como já explicitado anteriormente.

A referência para a performance de modelos de previsão (*forecast*) usualmente é o método da Persistência. A Persistência utiliza-se como premissa que o dado que aconteceu anteriormente irá se repetir no futuro. Explicando de forma prática com o caso de previsão da velocidade do vento, se nos dados medidos a velocidade medida ao meio dia foi de 4 m/s, a Persistência prevê que a velocidade do vento às 13h será de 4 m/s. Se a velocidade do vento medida às 13h for de 4,3 m/s, a velocidade do vento prevista às 14h será de 4,3 m/s, e assim sucessivamente. Apesar de simples, esse método funciona bem para horizontes de previsões curtos e é uma boa referência segundo a literatura.

Por conseguinte, as métricas de teste para os modelos gerados por dados *ARPEGE* e *GFS* serão comparados com o método da persistência, todos para um horizonte de previsão de 24 horas para as quatro localidades. Caso o resultado do teste não consiga superar a Persistência, a metodologia irá ser novamente regredida a *Feature Engineering* para uma melhor parametrização e treinamento do modelo.

Além disso, é esperado que a acurácia dos métodos diminua com o aumento do horizonte de previsão, pois segundo Clark, McCracken e Mertens (2017), ao desenvolverem um modelo para a variação da incerteza no tempo na previsão de múltiplos horizontes, a precisão das previsões tende a diminuir à medida que o horizonte de previsão aumenta, devido ao aumento das incertezas associadas.

Para a avaliação das previsões foram calculados o *RMSE* e o R^2 para cada horizone de previsão $h \in \{1, 2, 3 ..., H\}$, sendo as previsões feitas h passos à frente e o H sendo o total dos horizontes, 24. Para o valor do *RMSE* médio e do R^2 médio do modelo, foram calculadas as médias do *RMSE* e do R^2 por horizonte de previsão.

As equações 3, 4, 5 e 6 mostram o cálculo feito para o *RMSE* por horizonte e o valor médio e para o R^2 por horizonte e o valor médio, respectivamente.

$$RMSE_h = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{i,h} - \hat{y}_{i,h})^2} \tag{3}$$

$$RMSE_{médio} = \frac{1}{H}\sum_{h=1}^{H} RMSE_h \tag{4}$$

$$R_h^2 = 1 - \frac{\sum_{i=1}^{n}(y_{i,h} - \hat{y}_{i,h})^2}{\sum_{i=1}^{n}(y_{i,h} - \bar{y}_h)^2} \tag{5}$$

$$R_{médio}^2 = \frac{1}{H}\sum_{h=1}^{H} R^2{}_h \tag{6}$$

Em que:

$y_{i,h}$ são os valores reais,

$\hat{y}_{i,h}$ são os valores previstos,

$\bar{y}_h$ é a média dos valores reais,

n é o número de amostras,

H é o total de horizontes de previsão.

6 RESULTADOS E DISCUSSÃO

6.1 ANÁLISE PRELIMINAR DOS DADOS

Neste tópico houve a discussão acerca de como os dados se apresentam inicialmente, investigando a proximidade dos dados de previsão numérica com os dados medidos nas estações de cada localidade.

Primeiramente, foi realizado um tratamento dos dados no qual foram eliminados todos os dados duplicados e todos os dados faltantes do conjunto, com o intuito de deixar a base de dados idêntica para os dados do INMET, da *ARPEGE* e da *GFS*. Na Tabela 8 há uma sumarização do que se obteve após tratamento.

Tabela 8. Resumo do conjunto de dados após tratamento.

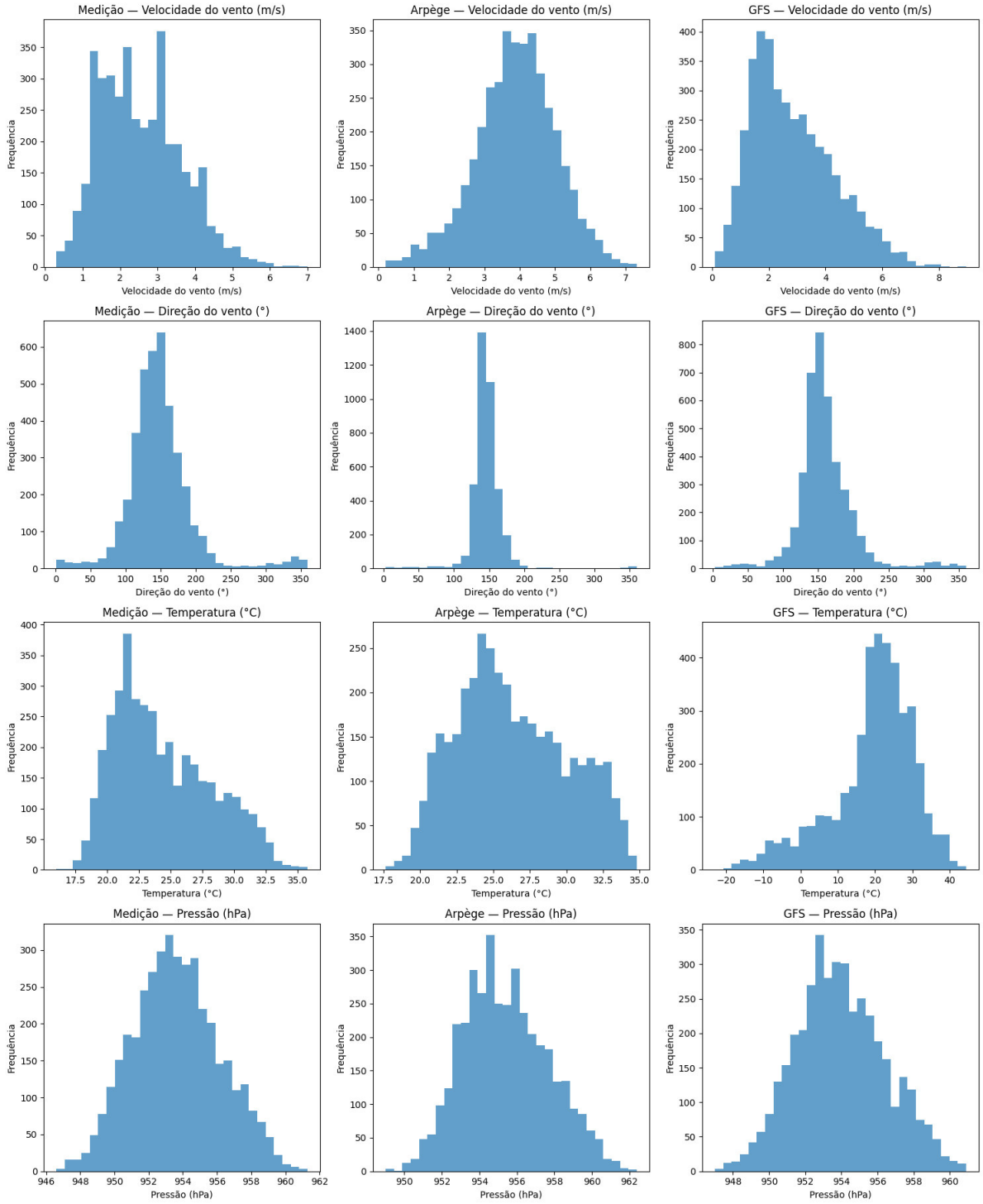
Localidade	**Data inicial**	**Data final**	**Qtd. de dados válidos por variável**	**Qtd. de dados totais**
Senhor do Bonfim	19/01/2025	01/07/2025	3589	14356
Conde	19/01/2025	01/07/2025	3736	14944
Mossoró	19/01/2025	01/07/2025	3926	15704
Rio Grande	19/01/2025	01/07/2025	3927	15708

Fonte: Elaborado pelo autor (2025).

Assim, pode-se observar que foi armazenado um pouco menos de 6 meses de dados, e se apresentam com mais de 3500 valores horários para cada variável. Como foram coletadas como principais e foco de análise a velocidade e direção do vento, temperatura e pressão atmosférica, multiplicando o valor por quatro chega-se no valor de dados válidos totais. Estes são mais de 14000 para todas as localidades, uma quantidade suficiente para analisar o comportamento local, pelo menos para a sazonalidade do primeiro semestre do ano, e munir um modelo de *machine learning*.

A Figura 18 ilustra os histogramas para cada variável meteorológica utilizada no trabalho no período de cerca de 6 meses para Senhor do Bonfim, na Bahia.

Figura 18. Histograma dos dados das variáveis meteorológicas analisadas para Senhor do Bonfim.



Fonte: Elaborado pelo autor (2025).

Pode-se observar que o histograma de velocidade do vento aproxima-se de uma distribuição normal, como o esperado, mas que o histograma de previsão da *GFS* é mais próximo do medido do que para *ARPEGE*. Analisando o histograma, *ARPEGE* tende a superestimar o vento, visto que os valores mais frequentes de velocidade estão na faixa de 3,5 a 4 m/s, enquanto que para o histograma dos dados medidos a maior frequência é próxima de 3 m/s, com muitos valores inferiores.

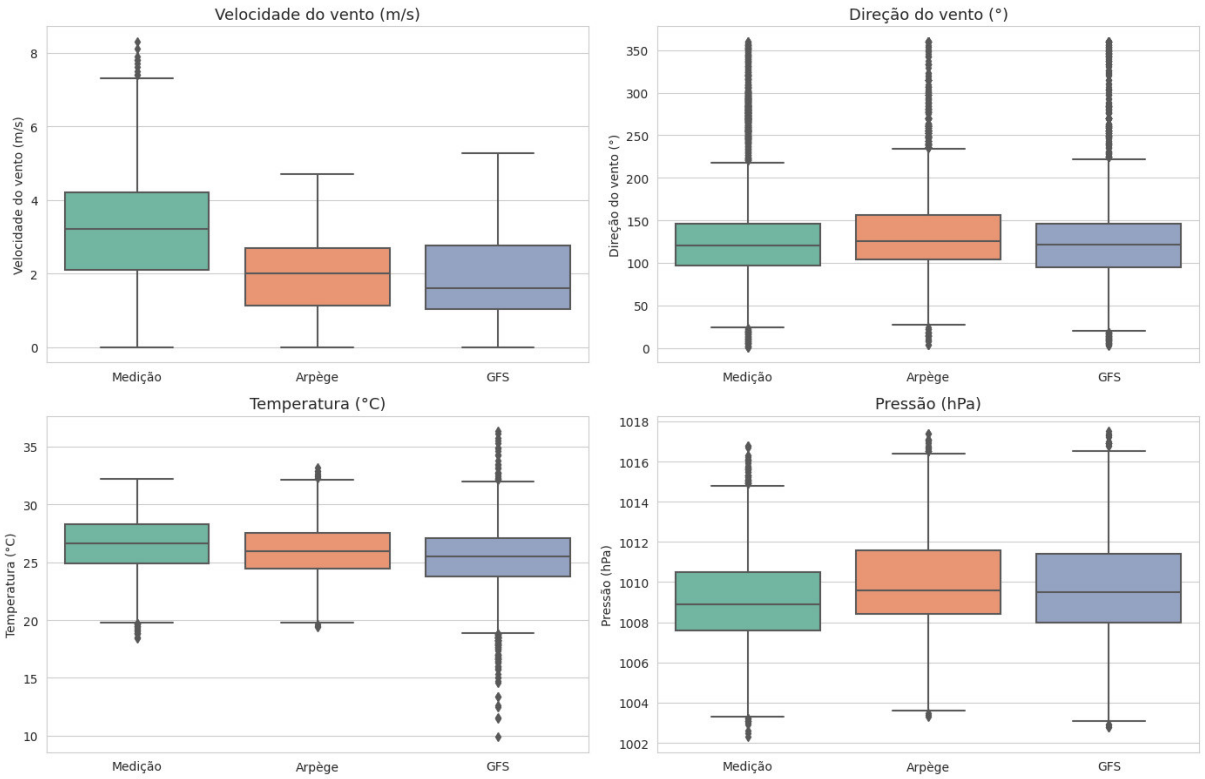
Referente ao histograma dos dados de direção do vento, os três são bem

semelhantes, com uma mediana bem próxima de 150°C. Para o histograma de temperatura, ocorre o inverso em relação a velocidade do vento, a *GFS* superestima muito os valores de temperatura enquanto que *ARPEGE* superestima um pouco menos. Por último, para os valores de pressão, os três histogramas são bem similares, mesmo sendo uma comparação de 10 metros de altitude para os dados observados com a pressão na superfície para os dados previstos.

Os histogramas de Conde, Mossoró e Rio Grande estão nos apêndices A.1 e é observado que os valores superestimados de velocidade do vento não são unânimes para os dados de previsão em relação aos dados medidos, todavia ocorrem com certa frequência. Os histogramas de direção do vento se mostraram semelhantes, apesar de também terem um erro associado. Para a temperatura também, apesar de *GFS* demonstrar possuir uma menor variabilidade de dados de temperatura. Para os dados de pressão, estes são os que demonstraram serem mais próximos, apesar da diferença de altitude.

A Figura 19 mostra o gráfico de box plot das variáveis meteorológicas para a cidade de Conde.

Figura 19. Box plot dos dados das variáveis meteorológicas analisadas para Conde.



Fonte: Elaborado pelo autor (2025).

Verifica-se que neste caso os dados de velocidade do vento previstos estão com os seus limites superiores bem abaixo do limite dos dados medidos, mostrando que a velocidade do vento está sendo subestimada. O box plot de direção do vento possui muitos *outliers* para os três conjuntos, isso é, porém, esperado, pois como a variável é circular, essa

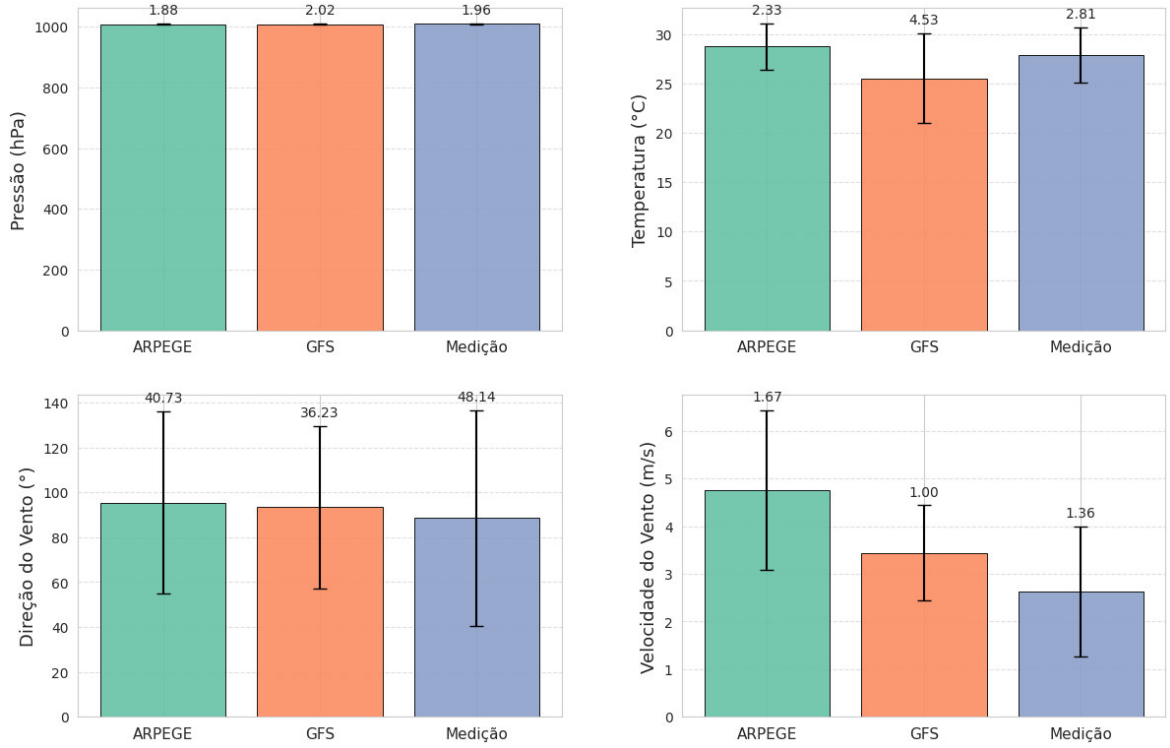
descontinuidade é lida como *outlier* no cálculo do *box plot*.

Para os dados de temperatura, destaque para uma variação grande em *GFS*, com um limite extenso e muitos *outliers*, comportamento incomum aos outros. Para a pressão, os *box plots* estão semelhantes, com um leve deslocamento entre a pressão prevista e medida.

Os gráficos de *box plot* para Senhor do Bonfim, Mossoró e Rio Grande possuem conclusão semelhante e podem ser encontrados no apêndice A.2.

Uma análise importante é ilustrada na Figura 20. Nela há um gráfico de barras que representa a média da variável no período e uma linha auxiliar que mostra o seu desvio padrão.

Figura 20. Gráfico de barras da média e desvio padrão dos dados medidos e previstos das variáveis meteorológicas para Mossoró.



Fonte: Elaborado pelo autor (2025).

Para Mossoró, observa-se valores muito baixos de desvio para a pressão atmosférica e uma média bem próxima entre os modelos. Para a temperatura, enquanto a média da *GFS* é menor que a medida, sendo um pouco maior que 25°C, a temperatura média da *ARPEGE* é levemente maior, em torno de 29°C. Nota-se um desvio padrão mais elevado para *GFS*.

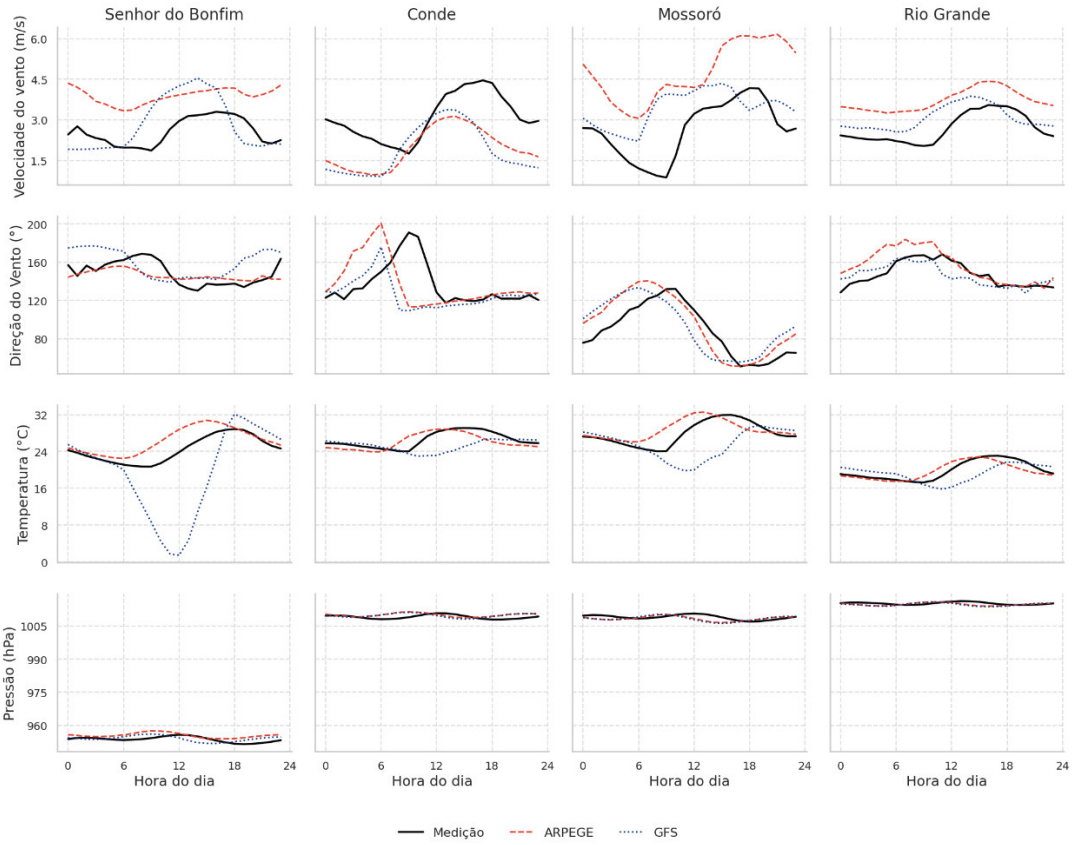
Em relação a direção do vento, as médias são próximas, em torno de 90°, e as previsões possuem um desvio padrão menor que o observado de 48,14, menor em aproximadamente 8 pontos para *ARPEGE* e 12 para *GFS*. Já para a velocidade, a média é superestimada para *ARPEGE* e *GFS*, com um valor em torno de 4,7 m/s e 3,4 m/s respectivamente. A média observada na estação é próxima a 2,6 m/s. Os desvios são

relativamente semelhantes, diferindo em torno de 0,30 negativamente para *GFS* e positivamente para *ARPEGE*.

Os gráficos para Senhor do Bonfim, Conde e Rio Grande se encontram no apêndice A.3.

Uma outra análise muito importante a se considerar, está na Figura 21. Nela está plotada a média por horizonte de tempo dos três conjuntos de dados.

Figura 21. Média por hora do dia das variáveis meteorológicas para as quatro localidades em estudo.



Fonte: Elaborado pelo autor (2025).

Verifica-se que para Senhor do Bonfim, a velocidade do vento tende a ser maior – ultrapassa os 3 m/s – depois do meio dia. Neste caso, *ARPEGE* superestima fortemente a média por hora, mostrando um enviesamento maior que *GFS*, que também possui um erro maior em horizontes de tempo entre 6 e 18 horas. Em relação à direção do vento, as curvas estão próximas e possuem um erro menor do que para a velocidade.

Para a temperatura, é visto um comportamento anômalo dos dados de *GFS*, que se aproximam de zero na média ao meio dia, indicando um erro importante nos dados de previsão e que pode ter sido influenciado por *outliers*. Apesar desse erro, foi escolhido não mexer no conjunto de dados original para não correr o risco de perder o padrão de comportamento entre as variáveis de entrada e a variável alvo e não influenciar nos dados coletados no passado, que podem ou não serem os mesmos do futuro.

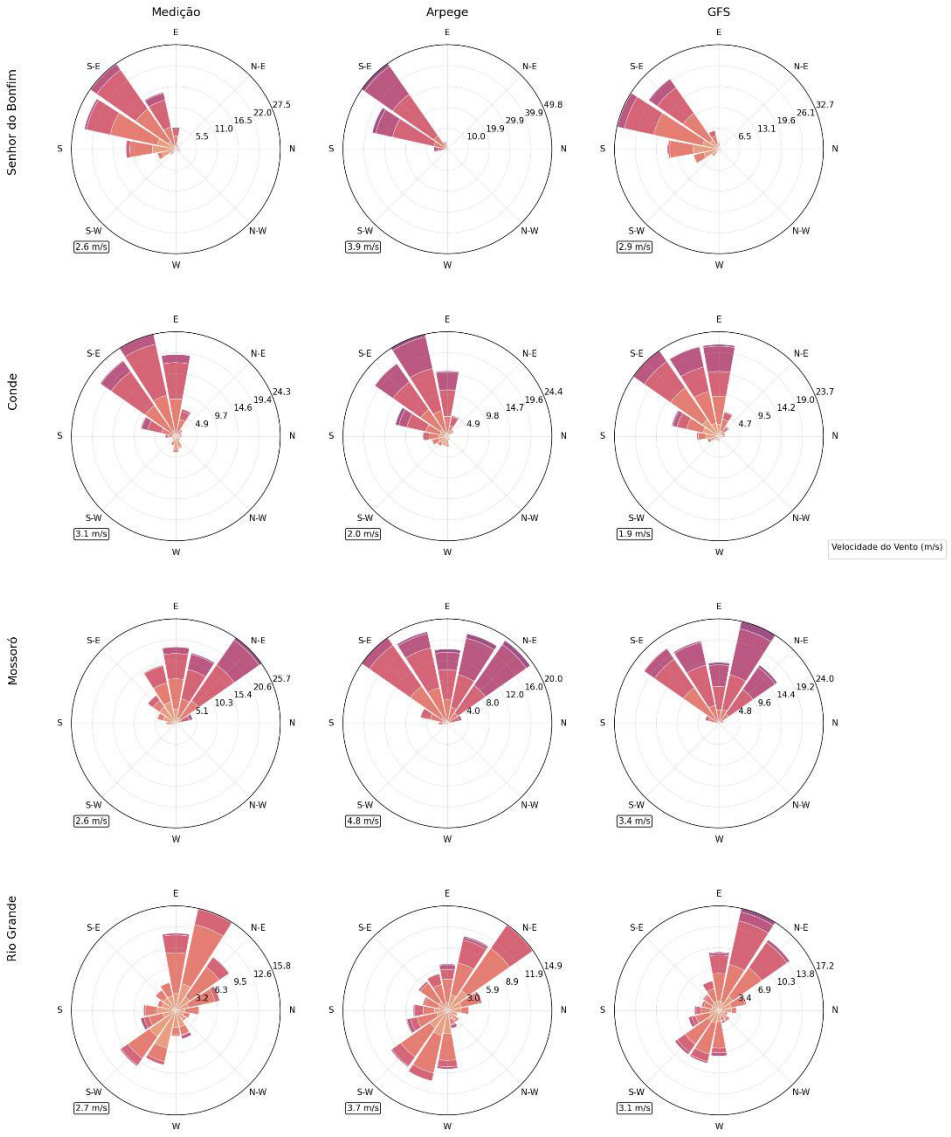
Para a pressão, em todas as localidades é confirmado mais uma vez que a previsão e os valores medidos possuem um erro pequeno.

Ao observar somente a média por hora da velocidade do vento da estação, é comum o comportamento de aumento no período da tarde, indicando um padrão de recurso eólico a partir de cerca de 12 horas. É importante destacar, também, que os dados de *ARPEGE* se mostraram mais enviesados do que *GFS* para todas as localidades com exceção de Conde.

A média horária da direção do vento possui um comportamento próximo para os três conjuntos de dados, apresentando um distanciamento maior da curva medida em Conde e Mossoró. Para a temperatura, o maior destaque é o comportamento falado anteriormente em Senhor do Bonfim, ademais é visto uma proximidade maior da curva dos dados da *ARPEGE* do que da *GFS*.

Na Figura 22 tem-se a rosa dos ventos para cada localidade e sua comparação entre valores medidos e previstos.

Figura 22. Comparação da rosa dos ventos com os dados de medição e dados de previsão para as quatro localidades.



Fonte: Elaborado pelo autor (2025).

Com os dados de velocidade e direção do vento de Senhor do Bonfim, *ARPEGE* e os valores medidos convergiram para uma predominância de ventos de Sudeste, enquanto que para *GFS*, apesar de estar próximo, obteve uma leve tendência ao Sul.

Em Conde, verifica-se o mesmo enviesamento por parte da previsão da *GFS*, enquanto que para Mossoró o erro de previsão foi conjunto. Em Rio Grande, *GFS* possui uma acurácia superior a *ARPEGE*, capturando uma direção de vento predominante entre Leste e Nordeste.

Os dados foram validados com o trabalho de Simão (2023), onde foi feita uma caracterização da direção predominante da velocidade média do vento no município de Mossoró. No estudo de Simão foi utilizado como fonte de dados a exata mesma estação meteorológica do INMET que foi utilizada no presente trabalho, com a diferença que o período de dados foi de 2008 a 2019 para Simão (2023).

Calculando a média da velocidade do vento coletada por Simão (2023) para os

meses de janeiro a junho, o resultado foi de aproximadamente 2,9 m/s, uma diferença de 0,3 m/s em relação ao que foi calculado neste trabalho – 2,6 m/s.

Em relação a rosa dos ventos, as direções foram bem semelhantes às calculadas e mostradas na Figura 22. Segundo Simão (2023, a direção predominante de ventos no primeiro semestre é de Nordeste (NE), corroborando com o resultado dos cálculos empregados no trabalho.

Após a análise preliminar, foi confirmado que as previsões numéricas possuem uma certa proximidade dos dados medidos, apesar dos erros associados, com alguns enviesamentos maiores em algumas variáveis e em algumas localidades específicas.

6.2 ENGENHARIA DE FEATURES

Após uma sólida análise preliminar, este tópico discorrerá sobre as técnicas de aprimoramento das features, a fim de criar um conjunto de dados mais informativo que crie um modelo com o menor erro possível para a previsão em um horizonte de 24 horas.

As variáveis de entrada (inputs) principais do modelo são a:

- direção do vento (10 metros de altitude) transformada para seno e cosseno;
- temperatura (10 metros de altitude) interpolada com os valores de temperatura nas altitudes fornecidas e informadas anteriormente;
- pressão atmosférica na superfície.

A variável alvo é a velocidade do vento a 10 metros de altitude.

Foram testadas algumas *features* temporais como o dia do ano e o mês do ano, porém não foram obtidos resultados satisfatórios. Sendo assim, uma das *features* temporais criadas e que serviu como base para a criação de outras foi a hora do dia. A hora é cíclica e varia de 0h a 24h, por isso ela também foi transformada em seno e cosseno utilizando as equações 7 e 8.

$$hora_{sin} = \sin(2\pi * \frac{hora}{24}) \tag{7}$$

$$hora_{cos} = \cos(2\pi * \frac{hora}{24}) \tag{8}$$

A hora parametrizada foi utilizada para a criação das outras features temporais como a sua interação com a temperatura. Para cada hora, existe um dado de temperatura, e essa temperatura normalmente armazena uma tendência de aumento ou de redução a depender se está amanhecendo ou anoitecendo. Essa relação foi calculada através da multiplicação entre o valor de temperatura e valor senoidal da hora associada. O mesmo cálculo foi repetido para

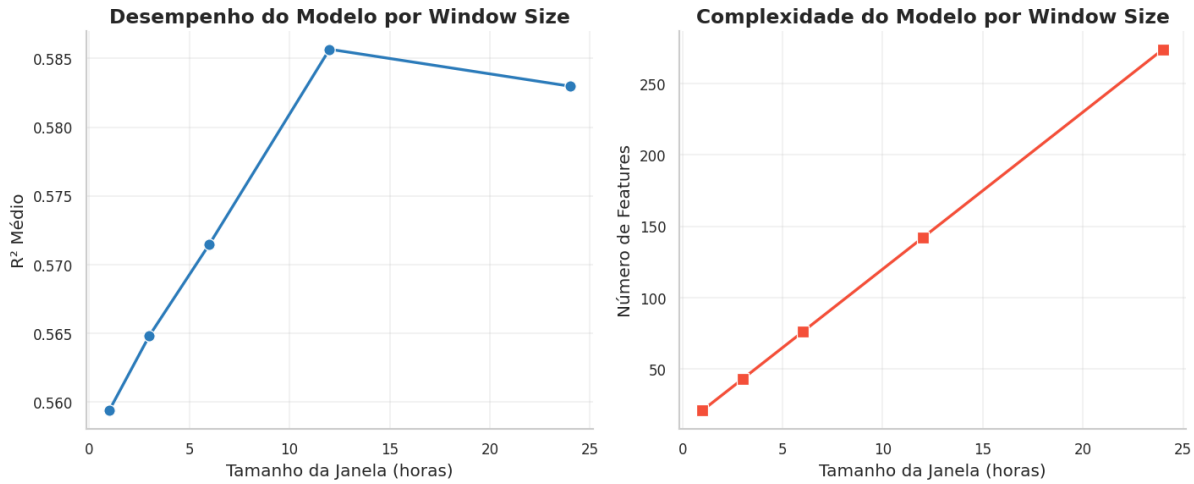
a interação entre a pressão atmosférica e a hora.

As interações entre os componentes cíclicos, direção do vento e hora do dia, visaram capturar a variação periódica da direção no ciclo diário. As componentes seno e cosseno da direção do vento foram multiplicadas somente por $hour_{sin}$, pois ela foi usada como base por questões empíricas, simplificando a dimensionalidade do problema e demonstrando bons resultados nos testes de validação.

Em resumo, isso adicionou mais cinco variáveis de entrada para o modelo, que agora conta com a direção do vento (seno e cosseno), a temperatura, a pressão atmosférica, a hora do dia (seno e cosseno), a interação entre a direção do vento e a hora do dia (seno e cosseno), a interação entre a temperatura e a hora do dia e a interação entre a pressão atmosférica e a hora do dia.

Após a escolha de quais as *features* temporais que seriam incrementadas, é necessário realizar a análise de qual o tamanho da janela (*window size*) das *lag features* que será utilizado no modelo. Como foi falado na metodologia, os valores de *window size* testados foram de 1, 3, 6, 12 e 24, e o resultado da iteração do treinamento do modelo com os dados de validação para cada valor é mostrado na Figura 23. Para cada tamanho da janela, era representado os valores passados, ou seja, 1 valor passado para cada variável para o tamaho de 1, 3 valores passados para o teste com a janela de 3, e assim sucessivamente.

Figura 23. Variação do R^2 do modelo de acordo com o tamanho das janelas de lag features.



Fonte: Elaborado pelo autor (2025).

No gráfico a esquerda pode-se observar o aumento da explicabilidade do modelo através do coeficiente de determinação (R^2) médio variando o tamanho da janela. O melhor coeficiente foi atingido para uma janela de 12 horas, obtendo um R^2 de 0,586. Com o aumento da janela para 24 horas, é visto que o R^2 começa a cair, piorando a performance do modelo mesmo com mais dados.

O gráfico a direita mostra o impacto na quantidade de inputs totais do modelo de

acordo com o tamanho da janela. Assim, verifica-se que o aumento do número de inputs não correspondeu ao aumento do coeficiente R^2, sendo a janela com 12 horas alcançando 142 features totais, quase 50% a menos que que a janela de 24 horas, que registrou 274 inputs criadas.

Em síntese, dentre as 142 inputs totais utilizadas para o treinamento do modelo, tem-se:

- As 4 variáveis principais;
- A hora parametrizada em seno e consseno;
- As interações entre os pares variável principal/hora do dia;
- Os dados das 12 horas anteriores para as 10 variáveis acima;
- Os dados das 12 horas anteriores da velocidade do vento, totalizando 142 inputs.

Com a etapa de *feature engineering* completa, pode-se passar a etapa de ajuste dos hiperparâmetros do *XGBoost*.

6.3 OTIMIZAÇÃO DOS HIPERPARÂMETROS

Esta etapa se concentra em apresentar os resultados que foram encontrados para a otimização dos hiperparâmetros do algortimo. Como foi falado anteriormente na metodologia, as combinações foram geradas através do método de pesquisa em grade *GridSearchCV*, disponível na biblioteca python *SciKit-Learn*.

Ao total foram geradas 60 combinações e, para o cálculo de todas as combinações possíveis, foi cronometrado um tempo total de 42 minutos e 10 segundos. A Tabela 9 mostra as 10 combinações com os menores valores de *RMSE* em ordem decrescente.

Tabela 9. Classificação do RMSE para as 10 combinações mais performantes de hiperparâmetros.

***RMSE* médio**	***learning_rate***	***max_depth***	***n_estimators***
0,776	0,01	4	500
0,778	0,01	4	1000
0,779	0,01	5	500
0,780	0,01	3	1000
0,780	0,01	3	500
0,783	0,01	5	1000
0,783	0,10	3	100
0,784	0,10	4	100
0,785	0,01	2	500
0,785	0,10	2	100

Fonte: Elaborado pelo autor (2025).

Observa-se que a maioria das combinações do ranking possuem uma taxa de aprendizado de 0,01, sendo a melhor combinação aliando o valor de 0,01 de *learning_rate* com um *max_depth* de 4 e 500 *n_estimators*, alcançando um *RMSE* médio de 0,776.

Para analisar o impacto de cada um individualmente, foi feito um agrupamento por parâmetro e por erro quadrático médio e calculado a média e desvio padrão associado, ou seja, para uma taxa de aprendizado de 0,01, foi calculado a média de todos os *RMSE* e desvios. A Tabela 10 ilustra todos os valores encontrados.

Tabela 10. RMSE e desvio padrão médio de cada hiperparâmetro.

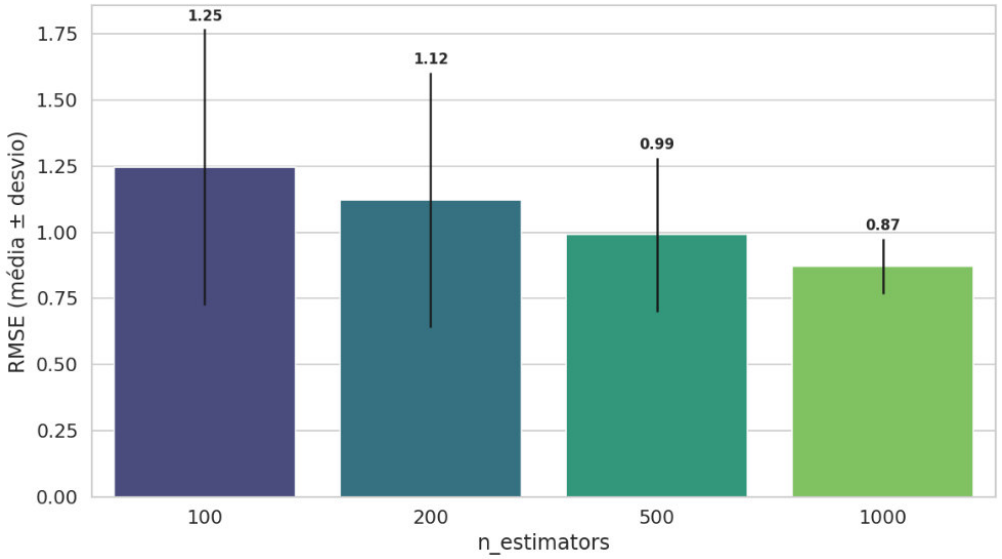
Hiperparâmetro	**Valor**	***RMSE***	**Desvio padrão**
n_estimators	100	1,248	+- 0,521
n_estimators	200	1,122	+- 0,482
n_estimators	500	0,991	+- 0,290
n_estimators	1000	0,870	+- 0,105
max_depth	1	1,056	+- 0,418
max_depth	2	1,053	+- 0,412
max_depth	3	1,057	+- 0,417
max_depth	4	1,059	+- 0,419
max_depth	5	1,063	+- 0,421
learning_rate	0,001	1,532	+- 0,372
learning_rate	0,01	0,842	+- 0,100
learning_rate	0,1	0,799	+- 0,015

Fonte: Elaborado pelo autor (2025).

Pode-se observar que o parâmetro *max_depth* possui uma influência menor no *RMSE* do modelo, visto que o mesmo não varia tanto com o aumento ou redução da profundidade das árvores. O mesmo ocorre para o desvio padrão, que também possui valores similares para diferentes valores, indicando uma baixa variabilidade.

De forma a poder observar a média e os desvios de uma forma mais visual, foi feito um gráfico de barras do valor do *RMSE* e do seu desvio padrão para *n_estimators* e *learning_rate*. A base da barra representa o valor médio do *RMSE*, e o valor exato da barra está ilustrado na linha. A ponta final da linha representa o valor do da soma entre o erro mais o seu desvio padrão. A Figura 24 ilustra o gráfico para o hiperparâmetro *n_estimators*.

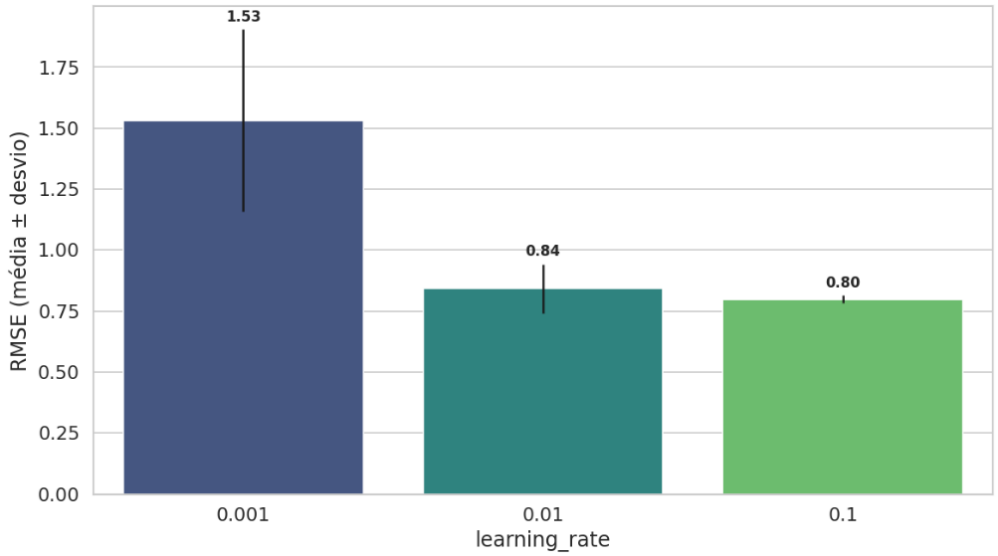
Figura 24. Impacto do n_estimators através da média e desvio padrão do RMSE.



Fonte: Elaborado pelo autor (2025).

Através dessa Figura pode-se observar que com o aumento do número de árvores, obtem-se um menor erro quadrático. O desvio padrão é menor para 1000 árvores, atingindo uma soma com o *RMSE* próximo a 1,0. A maior soma é para 100 árvores criadas, com um *RMSE* de 1,25 e uma linha extensa que atinge um pouco mais que 1,75, indicando um desvio mais alto e um maior nível de incerteza no resultado. A Figura 25 mostra o gráfico para a taxa de aprendizado.

Figura 25. Impacto da learning_rate através da média e desvio padrão do RMSE.



Fonte: Elaborado pelo autor (2025).

Para a taxa de aprendizado, observa-se a soma da média do *RMSE* e o seu desvio muito maior para 0,001, indicando que o modelo ficou lento para aprender o comportamento dos dados. Para *learning_rate* de 0,01 e 0,1 os valores da média do *RMSE* ficaram bem próximos, de 0,84 e 0,80, sendo que para 0,01 a linha está um pouco mais extensa, indicando

um desvio padrão levemente maior.

É importante destacar que a análise individual ajuda a ter uma noção de como os hiperparâmetros se comportam e influenciam o resultado, mas na prática, o modelo é treinado em conjunto. Isso significa que quando eles são combinados, não necessariamente o menor *RMSE* individual para *learning_rate* e *n_estimators* irá resultar em uma combinação final com melhor desempenho no teste ou na validação.

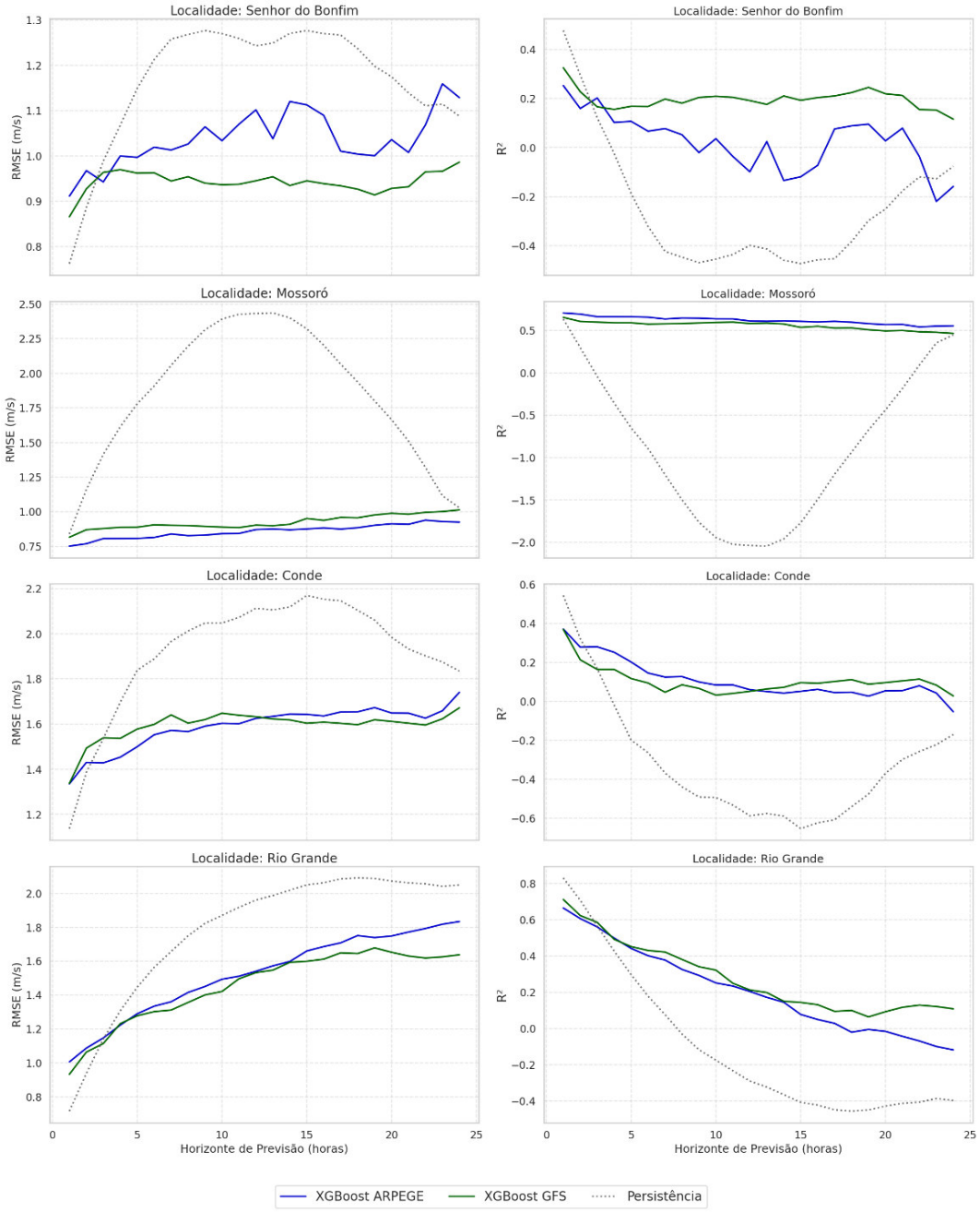
Enfim, com o menor *RMSE* para uma taxa de aprendizado de 0,01, um *n_estimators* de 500 e profundidade máxima de 4, o treinamento do XGBoost foi feito para *ARPEGE* e *GFS* para a avaliação final dos resultados.

6.4 AVALIAÇÃO DOS RESULTADOS DO MODELO

O treinamento foi realizado de forma definitiva com a sua avaliação sendo baseada no resultado dos 20% de dados finais do conjunto de teste, que não foram vistos pelo modelo em nenhum momento.

Como o horizonte de previsão é de 24 horas, para cada horizonte de tempo calcula-se um *RMSE* e assim pode-se criar um gráfico do erro quadrático médio por horizonte do tempo. A Figura 26 mostra essa análise para as quatro localidades, comparando os modelos de *XGBoost Arpege*, *XGBoost GFS* e Persistência.

Figura 26. Evolução do RMSE dos modelos com o horizonte de previsão para as quatro localidades do trabalho.



Fonte: Elaborado pelo autor (2025).

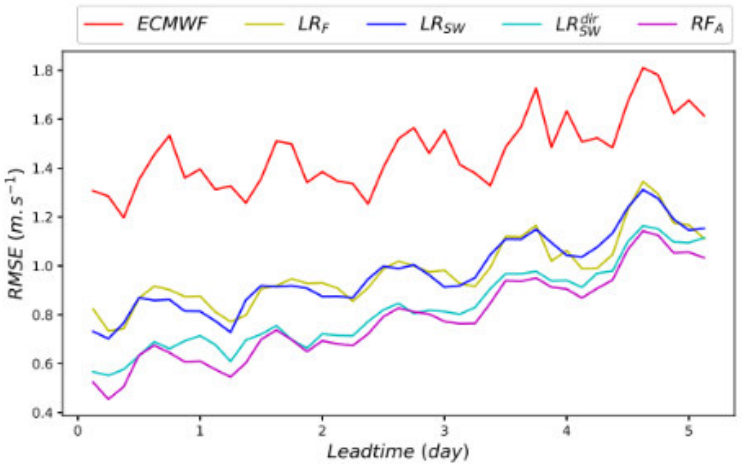
Pode-se observar que de fato, o *RMSE* aumenta com o aumento do horizonte de previsão. Assim como o erro, o R^2 também tende a diminuir com o horizonte, confirmando que o modelo perde explicabilidade quanto mais do futuro se tenta prever. O modelo *XGBoost GFS* possuiu um erro menor e uma explicabilidade maior que o *XGBoost Arpege* em Senhor do Bonfim e em Rio Grande, enquanto que para Conde o algortimo treinado com dados de Arpege foi superior. Em Mossoró, *XGBoost Arpege* possuiu erros menores para horizontes de previsões até em torno de 12 horas, após isso o erro foi maior que o do modelo treinado pelos dados da previsão *GFS*.

Em relação ao *benchmarking* padrão, é verificado que em todos os casos, tanto para o *RMSE* quando para o R^2, os algoritmos de *machine learning* obtiveram uma melhor

performance que o método da Persistência, com exceção de horizontes de previsão muito curtos de 1 ou 2 horas, em que o modelo da Persistência possui um bom resultado no geral, mas piora exponencialmente logo depois.

Para fins de *benchmarking*, a Figura 27 mostra o resultado de Dupré et al. (2017) para a previsão da velocidade do vento em 10 metros, onde foi utilizado o algortimo *Random Forest* para a melhora da previsão. O resultado é comparado com diferentes tipos de regressões lineares e dados do *NWP* europeu *ECMWF*.

Figura 27. Evolução do RMSE obtido com o passar dos dias para previsão da velocidade do vento na altitude de 10 metros.

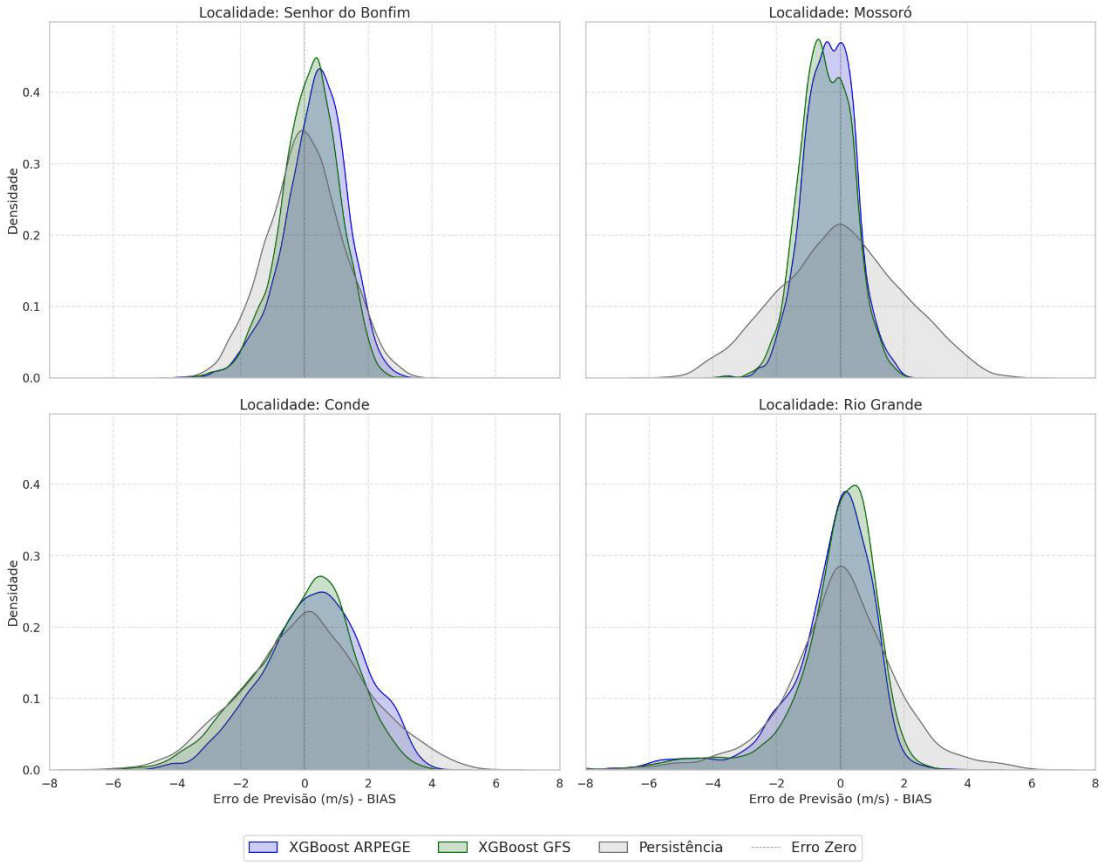


Fonte: Dupré et al. (2017).

Para o *leadtime* de 1 dia, o método *Random Forest* registrou em torno de 0,6 m/s de erro quadrático, enquanto que o *ECMWF* estava acima de 1,2 m/s. Comparando os modelos com o resultado da Figura 26 e utilizando como exemplo a cidade de Mossoró, em ambos os algoritmos o *RMSE* variou de 0,75 m/s a 1 m/s, se mostrando próximo ao resultado do algortimo obtido em Dupré et al. (2017) para o horizonte de previsão de 24 horas, mas menos performante.

A Figura 28 mostra os histogramas do viés (*Bias*) para os três métodos. O viés ou *Bias*, é um erro simples calculado pela subtração entre o valor predito e o valor medido.

Figura 28. Histograma do viés dos três métodos analisados para cada localidade estudada.



Fonte: Elaborado pelo autor (2025).

Através da Figura 28, pode-se ratificar a imprecisão do método da Persistência para Mossoró. O histograma está deveras achatado para Conde e para Mossoró, mostrando uma variabilidade alta de ocorrência dos erros, maior inclusive que para os algoritmos calculados. Em Senhor do Bonfim e em Rio Grande também pode-se observar alguns outliers para o método da Persistência, pois a curva atinge pontos mais extremos que os algoritmos.

Analisando o viés de *XGBoost Arpege*, há uma pequena superestimação da velocidade do vento em Senhor do Bonfim, Conde e Rio Grande. Para *XGBoost GFS*, a maior densidade de pontos erráticos também está superestimada. Os histogramas possuem um comportamento de distribuição normal, bem próximos a linha de erro zero, confirmando o sucesso do modelo.

No apêndice B.1 a análise feita é do erro viés em relação á velocidade do vento real para os dois algoritimos, onde pode ser verificado um destaque de outliers para a localidade de Rio Grande e erros mais próximos de zero para Mossoró.

Nas Tabelas 11 e 12 foram calculadas a média do *RMSE* e do R^2 para os três métodos discutidos até o momento, além do mesmo cálculo no mesmo período de teste para as previsões de *ARPEGE* e *GFS*. Foi optado pela exposição do melhor caso e do pior caso, sendo as outras duas tabelas expostas no apêndice B.2. Para a definição do melhor caso foi escolhido a localidade com o menor erro entre os algortimos e o pior caso reflete o erro mais

considerável deles.

Tabela 11. RMSE e R² médio de todos os métodos para Mossoró.

Método	***RMSE* (m/s)**	***R²***
XGBoost Arpege	0,86	0,62
XGBoost GFS	0,92	0,56
Persistência	1,85	-0,89
ARPEGE	3,20	-4,34
GFS	1,69	-0,49

Fonte: Elaborado pelo autor (2025).

Pode-se observar na Tabela 11 uma melhora expressiva das métricas das previsões numéricas *ARPEGE* e *GFS*. Principalmente para os dados de previsão de *ARPEGE*, que antes registraram um *RMSE* de 3,20 m/s, reduziram 73,13%, chegando a um *RMSE* de 0,86 m/s. A utilização do *XGBoost* também melhorou a explicabilidade do modelo, que era muito negativa, e saltou para 0,62, o maior R^2 dentre todas as localidades e todos os métodos.

A melhora também ocorreu para *GFS*, a qual teve uma melhora de 45,45% do seu *RMSE* (0,92 m/s) e registrou um alto coeficiente de determinação (0,56). Em relação a Persistência, *XGBoost GFS* performou 45,56% melhor utilizando como premissa a diminuição do erro quadrático médio. Um desempenho comparável com *XGBoost Arpege*, que obteve um *RMSE* 53,5% menor que o método da Persistência.

Tabela 12. RMSE e R² médio de todos os métodos para Conde.

Método	***RMSE* (m/s)**	***R²***
XGBoost Arpege	1,59	0,11
XGBoost GFS	1,59	0,11
Persistência	1,92	-0,32
Arpege	1,78	-0,11
GFS	1,94	-0,32

Fonte: Elaborado pelo autor (2025).

Em Conde foram observados os maiores erros para os algoritmos de aprendizado de máquina, como também erros altos e explicabilidades baixas para os outros métodos. A melhora entre a previsão original utilizando *machine learning* ainda é existente, reduzindo o *RMSE* de 1,78 m/s para 1,59 m/s para o conjunto de *ARPEGE* e de 1,94 m/s para 1,59 m/s para o conjunto de *GFS*.

Uma otimização em relação ao método da Persistência também foi alcançada para

os dois modelos, mostrando que mesmo com uma possível propensão da localidade à registrar incertezas mais significativas, o algortimo conseguiu desempenhar melhor que os outros métodos.

Em síntese, foi observado na grande maioria dos casos uma melhora significativa na previsão da velocidade do vento em um horizonte de previsão de 24 horas utilizando o algoritmo *XGBoost*, superando o método da Persistência na maioria dos horizontes de previsão e reduzindo o erro das previsões numéricas para praticamente todas as localidades, confirmando, portanto, o sucesso nos objetivos gerais e específicos citados anteriormente.

7 CONCLUSÃO

Através deste trabalho foi possível alcançar resultados importantes e cumpriu o objetivo geral e os objetivos específicos propostos. As investigações preliminares mostraram as tendências existentes nos modelos de Previsão Numérica do Tempo, que por muitas vezes superestimaram ou subestimaram a variável meteorológica analisada. Apesar dessa tendência na maioria das variáveis, o comportamento é assertivo principalmente para a pressão atmosférica.

Com a implementação de testes para as *features* temporais e uso exaustivo do conjunto de validação para validar o tamanho das janelas (*window size*) usadas, o resultado se mostrou significativo no resultado do modelo. Através da análise de validação foi encontrado um valor ótimo de 12 para o tamanho da janela que captura as variáveis passadas e a incrementa na predição do valor futuro, assim como o uso de variáveis cíclicas como a hora do dia se mostrou uma boa prática para resultados melhores no modelo. A análise da hora do dia foi crucial não só por ela em si, mas pela criação da sua interação com as outras variáveis, sendo possível capturar padrões diurnos e noturnos e fornecer mais informação no treinamento do algoritmo.

A otimização dos hiperparâmetros validou o fato de que a maior influência do modelo originou-se da taxa de aprendizado e do *n_estimators*, sendo a profundidade das árvores a menor influência na sua performance. Com o uso da metodologia *GridSearch*, foi feito o cálculo das 60 combinações geradas e a melhor configuração registrada foi para 0,01 de *learning_rate*, 500 *n_estimators* e 4 *max_depth*.

Os resultados cumpriram o objetivo geral e, uma aprimoração da previsão da velocidade do vento em relação aos modelos *NWP* originais foi alcançada. Além do *RMSE* e do R^2 ter sido melhor para os algoritmos *XGBoost*, este também se adequou melhor que o método da Persistência em praticamente todos os horizontes de previsão para todas as localidades estudadas. Mossoró registrou as melhores performances do modelo, alcançando uma melhora de 73,13% para *XGBoost Arpege* e de 45,45% para *XGBoost GFS* em relação às suas previsões originais. Alcançando um coeficiente de determinação (R^2) acima de 0,55, os dois algoritmos registraram uma redução de mais de 40% do *RMSE* médio se comparados ao método da Persistência no mesmo período.

Por fim, conclui-se que o uso do algoritmo *XGBoost* possibilitou a melhora da previsão da velocidade do vento para o horizonte de previsão de 24 horas. Os trabalhos futuros poderiam visar em reduzir ainda mais o erro do modelo através de uma maior base de dados para o treinamento do modelo, adicionando outros algortimos de aprendizado de máquina para comparação com o *XGBoost* e podendo também fomentar uma comparação entre a performance da previsão à altitude de 10 metros e da previsão à altitude de 100 metros.

8 REFERÊNCIAS

PAROLINI, A. *Weather, climate, and agriculture: Historical contributions and perspectives from agricultural meteorology*. *WIREs Climate Change*, v. 13, n. 3, 2022.

SHEN, D.; Shi, W.-F.; Tang, W.; Wang, Y.; Liao, J. *The Agricultural Economic Value of Weather Forecasting in China*. *Sustainability* 2022, *14*, 17026.

SHEN, D & Zuo, zhengyu & Zhang, Xiaofeng & Zhao, Xinyu. (2023). The impact of weather forecast accuracy on the economic value of weather-sensitive industries. 10.21203/rs.3.rs-3306307/v1.

ROBERTS, Justo & Cassula, Agnelo & Hauer, Ines. (2014). Electricity Consumption Characterization Of Different End-Use Sectors Of Brazil.

LYNCH, P. The emergence of numerical weather prediction: Richardson's dream. Cambridge: Cambridge University Press, 2006.

HAM, Y.-G.; KIM, J.-H.; LUO, J.-J. A Bayesian Deep Learning Approach to Near-Term Climate Prediction. Journal of Advances in Modeling Earth Systems, [s.l.], v. 15, n. 2, e2022MS003058, 2023. Disponível em: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS003058. Acesso em: 4 jul. 2025.

LEI, M. et al. A review on the forecasting of wind speed and generated power. Renewable and Sustainable Energy Reviews, v. 13, n. 4, p. 915-920, 2009.

PAPOULIS, A.; PILLAI, S. U. Probability, random variables, and stochastic processes. 4. ed. New York: McGraw-Hill, 2002.

WANG, H. et al. A review of deep learning for renewable energy forecasting. Energy Conversion and Management, v. 198, p. 111799, 2019.

AHRENS, C. D. Essentials of Meteorology: An Invitation to the Atmosphere. 8. ed. Boston, MA: Cengage Learning, 2018.

BRAGA, Rafael. Ventos, o que são, como se formam, tipos e importância. Conhecimento Científico, 20 maio 2022. Disponível em: https://conhecimentocientifico.r7.com/ventos-o-que-sao-como-se-formam/. Acesso em: 05 jul. 2025.

MANWELL, J. F.; MCGOWAN, J. G.; ROGERS, A. L. Wind Energy Explained: Theory, Design and Application. 2. ed. Chichester: John Wiley & Sons, 2009.

STENSRUD, D. J. Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models. Cambridge: Cambridge University Press, 2007.

de Oliveira-Júnior, José & Terassi, Paulo & Gois, Givanildo. (2017). Estudo da circulação dos ventos na Baía de Guanabara/RJ, entre 2003 e 2013. Revista Brasileira de climatologia. 21. 59-80.

CHAMMI REDDY, A. K. R. et al. Development of wind rose diagrams for coal used thermal power plant Kadapa, India. International Journal of ChemTech Research, v. 9, n. 2, p. 60–64, 2016. Disponível em: https://sphinxsai.com/2016/ch_vol9_no2/1/(60-64)V9N2CT.pdf. Acesso em: 6 jul. 2025.

SANTOS, F. S. et al. Análise estatística da velocidade do vento em Petrolina-PE utilizando as

distribuições Weibull e Burr. Journal of Environmental Analysis and Progress, v. 4, n. 1, p. 28-36, 2019. DOI: 10.24221/jeap.4.1.2019.2057.

CONCEIÇÃO, G. C. P. da; SOUZA, A. P. de; BARROS, N. F. de. Análise da velocidade e direção dos ventos em Cruz das Almas, Bahia, no período 1973-2001. Revista Semiárido De Visu, v. 1, n. 1, p. 41-47, 2011. DOI: 10.24864/rsdv.v1i1.107.

SOUZA, A. de F. F. de et al. Modelagem da velocidade do vento usando metodologias ARIMA, Holt-Winters e RNA na previsão de geração eólica no Nordeste Brasileiro. Revista Brasileira de Climatologia, v. 18, p. 200-218, 2016. DOI: 10.5380/abclima.v18i0.48565.

RAMACHANDRAN, G. et al. Techno-Economic Investigation of Wind Energy Potential in Selected Sites with Uncertainty Factors. Sustainability, v. 13, n. 4, p. 2182, 2020. Disponível em: <https://www.mdpi.com/2071-1050/13/4/2182>. Acesso em: 6 jul. 2025.

RENKEER. Meteorological Station Introduction and Types. [S. l.], 2024. Disponível em: https://www.renkeer.com/meteorological-station-introduction/. Acesso em: 6 jul. 2025.

INSTITUTO NACIONAL DE METEOROLOGIA (INMET). Rede de Estações Meteorológicas Automáticas do INMET. Nota Técnica nº 001/2011/SEGER/LAIME/CSC/INMET. Brasília: INMET, 2011. Disponível em: https://portal.inmet.gov.br/uploads/notas_tecnicas/NT_REDE_AUTOMATICA.pdf. Acesso em: 6 jul. 2025.

BRASIL. Ministério da Agricultura e Pecuária. INMET – 115 anos. Disponível em: https://www.gov.br/agricultura/pt-br/campanhas/inmet-115-anos. Acesso em: 6 jul. 2025.

PU, Z.; KALNAY, E. Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation. [S.l.]: Universidade de Utah, 2018. Disponível em: https://www.inscc.utah.edu/~pu/6500_sp12/Pu-Kalnay2018_NWP_basics.pdf. Acesso em: 6 jul. 2025.

NOAA. The History of Numerical Weather Prediction. In: Mariners Weather Log, v. 51, n. 3, dez. 2007. Washington, DC: NOAA. Disponível em: https://www.vos.noaa.gov/MWL/dec_07/weatherprediction.shtml. Acesso em: 6 jul. 2025.

LYNCH, Peter. Weather and climate forecasting: a chronicle of a revolution. World Meteorological Organization, 1 nov. 2010. Disponível em: https://public.wmo.int/media/magazine-article/weather-and-climate-forecasting-chronicle-of-revolution. Acesso em: 6 jul. 2025.

MARQUES, Clara. Brasil perde a chance de liderar debates sobre uso de inteligência artificial na transição energética. Climate Tracker, 25 jul. 2024. Disponível em: https://climatetrackerlatam.org/historias/brasil-perde-a-chance-de-liderar-debates-sobre-uso-de-inteligencia-artificial-na-transicao-energetica/. Acesso em: 6 jul. 2025.

LYNCH, P. Numerical Weather Prediction: The Basic Equations. In: Pryor, S. (Ed.), Numerical Weather and Climate Prediction. [S.l.]: University College Dublin, 2017. Cap. 2, p. 11–20. Disponível em: https://maths.ucd.ie/~plynch/Publications/pcam0159_proof_2.pdf. Acesso em: 6 jul. 2025.

MESINGER, F. Regional NWP Modeling and Predictability, Introduction. Trieste: ICTP, 11–22 Apr. 2005. Disponível em: https://indico.ictp.it/event/a04186/session/5/contribution/1/material/0/0.pdf. Acesso em: 6 jul. 2025.

RAMS—Regional Atmospheric Modeling System. Numerical Weather Prediction: lateral, top and lower boundary conditions, cap. 7.10. In: RAMS Documentation, 2003. Disponível em: https://rams.atmos.colostate.edu/at540/fall03/fall03Pt7.pdf. Acesso em: 6 jul. 2025.

MONTEIRO, C. et al. Wind Power Forecasting: State-of-the-Art 2009. Argonne: Argonne National Laboratory, 2009.

CHANG, W.-Y. A Literature Review of Wind Forecasting Methods. Journal of Power and Energy Engineering, v. 02, n. 04, p. 161–168, 2014.

ALKESAIBERI, A.; HARROU, F.; SUN, Y. Efficient Wind Power Prediction Using Machine Learning Methods: A Comparative Study. Energies, v. 15, n. 13, p. 5246, 2022. DOI: 10.3390/en15135246.

HANIFI, S. et al. A Critical Review of Wind Power Forecasting Methods—Past, Present and Future. Energies, v. 13, n. 15, p. 3784, 2020. DOI: 10.3390/en13153784.

FOLEY, A. M. et al. Current methods and advances in forecasting of wind power generation. Renewable Energy, v. 45, p. 143-154, 2011. DOI: 10.1016/j.renene.2012.02.016.

DUPRÉ, A. et al. Sub-hourly forecasting of wind speed and wind energy. Renewable Energy, v. 145, p. 2373-2379, 2020. DOI: 10.1016/j.renene.2019.07.161.

PIELKE, R. A. Mesoscale Meteorological Modeling. 2. ed. San Diego: Academic Press, 2002.

RUSSELL, S.; NORVIG, P. Inteligência artificial: uma abordagem moderna. 3. ed. São Paulo: Pearson Education do Brasil, 2013.

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

RATHOD, Jash. Underfitting, Overfitting and Regularization in Machine Learning. 2021. Disponível em: https://jashrathod.github.io/2021-09-30-underfitting-overfitting-and-regularization/. Acesso em: 8 jul. 2025.

BENSON, B.; PAN, W.; PRASAD, A.; GARY, G.; HU, Q. Forecasting Solar Cycle 25 Using Deep Neural Networks. Solar Physics, [S.l.], v. 295, 2020. DOI: 10.1007/s11207-020-01634-y.

LAZZERI, Francesca. Machine learning for time series forecasting with Python. Hoboken: John Wiley & Sons, 2020.

HAYKIN, Simon. Neural Networks: A Comprehensive Foundation. 2. ed. Upper Saddle River: Prentice Hall, 2001.

GROSSBERG, Stephen. Neural Networks and Natural Intelligence. Cambridge: MIT Press, 1988.

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. 3. ed. Sebastopol: O'Reilly Media, 2022.

GEEKSFORGEEKS. Layers in Artificial Neural Networks (ANN). 2024. Disponível em:

https://www.geeksforgeeks.org/deep-learning/layers-in-artificial-neural-networks-ann/. Acesso em: 8 jul. 2025.

GUNJI, Malleswara Rao. Intuition - Decision tree, Ensemble, Regression Ananlysis. LinkedIn, 25 jul. 2020. Disponível em: https://www.linkedin.com/pulse/intuition-decision-tree-ensemble-regression-ananlysis-gunji/. Acesso em: 8 jul. 2025.

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

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. In: PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2016, San Francisco. New York: ACM, 2016. p. 785-794.

REICHSTEIN, M. et al. Deep learning and process understanding for self-explaining earth system science. Nature, v. 566, n. 7743, p. 195-204, 2019. DOI: 10.1038/s41586-019-0912-1.

BARRIOS, A. F.; LORENZO, M. S.; RODRÍGUEZ, A. E. R. LSTM Model for Wind Speed and Power Generation Nowcasting. In: INTERNATIONAL ELECTRONIC CONFERENCE ON ATMOSPHERE, 2., 2022, [online]. Proceedings... Basel: MDPI, 2022. Art. 30.

SHIN, J.-Y.; MIN, B.; KIM, K. R. High-resolution wind speed forecast system coupling numerical weather prediction and machine learning for agricultural studies — a case study from South Korea. International Journal of Biometeorology, v. 66, p. 1429–1443, 2022. DOI: 10.1007/s00484-022-02287-1.

GIEBEL, G.; KARINIOTAKIS, G. Wind power forecasting: A review of the state of the art. In: KARINIOTAKIS, G. (Ed.). Renewable Energy Forecasting: From Models to Applications. Cambridge: Woodhead Publishing, 2017. p. 27-84.

FARIAS, J. G. de. Previsão de geração de energia eólica em múltiplos horizontes utilizando modelos de machine learning. 2020. 104 f. Dissertação (Mestrado em Engenharia Mecânica) – Universidade Federal de Santa Catarina, Florianópolis, 2020.

BAGGIO, Roberta; PUJOL, Killian; PANTILLON, Florian; LAMBERT, Dominique; FILIPPI, Jean-Baptiste; MUZY, Jean-François. Local wind speed forecasting at short time horizons relying on both Numerical Weather Prediction and observations from surrounding station. arXiv, [s.l.], 2025.

NOAA. GFS Documentation – EMC Virtual Lab. 2025. Disponível em: https://vlab.noaa.gov/web/emc/gfs. Acesso em: 8 jul. 2025.

GFS. NOAA GFS on AWS – Registry of Open Data. 2022. Disponível em: https://registry.opendata.aws/noaa-gfs-bdp-pds/. Acesso em: 8 jul. 2025.

NATIONAL WEATHER SERVICE. How Do We Use Models in Our Forecasting?. Silver Spring, 2023. Disponível em: https://www.weather.gov/ilx/about_models. Acesso em: 9 jul. 2025.

CNRM – CENTRE NATIONAL DE RECHERCHES MÉTÉOROLOGIQUES. ARPEGE – Global model description. Météo-France, 2022. Disponível em: https://www.umr-cnrm.fr/spip.php?article121=&lang=en. Acesso em: 8 jul. 2025.

RENEAU, Alex; HU, Jerry Yao-Chieh; GILANI, Ammar; LIU, Han. Feature programming for multivariate time series prediction. In: PROCEEDINGS OF MACHINE LEARNING

RESEARCH (PMLR), v. 202, p. 1–18, 2023.

GORDON, Aaron. A Practical Guide for Feature Engineering of Time Series Data. dotData Blog, 20 jun. 2023. Disponível em: https://dotdata.com/blog/practical-guide-for-feature-engineering-of-time-series-data/. Acesso em: 9 jul. 2025.

BANSAL, Aayam; BALAJI, Keertan; LALANI, Zeus. Temporal encoding strategies for energy time series prediction. arXiv, Mar. 2025.

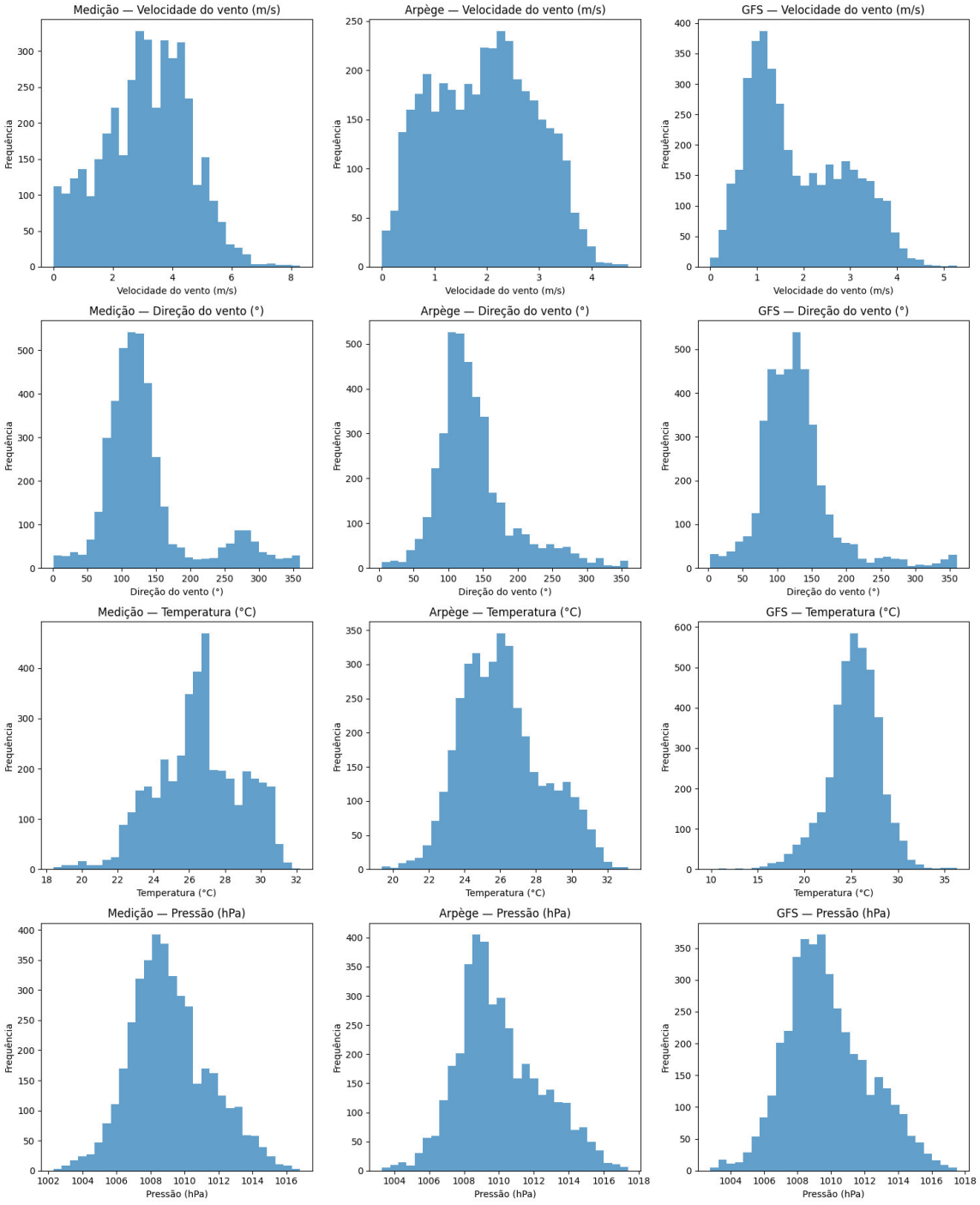
CLARK, T. E.; MCCRACKEN, M. W.; MERTENS, E. Modeling Time-Varying Uncertainty of Multiple-Horizon Forecast Errors. BIS Working Paper, n. 667, 2017. Disponível em: https://www.bis.org/publ/work667.htm. Acesso em: 4 jul. 2025.

SIMÃO, Kadidja Meyre Bessa. Caracterização da direção predominante e a velocidade média do vento no município de Mossoró-RN. 2023. Trabalho de Conclusão de Curso (Graduação em Engenharia Agrícola e Ambiental) – Universidade Federal Rural do Semi-Árido, Centro de Engenharias e Ciências Ambientais, Mossoró, 2023.

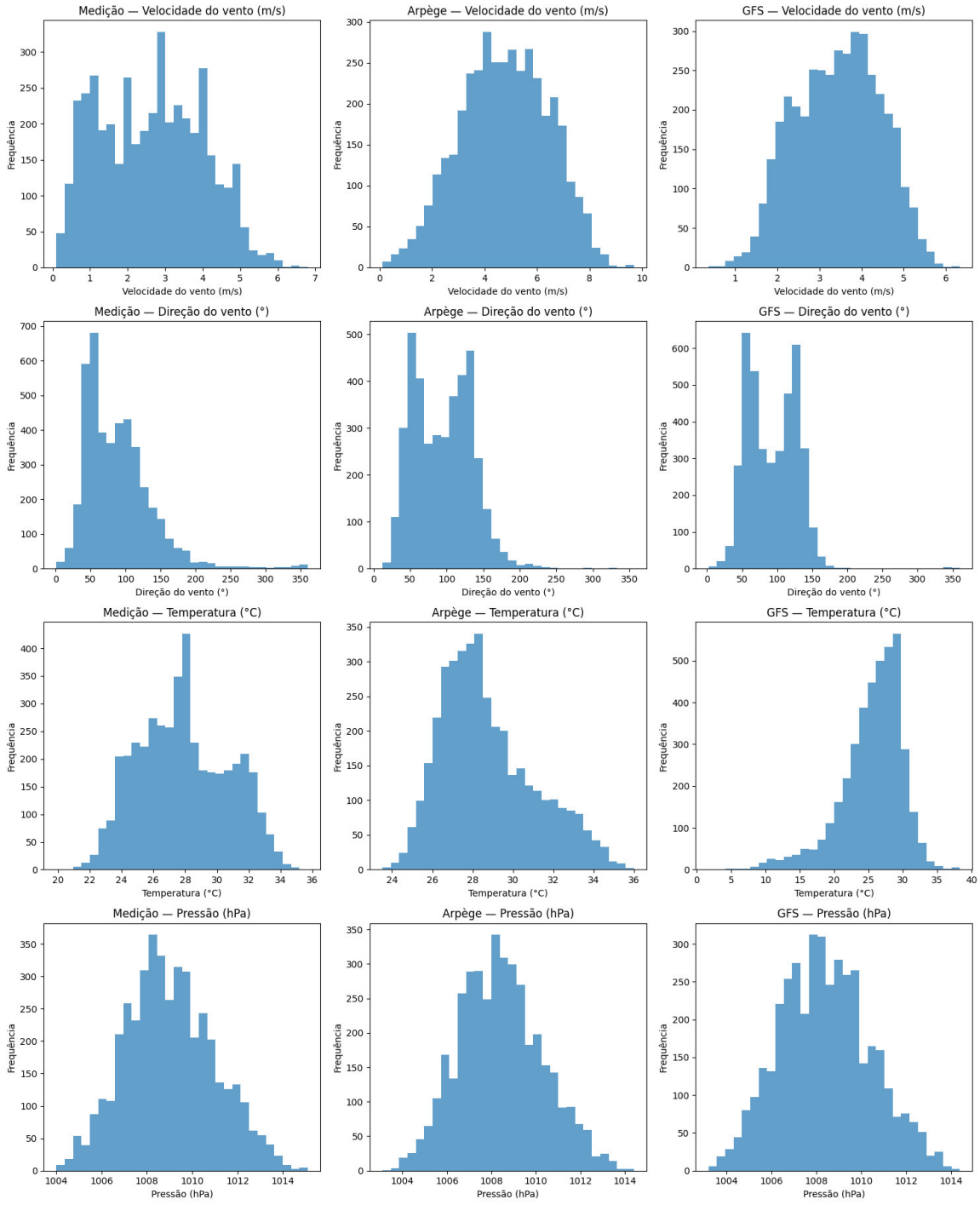
9 APÊNDICE A – ANÁLISE PRELIMINAR DOS DADOS

9.1 A.1 HISTOGRAMAS

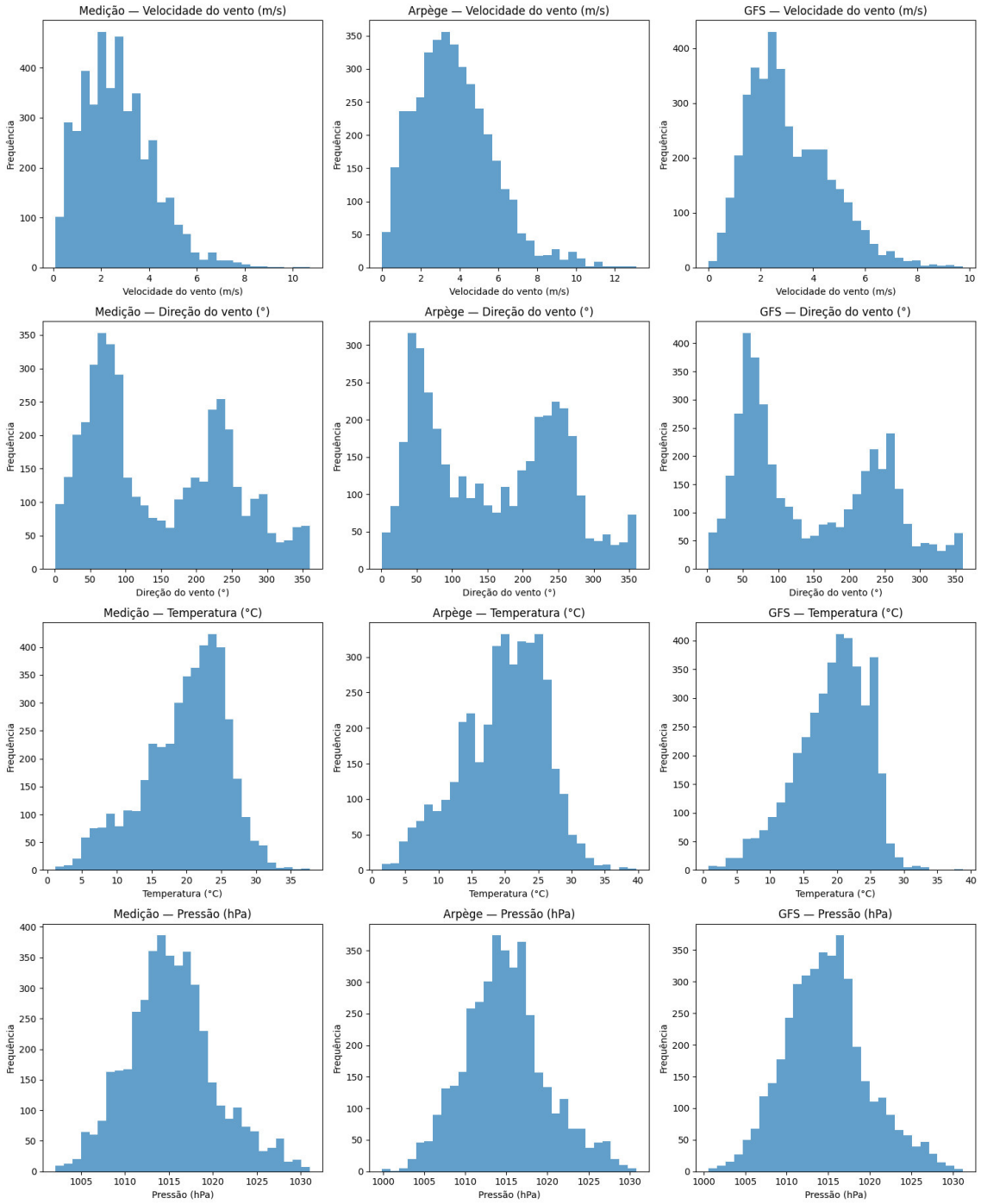
Histograma dos dados das variáveis meteorológicas analisadas para Conde.



Histograma dos dados das variáveis meteorológicas analisadas para Mossoró.

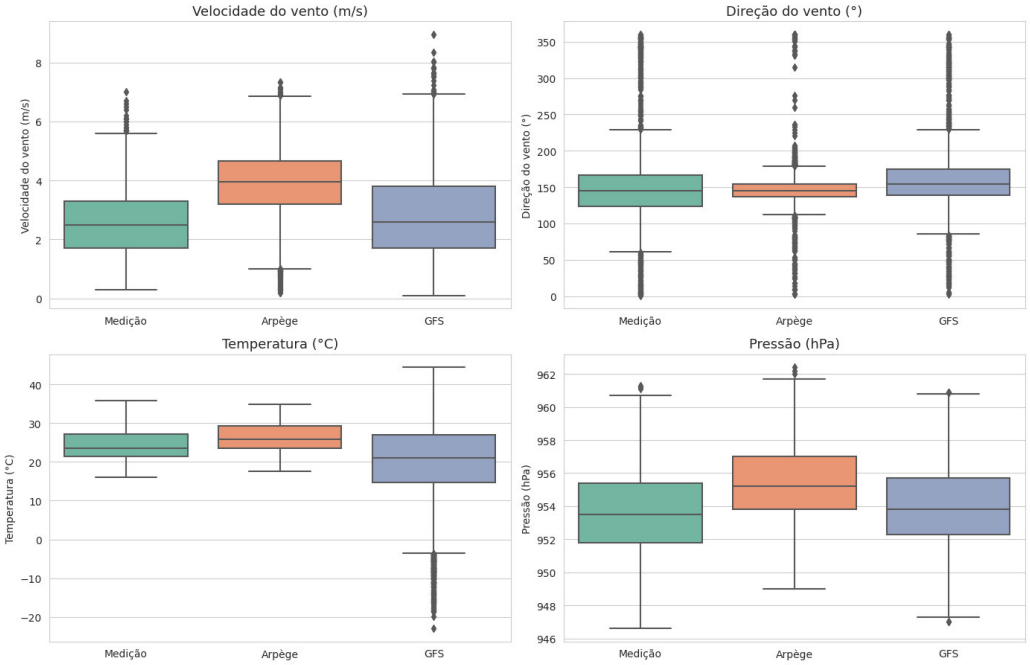


Histograma dos dados das variáveis meteorológicas analisadas para Rio Grande.

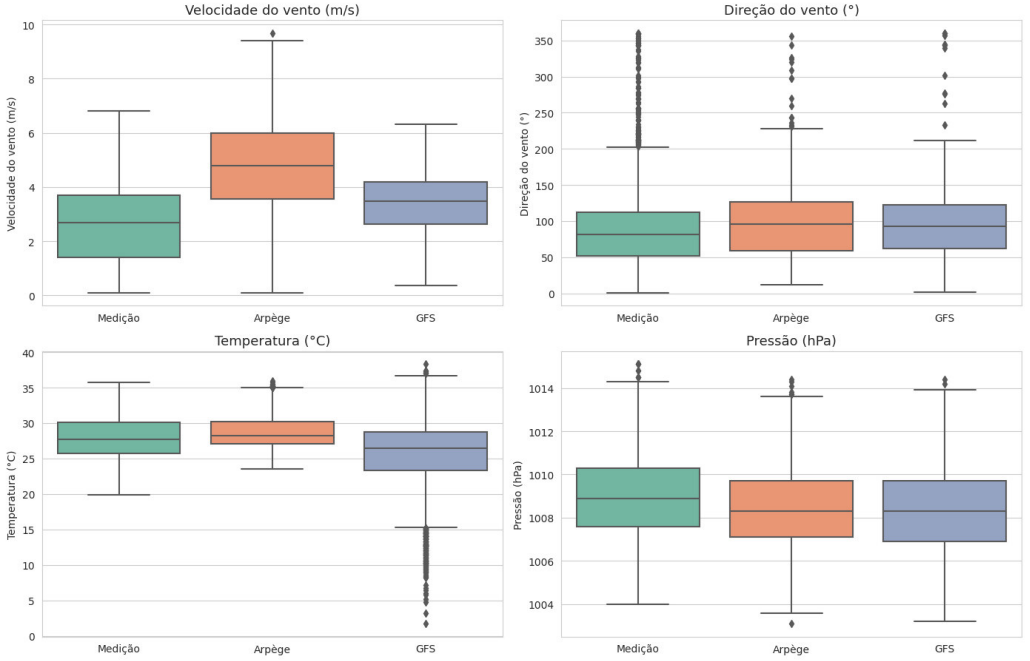


9.2 A.2 BOX PLOTS

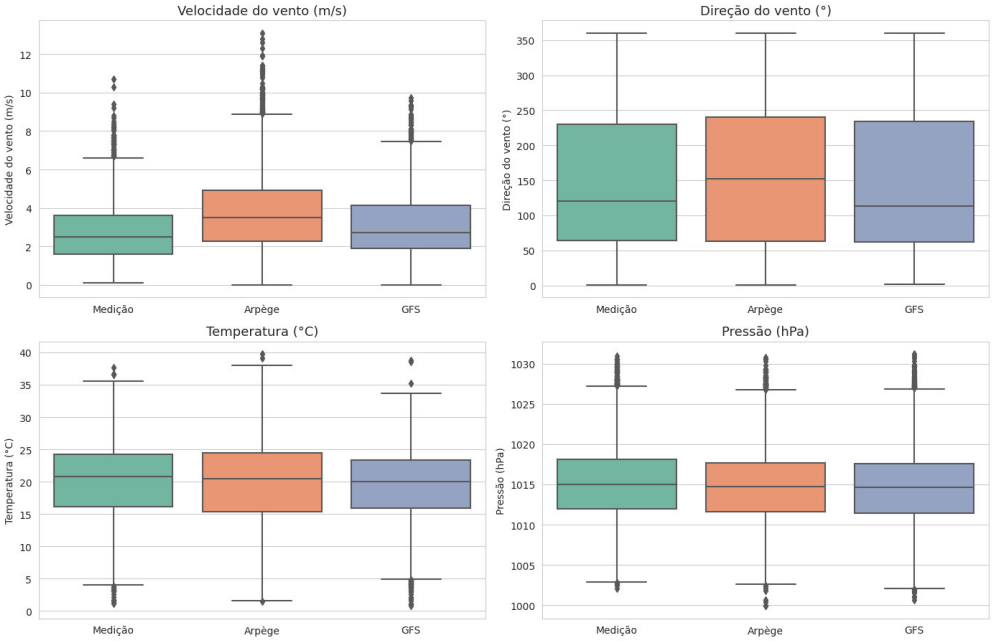
Box plot dos dados das variáveis meteorológicas analisadas para Senhor do Bonfim.



Box plot dos dados das variáveis meteorológicas analisadas para Mossoró.



Box plot dos dados das variáveis meteorológicas analisadas para Rio Grande.



9.3 A.3 GRÁFICOS COM AS ESTATÍSTICAS DESCRITIVAS

Gráfico de barras da média e desvio padrão dos dados medidos e previstos das variáveis meteorológicas para Senhor do Bonfim.

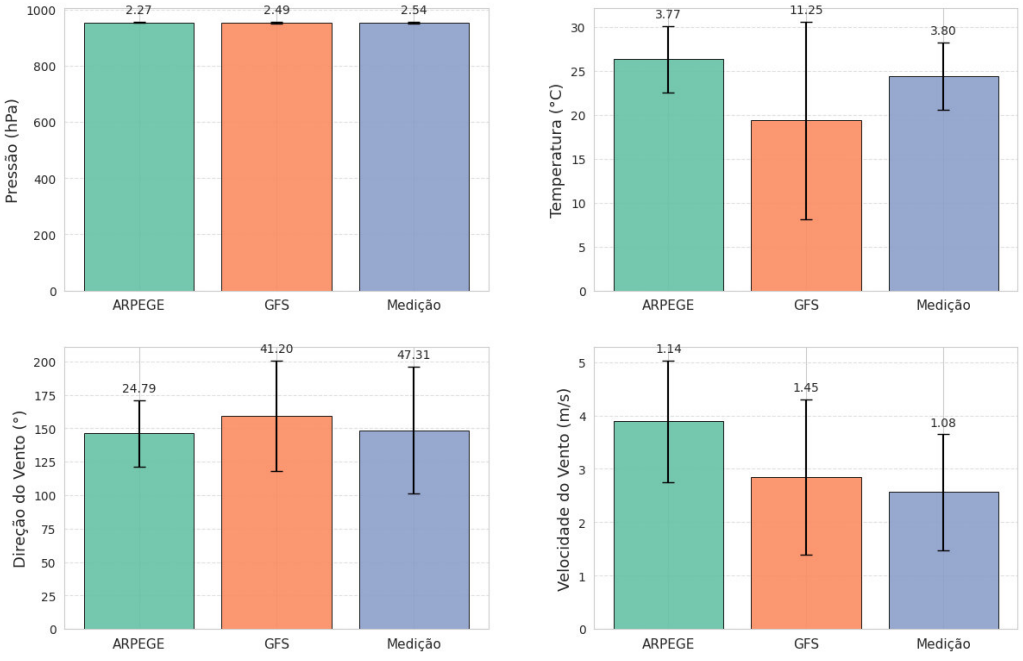


Gráfico de barras da média e desvio padrão dos dados medidos e previstos das variáveis meteorológicas para Conde.

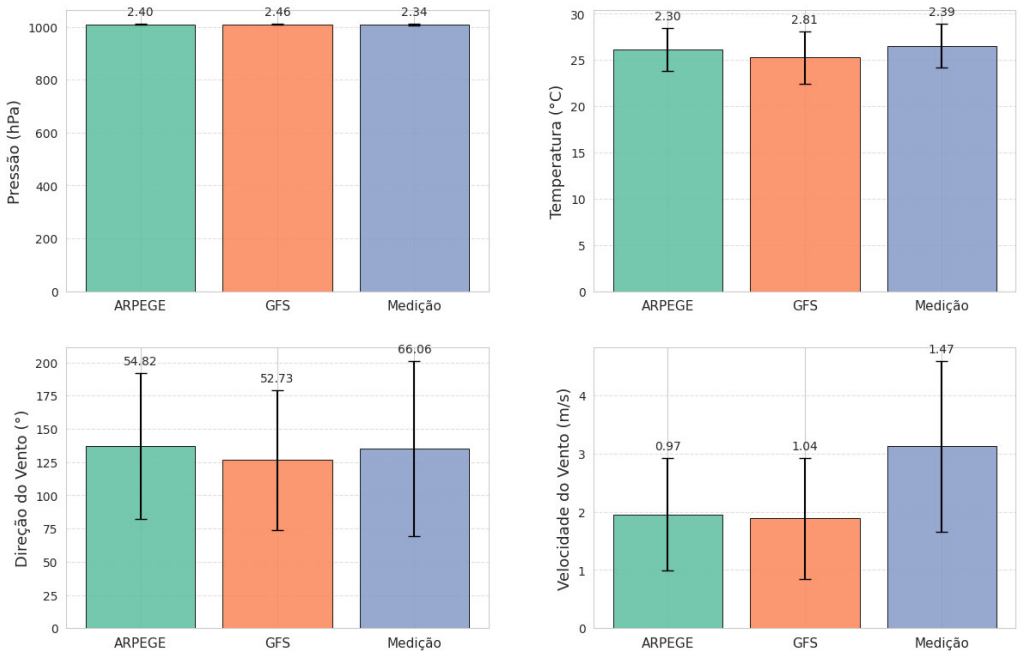
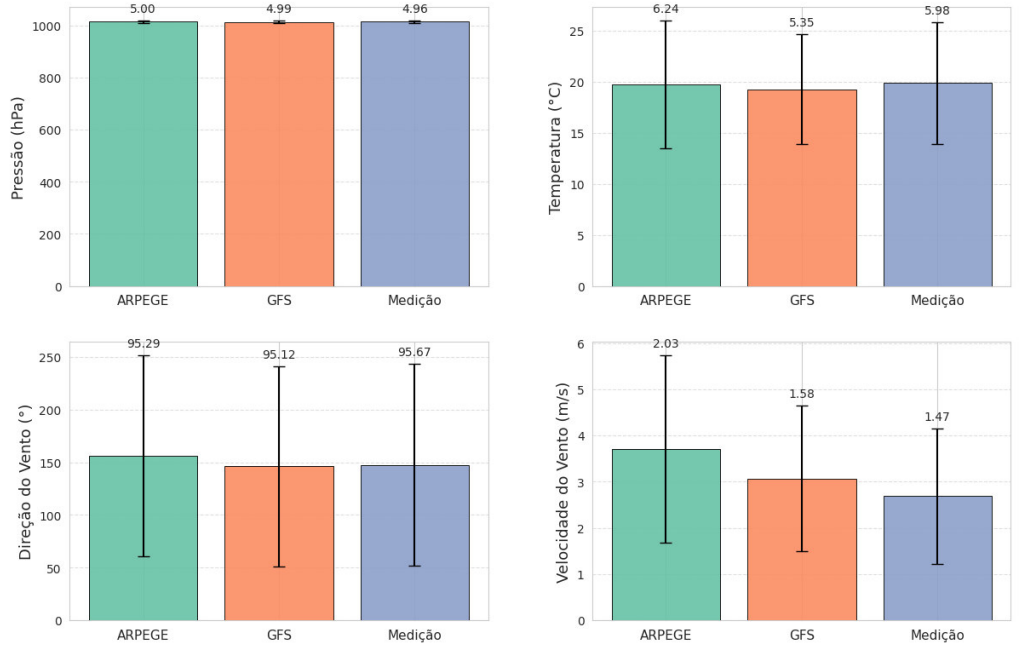


Gráfico de barras da média e desvio padrão dos dados medidos e previstos das variáveis meteorológicas para Rio Grande.



10 APÊNDICE B – AVALIAÇÃO DOS RESULTADOS DO MODELO

10.1 B.1 ANÁLISE DO VIÉS

Gráfico do viés pela velocidade do vento medida para *XGBoost ARPEGE* para cada localidade.

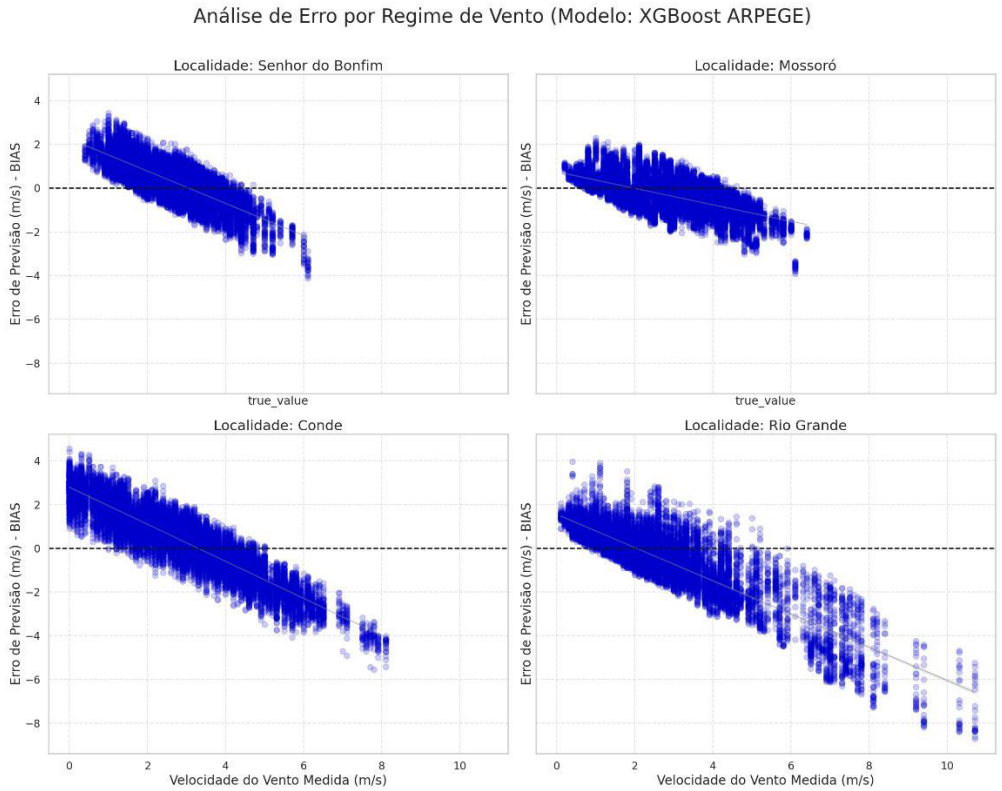
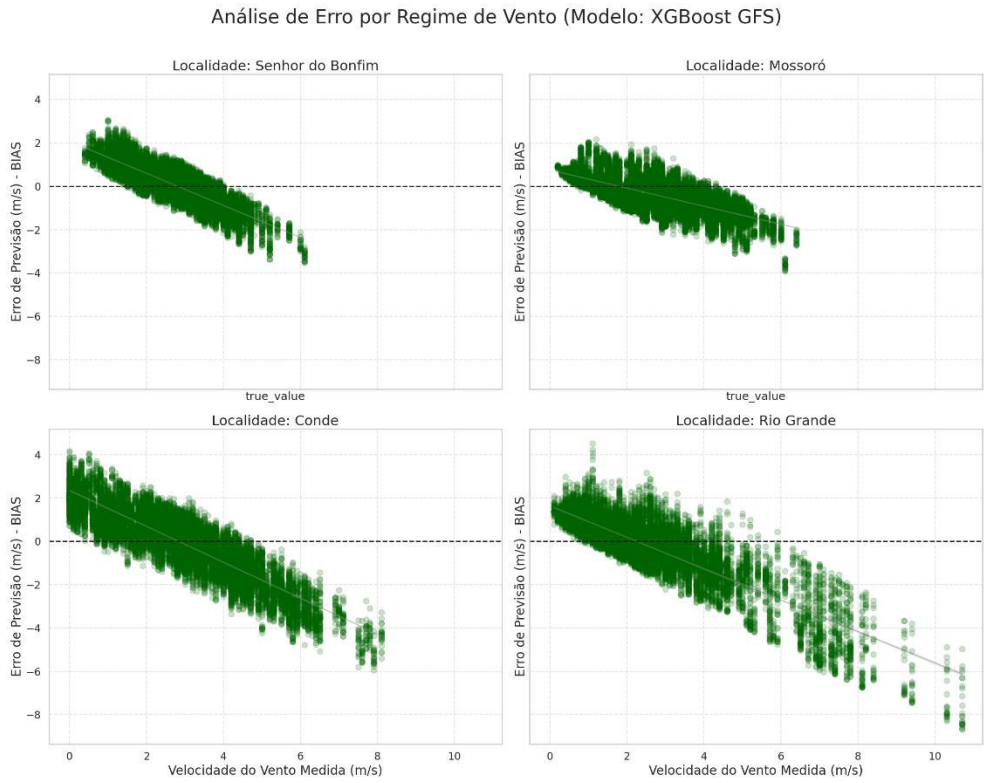


Gráfico do viés pela velocidade do vento medida para *XGBoost GFS* para cada localidade.



10.2 B.2 TABELA SÍNTESE DAS MÉTRICAS DOS MODELOS

Tabela de comparação das métricas entre os 5 métodos para a Senhor do Bonfim.

Método	***RMSE* (m/s)**	***R²***
XGBoost Arpege	1,04	0,02
XGBoost GFS	0,94	0,20
Persistência	1,17	-0,25
Arpege	1,82	-1,98
GFS	1,28	-0,47

Tabela de comparação das métricas entre os 5 métodos para a Rio Grande.

Método	***RMSE* (m/s)**	***R²***
XGBoost Arpege	1,52	0,20
XGBoost GFS	1,46	0,28
Persistência	1,78	-0,11
Arpege	1,75	-0,01
GFS	1,31	0,43