UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA MECÂNICA

DESENVOLVIMENTO DE UM MODELO DE MACHINE LEARNING PARA PREVISÃO DE POTÊNCIA EM AEROGERADORES

ERYCA FRANCYELE DE MOURA E SILVA

RECIFE
2025

ERYCA FRANCYELE DE MOURA E SILVA

DESENVOLVIMENTO DE UM MODELO DE MACHINE LEARNING PARA PREVISÃO DE POTÊNCIA EM AEROGERADORES

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Engenharia Mecânica da Universidade Federal de Pernambuco - UFPE, como requisito para obtenção do título de Bacharel em Engenharia Mecânica.

Área de Concentração: Engenharia Mecânica

Orientador: Prof. Francisco Fernando Roberto Pereira.

RECIFE

2025

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Silva, Eryca Francyele de Moura e.
Desenvolvimento de um modelo de machine learning para previsão de potência em aerogeradores / Eryca Francyele de Moura e Silva. - Recife, 2025.
52 p. : il., tab.

Orientador(a): Francisco Fernando Roberto Pereira
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia Mecânica - Bacharelado, 2025.
Inclui referências, anexos.

1. Energia Eólica. 2. Previsão de Potência. 3. Aprendizado de Máquina. 4. Regressão. 5. Sustentabilidade. 6. Eficiência Energética. I. Pereira, Francisco Fernando Roberto. (Orientação). II. Título.

620 CDD (22.ed.)

Universidade Federal de Pernambuco
Departamento de Engenharia Mecânica Centro de Tecnologia e Geociências- CTG/EEP

DEMEC

ATA DE SESSÃO DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO – TCC2

Ao 26° dia do mês de março do ano de dois mil e vinte e cinco, às 14:00 horas, reuniu-se a banca examinadora para a sessão pública de defesa do Trabalho de Conclusão de Curso em Engenharia Mecânica da Universidade Federal de Pernambuco, intitulado **Desenvolvimento de um modelo de machine learning para previsão de potência em aerogeradores**, elaborado pela aluna **Eryca Francyele de Moura e Silva**, matrícula 20190042008, sob a orientação do Prof. **Francisco Fernando Roberto Pereira**. A banca foi composta pelos avaliadores: Prof. **Francisco Fernando Roberto Pereira** (orientador), Profa. **Janaina Moreira de Meneses** (avaliadora), e Prof. **Marcelo José Alves de Santana** (avaliador). Após a exposição oral do trabalho, a candidata foi arguida pelos componentes da banca que em seguida reuniram-se e deliberaram pela sua aprovação, atribuindo-lhe a média 9,5, julgando-a apta(X) / inapta( ) à conclusão do curso de Engenharia Mecânica. Para constar, redigi a presente ata aprovada por todos os presentes, que vai assinada pelos membros da banca.

Orientador: Prof. Francisco Fernando Roberto Pereira Nota: 9,5

Assinatura gov.br Documento assinado digitalmente FRANCISCO FERNANDO ROBERTO PEREIRA Data: 26/03/2025 17:47:43-0300 Verifique em https://validar.iti.gov.br

Avaliadora Interna: Profa. Janaina Moreira de Meneses Nota: 9,5

Assinatura gov.br Documento assinado digitalmente JANAINA MOREIRA DE MENESES Data: 07/04/2025 00:20:20-0300 Verifique em https://validar.iti.gov.br

Avaliador Externo: Prof. Marcelo José Alves de Santana Nota: 9,5

Assinatura gov.br Documento assinado digitalmente MARCELO JOSE ALVES DE SANTANA Data: 08/04/2025 10:45:31-0300 Verifique em https://validar.iti.gov.br

Recife, 26 de março de 2025.

Prof. Marcus Costa de Araújo
Coordenador de Trabalho de Conclusão de curso - TCC
Curso de Graduação em Engenharia Mecânica – CTG/EEP-UFPE

RESUMO

A crescente demanda por energia renovável e a necessidade de otimizar a eficiência operacional de parques eólicos destacam a importância de modelos precisos de previsão de geração de energia. Este trabalho propõe o desenvolvimento de um modelo de aprendizado de máquina para prever a potência ativa gerada por aerogeradores, utilizando dados provenientes do sistema Aquisição e Supervisão de Dados (SCADA) de uma turbina eólica localizada na Turquia. Foram avaliados algoritmos como Regressão Linear, Árvores de Decisão, *Gradient Boosting*, *Extreme Gradient Boosting* (*XGBoost*) e *Random Forest*, com o objetivo de identificar o modelo mais adequado para capturar as complexidades dos dados e fornecer previsões confiáveis. A metodologia incluiu a preparação dos dados, a otimização de hiperparâmetros e a avaliação dos modelos com métricas como erro médio quadrático (RMSE), erro absoluto médio (MAE), erro absoluto percentual médio (MAPE) e coeficiente de determinação (R^2). O *XGBoost* destacou-se como o modelo com melhor desempenho, apresentando RMSE de 170,17 kW, R^2 de 0,9818 e MAPE de 12,88%, seguido pelo *Gradient Boosting*, *Random Forest*, Árvore de Regressão e a Regressão Linear. Os resultados demonstram que a previsão precisa da geração eólica é fundamental para mitigar incertezas, otimizar a eficiência operacional e contribuir para uma gestão energética mais sustentável e eficiente. Este trabalho alinha-se aos desafios globais de transição para fontes renováveis, oferecendo uma ferramenta valiosa para a gestão de parques eólicos.

Palavras-chave: energia eólica; previsão de potência; aprendizado de máquina; regressão; sustentabilidade; eficiência energética.

ABSTRACT

The increasing demand for renewable energy and the need to optimize the operational efficiency of wind farms highlight the importance of accurate energy generation forecasting models. This study proposes the development of a machine learning model to predict the active power generated by wind turbines, using data from the Supervisory Control and Data Acquisition (SCADA) system of a wind turbine located in Turkey. Algorithms such as Linear Regression, Decision Trees, Gradient Boosting, Extreme Gradient Boosting (XGBoost), and Random Forest were evaluated to identify the most suitable model for capturing data complexities and providing reliable predictions. The methodology included data preparation, hyperparameter optimization, and model evaluation using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). XGBoost stood out as the best-performing model, achieving an RMSE of 170.17 kW, an R^2 of 0.9818, and a MAPE of 12.88%, followed by Gradient Boosting, Random Forest, Decision Tree, and Linear Regression. The results demonstrate that accurate wind power forecasting is essential for mitigating uncertainties, optimizing operational efficiency, and contributing to more sustainable and efficient energy management. This study aligns with global challenges in transitioning to renewable energy sources, offering a valuable tool for wind farm management.

Keywords: wind energy; power forecasting; machine learning; Regression; Sustainability; energy efficiency.

SUMÁRIO

1 INTRODUÇÃO 8
1.1 OBJETIVOS 10
2 FUNDAMENTAÇÃO TEÓRICA 11
2.1 FUNDAMENTOS DA GERAÇÃO DE ENERGIA EÓLICA 11
2.2 ENERGIA EÓLICA NO BRASIL 12
2.3 AEROGERADORES 14
2.4 CONCEITOS DE INTELIGÊNCIA ARTIFICIAL 16
2.5 APLICAÇÕES EM CIÊNCIA DE DADOS 17
2.6 ALGORITMOS DE REGRESSÃO 17
2.6.1 Regressão linear 18
2.6.2 Árvore de decisão 20
2.6.3 Random forest 21
2.6.4 Gradient boosting regressor 23
2.6.5 XGBoost regressor 23
2.7 AVALIAÇÃO DE MODELOS 24
2.7.1 Viés e variância 24
2.7.2 Métricas de desempenho 25
3 METODOLOGIA 28
3.1 CONSTRUÇÃO DO MODELO 28
3.1.1 Descrição dos dados 29
3.1.2 Ferramentas computacionais 30
3.1.3 Análise exploratória dos dados 31
3.1.4 Pré-processamento 31
3.1.4.1 Divisão de dados 33
3.1.5 Modelagem e inferência 34
3.1.5.1 Validação cruzada 34
3.1.5.2 Hiperparâmetros 35
3.1.6 Pós-processamento 37
3.1.6.1 Interação com o usuário 37
4 RESULTADOS 38
4.1 ANÁLISE EXPLORATÓRIA DOS DADOS 38

4.2 PRÉ-PROCESSAMENTO 41
4.3 MODELAGEM E INFERÊNCIA 42
4.3.1 Hiperparâmetros e validação cruzada 42
4.4 PÓS- PROCESSAMENTO 43
4.4.1 Interação com o usuário 46
5 CONCLUSÃO.....47
REFERÊNCIAS BIBLIOGRÁFICAS 49

1 INTRODUÇÃO

A crescente preocupação global com a questão energética figura entre os grandes desafios da atualidade. O objetivo é atender a elevada demanda por energia, assegurar a integridade das fontes energéticas e, ao mesmo tempo, enfrentar a ameaça iminente da mudança climática. Contudo, o problema é agravado pelo fato de que a maior parte da eletricidade utilizada no planeta é proveniente de fontes não renováveis, seja de origem mineral, atômica ou térmica, que apresentam oferta limitada, desafios de segurança e sustentabilidade ecológica, custos elevados, e riscos de desastres (Nazir *et al*., 2020).

Neste contexto, muitos países têm se voltado para a implementação de políticas energéticas inovadoras, centradas no incremento, integração e avanço de tecnologias provenientes de fontes renováveis, como a energia eólica, solar, das marés, geotérmica, das águas, entre outras. Estas alternativas não apenas oferecem soluções mais seguras e limpas, mas também se revelam amplamente disponíveis, apresentando-se como opções viáveis em contraposição aos combustíveis convencionais. Em particular, a gratuidade de seu combustível (os ventos), os custos reduzidos de manutenção, o curto período necessário para instalação e operação, entre outros fatores, contribuem para solidificar a posição competitiva da energia eólica em meio às diversas fontes de energia disponíveis (Rodrigues, 2007).

Assim, devido à natureza intrínseca dessa forma de geração estar vinculada a parâmetros meteorológicos como direção e velocidade do vento, os sistemas de potência enfrentam variações imprevisíveis mais significativas nos fluxos de potência conforme a capacidade instalada desse tipo de geração intermitente cresce. Portanto, a integração substancial da geração eólica sem controles inteligentes pode acarretar instabilidades nos sistemas de potência e consequente desequilíbrio entre a produção e demanda elétrica, impondo penalidades que causam perdas de receita aos proprietários de parques eólicos. Nesse sentido, a previsão precisa da geração eólica desempenha um papel crucial, independentemente das escalas e horizontes temporais considerados, para mitigar as incertezas inerentes à produção de energia proveniente desses parques.

Neste trabalho, propôs-se projetar e desenvolver modelos de previsão baseados em aprendizado de máquina para geração de energia eólica, utilizando um

dataset proveniente do sistema SCADA (*Supervisory Control and Data Acquisition*) de uma turbina eólica localizada na Turquia. Nesse contexto, técnicas como Regressão Linear, Árvores de Decisão, *Gradient Boosting*, *XGBoost* e *Random Forest* têm se mostrado eficazes. Esses métodos destacam-se pela capacidade de analisar grandes quantidades de dados e fazer previsões com base em padrões e tendências, sendo amplamente utilizados para previsões de curto e médio prazo em energia eólica (Preethaa *et al*., 2023). Diante disso, foram avaliados esses modelos a fim de identificar aquele que melhor captura as complexidades dos dados e fornece previsões precisas.

A metodologia adotada incluiu a preparação dos dados, a seleção e otimização de hiperparâmetros, e a avaliação dos modelos por meio de métricas como RMSE (Raiz do Erro Quadrático Médio), MAE (Erro Absoluto Médio), MAPE (Erro Percentual Absoluto Médio) e R^2 (Coeficiente de Determinação). Ademais, a técnica de validação cruzada foi utilizada para garantir a robustez dos modelos, e o *Grid Search* foi aplicado para otimizar os hiperparâmetros de cada algoritmo.

Em suma, os resultados obtidos demonstraram que o *XGBoost* foi o modelo com melhor desempenho, apresentando o menor RMSE (170,17 kW), MAPE (12,88%) e o maior R^2 (0,9818), seguido pelo *Gradient Boosting Regressor* (RMSE de 171.20 kW; R^2 de 0.9816; MAPE de 13.16%) e pelo *Random Forest* (RMSE de 172.41 kW; R^2 de 0.9813; MAPE de 12.58%). O *XGBoost* também apresentou o menor MAE (90.03 kW), indicando que suas previsões estão, em média, mais próximas dos valores reais. Em contraste, a Árvore de Regressão e a Regressão Linear tiveram desempenhos inferiores, com maiores erros e menor capacidade de explicação da variância dos dados.

Com os resultados, os algoritmos demonstram a capacidade de prever com precisão a potência ativa dos aerogeradores, reduzindo incertezas e otimizando a gestão dos parques eólicos. Assim, a previsão de curto prazo reduz perdas mecânicas e otimiza a eficiência operacional, bem como fornece dados essenciais para sistemas de controle, contribuindo para um equilíbrio mais preciso entre produção e demanda de energia. Dessa forma, a implementação de modelos de previsão como o proposto neste trabalho fortalece a gestão energética sustentável e eficiente, alinhando-se aos desafios globais de transição para fontes renováveis.

1.1 OBJETIVOS

O objetivo geral deste trabalho é desenvolver um modelo de previsão baseado em conjuntos de dados de séries temporais para estimar energia eólica, utilizando algoritmos de aprendizado de máquina. Como objetivos específicos podem-se citar:

a) Realizar uma análise exploratória dos dados;
b) Limpar os dados utilizando técnicas estatísticas;
c) Pré-processar os dados;
d) Desenvolver um modelo de previsão baseado em aprendizado de máquina;
e) Avaliar o desempenho e examinar as capacidades de previsão do modelo utilizando métricas estatísticas.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão abordados temas relacionados aos fundamentos da geração da energia eólica, os avanços dessa fonte no Brasil, o princípio de funcionamento dos aerogeradores, além de introduzir conceitos de inteligência artificial e ciência de dados, com foco nos principais modelos de regressão aplicados a esse contexto.

2.1 FUNDAMENTOS DA GERAÇÃO DE ENERGIA EÓLICA

A energia eólica, resultado da conversão da energia cinética do vento em eletricidade ou trabalho mecânico, remonta a uma longa história de aplicação desde os primórdios da civilização, sendo utilizada em atividades agrícolas, navegação e bombeamento de água. No século XX, os dinamarqueses desempenharam um papel crucial nos primeiros desenvolvimentos para a geração de eletricidade por meio do vento. Por volta de 1930, empresas americanas já produziam e comercializavam geradores eólicos primitivos, principalmente destinados a fazendeiros, capazes de gerar até 1000 watts de corrente contínua em condições ventosas (Rodigues, 2007).

A competitividade de preços das energias renováveis, impulsionada pelo mercado energético e leilões de energia, emergiu como um fator crucial no cenário contemporâneo. Países como a China, Estados Unidos, Reino Unido, Alemanha, Canadá, Itália, França, Índia e Brasil têm desempenhado papéis significativos nesse cenário (Nazir, 2019). Essa competitividade resultou em uma rápida expansão da indústria eólica globalmente nos últimos anos.

Em uma perspectiva global, o relatório mais recente do Conselho Global de Energia Eólica (2024), o *Global Wind Report* (GWEC), revela um marco significativo para a indústria eólica mundial. De acordo com os dados apresentados, a capacidade instalada de energia eólica global atingiu um recorde de 117 GW em 2023, marcando este como o ano mais promissor para a energia eólica até o momento, representando um aumento de 50% em relação a 2022 (ABEEólica, 2024). O GWEC observa que, apesar de um cenário político e macroeconômico instável, a indústria eólica está ingressando em uma nova fase de expansão acelerada, impulsionada por uma crescente ambição política. Esta nova dinâmica é evidenciada pela adoção de uma meta histórica na COP28, 28ª conferência sobre mudanças climáticas da Organização

das Nações Unidas (ONU), que visa triplicar a participação das energias renováveis até 2030.

Nesse cenário, a Tabela 1 oferece uma visão em ranking da indústria eólica mundial, onde o Brasil mantém sua posição como o sexto maior país em capacidade instalada de energia eólica *onshore*.

Tabela 1. Ranking GWEC 2023: Capacidade total onshore instalada de energia eólica em 2023.

Posição	**País**	**Capacidade total instalada *onshore* (GW)**
1	China	403,33
2	USA	150,43
3	Alemanha	61,14
4	Índia	44,74
5	Espanha	30,56
6	Brasil	30,45
7	França	22,00
8	Canadá	16,99
9	Suécia	16,25
10	Reino Unido	14,87

Fonte: Adaptado de GWEC (2024).

De acordo com a ABEEólica, esse avanço representa um marco significativo desde 2012, quando ocupava o 15º lugar. Dessa maneira, evidencia-se a capacidade de resiliência e dinamismo do país, que aproveitou oportunidades de crescimento através de leilões privados e acordos bilaterais de aquisição de energia, mesmo diante de um contexto econômico desafiador nos últimos 10 anos. Adicionalmente, o relatório ressalta os recursos renováveis disponíveis no Brasil, especialmente os ventos de alta qualidade tanto em áreas terrestres quanto marítimas, considerados únicos globalmente. Essa particularidade abre caminho para iniciativas como o hidrogênio verde, contribuindo para a diversificação e a sustentabilidade da matriz energética brasileira.

2.2 ENERGIA EÓLICA NO BRASIL

O avanço substancial na matriz elétrica brasileira em 2023, somando 8.412,1 megawatts (MW) até 30 de novembro, ultrapassou em 11 meses o crescimento registrado em 2022, que já era a segunda maior marca desde 1997, ano de fundação da Agência Nacional de Energia Elétrica (ANEEL). Desse notável

progresso, as usinas eólicas e solares centralizadas foram responsáveis por expressivos 90,4%, totalizando a entrada em operação de 7.608,4 MW em 2023, segundo dados da Agência Nacional de Energia Elétrica (ANEEL, 2023).

Esses avanços representam uma significativa contribuição para a diversificação e expansão da matriz elétrica brasileira, fortalecendo o compromisso nacional com fontes de energia mais limpas e sustentáveis. Ainda, o Sistema de Informações de Geração da ANEEL (SIGA) confirma que o Brasil atingiu uma capacidade fiscalizada de 196.780,4 MW. Dentro desse panorama, 83,9% das usinas em operação, conforme registros do SIGA, são consideradas fontes renováveis.

Esses indicadores positivos na matriz elétrica coexistem com o histórico pioneiro da energia eólica no Brasil, exemplificado pela elaboração do Atlas Eólico do Nordeste em 1998 e do Atlas do Potencial Eólico Brasileiro pelo Centro de Pesquisas Elétricas (CEPEL), vinculado à ELETROBRÁS. O país, detentor de expressivo potencial eólico, respondeu a essa oportunidade com a criação do Programa de Incentivo às Fontes Alternativas de Energia (PROINFA) em 2002, evidenciando o compromisso governamental com a transição para fontes renováveis (Do Amarante, 2001).

No que tange o cenário atual, os primeiros três meses de 2024 reforçam a tendência de crescimento observada em 2023. A ANEEL reporta que o país concluiu o primeiro trimestre com 105 novas usinas e um incremento de 2,6 gigawatts (GW) na potência instalada. Este aumento foi impulsionado especialmente pelo mês de março, que registrou a entrada em operação de 41 usinas e um acréscimo de 906 megawatts (MW), decorrente em grande parte das 13 centrais solares fotovoltaicas (544,22 MW) e das 25 eólicas (316,30 MW) que passaram a operar no período. Além disso, é importante ressaltar que o Brasil ultrapassou a marca de 200 gigawatts de potência ainda em março. Em 10 de abril, a capacidade fiscalizada alcançou 201.108,7 MW, com 84,41% das usinas em operação sendo consideradas renováveis, conforme dados do SIGA.

Ademais, é evidente que a revolução da energia eólica no Brasil está em curso há alguns anos, consolidando-se como um dos setores de geração de energia mais robustos do país. Além do crescimento contínuo da energia eólica *onshore*, há também a expectativa de desenvolvimento da energia eólica *offshore* brasileira. De acordo com Elbia Gannoum (2023), Presidente da ABEEólica, Associação brasileira

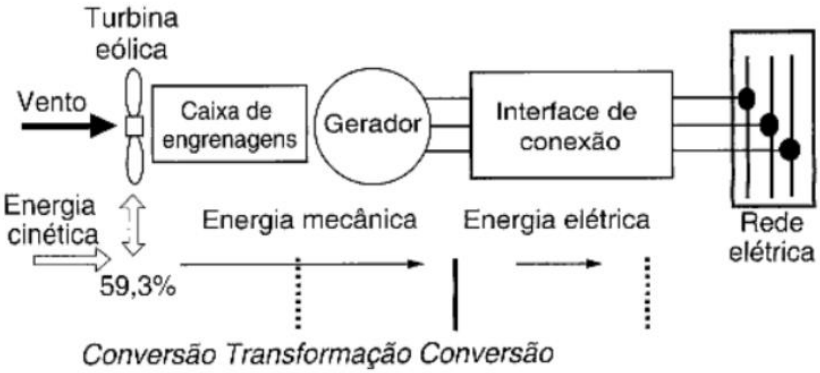
de energia eólica e vice-presidente do GWEC, o IBAMA - Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis, já recebeu propostas de projetos para mais de 170 GW de energia eólica *offshore*. Esse montante é praticamente equivalente a toda a matriz elétrica brasileira, evidenciando o amplo interesse dos investidores e o enorme potencial para a energia eólica offshore nas águas brasileiras.

Sendo assim, a escolha estratégica da geração eólica como principal fonte de crescimento renovável é respaldada pela sua vantajosa relação custo/benefício em comparação com outras fontes, como a hidrelétrica. Essa decisão é fortalecida por avanços tecnológicos, significativos investimentos proporcionando ganhos de escala na fabricação de equipamentos, uma crescente consciência ambiental e políticas governamentais eficazes, como as estabelecidas pelo PROINFA, que têm desempenhado um papel crucial na promoção da energia renovável e no desenvolvimento tecnológico.

2.3 AEROGERADORES

As turbinas eólicas desempenham um papel crucial na conversão da energia do vento em eletricidade, seguindo um processo dividido em duas etapas distintas. Inicialmente, na turbina, ocorre a extração de uma porção da energia cinética disponível, transformando-a em energia mecânica. Posteriormente, essa energia mecânica é direcionada para o gerador, onde é convertida em energia elétrica e, por conseguinte, transmitida para a rede elétrica da concessionária, conforme ilustrado na Figura 1.

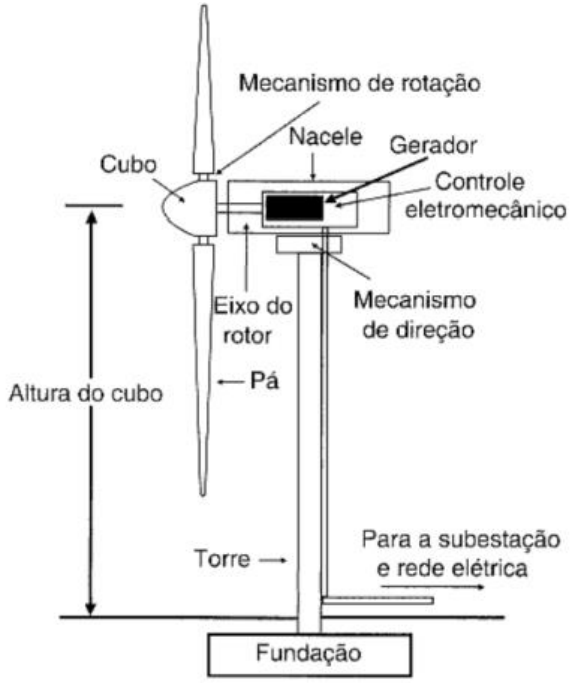
Figura 1. Princípio da conversão de energia cinética do vento em elétrica.



Fonte: Pinto (2014).

Em termos de constituição, é importante destacar que o aerogerador é o sistema completo responsável por essa conversão de energia eólica em eletricidade, possuindo diversas partes igualmente vitais para o seu funcionamento, conforme ilustrado na Figura 2. A torre, estrutura de sustentação, proporciona a elevação necessária para otimizar a captação do vento. A nacele, localizada no topo da torre, abriga o gerador e a caixa de engrenagens, sendo responsável por realizar a conversão da energia mecânica gerada pela turbina em eletricidade.

Figura 2. Principais partes de um aerogerador.



Fonte: Pinto (2014).

A caixa de engrenagens, por sua vez, desempenha um papel crucial ao ajustar a sintonia entre a baixa velocidade da turbina e a alta velocidade do gerador.

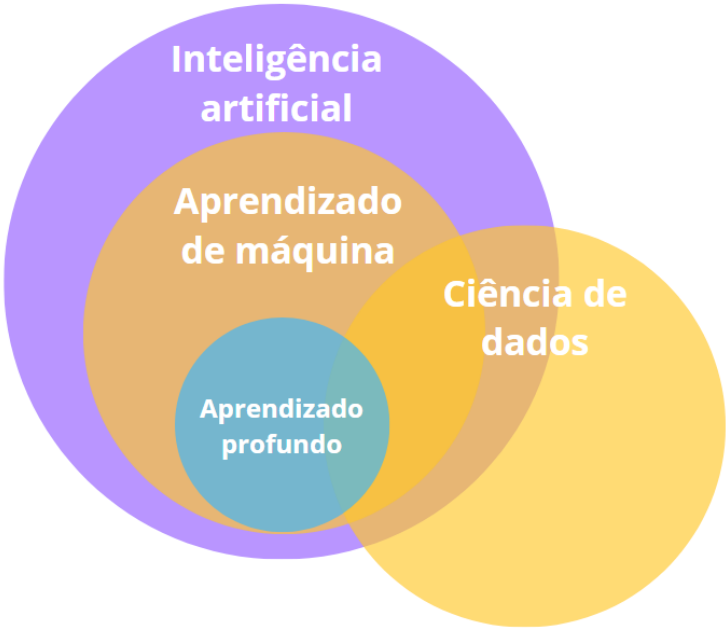
O cubo, construído em aço ou liga metálica de alta resistência, é a estrutura na qual as pás são fixadas, proporcionando a estabilidade necessária para o funcionamento adequado do aerogerador. As pás, movimentadas pelo vento, são fundamentais para a captação eficiente da energia cinética.

2.4 CONCEITOS DE INTELIGÊNCIA ARTIFICIAL

A Inteligência Artificial (IA) é o desenvolvimento de sistemas computacionais capazes de realizar tarefas que tradicionalmente exigiriam inteligência humana, como aprendizado, resolução de problemas e tomada de decisão. Esses sistemas utilizam técnicas avançadas, incluindo aprendizado de máquina, aprendizado profundo, processamento de linguagem natural e visão computacional, para analisar grandes volumes de dados e tomar decisões precisas. Como resultado, os exemplos de aplicação da IA são vastos, abrangendo desde assistentes virtuais pessoais e carros autônomos até sistemas de detecção de fraudes, reconhecimento de imagem e fala, e até mesmo diagnóstico e tratamento médico (Malik *et. al*., 2024).

A Figura 3 destaca a relação entre IA, aprendizado de máquina, aprendizado profundo e Ciência de Dados.

Figura 3. Relação entre Inteligência artificial, aprendizado de máquina, ciência de dados e aprendizado profundo.



Fonte: autoria própria.

Nesse contexto, o aprendizado de máquina, um subconjunto da IA, emprega algoritmos baseados em matemática e estatística para extrair padrões de dados

históricos, podendo ser classificado em supervisionado e não supervisionado (Escovedo e Koshiyama, 2023). No aprendizado supervisionado, o algoritmo é treinado com pares de entradas e saídas conhecidas, permitindo prever saídas para novas entradas. Já no não supervisionado, o algoritmo identifica padrões e agrupamentos nos dados sem respostas pré-definidas (VanderPlas, 2023). A IA também incorpora a Ciência de Dados, que envolve métodos e ferramentas para manipular e analisar grandes volumes de dados, visando extrair conhecimento útil e relevante (Bruce e Bruce, 2019).

2.5 APLICAÇÕES EM CIÊNCIA DE DADOS

A Ciência de Dados, ou *Data Science*, é essencial para a tomada de decisões, abordando problemas como Classificação, Regressão, Agrupamento e Associação. Logo, problemas de Classificação e Regressão geralmente empregam técnicas de aprendizado supervisionado, enquanto Agrupamento e Associação utilizam abordagens de aprendizado não supervisionado.

A Classificação prevê categorias, como identificar e-mails como "spam", e a Regressão estima valores contínuos, como a temperatura. Ademais, o Agrupamento segmenta dados em *clusters* por similaridades, e a Associação identifica relações entre variáveis, útil em análises de mercado. Com isso, a Ciência de Dados inclui análises retrospectivas (conhecida como *Business Intelligence*, BI) e preditivas, com técnicas de Aprendizado de Máquina que antecipam tendências futuras, apoiando decisões estratégicas (Escovedo e Koshiyama, 2023; Sharda, Delen e Turban, 2020).

2.6 ALGORITMOS DE REGRESSÃO

Marugán (2018) destaca a crescente relevância da inteligência artificial (IA) na previsão de energia eólica, área que demanda métodos precisos devido à complexidade e não linearidade das variáveis envolvidas, como condições climáticas. Para isso, são utilizadas diversas abordagens, incluindo modelos estatísticos, técnicas de regressão, séries temporais e algoritmos de aprendizado de máquina, visando previsões mais confiáveis e eficazes.

Como já abordado, o problema de regressão é um tipo de aprendizado supervisionado que busca prever valores numéricos contínuos com base em variáveis explicativas (independentes). Assim, diferente de problemas de classificação, que

atribuem categorias, a regressão estima um valor contínuo para a variável resposta (dependente), também chamada de *target* (*Y*), utilizando um vetor de características (*features*) (*X*), conforme esquematizado na Figura 4. Dessa forma, o objetivo do modelo é minimizar o erro entre os valores preditos e os reais (resíduos), avaliado por métricas específicas, detalhadas no tópico 2.7.2 deste trabalho. A eficácia do modelo depende de sua capacidade de reduzir esse erro, processo que ocorre durante o treinamento, onde o algoritmo ajusta os parâmetros para aproximar as previsões dos valores observados (Bruce e Bruce, 2019).

Figura 4. Variáveis para um problema de regressão.

Padrão	Explicativas			Resposta
$\mathbf{x}_1$	x_{11}	...	x_{1J}	y_1
$\mathbf{x}_2$	x_{21}	...	x_{2J}	y_2
$\mathbf{x}_3$	x_{31}	...	x_{3J}	y_3
...	...	...	...	...
$\mathbf{x}_i$	x_{i1}	...	x_{iJ}	y_i
...	...	...	...	...
$\mathbf{x}_n$	x_{n1}	...	x_{nJ}	y_n

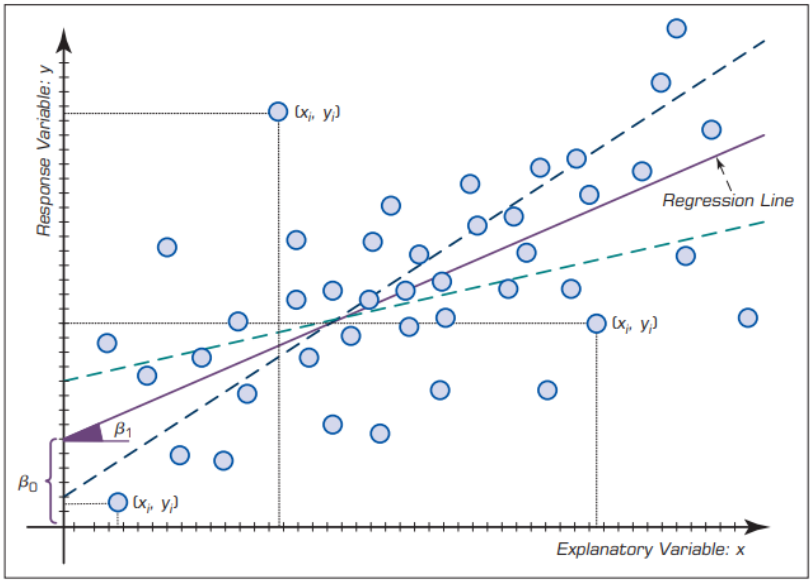
Fonte: Escovedo e Koshiyama (2023).

Dentre os algoritmos de regressão mais utilizados para previsão, pode-se citar: Regressão linear, Árvores de decisão/regressão, *Random Forest*, *Gradient Boosting* e *XGBoost*.

2.6.1 Regressão linear

A regressão linear é uma técnica estatística utilizada para modelar a relação entre variáveis explicativas e de resposta, expressando essa relação por meio de uma equação linear. Dessa forma, seu objetivo é estimar os coeficientes dessa equação para prever o comportamento de uma variável dependente (quantitativa) com base em uma (regressão linear simples) ou mais (regressão linear múltipla) variáveis independentes (Moreira *et al*., 2020). A Figura 5 ilustra um gráfico de dispersão, uma maneira simples de compreender a relação entre as variáveis em questão.

Figura 5. Gráfico de dispersão e linha de regressão linear.



Fonte: Sharda, Delen e Turban (2020).

A regressão linear múltipla, que incorpora múltiplos regressores, é expressa pela Equação 1.

$$y = \beta_0 + \beta_1 x + \cdots + \beta_\rho X_\rho + \varepsilon \quad (1)$$

Nesta equação, y representa a variável dependente (ou variável alvo, *target*); X_ρ denota as variáveis independentes ou explicativas (ou variável preditora, *features*); β_ρ são os parâmetros estimados da regressão; e ε refere-se ao erro aleatório do modelo, representando as influências não controladas, ou seja, fatores que afetam a variável dependente além daqueles incluídos no modelo.

Por outro lado, a regressão linear simples, que modela a relação entre uma única variável explicativa e a variável de resposta, é expressa pela Equação 2.

$$y = \beta_0 + \beta_1 x \quad (2)$$

Nesta equação, β_0 é denominado intercepto, representando o valor esperado de y quando $x = 0$, e β_1 é a inclinação, indicando a mudança esperada em y para um aumento unitário em x.

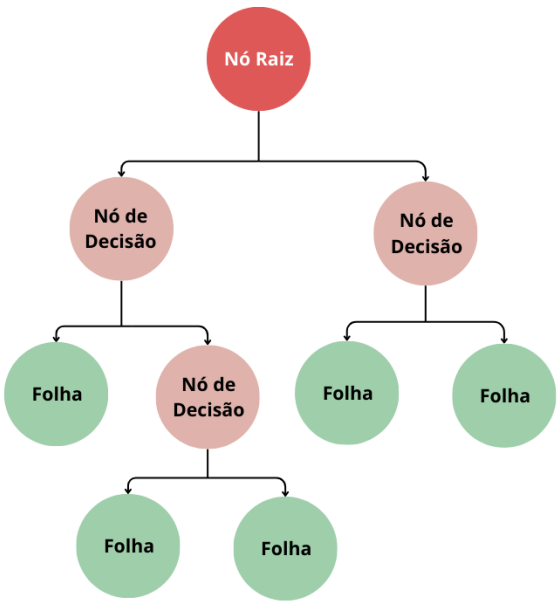
De acordo com Sharda, Delen e Turban (2020), o método mais utilizado para identificar a linha de regressão é o dos mínimos quadrados ordinários (*Ordinary Least Squares*, OLS), que minimiza a soma dos quadrados dos resíduos (diferenças entre os valores observados e previstos). Esse método garante que a linha de regressão se ajuste aos dados da melhor forma possível, revelando a direção e a força da relação entre as variáveis.

Ademais, a regressão linear assume que a relação entre as variáveis é linear e aditiva, e sua confiabilidade depende da validação de suposições como normalidade dos resíduos, homocedasticidade e independência dos erros. Caso essas suposições sejam violadas, técnicas de transformação de dados ou modelos alternativos podem ser necessários (Moreira et al., 2020).

2.6.2 Árvore de decisão

O algoritmo de Árvore de Decisão, ou *Decision Tree*, é uma ferramenta eficaz para problemas de classificação e regressão, e sua popularidade cresceu recentemente graças ao aumento do poder computacional, essencial para processar grandes volumes de dados. Assim, o método funciona através de sucessivas divisões nos dados, em que cada divisão define os próximos nós da árvore, caracterizando uma estrutura hierárquica, intuitiva e interpretável. Além disso, possui poucos parâmetros para configuração, o que torna seu uso simples e rápido, e tem a capacidade de aprender com o ruído dos dados, evitando o ajuste excessivo dos parâmetros (Rathan, Sai, and Manikanta, 2019).

Figura 6. Funcionamento da Árvore de decisão.



Fonte: Autoria própria.

Na Figura 6 é apresentada a estrutura básica de uma árvore de decisão, onde cada nó interno representa uma pergunta ou decisão sobre um atributo dos dados, determinando como os dados serão divididos nos nós filhos. Dessa forma, a partir da raiz, segue-se o caminho da árvore testando os valores dos atributos até chegar a um

nó folha, onde encontra-se a classificação ou valor predito.

As Árvores de Regressão foram introduzidas nos anos 1980 como parte do algoritmo CART (*Classification and Regression Tree*) e são projetadas para prever valores contínuos. Assim, sua construção segue o princípio de divisão e conquista, particionando os dados para maximizar a homogeneidade em cada divisão. Enquanto nas Árvores de Classificação, a homogeneidade é medida pela entropia, nas árvores de regressão, a homogeneidade é avaliada por métricas estatísticas como variância, desvio padrão ou desvio absoluto da média. A redução do desvio padrão (SDR – *Standard Deviation Reduction*) é um critério comum para selecionar estas divisões, expresso na Equação 3 (Escovedo e Koshiyama, 2023).

$$SDR = sd(T) - \sum_{i} \frac{|T_i|}{|T|} \times sd(T_i) \quad (3)$$

Onde $sd(T)$ é a função que se refere ao desvio padrão do conjunto *T*, em que $T_1, T_2 \ldots, T_n$ são os conjuntos de valores resultantes de uma divisão em uma característica. Ademais, $|T|$ refere-se ao número de observações no conjunto *T*.

Logo, o particionamento binário é utilizado para definir a estrutura da árvore, buscando minimizar a soma dos desvios quadrados da média em cada divisão. Essa abordagem permite lidar com relações lineares ou não lineares entre os atributos e o valor a ser predito. Contudo, para evitar problemas como o *overfitting*, que compromete a capacidade de generalização do modelo, é essencial empregar um conjunto de validação durante o processo de ajuste, garantindo maior robustez e precisão na predição (Delgado Filho, 2020).

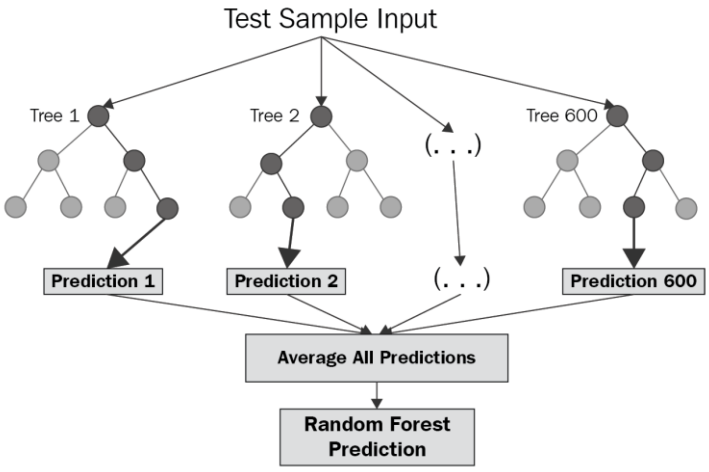
2.6.3 Random forest

O conceito de *ensemble learning* envolve a combinação de múltiplos modelos para gerar um desempenho superior ao dos modelos individuais. Com isso, entre as principais abordagens de *ensemble*, destacam-se o *Bagging* e o *Boosting*. O *Bagging*, ou *Bootstrap Aggregating*, visa reduzir a variância e melhorar a robustez dos modelos por meio da construção de múltiplos classificadores ou regressores a partir de amostras aleatórias dos dados. Por outro lado, o *Boosting* busca reduzir o viés, ajustando iterativamente os modelos para focar nos erros cometidos por predições anteriores, com o objetivo de criar um modelo final mais preciso (Silva, 2019; Sharda, Delen e Turban, 2020).

O modelo de *Random Forest* (RF), ou Floresta Aleatória, foi introduzido por Breiman (2001) como uma extensão do método de *bagging*, com o objetivo de melhorar a precisão e a robustez das previsões. Assim como no *bagging*, o algoritmo de RF começa com a construção de múltiplas Árvores de Decisão, cada uma treinada em uma amostra *bootstrap* dos dados de treinamento. No entanto, o RF incorpora uma camada adicional de aleatoriedade: em cada divisão da árvore, apenas um subconjunto aleatório de variáveis preditoras é considerado para a divisão, em vez de todas as variáveis disponíveis (Sharda, Delen e Turban, 2020).

A Figura 7 ilustra a estrutura do RF. Cada árvore é treinada em uma amostra com reposição dos dados de entrada, e a previsão final do modelo é obtida pela média das previsões das árvores individuais no caso de regressão, ou por votação no caso de classificação.

Figura 7. Estrutura *Random Forest*.



Fonte: Delgado Filho (2020).

A prática comum sugere que o número de variáveis consideradas em cada divisão seja aproximadamente igual à raiz quadrada do número total de preditores. Por exemplo, se houver 100 variáveis preditoras, cerca de 10 seriam selecionadas aleatoriamente para cada divisão. Isso garante que as árvores sejam diversas, pois é improvável que as mesmas variáveis sejam usadas repetidamente em divisões consecutivas. Nesse contexto, a diversidade é crucial para melhorar a generalização do modelo, já que árvores altamente correlacionadas (como no *bagging*) tendem a produzir previsões semelhantes. Isso, devido aos preditores dominantes serem usados repetidamente nas primeiras divisões, limitando a redução da variância (Sharda, Delen e Turban, 2020; James *et al*., 2023).

2.6.4 Gradient boosting regressor

O *Gradient Boosting Regressor* (GBR), ou aumento de gradiente, uma das variantes mais populares do *boosting*, foi desenvolvido para aprimorar a técnica original, tornando-a mais eficaz na minimização da função de custo. Esse método é compatível com qualquer função de custo diferenciável, o que possibilita sua aplicação em problemas de regressão e classificação. Além disso, o GBR apresenta vantagens notáveis, como o tratamento eficiente de dados heterogêneos e maior robustez contra *outliers* (Silva, 2019). A formulação matemática pode ser compreendida abaixo.

$$F(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{4}$$

Em que $h_m(x)$ representa as funções básicas do modelo $F(x)$, comumente denominadas “aprendizes fracos” na técnica de *boosting*, baseadas em pseudoresíduos. Nesse contexto, árvores de tamanho fixo desempenham esse papel, sendo combinadas iterativamente para aprimorar o modelo em uma abordagem aditiva, com γ_m sendo o peso (ou fator de aprendizado) atribuído para cada árvore (Equação 5). A cada etapa, o ajuste foca na minimização da função de custo, corrigindo os erros das previsões anteriores (Silva, 2019).

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{5}$$

Assim, em cada estágio, a árvore de decisão $h_m(x)$ é escolhida para minimizar a função de custo *L* dado o modelo atual F_{m-1} e seu ajustamento $F_{m-1}(x_i)$, tais como:

$$F_m(x) = F_{m-1}(x) + argmin_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + h(x)) \tag{6}$$

As funções de perda (ou custo) podem ser escolhidas de forma flexível, dependendo do tipo de variável de resposta. No caso de uma variável de resposta contínua, a função de perda mais comumente utilizada é o erro quadrático médio (MSE) (Bentéjac *et. al*., 2020).

2.6.5 XGBoost regressor

O *Extreme Gradient Boosting* (*XGBoost*) é um método que busca melhorar o desempenho do *Gradient Boosting*, oferecendo um pacote altamente escalável, com melhor velocidade de treinamento e capacidade de generalização. Dessa forma, assim como o GBR, o *XGBoost* constrói uma expansão aditiva da função objetivo

(Equação 7) ao minimizar uma função de perda. Além disso, o algoritmo utiliza técnicas avançadas de regularização para evitar o *overfitting* do modelo, o que o torna mais geral e robusto em relação a dados novos. Dessa forma, uma variação da função perda é usada para controlar a complexidade das árvores de decisão (Bentéjac *et. al*., 2020).

$$L_{xgb}(x) = \sum_{i=1}^{N} L(y_i, F(x_i)) + \sum_{m=1}^{M} \Omega(h_m) \tag{7}$$

Onde tem-se que $\sum_{i=1}^{N} L(y_i, F(x_i))$ é usado para medir a diferença entre o valor previsto $F(x_i)$ e o valor real y_i da função de perda; $\sum_{m=1}^{M} \Omega(h_m)$ é o termo de regularização aplicado à árvore h_m, em que a função penalidade é descrita conforme:

$$\Omega(h) = \gamma T + \frac{1}{2}\lambda||\omega||^2 \tag{8}$$

Na expressão, T representa o número de folhas da árvore, e ω são os valores de saída associados a essas folhas, γ é o coeficiente de penalidade da folha e λ garante que as pontuações dos nós folha não sejam muito altas. Assim, a função de perda pode ser incorporada ao critério de divisão das árvores de decisão, implementando uma estratégia de pré-poda. Logo, valores mais altos de γ resultam em árvores mais simples, já que este fator define o ganho mínimo de redução da perda necessário para dividir um nó interno (Bentéjac *et. al*., 2020).

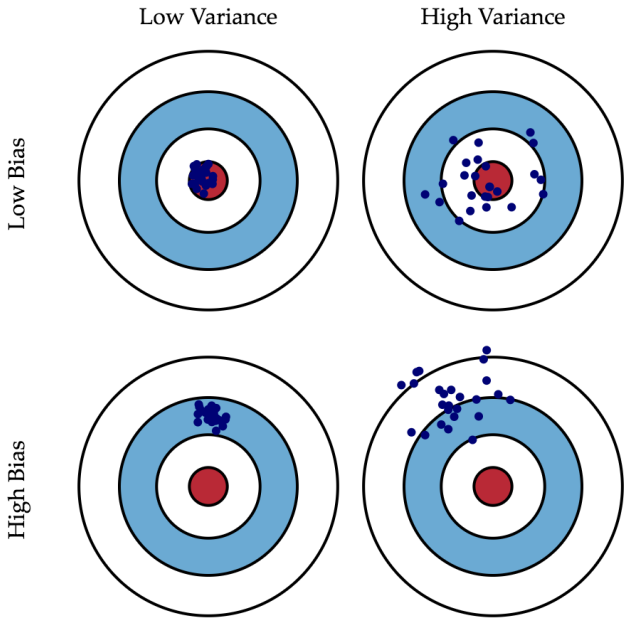
2.7 AVALIAÇÃO DE MODELOS

Nesta seção, são apresentadas as abordagens utilizadas para avaliar o desempenho dos modelos de regressão aplicados. Assim, serão discutidos os conceitos de viés e variância, bem como as principais métricas utilizadas para medir a precisão e a robustez das previsões.

2.7.1 Viés e variância

O *trade-off* entre viés e variância é fundamental no desenvolvimento de modelos de aprendizado de máquina. O viés refere-se ao erro introduzido ao simplificar um problema real, enquanto a variância mede a sensibilidade do modelo a flutuações nos dados de treinamento. Logo, modelos ideais equilibram baixo viés (alta precisão) e baixa variância (consistência), mas melhorar um aspecto geralmente piora o outro (Sharda, Delen e Turban, 2020).

Figura 8. Diagrama de alvo para viés e variância.



Fonte: Foreman-Roe, 2012.

A Figura 8 ilustra esse conceito com um diagrama de alvo, onde o centro representa o valor verdadeiro. Assim, modelos com alto viés erram consistentemente, enquanto alta variância resulta em previsões dispersas. O equilíbrio ideal é alcançado quando as previsões estão próximas do centro e consistentemente agrupadas.

Matematicamente, o Erro Quadrático Médio (MSE) esperado pode ser decomposto em variância, viés ao quadrado e erro irredutível (James *et al*., 2023):

$$\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2 + Var(\epsilon) \quad (9)$$

Onde $Var\left(\hat{f}(x_0)\right)$ é a variância do modelo; $\left[Bias\left(\hat{f}(x_0)\right)\right]^2$ é o viés ao quadrado; $Var(\epsilon)$ é a variância do erro irredutível, que não pode ser eliminada, independentemente do modelo utilizado.

A complexidade do modelo afeta diretamente esse *trade-off*: aumentar a complexidade reduz o viés, mas aumenta a variância, podendo levar a *overfitting*. O ponto ideal minimiza o erro de teste, e práticas como validação cruzada e modelos ensembles são recomendadas para lidar com esse equilíbrio (James *et al.*, 2023; Sharda, Delen e Turban, 2020).

2.7.2 Métricas de desempenho

Existem diversas métricas de desempenho para avaliar modelos de regressão, sendo algumas das mais comuns o RMSE (*Root Mean Squared Error*, ou erro

quadrático médio), o MAE (*Mean Absolute Error*, ou erro absoluto médio), o MAPE (*Mean Absolute Percentage Error*, ou erro percentual absoluto médio) e o coeficiente de determinação R^2. O RMSE quantifica a média das diferenças entre os valores previstos pelo modelo e os valores observados, medindo a precisão geral do modelo, calculado pela Equação 10:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{10}$$

Onde y_i são os valores observados, $\hat{y}_i$ são os valores previstos pelo modelo, n é o número de observações e $\overline{y}_i$ é a média das observações. Vale ressaltar que, em geral, menores valores de RMSE indicam melhor desempenho preditivo. No entanto, um RMSE muito baixo pode ser sinal de sobreajuste, especialmente se o modelo estiver excessivamente ajustado aos dados de treinamento, o que pode comprometer sua capacidade de generalização em novos dados.

O MAE mede a média das diferenças absolutas entre os valores observados e os valores preditos, sendo calculado pela Equação 11:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{11}$$

O MAE é menos sensível a *outliers* que o RMSE, pois não eleva as diferenças ao quadrado, o que o torna uma métrica útil para avaliar a precisão média do modelo.

Por sua vez, o MAPE expressa o erro percentual médio entre os valores observados e os preditos, sendo calculado pela Equação 12:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{12}$$

O MAPE é particularmente útil para interpretar o erro em termos percentuais, facilitando a comparação entre diferentes conjuntos de dados ou modelos.

Ademais, o coeficiente de determinação R^2 indica a proporção da variância dos dados que é explicada pelo modelo. Quanto mais próximo de 1 for o valor de R^2, melhor o ajuste do modelo aos dados, calculado pela Equação 13:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{13}$$

Conforme destacado por Escovedo e Koshiyama (2023), é importante ressaltar que não existe um único algoritmo de aprendizado que seja superior em todos os

casos. Logo, a escolha do melhor algoritmo depende das características específicas do problema e dos dados disponíveis, sendo recomendado experimentar diferentes abordagens para identificar aquela que proporciona o melhor desempenho em cada contexto.

3 METODOLOGIA

Esse trabalho trata-se de uma pesquisa classificada como aplicada quanto à finalidade, descritiva quanto aos objetivos, qualitativa e quantitativa quanto à sua abordagem, e pesquisa bibliográfica e experimental quanto aos procedimentos. O método científico utilizado é o hipotético-dedutivo.

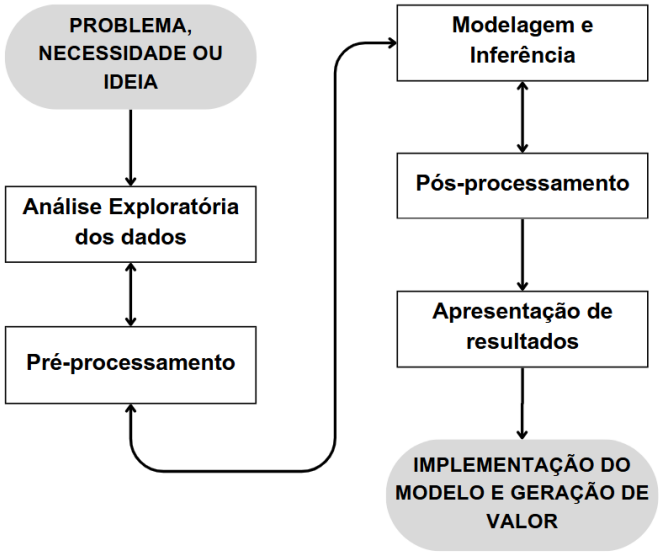
As próximas seções abordam as atividades e os procedimentos para a geração, a coleta e a organização dos dados dessa pesquisa.

3.1 CONSTRUÇÃO DO MODELO

A proposta central deste estudo envolve a construção de um modelo preditivo para estimar a potência ativa, utilizando dados do sistema SCADA (*Supervisory Control and Data Acquisition*) de uma turbina eólica localizada na Turquia. O *dataset*, publicado por Erisen (2018) na plataforma *Kaggle*, é composto por séries temporais coletadas em intervalos de 10 minutos e inclui variáveis como potência teórica, potência ativa, direção e velocidade do vento, além de informações temporais (mês e hora).

Nesse contexto, a metodologia deste trabalho segue uma sequência de etapas para a construção do modelo de previsão, esquematizado pela Figura 9. Inicia-se com a definição do problema, motivado pela necessidade de otimizar a operação de redes elétricas e garantir a estabilidade do sistema. Em seguida, na análise exploratória dos dados, são aplicados métodos estatísticos e técnicas de visualização para entender a estrutura e o comportamento dos dados. O pré-processamento trata problemas como dados faltantes, duplicados ou inconsistentes, preparando-os para a modelagem. Na etapa de modelagem e inferência, algoritmos de aprendizado de máquina são utilizados para desenvolver um modelo de regressão robusto e preciso. Em seguida, o pós-processamento avalia o desempenho do modelo, identificando pontos fortes e limitações. Os resultados são então apresentados por meio de visualizações e análises interpretativas. Por fim, a implantação do modelo visa gerar valor prático, integrando a solução ao sistema de geração de energia eólica.

Figura 9. Fluxograma dos procedimentos utilizados nesse trabalho.



Fonte: autoria própria.

Nas próximas seções, são apresentadas a descrição dos dados utilizados, bem como as ferramentas computacionais empregadas para o desenvolvimento do modelo. E, posteriormente, são detalhadas as etapas metodológicas para um entendimento completo da abordagem seguida neste trabalho.

3.1.1 Descrição dos dados

A partir do banco de dados do sistema SCADA, foi extraído um *dataset* em formato CSV com 5 atributos que são apresentados na Tabela 2. A base de dados contém um total de 50.530 instâncias, coletadas em intervalos de 10 minutos, cobrindo o período de 1º de janeiro de 2018 a 31 de dezembro de 2018.

Tabela 2. Atributos do *dataset*.

Atributo	**Descrição**
Data/Hora	Data e hora da coleta dos dados, em intervalos de 10 minutos.
Potência Ativa (KW)	Potência ativa gerada pela turbina no momento da medição, em quilowatt.
Velocidade do vento (m/s)	Velocidade do vento na altura do cubo da turbina, em metros por segundo.
Potência Teórica (KWh)	Valores teóricos de potência com a velocidade do vento fornecida, conforme fabricante da turbina, medidos em quilowatt-hora.
Direção do Vento (°)	Direção do vento na altura do cubo da turbina, medida em graus.

Fonte: autoria própria.

A Tabela 2 resume os atributos do *dataset*, fornecendo uma visão clara das

variáveis coletadas e suas unidades de medida.

3.1.2 Ferramentas computacionais

Os dados extraídos foram importados, transformados, analisados e treinados a partir do desenvolvimento de uma aplicação *web* no *Google Colab*, utilizando a linguagem de programação *Python*. Para isso, foram empregadas bibliotecas de manipulação de dados, como *Pandas* e *NumPy*, e pacotes específicos para a aplicação prática de Aprendizado de Máquina, como *scikit-learn*. As principais ferramentas são listadas e descritas na Tabela 3.

Tabela 3. Principais bibliotecas Python utilizadas para previsão.

Categoria	**Biblioteca**	**Descrição**
Manipulação de Dados	pandas (pd)	Manipulação e análise de dados tabulares
	numpy (np)	Computação numérica eficiente, operações matemáticas e manipulação de *arrays*
	missingno (ms)	Visualização e análise de dados faltantes
Visualização de Dados	matplotlib e matplotlib.pyplot (plt); seaborn (sns)	Criação de gráficos e visualizações estáticas
Pré-processamento	sklearn.preprocessing.StandardScaler; sklearn.preprocessing.MinMaxScaler	Padronização e normalização de dados
Seleção e Validação	sklearn.model_selection.train_test_split	Divisão do *dataset* em conjuntos de treino e teste
	sklearn.model_selection.KFold; sklearn.model_selection.GridSearchCV	Validação cruzada com *k folds* e busca de hiperparâmetros ótimos
Modelos de Regressão	sklearn.linear_model.LinearRegression; sklearn.tree.DecisionTreeRegressor; sklearn.ensemble.GradientBoostingRegressor sklearn.ensemble.RandomForestRegressor xgboost.XGBRegressor	Modelos de Regressão Linear, Árvore de Regressão, *Gradient Boosting Regressor, Random Forest Regressor, XGBoost*
Métricas de Avaliação	sklearn.metrics.mean_squared_error, mean_absolute_error, r2_score	Métricas de avaliação (erro quadrático médio, erro absoluto médio, R^2).

Fonte: autoria própria.

A Tabela 3 abrange ferramentas que vão desde a manipulação e visualização

de dados até a aplicação e avaliação de modelos de regressão. Adicionalmente, o cálculo da métrica MAPE (*Mean Absolute Percentage Error*) foi realizado manualmente, uma vez que o *Scikit-learn* não oferece uma função pronta para a mesma.

3.1.3 Análise exploratória dos dados

A análise exploratória de dados é uma etapa fundamental no processo de ciência de dados, cujo objetivo é compreender a estrutura, distribuição e características dos dados. Nessa fase, são realizadas técnicas estatísticas e visualizações para identificar padrões, tendências, *outliers* e possíveis problemas nos dados, como valores faltantes ou inconsistentes.

Outliers são valores atípicos que se destacam significativamente em relação à maioria dos dados, podendo ocorrer tanto nas *features* quanto na variável de resposta, sendo nestes casos conhecidos como pontos de alta alavancagem (Klosterman, 2020). No contexto do problema de geração de energia eólica, esses valores podem ser causados por erros de medição, falhas no sistema SCADA, condições de operação anômalas (como inicialização ou parada da turbina), problemas na integração com a rede elétrica, ou até mesmo tratamento inadequado dos dados durante a coleta.

Por isso, a identificação dessas inconsistências é crucial para garantir a qualidade dos dados, como a remoção ou correção desses registros. Além disso, a análise exploratória auxilia na definição de estratégias para pré-processamento e modelagem, garantindo que os dados estejam adequados para as etapas subsequentes.

3.1.4 Pré-processamento

O pré-processamento de dados é a etapa subsequente à análise exploratória, com o objetivo de preparar os dados brutos para análise e modelagem, garantindo qualidade, consistência e representatividade do conjunto. Inicialmente, foi criada uma cópia do *dataset* para preservar os dados originais antes da aplicação de qualquer filtro.

A etapa incluiu limpeza, tratamento de *outliers*, nulos, e seleção de características relevantes. Assim, neste trabalho optou-se por utilizar uma abordagem adaptativa baseada no método interquartil, aplicado de forma segmentada a diferentes

faixas de valores da coluna 'Potência Ativa [kW]'. Dessa forma, permitindo uma remoção mais precisa de *outliers*, considerando as particularidades de cada intervalo de potência.

O método interquartil tradicional utiliza os quartis Q_1 (primeiro quartil) e Q_3 (terceiro quartil) para definir um intervalo de valores considerados normais, excluindo aqueles que estão abaixo de $Q_1 - 1.5 * IQR$ ou acima de $Q_3 + 1.5 * IQR$, onde $IQR = Q_3 - Q_1$. No entanto, para aumentar a flexibilidade e adaptabilidade do filtro, optou-se por utilizar percentis diretamente, permitindo ajustes mais refinados conforme a distribuição dos dados em cada subconjunto.

O procedimento foi dividido em duas etapas principais: a criação de subconjuntos de dados com base em intervalos de potência ativa e a aplicação da filtragem de *outliers* na coluna de velocidade do vento para cada subconjunto. Inicialmente, os dados de Potência Ativa foram divididos em intervalos de tamanho fixo (por padrão, 50 unidades), começando em 20kW e indo até um limite superior definido (3400kW), sendo o último subconjunto todos os valores maiores ou iguais a 3300kW.

Em seguida, para cada subconjunto, foi aplicada uma filtragem baseada nos percentis da variável 'Velocidade do Vento [m/s]'. O critério adotado considerou um percentil inferior fixo, equivalente a 0.0001, com o intuito de remover eventuais valores anômalos extremamente baixos, que poderiam comprometer a análise. O percentil superior, por sua vez, foi ajustado dinamicamente para cada subconjunto, levando em consideração o limite inferior da potência ativa correspondente.

Assim, para subconjuntos da 'Potência Ativa [kW]' cujo limite inferior fosse menor ou igual a 3 kW, o percentil superior foi fixado em 0.95. Para limites entre 3 kW e 10 kW, o percentil superior foi ajustado para 0.90. Subconjuntos com limites entre 10 kW e 20 kW adotaram um percentil de 0.92, enquanto os intervalos entre 20 kW e 30 kW utilizaram um percentil de 0.96. Ademais, para os subconjuntos com limite inferior igual ou superior a 30 kW, o percentil superior foi definido como 0.985. Após a aplicação do filtro em todos os subconjuntos, os dados foram reunidos em um único conjunto, garantindo a reindexação sequencial das linhas.

Nesse sentido, cada subconjunto corresponde a um intervalo de potência crescente, permitindo capturar diferentes regimes operacionais da turbina. Logo, eliminaram-se valores anômalos sem comprometer dados válidos, preservando a coerência física do sistema, uma vez que, em baixas potências, pequenas variações

na velocidade do vento têm menor impacto na operação da turbina, e valores extremos são mais provavelmente *outliers*. Já em altas potências, a turbina opera próximo ao limite nominal, onde ventos mais fortes são esperados e, portanto, apenas valores extremamente altos foram considerados anômalos.

Além disso, a coluna "Data/Hora" foi convertida para o formato *datetime*, permitindo a extração de informações temporais como hora, dia da semana e mês, essenciais para a correta interpretação dos dados pelos modelos.

Para evitar *data leakage*, ou seja, a contaminação dos dados de treino com informações dos dados de teste, técnicas de normalização ou padronização não foram aplicadas diretamente nos dados brutos. No entanto, durante o *pipeline* de modelagem, utilizou-se a padronização via *StandardScaler*, essencial para algoritmos sensíveis à magnitude das variáveis, como regressão linear. Com isso, essas operações asseguraram que os dados estivessem adequados para análise, maximizando a performance preditiva e alinhando os resultados ao comportamento físico esperado.

3.1.4.1 Divisão de dados

A divisão dos dados é uma etapa essencial no desenvolvimento de modelos de aprendizado de máquina, sendo responsável por separar o *dataset* em subconjuntos distintos para treinamento e teste. Esse processo permite avaliar a capacidade de generalização do modelo, ou seja, sua habilidade de realizar previsões precisas em novos dados.

Neste estudo, utilizou-se o método *holdout*, no qual os dados foram particionados em dois conjuntos exclusivos: treinamento e teste. O conjunto de treinamento é utilizado para ajustar os parâmetros do modelo, enquanto o conjunto de teste serve para avaliar seu desempenho em dados inéditos. No entanto, a escolha da proporção ideal depende do tamanho do *dataset*.

Para este trabalho, optou-se pela divisão 80/20, onde 80% dos dados foram utilizados para treinar o modelo e 20% foram reservados para teste. Essa escolha foi motivada pelo fato de que originalmente o *dataset* continha 50.530 instâncias e após o pré-processamento, o *dataset* manteve um volume significativo de dados (38.839 instâncias), permitindo que o modelo fosse treinado com um conjunto amplo sem comprometer a representatividade do conjunto de teste. Assim, a divisão adotada garantiu um equilíbrio entre aprendizado e validação, assegurando uma avaliação

confiável do desempenho do modelo.

3.1.5 Modelagem e inferência

Nessa fase, são aplicados algoritmos de aprendizado de máquina que aprendem padrões a partir dos dados e permitem fazer previsões ou inferências sobre novos dados. Para isso, tais modelos buscam estabelecer uma relação matemática entre as variáveis independentes e a variável dependente (Potência Ativa). A escolha dos modelos e a avaliação de seu desempenho são fundamentais para garantir previsões precisas e confiáveis.

Neste trabalho, foram aplicados diversos algoritmos de regressão, incluindo Regressão Linear, *Gradient Boosting*, Árvore de Regressão, *XGBoost* e *Random Forest*, com o objetivo de identificar o modelo com o melhor desempenho.

Ainda, para garantir uma avaliação robusta e confiável, aplicou-se a técnica de Validação Cruzada *5-folds*, aliada à otimização de hiperparâmetros, visando aprimorar o desempenho dos modelos.

3.1.5.1 Validação cruzada

A validação cruzada é uma técnica essencial para avaliar a precisão de modelos, reduzindo o viés associado a uma única divisão dos dados em treinamento e teste. De acordo com Sharda, Delen e Turban (2020), a abordagem mais comum é a validação cruzada *k-fold*, onde o *dataset* é dividido aleatoriamente em *k* subconjuntos mutuamente exclusivos (ou *folds*). O modelo é então treinado e testado *k* vezes, usando a cada iteração todos os subconjuntos, exceto um (*k-1*), para treinamento, e o restante para teste. Ao final, a precisão geral do modelo é calculada pela média das métricas de avaliação de cada iteração.

Segundo Bruce e Bruce (2019), a validação cruzada supera as limitações das análises "na amostra" ao introduzir a validação "fora da amostra". Isso permite que todos os registros sejam usados tanto para ajuste quanto para avaliação, fornecendo uma estimativa mais confiável do desempenho do modelo.

Portanto, o processo básico para a validação cruzada *5-folds* utilizada neste trabalho envolveu os seguintes passos:

1. Dividir os dados em *5 folds* (5 partes).
2. Reservar 1/5 dos dados como amostra de retenção (teste) e treinar o

modelo nos dados restantes.

3. Avaliar o desempenho do modelo no *fold* de retenção e registrar as métricas necessárias.
4. Repetir o processo para cada *fold*, garantindo que todos os registros participem tanto do treinamento quanto da retenção.
5. Calcular a média das métricas para obter uma avaliação geral do modelo.

3.1.5.2 Hiperparâmetros

Os hiperparâmetros são ajustes externos que determinam o comportamento e o desempenho de modelos de aprendizado de máquina, mas diferem dos parâmetros do modelo por não serem aprendidos diretamente dos dados durante o treinamento. Assim, os hiperparâmetros podem variar conforme o algoritmo e o tipo de modelo utilizado, e devem ser definidos antes do treinamento, influenciando como o modelo é treinado e como realiza previsões (Klosterman, 2020).

Nesse sentido, a escolha adequada desses valores exige experimentação, sendo comum o uso de métodos automatizados, como o *grid search*, para otimização. Entretanto, para uma avaliação mais precisa, os dados podem ser divididos em treinamento, validação e teste, permitindo ajustes antes da avaliação final. Além disso, a validação cruzada reduz a variabilidade, tornando a análise mais confiável (Escovedo e Koshiyama, 2023; Klosterman, 2020).

Os hiperparâmetros submetidos à otimização podem ser verificados na Tabela 4.

Tabela 4. Descrição dos hiperparâmetros e valores padrão na busca em *grid search*.

Modelo	**Nome**	**Default**	**Descrição**
Árvore de Regressão	max_depth	None	A profundidade máxima da árvore. Se for None, os nós serão expandidos até que todas as folhas sejam puras ou contenham menos do que min_samples_split amostras.
	min_samples_split	2	O número mínimo de amostras necessário para dividir um nó interno
Random Forest	max_depth	None	A profundidade máxima da árvore. Se for None, então os nós são expandidos até que todas as folhas sejam puras ou até que todas as folhas contenham menos do que min_samples_split amostras.
	min_samples_split	2	O número mínimo de amostras necessárias para dividir um nó interno.
	n_estimators	100	O número de árvores na floresta.
Gradient Boosting	max_depth	3	A profundidade máxima dos estimadores de regressão individuais. Esse parâmetro limita o número de nós na árvore.
	n_estimators	100	O número de etapas de *boosting* a serem realizadas. O *gradient boosting* é relativamente robusto contra overfitting, então um número maior de estimadores geralmente melhora o desempenho. Os valores devem estar no intervalo [0.0,∞).
	learning_rate	0.1	A taxa de aprendizado reduz a contribuição de cada árvore pelo fator learning_rate. Há um equilíbrio entre learning_rate e n_estimators. Os valores devem estar no intervalo [0.0,∞).
XGBoost	max_depth	6	Profundidade máxima das árvores.
	n_estimators	100	Número de árvores.
	learning_rate	0.3	Taxa de aprendizado.

Fonte: Documentação da biblioteca *scikit-learn* e XGBoost.

A escolha dos hiperparâmetros para otimização dos modelos depende de fatores como a complexidade do problema, o tamanho do *dataset* e os recursos computacionais disponíveis. Dessa forma, para os modelos de regressão baseados em árvores, como RF, *Gradient Boosting* e *XGBoost*, os principais hiperparâmetros ajustados foram o número de árvores, a profundidade máxima das árvores e critérios que influenciam a divisão dos nós. No caso do RF, um número maior de árvores pode melhorar a generalização, enquanto a limitação da profundidade controla o risco de

overfitting. Entretanto, no *Gradient Boosting* e no *XGBoost*, a taxa de aprendizado regula a influência de cada árvore, equilibrando viés e variância do modelo.

A árvore de regressão simples também permite ajustes na profundidade e no número mínimo de amostras para divisão, sendo útil para capturar relações não lineares sem o custo computacional do *boosting*. Em contraste, a regressão linear não possui hiperparâmetros ajustáveis para configuração básica, pois seu desempenho depende diretamente das variáveis de entrada e da relação linear entre elas.

3.1.6 Pós-processamento

O pós-processamento é a etapa onde os resultados são analisados, interpretados e utilizados para tomar decisões informadas. Nesta fase, os modelos são comparados com base em métricas de desempenho, o melhor modelo é selecionado e, finalmente, ele é salvo para uso futuro. Além disso, é comum criar interfaces para permitir que o modelo seja utilizado de forma prática, como fazer previsões com base em novos dados fornecidos pelo usuário.

3.1.6.1 Interação com o usuário

Em projetos de ciência de dados, após o desenvolvimento e validação de um modelo, é comum implementar uma interface que permita a interação com o usuário, possibilitando a aplicação prática do modelo em cenários reais. Essa etapa é crucial para demonstrar a utilidade do modelo e sua capacidade de gerar previsões com base em novos dados. Assim, a interação pode ser feita por meio de interfaces gráficas, aplicações *web* ou, em casos mais simples, como o deste trabalho, por meio de scripts que solicitam entradas do usuário diretamente no terminal. A escolha da abordagem depende da complexidade do projeto e do público-alvo. Portanto, independentemente do método, o objetivo é tornar o modelo acessível e funcional, permitindo que usuários sem conhecimento técnico possam utilizá-lo de forma intuitiva.

4 RESULTADOS

Neste capítulo serão apresentados os resultados gerados a partir da aplicação dos modelos de regressão, bem como as análises correspondentes. A estruturação segue a mesma sequência da metodologia adotada, priorizando a exibição de gráficos, métricas de desempenho e considerações pós-processamento.

4.1 ANÁLISE EXPLORATÓRIA DOS DADOS

No presente trabalho, a análise exploratória foi iniciada com a inspeção dos dados por meio dos métodos dataset.shape(), dataset.info(), dataset.head(), dataset.tail(), a fim de exibir a estrutura do *dataset*, incluindo o número de linhas e colunas, os tipos de variáveis e uma prévia dos primeiros e últimos registros.

Nessa etapa, o *dataset* possuía dimensões com 50.530 amostras (linhas) e 5 variáveis (colunas). E assim, foi realizado um resumo estatístico do *dataset*, evidenciando métricas como média, desvio padrão, valores mínimo e máximo, e os quartis (25%, 50% e 75%) para cada atributo numérico, com exceção da coluna Data/Hora, como pode ser visualizado na Tabela 5. Durante essa análise, foi observado que o valor mínimo da variável Potência Ativa apresentou um valor negativo, além de valores anômalos da velocidade do vento, o que é inconsistente com o contexto físico de geração de energia. Essa inconsistência evidenciou a necessidade de tratamento de dados para corrigir ou remover tais anomalias.

Tabela 5. Resumo estatístico do *dataset*.

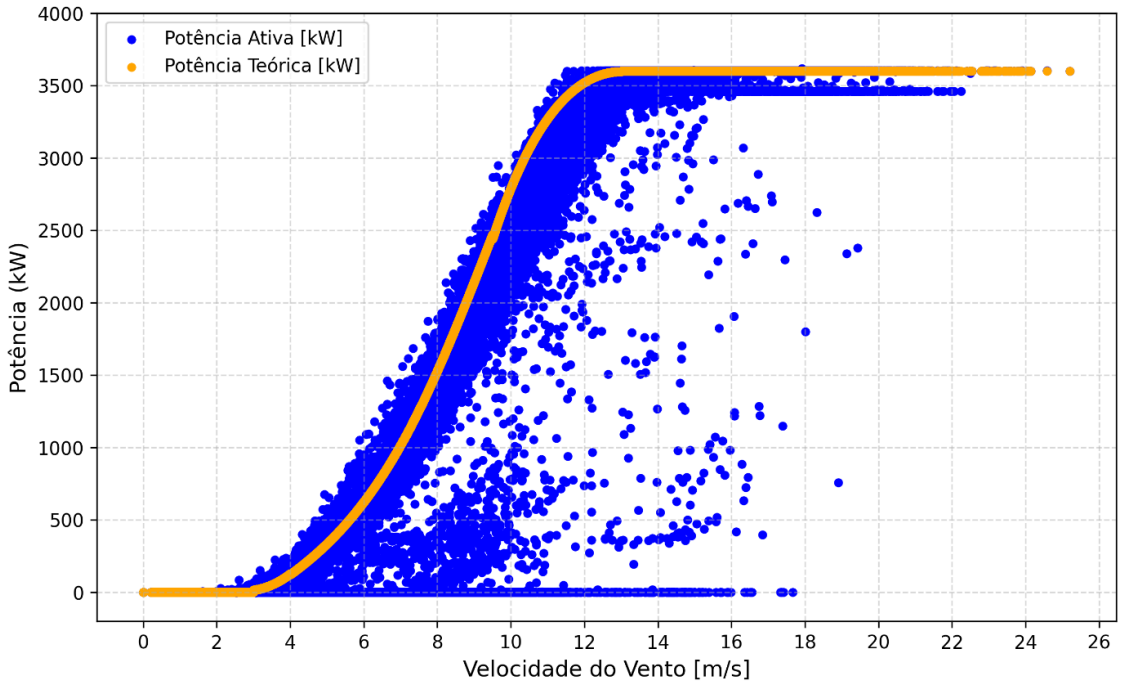
Estatística	**Potência Ativa [kW]**	**Velocidade do Vento [m/s]**	**Potência Teórica [kW]**	**Direção do Vento [°]**
Contagem	50530.000	50530.000	50530.000	50530.000
Média	1307.684	7.557	1492.175	123.687
Desvio Padrão	1312.459	4.227	1368.018	93.443
Mínimo	-2.471	0.000	0.000	0.000
25%	50.677	4.201	161.328	49.315
Mediana	825.838	7.104	1063.776	73.712
75%	2482.507	10.300	2964.972	201.696
Máximo	3618.732	25.206	3600.000	359.997

Fonte: autoria própria.

Ademais, também foram utilizadas visualizações gráficas para compreender a distribuição das variáveis e suas relações. Histogramas e gráficos de densidade

permitiram analisar a distribuição de cada variável, enquanto gráficos de dispersão evidenciaram a relação entre a potência ativa e outras variáveis. A Figura 10 evidencia a relação entre variáveis como Potência Ativa, Velocidade do Vento e Potência Teórica.

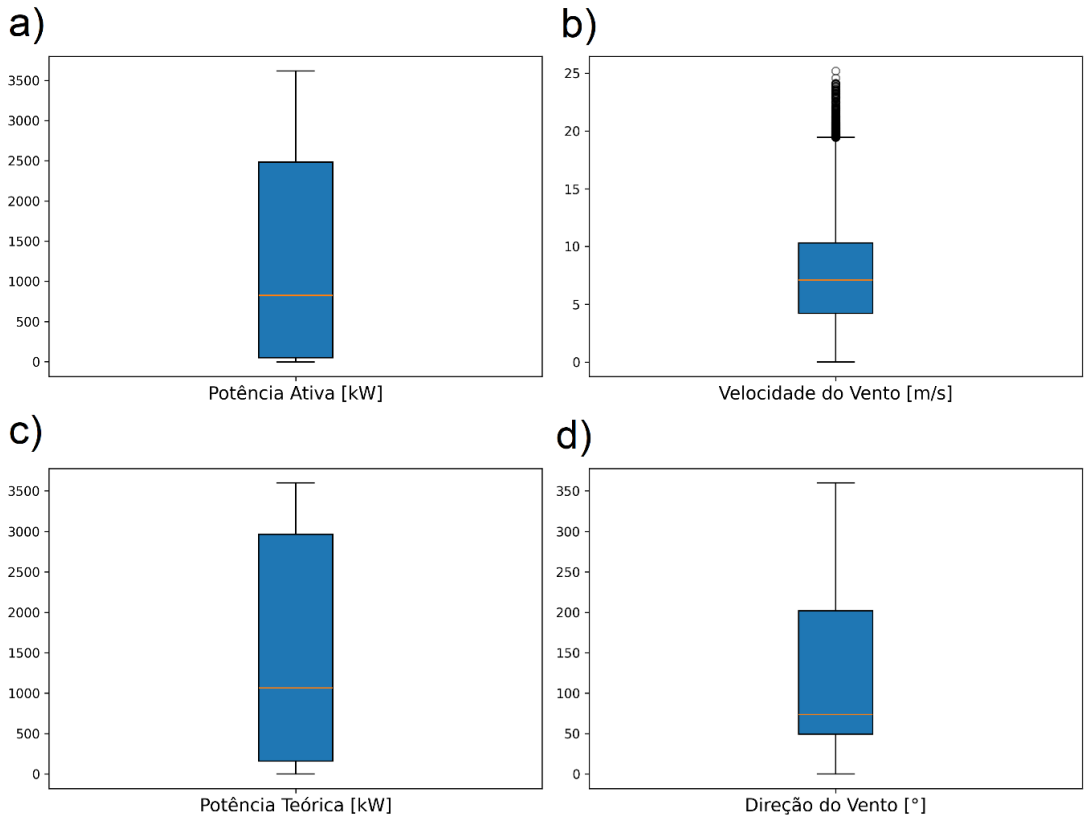
Figura 10. Curva de Potência *vs* Velocidade do Vento.



Fonte: autoria própria.

Nesse contexto, também foram verificados e constatados a ausência de dados faltantes. Além disso, utilizou-se a técnica de Box plot para identificar possíveis *outliers* no *dataset*.

Figura 11. *Box plot* a) para a potência ativa; b) velocidade do vento; c) potência teórica; d) direção do vento.



Fonte: autoria própria.

Ao analisar os *Box plots* da Figura 11, é possível concluir que a distribuição de potência ativa e teórica mostram uma ampla variação, de 0 a 3.500kW, e medianas próximas do limite inferior da caixa. Os dados apontam que a potência ativa apresenta *outliers* próximo a 0kW, indicando momentos em que a turbina não estava gerando energia, possivelmente devido a condições de vento insuficientes. Entretanto, o valor negativo está muito próximo de zero em comparação com a escala geral dos dados e não aparece no *box plot*. Por sua vez, a velocidade do vento apresenta uma distribuição mais concentrada, variando de 0 a 25m/s e mediana em torno de 5-10 m/s. No entanto, há alguns *outliers* na parte superior, com velocidades de vento muito altas, possivelmente justificado por eventos de ventos extremos. Por fim, a direção do vento apresenta uma assimetria leve à direita, com maior concentração de valores abaixo da mediana. Ainda assim, os dados permanecem dentro de um intervalo esperado para esse tipo de variável, que varia entre 0° e 360°.

4.2 PRÉ-PROCESSAMENTO

Para esta etapa, foi gerado mais um resumo estatístico (Tabela 6) para verificar a distribuição dos dados após a aplicação dos métodos de limpeza. O *dataset* passou então a conter 38.839 instâncias, além de valores de Potência Ativa e Velocidade do Vento dentro das faixas esperadas.

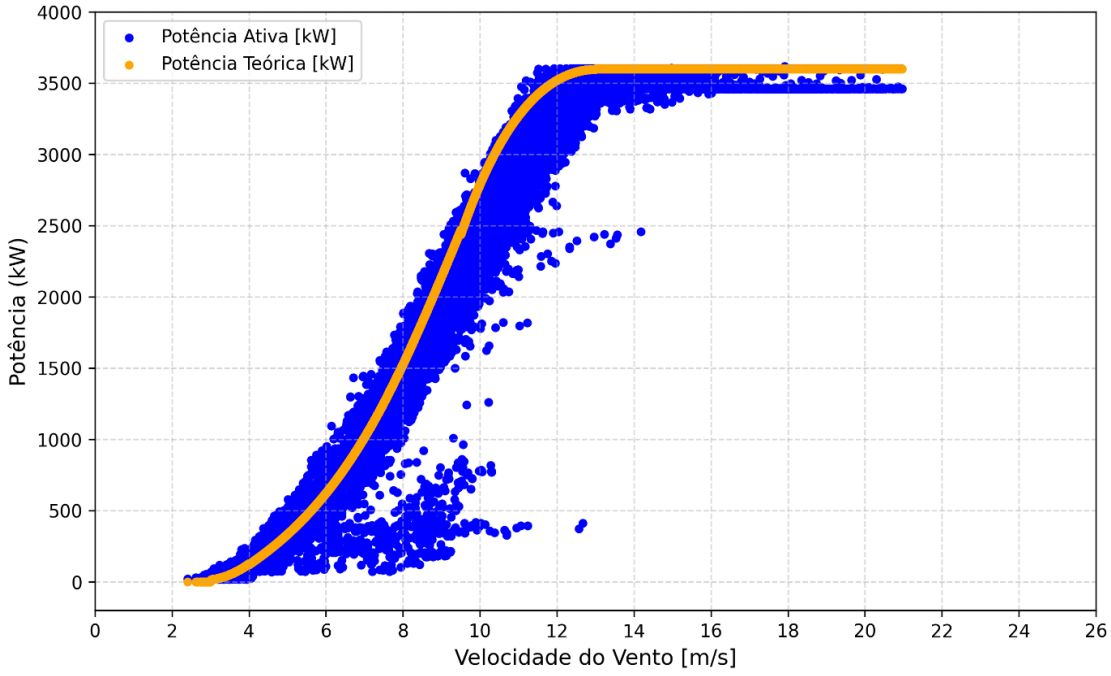
Tabela 6. Resumo estatístico do *dataset* após limpeza.

Estatística	**Potência Ativa [kW]**	**Velocidade do Vento [m/s]**	**Potência Teórica [kW]**	**Direção do Vento [°]**
Contagem	38839.000	38839.000	38839.000	38839.000
Média	1744.570	8.910	1892.184	114.859
Desvio Padrão	1261.763	3.638	1297.773	85.331
Mínimo	20.047	2.404	0.000	0.015
25%	549.767	6.097	657.484	49.395
Mediana	1520.890	8.275	1691.216	69.855
75%	3057.109	11.350	3382.517	196.770
Máximo	3618.732	20.969	3600.000	359.997

Fonte: autoria própria.

Em seguida, foi gerado um gráfico de dispersão das curvas de potência ativa e teórica em função da velocidade do vento (Figura 12).

Figura 12. Curva de potência *vs* Velocidade do Vento após limpeza.



Fonte: autoria própria.

O gráfico acima permite analisar a relação entre as variáveis e verificar a

remoção dos *outliers*, evidenciando uma melhoria significativa quando comparado com a distribuição original da Figura 10.

4.3 MODELAGEM E INFERÊNCIA

A modelagem foi realizada utilizando os algoritmos discutidos no tópico 3.1.5: Regressão Linear, *Gradient Boosting*, Árvore de Regressão, *XGBoost* e *Random Forest.* Logo, para cada modelo, aplicou-se uma busca em grade (*GridSearch*) com validação cruzada de 5 *folds*, visando a otimização dos hiperparâmetros. A seção abaixo apresenta os melhores resultados encontrados para cada modelo.

4.3.1 Hiperparâmetros e validação cruzada

A escolha dos melhores hiperparâmetros foi realizada por meio de uma busca em grade (*Grid Search*) combinada com validação cruzada de 5 pastas (*5-fold cross-validation*), garantindo que os hiperparâmetros selecionados fossem robustos e generalizassem bem para diferentes subconjuntos dos dados.

A Tabela abaixo lista os valores testados para os hiperparâmetros e também evidencia os melhores hiperparâmetros na busca do *Gridsearch*.

Tabela 7. Valores utilizados por hiperparâmetros na busca do Gridsearch e melhores hiperparâmetros encontrados.

Modelo	**Hiperparâmetro**	**Valores Testados Gridsearch**	**Melhor hiperparâmetro**
Árvore de Regressão	max_depth	None, 5, 10	5
	min_samples_split	2, 5, 10	2
Gradient Boosting	n_estimators	50, 100, 200	200
	learning_rate	0.01, 0.1, 0.2	0.1
	max_depth	3, 5, 7	3
Random Forest	n_estimators	50, 100, 200	200
	max_depth	None, 10, 20	10
	min_samples_split	2, 5, 10	10
XGBoost	n_estimators	50, 100, 200	200
	learning_rate	0.01, 0.1, 0.2	0.1
	max_depth	3, 5, 7	3
Regressão Linear	-	-	N/A

Fonte: autoria própria.

Nesse contexto, os hiperparâmetros foram definidos com base em valores comumente utilizados na literatura e em práticas recomendadas para problemas de

regressão, conforme destacado por Bentéjac *et al*. (2020). Para a Árvore de Regressão, foram ajustados a profundidade máxima da árvore e o número mínimo de amostras para dividir um nó, visando equilibrar a complexidade do modelo e evitar *overfitting*. No *Gradient Boosting Regressor*, foram selecionados o número de árvores, a taxa de aprendizado e a profundidade máxima para controlar a capacidade de generalização. Para o *Random Forest*, além desses parâmetros, foi incluído o número mínimo de amostras para dividir um nó para regularizar a divisão dos nós. No *XGBoost*, os mesmos parâmetros foram ajustados, com ênfase na taxa de aprendizado para garantir um aprendizado eficiente. Por fim, a Regressão Linear foi utilizada como *baseline*, uma vez que não possui hiperparâmetros ajustáveis.

4.4 PÓS- PROCESSAMENTO

Após a otimização dos hiperparâmetros e a validação dos modelos, os resultados foram consolidados como dispostos na Tabela 8 para facilitar a comparação. Assim, as métricas utilizadas para avaliação foram o Erro Absoluto Médio (MAE), Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R^2) e erro absoluto percentual médio (MAPE).

Tabela 8. Comparação dos modelos otimizados.

Modelo	**MAE (kW)**	**RMSE (kW)**	**R^2**	**MAPE (%)**
XGBoost	90.0251	170.1738	0.9818	12.8824
Gradient Boosting Regressor	90.6475	171.2093	0.9816	13.1629
Random Forest	90.4044	172.2621	0.9814	12.5558
Árvore de Regressão	99.4215	180.0899	0.9796	15.1225
Regressão Linear	117.3877	192.0331	0.9768	21.2535

Fonte: autoria própria.

Como evidenciado, o modelo *XGBoost* obteve os melhores resultados gerais, com um MAE de aproximadamente 90.03 kW, RMSE de 170.17 kW, um coeficiente de determinação R^2 de 0.9818 e um MAPE de 12.88%. Esses resultados são consistentes com as características do algoritmo, que utiliza o método de *boosting* com regularização para corrigir iterativamente os erros de previsões anteriores, capturando complexidades não lineares na geração de potência eólica. Assim, a regularização previne sobreajuste, garantindo robustez frente a variações sazonais e

ruídos nos dados. Contudo, destaca-se a perda de interpretabilidade, uma vez que a combinação de várias árvores encadeadas dificulta a análise direta do impacto individual das variáveis no resultado final, o que pode ser um obstáculo em situações onde a transparência do modelo é necessária para controle operacional ou análise de falhas.

O *Gradient Boosting Regressor* apresentou resultados muito próximos aos do *XGBoost*, com MAE de 90.65 kW, RMSE de 171.20 kW, R^2 de 0.9816 e MAPE de 13.16%. Apesar de levemente inferior ao *XGBoost*, seu desempenho é esperado devido ao uso do mesmo princípio básico de *boosting*. Entretanto, o *Gradient Boosting* convencional não possui otimizações como regularização explícita e tratamento eficiente de valores ausentes, o que explica a pequena diferença de desempenho. Como no *XGBoost*, a principal limitação está na interpretabilidade reduzida, já que o modelo se constrói a partir da soma de várias árvores pequenas, tornando-se menos transparente para operadores ou engenheiros interessados em entender detalhadamente as condições que levam a determinada previsão.

O *Random Forest* também apresentou desempenho muito sólido, com MAE de 90.40 kW, RMSE de 172.26 kW, R^2 de 0.9814 e MAPE de 12.55%, sendo este último o menor entre os modelos, indicando bom desempenho em diferentes faixas de potência. O método de *bagging*, que constrói múltiplas árvores com subconjuntos aleatórios de dados e variáveis, reduz a variância e aumenta a resistência a ruídos, como os causados por turbulência do vento ou medições imprecisas. Além disso, o modelo permite análises de importância das variáveis, oferecendo um equilíbrio entre precisão e interpretabilidade, embora o modelo completo, com centenas de árvores, não seja intuitivamente compreensível.

A Árvore de Regressão Simples apresentou métricas inferiores, com MAE de 99,42 kW, RMSE de 180,09 kW, R^2 de 0,9797 e MAPE de 15,12%. Esses resultados eram esperados, pois o modelo, baseado em uma única estrutura de decisão, não consegue corrigir erros ou reduzir variações como os métodos *ensemble*. Essa limitação impede a captura de transições suaves entre regimes operacionais, como mudanças do regime laminar para turbulento do vento ou a entrada em potência nominal, especialmente em faixas intermediárias, onde a variabilidade é maior. Logo, isso justifica os erros maiores nessas regiões, refletidos no RMSE e MAPE, que penalizam desvios significativos. Apesar disso, sua principal vantagem é a interpretabilidade, permitindo visualizar claramente como faixas de velocidade do

vento e direção influenciam a potência, sendo útil para diagnósticos e manutenção, mesmo com desempenho preditivo inferior.

O modelo de Regressão Linear apresentou o pior desempenho dentre os testados, com MAE de 117.39 kW, RMSE de 192.03 kW, R^2 de 0.9768 e MAPE de 21.25%. Essa limitação é coerente com a própria natureza do modelo, que pressupõe uma relação linear direta entre as variáveis independentes e a variável alvo. No contexto físico de uma turbina eólica, tal simplificação é insuficiente, dado que a potência ativa não responde de forma linear à velocidade do vento, sendo influenciada por efeitos não lineares como o estol aerodinâmico e regimes de controle específicos do equipamento. Entretanto, a simplicidade e facilidade de implementação deste modelo podem justificar seu uso em cenários de análise preliminar ou onde restrições computacionais sejam um fator relevante.

De maneira geral, observa-se um compromisso claro entre precisão preditiva e interpretabilidade dos modelos. As técnicas baseadas em *boosting*, como o *XGBoost* e o *Gradient Boosting Regressor*, oferecem maior precisão, sendo recomendadas para aplicações que demandam alta confiabilidade na previsão de potência, como no planejamento de oferta para redes elétricas. No entanto, essa precisão vem acompanhada de uma maior complexidade, que pode ser um fator limitante em situações onde a explicação detalhada do comportamento do sistema é necessária para fins operacionais ou regulatórios. Ademais, o RF oferece uma alternativa interessante, conciliando boa precisão com razoável capacidade de análise das variáveis mais relevantes. Por outro lado, modelos como a Árvore de Regressão Simples e a Regressão Linear, apesar de menos precisos, apresentam vantagens em termos de simplicidade e clareza interpretativa, sendo úteis para aplicações em operações locais, manutenção preditiva ou análise de falhas específicas.

No contexto deste trabalho, que visa otimizar a previsão de potência ativa em parques eólicos, o *XGBoost* destacou-se como o modelo mais adequado por sua confiabilidade e eficiência no tratamento de grandes volumes de dados complexos, além de sua capacidade de minimizar erros e adaptar-se a diferentes cenários. No entanto, em cenários mais amplos, a escolha do modelo ideal dependerá dos critérios específicos de cada problema, devendo-se ponderar a tolerância ao erro, a necessidade de transparência e a complexidade operacional exigida pelo setor energético.

4.4.1 Interação com o usuário

No presente trabalho, foi desenvolvido um *loop* iterativo que permite ao usuário inserir dados manualmente e obter previsões da potência ativa gerada pela turbina eólica. Logo, a ferramenta utilizada para essa interação foi a função *input()* do *Python*, que captura as entradas do usuário, como a velocidade do vento, a potência teórica e a direção do vento.

No código, esses dados são validados para garantir que estejam dentro de faixas aceitáveis (por exemplo, a velocidade do vento deve ser um valor positivo, e a direção do vento deve estar entre 0° e 360°). Após a validação, os dados são convertidos em uma estrutura tabular e passados para o modelo carregado, que retorna a previsão da potência ativa. O *loop* continua até que o usuário decida encerrar as previsões. Ainda, foi utilizada a biblioteca *joblib* para salvar e carregar o modelo, garantindo que o processo seja ágil e que o modelo possa ser reutilizado sem a necessidade de um novo treinamento. A Figura 13 ilustra a saída do sistema com um exemplo.

Figura 13. Exemplo de entrada de dados na tela de interação com o usuário.

```
Por favor, insira os detalhes para a previsão da potência ativa:
Velocidade do Vento (m/s): 5.79
Potência Teórica (kW): 560.00
Direção do Vento (°): 258.7

A previsão da potência ativa é: 494.11 kW

Gostaria de inserir novos dados? (s/n): n
Encerrando previsões.
```

Fonte: autoria própria.

Com essa interação, é possível que o modelo possa ser utilizado por operadores ou gestores de usinas eólicas para prever a geração de energia com base em condições específicas do vento, contribuindo para uma gestão mais eficiente dos recursos energéticos.

5 CONCLUSÃO

O principal objetivo deste trabalho foi desenvolver um modelo preditivo de potência ativa em aerogeradores, utilizando técnicas de aprendizado de máquina aplicadas a dados provenientes do sistema SCADA de uma turbina eólica localizada na Turquia. Para isso, foram estabelecidos objetivos específicos, os quais foram cumpridos conforme descrito a seguir.

A análise exploratória permitiu compreender o comportamento das principais variáveis (velocidade do vento, potência ativa, potência teórica e direção do vento), identificando tendências, padrões e inconsistências. Assim, análise gráfica e estatística evidenciou a forte correlação entre a velocidade do vento e a potência gerada, alinhada ao esperado para sistemas eólicos, direcionando as etapas seguintes.

A limpeza e filtragem dos dados foi realizada junto ao pré-processamento, incluindo remoção de registros nulos e *outliers* utilizando métodos estatísticos como desvio-padrão e análise interquartil. Ressalta-se ainda que, por se tratar de um *dataset* disponível na internet, não é possível garantir a total fidedignidade da origem dos dados, podendo conter alterações ou omissões que impactem os resultados. Logo, essa etapa garantiu que o *dataset* utilizado nos modelos fosse robusto e representativo da realidade física do sistema analisado, minimizando ruídos e erros provenientes de medições.

No pré-processamento dos dados, também foi aplicada a normalização das variáveis, conversão de datas para formatos apropriados e separação entre conjunto de treino e teste. Nesse sentido, tais ações possibilitaram um ambiente adequado para o aprendizado dos algoritmos, evitando viés e garantindo maior generalização dos modelos.

O desenvolvimento de modelos de previsão foi realizado utilizando diferentes algoritmos de aprendizado de máquina, como Regressão Linear, Árvore de Regressão, *Random Forest*, *Gradient Boosting Regressor* e *XGBoost*. Cada um deles foi submetido a uma otimização de hiperparâmetros, garantindo que o desempenho fosse extraído ao máximo dentro das limitações de cada modelo.

Por fim, o desempenho dos modelos foi avaliado com base nas métricas MAE, RMSE, R^2 e MAPE. Logo, foi possível verificar que os modelos mais complexos, como *XGBoost* e *Gradient Boosting*, apresentaram maior precisão, alcançando valores de

R^2 acima de 0,98 e menores erros médios absolutos. Contudo, essa elevação na performance traz consigo uma redução significativa na interpretabilidade. À vista disso, modelos mais simples, como a Regressão Linear e a Árvore de Regressão, embora tenham apresentado métricas inferiores, oferecem maior transparência e facilidade de interpretação, características muitas vezes desejáveis em ambientes industriais onde a compreensão do comportamento físico do sistema é fundamental.

Conclui-se, portanto, que este trabalho oferece uma contribuição relevante para a previsão de energia eólica, apresentando um modelo robusto e preciso para estimar a potência ativa de aerogeradores. A combinação de técnicas de aprendizado de máquina com uma análise e pré-processamento rigorosos dos dados permitiu desenvolver soluções que otimizam a operação de parques eólicos, reduzem perdas mecânicas e melhoram o equilíbrio entre produção e demanda. Dentre os modelos avaliados, o *XGBoost* foi selecionado como o mais adequado, devido à sua alta capacidade preditiva, confiabilidade frente a grandes volumes de dados e eficiência na redução de erros, características essenciais para aplicações industriais. Embora modelos complexos apresentem limitações, como a menor interpretabilidade, os resultados evidenciam seu potencial para impulsionar a eficiência e sustentabilidade na geração eólica, alinhando-se aos desafios globais de transição para energias renováveis.

REFERÊNCIAS BIBLIOGRÁFICAS

ABEEólica. Brasil permanece em 6º lugar no ranking mundial de energia eólica. Disponível em: <https://abeeolica.org.br/brasil-permanece-em-6o-lugar-no-ranking-mundial-de-energia-eolica/#:~:text=O%20relat%C3%B3rio%20destaca%20o%20impulso,anterior%20em%20rela%C3%A7%C3%A3o%20a%202022>. Acesso em 15 de maio 2024.

Agência Nacional de Energia Elétrica (ANEEL). (2024). Primeiro trimestre de 2024 tem expansão de 2,6 GW na matriz elétrica. Disponível em: <https://www.gov.br/aneel/pt-br/assuntos/noticias/2024/primeiro-trimestre-de-2024-tem-expansao-de-2-6-gw-na-matriz-eletrica>. Acesso em 15 de maio 2024.

Agência Nacional de Energia Elétrica (ANEEL). (2023). Sistema de Informações de Geração - SIGA. Disponível em: <https://app.powerbi.com/view?r=eyJrIjoiNjc4OGYyYjQtYWM2ZC00YjllLWJlYmEtYzdkNTQ1MTc1NjM2IiwidCI6IjQwZDZmOWI4LWVjYTctNDZhMi05MmQ0LWVhNGU5YzAxNzBlMSIsImMiOjR9>. Acesso em 22 dez. 2023.

BENTÉJAC, C.; CSÖRGŐ, A.; MARTÍNEZ-MUÑOZ, G.. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, [S.l.], v. 53, n. 1, p. 1-35, 2020. Springer Nature B.V. Disponível em: <https://doi.org/10.1007/s10462-020-09896-5>. Acesso em: 23 de fevereiro de 2025.

BRUCE P.; BRUCE A. (2019). Estatística Prática para Cientistas de Dados – 50 Conceitos Essenciais. 1ª Edição. Starlin Alta Editora e Consultoria Eireli.

DELGADO FILHO, A. J. F. (2020). Análise de métodos de regressão para previsão de demanda de curto prazo. Dissertação de Mestrado, Universidade Federal de Pernambuco - CIn . Ciência da Computação, Recife.

DO AMARANTE, O. A. C., BROWER, M., ZACK, J., & LEITE DE SÁ, A. (2001). Atlas do Potencial Eólico Brasileiro. Camargo-Schubert, TrueWind Solutions, CEPEL.

ERIŞEN, B. Wind Turbine Scada Dataset. 2018. Disponível em: <https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-

dataset?resource=download&select=T1.csv>. Acesso em: 10 de outubro de 2023.

ESCOVEDO, T.; KOSHIYAMA, A. (2023). Introdução a Data Science: Algoritmos de Machine Learning e Métodos de Análise. Casa do Código, Alura, São Paulo.

FORTMANN-ROE, S. Understanding the Bias-Variance Tradeoff. 2012. Disponível em: <https://scott.fortmann-roe.com/docs/BiasVariance.html>. Acesso em: 2 de fevereiro de 2025.

Global Wind Energy Council (GWEC). (2023). Global Wind Report 2023. Disponível em: <https://gwec.net/wp-content/uploads/2023/03/GWR-2023_interactive.pdf>. Acesso em 15 de maio 2024.

Global Wind Energy Council (GWEC). (2022). Global Wind Report 2022.

JAMES, G., WITTEN, D., HASTIE, T., & TIBSHIRANI, R. (2023). An Introduction to Statistical Learning: With Applications in R. 2nd Edition. Springer.

KLOSTERMAN S. (2020). Projetos de Ciência de Dados usando Python: Abordagem de estudo de caso para a criação de projetos de ciência de dados bem-sucedidos usando Python, pandas e scikit-learn. Novatec Editora, São Paulo.

MALIK, S.; MUHAMMAD, K.; WAHEED, Y.. Artificial intelligence and industrial applications – A revolution in modern industries. Ain Shams Engineering Journal, 2024. Disponível em: <https://doi.org/10.1016/j.asej.2024.102886>. Acesso em: 23 de fevereiro de 2025.

MOREIRA, M. S.; RODRIGUES, M. P.; FERREIRA, C. F.; NIENOV, O. H. Bioestatística quantitativa aplicada. Porto Alegre: UFRGS, 2020. Cap. 10, p. 197-216. Disponível em: <http://hdl.handle.net/10183/213489>. Acesso em: 02 de fevereiro de 2025.

NAZIR, M. S. (2019). Environmental impact and pollution-related challenges of renewable wind energy paradigm—A review. Sci. Total Environ., 683, 436–444.

NAZIR, M. S., ALTURISE, F., ALSHMRANY, S., NAZIR, H., BILAL, M., ABDALLA, A. N., SANJEEVIKUMAR, P., & ALI, Z. M. (2020). Wind generation forecasting methods and proliferation of artificial neural network: a review of five years research trend.

Sustainability, 12(9), 3778.

NAZIR, G. C. (2007). Utilização de redes neurais para previsão de ventos no horizonte de 24 horas. Dissertação de mestrado, Rio de Janeiro, RJ, Brasil.

NOVAIS, P., & FREITAS, P. M. (2018). Inteligência Artificial e regulação de algoritmos. Disponível em: <http://www.sectordialogues.org/projetos/inteligencia artificial-e-regulacao-de-algoritmos>. Acesso em 22 dez. 2023.

PINTO, M. O. (2014). Fundamentos de Energia Eólica. Rio de Janeiro: LTC.

PREETHAA, K. R. Sri; MUTHURAMALINGAM, A.; NATARAJAN, Y.; WADHWA, G.; ALI, A. A. Y. A Comprehensive Review on Machine Learning Techniques for Forecasting Wind Flow Pattern. Sustainability, v. 15, n. 17, p. 12914, 2023. Disponível em: <https://doi.org/10.3390/su151712914>. Acesso em: 2 de fevereiro de 2025.

RATHAN, K.; SAI, S.; MANIKANTA, T. Crypto-Currency price prediction using Decision Tree e Regression techniques. 3rd International Conference on Trends in Electronics e Informatics, 2019.

RODRIGUES, G. C. (2007). Utilização de redes neurais para previsão de ventos no horizonte de 24 horas. Dissertação de mestrado, Rio de Janeiro, RJ, Brasil.

SHARDA, R., DELEN, D., & TURBAN, E. (2020). Analytics, Data Science, & Artificial Intelligence. Hoboken, NJ: Pearson Education, Inc.

SILVA, R. A. (2019). Determinação de Recursos Minerais com Utilização de Otimização e Algoritmos de Aprendizado de Máquina. Tese de Doutorado, Universidade Federal de Pernambuco, Recife.

SINGH, U., & RIZWAN, M. (2022). SCADA system dataset exploration and machine learning based forecast for wind turbines. Results in Engineering.

VANDERPLAS, J. Python Data Science Handbook. Sebastopol: O’Reilly Media, 2023.

XGBOOST DEVELOPERS. XGBoost Documentation. Disponível em: <https://xgboost.readthedocs.io/en/stable/>. Acesso em: 3 de março de 2025.