



DURVAL CAMARGO RANGEL DE SOUZA BRITO

**USO DE APRENDIZAGEM DE MÁQUINA E OTIMIZAÇÃO
MULTIOBJETIVO PARA ESTIMAR ENERGIA POTENCIAL
DE AEROGERADORES**

LAVRAS – MG

2026

DURVAL CAMARGO RANGEL DE SOUZA BRITO

**USO DE APRENDIZAGEM DE MÁQUINA E OTIMIZAÇÃO MULTIOBJETIVO
PARA ESTIMAR ENERGIA POTENCIAL DE AEROGERADORES**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

Prof. DSc. Bruno Henrique Groenner Barbosa
Orientador

**LAVRAS – MG
2026**

**Ficha Catalográfica elaborada pelo Sistema de Geração
de Ficha Catalográfica da Biblioteca Universitária da UFLA, com
dados informados pelo(a) próprio(a) autor(a).**

Brito, Durval Camargo Rangel de Souza.

Uso de aprendizagem de máquina e otimização multiobjetivo para estimar energia
potencial de aerogeradores / Durval Camargo Rangel de Souza Brito. - 2026.
106 p. : il.

Orientador: Bruno Henrique Groenner Barbosa

Dissertação (Mestrado Acadêmico) - Universidade Federal de Lavras, 2026.
Bibliografia.

1. Energia Eólica. 2. Curva de Potência. 3. Aprendizagem de Máquina. 4.
Otimização Multiobjetivo. 5. Algoritmo Genético. I. Barbosa, Bruno Henrique
Groenner. II. Universidade Federal de Lavras. III. Título.

DURVAL CAMARGO RANGEL DE SOUZA BRITO

**USO DE APRENDIZAGEM DE MÁQUINA E OTIMIZAÇÃO MULTIOBJETIVO
PARA ESTIMAR ENERGIA POTENCIAL DE AEROGERADORES**

**MACHINE LEARNING AND MULTIOBJECTIVE OPTIMIZATION FOR WIND
TURBINE POTENTIAL ENERGY ESTIMATION**

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Engenharia de Sistemas e Automação, para a obtenção do título de Mestre.

APROVADA em 13 de julho de 2026.

Prof. DSc. Bruno Henrique Groenner Barbosa	UFLA
Prof. DSc. Wilian Soares Lacerda	UFLA
Profa. DSc. Silvia Costa Ferreira	UFLA
Dr. Caio dos Santos	Serena

Documento assinado digitalmente
 **BRUNO HENRIQUE GROENNER BARBOSA**
Data: 16/10/2026 17:20:28-0300
Verifique em <https://validar.iti.gov.br>

Prof. DSc. Bruno Henrique Groenner Barbosa
Orientador

**LAVRAS – MG
2026**

Dedico este trabalho à visão de um mundo sustentável.

AGRADECIMENTOS

Agradeço à minha esposa, minha mãe, meus sogros e meu padrinho, que sempre me apoiaram em todos os momentos. Agradeço à minha família. Agradeço à todos os envolvidos no projeto. Agradeço à Serena Energia.

"O saber não se transfere, se constrói."

(Paulo Freire)

RESUMO

A transição energética global tem impulsionado o crescimento da energia eólica, especialmente no Brasil, onde a fonte representa 12,8% da matriz elétrica. A operação eficiente de parques eólicos requer estratégias sofisticadas para maximizar a extração de energia, considerando a intermitência do recurso eólico e os desafios operacionais inerentes a ambientes de campo. Este trabalho propõe uma metodologia baseada em inteligência artificial para estimativa da energia potencial de aerogeradores, utilizando exclusivamente dados meteorológicos de torres de medição eólica (TME) e dados operacionais de sistemas SCADA. A abordagem integra filtragem multimodal de dados e otimização multiobjetivo via NSGA-II para seleção simultânea de fontes de dados, variáveis meteorológicas e hiperparâmetros de modelos de aprendizado de máquina, explorando o compromisso entre acurácia preditiva e complexidade operacional. A metodologia foi validada em dois aerogeradores GE de 2,5 MW localizados em Xique-Xique, Bahia, com condições operacionais distintas: o W301, em operação com vento relativamente livre, e o W310, sujeito a efeito de esteira pronunciado. Os dados compreendem 31 variáveis SCADA e medições meteorológicas de 8 TME próximas, totalizando 51.564 amostras entre janeiro de 2024 e janeiro de 2025. A filtragem de dados resultou em 65% de amostras aproveitáveis para o W301 e 63,5% para o W310, removendo períodos de paradas, limitações e operação degradada. O algoritmo NSGA-II revelou uma hierarquia geoespacial consistente na seleção de fontes, convergindo autonomamente para configurações com três TME que cobrem as condições de entrada, evolução intermediária e condições locais do escoamento. Para o W301, os modelos otimizados alcançaram R^2 de 0,984 (XGBoost) e 0,982 (MLP), frente a 0,973 do modelo de referência. Para o W310, os ganhos foram ainda mais expressivos, com R^2 de 0,973 e 0,972 contra 0,933 do modelo base, demonstrando que a metodologia agrega valor proporcionalmente maior em condições operacionais complexas. A aplicação dos modelos ao histórico completo de dados permitiu estimar perdas energéticas entre 4,7% e 6,5% do potencial disponível, com implicações diretas para gestão de manutenção e negociações contratuais. A abordagem proposta supera limitações dos métodos tradicionais baseados na norma IEC 61400-12-1, oferecendo estimativas do potencial físico real do aerogerador sem incorporar restrições operacionais como comportamento normal.

Palavras-chave: Energia Eólica; Aprendizagem de Máquina; Curva de Potência; Torres de Medição Eólica; Otimização Multiobjetivo; Algoritmo Genético.

ABSTRACT

The global energy transition has driven the growth of wind energy, particularly in Brazil, where it accounts for 12.8% of the national electricity mix. The efficient operation of wind farms requires sophisticated strategies to maximize energy extraction, given the inherent intermittency of the wind resource and the operational challenges of field environments. This work proposes an artificial-intelligence-based methodology for estimating the potential energy of wind turbines, using exclusively meteorological data from met masts and operational data from SCADA systems. The approach integrates multimodal data filtering and multi-objective optimization via NSGA-II for the simultaneous selection of data sources, meteorological variables, and machine learning model hyperparameters, exploring the trade-off between predictive accuracy and operational complexity. The methodology was validated on two GE 2.5 MW wind turbines located in Xique-Xique, Bahia, under distinct operational conditions: W301, operating under relatively free-stream wind, and W310, subject to pronounced wake effects. The dataset comprises 31 SCADA variables and meteorological measurements from 8 nearby met masts, totaling 51,564 samples collected between January 2024 and January 2025. Data filtering yielded 65% of usable samples for W301 and 63.5% for W310, removing periods of shutdowns, curtailments, and degraded operation. The NSGA-II algorithm revealed a consistent geospatial hierarchy in source selection, autonomously converging to three-met-mast configurations covering inflow conditions, intermediate flow evolution, and local conditions near the target turbine. For W301, optimized models achieved R^2 of 0.984 (XGBoost) and 0.982 (MLP), compared to 0.973 for the reference model. For W310, gains were even more expressive, with R^2 of 0.973 and 0.972 against 0.933 for the baseline, demonstrating that the methodology delivers proportionally greater value under complex operational conditions. Applying the final models to the complete historical dataset enabled estimation of energy losses between 4.7% and 6.5% of the available potential, with direct implications for maintenance planning and contractual negotiations. The proposed approach overcomes limitations of traditional methods based on IEC 61400-12-1, providing estimates of the true physical potential of the turbine without embedding operational restrictions as normal behavior.

Keywords: Wind Energy; Machine Learning; Power Curve; Met Mast; Multi-Objective Optimization; Genetic Algorithm

INDICADORES DE IMPACTO

Esta dissertação propõe uma metodologia baseada em inteligência artificial para estimar a energia potencial de aerogeradores a partir de dados meteorológicos de torres de medição eólica (TME) e de dados operacionais de sistemas SCADA, validada em um parque eólico real no semiárido baiano (Xique-Xique, Bahia), em parceria com a Serena Energia.

Impactos tecnológicos e econômicos: a metodologia quantifica perdas energéticas associadas a indisponibilidades, limitações e degradação operacional (estimadas entre 4,7% e 6,5% do potencial nos casos analisados), oferecendo base técnica para o planejamento de manutenção e para negociações e contestações de contratos de fornecimento de energia. Por dispensar a instrumentação dedicada exigida pela norma IEC 61400-12-1 e selecionar automaticamente o conjunto mínimo de fontes de dados, contribui para a redução de custos de instrumentação e de operação na gestão de ativos de geração renovável.

Impactos ambientais e sociais: ao maximizar o aproveitamento do recurso eólico e reduzir perdas evitáveis, a pesquisa contribui para a eficiência da geração de energia limpa e para a transição energética brasileira.

Alinhamento com os Objetivos de Desenvolvimento Sustentável (ODS) da ONU: os resultados alinham-se principalmente ao ODS 7 (Energia limpa e acessível), ao ODS 9 (Indústria, inovação e infraestrutura) e ao ODS 13 (Ação contra a mudança global do clima). Quanto às áreas temáticas da Política Nacional de Extensão, o trabalho enquadra-se em Tecnologia e Produção e em Meio Ambiente. Os impactos descritos são, neste momento, predominantemente de caráter potencial.

IMPACT INDICATORS

This dissertation proposes an artificial-intelligence-based methodology to estimate the potential energy of wind turbines from meteorological data of met masts (TME) and operational data of SCADA systems, validated on a real wind farm in the Brazilian semi-arid region (Xique-Xique, Bahia), in partnership with Serena Energia.

Technological and economic impacts: the methodology quantifies energy losses associated with unavailability, curtailment and degraded operation (estimated between 4.7% and 6.5% of the potential in the analyzed cases), providing a technical basis for maintenance planning and for the negotiation and contestation of energy supply contracts. By dispensing with the dedicated instrumentation required by the IEC 61400-12-1 standard and automatically selecting the minimal set of data sources, it contributes to reducing instrumentation and operation costs in renewable generation asset management.

Environmental and social impacts: by maximizing the use of the wind resource and reducing avoidable losses, the research contributes to the efficiency of clean energy generation and to the Brazilian energy transition.

Alignment with the UN Sustainable Development Goals (SDGs): the results align mainly with SDG 7 (Affordable and clean energy), SDG 9 (Industry, innovation and infrastructure) and SDG 13 (Climate action). Regarding the thematic areas of the National Extension Policy, the work falls under Technology and Production and under Environment. The described impacts are, at this stage, predominantly potential in nature.

LISTA DE FIGURAS

Figura 1.1 – Matriz Elétrica Brasileira	18
Figura 2.1 – Esquema de uma TME típica	24
Figura 2.2 – Esquema de um Aerogerador	25
Figura 2.3 – Curva de Potência Típica de Aerogeradores	27
Figura 2.4 – Estrutura de Perceptron Multicamadas	34
Figura 2.5 – Estrutura de Árvore de Decisão	35
Figura 2.6 – Cruzamento uniforme	37
Figura 2.7 – Mutação em bit a bit	37
Figura 2.8 – Fluxograma NSGA-II	38
Figura 3.1 – Fluxograma da metodologia implementada.	41
Figura 3.2 – Disposição de WTGs e TMEs na região.	42
Figura 3.3 – Disposição de WTGs e TMEs próximas do W301.	43
Figura 3.4 – Disposição de WTGs e TMEs próximas do W310.	43
Figura 4.1 – Vento característico da TME T103.	67
Figura 4.2 – Correlação entre variáveis da TME T103.	68
Figura 4.3 – Características operacionais do W301.	69
Figura 4.4 – Visualização do Filtro <i>05-BLADE</i> do W301.	70
Figura 4.5 – Visualização dos Filtros <i>06-DIFF_WTG_SMF</i> e <i>07-2DESVPAD</i> do W301.	70
Figura 4.6 – Visualização do Filtro <i>08-GENERATOR</i> do W301.	71
Figura 4.7 – Dados filtrados do W301.	71
Figura 4.8 – Modelo Base do W301.	72
Figura 4.9 – Fronteiras de Pareto do W301.	74
Figura 4.10 – Evolução do treinamento do W301.	76
Figura 4.11 – Fronteira de Pareto e distribuição de fontes e características do W301.	79
Figura 4.12 – Mapa com direções predominantes das TMEs do W301.	80
Figura 4.13 – Distribuição das predições para os modelos escolhidos do W301.	83
Figura 4.14 – Predição de Fator de Capacidade durante um dia de operação com falhas do W301.	85
Figura 4.15 – Mapa com direções predominantes das TMEs do W310.	87

Figura 4.16 – Dados filtrados do W310.	87
Figura 4.17 – Modelo Base do W310.	88
Figura 4.18 – Evolução do treinamento do W310.	89
Figura 4.19 – Fronteira de Pareto e distribuição de fontes e características do W310. . . .	91
Figura 4.20 – Distribuição das predições para os modelos escolhidos do W310.	92
Figura 4.21 – Predição de Fator de Capacidade durante um dia de operação com falhas do W310.	94

LISTA DE TABELAS

Tabela 3.1 – Variáveis SCADA	45
Tabela 3.2 – Descrição das Variáveis da TME	46
Tabela 3.3 – Informações das Variáveis da TME	46
Tabela 3.4 – Exemplo de Cromossomo XGBoost	56
Tabela 3.5 – Exemplo de Cromossomo MLP	58
Tabela 4.1 – Métricas de validação e características dos modelos selecionados pelo critério do ponto utópico nas cinco execuções independentes do NSGA-II.	73
Tabela 4.2 – Contagem de execuções por variável selecionada	75
Tabela 4.3 – Comparação de Acurácia entre modelos do W301	82
Tabela 4.4 – Variáveis Escolhidas dos Modelos do W301	83
Tabela 4.5 – Energia gerada relativa ao potencial do W301.	86
Tabela 4.6 – Comparação de Acurácia entre modelos do W310	92
Tabela 4.7 – Variáveis Escolhidas dos Modelos do W310	93
Tabela 4.8 – Energia gerada relativa ao potencial do W310.	94

LISTA DE SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
ANN	Rede Neural Artificial
BHM	Modelo Hierárquico Bayesiano
CFD	<i>Computational Fluid Dynamics</i>
DT	Árvore de Decisão
ELM	Máquina de Aprendizado Extremo
EMD	Earth-Mover Distance
EPE	Empresa de Pesquisa Energética
FC	Fator de Capacidade
FIS	Sistemas de Inferência Fuzzy
GP	Processo Gaussiano
HV	<i>Hipervolume</i>
IA	Inteligência Artificial
IEC 61400-12-1	Norma 61400-12-1 da Comissão Eletrotécnica Internacional
IGD	<i>Distância Geracional Invertida</i>
MAE	Erro Médio Absoluto
MEASNET	Measuring Network of Wind Energy Institutes
ML	Aprendizagem de Máquina
MLP	Perceptron Multicamadas
MMGD	Micro e Mini Geração Distribuída
NSGA-II	Algoritmo Genético de Ordenação Não Dominada II
OCSVM	One-Class SVM
ONS	Operador Nacional do Sistema Elétrico
RBFNN	Rede Neural de Função de Base Radial
ReLU	<i>Rectified Linear Unit</i>
RF	Floresta Aleatória
RMSE	Raiz do Erro Quadrático Médio
RT	Árvores de Regressão
SCADA	Sistemas de Supervisão e Aquisição de Dados
SOM	Mapas Auto-Organizáveis

SVM	Máquinas de Vetores de Suporte
TME	Torre de Medição Eólica
WTG	<i>Wind Turbine Generator</i> , Aerogerador, Turbina Eólica
XGBoost	<i>Extreme Gradient Boosting</i>

LISTA DE SÍMBOLOS

T_i Intensidade de turbulência

μ Média

σ Desvio Padrão

SUMÁRIO

1	INTRODUÇÃO	18
1.1	Objetivos	21
1.1.1	Objetivos Específicos	21
2	REVISÃO BIBLIOGRÁFICA	23
2.1	Parques Eólicos	23
2.1.1	Torres de Medição Eólica	23
2.1.2	Aerogeradores	25
2.2	Modelagem de Curvas de Potência em Aerogeradores	26
2.2.1	Efeitos de Interação em Parques Eólicos	28
2.2.1.1	Métodos Paramétricos vs. Não Paramétricos	29
2.2.2	Aprendizado de Máquina aplicado em Eólicas	30
2.2.3	Qualidade de dados	32
2.3	Algoritmos de Otimização e Aprendizado de Máquina	33
2.3.1	Perceptron Multicamadas (MLP)	33
2.3.2	<i>Extreme Gradient Boosting</i> (XGBoost)	34
2.3.3	Algoritmo Genético Multiobjetivo (NSGA-II)	35
2.4	Lacunas na literatura	39
3	MATERIAIS E MÉTODOS	41
3.1	Análise Exploratória e Compreensão dos Dados	41
3.1.1	Caracterização dos Dados	42
3.1.2	Análise de Correlações e Seleção Preliminar de Variáveis	46
3.2	Pré-processamento dos Dados	47
3.3	Filtragem de Dados	48
3.3.1	Normalização dos Dados	50
3.4	Desenvolvimento de Modelos	50
3.4.1	Modelo de Referência (Base)	51
3.4.2	Algoritmo de Otimização NSGA-II	52
3.4.2.1	XGBoost: Espaço de Busca	55
3.4.2.2	MLP: Espaço de Busca	56
3.5	Métricas de Avaliação e Validação dos Modelos	58
3.5.1	Validação da Otimização Multiobjetivo	59

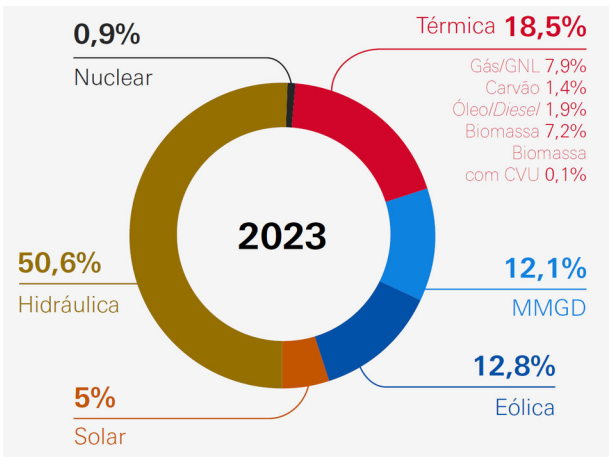
3.5.2	Métricas de Desempenho dos Modelos	60
3.6	Avaliação Final e Cálculo de perdas	63
4	RESULTADOS E DISCUSSÃO	66
4.1	Resultados para o Aerogerador W301	66
4.1.1	Caracterização dos Dados	66
4.1.1.1	Avaliação de Variáveis Disponíveis	67
4.1.1.2	Tratamento dos Dados	69
4.1.2	Modelo Base	72
4.1.3	Modelagem NSGA-II	72
4.1.3.1	Validação por meio de Múltiplas Execuções	73
4.1.3.2	Evolução do Processo de Otimização	75
4.1.3.3	Caracterização da Fronteira de Pareto	77
4.1.3.4	Modelos Seleccionados e Desempenho Preditivo	81
4.1.3.5	Estimativa de Perdas Energéticas	84
4.2	Resultados para o Aerogerador W310	86
4.3	Discussão Geral	94
5	CONCLUSÃO	98
	REFERÊNCIAS	101

1 INTRODUÇÃO

A transição energética global em direção a fontes renováveis tem se intensificado nas últimas décadas, impulsionada pela necessidade de reduzir as emissões de gases de efeito estufa e promover um desenvolvimento sustentável. No contexto brasileiro, esta transição assume características particulares devido à matriz energética já predominantemente renovável do país (ONS, 2023).

O Brasil possui uma das matrizes elétricas mais limpas do mundo, com as fontes renováveis representando aproximadamente 87% da capacidade instalada em 2023. Conforme dados do Operador Nacional do Sistema Elétrico (ONS), a energia hidráulica mantém sua posição dominante com 50,6% da matriz, seguida pelas térmicas (18,5%) das quais Biomassa (7,3%) é considerado fonte renovável, eólica (12,8%), mini e micro geração distribuída (MMGD) (12,1%), solar (5%) e nuclear (0,9%) como visto na Figura 1.1. Esta configuração coloca o Brasil em posição privilegiada no cenário energético mundial, onde a média global de participação renovável ainda é significativamente menor.

Figura 1.1 – Matriz Elétrica Brasileira



Fonte: (ONS, 2023)

Dentro deste panorama, aerogeradores têm se tornado elementos cada vez mais presentes na matriz energética brasileira. Esta expansão é estratégica não apenas para diversificar a matriz energética, mas também para reduzir a dependência da geração hidrelétrica, especialmente considerando os impactos climáticos que afetam os níveis dos reservatórios (Agência Nacional de Águas e Saneamento Básico, 2024). Ao mesmo tempo, o recurso eólico traz consigo um desafio central: operar uma fonte cuja geração é, por natureza, intermitente. A vari-

abilidade temporal, espacial e fenomenológica do vento cria complexidades operacionais que demandam estratégias sofisticadas de gestão e otimização.

A gestão eficiente de um parque eólico busca maximizar a extração de energia evitando paradas durante períodos de alta produção potencial. Isso implica na necessidade de garantir que, quando o recurso eólico está disponível, o aerogerador (WTG) esteja operacional e capaz de extrair energia em sua capacidade ótima. A intermitência do recurso natural exige estratégias de otimização para maximizar o desempenho e garantir a disponibilidade das turbinas (Rodrigues; Aquino; Thomaz, 2014).

Para alcançar essa otimização operacional e garantir a máxima extração de energia, torna-se fundamental o entendimento preciso do desempenho dos aerogeradores por meio de sua curva de potência, que representa a relação fundamental entre a velocidade do vento incidente e a potência elétrica gerada pelo equipamento. Na prática, as curvas de potência teóricas não representam adequadamente o desempenho operacional real dos aerogeradores em campo. Esta discrepância entre o desempenho teórico e prático motivou o desenvolvimento de curvas de potência específicas para aerogeradores *in loco*, onde duas abordagens principais são exploradas: modelos paramétricos, baseados em condições e modelos físicos; e modelos não paramétricos, baseados em aprendizado de máquina que utilizam primordialmente dados operacionais reais para suas modelagens.

Dentro dessa perspectiva, os modelos não paramétricos baseados em aprendizado de máquina oferecem uma aplicação promissora de inteligência artificial (IA) para aprimorar a eficiência energética e a operação dos parques eólicos. Essas técnicas de IA são particularmente adequadas para modelar as complexidades e interações não lineares presentes na geração e aproveitamento do recurso eólico, uma vez que conseguem capturar padrões complexos nos dados operacionais reais sem depender de premissas físicas pré-estabelecidas. Dessa forma, essas técnicas permitem capturar com maior fidelidade a relação entre condições meteorológicas e geração de energia, viabilizando a estimativa do potencial energético dos aerogeradores mesmo em cenários operacionais complexos. Diversos modelos têm sido aplicados com sucesso, incluindo Árvores de Decisão, Florestas Aleatórias, *Extreme Gradient Boosting* (XGBoost), Máquinas de Vetores de Suporte (SVM), Redes Neurais Artificiais (ANN) e métodos mais avançados como *Gaussian Process* e abordagens *ensemble*, frequentemente superando métodos tradicionais em precisão e capacidade de generalização (Wang *et al.*, 2019).

No entanto, a eficácia destes modelos de aprendizado de máquina não depende apenas da escolha do algoritmo, mas também de decisões metodológicas críticas que precedem a modelagem: quais variáveis incluir e como configurar adequadamente os hiperparâmetros do modelo. Tradicionalmente, essas decisões são tomadas de forma sequencial e isolada, ou seja, primeiro selecionam-se as variáveis baseando-se em conhecimento especialista, depois otimizam-se os hiperparâmetros do modelo escolhido. Esta abordagem sequencial ignora as interdependências entre essas escolhas e frequentemente resulta em soluções subótimas.

A literatura apresenta uma lacuna significativa na consideração simultânea de múltiplos objetivos conflitantes inerentes ao problema. Enquanto a maioria dos estudos busca maximizar exclusivamente a acurácia preditiva dos modelos, aplicações práticas em parques eólicos demandam o equilíbrio entre precisão, complexidade computacional, custo de instrumentação e simplicidade operacional. Um modelo altamente preciso que requer dezenas de sensores distribuídos por múltiplas torres pode ser economicamente inviável ou operacionalmente impraticável, enquanto um modelo excessivamente simplificado pode não capturar adequadamente a variabilidade das condições reais. Métodos de otimização multiobjetivo, que permitem explorar explicitamente estes *trade-offs* e identificar soluções de compromisso na fronteira de Pareto, permanecem praticamente inexplorados para modelagem de desempenho de aerogeradores.

Neste contexto, o conceito de sensores virtuais representa uma abordagem particularmente relevante para o monitoramento e estimativa de desempenho em parques eólicos. Sensores virtuais possuem a capacidade de estimar variáveis de interesse baseando-se em medições indiretas de outros parâmetros disponíveis, oferecendo alternativas econômicas e práticas à instrumentação física extensiva (Dimitrov; Göçmen, 2022; Li; Shen, 2022). No âmbito específico deste trabalho, o foco está no desenvolvimento de um sensor virtual para estimativa direta da energia potencial do aerogerador, utilizando dados meteorológicos regionais provenientes de torres de medição eólica (TME) e dados operacionais do próprio aerogerador coletados via sistemas SCADA.

Esta abordagem de estimativa direta é mais prática que a criação tradicional de curvas de potência como etapa intermediária, pois elimina processos adicionais e foca diretamente no objetivo final: quantificar a capacidade de geração do aerogerador sob diferentes condições operacionais. A metodologia proposta considera que aerogeradores em um parque exercem influência mútua através de efeitos de bloqueio e esteira (Nygaard *et al.*, 2020), fenômenos que impactam significativamente o desempenho e devem ser adequadamente incorporados na mo-

delagem. Estudos demonstram que dados da região próxima ao aerogerador (tipicamente entre 25 e 50 diâmetros de rotor) proporcionam melhores resultados de modelagem do que utilizar apenas dados da própria turbina ou do parque completo (Lyons; Göçmen, 2021), justificando a estratégia de utilizar múltiplas fontes de dados distribuídas espacialmente.

Portanto, este trabalho propõe uma metodologia integrada que combina otimização evolutiva multiobjetivo com técnicas de aprendizado de máquina para realizar simultaneamente a seleção de características meteorológicas, a escolha de fontes de dados dentre as torres disponíveis, e a configuração de hiperparâmetros do modelo preditivo. Através do algoritmo NSGA-II, busca-se explorar sistematicamente o espaço de soluções considerando explicitamente o compromisso entre acurácia preditiva e complexidade do modelo, permitindo identificar configurações que equilibrem desempenho e viabilidade prática para aplicação em ambiente operacional de parques eólicos.

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver uma metodologia baseada em inteligência artificial para estimativa da energia potencial de aerogeradores utilizando dados operacionais de sistemas supervisórios (SCADA) e dados meteorológicos de TMEs distribuídas na região. O intuito é gerar um sensor virtual capaz de estimar o fator de capacidade do aerogerador baseando-se em condições meteorológicas regionais e características operacionais específicas da máquina, permitindo quantificar perdas energéticas durante períodos de indisponibilidade, limitações operacionais ou baixo desempenho, com equilíbrio otimizado entre acurácia preditiva e complexidade do modelo.

1.1.1 Objetivos Específicos

Os objetivos específicos são:

1. realizar uma análise exploratória abrangente dos dados de sistemas SCADA de aerogeradores e torres de medição eólica, identificando relações entre variáveis meteorológicas, operacionais e de geração de energia;

2. desenvolver e implementar metodologia integrada de filtragem de dados para identificar e remover períodos de paradas, limitações e operação degradada, garantindo que apenas amostras representativas da operação ótima sejam utilizadas na modelagem;
3. implementar algoritmos de aprendizagem de máquina para estimativa de energia potencial de aerogeradores, utilizando dados filtrados de sistemas SCADA e torres de medição eólica;
4. desenvolver e aplicar metodologia de otimização evolutiva multiobjetivo para seleção conjunta de características, fontes de dados e hiperparâmetros de modelo;
5. analisar o compromisso entre acurácia preditiva, complexidade do modelo e simplicidade operacional por meio da fronteira de Pareto, estabelecendo critérios de seleção de modelos adequados para diferentes aplicações práticas em parques eólicos;
6. demonstrar a aplicação prática da metodologia por meio da quantificação de perdas energéticas em períodos de indisponibilidade e limitações operacionais, demonstrando a capacidade do modelo de estimar o potencial energético real do aerogerador.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo apresenta a revisão bibliográfica que fundamenta os conceitos, métodos e desafios relacionados à caracterização do recurso eólico e à modelagem de curvas de potência em parques eólicos. Inicialmente, são discutidas definições técnicas e regulatórias do contexto brasileiro, bem como os principais elementos de infraestrutura de medição e geração (torres de medição e aerogeradores). Em seguida, são revisadas abordagens tradicionais e recentes para modelagem de curvas de potência, incluindo limitações do procedimento normativo, efeitos de interação em parques, métodos paramétricos e não paramétricos e a influência da qualidade dos dados no desempenho de modelos baseados em aprendizado de máquina. Por fim, são destacadas lacunas na literatura que motivam e delimitam a contribuição deste trabalho.

2.1 Parques Eólicos

A definição técnica e regulatória de parque eólico no contexto brasileiro é estabelecida pela Portaria MME 21/2008 (Ministério de Minas e Energia, 2008; Ministério de Minas e Energia, 2011), conforme detalhado na Nota Técnica:

o conjunto de aerogeradores interligados eletricamente, situados nas áreas circulares com raio de até dez quilômetros em torno das torres de medição anemométrica, no caso de terrenos de superfície plana com rugosidade homogênea, e com raio de até seis quilômetros, no caso de terrenos complexos, identificados os aerogeradores e as torres de medição por suas coordenadas UTM (Universal Transversa de Mercator), sujeita à validação da EPE a definição do raio quanto à adequação com a topografia. ((ANEEL), 2013, 049/2013-SCG § 1º do art. 6º-A).

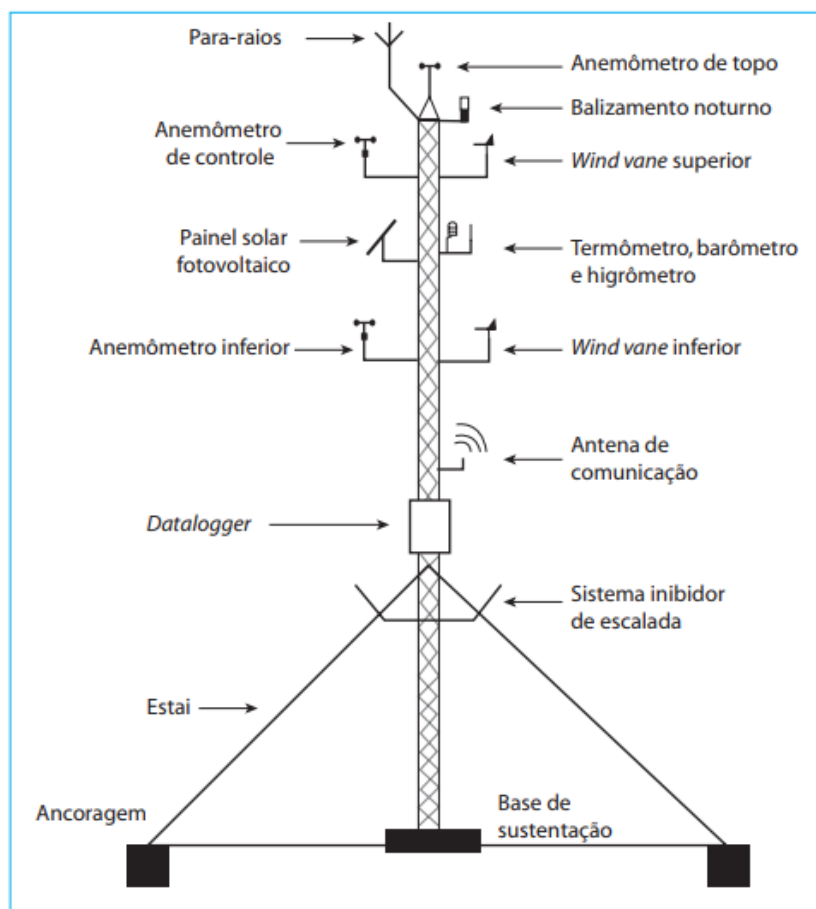
Esta definição é fundamental pois estabelece não apenas a configuração física dos parques, mas também a importância das TME como elementos centrais na caracterização do recurso eólico regional. A definição está alinhada com normas internacionais, especificamente a MEASNET (Rede Internacional de Institutos de Medição de Energia Eólica).

2.1.1 Torres de Medição Eólica

As TME costumeiramente são instaladas como na Figura 2.1 e desempenham papel duplo e crítico tanto na fase pré-operacional quanto operacional dos parques eólicos. Durante a

fase de desenvolvimento, as TME são responsáveis pela caracterização inicial do recurso eólico, fornecendo dados fundamentais para a determinação da viabilidade técnica e econômica do empreendimento, o posicionamento ótimo das turbinas conforme o regime de ventos local, a estimativa de efeitos de esteira e interferências topográficas, além da comprovação da capacidade de geração necessária para a participação em leilões de energia.

Figura 2.1 – Esquema de uma TME típica



Fonte: Empresa de Pesquisa Energética (2014)

Já na fase operacional, as TME continuam essenciais ao permitirem o acompanhamento das condições meteorológicas em tempo real e oferecerem suporte aos modelos de previsão de recurso de curto e médio prazo. Adicionalmente, sua atuação é fundamental para a identificação de degradação de desempenho ou falhas operacionais, garantindo, simultaneamente, o atendimento às exigências de monitoramento impostas pela ANEEL e pela EPE.

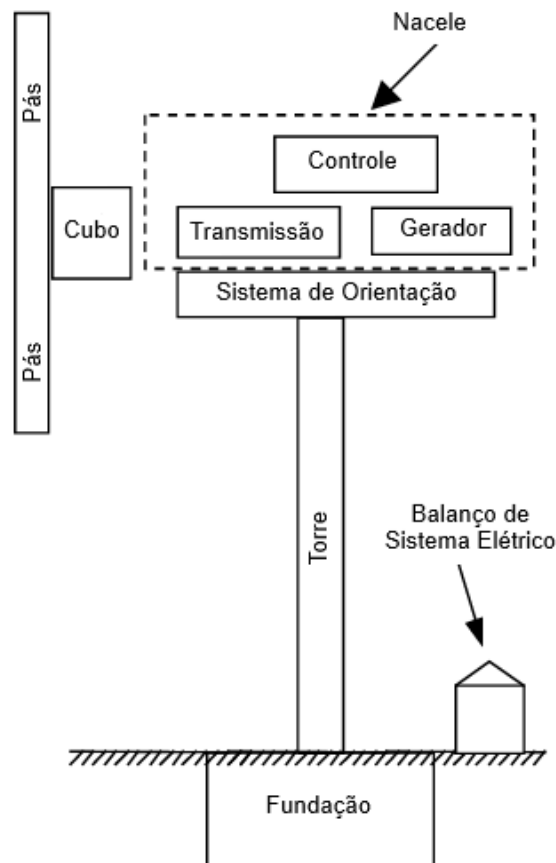
A necessidade de múltiplas TME distribuídas estrategicamente em grandes aglomerados de parques eólicos decorre da variabilidade espacial do recurso vento, especialmente em terrenos complexos como os encontrados no Nordeste brasileiro. Esta distribuição permite capturar

adequadamente os gradientes de vento e efeitos topográficos que influenciam significativamente o desempenho dos aerogeradores.

2.1.2 Aerogeradores

Os aerogeradores (WTG) modernos são sistemas complexos compostos por múltiplos subsistemas interconectados que trabalham em conjunto para converter a energia cinética do vento em energia elétrica. A compreensão da arquitetura destes sistemas é fundamental para o desenvolvimento de modelos precisos de estimativa de energia potencial.

Figura 2.2 – Esquema de um Aerogerador



Fonte: Adaptado de Manwell, McGowan & Rogers (2010)

A Figura 2.2 ilustra a configuração típica de um aerogerador de eixo horizontal, que representa a tecnologia predominante em aplicações comerciais. Nela estão os principais componentes:

- pás: elementos aerodinâmicos responsáveis pela captação da energia do vento;

- cubo: estrutura central que conecta o rotor ao eixo principal;
- rotor: parte giratória frontal, composta pelas pás e pelo cubo;
- transmissão: sistema multiplicador que conecta e eleva a rotação do rotor ao gerador;
- gerador: converte energia mecânica rotacional em energia elétrica;
- controle: sistema de supervisão e controle que gerencia a operação da turbina, incluindo partida, parada e otimização de desempenho;
- sistema de orientação: mecanismo de rotação da nacele para alinhamento com o vento;
- torre: estrutura que eleva o rotor à altura adequada para capturar ventos de melhor qualidade;
- fundação: base estrutural que transmite todas as cargas da turbina ao solo;
- balanço do sistema elétrico: equipamentos de condicionamento de energia, transformadores e sistemas de proteção.

Os aerogeradores modernos incorporam sistemas de controle sofisticados que otimizam automaticamente a operação baseando-se em condições de vento em tempo real (Manwell; McGowan; Rogers, 2010). Destacam-se três estratégias principais de controle:

- controle por estol¹: limitação de potência por meio de características aerodinâmicas das pás;
- controle ativo do ângulo de ataque das pás;
- controle por orientação da nacele.

2.2 Modelagem de Curvas de Potência em Aerogeradores

A curva de potência é uma relação do comportamento da máquina perante o recurso incidente que o traduz em energia potencial. Teoricamente, a potência disponível no vento pode ser calculada pela Equação 2.1 (Clifton *et al.*, 2013):

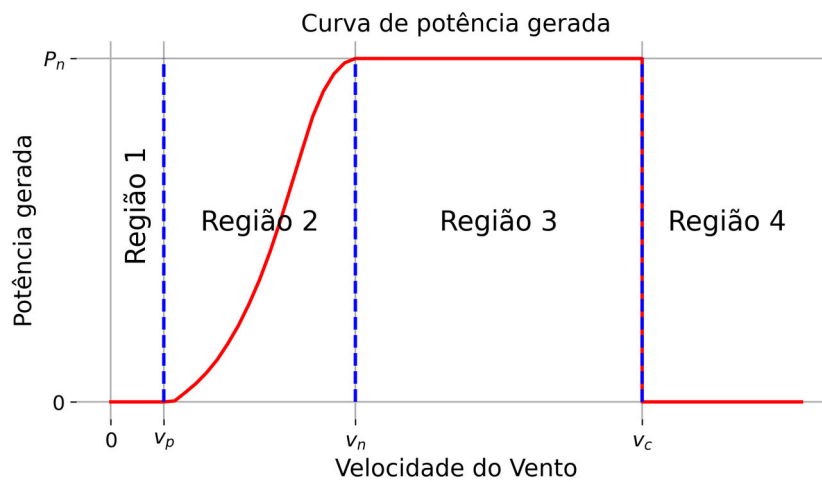
¹ Fenômeno aerodinâmico que ocorre quando o ângulo de ataque da pá (o ângulo entre o vento relativo e a pá) se torna muito alto, fazendo com que o fluxo de ar se descole da superfície da pá. Isso resulta em turbulência, perda de sustentação e um aumento significativo no arrasto.

$$P_e(v) = \frac{1}{2} C_p \rho A v^3, \quad (2.1)$$

onde P_e é a potência elétrica disponível, C_p é o coeficiente de potência, ρ é a densidade do ar, A é a área varrida pelas pás e v é a velocidade do vento. O coeficiente C_p depende de características mecânicas e operacionais complexas, como o ângulo de ataque e velocidade de rotação das pás, sendo fundamental na otimização do desempenho do aerogerador.

A curva de potência fornecida pelos fabricantes de turbina usualmente possui o formato apresentado na Figura 2.3, onde é possível identificar 4 regiões. A região 1 apresenta um limiar mínimo de velocidade V_p (*cut-in*) para que haja produção de energia potencial. Na região 2 há crescimento de geração de energia entre V_p e V_n . Já a região 3 indica que o WTG está operando em seu máximo. Para velocidades acima de V_c (*cut-out*), o aerogerador adentra a região 4, onde há uma interrupção no funcionamento do aerogerador para que seus componentes sejam protegidos (Sohoni; Gupta; Nema, 2016). No entanto, a curva frequentemente não representa adequadamente o desempenho operacional da turbina nas condições meteorológicas e geoespaciais específicas do local de instalação (Astolfi, 2021).

Figura 2.3 – Curva de Potência Típica de Aerogeradores



Fonte: Adaptado de Wang *et al.* (2019)

A curva de potência desempenha papel crucial na previsão de geração, monitoramento da saúde da turbina e estimativa de produção energética. O método padrão da indústria, descrito na norma IEC 61400-12-1 (International Electrotechnical Commission, 2017), estabelece um procedimento rigoroso que requer a instalação de uma TME dedicada a uma distância entre 2 e 4 diâmetros de rotor da turbina avaliada. O método consiste em coletar dados simultâneos

de velocidade do vento na TME e potência gerada na turbina eólica (WTG), segmentá-los em faixas de velocidade de vento e calcular a média de potência para cada faixa (Lydia *et al.*, 2014). A norma estabelece critérios rigorosos de filtragem, descartando dados coletados durante condições de operação anômala, falhas, manutenção ou quando a direção do vento resulta em efeito de esteira. Embora este método garanta padronização e rastreabilidade, apresenta alta propensão a erros devido à existência de diversos fatores externos como turbulência, densidade do ar, calibração de sensores e influência do terreno, além do descarte significativo de dados operacionais potencialmente informativos. A modelagem precisa de curvas de potência de aerogeradores representa um dos principais desafios na operação eficiente de parques eólicos. Uma revisão abrangente das abordagens para modelagem de curvas de potência é apresentada por Wang *et al.* (2019), destacando a evolução dos métodos tradicionais baseados na norma IEC 61400-12-1 para técnicas mais sofisticadas.

Entretanto, a modelagem de curvas de potência em ambiente de parque eólico enfrenta complexidades adicionais que extrapolam as condições controladas previstas pela norma IEC 61400-12-1. A presença de múltiplas turbinas operando simultaneamente introduz fenômenos aerodinâmicos de interação que afetam significativamente o desempenho individual das máquinas, tornando fundamental a compreensão destes efeitos para o desenvolvimento de modelos preditivos mais realistas e precisos.

2.2.1 Efeitos de Interação em Parques Eólicos

A operação de turbinas eólicas no parque introduz fenômenos aerodinâmicos complexos que afetam significativamente o desempenho individual das máquinas. Nygaard *et al.* (2020) demonstram o impacto que o efeito de bloqueio (redução da velocidade do vento a montante da turbina) e o efeito esteira (zona de velocidade reduzida e turbulência intensificada a jusante) exercem sobre a geração de energia em parques eólicos. Estes efeitos manifestam-se de forma espacialmente distribuída, com diferentes intensidades dependendo do espaçamento entre turbinas, direção do vento e características topográficas do terreno.

Para incorporar adequadamente esses fenômenos em modelos de curva de potência, Lyons & Göçmen (2021) avaliaram diferentes estratégias de seleção de fontes de dados, comparando o uso de medições da própria turbina, de turbinas vizinhas próximas e do parque completo. Os resultados indicaram que dados provenientes da região próxima (entre 25 e 50 diâme-

tros de rotor) proporcionam os melhores resultados preditivos, equilibrando a captura de efeitos locais sem incorporar ruído de turbinas distantes onde a correlação é menos significativa.

Além do impacto na produção energética, o efeito esteira induz carregamentos mecânicos adicionais nas turbinas afetadas devido à turbulência intensificada. Avendaño-Valencia, Abdallah & Chatzi (2021) desenvolveram sensores virtuais capazes de relacionar a intensidade de turbulência ao estresse mecânico dos componentes, permitindo prever e estimar danos estruturais nas pás das máquinas operando sob influência de esteira. Complementarmente, Schreiber, Bottasso & Bertelè (2020) propuseram um sistema integrado de estimação de fluxo de vento local e detecção de esteira utilizando dados combinados de TME e WTG, abrindo caminho para estratégias de otimização da geração através do controle ativo destes efeitos aerodinâmicos.

2.2.1.1 Métodos Paramétricos vs. Não Paramétricos

A literatura distingue claramente entre duas abordagens principais para modelagem de curvas de potência: métodos paramétricos e não paramétricos. Os métodos paramétricos, baseados em condições e modelos físicos, oferecem maior interpretabilidade mas frequentemente falham em capturar a complexidade das condições operacionais reais (Sohoni; Gupta; Nema, 2016).

Em contraste, os métodos não paramétricos, fundamentados em aprendizado de máquina (ML), utilizam primordialmente dados reais para suas modelagens, embora apresentem menor transparência devido à complexidade e à não-linearidade dos modelos. Estudos recentes demonstram que é possível desenvolver estruturas explicáveis para modelos de ML em turbinas eólicas (Letzgus; Müller, 2024).

Benabdesselam *et al.* (2024) apresentam uma revisão sistemática das técnicas de medição indireta em energias renováveis, classificando as abordagens em três categorias: baseadas em dados, baseadas em física e híbridas. Os autores destacam que métodos puramente baseados em dados, como as redes neurais discutidas anteriormente, oferecem alta flexibilidade e capacidade de capturar relações complexas, mas enfrentam desafios relacionados à explicabilidade, generalização e potencial obsolescência quando o sistema encontra condições operacionais não observadas durante o treinamento. Em contraste, abordagens híbridas que combinam conhecimento físico com aprendizado de máquina são alternativas promissoras, permitindo maior robustez e interpretabilidade. A revisão demonstra que há uma lacuna metodológica significativa

na literatura: a ausência de métodos sistemáticos de otimização multiobjetivo que considerem simultaneamente a seleção de características, fontes de dados e hiperparâmetros de modelo, especialmente no contexto de sistemas eólicos onde múltiplas fontes de medição distribuídas espacialmente estão disponíveis.

2.2.2 Aprendizado de Máquina aplicado em Eólicas

Pandit, Infield & Kolios (2019) realizam uma comparação abrangente entre métodos não paramétricos avançados, incluindo Processo Gaussiano (GP), Floresta Aleatória (RF) e Máquina de Vetores de Suporte (SVM), utilizando dados SCADA com pré-processamento por filtragem e correção por densidade do ar pela Equação 2.2:

$$v_c = v_m \left(\frac{\rho}{\rho_{ref}} \right)^{\frac{1}{3}}, \quad (2.2)$$

em que v_c é a velocidade do vento normalizada à densidade de referência, v_m é a velocidade do vento medida, ρ é a densidade do ar medida, ρ_{ref} é a densidade de referência. Os autores demonstraram que o GP obteve maior acurácia, seguido pelo RF e SVM. Vale acrescentar que há vantagem em diminuir a dimensionalidade do modelo ao corrigir a velocidade do vento pela densidade, mas perde-se a capacidade do modelo de interpretar o impacto real que a densidade pode ter na geração, aderindo ao erro intrínseco da Equação 2.2.

O uso de Árvores de Regressão (RT) para melhorar a precisão de curvas de potência específicas do local é explorado em (Barber; Nordborg, 2020), incorporando dados de turbulência e cisalhamento do vento. Os resultados mostraram raiz do erro médio quadrático (RMSE) três vezes menor que o do método tradicional de faixas, evidenciando o potencial dos métodos baseados em árvores de decisão.

Albuquerque (2021) investiga o uso de ferramentas de inteligência artificial, incluindo Redes Neurais Artificiais (ANN) e Sistemas de Inferência Fuzzy (FIS), para modelagem de curvas de potência. O estudo demonstra que redes neurais, especialmente arquiteturas como Perceptron Multicamadas (MLP), são capazes de capturar relações não lineares complexas entre variáveis meteorológicas e geração de energia.

Estudos que utilizam os dados de anemômetro, como Virgolino *et al.* (2020), demonstram que é possível criar a curva de potência do aerogerador. Porém, falhas e descalibração em

anemômetros são bastante comuns em WTGs; por isso, Cascianelli *et al.* (2022) desenvolveram uma abordagem de monitoramento de curvas de potência baseada em dados ambientais e operacionais, utilizando um conjunto de modelos que considera a existência ou não de dados de anemômetros.

Estudos mais recentes exploram abordagens probabilísticas e métodos *ensemble*. Wang, Li & Zhang (2023) propõem uma estimativa probabilística de curvas de potência baseada em fatores meteorológicos, utilizando distribuições Laplace e Gamma para melhor representar a variabilidade da geração eólica, já que distribuições gaussianas não representam adequadamente o fenômeno. Prakash *et al.* (2024) também buscaram a inclusão de informações diferentes do padrão da literatura, apresentando um modelo hierárquico bayesiano (BHM) que conecta características do terreno aos parâmetros da curva de potência, incorporando efeitos topográficos como inclinação, índice de rugosidade e altura máxima de cume. O método demonstrou melhoria de 7-10% no erro médio absoluto (MAE) em comparação ao método de faixas.

Abordagens híbridas e baseadas em aprendizado de máquina também são utilizadas na modelagem de curvas de potência. Wang, Niu & Yang (2024) propõem um modelo segmentado que combina regressão polinomial multivariada com *Extreme Gradient Boosting* (XGBoost) para predição de potência em turbinas eólicas. A abordagem segmenta a curva de potência em diferentes regiões operacionais, aplicando modelos específicos para cada segmento, o que permite capturar de forma mais precisa as não-linearidades presentes em diferentes faixas de velocidade do vento. De forma complementar, Purohit, Ng & Syed Ahmed Kabir (2022) avaliam os algoritmos SVM, ANN e XGBoost para predição de velocidade e intensidade de turbulência na esteira de turbinas eólicas. Utilizando simulações de dinâmica dos fluidos computacional (CFD) de alta fidelidade para gerar os conjuntos de treinamento e teste, os autores demonstram que os modelos baseados em aprendizado de máquina conseguem prever características da esteira com acurácia equivalente às simulações CFD, porém com tempo computacional similar aos modelos analíticos de baixa fidelidade. Entre os três algoritmos avaliados, o XGBoost apresentou o melhor desempenho combinando alta precisão e menor tempo de treinamento, evidenciando a viabilidade de métodos de aprendizado de máquina como alternativa aos modelos semi-empíricos tradicionais. Contudo, a aplicação bem-sucedida destes algoritmos depende fundamentalmente da qualidade dos dados utilizados em seu treinamento.

2.2.3 Qualidade de dados

Embora os avanços em algoritmos de aprendizado de máquina demonstrem crescente capacidade preditiva na modelagem de curvas de potência, a qualidade dos dados de entrada permanece como fator crítico para o desempenho dos modelos. O princípio fundamental de que dados de baixa qualidade resultam em previsões não confiáveis evidencia a necessidade de técnicas robustas de filtragem e validação de dados operacionais de turbinas eólicas (Lydia *et al.*, 2014). Neste contexto, a metodologia de pré-processamento de dados torna-se tão relevante quanto a escolha do algoritmo de modelagem.

Como descrito na Seção 2.2, o método tradicional da IEC 61400-12-1 para determinação de curvas de potência apresenta limitações significativas relacionadas à qualidade e aproveitamento dos dados. A exigência de instalação de uma TME dedicada exclusivamente para a medição representa investimento considerável em equipamentos, instalação e manutenção. Adicionalmente, os critérios de filtragem estabelecidos pela norma, embora garantam padronização, resultam no descarte de volumes substanciais de dados operacionais eliminando não apenas dados potencialmente ruidosos, mas também informações relevantes sobre o comportamento real da turbina em diferentes condições operacionais, comprometendo a capacidade de detectar padrões de degradação ou comportamentos anômalos. Estas limitações motivaram o desenvolvimento de técnicas mais sofisticadas de filtragem de dados que buscam maximizar o aproveitamento das informações disponíveis (Guo; Infield, 2013).

Bai *et al.* (2019) propõem uma solução estruturada para filtragem de dados por meio de três etapas: (1) redução de dimensionalidade utilizando Mapas Auto-Organizáveis (SOM) para clusterização; (2) aplicação de Máquina de Aprendizado Extremo (ELM) e *Earth-Mover Distance* (EMD) para filtrar dados que não representam o desempenho ótimo; e (3) modelagem final com Rede Neural de Função de Base Radial (RBFNN).

A aplicação de técnicas de detecção de anomalias, incluindo Floresta de Isolamento, *One-Class SVM* (OCSVM) e Autoencoders, permite separar dados que representam operação normal daqueles que indicam problemas operacionais ou paradas de máquina (Nascimento *et al.*, 2022; Nascimento, 2023).

2.3 Algoritmos de Otimização e Aprendizado de Máquina

Esta seção apresenta os principais algoritmos de aprendizado de máquina e otimização utilizados neste trabalho, destacando seus fundamentos, hipóteses, vantagens e limitações. O objetivo é fornecer a base conceitual necessária para compreender as escolhas metodológicas adotadas, bem como os critérios que orientam o ajuste de hiperparâmetros e a comparação de desempenho entre modelos.

2.3.1 Perceptron Multicamadas (MLP)

O *Perceptron*, inspirado no funcionamento biológico dos neurônios, representa o modelo mais fundamental de neurônio artificial e constitui a base conceitual das redes neurais modernas (Rosenblatt, 1958). O *Perceptron* recebe um conjunto de entradas, cada qual associada a um peso sináptico que representa a “força” da conexão, e produz uma saída binária por meio da aplicação de uma função de ativação a uma combinação linear ponderada das entradas (Bishop, 1995). Matematicamente, a saída de um *Perceptron* é dada pela Equação 2.3:

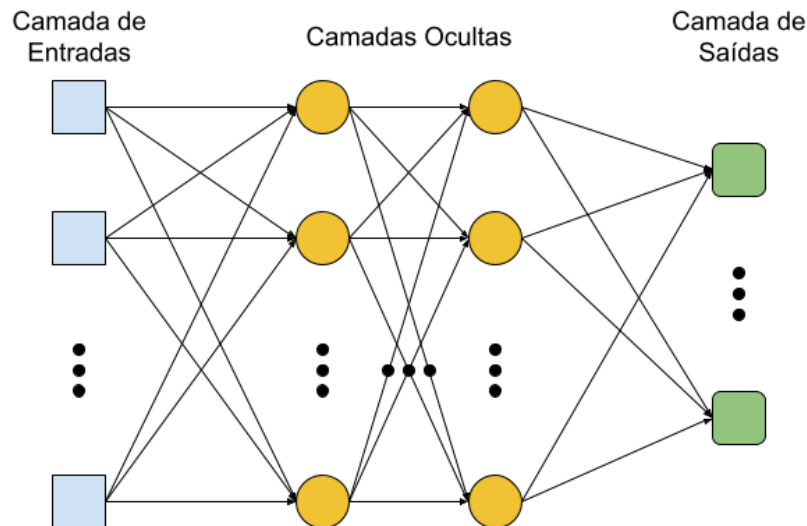
$$y = f \left(\sum_{i=1}^n w_i x_i + b \right), \quad (2.3)$$

onde w_i são os pesos sinápticos, x_i são as entradas, b é um termo de viés que permite ajustar o limiar de ativação e $f(\cdot)$ é uma função de ativação, tipicamente a função degrau. O processo de aprendizado do *Perceptron* consiste em ajustar iterativamente os pesos sinápticos e o viés, com base em exemplos de treinamento, permitindo que o modelo aprenda a classificar padrões linearmente separáveis (Haykin, 2009).

O *Multi Layer Perceptron* (MLP) supera a restrição a padrões linearmente separáveis ao organizar múltiplos *Perceptrons* em camadas sucessivas, criando uma arquitetura de rede neural capaz de aproximar funções não lineares arbitrariamente complexas (Bishop, 1995). Um MLP típico consiste em três tipos de camadas: uma camada de entrada que recebe os dados brutos, uma ou mais camadas ocultas que realizam transformações não lineares progressivas dos dados, e uma camada de saída que produz a predição final, como visto na Figura 2.4. Cada neurônio em uma camada está conectado a todos os neurônios da camada subsequente e cada conexão possui um peso associado que é ajustado durante o treinamento. A chave para a capacidade de

representação não-linear do MLP está no uso de funções de ativação não lineares nas camadas ocultas, como a sigmoide, tangente hiperbólica ou, mais comumente em aplicações modernas, a *Rectified Linear Unit* (ReLU), representada por $ReLU(z) = \max(0, z)$ (Goodfellow; Bengio; Courville, 2016).

Figura 2.4 – Estrutura de Perceptron Multicamadas



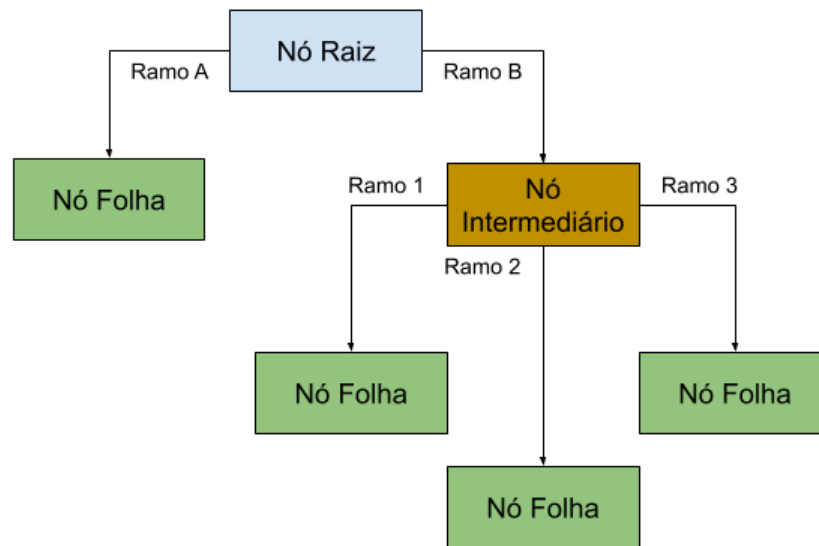
Fonte: Adaptado de Bishop (1995)

2.3.2 *Extreme Gradient Boosting* (XGBoost)

Árvores de decisão (DT) são modelos de aprendizado supervisionado que realizam previsões através de uma estrutura hierárquica de regras de decisão, organizadas em formato de árvore (Breiman *et al.*, 1984; Quinlan, 1986). A estrutura consiste em nós internos, que representam testes sobre atributos dos dados, ramos, que correspondem aos possíveis resultados desses testes, e nós folha, que contêm as previsões finais (Hastie; Tibshirani; Friedman, 2009), como na Figura 2.5. Embora árvores de decisão individuais sejam modelos intuitivos e interpretáveis, elas tendem a sofrer de alto viés quando muito rasas (*underfitting*) ou alta variância quando muito profundas (*overfitting*), tornando-se sensíveis a pequenas variações nos dados de treinamento (James *et al.*, 2013).

Para superar essas limitações, técnicas de aprendizagem *ensemble* combinam múltiplas árvores de decisão para produzir modelos mais robustos e precisos (Dietterich, 2000). Uma estratégia é o treinamento de múltiplas árvores em diferentes subconjuntos aleatórios de amostras e suas previsões são agregadas por votação ou média, como é o caso do *Random Forest*

Figura 2.5 – Estrutura de Árvore de Decisão



Fonte: Do Autor (2026)

(Breiman, 2001). Outra estratégia é o *boosting*, onde árvores são treinadas sequencialmente, com cada nova árvore focando em corrigir os erros das árvores anteriores (Freund; Schapire, 1997; Friedman, 2001). O *gradient boosting* representa uma abordagem particularmente poderosa de *boosting*, onde cada nova árvore é ajustada aos resíduos (erros) do modelo acumulado até o momento, seguindo a direção do gradiente negativo de uma função de perda diferenciável, de forma análoga aos métodos de otimização por gradiente em espaços funcionais (Friedman, 2001). Esse processo iterativo permite que o *ensemble* capture padrões complexos e não lineares nos dados de forma incremental, reduzindo tanto o viés quanto a variância (Hastie; Tibshirani; Friedman, 2009). O *Extreme Gradient Boosting* (XGBoost) é uma implementação otimizada e escalável do algoritmo de *gradient boosting* com árvores de decisão, que introduz diversas inovações técnicas em relação às implementações tradicionais de *gradient boosting* (Chen; Guestrin, 2016).

2.3.3 Algoritmo Genético Multiobjetivo (NSGA-II)

Algoritmos Genéticos são uma abordagem inspirada e fundamentada nos princípios da seleção natural e da genética evolutiva de Darwin (1859), que não buscam uma solução ótima de forma determinística, mas sim através da evolução de uma população de soluções candidatas ao longo de gerações sucessivas (Holland, 1975). Essa abordagem probabilística permite explorar

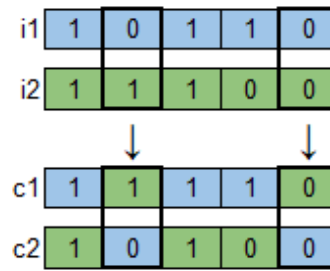
um espaço de busca amplo e complexo de maneira robusta, sendo particularmente eficaz em problemas onde métodos tradicionais baseados em gradientes ou busca exaustiva se mostram ineficientes ou intratáveis.

O funcionamento de um algoritmo genético clássico baseia-se na codificação de soluções candidatas como cromossomos, tipicamente representados por vetores binários, inteiros ou reais que constituem os “indivíduos” de uma população (Goldberg, 1989; Mitchell, 1998). Cada indivíduo possui um valor de aptidão (*fitness*) que quantifica a qualidade da solução que ele representa em relação ao problema. O ciclo evolutivo inicia com a geração de uma população inicial, frequentemente de forma aleatória, e então prossegue iterativamente por meio de três operadores genéticos principais:

1. **Seleção:** escolhe indivíduos da população atual para reprodução, favorecendo aqueles com maior aptidão.
2. **Cruzamento** (*crossover*): combina material genético de dois indivíduos pais para gerar descendentes, por meio de cruzamento uniforme bit a bit para representações binárias (Figura 2.6) ou cruzamento aritmético para representações reais.
3. **Mutação:** introduz variações aleatórias nos cromossomos dos descendentes, como inversão de bits em representações binárias (Figura 2.7) ou adição de ruído gaussiano em representações de números reais, garantindo diversidade genética e permitindo a exploração de novas regiões do espaço de busca (Jong, 2006).

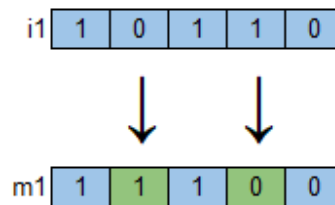
Após a aplicação desses operadores, forma-se uma nova geração que substitui total ou parcialmente a população anterior e o processo se repete até que um critério de parada seja satisfeito. Essa dinâmica de exploração (busca em novas regiões via mutação e cruzamento diversificado) e exploração (refinamento de boas soluções via seleção e cruzamento de indivíduos aptos) permite aos algoritmos genéticos navegar de modo eficaz por espaços de busca complexos, multimodais e descontínuos (Eiben; Smith, 2015).

Figura 2.6 – Cruzamento uniforme



Fonte: Do Autor (2026)

Figura 2.7 – Mutaç o em bit a bit

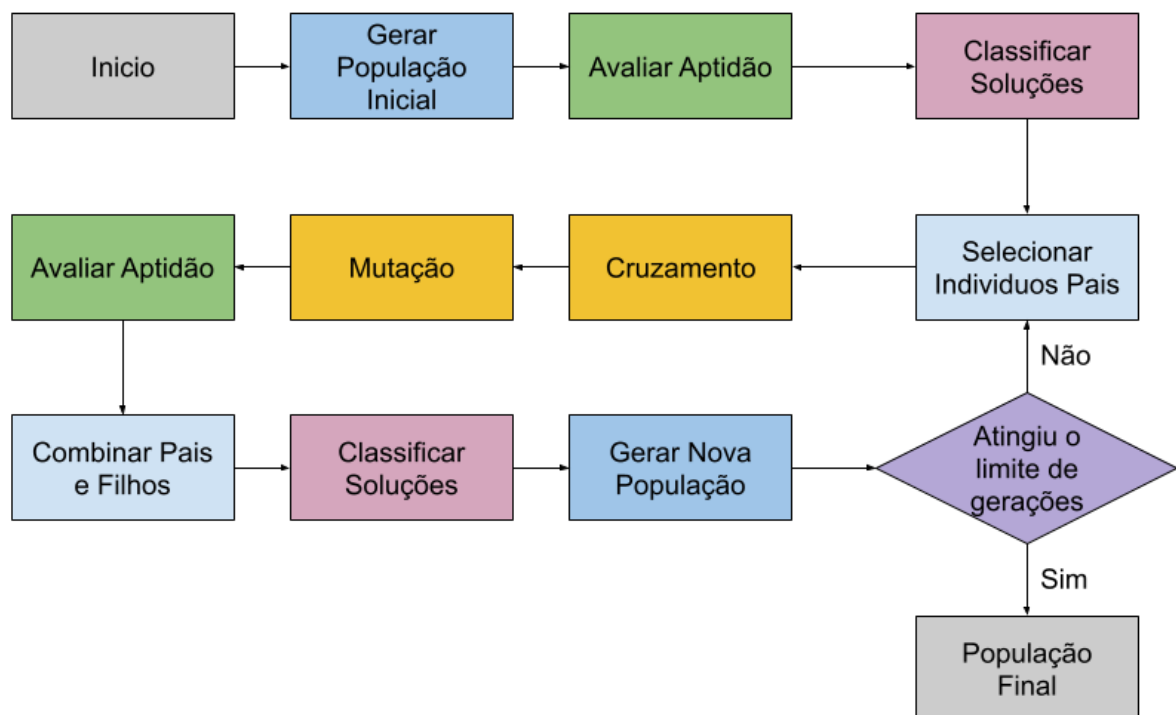


Fonte: Do Autor (2026)

O *Non-dominated Sorting Genetic Algorithm II* (NSGA-II) estende a estrutura dos algoritmos gen ticos tradicionais para o contexto de otimiza o multiobjetivo, onde m ltiplos crit rios conflitantes devem ser simultaneamente otimizados (Deb *et al.*, 2002). Ao contr rio dos algoritmos gen ticos mono-objetivo, que buscam convergir para uma  nica solu  o  tima por meio da maximiza o de uma fun  o de aptid o escalar, o NSGA-II busca identificar um conjunto de solu  es que representam os melhores compromissos poss veis entre os objetivos, conhecido como fronteira de Pareto. Para alcan ar esse objetivo, o algoritmo incorpora mecanismos espec ficos ao ciclo evolutivo tradicional. A principal modifica  o est  no crit rio de sele  o, no qual utiliza o conceito de domin ncia de Pareto, onde uma solu  o A domina uma solu  o B se A   pelo menos t o boa quanto B em todos os objetivos e estritamente melhor em pelo menos um objetivo. Com base nesse conceito, o algoritmo realiza a classifica  o por n o-domin ncia, organizando a popula  o em fronteiras de n veis hier rquicos. Durante o processo de sele  o para reprodu  o e forma  o da pr xima gera  o, indiv duos em fronteiras melhores s o priorizados. Adicionalmente, para garantir diversidade de solu  es ao longo da fronteira de Pareto e evitar converg ncia prematura para regi es espec ficas, o NSGA-II calcula a dist ncia de aglomera  o para cada indiv duo, que estima o qu o isolado ele est  em rela  o aos seus vizinhos no espa o dos objetivos, assim solu  es em regi es menos povoadas rece-

bem maior distância de aglomeração e são favorecidas quando há empate de nível hierárquico. Essa estratégia elitista, onde os melhores indivíduos da geração atual podem ser mantidos na próxima geração, combinada com a preservação de diversidade através da distância de aglomeração, permite ao NSGA-II convergir para um conjunto diversificado de soluções ótimas que representam diferentes compromissos entre os objetivos conflitantes do problema. O algoritmo iterativo simplificado pode ser visto na Figura 2.8.

Figura 2.8 – Fluxograma NSGA-II



Fonte: Do Autor (2026)

A adoção do NSGA-II neste trabalho decorre da natureza do problema de otimização formulado. A seleção de fontes de dados e de variáveis meteorológicas é essencialmente combinatória e codificada por genes binários, enquanto os hiperparâmetros dos modelos assumem valores inteiros e reais, resultando em um cromossomo misto e em uma função objetivo não diferenciável, avaliada por meio do treinamento de modelos de aprendizagem de máquina. Algoritmos genéticos operam diretamente sobre representações dessa natureza, admitindo operadores de cruzamento e mutação específicos para cada tipo de gene, sem exigir relaxamentos ou aproximações contínuas do espaço de busca. Entre os algoritmos evolutivos multiobjetivo, o NSGA-II foi escolhido por ser adequado a problemas com dois objetivos conflitantes, por dispensar a definição a priori de pesos entre acurácia e complexidade, que seriam arbitrários neste

contexto, e por combinar elitismo e preservação de diversidade em uma formulação consolidada e amplamente validada na literatura (Deb *et al.*, 2002), com implementação madura disponível na biblioteca DEAP.

2.4 Lacunas na literatura

Apesar dos avanços em ML para energia eólica, existe uma lacuna significativa na aplicação de métodos de otimização multiobjetivo que considerem simultaneamente aspectos técnicos e operacionais. Enquanto existem abordagens baseadas em inspeção visual (Barreto; Brasil; Souza, 2022), a maioria dos estudos foca exclusivamente na maximização da precisão preditiva, negligenciando fatores como:

- robustez operacional: capacidade de funcionar sem falhas na predição;
- custo de implementação: balanceamento entre número de fontes de dados, precisão e complexidade computacional;
- manutenibilidade: simplicidade para operação e manutenção em campo.

A literatura apresenta poucos estudos que abordem sistematicamente a seleção de fontes de dados em parques eólicos. Astolfi (2021) menciona que o uso de TME não resulta em melhora das predições, provavelmente devido ao fato de que poucos instrumentos altamente precisos, porém concentrados em poucos pontos, têm menor influência e potencialmente não conseguem reproduzir a complexidade dos parques eólicos. Porém, os modelos multivariáveis de curva de potência que utilizam dados de componentes da máquina apresentam uma limitação conceitual crítica: tendem a replicar o padrão operacional atual da turbina ao invés de estimar seu potencial real dadas as condições meteorológicas. Quando um modelo usa parâmetros controlados pelo sistema de controle da turbina como variáveis de entrada, ele aprende a imitar decisões de controle em vez de quantificar a capacidade física da máquina, incorporando limitações operacionais como comportamento “normal”. Isso perpetua estratégias de limitação intencional de potência, o que acaba camuflando o potencial real da turbina sob condições específicas. Além disso, estes modelos são inerentemente instáveis sob qualquer alteração no padrão operacional, seja por atualizações de *software* de controle, mudanças em estratégias operacionais, desgaste de componentes ou mesmo ajustes de manutenção preventiva. Qualquer

modificação no comportamento operacional compromete a validade do modelo, exigindo re-treinamento com novos dados, por sua vez, exigindo tempo suficiente de operação para coleta de dados.

Estudos que buscam reproduzir a curva de potência baseiam-se apenas na velocidade do vento, ignorando o fato de que a direção gera diferenças no efeito esteira e que a densidade do ar impacta a capacidade de extração de energia do vento. Além disso, a maioria dos estudos na literatura utiliza dados de parques eólicos europeus ou norte-americanos, com características meteorológicas e topográficas distintas das condições brasileiras. Existe uma lacuna na validação de metodologias de ML para condições específicas do Brasil, incluindo características climáticas, topográficas, tecnológicas e regulatórias.

Embora diversos estudos abordem filtragem de dados SCADA, poucos integram de forma sistemática múltiplas técnicas de detecção de anomalias para garantir qualidade dos dados de treinamento. A combinação de métodos paramétricos com técnicas não paramétricas permanece pouco explorada.

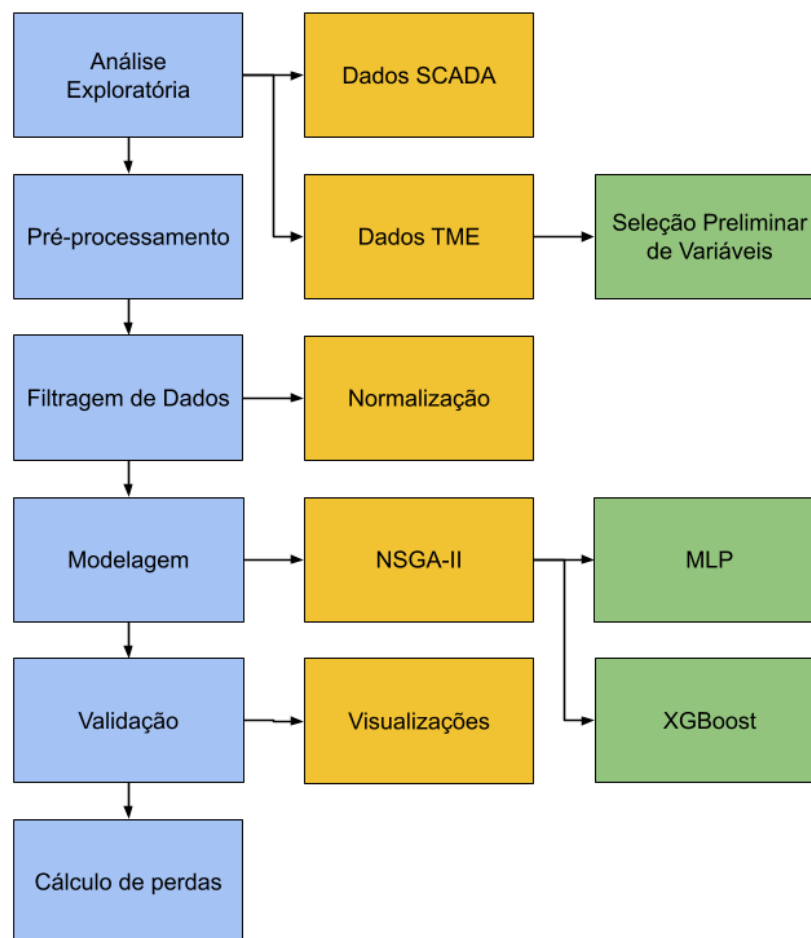
Outra dificuldade está na explicabilidade dos modelos de IA. Ainda é necessário o desenvolvimento de trabalhos que equilibrem o desempenho preditivo com a interpretabilidade, especialmente para aplicações críticas onde decisões operacionais dependem das predições do modelo (Letzgus; Müller, 2024).

Estas lacunas justificam a necessidade de pesquisas que integrem otimização multiobjetivo, seleção sistemática de dados, validação com condições brasileiras e consideração de aspectos práticos de implementação, configurando o contexto no qual este trabalho se insere.

3 MATERIAIS E MÉTODOS

A metodologia desenvolvida neste trabalho é estruturada em cinco etapas sequenciais que abordam desde a análise exploratória dos dados até a aplicação prática para estimativa de perdas energéticas. Além disso, será aplicado o método desenvolvido em outra turbina, para demonstrar sua eficácia.

Figura 3.1 – Fluxograma da metodologia implementada.



Fonte: Do Autor (2026)

3.1 Análise Exploratória e Compreensão dos Dados

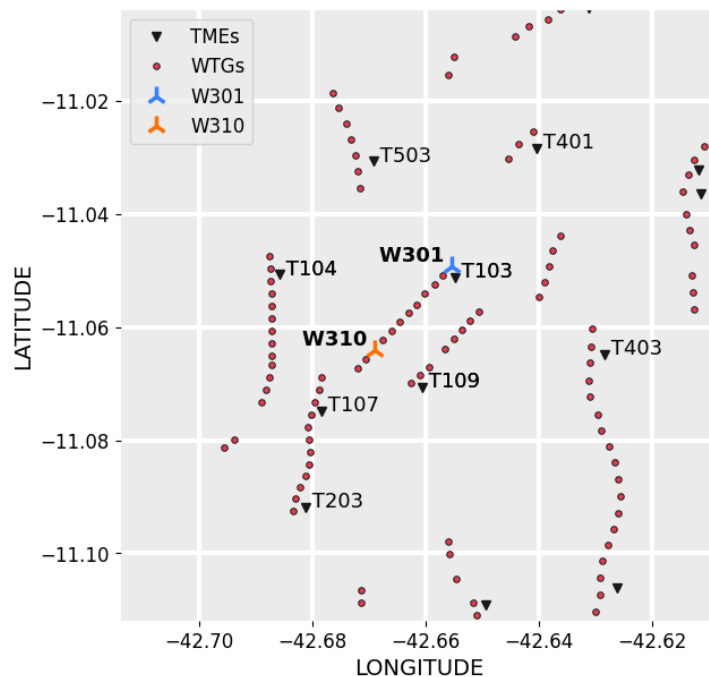
A primeira etapa da metodologia consiste em conhecer a origem, a estrutura e as características dos dados disponíveis, condição necessária para as decisões de modelagem tomadas nas etapas seguintes. Inicialmente, descrevem-se os conjuntos de dados fornecidos, sua localização, instrumentação e variáveis registradas, tanto do sistema SCADA dos aerogeradores quanto

das torres de medição eólica. Em seguida, analisam-se as correlações entre as variáveis, com o objetivo de avaliar redundâncias entre medições e orientar uma seleção preliminar que reduza a dimensionalidade do problema. Essa compreensão inicial fundamenta a definição das variáveis de referência empregadas na filtragem dos dados e na construção dos modelos preditivos.

3.1.1 Caracterização dos Dados

As turbinas chamadas de “W301” e “W310” estão localizadas próximas à cidade de Xique-Xique, Bahia, Brasil, em uma região de alta concentração de aerogeradores, com terreno montanhoso, onde as turbinas se encontram em diferentes elevações relativas entre si. A disposição da região pode ser vista na Figura 3.2, onde a região próxima está em um raio de 30 vezes o diâmetro de rotor do WTG, visto em Figura 3.3 e Figura 3.4, dentro do mencionado na Subseção 2.2.1. Os aerogeradores em questão são máquinas da fabricante GE de 2,5 MW, com 80 metros de altura de nacele e 116 metros de diâmetro de rotor.

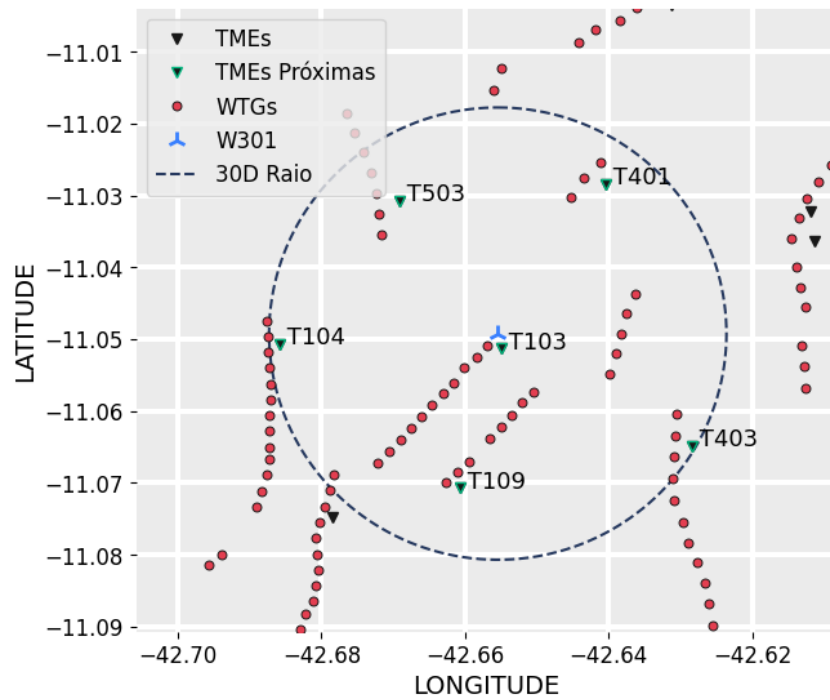
Figura 3.2 – Disposição de WTGs e TMEs na região.



Fonte: Do Autor (2026)

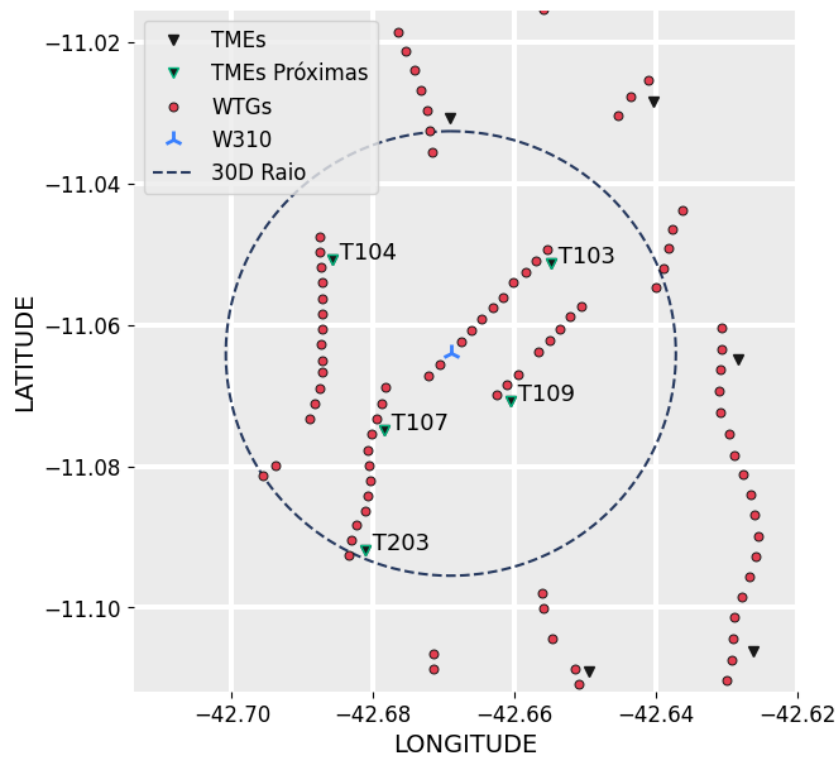
Foi fornecido pela empresa Serena Energia, um conjunto de dados obtidos do sistema SCADA para os WTGs “W301” e “W310”, tendo 31 variáveis e 51564 amostras entre janeiro de 2024 e janeiro de 2025. O sistema SCADA registra os dados em 1Hz e compila em uma

Figura 3.3 – Disposição de WTGs e TMEs próximas do W301.



Fonte: Do Autor (2026)

Figura 3.4 – Disposição de WTGs e TMEs próximas do W310.



Fonte: Do Autor (2026)

amostra a cada 10 minutos. Para cada variável, são registrados na amostra os valores da média, máximo, mínimo, desvio padrão e contagem de registros coletados. As variáveis contidas no SCADA estão descritas na Tabela 3.1. Além disso, também foram concedidos os registros de eventos de parada e limitação de máquina, contendo nome do evento, instante de início e fim de cada ocorrência.

Para os dados meteorológicos, foram utilizadas 8 TMEs chamadas: T401, T403, T503, T104, T107, T203, T103 e T109. Estas estão a uma distância de $30D$, onde D é o diâmetro do rotor do WTG, marcadas como “TMEs Próximas” nas Figura 3.3 e Figura 3.4. As variáveis também possuem período de amostragem de 10 minutos e são padronizadas entre cada torre, como é mostrado na Tabela 3.2 e Tabela 3.3. Tais dados advêm de um sistema que já filtra os dados meteorológicos vindos do *datalogger* de cada TME, para se ter somente dados de alta confiabilidade. Cada TME tem 1 termo-higrômetro, medindo temperatura e umidade, 1 barômetro, medindo a pressão, 2 sensores de direção do vento e 3 anemômetros, medindo velocidade do vento em diferentes alturas da torre. Para cada anemômetro é calculada a sua intensidade de turbulência respectiva, T_i que é calculada pela Equação 3.1:

$$T_i = \frac{V_{std}}{V_{avg}}, \quad (3.1)$$

em que T_i é a intensidade da turbulência, V_{std} é o desvio padrão da velocidade do vento da amostra e V_{avg} é a média da velocidade do vento da amostra (Manwell; McGowan; Rogers, 2010). Além disso, é calculada a densidade do ar, por meio da Equação 3.2 (Monteith; Unsworth, 2013; CIPM, 2007):

$$\begin{cases} p_{sat} = 610,78 \exp\left(\frac{17,27(T-273,15)}{T-35,85}\right), \\ p_v = H \times p_{sat}, \\ p_d = p - p_v, \\ \rho = \frac{p_d}{R_d T} + \frac{p_v}{R_v T}, \end{cases} \quad (3.2)$$

em que T é a temperatura em K, H é a umidade relativa do ar, p é a pressão atmosférica, p_{sat} é a pressão de saturação da água, p_v é a pressão de vapor da água, p_d é a pressão do ar seco, $R_d = 287,058 J/(kgK)$ é a constante específica do gás para o ar seco, $R_v = 461,495 J/(kgK)$ é a constante específica do gás para o vapor de água e ρ é a densidade do ar em kg/m^3 .

Tabela 3.1 – Variáveis SCADA

Variável SCADA	Descrição	Unidade
BL1_ACT	Ângulo de ataque da pá 1	°
BL2_ACT	Ângulo de ataque da pá 2	°
BL3_ACT	Ângulo de ataque da pá 3	°
DrvTr_DrvTrOscAccel	Aceleração do sistema de transmissão	m/s^2
FREQ	Frequência da linha	Hz
I_A	Corrente eficaz fase A	A
I_B	Corrente eficaz fase B	A
I_C	Corrente eficaz fase C	A
N_GEN_CCU	Velocidade do gerador	RPM
N_ROT_PLC	Velocidade do rotor	RPM
PF	Fator de potência	-
POS_NAC	Direção da Nacele	°
P_ACT	Potência ativa	kW
Q	Potência reativa	kVar
T_AMB	Temperatura ambiente	°C
T_BEAR_A	Temperatura rolamento A	°C
T_BEAR_B	Temperatura rolamento B	°C
T_BEAR_SHAFT	Temperatura rolamento do eixo	°C
T_GEAR	Temperatura caixa multiplicadora	°C
T_GEAR_BEAR	Temperatura rolamento da caixa multiplicadora	°C
T_GEN_1	Temperatura gerador 1	°C
T_GEN_2	Temperatura gerador 2	°C
T_GEN_COOL	Temperatura ar de refrigeração do gerador	°C
T_MAIN_BOX	Temperatura caixa principal	°C
T_NAC	Temperatura Nacelle	°C
T_TOP_BOX	Temperatura caixa superior	°C
U_A_N	Tensão eficaz Fase A-Neutro	V
U_B_N	Tensão eficaz Fase B-Neutro	V
U_C_N	Tensão eficaz Fase C-Neutro	V
V_WIN	Velocidade do Vento	m/s
WIND_DEV_1SEC	Desvio do vento incidente	°

Fonte: Do Autor (2026)

Também foram concedidos dados do sistema de faturamento (SMF¹), contendo a geração do parque e o dado de restrição de geração (*curtailment*²) com motivo da restrição, início e fim do período restrito, para o mesmo período que os aerogeradores.

¹ Sistema de Medição para Faturamento é um conjunto de equipamentos e tecnologias que registra e envia dados de consumo de energia elétrica para a Câmara de Comercialização de Energia Elétrica (CCEE) (Operador Nacional do Sistema Elétrico (ONS), 2015)

² Evento de redução ou corte da geração de energia elétrica, especialmente de fontes renováveis (solar e eólica), mesmo quando há capacidade de produzir, para manter a estabilidade e segurança do sistema,

Tabela 3.2 – Descrição das Variáveis da TME

Variável TME	Descrição	Unidade
ATM_PRESSURE	Pressão Atmosférica	hPa
TEMP_AIR	Temperatura do ar	°C
HUMIDITY_RELATIVE	Umidade relativa	%
WIND_DIR	Direção do vento	°
WIND_SPD	Velocidade do vento	m/s

Fonte: Do Autor (2026)

Tabela 3.3 – Informações das Variáveis da TME

Variável TME	Informação Disponível			
	<i>Média</i>	<i>Máximo</i>	<i>Mínimo</i>	<i>Desv Pad</i>
ATM_PRESSURE	X			
TEMP_AIR	X			
HUMIDITY_RELATIVE	X			
WIND_DIR	X			X
WIND_SPD	X	X	X	X

O X indica que a informação existe para a variável

Fonte: Do Autor (2026)

3.1.2 Análise de Correlações e Seleção Preliminar de Variáveis

Cada TME tem múltiplos instrumentos de medição de vento em alturas distintas, como visto na Figura 2.1. A fim de determinar se seria necessário manter várias medições, foi realizada uma avaliação da matriz de correlação e foi avaliada a possibilidade de descartar algumas informações previamente para reduzir a dimensionalidade do modelo.

Considerando a variedade de medidas provenientes do sistema SCADA do aerogerador, foram selecionadas variáveis específicas para compreender as relações entre elas e a potência gerada. O objetivo principal consiste em filtrar as amostras nas quais a potência do aerogerador não representa seu melhor desempenho operacional, ou seja, amostras nas quais o aerogerador é incapaz de extrair a energia potencial do vento incidente. Para isso, torna-se necessário identificar os pontos de dados que não apresentam o comportamento esperado de uma curva de potência típica, isto é, amostras que se desviam significativamente do padrão apresentado na

ocorrendo por excesso de oferta, falta de demanda ou limitações da rede de transmissão (Operador Nacional do Sistema Elétrico (ONS), 2026)

Figura 2.3. Mas, como descrito na Seção 2.2, o sensor de velocidade do vento apresenta problemas recorrentes de calibração, isso acontece devido à turbulência intensa a que está submetido por estar na nacelle; além disso, as manutenções deste sensor são de custo elevado, impossibilitando que sejam feitas recorrentemente. Essa limitação compromete a confiabilidade da velocidade do vento como variável de referência, resultando em uma filtragem inadequada dos dados e adicionando complexidade desnecessária ao processo. Diante dessa condição, faz-se necessário utilizar dados provenientes de outros componentes do sistema que não apresentam os mesmos problemas de calibração. Essa abordagem permite identificar e remover as amostras que não se adequam ao comportamento típico da curva de potência, utilizando variáveis mais confiáveis do sistema SCADA.

3.2 Pré-processamento dos Dados

A potência ativa é convertida em Fator de Capacidade (FC), representando a porcentagem da potência nominal pela Equação 3.3:

$$FC = \frac{P_{act}}{P_{nom}}, \quad (3.3)$$

em que P_{act} é a potência ativa, P_{nom} é a potência nominal e FC é a porcentagem da potência nominal sendo gerada e saturada entre 0 e 1. O ajuste é devido ao consumo de energia do WTG quando não há condições de geração e há um pequeno excedente acima da nominal para compensar perdas elétricas.

Para a análise de consistência entre pares, define-se a diferença de fator de capacidade como o desvio entre a referência e o aerogerador analisado pela Equação 3.4:

$$\Delta FC = FC_{ref} - FC_{WTG}, \quad (3.4)$$

em que FC_{WTG} é o fator de capacidade do aerogerador analisado, FC_{ref} é o fator de capacidade do aerogerador vizinho ou parque (SMF) e ΔFC é o desvio do fator de capacidade.

As direções do vento são decompostas em componentes vetoriais cartesianas u (zonal) e v (meridional) para facilitar o processamento estatístico (Stull, 1988), seguindo a convenção meteorológica padrão, na qual a direção representa a origem do fluxo (Stull, 1988), pela

Equação 3.5:

$$\begin{cases} u = -\sin(\theta_{rad}) \\ v = -\cos(\theta_{rad}) \end{cases} \quad (3.5)$$

onde θ_{rad} representa a direção do vento medida em radiano. Isso é importante para evitar distorções ao normalizar o dado, em que 0° e 360° são a mesma direção, porém numericamente distantes.

3.3 Filtragem de Dados

Os dados operacionais coletados por sistemas SCADA de aerogeradores representam o comportamento real da máquina sob todas as condições operacionais, incluindo períodos de funcionamento ótimo, mas também eventos de parada programada ou não programada, limitações de potência e baixo desempenho. A presença dessas amostras anômalas nos dados de treinamento pode comprometer significativamente a capacidade dos modelos de aprendizado de máquina de capturar adequadamente a relação entre condições meteorológicas e potencial de geração, uma vez que o modelo aprenderia padrões associados a estados degradados ou limitados ao invés da operação ótima do equipamento. Portanto, torna-se fundamental a implementação de metodologias robustas de filtragem de dados que identifiquem e removam amostras correspondentes a condições operacionais anômalas, preservando apenas aquelas representativas do desempenho esperado do aerogerador em operação normal.

Embora algoritmos de detecção de anomalias baseados em aprendizado de máquina, como o *Isolation Forest*, sejam amplamente utilizados para identificação de *outliers*, a aplicação direta para filtragem de dados operacionais de aerogeradores apresenta limitações fundamentais que comprometem sua adequação para este propósito específico. O *Isolation Forest* requer a especificação *a priori* da taxa de contaminação esperada nos dados, ou seja, a proporção de amostras que devem ser classificadas como anômalas. Esta característica representa um desafio significativo no contexto de dados SCADA de aerogeradores, onde a proporção real de amostras anômalas varia consideravelmente entre diferentes períodos de operação e entre aerogeradores e não é conhecida previamente com precisão. Mais fundamentalmente, o algoritmo identifica como anomalias quaisquer amostras que ocorrem com baixa frequência no conjunto de dados, independentemente de representarem condições operacionais, porém raras. Esta característica

resulta na remoção inadequada de dados correspondentes a velocidades de vento extremas, tanto baixas quanto altas, que naturalmente ocorrem com menor frequência devido à distribuição estatística típica do recurso eólico, mas que são essenciais para a correta caracterização do comportamento do aerogerador em toda sua faixa operacional. Como pode ser observado em Brito, Souza & Barbosa (2025) (reproduzido integralmente no ?? para conveniência do leitor), a aplicação do *Isolation Forest* aos dados deste estudo resultou na eliminação sistemática de amostras nas extremidades da distribuição de velocidades de vento, potencialmente comprometendo a representatividade do conjunto de dados filtrado. Por estas razões, optou-se por desenvolver uma metodologia de filtragem baseada em critérios físicos, operacionais, regulatórios e estatísticos para remover amostras inadequadas:

1. Filtro de Eventos (02-EVENTS): remove amostras com eventos de parada de máquina, limitação e *curtailment*;
2. Filtro de Contagem de Dados (03-NAN_COUNT): remove amostras com menos de 600 registros coletados no período de 10 minutos, indicando problemas de comunicação ou aquisição e elimina instantes onde há ausência de dados;
3. Filtro de Frequência da Rede (04-FREQ): remove amostras onde a frequência da rede elétrica está fora dos padrões estabelecidos pela ANEEL (59,9 - 60,1 Hz), indicando condições anômalas de operação ((ANEEL), 2022).
4. Filtro de Ângulo de Ataque da Pá (05-BLADE): remove amostras que satisfaçam simultaneamente $FC \in (0, 1)$ e $\alpha \geq 7,5^\circ$, onde α representa o ângulo de ataque da pá, indicando que o aerogerador se encontra operando fora do ângulo de ataque definido pelo fabricante;
5. Filtro de Diferenças para WTG vizinhos e SMF (06-DIFF_WTG_SMF): elimina amostras onde $\Delta FC > 0,25$, indicando que há grande diferença de desempenho entre a turbina de referência e as turbinas vizinhas ou SMF (International Electrotechnical Commission, 2017);
6. Filtro de 2 Desvios Padrão (07-2DESVPAD): remove amostras cujas diferenças de FC em relação aos vizinhos ou SMF excedam o limite de $\mu_\Delta + 2\sigma_\Delta$, calculado por faixa de 10% do fator de capacidade (FC arredondado) do WTG e interpolado linearmente a cada 1% de FC, gerando uma curva (tabela) de limite de desvio para cada 1% de FC arredondado;

7. Filtro de Gerador (08-GENERATOR): descarta amostras cuja velocidade do gerador (v_{gen}) esteja fora do intervalo $[\mu_{gen} \pm 2\sigma_{gen}]$ para a respectiva faixa de 1% FC ;
8. Filtro de Sequência Temporal de Filtros (09-FILTER_SEQUENCE): remove x_t se x_{t-1} e x_{t+1} forem amostras filtradas;

O cálculo da média (μ_{Δ}) e do desvio padrão (σ_{Δ}) baseia-se exclusivamente nas amostras onde $\Delta FC > 0$. Esta premissa justifica-se pelo fato de que diferenças negativas indicam que o aerogerador sob análise apresenta geração superior à sua referência, sugerindo restrições de disponibilidade ou baixo desempenho na própria referência (vizinhos ou SMF), e não no WTG avaliado.

3.3.1 Normalização dos Dados

Para melhorar a capacidade preditiva da modelagem, é necessário normalizar as variáveis de entrada. As características escolhidas para normalização foram as variáveis meteorológicas de direção e velocidade do vento e densidade do ar medidos pela TME. Para estes dados, foi aplicada a normalização entre máximo e mínimo, conforme a Equação 3.6:

$$X_n = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (3.6)$$

em que X_n é o dado normalizado, X é o dado a ser normalizado, X_{min} e X_{max} são os valores mínimo e máximo respectivamente, entre os quais o dado deve ser normalizado. Faixas de normalização específicas:

- velocidade do vento: [0 ; 32,5] m/s;
- direção do vento (u e v): [-1 ; 1];
- densidade do ar: [1,0 ; 1,225] kg/m³;

3.4 Desenvolvimento de Modelos

A estimativa do fator de capacidade do aerogerador foi abordada por meio de três estratégias de modelagem: um modelo de referência baseado em média por faixas de velocidade de

vento, similar ao método prescrito pela norma IEC 61400-12-1; e dois modelos baseados em aprendizado de máquina: o MLP e o XGBoost. Para os modelos de aprendizado de máquina, a configuração adequada envolve múltiplas decisões interdependentes: quais torres de medição utilizar como fontes de dados, quais variáveis meteorológicas incluir de cada torre e como ajustar os hiperparâmetros específicos de cada algoritmo.

Com esse propósito, foi implementado o algoritmo de otimização evolutiva multiobjetivo NSGA-II (*Non-dominated Sorting Genetic Algorithm II*), que realiza simultaneamente a seleção de fontes de dados (torres de medição), a escolha de características meteorológicas de cada fonte e a otimização de hiperparâmetros dos modelos MLP e XGBoost. Esta estratégia permite explorar sistematicamente o espaço de soluções considerando explicitamente o compromisso entre dois objetivos conflitantes: maximização da acurácia preditiva e minimização da complexidade do modelo.

A otimização multiobjetivo resulta em um conjunto de soluções não-dominadas (fronteira de Pareto), cada uma representando um equilíbrio diferente entre acurácia e complexidade. Esta abordagem oferece vantagens práticas significativas sobre métodos de objetivo único: permite ao operador do parque eólico escolher a configuração mais adequada às suas restrições operacionais e econômicas específicas, sem impor arbitrariamente pesos relativos entre os objetivos ou estabelecer limites rígidos de complexidade que podem descartar soluções potencialmente úteis.

3.4.1 Modelo de Referência (Base)

O modelo de referência utiliza exclusivamente a velocidade do vento medida pela torre meteorológica mais próxima ao aerogerador. Para cada faixa de velocidade de vento de 0,5 m/s, calcula-se a média aritmética dos valores observados de fator de capacidade. Os valores intermediários são obtidos por interpolação linear a cada 0,1 m/s. Este método aproxima-se do procedimento descrito pela norma IEC 61400-12-1 para construção de curvas de potência e serve como linha de base para avaliação dos ganhos obtidos pelos modelos de aprendizado de máquina.

3.4.2 Algoritmo de Otimização NSGA-II

O algoritmo genético multiobjetivo, desenvolvido para otimizar o MLP e XGBoost, visa a minimização simultânea das funções de aptidão da Equação 3.7, ajustando a escolha de fontes, características e hiperparâmetros. Para garantir que não há interferência nos dados escolhidos para treino e teste, foi utilizado o método *K-fold* (Wong, 2015) para validação cruzada com 5 *folds*. Os diversos modelos da fronteira de Pareto gerada pelo NSGA-II foram comparados pela função de aptidão e ao modelo base.

O problema de otimização multiobjetivo para os modelos MLP e XGBoost é formulado como a minimização simultânea de duas funções objetivo, conforme a Equação 3.7:

$$\min_{x \in \Omega} F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \end{pmatrix}, \quad (3.7)$$

onde Ω representa o espaço de busca definido pelas possíveis combinações de fontes de dados, características meteorológicas e hiperparâmetros do modelo; $f_1(x)$ quantifica o erro de predição por meio da raiz do erro quadrático médio (RMSE), o valor adotado é a média dos RMSE calculados individualmente em cada uma das $k = 5$ partições da validação cruzada, conforme a Equação 3.8; e $f_2(x)$ representa a complexidade do modelo por meio do produto entre o número de fontes utilizadas e o número total de características selecionadas, como definido na Equação 3.9:

$$f_1(x) = \frac{1}{k} \sum_{j=1}^k \sqrt{\frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - \hat{y}_{ij})^2}, \quad (3.8)$$

$$f_2(x) = N_f \cdot n_c, \quad (3.9)$$

sendo n_j o número de amostras na partição de validação j , y_{ij} o valor observado do fator de capacidade para a amostra i na partição j , $\hat{y}_{ij}$ a predição do fator de capacidade para a amostra i correspondente do modelo treinado exclusivamente na partição j , N_f o número de torres de medição (TME) utilizadas como fontes de dados e n_c o número total de características meteorológicas selecionadas de todas as fontes. A função $f_1(x)$ representa, assim, a média aritmética dos valores de RMSE calculados independentemente em cada uma das cinco partições de va-

lidação, assegurando estimativa robusta da capacidade de generalização do modelo para dados não observados durante o treinamento.

A métrica de complexidade $f_2(x)$ foi projetada para penalizar tanto o uso de múltiplas torres quanto a seleção de muitas características, refletindo os custos práticos associados: cada torre adicional representa investimento em instrumentação e manutenção; cada característica adicional aumenta o risco de dependência de sensores específicos e a vulnerabilidade a falhas de equipamento. O produto entre N_f e n_c penaliza especialmente configurações que utilizam poucas características de muitas torres, resultando em dependência distribuída.

O algoritmo NSGA-II foi implementado utilizando a biblioteca DEAP (*Distributed Evolutionary Algorithms in Python*) (Fortin, 2012), com os seguintes parâmetros operacionais, calibrados empiricamente em experimentos preliminares para balancear exploração e a convergência para boas soluções:

- Tamanho da população: 100 indivíduos;
- Critério de parada: 100 gerações;
- Operador de seleção: seleção por torneio com pressão seletiva de tamanho 3, aplicando o conceito de dominância de Pareto e distância de aglomeração (*crowding distance*) do NSGA-II;
- Operador de cruzamento (*crossover*): probabilidade de 0,8 para o indivíduo, utilizando o método uniforme com probabilidade de troca de 0,5 por gene;
- Operador de mutação: probabilidade de 0,3 para o indivíduo, com uma taxa de mutação pontual de 0,08 por gene;
- Estratégia elitista: preservação integral da fronteira de Pareto entre gerações sucessivas.

A codificação das soluções candidatas em cromossomos segue uma estrutura de comprimento variável que representa simultaneamente três níveis de decisão: seleção de torres de medição como fontes de dados, escolha de características meteorológicas de cada torre selecionada e configuração dos hiperparâmetros específicos do modelo de aprendizado de máquina.

Cada cromossomo é composto por $N_{TME} + N_{hiper}$ genes, onde N_{TME} corresponde ao número total de torres de medição próximas ao WTG analisado e N_{hiper} representa o número de hiperparâmetros do modelo sendo otimizado (XGBoost ou MLP). Os primeiros N_{TME} genes

codificam a seleção de características de cada torre, enquanto os genes subsequentes codificam os valores dos hiperparâmetros do modelo. Para cada torre de medição T , o gene correspondente é um vetor binário de dimensão 4: $\mathbf{g}_t = [b_{\text{vel}}, b_{\text{dir}}, b_{\text{turb}}, b_{\rho}]$, onde cada elemento binário $b_i \in \{0, 1\}$ indica a inclusão (1) ou exclusão (0) da respectiva variável meteorológica:

- b_{vel} : velocidade do vento;
- b_{dir} : direção do vento, representada pelas componentes u e v ;
- b_{turb} : intensidade de turbulência;
- b_{ρ} : densidade do ar.

Uma característica importante da codificação é o tratamento da direção do vento: embora seja representada por um único *bit* no cromossomo, sua ativação ($b_{\text{dir}} = 1$) resulta na inclusão de duas variáveis no modelo, as componentes u (zonal) e v (meridional) da direção do vento. Esta abordagem garante que a informação direcional seja sempre utilizada de forma completa, evitando configurações inválidas onde apenas uma das componentes seria incluída, o que resultaria em perda de informação direcional ou interpretação física incorreta dos dados de vento.

A estrutura do cromossomo permite representar implicitamente tanto a seleção de fontes quanto a seleção de características: uma torre é considerada uma fonte ativa se e somente se ao menos uma de suas características for selecionada ($\exists b_i = 1$ em $\mathbf{g}_t$); caso contrário ($\mathbf{g}_t = [0, 0, 0, 0]$), a torre não contribui para o modelo. Esta codificação possui duas vantagens importantes: primeiro, reduz o espaço de busca ao eliminar a necessidade de um gene adicional para ativação/desativação explícita de cada torre; segundo, previne naturalmente a criação de soluções inválidas onde uma torre seria marcada como ativa sem nenhuma característica selecionada.

Os genes de hiperparâmetros têm sua codificação atrelada ao tipo de parâmetro. Hiperparâmetros inteiros são codificados diretamente como valores inteiros dentro de seus intervalos permitidos. Hiperparâmetros contínuos são codificados como valores de ponto flutuante. Hiperparâmetros em escala logarítmica são codificados no espaço logarítmico e convertidos para escala linear durante a avaliação do modelo.

Durante os operadores genéticos de cruzamento e mutação, os genes de características e hiperparâmetros são tratados com estratégias distintas: genes binários de características utilizam cruzamento uniforme e mutação *bit a bit* como explicado em Subseção 2.3.3; genes

de hiperparâmetros inteiros sofrem mutação por incremento/decremento limitado aos intervalos permitidos; genes de hiperparâmetros contínuos utilizam mutação gaussiana com desvio padrão adaptativo proporcional ao tamanho do intervalo de busca. Esta diferenciação entre tipos de genes melhora a eficiência do processo evolutivo ao aplicar operadores geneticamente apropriados a cada tipo de parâmetro.

Além disso, todas as avaliações de modelos durante o processo evolutivo foram executadas com aceleração por GPU para viabilizar a exploração das configurações candidatas dentro de prazo computacional razoável. Os experimentos foram conduzidos em um sistema equipado com GPU NVIDIA GeForce RTX 3060 com 12 GB de memória VRAM, processador Intel Core i7-12700K e 32 GB de RAM.

3.4.2.1 XGBoost: Espaço de Busca

O modelo XGBoost foi implementado em Python utilizando a biblioteca XGBoost (Chen; Guestrin, 2016), executado em modo de regressão com função de perda quadrática (*squared error*). Para acelerar o treinamento dos modelos durante o processo evolutivo, os dados foram transferidos para memória da GPU utilizando a biblioteca CuPy (Okuta *et al.*, 2017), permitindo o uso da implementação CUDA-acelerada do XGBoost (*device='cuda'*). O espaço de busca definido para a otimização de seus hiperparâmetros compreende os seguintes intervalos:

- Profundidade máxima (*max_depth*): valores inteiros entre 3 e 10;
- Taxa de aprendizado (*learning_rate*): valores contínuos entre 0,01 e 0,3;
- Número de estimadores (*n_estimators*): valores inteiros no intervalo [50, 500];
- Subamostragem (*subsample*): fração de amostras por árvore entre 0,5 e 1,0;
- Fração de colunas (*colsample_bytree*): proporção de atributos por árvore entre 0,5 e 1,0;
- Peso mínimo do nó filho (*min_child_weight*): valores inteiros de 1 a 10;
- Gama (γ): parâmetro de regularização para redução de perda entre 0 e 5;
- Termos de regularização L1 (α) e L2 (λ): busca em escala logarítmica no intervalo de 10^{-8} a 10.

A combinação destes hiperparâmetros define o compromisso entre viés e variância do modelo: profundidade, número de estimadores e taxa de aprendizado controlam a capacidade expressiva; subamostragem e fração de colunas promovem generalização por meio de diversidade; peso mínimo, gama e termos de regularização limitam a complexidade para prevenir sobreajuste. O algoritmo NSGA-II incorpora estes hiperparâmetros aos genes de acordo com a Tabela 3.4.

Tabela 3.4 – Exemplo de Cromossomo XGBoost

Gene	Característica
TME ₁	b_{vel}
	b_{dir}
	b_{turb}
	b_{ρ}
$\vdots$	$\vdots$
TME _n	b_{vel}
	b_{dir}
	b_{turb}
	b_{ρ}
Hiperparâmetros	max_depth
	learning_rate
	n_estimators
	subsample
	colsample_bytree
	min_child_weight
	γ
	α (regularização L1)
	λ (regularização L2)

Fonte: Do Autor (2026)

3.4.2.2 MLP: Espaço de Busca

O modelo MLP foi implementado utilizando a biblioteca PyTorch (Paszke *et al.*, 2019), com arquitetura *feed-forward* totalmente conexa, executado em modo de regressão com função de perda de erro quadrático médio (*mean squared error*). O treinamento foi realizado utilizando o otimizador Adam (Kingma; Ba, 2014) com taxa de aprendizado variável, aplicando normali-

zação em lote (*batch normalization*) entre camadas e regularização por *dropout* e decaimento de pesos. Todos os tensores (dados de entrada, pesos da rede e gradientes) foram alocados diretamente na memória da GPU pelo método `.to('cuda')`, aproveitando a computação paralela massiva para acelerar tanto a propagação direta quanto a retropropagação durante o treinamento. O espaço de busca definido para otimização de seus hiperparâmetros compreende os seguintes intervalos:

- Arquitetura de camadas ocultas: configuração variando de 1 a 3 camadas, com o número de neurônios por camada compreendido entre 16 e 128;
- Taxa de aprendizado (*learning_rate*): busca em escala logarítmica no intervalo de 10^{-5} a 10^{-2} ;
- Taxa de *dropout*: valores entre 0,0 (sem descarte) e 0,4;
- Função de ativação: utilização da função retificadora linear (ReLU) em todas as camadas ocultas;
- Decaimento de pesos (*weight_decay*): regularização L2 com busca logarítmica entre 10^{-6} e 10^{-2} ;
- Critério de parada precoce (*early stopping*): monitoramento com paciência (*patience*) entre 10 e 50 épocas e variação mínima (Δ_{min}) de 10^{-5} a 10^{-3} ;
- Limite de treinamento: fixado em 500 épocas.

A arquitetura variável do MLP resulta em cromossomos de comprimento não uniforme: redes com uma camada oculta requerem especificação de apenas um hiperparâmetro de neurônios, enquanto redes com três camadas requerem três especificações. Esta codificação de comprimento variável é naturalmente suportada pelo algoritmo NSGA-II por meio da utilização de genes condicionais, onde os genes correspondentes aos neurônios das camadas 2 e 3 são ignorados durante a construção do modelo caso o gene de número de camadas especifique arquitetura com menos de 2 ou 3 camadas, respectivamente.

A combinação destes hiperparâmetros determina o compromisso entre expressividade e generalização: arquitetura e taxa de aprendizado controlam a capacidade de aproximação de funções complexas; *dropout* e decaimento de pesos implementam regularização explícita;

parada precoce atua como regularização implícita baseada em validação. O NSGA-II otimiza simultaneamente estes hiperparâmetros junto à seleção de fontes e características, explorando configurações que minimizem erro preditivo (Equação 3.8) e complexidade (Equação 3.9). O algoritmo NSGA-II incorpora estes hiperparâmetros aos genes de acordo com a Tabela 3.5.

Tabela 3.5 – Exemplo de Cromossomo MLP

Gene	Característica
TME ₁	b_{vel}
	b_{dir}
	b_{turb}
	b_{ρ}
$\vdots$	$\vdots$
TME _n	b_{vel}
	b_{dir}
	b_{turb}
	b_{ρ}
Hiperparâmetros	Número de camadas ocultas
	Neurônios na camada 1
	$\vdots$
	Neurônios na camada n
	<i>learning_rate</i>
	<i>dropout</i>
	<i>weight_decay</i>
	<i>patience</i>
	Δ_{min}

Fonte: Do Autor (2026)

3.5 Métricas de Avaliação e Validação dos Modelos

A metodologia proposta é avaliada em dois níveis complementares: validação da robustez e convergência do processo de otimização evolutiva por meio de métricas específicas para algoritmos multiobjetivo; e análise de desempenho preditivo dos modelos selecionados da fronteira de Pareto por métricas de regressão.

3.5.1 Validação da Otimização Multiobjetivo

Para verificar a robustez do algoritmo NSGA-II, o processo de otimização foi executado cinco vezes de forma independente para cada modelo (XGBoost e MLP), utilizando sementes aleatórias distintas para inicialização da população. Esta replicação permite verificar se o algoritmo converge consistentemente para regiões similares do espaço de objetivos, independentemente das condições iniciais, ou se apresenta alta sensibilidade à inicialização. As fronteiras de Pareto obtidas nas diferentes execuções são comparadas utilizando duas métricas complementares:

- Hipervolume (HV): quantifica o volume do espaço de objetivos dominado por uma fronteira de Pareto em relação a um ponto de referência predefinido (While *et al.*, 2006; Zitzler *et al.*, 2003). Para um conjunto de soluções não-dominadas $\mathcal{P}$ em um problema de minimização com dois objetivos, o hipervolume é definido pela Equação 3.10:

$$HV(\mathcal{P}, \mathbf{r}) = \Gamma_2 \left(\bigcup_{p_i \in \mathcal{P}} [p_i; \mathbf{r}] \right), \quad (3.10)$$

onde $\Gamma_2(\cdot)$ denota a medida de Lebesgue bidimensional (área), $p_i = (f_1(p_i); f_2(p_i))$ representa as coordenadas da solução i no espaço de objetivos, $[p_i; \mathbf{r}]$ denota o hiper-retângulo delimitado pela solução p_i e o ponto de referência $\mathbf{r}$, sendo $\mathbf{r} = (r_1; r_2) = (1,0; 144)$, que é o ponto de referência escolhido. O valor $r_1 = 1,0$ corresponde ao erro máximo teoricamente possível para predição de fator de capacidade, enquanto $r_2 = 144$ representa a complexidade máxima do problema para o WTG “W301”, calculada como o produto do número de torres disponíveis pelo número total de variáveis que podem ser seleccionadas. Valores maiores de hipervolume indicam fronteiras que dominam regiões mais extensas do espaço de objetivos, sendo portanto preferíveis.

- Distância Geracional Invertida (IGD): avalia a proximidade entre uma fronteira aproximada e uma fronteira de referência, funcionando como indicador de convergência para aplicações de otimização multiobjetivo (Coello; Sierra, 2004). Diferentemente do hipervolume, que mede apenas a região dominada, o IGD quantifica a proximidade média entre cada ponto da fronteira de referência e o ponto mais próximo da fronteira aproximada. Para um conjunto de soluções não-dominadas $\mathcal{P}$ e uma fronteira de referência $\mathcal{P}^*$, o

IGD é definido pela Equação 3.11:

$$IGD(\mathcal{P}, \mathcal{P}^*) = \frac{1}{|\mathcal{P}^*|} \sum_{\mathbf{k} \in \mathcal{P}^*} d_{\mathbf{k}}, \quad (3.11)$$

onde $d_{\mathbf{k}}$ pode ser calculado pela Equação 3.12 e representa a distância euclidiana mínima entre o ponto $\mathbf{k} \in \mathcal{P}^*$ da fronteira de referência e todos os pontos $\mathbf{i} \in \mathcal{P}$ da fronteira avaliada:

$$d_{\mathbf{k}} = \min_{\mathbf{i} \in \mathcal{P}} \left(\sum_{j=1}^m \left(\mathcal{P}_{\mathbf{k}_j}^* - \mathcal{P}_{\mathbf{i}_j} \right)^2 \right)^{\frac{1}{2}}, \quad (3.12)$$

sendo $m = 2$ o número de funções objetivo, $\mathcal{P}_{\mathbf{k}_j}^*$ o valor do objetivo j para o ponto $\mathbf{k}$ da fronteira de referência, e $\mathcal{P}_{\mathbf{i}_j}$ o valor correspondente do objetivo j para o ponto $\mathbf{i}$ da fronteira avaliada. A fronteira de referência $\mathcal{P}^*$ é construída pela união de todas as soluções não-dominadas obtidas nas cinco execuções independentes, seguida de filtragem para reter apenas as soluções mutuamente não-dominadas. Valores menores de IGD indicam maior proximidade à fronteira de referência, refletindo tanto convergência quanto cobertura adequada.

A análise conjunta de hipervolume e IGD fornece compreensão complementar da qualidade das fronteiras obtidas: o hipervolume avalia a qualidade absoluta em termos de dominância, enquanto o IGD avalia a proximidade relativa entre execuções. Baixa variação destas métricas entre as cinco execuções indica convergência robusta e consistente do algoritmo.

3.5.2 Métricas de Desempenho dos Modelos

Após validação da convergência do processo evolutivo, a análise concentra-se na fronteira de Pareto obtida na execução que apresentou hipervolume mediano entre as cinco replicações, garantindo representatividade sem viés de seleção para casos extremos. As soluções desta fronteira são avaliadas mediante a média da validação cruzada para as seguintes métricas:

- Raiz do Erro Quadrático Médio (RMSE): conforme definido na Equação 3.8 para a função objetivo $f_1(x)$ durante a otimização, o RMSE quantifica a magnitude típica dos erros de predição. Esta métrica penaliza erros grandes de forma quadrática, sendo particularmente sensível a predições com grandes desvios e é reportada em fator de capacidade.

- Coeficiente de Determinação (R^2): quantifica a proporção da variância total da variável resposta explicada pelo modelo, calculado pela Equação 3.13:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3.13)$$

onde $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ representa a média dos valores observados no conjunto de teste. Valores de R^2 próximos a 1 indicam que o modelo explica grande parte da variabilidade dos dados, enquanto valores próximos a 0 sugerem baixo poder preditivo. Diferentemente do RMSE, que fornece erro em unidades absolutas, o R^2 oferece medida normalizada de qualidade de ajuste, facilitando comparações entre diferentes conjuntos de dados ou escalas de variáveis.

Para avaliar a convergência do algoritmo NSGA-II durante a execução principal, três aspectos são monitorados ao longo das 100 gerações:

1. Evolução do erro preditivo (Equação 3.8): acompanhamento da distribuição dos valores de RMSE de todos os indivíduos da população a cada geração, apresentando a faixa mínimo-máximo e a média. Este gráfico permite identificar a progressão da população em direção a soluções com menor erro de predição e verificar a estagnação ou melhoria contínua ao longo do processo evolutivo;
2. Evolução da complexidade (Equação 3.9): monitoramento da distribuição dos valores de complexidade (função objetivo $f_2 = N_f \times n_c$) ao longo das gerações, também apresentando faixa e média. Diferentemente de f_1 , que tende a diminuir monotonicamente, f_2 pode apresentar dinâmica mais complexa conforme o algoritmo explora diferentes regiões do espaço de soluções, alternando entre configurações mais simples e mais complexas;
3. Tamanho da fronteira de Pareto: contagem do número de soluções não-dominadas presentes em cada geração. O crescimento ou estabilização deste número ao longo das gerações indica o grau de diversidade mantido pelo algoritmo e pode revelar se múltiplas regiões distintas de *trade-off* estão sendo adequadamente exploradas.

Estes gráficos permitem diagnosticar possíveis problemas no processo evolutivo, como convergência prematura com a estagnação precoce de f_1 com baixa diversidade, perda de diversidade, ou exploração inadequada do espaço de objetivos, ou seja, ausência de progressão em

f_1 ou f_2 . A fronteira de Pareto obtida ao final da execução principal é analisada recorrendo a um conjunto integrado de visualizações que revelam diferentes aspectos das soluções:

1. Gráfico de dispersão no espaço de objetivos ($f_1 \times f_2$): visualização bidimensional apresentando o RMSE *versus* o número de variáveis para todas as soluções da fronteira de Pareto, identificadas por cores que indicam o número de fontes de dados utilizadas. O modelo de referência, baseado em média por faixas, é incluído como ponto de comparação, representado por um marcador distinto. Este gráfico permite identificar visualmente o *trade-off* entre precisão, número de variáveis e número de fontes de dados utilizadas;
2. Distribuição cumulativa de seleção de torres: gráfico de barras empilhadas apresentando, para cada torre de medição disponível, a contagem do número de vezes que a respectiva torre foi selecionada nas soluções da fronteira, estratificada pelo número total de fontes utilizadas em cada solução. Cada segmento da barra corresponde às soluções com N fontes, permitindo identificar se determinadas torres são preferencialmente selecionadas em configurações com poucas fontes ou apenas em configurações complexas. Esta visualização revela a hierarquia de relevância das torres: torres com alta contagem total são consideradas mais informativas pelo algoritmo, enquanto a distribuição entre os segmentos indica se a torre é essencial ou complementar;
3. Distribuição cumulativa da seleção de características por torre: visualização bidimensional em grade cartesiana onde cada posição representa uma variável meteorológica específica de uma torre. Cada posição é representada por círculos concêntricos cujo raio total é proporcional ao número de vezes que aquela característica específica foi selecionada nas soluções da fronteira, com camadas coloridas concêntricas estratificando esta contagem pelo número de fontes utilizadas nas respectivas soluções. Este diagrama permite identificar simultaneamente quatro dimensões de informação: qual torre fornece a variável, qual variável meteorológica é selecionada, com que frequência esta combinação específica torre-variável aparece nas soluções e com quais distribuições de configuração esta seleção ocorre. A interpretação integrada revela não apenas quais variáveis são globalmente importantes, mas também se determinadas características são essenciais mesmo em modelos minimalistas ou apenas complementares em modelos complexos.

A análise conjunta destas visualizações permite compreender não apenas o *trade-off* quantitativo entre erro e complexidade, mas também os mecanismos qualitativos pelos quais o

algoritmo constrói este *trade-off*: quais torres são priorizadas, quais variáveis de cada torre são selecionadas e como esses padrões de seleção mudam ao longo da fronteira de Pareto conforme diferentes níveis de complexidade são aceitos. Adicionalmente, a comparação entre as fronteiras obtidas para XGBoost e MLP permite identificar se os dois algoritmos convergem para estratégias similares de seleção de fontes e características ou se exploram diferentes regiões do espaço de soluções.

Para aplicação prática, uma solução específica deve ser selecionada dentre a fronteira de Pareto. A seleção é realizada por meio do método do ponto utópico, que identifica a solução mais próxima do cenário ideal, mas inatingível, de erro nulo e complexidade mínima. Matematicamente, os valores dos objetivos f_1 e f_2 de todas as soluções da fronteira são primeiro normalizados utilizando a transformação linear da Equação 3.6, onde X é f_1 ou f_2 . Sendo $f_{1\min} = 0$, indicando erro inexistente; $f_{1\max} = 0,1$, como o limite de aceitabilidade prática para o RMSE; $f_{2\min} = 1$, onde a complexidade indica o uso de apenas uma variável; por fim, $f_{2\max}$ é a complexidade máxima do problema para o WTG em questão. Após normalização, o ponto utópico é definido como $\mathbf{u} = (0, 0)$ no espaço normalizado, representando simultaneamente erro nulo e complexidade mínima. A solução selecionada é aquela que minimiza a distância euclidiana ao ponto utópico. Esta abordagem pondera implicitamente os dois objetivos de forma equivalente após normalização, priorizando soluções que oferecem bom equilíbrio entre acurácia preditiva e simplicidade operacional. A solução selecionada tende a localizar-se na região de transição da fronteira onde ganhos adicionais de acurácia começam a exigir aumentos desproporcionais de complexidade, chamada de região de retornos decrescentes, adequando-se a aplicações onde tanto o desempenho preditivo quanto a parcimônia do modelo são igualmente valorizados.

3.6 Avaliação Final e Cálculo de perdas

A partir dos modelos selecionados pelo critério do ponto utópico para cada algoritmo (XGBoost e MLP), é realizado um retreinamento utilizando o conjunto completo de dados filtrados com a configuração ótima identificada (torres de medição, características meteorológicas e hiperparâmetros). Diferentemente do processo de validação cruzada utilizado durante a otimização, onde os dados são repetidamente particionados em k subconjuntos, o retreinamento final utiliza uma única divisão de 70% dos dados para conjunto de treinamento e 30% para conjunto de teste.

Este procedimento de retreinamento serve a três propósitos principais: primeiro, permite que o modelo final seja treinado com maior volume de dados comparado a cada *fold* individual da validação cruzada, potencialmente melhorando a qualidade do ajuste; segundo, fornece estimativa independente do desempenho do modelo em dados completamente isolados do processo de otimização, validando que a configuração selecionada generaliza adequadamente; terceiro, gera o modelo persistente que será utilizado para previsões operacionais e análises subsequentes.

As métricas de desempenho (RMSE e R^2) são calculadas sobre o conjunto de teste e comparadas com aquelas obtidas durante a validação cruzada no processo de otimização. Concordância entre estas métricas valida a robustez da configuração selecionada, enquanto divergências significativas podem indicar sensibilidade à partição específica dos dados ou *overfitting* durante a otimização. Adicionalmente, o modelo de referência baseado em média por faixas é retreinado sobre a mesma divisão treino-teste, permitindo comparação direta entre o desempenho do modelo otimizado e do modelo base sob condições idênticas de avaliação. O desempenho preditivo dos modelos finais é visualizado através de gráficos de dispersão dos valores observados *versus* valores preditos de fator de capacidade, além disso, a linha de identidade ($y = x$) é incluída como referência: pontos próximos a esta linha indicam previsões precisas, enquanto desvios revelam regiões do espaço de entrada onde o modelo apresenta viés sistemático ou maior variância.

Uma vez treinados e validados, os modelos finais selecionados são aplicados ao conjunto completo de dados históricos do aerogerador, incluindo períodos previamente removidos durante a etapa de filtragem descrita na Seção 3.3. Esta aplicação aos dados não filtrados permite estimar o fator de capacidade que o aerogerador teria alcançado caso estivesse operando em condições ótimas. A comparação entre a energia potencial estimada pelo modelo (correspondente à operação ótima) e a energia efetivamente gerada (registrada nos dados SCADA) quantifica as perdas energéticas atribuíveis a condições operacionais subótimas. A geração relativa ao potencial disponível, expressa como a razão entre a energia efetivamente gerada e a energia total esperada, é calculada pela Equação 3.14:

$$E_l = \frac{E_g}{E_e}, \quad (3.14)$$

em que E_l é a geração relativa ao potencial, E_e é a energia total esperada e E_g é a energia total gerada no período.

4 RESULTADOS E DISCUSSÃO

A aplicação da metodologia de otimização multiobjetivo proposta é conduzida em duas etapas com objetivos complementares: primeiro, aplicação detalhada e abrangente ao aerogerador W301, incluindo análise exploratória dos dados, validação da convergência do algoritmo NSGA-II, caracterização das fronteiras de Pareto obtidas e avaliação de desempenho dos modelos selecionados; segundo, aplicação ao aerogerador W310, avaliando a capacidade de generalização da abordagem para diferentes equipamentos e condições operacionais do mesmo parque eólico.

4.1 Resultados para o Aerogerador W301

O estudo de caso do aerogerador W301 permite investigar em profundidade o comportamento do processo de otimização evolutiva, identificar padrões de seleção de fontes de dados e de características meteorológicas e validar a robustez das soluções obtidas por meio de múltiplas execuções independentes. A análise detalhada deste caso fornece compreensão dos mecanismos pelos quais os algoritmos XGBoost e MLP equilibram acurácia preditiva e complexidade operacional, revelando quais combinações de torres e variáveis meteorológicas são consistentemente priorizadas e como os *trade-offs* manifestam-se ao longo das fronteiras de Pareto.

4.1.1 Caracterização dos Dados

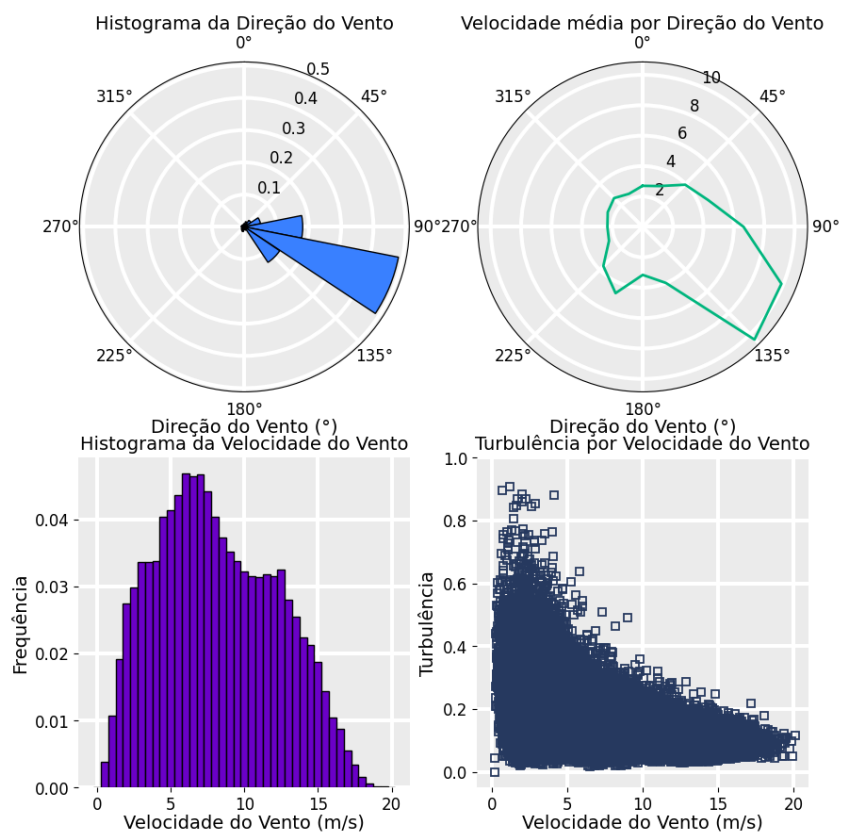
Antes de proceder à modelagem, é essencial compreender as características dos dados disponíveis e o impacto das transformações aplicadas durante o pré-processamento. Primeiro, executa-se uma análise exploratória das variáveis meteorológicas das torres de medição e das variáveis operacionais SCADA, avaliando distribuições, correlações e relevância física, com o objetivo de identificar o regime de vento característico da região, efeitos de micro-escala presentes e redundâncias entre variáveis que justificam a redução do espaço de características. Também é realizada a quantificação do impacto da filtragem de dados descrita na Seção 3.3, visualizando como os critérios aplicados eliminam regiões específicas do espaço operacional correspondentes a paradas, limitações e degradação de desempenho. A integração destas aná-

lises estabelece a base de dados limpa e representativa utilizada para treinamento, validação e teste dos modelos de aprendizado de máquina.

4.1.1.1 Avaliação de Variáveis Disponíveis

O regime de vento regional tem característica unidirecional: a maior parte do tempo o vento sopra de uma única direção e os ventos de alta velocidade também são majoritariamente vindos da direção predominante, conforme ilustrado nos gráficos de direção do vento da Figura 4.1. Isso evidencia um efeito esteira que afeta diretamente o WTG, ao considerar os outros aerogeradores da região, vistos na Figura 3.2, ou seja, é essencial considerar o impacto da turbulência T_i na operação do aerogerador. A Figura 4.1 também ilustra que as maiores intensidades de turbulência ocorrem em ventos de baixa velocidade; contudo, seu impacto é mais relevante em ventos de média velocidade (5 - 10m/s), pois, além dessa faixa de vento apresentar maior frequência de ocorrência, está associada a maiores variações na energia potencial, conforme a região 2 da curva de potência esperada na Figura 2.3.

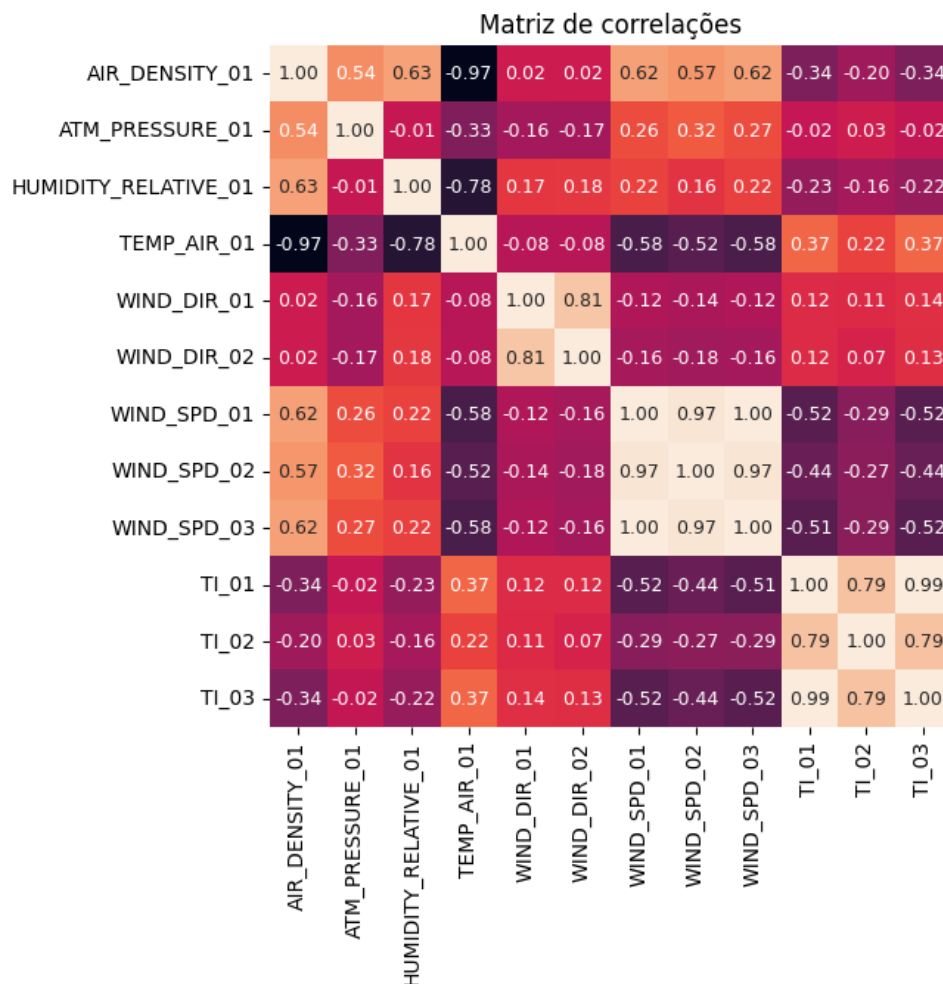
Figura 4.1 – Vento característico da TME T103.



Fonte: Do Autor (2026)

Como exemplo, tem-se a TME mais próxima, chamada “T103”, onde os instrumentos são numerados a partir do topo, ou seja, WIND_SPD_1 é o anemômetro de topo da TME. Isso permite descartar todos os instrumentos secundários (em alturas mais baixas) ao avaliar a alta correlação entre os instrumentos na Figura 4.2. Além disso, é visível a relação entre AIR_DENSITY_01 e TEMP_AIR_01, HUMIDITY_RELATIVE_01 e ATM_PRESSURE_01 na Equação 3.2, tornando possível removê-las das variáveis de entrada do modelo. Todas as TMEs obtiveram o mesmo comportamento, portanto, os dados de TME a serem utilizados para a modelagem serão apenas WIND_SPD_1, WIND_DIR_1, AIR_DENSITY_01 e TI_01.

Figura 4.2 – Correlação entre variáveis da TME T103.

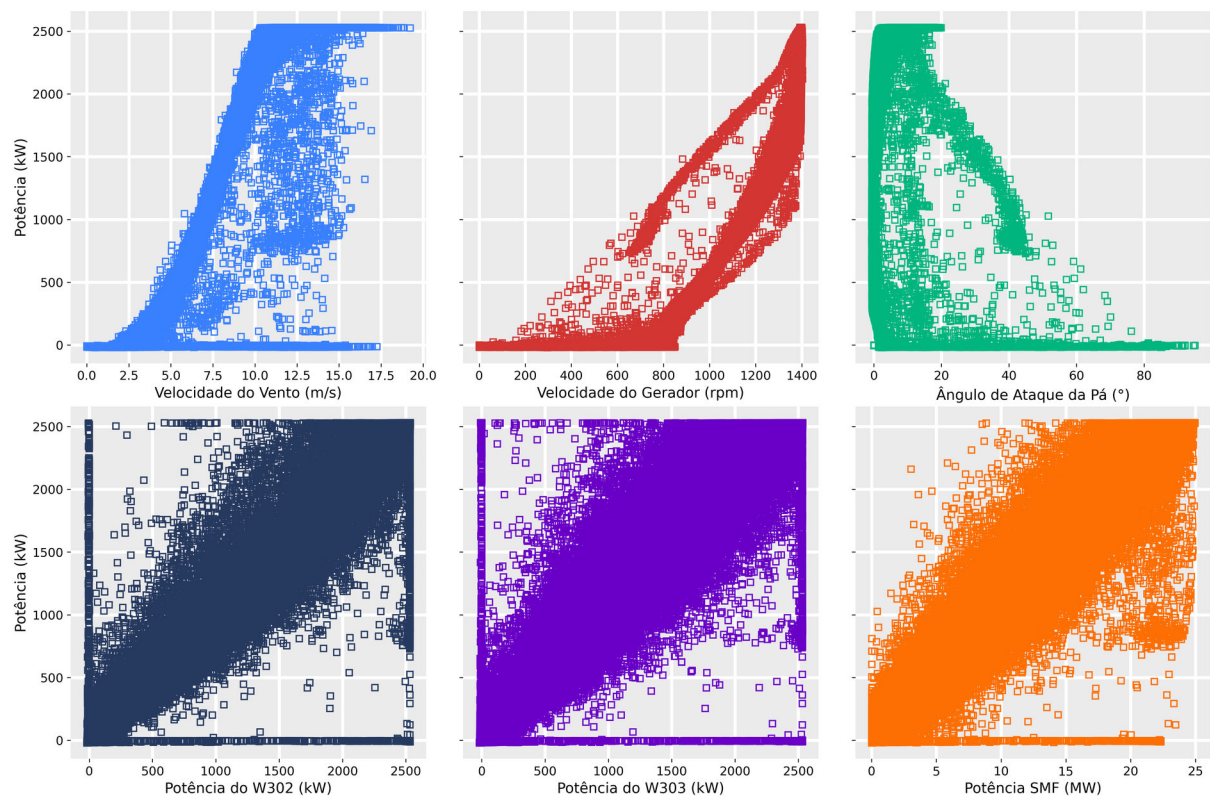


Fonte: Do Autor (2026)

No gráfico de vento por potência da Figura 4.3, é possível reconhecer pontos que representam paradas de máquina ou limitações durante a operação, avaliando as relações entre as variáveis e a potência. A partir da relação entre o ângulo de ataque da pá e a potência na Figura 4.3, verifica-se uma região dos dados que representam as paradas e limitações, já que a

região de operação da máquina tem um formato em “C” entre, aproximadamente, -5° e 20° . Da mesma forma, a operação normal na relação entre a velocidade do gerador e a potência deveria se manter em 0kW entre 0 e 800rpm, onde se mantém até 300kW, crescendo a potência e velocidade do gerador em uma reta até atingir 1500kW e 1400rpm, a partir do qual a potência segue até a nominal, mas a velocidade do gerador fica constante. Além disso, a relação da potência do aerogerador avaliado e seus vizinhos mais próximos, bem como em relação ao parque como um todo, devem ser lineares, já que características regionais se mostram presentes, permitindo destacar instantes nos quais a potência do aerogerador é muito menor que as referências. Pode-se observar a projeção dessas amostras a serem descartadas entre as relações de cada uma das variáveis e a potência.

Figura 4.3 – Características operacionais do W301.



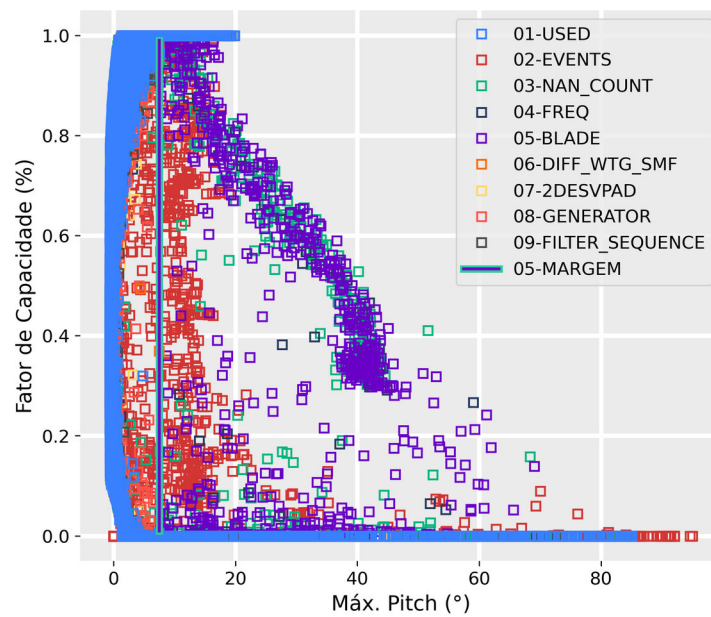
Fonte: Do Autor (2026)

4.1.1.2 Tratamento dos Dados

A aplicação dos filtros apresentados na Seção 3.3 resultou nos dados apresentados na Figura 4.7. Pode-se observar como essa filtragem se apresentou na relação da curva de vento por

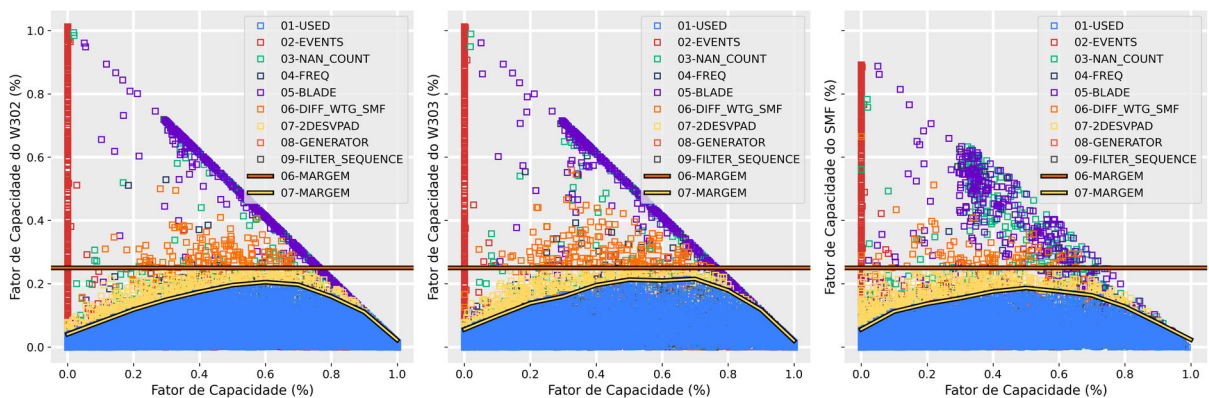
potência. Foram coletadas 51.564 amostras no total, deste conjunto, 65% das amostras foram utilizadas para o treino e teste dos modelos (01-USED), restando 18.047 que não foram aproveitadas. O impacto visual da limpeza de dados pode ser observado em Figura 4.4, Figura 4.5 e Figura 4.6. Cada gráfico isola a variável de controle do respectivo filtro em relação à potência, permitindo identificar as regiões de operação anômala que foram removidas e a linha de corte para a remoção de tais dados.

Figura 4.4 – Visualização do Filtro 05-BLADE do W301.



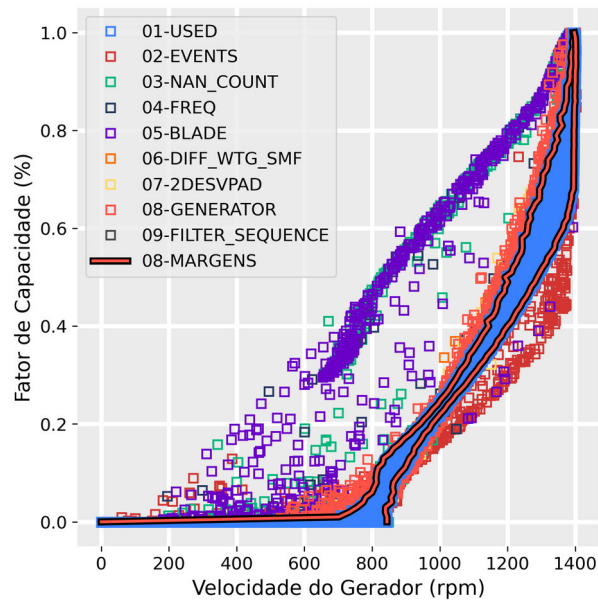
Fonte: Do Autor (2026)

Figura 4.5 – Visualização dos Filtros 06-DIFF_WTG_SMF e 07-2DESVPAD do W301.



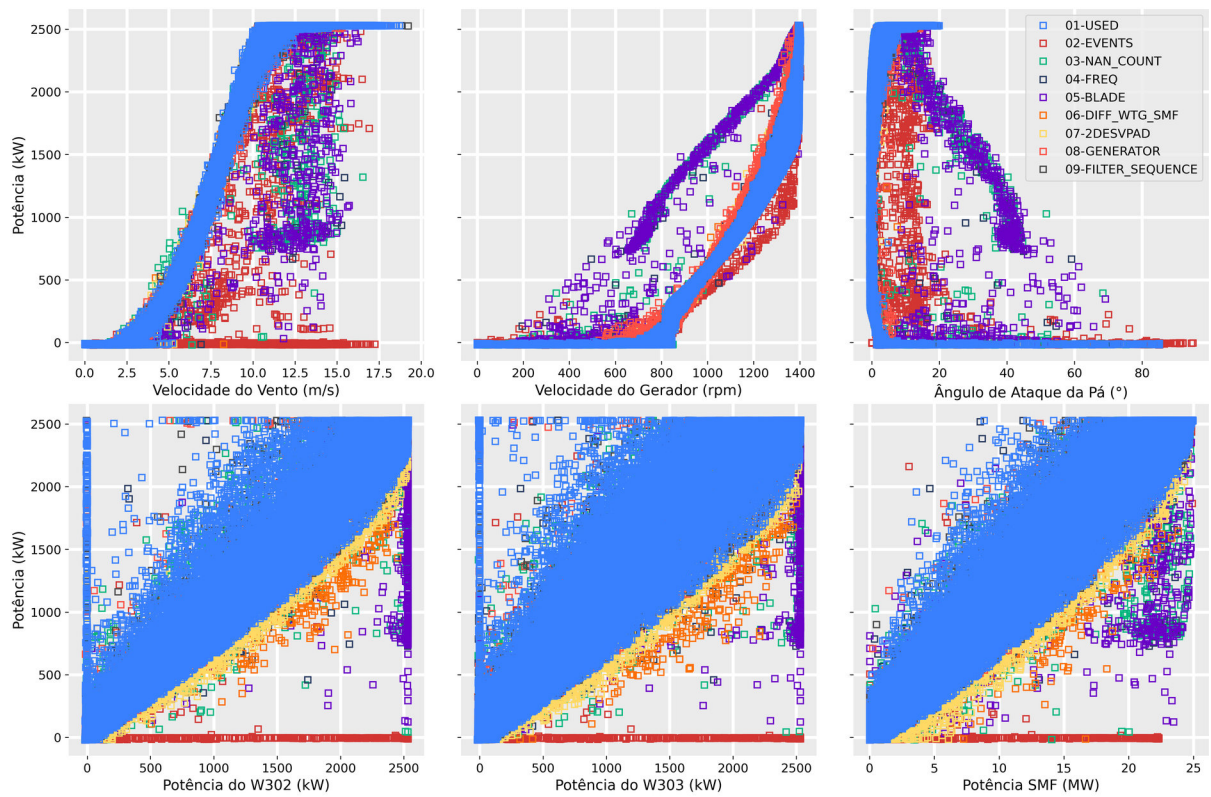
Fonte: Do Autor (2026)

Figura 4.6 – Visualização do Filtro 08-GENERATOR do W301.



Fonte: Do Autor (2026)

Figura 4.7 – Dados filtrados do W301.



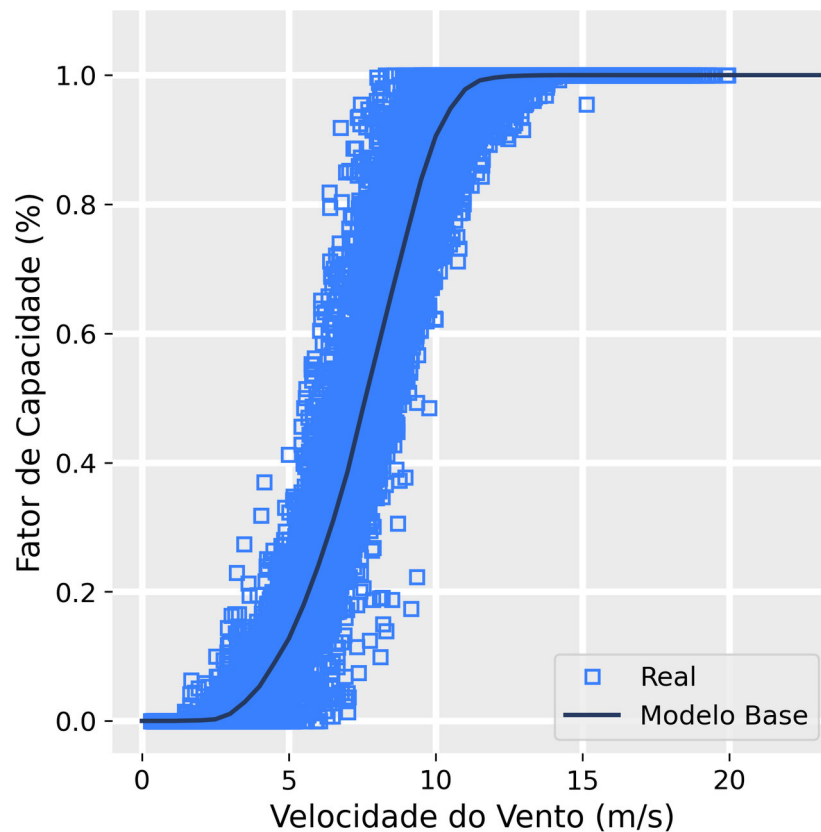
Fonte: Do Autor (2026)

Nota: As diferentes cores representam as implementações dos filtros e a cor azul os dados filtrados que são utilizados para treino e teste dos modelos de predição.

4.1.2 Modelo Base

Com o objetivo de prever o FC do WTG W301, foram aplicados os algoritmos descritos na Seção 3.4. Para o método base descrito na Subseção 3.4.1, foram utilizadas faixas de 0,5m/s da TME T103 e calculada a média do fator de capacidade por faixa, com os dados filtrados. Em seguida, o valor foi interpolado linearmente a cada 0,1m/s, obtendo a curva “Modelo Base” da Figura 4.8. Este modelo será utilizado como base comparativa para a validação da capacidade dos modelos de aprendizagem de máquina.

Figura 4.8 – Modelo Base do W301.



Fonte: Do Autor (2026)

4.1.3 Modelagem NSGA-II

A aplicação de algoritmos de aprendizado de máquina otimizados por meio do NSGA-II representa um avanço significativo em relação ao modelo de referência. Além de capturar relações não-lineares entre variáveis meteorológicas e fator de capacidade, a abordagem multiobjetivo permite explorar sistematicamente o compromisso entre acurácia preditiva e complexidade

operacional, identificando configurações que equilibram desempenho superior com requisitos práticos de instrumentação. Conforme descrito na Seção 3.4, o algoritmo realiza simultaneamente a seleção de torres de medição, escolha de características meteorológicas de cada torre, e otimização de hiperparâmetros dos modelos XGBoost e MLP.

Para o aerogerador W301, cinco execuções independentes do NSGA-II foram realizadas para cada modelo, permitindo validar a robustez da convergência e identificar padrões consistentes de seleção. Os resultados são apresentados progressivamente: validação por múltiplas execuções e seleção da execução principal; evolução dos objetivos ao longo das gerações; caracterização detalhada das fronteiras de Pareto; e desempenho dos modelos finais com estimativa de perdas energéticas.

4.1.3.1 Validação por meio de Múltiplas Execuções

Para avaliar a robustez do algoritmo NSGA-II, cinco execuções independentes foram realizadas para cada modelo (XGBoost e MLP) com sementes aleatórias distintas. A Tabela 4.1 apresenta as métricas de convergência (HV e IGD) e as características do modelo selecionado pelo critério do ponto utópico em cada execução.

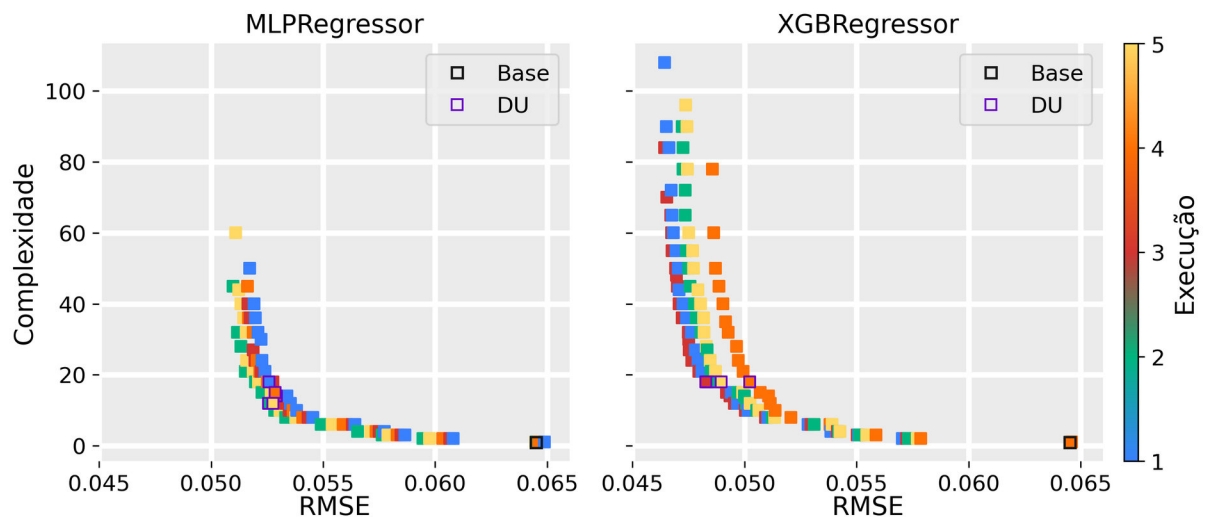
Tabela 4.1 – Métricas de validação e características dos modelos selecionados pelo critério do ponto utópico nas cinco execuções independentes do NSGA-II.

Modelo	Exec.	HV	IGD	Solução do Ponto Utópico		
				Compl.	RMSE	R ²
MLP	1	135,5	0,41	18	0,053	0,982
	2	135,6	0,89	12	0,053	0,982
	3	135,5	0,86	14	0,053	0,982
	4	135,5	0,56	15	0,053	0,982
	5	135,6	0,24	12	0,053	0,982
XGB	1	136,2	0,33	18	0,048	0,985
	2	136,1	0,46	18	0,049	0,984
	3	136,2	0,68	18	0,048	0,985
	4	135,9	1,46	18	0,050	0,983
	5	136,1	0,66	18	0,049	0,984

Fonte: Do Autor (2026)

A análise das métricas de convergência revela alta consistência entre as execuções: para o MLP, o hipervolume varia apenas 0,1 entre as cinco execuções, enquanto para o XGBoost a variação é de 0,3, determinada principalmente pela execução 4, que se comporta como *outlier*; desconsiderando essa execução, as quatro restantes variam apenas 0,1. Os valores de IGD são mais variáveis, refletindo diferenças na cobertura e distribuição das soluções ao longo da fronteira, mas todas as execuções convergem para fronteiras de qualidade comparável em termos de dominância, o que é visível na Figura 4.9. A baixa variação nos valores de RMSE e R^2 dos pontos utópicos indica que, independentemente das diferenças na estrutura das fronteiras obtidas, o algoritmo identifica consistentemente soluções com desempenho preditivo muito similar na região de melhor compromisso entre acurácia e complexidade. Notavelmente, todas as cinco execuções do XGBoost convergem para soluções com complexidade idêntica ($f_2 = 18$), sugerindo que esta configuração representa um equilíbrio robusto e consistente para este algoritmo. Em contraste, o MLP apresenta maior variabilidade na complexidade dos pontos utópicos (12 a 18), indicando que múltiplas configurações distintas podem oferecer *trade-offs* comparáveis entre acurácia e simplicidade para este modelo.

Figura 4.9 – Fronteiras de Pareto do W301.



Fonte: Do Autor (2026)

Nota: DU é o ponto de menor distância ao ponto utópico de cada execução.

A Tabela 4.2 apresenta a frequência de seleção de cada característica nos pontos utópicos das cinco execuções, revelando padrões consistentes e diferenças entre os algoritmos. Para ambos os modelos, as torres T503 e T103 são priorizadas, com velocidade do vento selecionada em todas as execuções. A direção de T103 também é altamente consistente, enquanto direção de

Tabela 4.2 – Contagem de execuções por variável selecionada

Algoritmo	TME	Velocidade	Direção	Turbulência	Densidade
MLP	T503	5	4	1	2
	T103	5	5	5	1
	T109	2	0	0	0
XGB	T403	4	4	1	0
	T503	5	0	0	0
	T103	5	5	4	0
	T109	1	0	0	1

Fonte: Autor (2026)

T503 aparece apenas no MLP. A turbulência de T103 é frequentemente selecionada, refletindo sua relevância para capturar variabilidade do escoamento e efeitos de micro-escala. Diferenças notáveis entre algoritmos incluem: XGBoost utiliza T403, enquanto MLP nunca a seleciona; densidade do ar apresenta baixa frequência em ambos, sugerindo contribuição marginal insuficiente para justificar o aumento de complexidade.

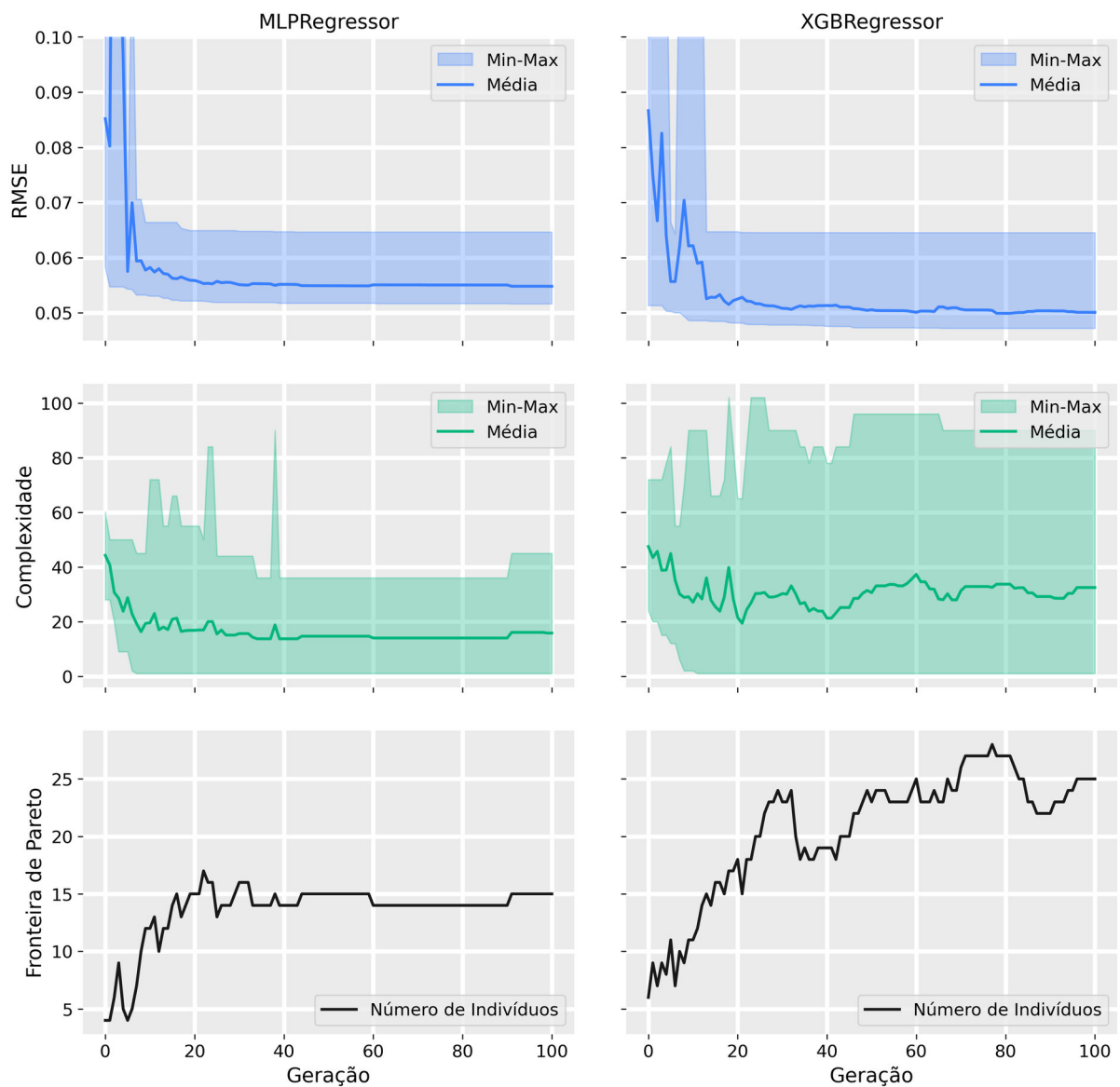
Baseado na análise de robustez das múltiplas execuções, a execução 4 do MLP e a execução 2 do XGBoost foram selecionadas para análise detalhada. A escolha de execuções foi baseada na proximidade ao hipervolume mediano, ao invés de selecionar aquelas com máximo HV ou mínimo IGD, isto garante que a análise concentra-se em casos representativos do comportamento típico do algoritmo, evitando viés de seleção para execuções excepcionalmente favoráveis ou desfavoráveis. Esta abordagem fornece compreensão mais realista do desempenho esperado da metodologia em aplicações práticas, onde apenas uma execução seria tipicamente realizada devido a restrições de custo computacional.

4.1.3.2 Evolução do Processo de Otimização

A Figura 4.10 apresenta a evolução das funções objetivo (f_1 e f_2) e o número de indivíduos não-dominados (fronteira de Pareto) ao longo das 100 gerações para a execução selecionada. Observa-se que ambos os modelos convergem rapidamente para regiões de baixo erro nas primeiras gerações, enquanto a fronteira de Pareto continua a se alterar ao longo do processo evolutivo, evidenciando que o refinamento ocorre predominantemente pela diversificação

das configurações de características e não pela redução adicional do erro. Este comportamento indica que o NSGA-II explora efetivamente o espaço de compromisso entre acurácia e complexidade mesmo após a estabilização do erro. Para o MLP, a fronteira estabiliza em torno de 15 indivíduos, com menor dispersão na complexidade, sugerindo convergência para configurações mais compactas. O XGBoost, por sua vez, exibe crescimento contínuo da fronteira até aproximadamente 27 indivíduos, acompanhado de maior amplitude na complexidade, o que reflete maior capacidade de exploração do espaço de soluções não-dominadas.

Figura 4.10 – Evolução do treinamento do W301.



Fonte: Do Autor (2026)

4.1.3.3 Caracterização da Fronteira de Pareto

A fronteira de Pareto obtida ao final da execução selecionada representa o conjunto de soluções não-dominadas que definem o *trade-off* entre acurácia preditiva e complexidade para o aerogerador W301. A análise desta fronteira é conduzida por meio do conjunto integrado de visualizações descrito na Seção 3.5, permitindo caracterizar simultaneamente a distribuição das soluções no espaço de objetivos, os padrões de seleção de fontes e variáveis e a justificativa física para as hierarquias observadas.

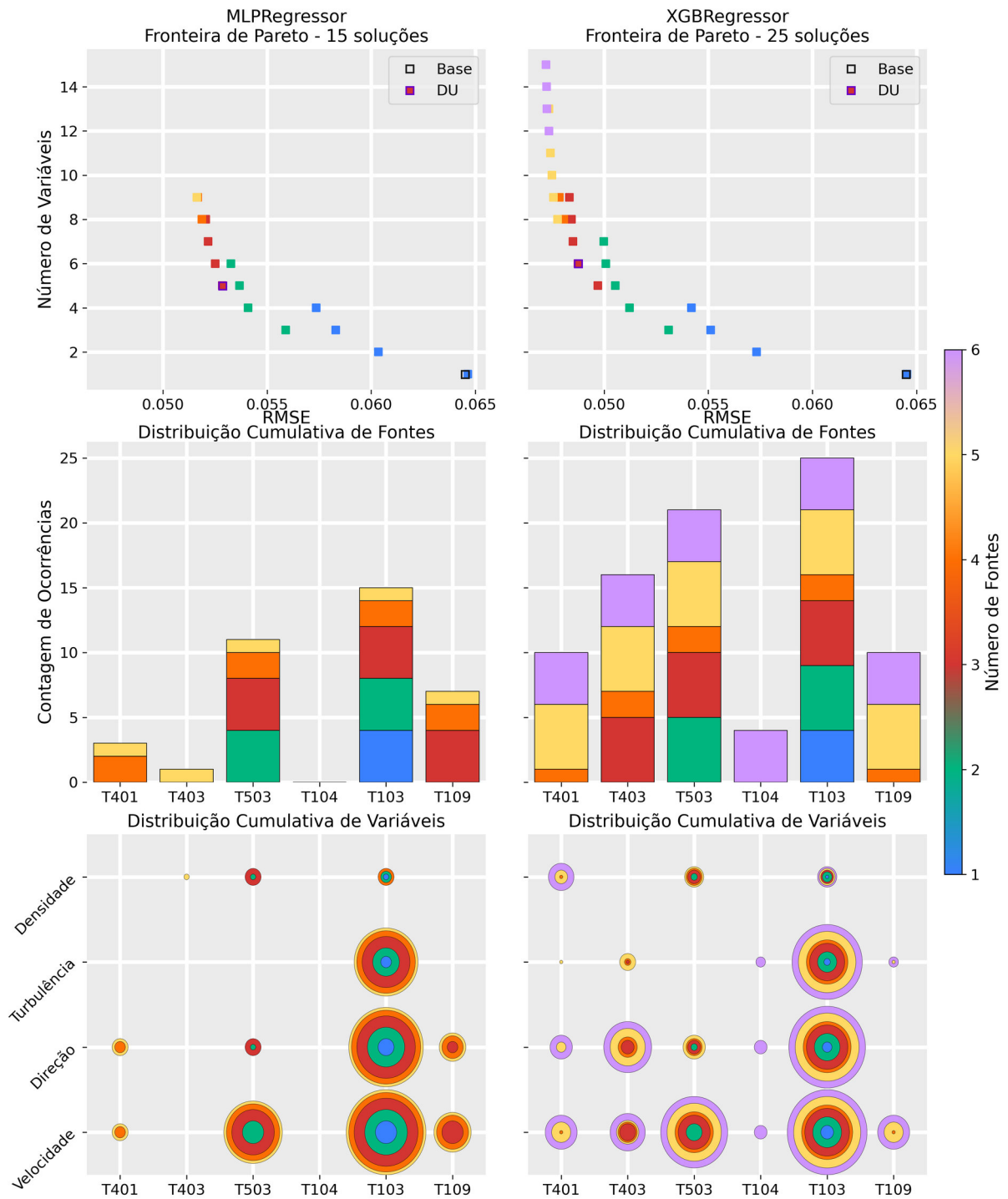
A Figura 4.11 apresenta, para cada modelo, o gráfico de dispersão no espaço de objetivos segregando o número de fontes utilizadas da quantidade total de características (variáveis), a distribuição cumulativa de seleção de fontes e a distribuição cumulativa de seleção de características por torre. O XGBoost resultou em uma fronteira com 25 soluções, enquanto o MLP produziu 15, refletindo a maior capacidade de exploração do espaço de soluções não-dominadas observada durante a evolução. Em ambos os casos, a fronteira exibe comportamento de retornos decrescentes: a redução no RMSE torna-se marginal após a inclusão de aproximadamente três fontes de dados, indicando que ganhos adicionais de acurácia exigem aumentos desproporcionais de complexidade. A solução de menor distância ao ponto utópico (DU), identificada na Figura 4.11, localiza-se na região de transição desta curva, operacionalizando o compromisso entre desempenho preditivo e parcimônia operacional conforme definido na Subseção 3.4.2.

A dinâmica de seleção das fontes revela uma hierarquia geoespacial consistente entre os modelos. As fontes T103 e T503 são selecionadas prioritariamente em ambos, constituindo o núcleo mínimo de informação necessário. A divergência manifesta-se a partir da terceira fonte: o MLP prioriza a T109, enquanto o XGBoost opta pela T403. Nas configurações com quatro fontes, o XGBoost incorpora preferencialmente a T401, com variações ocasionais para a T109; o MLP converge para a T401. A convergência entre os modelos ocorre na configuração com cinco fontes, quando o MLP inclui a T403, igualando a composição do XGBoost. A configuração com todas as seis fontes disponíveis foi atingida exclusivamente pelo XGBoost.

No que diz respeito às variáveis individuais, a distribuição cumulativa por torre evidencia que velocidade do vento e as componentes de direção (U e V) apresentam as maiores frequências de seleção em ambos os modelos, consolidando-se como preditores primários independentemente do nível de complexidade aceito. As variáveis de turbulência (T_i) se apresentam como complemento à capacidade preditiva, enquanto a densidade do ar traz ganhos marginais

mínimos. O XGBoost exibe maior presença de camadas externas nas variáveis de turbulência e densidade do ar, indicando que seu ganho de desempenho em configurações complexas advém da captura de nuances micrometeorológicas distribuídas entre múltiplas fontes. O MLP, por sua vez, concentra-se em um conjunto mais restrito de variáveis-chave, com uso da variável de densidade especialmente limitado. Em nível de pares fonte-variável, observa-se que mesmo fontes de seleção universal contribuem apenas com parte de suas variáveis. A T503 exemplifica esse comportamento: sua velocidade do vento é selecionada em todas as soluções, ao passo que sua turbulência e sua densidade do ar são raramente ou nunca retidas. O padrão é fisicamente coerente, pois, de uma fonte posicionada a montante, interessa sobretudo a magnitude e a direção do vento incidente, enquanto a turbulência determinante para o desempenho do W301 é aquela desenvolvida localmente nas imediações do aerogerador, capturada pela T103, principal fonte de turbulência retida pelos modelos. A densidade do ar, por variar lentamente na escala espacial do complexo e manter-se fortemente correlacionada entre as torres, mostra-se redundante quando medida em múltiplas fontes, o que explica sua baixa frequência de seleção em todos os pares.

Figura 4.11 – Fronteira de Pareto e distribuição de fontes e características do W301.

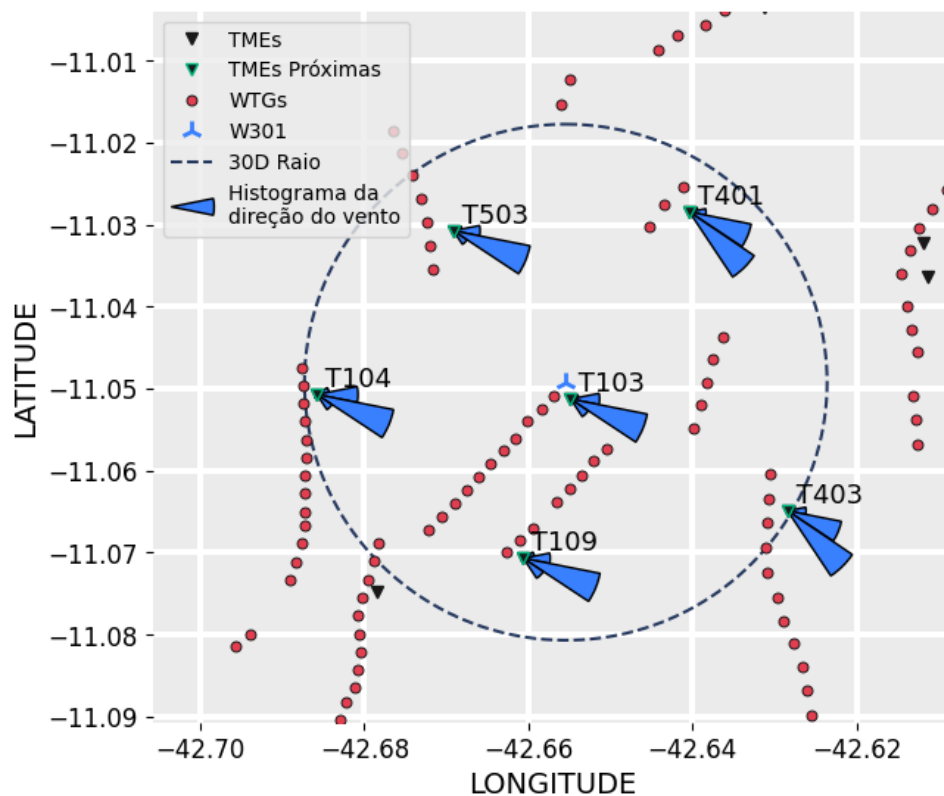


Fonte: Do Autor (2026)

A análise espacial das TME, integrada aos histogramas de direção predominante do vento apresentados na Figura 4.12, oferece justificativa física para a hierarquia de seleção observada. O regime de ventos é marcadamente de Sudeste (SE), definindo claramente as posições de montante e jusante em relação ao W301. A T103, em proximidade imediata ao aerogerador

e a T503, posicionada em linha de fluxo que intercepta o complexo a montante, configuram-se como pontos de medição críticos, o que explica sua seleção universal. A divergência na terceira fonte reflete estratégias distintas de captura do escoamento: a T109, preferida pelo MLP, encontra-se alinhada à T103 sob o regime de Sudeste, capturando variações laterais do escoamento; a T403, preferida pelo XGBoost, atua como estação de referência de vento livre na extremidade de entrada do fluxo, sem obstruções significativas por outros aerogeradores. Esses padrões sugerem que configurações preditivamente eficazes tendem a incorporar informações sobre o trajeto completo do vento no complexo: condições de entrada, evolução da esteira e turbulência interna, e condições locais próximas ao alvo.

Figura 4.12 – Mapa com direções predominantes das TMEs do W301.



Fonte: Do Autor (2026)

A comparação entre os dois algoritmos evidencia desempenho superior do XGBoost ao longo de toda a fronteira: o modelo atinge, com apenas duas fontes, valores de RMSE inferiores aos obtidos pelo MLP mesmo em suas configurações com maior número de fontes, indicando maior eficiência na extração de informação a partir de um número reduzido de variáveis. Três fatores ajudam a explicar essa diferença. Primeiro, o conjunto de dados analisado tem natureza tabular: cada amostra corresponde a uma única linha, composta por variáveis meteorológicas

heterogêneas (velocidade, direção, turbulência e densidade do ar das diferentes torres), sem que exista entre as colunas uma ordem ou estrutura espacial ou temporal a ser explorada, ao contrário do que ocorre em imagens ou séries temporais. Embora os dados tenham origem em uma série temporal, o uso de uma única janela de dez minutos como entrada, sem variáveis defasadas que conectem instantes consecutivos, descarta essa estrutura temporal e configura o problema como tabular. Nesse formato, adicionalmente favorecido pelo número reduzido de características informativas e pela dimensão moderada do conjunto, modelos baseados em árvores de decisão com *gradient boosting* mantêm desempenho consistentemente superior ao de redes neurais (Grinsztajn; Oyallon; Varoquaux, 2022; Shwartz-Ziv; Armon, 2022). Segundo, a resposta do fator de capacidade às variáveis meteorológicas é marcadamente não-suave, com transições abruptas entre os regimes da curva de potência e saturação nos extremos; modelos baseados em árvores aproximam naturalmente funções irregulares por partes, enquanto redes neurais tendem a soluções mais suaves (Grinsztajn; Oyallon; Varoquaux, 2022). Terceiro, dentro do orçamento fixo do NSGA-II, o treinamento de cada MLP candidato depende fortemente da inicialização, da arquitetura e da taxa de aprendizado, ao passo que o XGBoost é menos sensível a hiperparâmetros e converge de forma mais estável; essa diferença é coerente com a maior variabilidade da complexidade do ponto utópico observada para o MLP entre as cinco execuções (Tabela 4.1).

Cabe ressaltar, contudo, que a vantagem associada à melhor representação de funções irregulares decorre, em parte, da própria forma como os dados são observados. O fenômeno físico subjacente é contínuo no espaço e no tempo, e o caráter descontínuo dos dados resulta do registro em médias de janelas de dez minutos, e não de uma descontinuidade intrínseca do processo. Dessa forma, o melhor desempenho do XGBoost não invalida a abordagem por redes neurais, mas indica que esta requer refinamento adicional, por exemplo na arquitetura, na regularização ou na incorporação de informação temporal, para representar adequadamente a continuidade do escoamento e, potencialmente, superar os modelos baseados em árvores.

4.1.3.4 Modelos Seleccionados e Desempenho Preditivo

Identificadas as soluções ótimas pelo critério do ponto utópico, procede-se ao retreinamento dos modelos seleccionados sobre o conjunto completo de dados filtrados, conforme descrito na Seção 3.6. A Figura 4.13 apresenta os gráficos de dispersão entre os valores ob-

servados e preditos de fator de capacidade para os conjuntos de treino e teste, comparando o modelo de referência com o MLP e o XGBoost otimizados. As métricas de desempenho e as configurações selecionadas são consolidadas nas Tabela 4.3 e Tabela 4.4.

Ambos os modelos otimizados convergiram para configurações com três TME, sendo T103 e T503 selecionadas por ambos, confirmando a hierarquia observada na fronteira de Pareto e divergindo na terceira fonte: o MLP incorporou a T109, enquanto o XGBoost optou pela T403. Em ambos os casos, a T103 concentra o maior número de variáveis selecionadas, incluindo velocidade, direção e turbulência, coerentemente com sua proximidade imediata ao aerogerador alvo. A ausência da variável de densidade em ambos os modelos é consistente com o padrão observado na distribuição cumulativa da fronteira de Pareto, onde esta variável apresentou baixa frequência de seleção.

A superioridade dos modelos otimizados em relação ao modelo de referência é evidente tanto nas métricas quanto na maior concentração de pontos ao longo da linha de identidade ($y = x$). O modelo de referência, com RMSE de 0,064 e $R^2 = 0,973$, apresenta dispersão consideravelmente maior em toda a faixa de valores, refletindo a limitação inerente de uma abordagem baseada em média por faixas de velocidade do vento. O MLP e o XGBoost otimizados alcançaram RMSE de 0,053 e 0,049, com R^2 de 0,982 e 0,984, respectivamente, operando com complexidade de 15 e 18, equivalente a 5 e 6 variáveis distribuídas entre 3 fontes. A dispersão residual observada nos gráficos concentra-se predominantemente na região de fator de capacidade médio, entre 20% e 80%, onde a variabilidade do recurso eólico traduz-se em grandes variações de geração. A concordância entre os conjuntos de treino e teste, sem indícios de colapso de generalização, valida que as configurações selecionadas pelo ponto utópico generalizam adequadamente para dados não vistos durante a otimização.

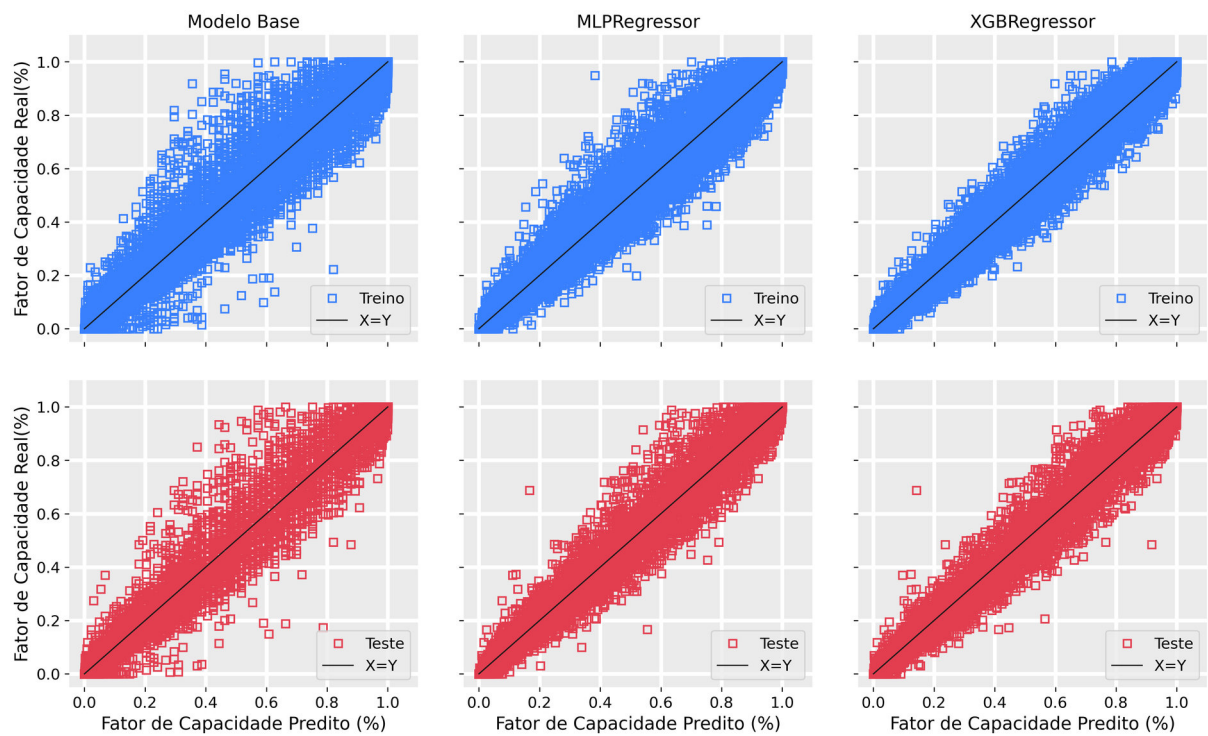
Tabela 4.3 – Comparação de Acurácia entre modelos do W301

Modelo	TMEs	Variáveis	Complexidade	RMSE	R^2
Base	1	1	1	0,064	0,973
MLP	3	5	15	0,053	0,982
XGB	3	6	18	0,049	0,984

Fonte: Do Autor (2026)

Nota: Complexidade calculada conforme Equação 3.9.

Figura 4.13 – Distribuição das predições para os modelos escolhidos do W301.



Fonte: Do Autor (2026)

Tabela 4.4 – Variáveis Escolhidas dos Modelos do W301

Modelo	TME	Velocidade	Direção	Turbulência	Densidade
Base	T103	X			
MLP	T503	X			
	T103	X	X	X	
	T109	X			
XGB	T403	X	X		
	T503	X			
	T103	X	X	X	

Fonte: Do Autor (2026)

A seleção da variável de direção pelo XGBoost merece atenção do ponto de vista físico. Nas soluções identificadas pelo critério do ponto utópico, a direção do vento é incorporada nas fontes T403 e T103: a primeira, posicionada na extremidade de entrada do fluxo predominante de Sudeste, fornece a direção das condições de vento livre, não perturbadas pela esteira dos aerogeradores; a segunda, em proximidade imediata ao W301, captura a direção local após

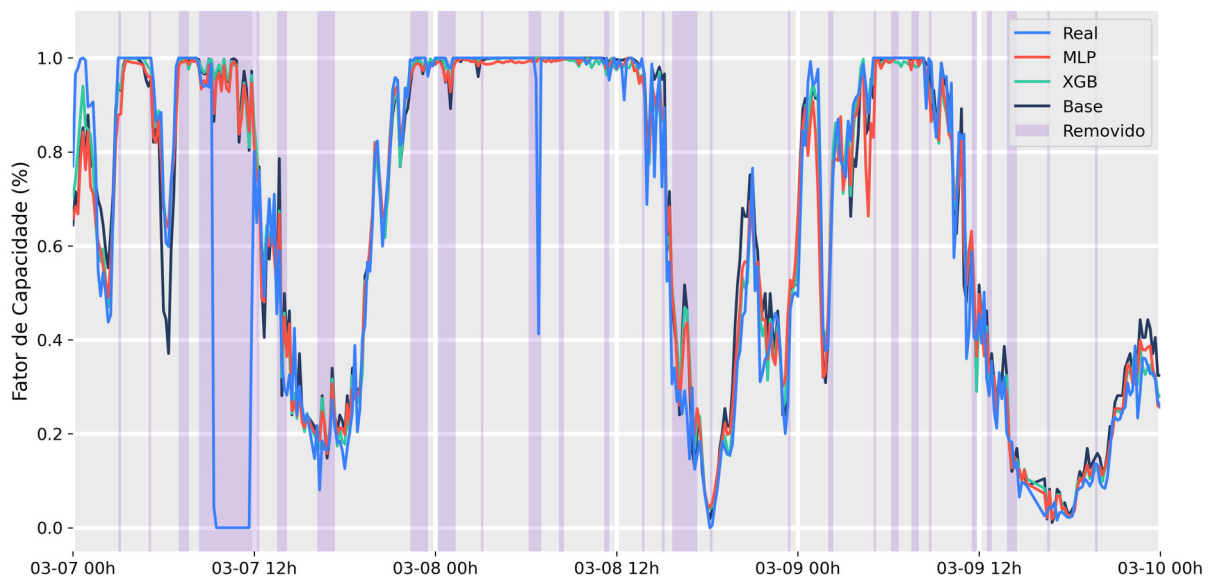
a evolução do escoamento ao longo do complexo. A ausência de direção na T503, apesar de esta torre ser selecionada por ambos os modelos, é fisicamente coerente: posicionada a montante e sob influência do fluxo já modificado pela operação dos aerogeradores a jusante, a direção medida neste ponto carrega menor poder discriminativo sobre as condições de chegada ao W301 do que a velocidade, que integra o déficit energético acumulado ao longo da esteira. Dessa forma, o XGBoost constrói implicitamente uma representação do trajeto completo do escoamento, as condições de entrada via T403, evolução da esteira via velocidade em T503 e condições locais via T103.

4.1.3.5 Estimativa de Perdas Energéticas

A estimativa de perdas energéticas apresentada nesta seção não constitui o objetivo central da pesquisa, voltada à seleção otimizada de fontes e variáveis e à modelagem do potencial de geração, mas um objetivo específico destinado a demonstrar uma aplicação prática dos modelos desenvolvidos. Validado o desempenho preditivo dos modelos finais, estes são aplicados ao conjunto completo de dados históricos do aerogerador, incluindo os períodos removidos durante a etapa de filtragem, conforme descrito na Seção 3.6. Nos períodos filtrados, o dado operacional existe mas não foi utilizado no treinamento: eventos como paradas por falha, limitação de potência ou manutenção resultam em geração real inferior ao potencial meteorológico e é exatamente nesta discrepância que reside a perda energética estimada. A Figura 4.14 ilustra este comportamento em uma janela de três dias, onde os intervalos destacados em roxo correspondem a períodos removidos do treinamento. Observa-se um evento de indisponibilidade completa imediatamente antes de “03-07 12h”, com geração real nula enquanto os modelos estimam capacidade de geração positiva, e um evento de limitação entre “03-08 00h” e “03-08 12h”, caracterizado por queda abrupta da potência real enquanto as previsões permanecem elevadas.

Ao acumular as diferenças entre energia estimada e energia gerada ao longo de todo o período histórico, obtém-se a geração relativa ao potencial disponível, apresentada na Tabela 4.5. Nos dados utilizados para treinamento e teste, modelos não enviesados tendem a produzir erros que se cancelam ao longo do tempo, de modo que a contribuição líquida deste subconjunto para a perda acumulada é marginal. A perda estimada é, portanto, majoritaria-

Figura 4.14 – Predição de Fator de Capacidade durante um dia de operação com falhas do W301.



Fonte: Do Autor (2026)

mente determinada pelo comportamento dos modelos nos períodos removidos, onde a geração real diverge sistematicamente do potencial meteorológico.

Os três modelos convergem para estimativas próximas, entre 93,55% e 94,32%, indicando que aproximadamente 6% da energia potencial disponível não foi convertida no período analisado. Essa proximidade decorre de a maior parte da energia potencial ser acumulada em períodos de operação normal, nos quais as condições de vento situam-se dentro da distribuição de treinamento e os três modelos, incluindo o de referência, reproduzem com alta acurácia a mesma relação dominante da curva de potência, produzindo estimativas muito parecidas. Como a energia gerada é a mesma medição em todos os casos, apenas o potencial estimado varia e suas diferenças restringem-se aos períodos removidos, que respondem por parcela reduzida do total. Os ganhos de acurácia obtidos pela otimização manifestam-se sobretudo na predição instantânea, ao passo que, no agregado de longo prazo, em que predomina a relação comum da curva de potência, seu efeito é modesto, o que explica a proximidade observada inclusive em relação ao modelo base. A divergência entre modelos reflete diferenças na extrapolação para condições fora da distribuição de treinamento: nos períodos de parada ou limitação, cada modelo estima um potencial ligeiramente distinto, resultando em perdas acumuladas diferentes. Esta sensibilidade à extrapolação é inerente a qualquer modelo de aprendizado de máquina treinado sobre dados de operação normal e deve ser considerada na interpretação dos resultados.

A convergência das estimativas indica que o valor agregado de perda é robusto à escolha do modelo, conferindo confiabilidade ao diagnóstico, enquanto o ganho da otimização se traduz em maior fidelidade das predições instantâneas, base para a futura decomposição das perdas por evento.

Tabela 4.5 – Energia gerada relativa ao potencial do W301.

Modelo	Geração*
Base	93,55%
MLP	94,32%
XGB	93,64%

Fonte: Do Autor (2026)

**A geração indica a porcentagem da energia total disponível*

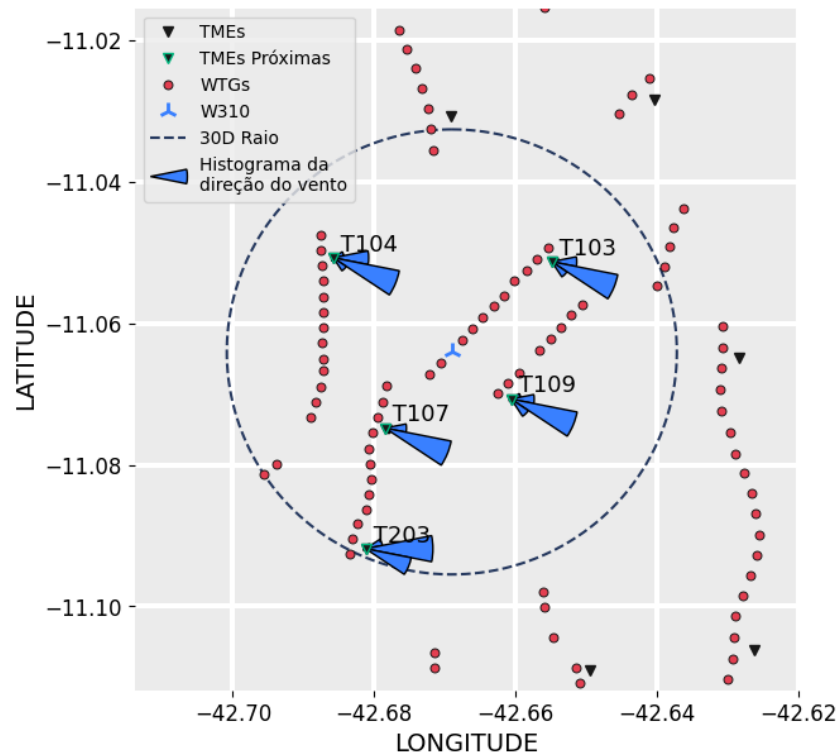
A métrica E_l definida na Equação 3.14 permite decompor temporalmente estas perdas, identificando períodos ou condições meteorológicas onde a discrepância entre potencial e geração é mais acentuada. Esta informação tem aplicação direta no planejamento de manutenção, na negociação de contratos de fornecimento de energia e na identificação de oportunidades de otimização operacional do parque eólico.

4.2 Resultados para o Aerogerador W310

A aplicação da metodologia ao aerogerador W310 valida a capacidade de generalização em contexto operacional distinto. Diferentemente de W301, que opera livre de interferências aerodinâmicas significativas, W310 está sujeito a efeito de esteira pronunciado de turbinas posicionadas a montante, resultando em condições de vento caracterizadas por velocidades reduzidas e turbulência intensificada, como visto na Figura 4.15. Apesar de compartilhar o mesmo regime climático regional e utilizar infraestrutura de medição parcialmente coincidente, as diferenças de micro-escala entre os dois equipamentos permitem avaliar se a metodologia adapta-se apropriadamente às particularidades operacionais específicas de cada aerogerador.

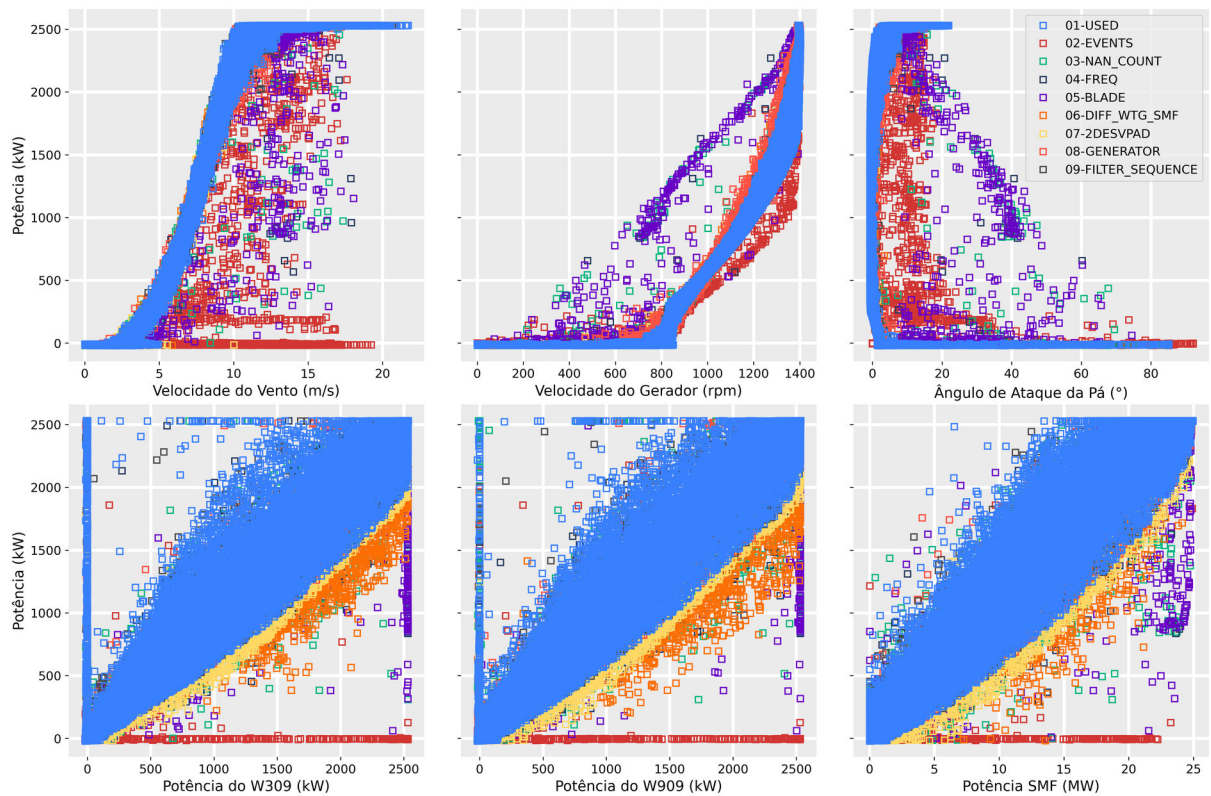
Os filtros apresentados na Seção 3.3 foram aplicados aos dados do W310 e apresentados visualmente na Figura 4.16, resultando em 63,5% de amostras úteis para treinamento, validação e teste. Ao utilizar a TME mais próxima “T109”, pode-se obter o modelo base visto na Figura 4.17.

Figura 4.15 – Mapa com direções predominantes das TMEs do W310.



Fonte: Do Autor (2026)

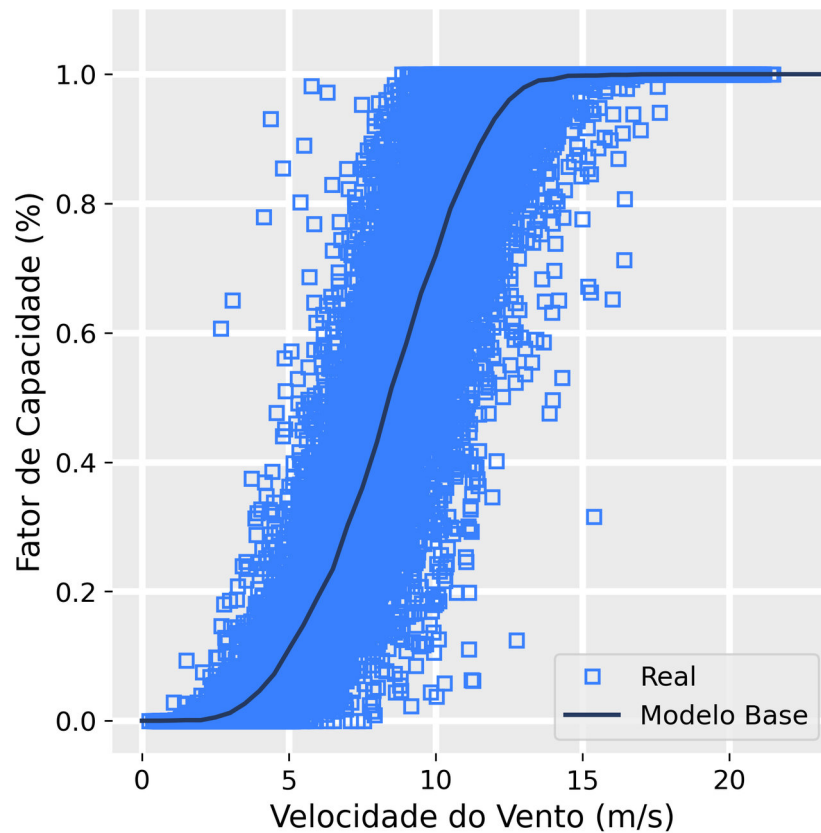
Figura 4.16 – Dados filtrados do W310.



Fonte: Do Autor (2026)

Nota: As diferentes cores representam as implementações dos filtros e a cor azul os dados filtrados que são utilizados para treino e teste dos modelos de predição.

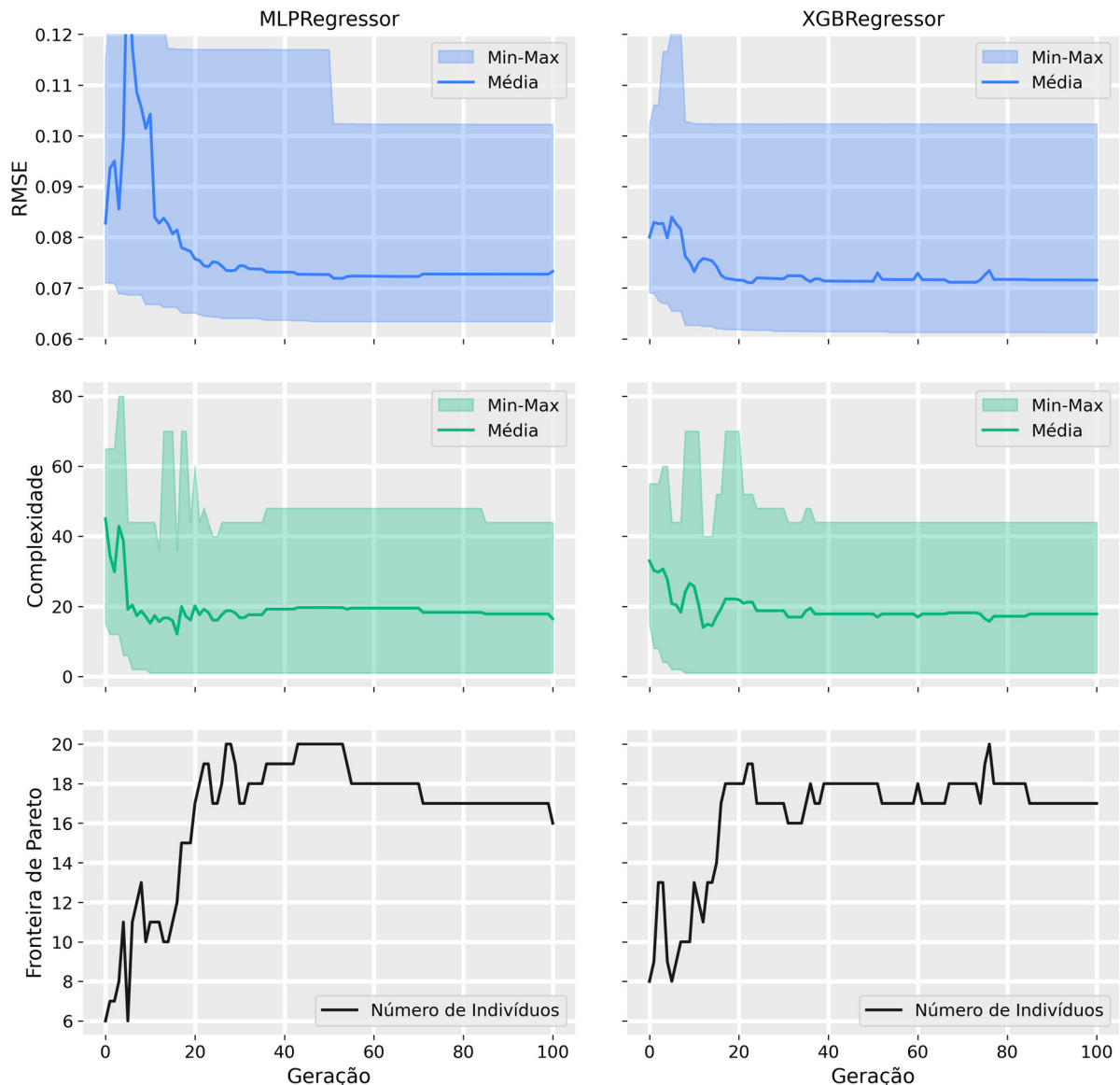
Figura 4.17 – Modelo Base do W310.



Fonte: Do Autor (2026)

Para o W310, foi realizada uma única execução do algoritmo NSGA-II, com as mesmas configurações mostrados na Subseção 3.4.2, para cada modelo (XGBoost e MLP), cujos resultados são apresentados na mesma sequência adotada para W301, facilitando comparação direta entre os casos. A Figura 4.18 apresenta a evolução das funções objetivo e do tamanho da fronteira de Pareto ao longo das 100 gerações. O padrão geral repete o observado para o W301: convergência rápida do RMSE médio nas primeiras gerações, seguida de expansão contínua da fronteira de Pareto ao longo de grande parte do processo evolutivo, indicando que o refinamento ocorre predominantemente pela diversificação das configurações de características. Ambos os modelos estabilizam a fronteira em torno de 16 a 17 indivíduos, com crescimento mais acentuado nas primeiras 20 gerações e flutuações posteriores que refletem a exploração contínua de regiões distintas do espaço de *trade-off*.

Figura 4.18 – Evolução do treinamento do W310.



Fonte: Do Autor (2026)

A fronteira de Pareto obtida para o W310 é apresentada na Figura 4.19, com 16 soluções para o MLP e 17 para o XGBoost, portanto, fronteiras de tamanho similar, contrastando com a assimetria observada no W301. O comportamento de retornos decrescentes é novamente verificado: a redução no RMSE torna-se marginal após a inclusão de aproximadamente três fontes, e o ponto de distância mínima ao utópico (DU) localiza-se consistentemente nessa região de transição para ambos os modelos.

Um resultado de particular relevância é a ausência completa da T203 nas soluções de ambas as fronteiras. Embora esta torre esteja dentro do raio de 30D e registre condições de vento livre, sua direção predominante é de Leste, distinta do regime de Sudeste que caracteriza

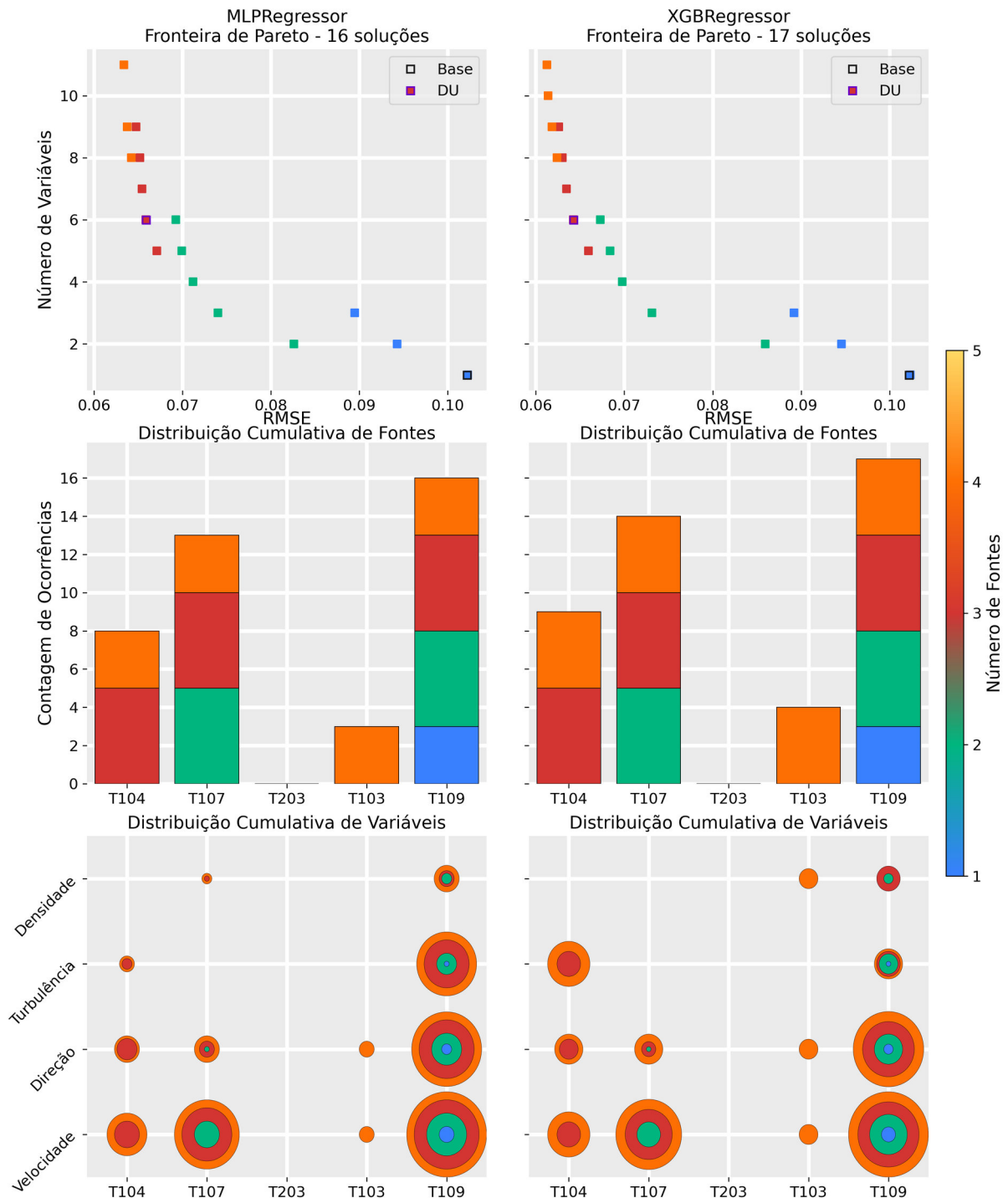
o escoamento no W310. Esta discordância entre o regime local da T203 e o regime do aerogerador alvo a torna redundante em relação às demais fontes disponíveis, e o NSGA-II a exclui sistematicamente, demonstrando que o algoritmo não apenas seleciona as fontes mais informativas, mas também descarta aquelas cuja representatividade meteorológica é inadequada para o contexto específico do aerogerador.

A hierarquia de seleção de fontes revela padrões fisicamente coerentes com a posição do W310 no interior do parque, como pode ser observado na Figura 4.15. A T109, torre mais próxima ao aerogerador, é a mais selecionada em ambos os modelos, sendo incorporada desde as configurações mais simples. A T104 é a segunda escolha, posicionada no lado de saída do fluxo em relação ao W310, enquanto a T107, situada linearmente na mesma fileira do aerogerador e em condições de vento mais livre que a T109, completa a configuração. Esta estrutura de seleção é análoga à identificada para o W301, onde o algoritmo também construiu uma representação do trajeto completo do escoamento, adaptada às condições específicas de cada aerogerador.

A distribuição cumulativa de variáveis evidencia estratégias distintas entre os modelos na captura dos efeitos de esteira. Para a turbulência, o MLP seleciona T_i da T109, privilegiando a caracterização das condições de vento no momento de impacto sobre o aerogerador. O XGBoost, por sua vez, opta por T_i da T104, capturando o estado do escoamento após a interação com o W310, uma perspectiva complementar que reflete como o vento foi modificado pela operação da turbina. A T107 é selecionada exclusivamente pela variável de velocidade em ambos os modelos, consistente com seu papel de referência de intensidade do vento em condições menos perturbadas.

As configurações identificadas pelo critério do ponto utópico para ambos os modelos convergiram para três TME e seis variáveis, resultando em complexidade 18, configuração idêntica entre MLP e XGBoost, ao contrário do W301, onde havia assimetria. As métricas de desempenho e as variáveis selecionadas são apresentadas nas Tabela 4.6 e Tabela 4.7. Assim como discutido sobre a distribuição cumulativa, o ponto utópico reflete a priorização de escolhas de cada modelo: enquanto o MLP utiliza a T_i de T109, o XGBoost opta pela de T104.

Figura 4.19 – Fronteira de Pareto e distribuição de fontes e características do W310.

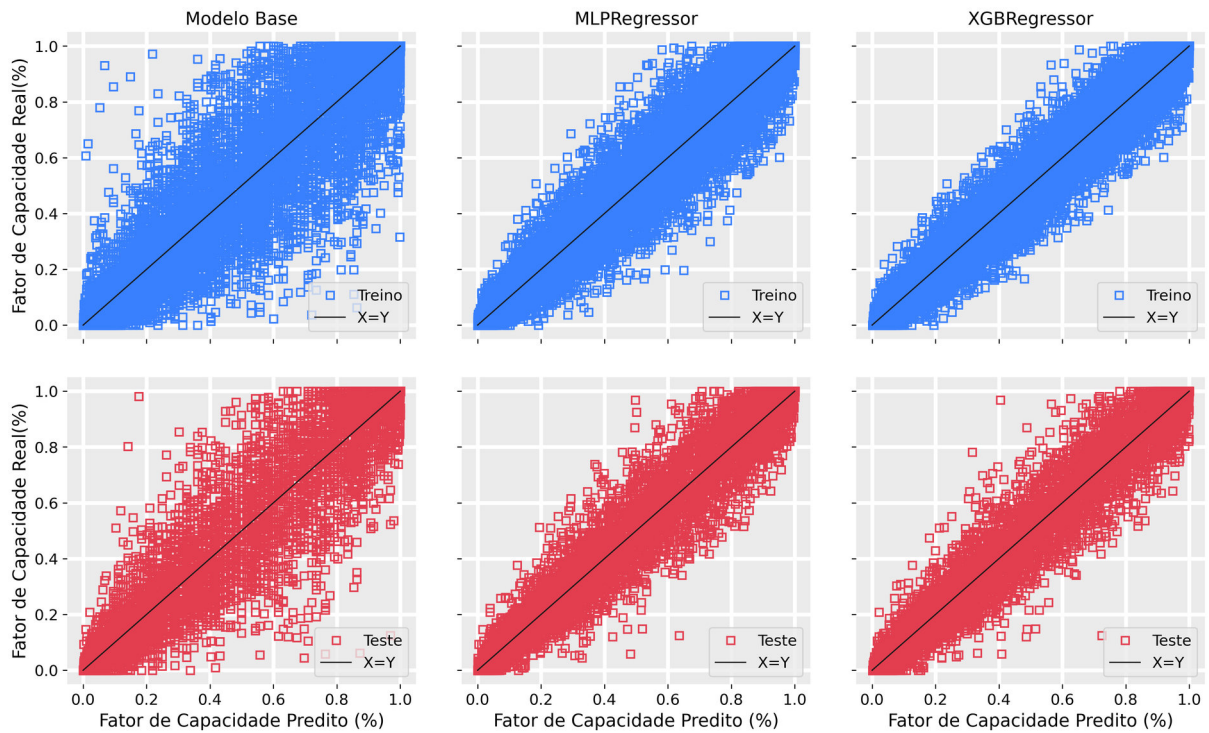


Fonte: Do Autor (2026)

O ganho de desempenho em relação ao modelo de referência é substancialmente maior que o observado para o W301. O modelo base apresenta RMSE de 0,102 e $R^2 = 0,933$, refletindo a maior dificuldade de modelar um aerogerador sob efeito de esteira pronunciado com apenas uma variável de velocidade. Os modelos otimizados reduzem o RMSE para 0,066 (MLP) e

0,064 (XGBoost), com R^2 de 0,972 e 0,973, respectivamente. Estas reduções relativas de erro são superiores às obtidas para o W301, evidenciando que a metodologia agrega valor proporcionalmente maior em condições operacionais mais complexas. A Figura 4.20 confirma a melhoria qualitativa na aderência à linha de identidade, com redução expressiva da dispersão sistemática presente no modelo base, especialmente nas faixas intermediárias do fator de capacidade.

Figura 4.20 – Distribuição das predições para os modelos escolhidos do W310.



Fonte: Do Autor (2026)

Tabela 4.6 – Comparação de Acurácia entre modelos do W310

Modelo	TMEs	Variáveis	Complexidade	RMSE	R^2
Base	1	1	1	0,102	0,933
MLP	3	6	18	0,066	0,972
XGB	3	6	18	0,064	0,973

Fonte: Do Autor (2026)

Nota: Complexidade calculada conforme Equação 3.9.

Tabela 4.7 – Variáveis Escolhidas dos Modelos do W310

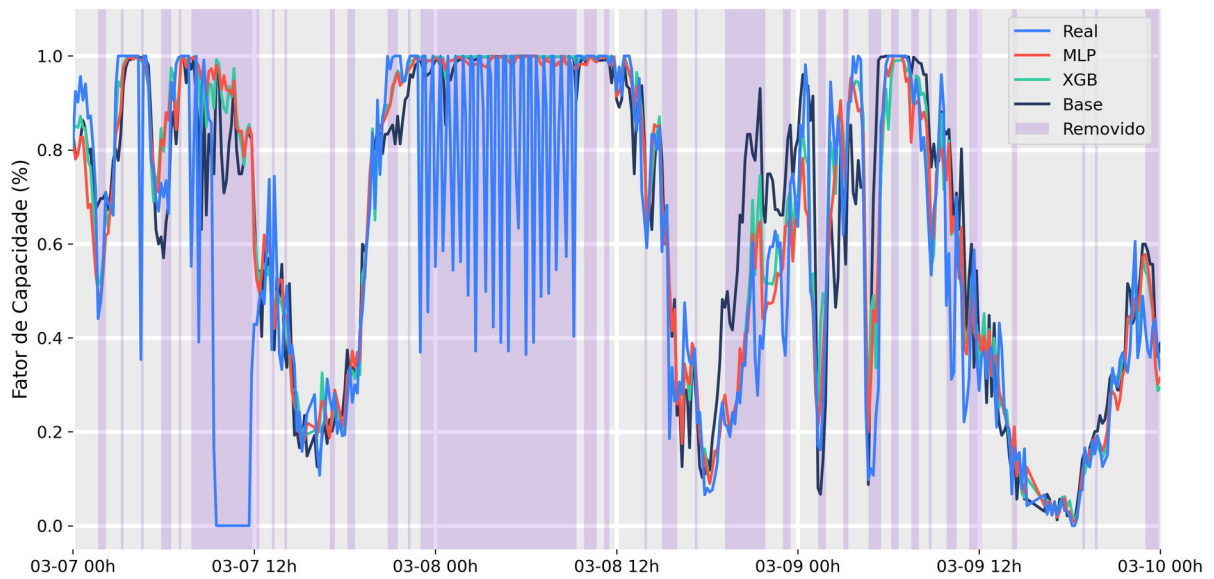
Modelo	TME	Velocidade	Direção	Turbulência	Densidade
Base	T109	X			
MLP	T104	X	X		
	T107	X			
	T109	X	X	X	
XGB	T104	X	X	X	
	T107	X			
	T109	X	X		

Fonte: Do Autor (2026)

Aplicados ao conjunto completo de dados históricos, os modelos finais permitem estimar o potencial de geração nos períodos removidos durante a filtragem. A Figura 4.21 ilustra o comportamento dos modelos em uma janela de três dias com elevada concentração de eventos anômalos. O período entre “03-08 00h” e “03-08 12h” evidencia um padrão de falhas sequenciais curtas, caracterizado por paradas e retomadas sucessivas da operação dentro da resolução de 10 minutos dos dados SCADA, gerando oscilações abruptas na geração real enquanto as previsões permanecem estáveis, demonstrando a capacidade dos modelos de estimar o potencial de geração subjacente independentemente das interrupções operacionais, assim como ocorre com o período de indisponibilidade que antecede o instante “03-07 12h”.

Os resultados acumulados para o período histórico completo são apresentados na Tabela 4.8. Os três modelos estimam geração entre 94,53% e 95,34% do potencial disponível, faixa ligeiramente superior à observada para o W301. Pelos mesmos motivos discutidos para o W301, as estimativas mantêm-se próximas entre si: a maior parte da energia potencial concentra-se em períodos de operação normal, restringindo-se as diferenças aos períodos removidos. A divergência residual entre modelos segue o padrão já identificado, refletindo diferenças na extrapolação para condições fora da distribuição de treinamento e a convergência das estimativas reforça a robustez do valor agregado de perda à escolha do modelo.

Figura 4.21 – Predição de Fator de Capacidade durante um dia de operação com falhas do W310.



Fonte: Do Autor (2026)

Tabela 4.8 – Energia gerada relativa ao potencial do W310.

Modelo	Geração*
Base	94,57%
MLP	95,34%
XGB	94,53%

Fonte: Do Autor (2026)

*A geração indica a porcentagem da energia total disponível

4.3 Discussão Geral

A análise comparativa entre os dois aerogeradores revela padrões consistentes de convergência do NSGA-II, validando a robustez da abordagem em contextos operacionais distintos. Em ambos os casos, o erro preditivo estabiliza-se nas primeiras gerações, enquanto a fronteira de Pareto continua a se expandir ao longo do processo evolutivo, indicando que a exploração do espaço de *trade-off* entre acurácia e complexidade ocorre de forma contínua e não se limita à fase inicial de otimização.

As cinco execuções independentes realizadas para o W301 demonstram que o NSGA-II converge de forma reprodutível: a variação do hipervolume entre execuções é de apenas 0,1 para o MLP e de 0,3 para o XGBoost — neste caso determinada pela execução 4, que se comporta como *outlier* (HV = 135,9; IGD = 1,46); excluía essa réplica, a variação cai para 0,1. Os pontos utópicos identificados apresentam RMSE e R^2 praticamente idênticos entre execuções. O XGBoost exibe maior consistência na complexidade do ponto utópico, convergindo para $f_2 = 18$ em todas as cinco execuções, enquanto o MLP apresenta maior variabilidade (f_2 entre 12 e 18), sugerindo que múltiplas configurações distintas oferecem *trade-offs* comparáveis para este modelo. Esta característica do MLP pode ser interpretada como maior flexibilidade na representação do problema, mas também como maior sensibilidade às condições iniciais do processo evolutivo. Para o W310, a única execução realizada por modelo produziu fronteiras de tamanho similar entre MLP e XGBoost (16 e 17 soluções), em contraste com a assimetria observada no W301 (15 e 25 soluções). Esta diferença pode estar associada à maior complexidade do espaço de estados do W310, sujeito a efeitos de esteira que reduzem a diferenciação entre configurações de fontes distintas, comprimindo a fronteira de Pareto em uma região mais estreita do espaço de objetivos.

Um achado transversal aos dois aerogeradores é que o NSGA-II sistematicamente constrói uma representação do trajeto completo do escoamento, selecionando fontes que cobrem as condições de entrada do fluxo, a evolução ao longo do parque e as condições locais próximas ao aerogerador alvo. Este padrão foi alcançado independentemente das condições operacionais distintas de cada equipamento, sugerindo que a abordagem captura a estrutura física real do problema e não apenas correlações espúrias nos dados de treinamento. A exclusão sistemática da T203 para o W310 ilustra uma capacidade adicional da metodologia: além de selecionar as fontes mais informativas, o algoritmo descarta aquelas cuja representatividade meteorológica é inadequada para o contexto específico do aerogerador, mesmo que estas estejam dentro do raio de busca definido. Esta propriedade reduz a dependência de critérios de pré-seleção manual de fontes, aumentando a autonomia da abordagem.

O limiar de retornos decrescentes, identificado consistentemente em torno de três fontes de dados para ambos os aerogeradores, representa um resultado de relevância prática: configurações com uma ou duas fontes adicionais além deste limiar produzem ganhos marginais de acurácia que não justificam o aumento de complexidade operacional. O critério do ponto utópico operacionaliza este equilíbrio de forma sistemática e reprodutível, identificando auto-

maticamente a região de transição da fronteira sem necessidade de limiares definidos *a priori*. Os modelos otimizados superaram consistentemente o modelo de referência em ambos os aerogeradores. A maior magnitude do ganho para o W310 evidencia que a metodologia agrega valor proporcionalmente maior em condições operacionais mais complexas, precisamente onde abordagens univariadas baseadas em curva de potência apresentam maiores limitações.

Do ponto de vista operacional, a estimativa de perdas energéticas em períodos de indisponibilidade e limitação tem implicações diretas para a gestão de parques eólicos. A capacidade de quantificar a geração potencial em períodos específicos permite identificar as causas e a magnitude das perdas associadas a cada evento, subsidiando decisões de manutenção preventiva e corretiva com maior precisão. Adicionalmente, em contextos onde contratos de fornecimento de energia preveem penalidades ou ressarcimentos por desvios de geração, estimativas mais precisas do potencial disponível fornecem base técnica mais robusta para negociações e contestações, reduzindo a assimetria de informação entre o operador do parque e as contrapartes contratuais.

A principal limitação operacional da metodologia proposta é o custo computacional do processo de otimização. O NSGA-II com 100 gerações e população de 100 indivíduos, avaliados por validação cruzada em 5 folds, implica em tempo de treinamento de mais de 6 horas por aerogerador. Este custo torna a aplicação direta da metodologia pouco escalável para parques com grande número de equipamentos, onde a otimização individual de cada aerogerador seria proibitiva em termos de tempo e recursos computacionais. Este problema poderia ser abordado por meio de redes neurais de múltiplas saídas (*multi-output neural networks*), que permitiriam o treinamento simultâneo para múltiplos equipamentos, amortizando o custo computacional da otimização entre um conjunto de aerogeradores com características operacionais similares.

A ausência de componentes temporais nas variáveis de entrada representa uma segunda limitação. Os modelos treinados utilizam apenas o estado instantâneo das variáveis meteorológicas, sem considerar a evolução temporal do escoamento. A inclusão de atrasos temporais (*lags*) como variáveis de entrada poderia melhorar consideravelmente a capacidade preditiva, especialmente em transições rápidas de regime de vento, onde o estado atual do aerogerador depende das condições dos instantes anteriores. A incorporação de atrasos temporais no espaço de busca do NSGA-II representa uma extensão natural da metodologia proposta. O cromossomo poderia ser estendido para incluir, além da seleção de torres e variáveis, a seleção do horizonte

de atraso ótimo para cada variável, permitindo que o algoritmo identifique automaticamente quais combinações de defasagem temporal são mais informativas para cada aerogerador.

Por fim, a análise de perdas energéticas realizada fornece estimativas acumuladas para o período histórico completo, sem decomposição por tipo de evento, direção de vento ou condição meteorológica. Esta granularidade adicional seria necessária para traduzir os resultados em recomendações operacionais específicas, transformando estimativas acumuladas em diagnósticos acionáveis para equipes de operação e manutenção.

5 CONCLUSÃO

Os resultados apresentados demonstram a eficácia da aplicação de técnicas de aprendizado de máquina para a estimativa da energia potencial em aerogeradores. A metodologia proposta, que integrou filtragem multimodal de dados e otimização multiobjetivo via NSGA-II, mostrou-se robusta para lidar com as complexidades micrometeorológicas e operacionais de um parque eólico real. A etapa de filtragem foi determinante para o sucesso da modelagem: ao isolar exclusivamente o comportamento de operação ótima, produziu conjuntos de treinamento representativos e eliminou o ruído introduzido por paradas e limitações operacionais.

A otimização multiobjetivo revelou uma hierarquia geoespacial consistente na seleção de fontes de dados, que se manifestou de forma autônoma em ambos os aerogeradores avaliados. Em ambos os casos, o NSGA-II convergiu para configurações que cobrem o trajeto completo do escoamento: uma fonte representando as condições de entrada do fluxo, uma capturando as condições locais próximas ao aerogerador e uma terceira caracterizando a evolução intermediária do vento ao longo do parque. Este padrão foi identificado independentemente das condições operacionais distintas de cada equipamento, sendo o W301 operando em vento relativamente livre e o W310 sob efeito de esteira pronunciado, o que sugere que a metodologia captura a estrutura física real do problema. Adicionalmente, a exclusão sistemática da T203 para o W310, torre cuja direção predominante diverge do regime do aerogerador, demonstra que o algoritmo não apenas seleciona as fontes mais informativas, mas também descarta automaticamente aquelas de representatividade meteorológica inadequada.

A fronteira de Pareto obtida em ambos os casos exibiu comportamento de retornos decrescentes, com ganhos marginais de acurácia além de três fontes de dados. O critério do ponto utópico identificou automaticamente esta região de transição, selecionando configurações com três TME e cinco a seis variáveis meteorológicas. Os modelos otimizados superaram consistentemente o modelo de referência: para o W301, o XGBoost e o MLP alcançaram R^2 de 0,984 e 0,982, respectivamente, frente a 0,973 do modelo base; para o W310, os ganhos foram ainda mais expressivos, com R^2 de 0,973 e 0,972 contra 0,933 do modelo base, evidenciando que a metodologia agrega valor proporcionalmente maior em condições operacionais mais complexas.

Em ambos os aerogeradores, o XGBoost apresentou desempenho ligeiramente superior ao MLP em termos de RMSE e explorou fronteiras de Pareto mais amplas, incorporando

nuances micrometeorológicas como turbulência e densidade do ar em configurações de maior complexidade. O MLP, por sua vez, convergiu para configurações mais compactas e apresentou maior variabilidade na complexidade do ponto utópico entre execuções independentes, indicando que múltiplas configurações distintas oferecem *trade-offs* comparáveis para este modelo.

A aplicação dos modelos finais ao histórico completo de dados, incluindo os períodos removidos durante a filtragem, permitiu estimar as perdas energéticas atribuíveis a eventos operacionais subótimos. Para o W301, a geração estimada situou-se entre 93,55% e 94,32% do potencial disponível; para o W310, entre 94,53% e 95,34%. A capacidade de quantificar estas perdas em períodos específicos tem aplicação direta na gestão operacional de parques eólicos, subsidiando decisões de manutenção e fornecendo base técnica para negociações contratuais em situações de penalidade ou ressarcimento por desvios de geração.

A metodologia proposta endereça diretamente as lacunas identificadas na revisão da literatura (Seção 2.4). Enquanto a maioria dos estudos foca exclusivamente na maximização da acurácia preditiva, esta pesquisa incorporou explicitamente a complexidade operacional como objetivo de otimização, produzindo soluções que equilibram desempenho preditivo com parcimônia de instrumentação. A abordagem baseada exclusivamente em dados meteorológicos de TME supera a limitação conceitual dos modelos multivariáveis que utilizam parâmetros de controle da turbina: ao invés de imitar decisões operacionais, os modelos propostos estimam o potencial físico real do aerogerador dado o recurso eólico disponível. A seleção automática de fontes e variáveis pelo NSGA-II demonstra que TMEs efetivamente melhoram as previsões quando a escolha das fontes é conduzida de forma sistemática e orientada pelas condições meteorológicas específicas de cada equipamento. Por fim, a validação com dados de um parque eólico brasileiro preenche a lacuna de representatividade geográfica predominante na literatura, majoritariamente restrita a condições europeias e norte-americanas.

Como trabalhos futuros, conforme discutido em maior profundidade na Seção 4.3, algumas direções promissoras podem ser destacadas. A principal limitação da abordagem é o custo computacional da otimização conduzida individualmente por aerogerador, o que motiva a investigação de redes neurais de múltiplas saídas, capazes de treinar simultaneamente modelos para conjuntos de aerogeradores com características operacionais similares e, assim, amortizar o custo da otimização. Outra extensão natural é a incorporação de informação temporal ao espaço de busca por meio de variáveis defasadas (*lags*), permitindo que o NSGA-II selecione

automaticamente o horizonte de atraso mais informativo para cada variável e capture a dependência do estado do aerogerador em relação aos instantes anteriores, especialmente relevante em transições rápidas de regime de vento. Adicionalmente, o uso de dados com maior frequência de amostragem poderia revelar dinâmicas do escoamento obscurecidas pela agregação em janelas de dez minutos e favorecer representações contínuas do fenômeno, com potencial benefício particular para os modelos baseados em redes neurais. Por fim, no plano da aplicação, a decomposição das perdas estimadas por evento e por causa, distinguindo falhas, manutenções e *curtailment*, permitiria separar perdas controláveis de não controláveis, requisito direto para a atribuição de responsabilidades em contratos de disponibilidade e para o direcionamento das ações de manutenção.

REFERÊNCIAS

Agência Nacional de Águas e Saneamento Básico. **Impacto da mudança climática nos Recursos Hídricos do Brasil**. 2024. Disponível em: https://metadados.snirh.gov.br/geonetwork/srv/api/records/31604c98-5bbe-4dc9-845d-998815607b33/attachments/Mudancas_Climaticas_25012024.pdf.

ALBUQUERQUE, J. C. d. **Desenvolvimento de modelos de curvas de potência via IA para previsão de geração eólica em curto prazo**. Tese (doctoralThesis) — Universidade Federal de Pernambuco, 2021. Accessed: 2024-11-10. Disponível em: <https://repositorio.ufpe.br/handle/123456789/42252>.

(ANEEL), A. N. de E. E. Nota Técnica, **Análise das contribuições recebidas na Consulta Pública 08/2013, bem como proposta de abertura de Audiência Pública para recebimento de contribuições referentes à minuta de Resolução Normativa que visa estabelecer condições e procedimentos a serem observados para instalação, operação e manutenção de estações de medição anemométricas, solarimétricas e climatológicas**. Brasília, DF: [s.n.], 2013. Processo nº 48500.001759/2013-28. Disponível em: https://antigo.aneel.gov.br/web/guest/audiencias-publicas-antigas?p_p_id=participacaopublica_WAR_participacaopublicaportlet&p_p_lifecycle=2&p_p_state=normal&p_p_mode=view&p_p_cacheability=cacheLevelPage&p_p_col_id=column-2&p_p_col_pos=1&p_p_col_count=2&_participacaopublica_WAR_participacaopublicaportlet_idDocumento=5118&_participacaopublica_WAR_participacaopublicaportlet_tipoFaseReuniao=fase&_participacaopublica_WAR_participacaopublicaportlet_jspPage=%2Fhtml%2Fpp%2Fvisualizar.jsp.

(ANEEL), A. N. de E. E. **PROCEDIMENTOS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA NO SISTEMA ELÉTRICO NACIONAL**. 2022. Acessado em: 6 de abril de 2025. Disponível em: <https://www.gov.br/aneel/pt-br/centrais-de-conteudos/procedimentos-regulatorios/prodist>.

ASTOLFI, D. Perspectives on SCADA data analysis methods for multivariate wind turbine power curve modeling. **Machines**, v. 9, n. 5, p. 100, 2021. ISSN 2075-1702. Disponível em: <https://www.mdpi.com/2075-1702/9/5/100>.

AVENDAÑO-VALENCIA, L. D.; ABDALLAH, I.; CHATZI, E. Virtual fatigue diagnostics of wake-affected wind turbine via gaussian process regression. **Renewable Energy**, v. 170, p. 539–561, 2021. ISSN 0960-1481. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960148121001671>.

BAI, L. *et al.* Wind turbine power curve estimation based on earth mover distance and artificial neural networks. **IET Renewable Power Generation**, v. 13, n. 15, p. 2939–2946, 2019. ISSN 1752-1424. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2019.0530>.

BARBER, S.; NORDBORG, H. Improving site-dependent power curve prediction accuracy using regression trees. **Journal of Physics: Conference Series**, v. 1618, n. 6, p. 062003, 2020. ISSN 1742-6596. Disponível em: <https://dx.doi.org/10.1088/1742-6596/1618/6/062003>.

BARRETO, G. A.; BRASIL, I. S.; SOUZA, L. G. M. Revisiting the modeling of wind turbine power curves using neural networks and fuzzy models: an application-oriented

evaluation. **Energy Systems**, v. 13, n. 4, p. 983–1010, 2022. ISSN 1868-3975. Disponível em: <https://doi.org/10.1007/s12667-021-00449-5>.

BENABDESSELAM, A. *et al.* On the use of indirect measurements in virtual sensors for renewable energies: A review. **Electronics**, v. 13, n. 8, p. 1545, 2024. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/13/8/1545>.

BISHOP, C. M. **Neural Networks for Pattern Recognition**. Oxford, UK: Oxford University Press, 1995. ISBN 978-0198538646.

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

BREIMAN, L. *et al.* **Classification and Regression Trees**. Boca Raton, FL: CRC Press, 1984. ISBN 978-0412048418.

BRITO, D. C. R. de S.; SOUZA, A. A. P. de; BARBOSA, B. H. G. Uso de aprendizagem de máquina para estimar a energia potencial de aerogeradores. In: AES, F. G.; FERREIRA, D.; BARRETO, G. (Ed.). **Anais do XVII Congresso Brasileiro de Inteligência Computacional (CBIC 2025)**. Belo Horizonte, MG: SBIC, 2025. p. 1–8.

CASCIANELLI, S. *et al.* Wind turbine power curve monitoring based on environmental and operational data. **IEEE Transactions on Industrial Informatics**, v. 18, n. 8, p. 5209–5218, 2022. ISSN 1941-0050. Disponível em: <https://ieeexplore.ieee.org/document/9616423>.

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**. New York, NY, USA: ACM, 2016. p. 785–794.

CIPM. **Report of the Meeting of the Comité International des Poids et Mesures**. Sèvres, France: [s.n.], 2007. BIPM Technical Report. Disponível em: <https://www.bipm.org/en/committees/cipm>.

CLIFTON, A. *et al.* Using machine learning to predict wind turbine power output. **Environmental Research Letters**, v. 8, n. 2, p. 024009, 2013. ISSN 1748-9326. Disponível em: <https://dx.doi.org/10.1088/1748-9326/8/2/024009>.

COELLO, C. A. C.; SIERRA, M. R. A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm. In: **MICAI 2004: Advances in Artificial Intelligence**. [S.l.]: Springer, 2004. p. 688–697.

DARWIN, C. **On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life**. London: John Murray, 1859.

DEB, K. *et al.* A fast and elitist multiobjective genetic algorithm: Nsga-ii. **IEEE Transactions on Evolutionary Computation**, v. 6, n. 2, p. 182–197, 2002.

DIETTERICH, T. G. Ensemble methods in machine learning. In: **Multiple Classifier Systems**. [S.l.]: Springer, 2000. (Lecture Notes in Computer Science, v. 1857), p. 1–15.

DIMITROV, N.; GÖÇMEN, T. Virtual sensors for wind turbines with machine learning-based time series models. **Wind Energy**, v. 25, n. 9, p. 1626–1645, 2022. ISSN 1099-1824. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/we.2762>.

EIBEN, A. E.; SMITH, J. E. **Introduction to Evolutionary Computing**. 2nd. ed. Berlin, Heidelberg: Springer, 2015. (Natural Computing Series). ISBN 978-3-662-44871-7.

Empresa de Pesquisa Energética. Nota Técnica, **Instruções para Medições Anemométricas**. Rio de Janeiro, RJ: [s.n.], 2014. Disponível em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-94/DEA%2008%20-%202014%20Instru%C3%A7%C3%B5es%20Medi%C3%A7%C3%B5es%20Anemom%C3%A9tricas.pdf>.

FORTIN, F.-A. DEAP: Evolutionary algorithms made easy. **Journal of Machine Learning Research**, v. 13, p. 2171–2175, 2012. Disponível em: <https://jmlr.org>.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. **Journal of Computer and System Sciences**, Elsevier, v. 55, n. 1, p. 119–139, 1997.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 2001.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Reading, MA: Addison-Wesley Professional, 1989. ISBN 978-0201157673.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. <https://www.deeplearningbook.org>. ISBN 978-0262035613.

GRINSZTAJN, L.; OYALLON, E.; VAROQUAUX, G. Why do tree-based models still outperform deep learning on typical tabular data? In: KOYEJO, S. *et al.* (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2022. v. 35, p. 507–520. Datasets and Benchmarks Track.

GUO, P.; INFELD, D. Wind turbine power curve modeling and monitoring with gaussian process and sprt. **Wind Energy**, 2013. Disponível em: <https://scispace.com/pdf/wind-turbine-power-curve-modeling-and-monitoring-with-313gean1sp.pdf>.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York, NY: Springer Science & Business Media, 2009. (Springer Series in Statistics). ISBN 978-0387848570.

HAYKIN, S. S. **Neural Networks and Learning Machines**. 3rd. ed. Upper Saddle River, NJ: Pearson, 2009. ISBN 978-0131471399.

HOLLAND, J. H. **Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence**. Ann Arbor, Michigan: University of Michigan Press, 1975.

International Electrotechnical Commission. **IEC 61400-12-1: Wind energy generation systems – Part 12-1: Power performance measurements of electricity producing wind turbines**. Geneva, Switzerland: [s.n.], 2017. International Standard.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. New York: Springer, 2013. (Springer Texts in Statistics). ISBN 978-1461471370.

JONG, K. A. D. **Evolutionary Computation: A Unified Approach**. Cambridge, MA: MIT Press, 2006. ISBN 978-0262041942.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. **arXiv preprint arXiv:1412.6980**, 2014.

LETZGUS, S.; MÜLLER, K.-R. An explainable AI framework for robust and transparent data-driven wind turbine power curve models. **Energy and AI**, v. 15, p. 100328, 2024. ISSN 2666-5468. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2666546823001003>.

LI, Y.; SHEN, X. A novel wind speed-sensing methodology for wind turbines based on digital twin technology. **IEEE Transactions on Instrumentation and Measurement**, v. 71, p. 1–13, 2022. ISSN 1557-9662. Disponível em: <https://ieeexplore.ieee.org/document/9666785>.

LYDIA, M. *et al.* A comprehensive review on wind turbine power curve modeling techniques. **Renewable and Sustainable Energy Reviews**, v. 30, p. 452–460, 2014. ISSN 1364-0321. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1364032113007296>.

LYONS, J. T.; GÖÇMEN, T. Applied machine learning techniques for performance analysis in large wind farms. **Energies**, v. 14, n. 13, p. 3756, 2021. ISSN 1996-1073. Disponível em: <https://www.mdpi.com/1996-1073/14/13/3756>.

MANWELL, J. F.; MCGOWAN, J. G.; ROGERS, A. L. **Wind Energy Explained: Theory, Design and Application**. 2nd. ed. Chichester, UK: John Wiley & Sons, 2010. ISBN 978-0-470-01500-1.

Ministério de Minas e Energia. **Portaria nº 21, de 18 de janeiro de 2008**. Brasília, DF: [s.n.], 2008. Diário Oficial da União. Alterada pela Portaria MME nº 029, de 28 de janeiro de 2011.

Ministério de Minas e Energia. **Portaria nº 029, de 28 de janeiro de 2011**. Brasília, DF: [s.n.], 2011. Diário Oficial da União. Altera a Portaria MME nº 21, de 18 de janeiro de 2008.

MITCHELL, M. **An Introduction to Genetic Algorithms**. Cambridge, MA: MIT Press, 1998. ISBN 978-0262631853.

MONTEITH, J. L.; UNSWORTH, M. H. **Principles of Environmental Physics: Plants, Animals, and the Atmosphere**. 4th. ed. Amsterdam: Academic Press, 2013. ISBN 978-0123869937.

NASCIMENTO, M. V. R. **Análise e predição de falhas em aerogeradores utilizando deep learning**. Guaratinguetá, SP, Brasil, 2023. Disponível em: <https://repositorio.unesp.br/entities/publication/ba10fe3f-47d2-487f-9448-6457a3e11bb4> (Acesso em: 28/05/2025).

NASCIMENTO, R. *et al.* DETECÇÃO DE ANOMALIAS EM POÇOS DE PRODUÇÃO DE PETRÓLEO OFFSHORE COM A UTILIZAÇÃO DE autoencoders E TÉCNICAS DE RECONHECIMENTO DE padrões. **UFLA**, mar. 2022. Disponível em: <http://repositorio.ufla.br/jspui/handle/1/49556>.

NYGAARD, N. G. *et al.* Modelling cluster wakes and wind farm blockage. **Journal of Physics: Conference Series**, v. 1618, n. 6, p. 062072, 2020. ISSN 1742-6596. Disponível em: <https://dx.doi.org/10.1088/1742-6596/1618/6/062072>.

OKUTA, R. *et al.* Cupy: A numpy-compatible library for nvidia gpu calculations. In: **Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)**. [s.n.], 2017. Disponível em: http://learningsys.org/nips17/assets/papers/paper_16.pdf.

ONS. **Relatório de Sustentabilidade - 2023**. 2023. Disponível em: https://www.ons.org.br/relatorio_anual/assets/pdf/ONS_RS2023.pdf.

Operador Nacional do Sistema Elétrico (ONS). **Cartilha do Sistema de Medição para Faturamento**. Rio de Janeiro, 2015. Disponível em: https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/Cartilha_SistemaMedicaoFaturamento.pdf. Acesso em: 7 jan. 2026.

Operador Nacional do Sistema Elétrico (ONS). **Perguntas Frequentes sobre Curtailment**. 2026. https://www.ons.org.br/Paginas/faq_curtailment.aspx. Acessado em: 6 de janeiro de 2026. Disponível em: https://www.ons.org.br/Paginas/faq_curtailment.aspx.

PANDIT, R. K.; INFELD, D.; KOLIOS, A. Comparison of advanced non-parametric models for wind turbine power curves. **IET Renewable Power Generation**, v. 13, n. 9, p. 1503–1510, 2019. ISSN 1752-1424. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1049/iet-rpg.2018.5728>.

PASZKE, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. **Advances in neural information processing systems**, v. 32, 2019.

PRAKASH, A. *et al.* A bayesian hierarchical model to understand the effect of terrain on wind turbine power curves. **IEEE Transactions on Sustainable Energy**, v. 15, n. 2, p. 1127–1137, 2024. ISSN 1949-3037. Disponível em: <https://ieeexplore.ieee.org/document/10301539>.

PUROHIT, S.; NG, E.; Syed Ahmed Kabir, I. F. Evaluation of three potential machine learning algorithms for predicting the velocity and turbulence intensity of a wind turbine wake. **Renewable Energy**, v. 184, p. 405–420, 2022. ISSN 0960-1481. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960148121016827>.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

RODRIGUES, M. V. S.; AQUINO, M. D. de; THOMAZ, A. C. F. Análise da eficiência na produção de energia eólica nos principais estados brasileiros produtores, por meio da aplicação de análise por envoltória de dados. **Anais do Congresso Brasileiro de Iniciação Científica (CBIC)**, 2014.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, American Psychological Association, v. 65, n. 6, p. 386–408, 1958.

SCHREIBER, J.; BOTTASSO, C. L.; BERTELÈ, M. Field testing of a local wind inflow estimator and wake detector. **Wind Energy Science**, v. 5, n. 3, p. 867–884, 2020. ISSN 2366-7443. Disponível em: <https://wes.copernicus.org/articles/5/867/2020/>.

SHWARTZ-ZIV, R.; ARMON, A. Tabular data: Deep learning is not all you need. **Information Fusion**, v. 81, p. 84–90, 2022.

SOHONI, V.; GUPTA, S. C.; NEMA, R. K. A critical review on wind turbine power curve modelling techniques and their applications in wind based energy systems. **Department of Electrical Engineering, Maulana Azad National Institute of Technology**, Bhopal 462051, India, jun. 2016. Received 27 March 2016; Revised 8 June 2016; Accepted 12 June 2016. Academic Editor: Kamal Aly. Disponível em: <https://doi.org/10.1155/2016/8519785>.

STULL, R. B. **An Introduction to Boundary Layer Meteorology**. Netherlands: Springer Science & Business Media, 1988.

VIRGOLINO, G. C. *et al.* Gaussian processes with logistic mean function for modeling wind turbine power curves. **Renewable Energy**, v. 162, p. 458–465, 2020. ISSN 0960-1481. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0960148120309150>.

WANG, J.; NIU, W.; YANG, Y. Wind turbine output power prediction by a segmented multivariate polynomial-xgboost model. **Energy Sources, Part A: Recovery, Utilization, and Environmental Effects**, Taylor & Francis, v. 46, n. 1, p. 505–521, 2024. Disponível em: <https://doi.org/10.1080/15567036.2023.2284840>.

WANG, P.; LI, Y.; ZHANG, G. Probabilistic power curve estimation based on meteorological factors and density LSTM. **Energy**, v. 269, p. 126768, 2023. ISSN 0360-5442. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360544223001627>.

WANG, Y. *et al.* Approaches to wind power curve modeling: A review and discussion. **Renewable and Sustainable Energy Reviews**, v. 116, p. 109422, 2019. ISSN 1364-0321. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1364032119306306>.

WHILE, L. *et al.* A faster algorithm for calculating hypervolume. **IEEE transactions on evolutionary computation**, v. 10, n. 1, p. 29–38, 2006.

WONG, T.-T. Performance evaluation of classification algorithms by k -fold and leave-one-out cross validation. **Pattern Recognition**, Elsevier, v. 48, n. 9, p. 2839–2846, 2015. Available online 17 March 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320315001070>.

ZITZLER, E. *et al.* Performance assessment of multiobjective optimizers: An analysis and review. **IEEE Transactions on Evolutionary Computation**, v. 7, n. 2, p. 117–132, 2003.