Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco

Campus Garanhuns

Curso de Bacharelado em Engenharia Elétrica

MATHEUS ALBUQUERQUE DE SATURNO

APLICAÇÃO DO ALGORITMO RANDOM FOREST PARA DETERMINAÇÃO DA GERAÇÃO FRUSTRADA EM PARQUES EÓLICOS EM FUNÇÃO DE RESTRIÇÕES OPERACIONAIS DO ONS

Garanhuns – PE

2023

MATHEUS ALBUQUERQUE DE SATURNO

APLICAÇÃO DO ALGORITMO RANDOM FOREST PARA DETERMINAÇÃO DA GERAÇÃO FRUSTRADA EM PARQUES EÓLICOS EM FUNÇÃO DE RESTRIÇÕES OPERACIONAIS DO ONS

Trabalho de conclusão de curso apresentado a Coordenação do Curso de Engenharia Elétrica do Instituto Federal de Pernambuco, campus Garanhuns, como requisito para obtenção do título de Bacharel em Engenharia Elétrica.

Orientador: Prof. Dr. Rafael Mendonça Rocha Barros

Coorientador: Prof. Dr. Diego Soares Lopes

Garanhuns – PE
2023

S254a Saturno, Matheus Albuquerque de.
Aplicação do algoritmo Random Forest para determinação da geração frustrada em parques eólicos em função de restrições operacionais do ONS. / Matheus Albuquerque de Saturno ; orientador Rafael Mendonça Rocha Barros ; Coorientador Diego Soares Lopes, 2023.
57 f. : il.

Orientador: Rafael Mendonça Rocha Barros.
Coorientador: Diego Soares Lopes
Trabalho de Conclusão de Curso (Graduação) – Instituto Federal de Pernambuco. Pró-Reitoria de Ensino. Diretoria de Ensino. Campus Garanhuns. Coordenação do Curso Superior em Engenharia. Curso de Bacharelado em Engenharia Elétrica, 2023.

1. Energia eólica - Produção. 2. Energia eólica – Modelos matemáticos. 3. Algoritmos. I. Título.

CDD 621.312136

Riane Melo de Freitas Alves –CRB4/1897

MATHEUS ALBUQUERQUE DE SATURNO

APLICAÇÃO DO ALGORITMO RANDOM FOREST PARA DETERMINAÇÃO DA GERAÇÃO FRUSTRADA EM PARQUES EÓLICOS EM FUNÇÃO DE RESTRIÇÕES OPERACIONAIS DO ONS

Trabalho de conclusão de curso apresentado a Coordenação do Curso de Engenharia Elétrica do Instituto Federal de Pernambuco, campus Garanhuns, como requisito para obtenção do título de Bacharel em Engenharia Elétrica.

Trabalho aprovado em: ____/____/____

_____

Prof. Dr. Rafael Mendonça Rocha Barros (IFPE, *campus* Garanhuns)
Orientador

_____

Prof. Dr. Diego Soares Lopes (IFPE, *campus* Garanhuns)
Coorientador

_____

Prof. Dr. Márcio Severino da Silva (IFPE, *campus* Garanhuns)
Avaliador interno

_____

Eng. Me. Vladmir Reis Pontes (Engeform Energia)
Avaliador externo

Dedico este trabalho a todos os ávidos buscadores do conhecimento, cuja sede por respostas profundas transcende a superficialidade. Que este esforço seja uma homenagem àqueles que, incansáveis em sua busca, recusam-se a serem saciados por respostas rasas. Que inspire a contínua exploração intelectual e o questionamento constante, construindo uma comunidade de aprendizado enriquecedora e dedicada à busca incansável pelo entendimento.

AGRADECIMENTOS

Agradeço à minha mãe, Thayz por sempre ter priorizado minha educação, me ensinando que o único caminho capaz de transformar vidas é através dos estudos.

À minha esposa, Máyra por todo suporte, carinho e atenção dados durante minha carreira profissional e acadêmica.

Aos meus irmãos, Israel, Nícolas e Lucas por sempre me incentivarem a ser uma pessoa melhor.

Aos meus familiares pela torcida e, em especial, ao meu tio Joventino, pelo abrigo e acolhimento nos anos iniciais do curso, quando não tive condições de me manter.

Aos amigos do SM98, em especial, Jackson, Fernando, Saulo, João Vitor, José Vitor e Marcos, por todos os momentos de descontração, ajuda e o companheirismo durante o curso.

Aos meus amigos, Janailson Almeida e Kleber Carvalho, por toda torcida durante meu período de formação.

Ao meu orientador, Professor Rafael Barros, pela orientação, incentivo e paciência durante o desenvolvimento deste trabalho.

Ao meu coorientador, Professor Diego Lopes, pela amizade e os ensinamentos passados desde o sexto período.

À minha orientadora de estágio e projetos desenvolvidos no IFPE, Manuelle Regina, pela amizade e todo suporte prestado nesses cinco anos de curso.

Aos amigos da Eólica Serra das Vacas, pelos ensinamentos, desafios e pela confiança depositada em mim durante os anos como estagiário de operação e manutenção.

Por fim, agradeço ao corpo docente do IFPE Garanhuns por todo ensinamento passado no período da minha formação.

“Não creio que haja uma emoção mais intensa para um inventor do que ver suas criações funcionando. Essas emoções fazem você esquecer de comer, de dormir, de tudo.”

Nikola Tesla

RESUMO

Este trabalho apresenta uma metodologia para a construção de modelo de previsão de geração eólica em momentos de restrições operacionais estabelecidas pelo Operador Nacional do Sistema (ONS). A abordagem adotada envolve a aplicação de técnicas de *Advanced Analytics*, e se destaca pela utilização do algoritmo *Random Forest*, proporcionando a capacidade preditiva necessária para estimar a energia gerada em intervalos de uma hora, com base em variáveis conhecidas do parque eólico. Para a validação e treinamento do modelo, foram coletadas 8.670 amostras reais, abrangendo dados de disponibilidade das unidades geradoras, potência e energia gerada e dados anemométricos. A aplicação de um procedimento de pré-processamento dos dados obtidos permitiu a seleção criteriosa das variáveis a serem utilizadas no treinamento do modelo, garantindo a representatividade e relevância das entradas. Todo processo computacional foi realizado através da plataforma computacional KNIME Analytics. Os resultados alcançados com a metodologia indicaram que a previsão da energia gerada apresenta Raiz do Erro Quadrático Médio (RMSE) de 0,354 MWh – para uma variável cujo valor médio é 11,5 MWh – e Erro Percentual Absoluto Médio (MAPE) de 4,5%, demonstrando uma boa acurácia do modelo desenvolvido para prever a variável *target*, reforçando a confiança na sua utilização para prever a geração eólica em momentos de restrição operacional. Com a utilização do modelo, foi possível determinar que para o parque eólico analisado, a geração frustrada no período de jan/2022 a dez/2022 foi de 228,73 MWh, que representa uma perda financeira de R$ 48 mil para o agente gerador. Dessa forma, a metodologia proposta pode ser utilizada por agentes de geração eólica para estimar a quantidade de geração frustrada em um dado período e utilizar esta informação para decisões gerenciais ou para pleitear compensações por parte do poder concedente.

Palavras-chave: Advanced Analytics. *Constrained-off*. Energia Eólica. Geração Frustrada. SIN.

ABSTRACT

This work presents a methodology for building a wind power generation prediction model during operational restrictions established by the National System Operator (ONS). The adopted approach involves the application of Advanced Analytics techniques and stands out for the use of the Random Forest algorithm, providing the predictive capability needed to estimate the generated energy in one-hour intervals based on known variables of the wind farm. For model validation and training, 8,670 real samples were collected, covering data on the availability of generating units, power and generated energy, and anemometric data. A data pre-processing procedure was applied to carefully select the variables to be used in the model training, ensuring the representativeness and relevance of the inputs. The entire computational process was carried out using the KNIME Analytics computational platform. The results achieved with the methodology indicated that the predicted energy generation has a Root Mean Square Error (RMSE) of 0.354 MWh – for a variable whose average value is 11.5 MWh – and a Mean Absolute Percentage Error (MAPE) of 3.8%, demonstrating a good accuracy of the developed model in predicting the target variable. This reinforces confidence in its use for predicting wind generation during operational restrictions. With the model, it was possible to determine that for the analyzed wind farm, the frustrated generation from Jan/2022 to Dec/2022 was 557.94 MWh, representing a financial loss of R$ 48.000 for the agent. Thus, the proposed methodology can be used by wind generation agents to estimate the amount of frustrated generation in each period and use this information for managerial decisions or to request compensation for the energy lost.

Keywords: Advanced Analytics. Constrained-off. Machine Learning. Wind Energy. SIN.

LISTA DE ILUSTRAÇÕES

Figura 1 – Evolução da Capacidade Instalada. 16
Figura 2 - Dinâmica dos ventos. 18
Figura 3 – Componentes básicos de uma turbina eólica. 20
Figura 4 - Curva de potência teórica 21
Figura 5 - Elementos de um *box-plot.* 25
Figura 6 – Ilustração do funcionamento do algoritmo *Random Forest.* 27
Figura 7 – Ilustração validação cruzada 28
Figura 8 – Fluxograma das etapas da metodologia utilizada no trabalho. 31
Figura 9 – Aerogerador Plataforma 2.X. 32
Figura 10 – Distribuição do conjunto de aerogeradores. 33
Figura 11 – Rosa dos Ventos. 33
Figura 12 – Ilustração ambiente KNIME. 34
Figura 13 – Fluxograma de coleta de dados. 35
Figura 14 – Ilustração Torre de Medição Anemométrica. 36
Figura 15 – *Data Logger* EOL Zenith. 37
Figura 16 – Medidor ION8650c. 38
Figura 17 – Fluxo do tratamento de dados. 39
Figura 18 – Fluxograma Tratamento Dados Anemométricos. 41
Figura 19 – Agrupamento por hora. 41
Figura 20 – Fluxograma de treinamento do modelo. 43
Figura 22 – Velocidade do vento e potência 45
Figura 23 – Curva Potência x Vento 46
Figura 24 – Curva de valores previstos e valores reais. 49

LISTA DE TABELAS

Tabela 1 – Especificações do *hardware*.....35
Tabela 2 – Relação dos sensores. 37
Tabela 3 – Limiares EPE. 40
Tabela 4 – Parâmetros utilizados no método *Brute Force.* 43
Tabela 4 – Correlação definida pelo coeficiente de relação Spearman para variável MED_G.....47
Tabela 5 – Hiperparâmetros ótimos para o algoritmo. 48
Tabela 6 – Resultado da validação cruzada para o modelo. 48
Tabela 7 – Resultado dos valores estimados de geração frustrada. 50

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ANEEL	Agência Nacional de Energia Elétrica
CCEE	Comercialização de Energia Elétrica
EPE	Empresa de Pesquisa Energética
FNEN	Fluxo De Energia Nordeste-Norte
FNESE	Fluxo De Energia Nordeste-Sudeste/Centro-Oeste
GWEC	Global Wind Energy Council
IA	Inteligência Artificial
IFPE	Instituto Federal de Pernambuco
KNIME	*Konstanz Information Miner*
MAPE	*Mean Absolute Percentage Error*
ML	*Machine learning*
ONS	Operador Nacional do Sistema Elétrico
PROREDE	Procedimentos de Rede
RMSE	*Root Mean Square Error*
RMT	Rede de Média Tensão
SAGER	Sistema de Apuração da Geração
SCDE	Sistema de Coleta de Dados de Energia
SIN	Sistema Interligado Nacional
SMF	Sistema de Medição de Faturamento
TMA	Torre de Medição Anemométrica
WFMS	Wind Farm Management System

LISTA DE SÍMBOLOS

P_w	Potência média dos ventos
ρ	Densidade do ar
A	Área de varredura do rotor
V^3	Velocidade do vento
P	Pressão absoluta
T	Temperatura
η	Rendimento
X'	Valor normalizado
X	Valor original
$X_{máx}$	Valor máximo
$X_{mín}$	Valor mínimo
ρ_s	Coeficiente de Spearman
rg_X	Ordem de classificação de X
rg_Y	Ordem de classificação de Y
σ_{rgX}	Desvio padrão de rg_X
σ_{rgX}	Desvio padrão de rg_Y
n	Quantidade de amostras
Y_i	Valor real
$\hat{Y}_i$	Valor previsto pelo modelo
R^2	Coeficiente de Determinação

SUMÁRIO

1. INTRODUÇÃO 15
   1.1. Panorama da Geração Eólica 15
   1.2. *Constrained-off* e Geração Frustrada 16
   1.3. Objetivos 17
2. FUNDAMENTAÇÃO TEÓRICA 18
   2.1. A Energia Eólica 18
      2.1.1. Os Ventos 18
      2.1.2. As Turbinas Eólicas 19
   2.2. O Sistema Interligado Nacional – SIN 22
   2.3. O processo de *Advanced Analytics* 24
      2.3.1. Pré-processamento de Dados 24
      2.3.2. Valores Faltantes 24
      2.3.3. Outliers 25
      2.3.4. Normalização 25
      2.3.5. Análise de Correlação 26
   2.4. Machine Learning 26
      2.4.1. Random Forest 26
      2.4.2. Otimização de Hiperparâmetros 27
   2.5. Validação Cruzada 28
   2.6. Medidas de Desempenho 29
      2.6.1. Root Mean Square Error 29
      2.6.1. Coeficiente de Determinação 29
      2.6.1. MAPE 30
3. MATERIAL E MÉTODOS 31
   3.1. Descrição do Parque Eólico Analisado 31
   3.2. Ambiente Computacional 34
   3.3. Pré-processamento 35
      3.3.1. Coleta dos Dados 35
      3.3.2. Tratamento dos Dados 39
      3.3.3. Consolidação da Base de Dados 41
   3.4. Seleção de Variáveis 42
   3.5. Treinamento do Modelo 42

3.6. Avaliação do Modelo 43
4. RESULTADOS E DISCUSSÕES 45
4.1. Análise Preliminar dos Dados 45
4.1.1. Seleção de Variáveis 47
4.2. Treinamento e Avaliação do Modelo 47
4.3. Determinação da Geração Frustrada 49
5. CONCLUSÕES 51
REFERÊNCIAS 53
APÊNDICE 55

1. INTRODUÇÃO

1.1. Panorama da Geração Eólica

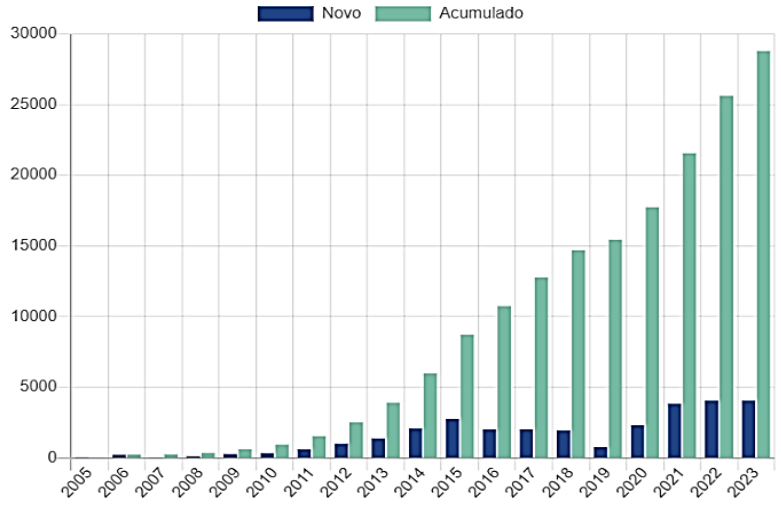
A Energia Elétrica se tornou indispensável para o funcionamento da sociedade, sendo necessária para os mais diversos tipos de atividades cotidianas. No princípio, o avanço da matriz elétrica foi lastreado em fontes hídricas e térmicas. As Fontes térmicas contribuíram paras emissões de CO2 na atmosfera, assim, fez-se necessário a busca por outras fontes de energias que fossem menos poluentes.

Nesse contexto, a demanda crescente de energia levou a expansão das fontes renováveis, sendo a eólica umas das principais fontes em expansão no Brasil. Embora as fontes hídricas, que também são fontes renováveis, tenham parcela expressiva na matriz elétrica brasileira somente sua geração não é o suficiente para suprir a demanda do Sistema Interligado Nacional (SIN). Em 2021, 93,6 GW de capacidade instalada foram adicionados ao portifólio global de potência instalada, somando um total de 837 GW de potência disponível, e até 2026, mais 110 GW serão instalados (GWEC, 2022).

Na geração eólica, a obtenção da energia elétrica se dá pela transformação da energia cinética dos ventos em energia mecânica através de grandes rotores, onde pás são acopladas e, por fim, através de um gerador, são convertidas em energia elétrica (REIS, 2011). As fontes renováveis representam 83,85% da Matriz Elétrica Brasileira, sendo 14% fontes eólicas, que somam um total de 28 GW (ANEEL, 2023).

Na Figura 1, pode-se observar a expansão da geração eólica no brasil ao longo dos últimos 18 anos.

Figura 1 – Evolução da Capacidade Instalada.



Fonte: (ABEEÓLICA, 2023).

Em 2006, pode-se observar uma capacidade instalada de apenas 230 MW tendo alcançado, em 2023, um valor acumulado de 28.811 MW (ABEEÓLICA, 2023).

1.2. *Constrained-off* e Geração Frustrada

Para manter a geração de energia otimizada, o Operador Nacional do Sistema (ONS) realiza o controle da geração elétrica em tempo real conforme o Procedimentos de Rede (ONS, 2021).

Para controlar o despacho de energia, o ONS realiza o fluxo de carga baseado nos montantes de geração e de Intercâmbio de energia. A fim de manter a confiabilidade do SIN, como por exemplo sobre a frequência e controle de carregamento em linhas de transmissão, é necessária a restrição de geração, solicitadas pelo ONS, para as usinas integradas ao SIN (ONS, 2022).

Estas restrições são nomeadas de *constrained-off*, e são aplicadas para limitar a potência máxima da planta que, no caso de fontes eólicas em momentos de ventos favoráveis, frusta a geração esperada, causando prejuízos monetários ao produtor. Estima-se que, desde o começo destas restrições, 75 milhões de reais tenham sido perdidos em termos de receita para os agentes geradores (AGÊNCIA INFRA, 2023).

Após o evento do dia 15 de Agosto de 2023, o apagão que deixou mais de 30 milhões de pessoas sem energia e afetou todos os estados brasileiros, o ONS reduziu o intercâmbio de energia excedente gerada no Nordeste, estes limites foram reduzidos para 5.000 MW para o fluxo Nordeste-Sudeste/Centro-Oeste (FNESE) e, 6.000 MW no fluxo Nordeste-Norte (FNEN). Durante esse período de contingência, alguns

parques tiveram seu limite de geração zerado, permanecendo impossibilitados de gerar energia.

Alguns agentes do setor eólico queixam-se pela falta de clareza na metodologia proposta pela ANEEL para cálculo da geração frustrada (SILVA, 2023). Neste contexto, uma metodologia bem definida é necessária para garantir aos agentes a possibilidade de estimar sua geração frustrada.

1.3. Objetivos

O objetivo geral deste trabalho é propor uma metodologia baseada em inteligência artificial para cálculo da geração frustrada em parques de geração eólica, pelas restrições operacionais estabelecidas pelo ONS.

Para alcançar o objetivo geral, também devem ser alcançados os seguintes objetivos específicos:

- Levantar o histórico da operação do parque eólico real, com as informações de velocidade do vento, temperatura, umidade, direção do vento, pressão, disponibilidade das unidades geradores e potência gerada;
- Realizar pré-processamento na base de dados levantada para identificar correlações entre as varáveis;
- Aplicar técnica de inteligência artificial para construção de modelo preditivo capaz de prever a potência de saída, baseada nas informações anemométricas;
- Determinar a geração frustrada no período de 2022 para o parque eólico em questão.

2. FUNDAMENTAÇÃO TEÓRICA

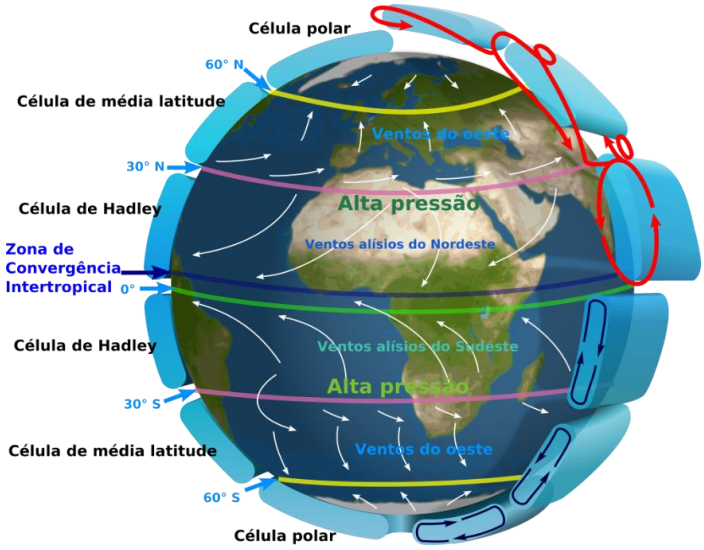
Neste capítulo, inicialmente será apresentado uma revisão sobre a dinâmica dos ventos e os aspectos intrínsecos à geração de energia eólica. Em seguida, é feita uma introdução ao modelo de previsão adotado e as métricas de desempenho para avaliação dos resultados obtidos.

2.1. A Energia Eólica

2.1.1. Os Ventos

Os ventos são resultado de uma movimentação de ar, movimentação essa que é causada pela diferença de pressão e temperatura entre pontos na superfície. Assim, em um movimento que vai em direção a zonas de baixa pressão e no sentido oposto de zonas de alta pressão (FERREIRA, 2006). A diferença de pressão é originada em razão do aquecimento desigual da superfície terrestre pela irradiância solar. Não obstante, outros fatores também influenciam a dinâmica do vento, uma vez que, o efeito Coriolis afeta o movimento das massas de ar no planeta. A diferença de temperatura entre os polos e a Linha do Equador, em conjunto com o efeito Coriolis, explicam o movimento dos ventos alísios e os ventos do oeste. Na Figura 2, observamos a dinâmica das correntes de ar no Globo.

Figura 2 - Dinâmica dos ventos.



Fonte: Adaptado de Kaidor (2013).

Além desses padrões observados na Figura 2, existem padrões regionais de circulação de vento, por exemplo, entre a terra e o oceano, como resultado dessa interação é a brisa marítima e a brisa terrestre. A brisa marítima ocorre durante o dia, enquanto a superfície da terra está mais quente que a superfície marítima, pela diminuição de pressão na terra, o ar mais denso sob o oceano se move em direção à terra. Durante a noite, o cenário se inverte, e temos a brisa terrestre.

A energia cinética disponível nos ventos, representada por P_w, em Watts, pode ser definida pela Equação (1).

$$P_w = \frac{1}{2}\rho A V^3 \tag{1}$$

Em que P_w é potência média do vento em Watts (W), ρ densidade do ar, A é área de varredura do rotor (m²) V velocidade do vento (m/s). Então, podemos observar que um aumento de 10% no vento, resulta em um aumento de 33% na potência disponível no vento, assim como um aumento na área do rotor, também aumenta na potência média do vento.

A densidade ρ, pode ser calculada através da Equação (2).

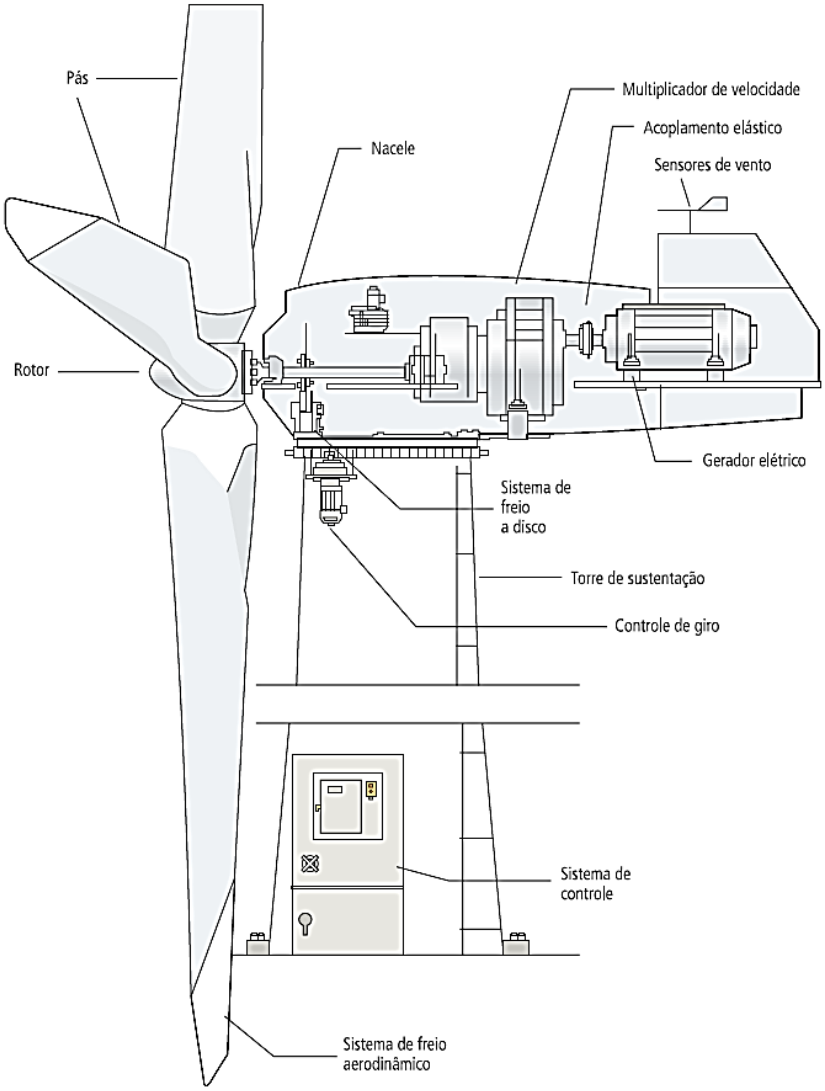
$$\rho = \frac{P}{RT} \tag{2}$$

Em que P é a pressão absoluta em (Pa), R é a constante específica do gás para o ar seco ($J/(kg * K)$) e T é a temperatura absoluta (K). Analisando as Equações (1) e (2), podemos observar uma relação inversamente proporcional entre a potência disponível P_w e a Temperatura T.

2.1.2. As Turbinas Eólicas

Para entendimento do funcionamento do sistema eólico, é necessário compreender tantos os aspectos mecânicos quanto os elétricos de um aerogerador. Assim, definimos a estrutura básica de um aerogerador na Figura 3.

Figura 3 – Componentes básicos de uma turbina eólica.



Fonte: (CBBE, 2000).

Na Figura 3, pode-se destacar alguns componentes principais que compõem a dinâmica de funcionamento de um aerogerador, são eles:

- Rotor e Pá: estes componentes são responsáveis por capturar a energia cinética dos ventos e pela frenagem aerodinâmica;
- Multiplicador de velocidade ou *gearbox*: tem como função o aumento da velocidade de rotação, assim diminuindo a quantidade de polos necessários no gerador;
- Gerador: é o elemento final de conversão, tem como função a conversão da energia mecânica em energia elétrica.

Não se limitando aos componentes supracitados, ainda existem diversos equipamentos de instrumentação, com objetivo de proteção e orientação do aerogerador.

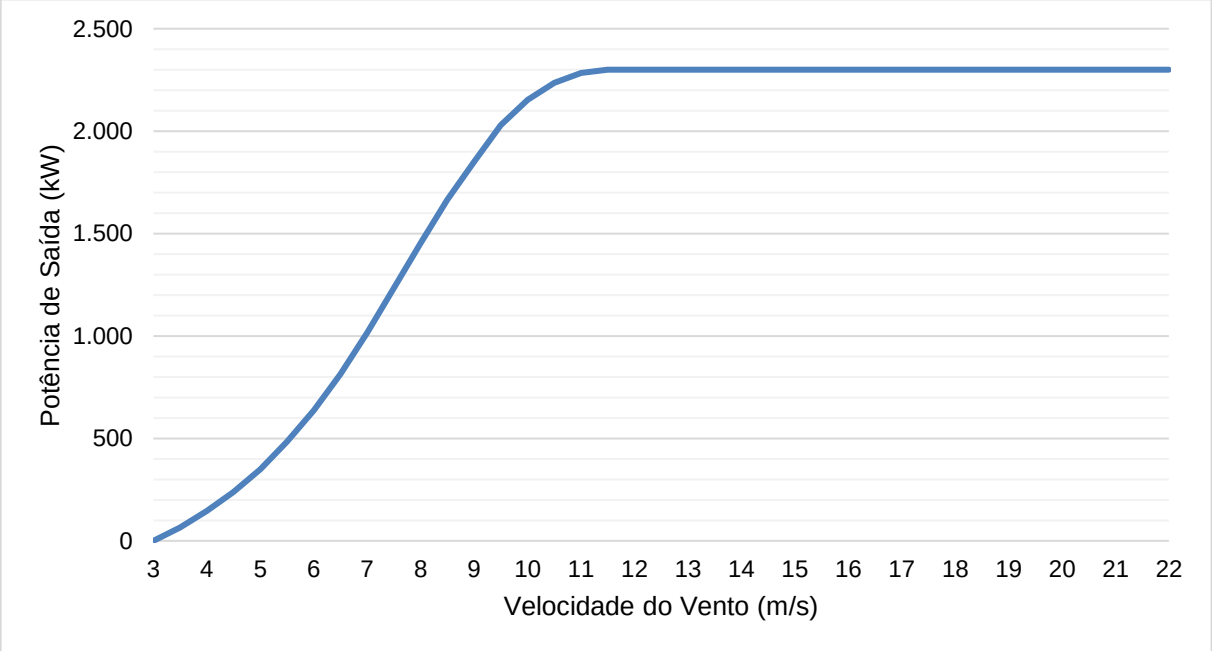
Devido às características aerodinâmicas, Albert Betz (1920) determinou a eficiência aerodinâmica em 59,3%, ou seja, apenas 59,3% da energia que passa pela área pode ser captura pelo rotor. Em aerogeradores reais, são encontradas eficiências próximas a 35%, devido as perdas de conversão. De forma geral, a eficiência de uma turbina eólica pode ser explicada pela Equação (3).

$$\eta = \eta_B * \eta_A * \eta_M * \eta_r * \eta_G \tag{3}$$

Em que η_B é a eficiência teórica (Betz), η_A é o rendimento aerodinâmico (pás), η_M é o rendimento do multiplicador de velocidades, η_r é o rendimento do rotor e η_G é o rendimento do gerador.

Cada aerogerador possui uma curva característica de desempenho de energia, essa curva representa o comportamento da potência de uma turbina em relação a velocidade do vento, indicando a produção máxima em diferentes velocidades de vento. A curva de potência de um aerogerador é apresentada conforme a Figura 4.

Figura 4 - Curva de potência teórica.



Fonte: Autoria própria.

Na Figura 4, é possível observar o comportamento da potência elétrica no aerogerador em relação à velocidade do vento. Esta curva, trata-se da curva ideal, para uma densidade do ar de $\rho = 1{,}225\ kg/m^3$, podendo sofrer alterações conforme a mudança da densidade do ar e outras variáveis supracitadas. A curva de potência possui três pontos principais:

- Velocidade mínima: que corresponde a mínima do vento para que a turbina entre em operação;
- Velocidade nominal: é a velocidade onde a turbina atinge sua potência nominal;
- Velocidade máxima: velocidade em que a turbina cessa sua operação devido aos ventos altos.

As velocidades mínima e máxima são os limiares extremos de operação do aerogerador e também podem ser chamados de velocidade de *cut-in e cut-off.*

2.2. O Sistema Interligado Nacional – SIN e Restrições Operacionais

O Sistema Interligado Nacional (SIN) representa uma infraestrutura crítica que contempla a geração, transmissão e distribuição de energia elétrica em todo país (GSI, 2022). O SIN é constituído de quatro subsistemas: Sul, Sudeste/Centro-Oeste, Nordeste e Norte.

A interligação dos sistemas elétricos, viabilizada pela rede de transmissão, possibilita a transferência de energia entre subsistemas, facilitando a realização de ganhos sinérgicos e explorando as variações nos regimes hidrológicos das bacias. A integração dos recursos de geração e transmissão emerge como um meio eficaz para atender ao mercado de forma segura e economicamente viável.

A capacidade instalada de geração no SIN é predominantemente composta por usinas hidrelétricas distribuídas em dezesseis bacias hidrográficas nas distintas regiões do país. Nos últimos anos, observa-se um expressivo crescimento na instalação de usinas eólicas, especialmente nas regiões Nordeste e Sul, destacando a crescente importância dessa modalidade na oferta ao mercado.

As usinas térmicas, geralmente localizadas nas proximidades dos principais centros de carga, desempenham um papel estratégico significativo, contribuindo para a segurança do SIN. O controle de despacho e intercâmbio de energia é uma atividade

fundamental para a operação segura e eficiente do SIN. O controle de despacho consiste na determinação da quantidade de energia que cada usina geradora deve produzir em cada instante. O ONS leva em consideração diversos fatores para realizar o despacho, incluindo a demanda de energia, as condições hidrológicas, as restrições da rede de transmissão e os custos de geração. (ANEEL, 2023). O controle de despacho e intercâmbio de energia é realizado por meio de um software chamado CAG. O CAG é um sistema complexo que utiliza modelos matemáticos e técnicas de inteligência artificial para tomar decisões de despacho e intercâmbio (ONS, 2021).

O ONS realiza o controle de despacho e intercâmbio de energia em duas etapas:

- Previsão de demanda: O ONS realiza uma previsão da demanda de energia para as próximas 24 horas. Essa previsão é utilizada para determinar a quantidade de energia que deve ser produzida pelas usinas geradoras.
- Despacho: O ONS realiza o despacho das usinas geradoras com base na previsão de demanda e nas condições hidrológicas.

As restrições operacionais são classificadas em 3 tipos:

- **Razão Energética**: motivada pela impossibilidade de alocação de geração na carga.
- **Razão de indisponibilidade externa:** originado por eventos de indisponibilidade em instalações externas às usinas correspondentes, tanto nas instalações de transmissão categorizadas como Rede Básica quanto nas Demais Instalações de Transmissão dentro do escopo da distribuição. Essa categorização exclui instalações destinadas ao uso exclusivo ou compartilhado pelo gerador, sob sua administração ou de terceiros. Engloba, portanto, as situações de indisponibilidade de linhas de transmissão, transformadores, disjuntores e instalações de subestações em geral.
- **Razão de atendimento a requisitos de confiabilidade elétrica:** gerada por motivos de confiabilidade elétrica que não derivam de falhas nos equipamentos do sistema de transmissão. Esta categoria abrange

casos como a redução da geração devido ao alcance de limites em linhas de transmissão, sobrecarga de equipamentos, requisitos de estabilidade dinâmica, entre outros.

-

2.3. O processo de *Advanced Analytics*

Advanced Analytics refere-se a um conjunto avançado de métodos analíticos e técnicas de processamento de dados que vão além das análises estatísticas tradicionais (LEVENTHAL, 2010). Essas técnicas são projetadas para descobrir *insights* mais profundos, padrões complexos e relacionamentos em conjuntos de dados extensos e muitas vezes complexos. O termo engloba diversas abordagens, incluindo técnicas de aprendizado de máquina, mineração de dados, análise preditiva e outras formas de análise estatística avançada. Na sequência serão discutidos os principais pontos dessa abordagem.

2.3.1. Pré-processamento de Dados

O tratamento de dados é um processo fundamental na construção de um modelo de aprendizagem de máquina. “Os dados do mundo real são sujos.” (GRUS, 2016, p. 202), assim, algumas técnicas devem ser aplicadas a fim de remover dados faltantes ou duvidosos, para garantir um melhor funcionamento do modelo de aprendizado. Logo, após coleta dos dados, é necessário eliminar *outliers,* remover dados faltantes e realizar a normalização de valores numéricos. Essa etapa é crucial e caso ela não seja bem-sucedida, é provável que ocorra um problema no treinamento do modelo. Após o tratamento, os dados são divididos em dois conjuntos: treinamento, que será usado para “ensinar” o modelo determinado padrão; e teste, que servirá para a validação do modelo treinado.

2.3.2. Valores Faltantes

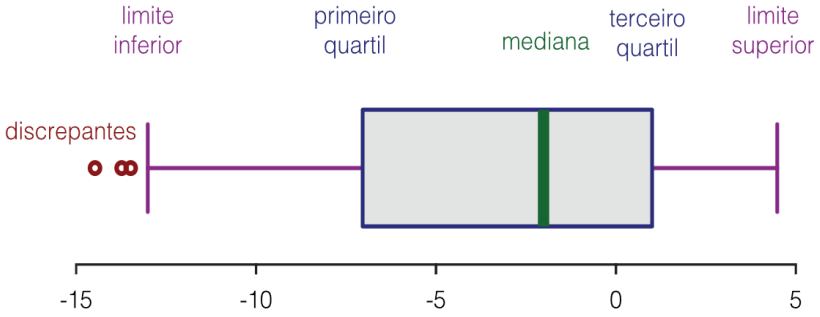
É normal que, em algumas linhas ou colunas dos bancos de dados, existam valores faltantes, essa situação é conhecida como *missing values*. Para tratar as lacunas, uma regra ou algoritmo podem ser aplicados para se obter um valor sintético próximo ao esperado (GRUS, 2016).

2.3.3. Outliers

Segundo (HAWKINS, 1980, p. 1) um *outlier* é “um valor observado que desvia tanto das outras observações ao ponto de ser considerada suspeito”. Um *outlier* é uma observação atípica que difere do restante do conjunto, podendo influenciar de maneira significativa as análises estatísticas.

Um dos métodos que podem ser utilizados para a identificação de *outliers* é o *box-plot*. Na Figura 5, podemos observar uma representação gráfica de *box-plot.*

Figura 5 - Elementos de um *box-plot.*



Fonte: NeuroMat (2017).

O *box-plot* exibe a mediana, quartis e possíveis outliers de maneira compacta e informativa. O gráfico consiste em uma caixa retangular que representa o intervalo interquartil, enquanto a linha no interior da caixa representa a mediana. As *whiskers* (as linhas que se estendem para fora da caixa), indicam a dispersão dos dados, e pontos fora das *whiskers* podem ser considerados outliers.

2.3.4. Normalização

No processo de análise de dados, a normalização refere-se ao processo de ajuste de valores das varáveis para uma escala comum, normalmente entre 0 e 1. Isso é feito para evitar variáveis com magnitudes com pesos desproporcionais, assim evitando o enviesamento no comportamento do modelo de aprendizado. A normalização na base unitária pode ser obtida através da Equação (4)

$$X' = \frac{X - X_{mín}}{X_{máx} - X_{mín}} \tag{4}$$

Em que X'é o valor normalizado, X é o valor original, $X_{máx}$ e $X_{mín}$ são os valores máximo e mínimo da coluna a ser normalizada.

2.3.5. Análise de Correlação

O coeficiente ρ de Spearman mede a intensidade de correlação entre duas variáveis. Este valor de correlação pode ser determinado conforme a Equação (5).

$$\rho = \frac{cov(rg_X, rg_Y)}{\sigma_{rgX}\sigma_{rgY}} \quad (5)$$

Em que rg_X e ,rg_Y são as ordens de classificações das variáveis X e Y; σ_{rgX} e σ_{rgY} são os desvios padrão das variáveis rg_X e ,rg_Y.

Esse coeficiente de determinação varia entre -1 e 1. Quanto mais próximo destes extremos, maior será a correlação entre as variáveis, seja ela inversa ou direta.

2.4. Machine Learning

O termo *Machine learning* (ML), ou Aprendizagem de Máquina, é um campo da inteligência artificial (IA) que se concentra no desenvolvimento de algoritmos e modelos que permitem aos sistemas computacionais aprenderem padrões e realizar tarefas específicas sem serem explicitamente programados. (IZBICKI; SANTOS, 2020).

Existem diversas aplicações da Aprendizagem de Máquina, dentre estas, podemos citar os problemas de regressão, que buscam, através de variáveis conhecidas, prever uma variável desconhecida. (SMOLA; VISHWANATHAN, 2008).

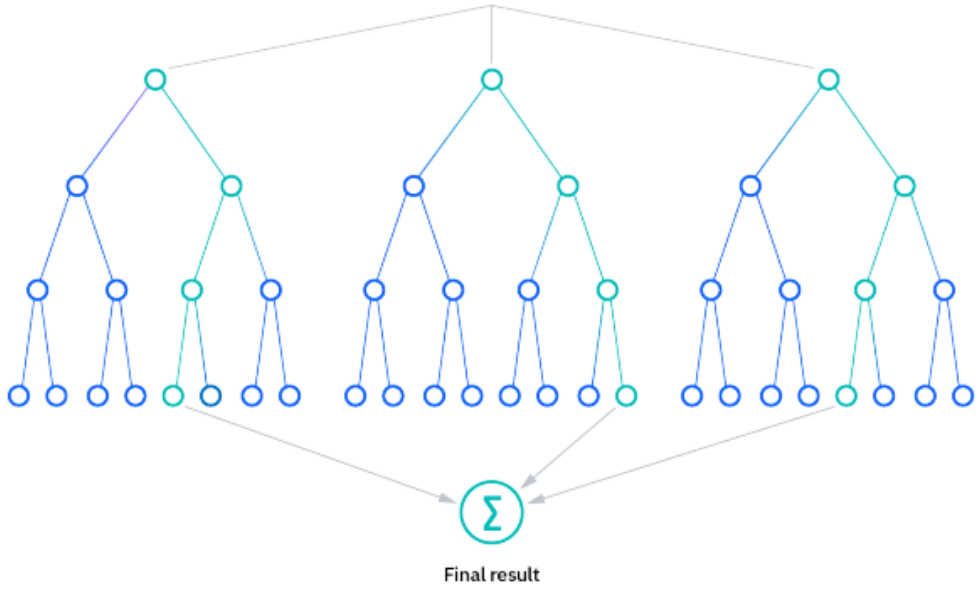
Para obter o modelo pode-se usar o aprendizado supervisionado. Neste tipo de aprendizado o algoritmo é treinado em um conjunto de dados. O modelo tenta aprender a estrutura e as relações entre os dados de entrada e as saídas desejadas. Uma vez treinado, o modelo pode então fazer as previsões com base no que aprendeu no treinamento (ESCOVEDO; KOSHIYAMA, 2020).

2.4.1. Random Forest

Random Forest, ou Floresta Aleatória em português, é um algoritmo de aprendizado de máquina que pertence à categoria de métodos ensemble. Esses métodos constroem vários modelos individuais e os combinam para obter um modelo mais robusto e preciso. A *Random Forest* é particularmente eficaz para tarefas de classificação e regressão.

A ideia principal por trás da Random Forest é criar várias árvores de decisão durante o treinamento e combiná-las para reduzir o *overfitting* e melhorar a generalização. Cada árvore na floresta é treinada em uma subamostra aleatória do conjunto de dados, e as previsões são feitas combinando previsões de todas as árvores. Sendo assim, a média para a regressão e votação, para a classificação. A Figura 6 apresenta uma ilustração sobre o funcionamento do algoritmo.

Figura 6 – Ilustração do funcionamento do algoritmo *Random Forest.*



Fonte: IBM (2023).

2.4.2. Otimização de Hiperparâmetros

A etapa de otimização de hiperparâmetros desempenha um papel fundamental na aprimoração do desempenho do modelo de modelos preditivos. O principal objetivo dessa etapa é encontrar a combinação mais apropriada de hiperparâmetros que otimize o desempenho do algoritmo. Isso envolve a exploração de diferentes valores para parâmetros-chave, como o número de árvores na floresta, a profundidade máxima de cada árvore. A escolha acertada desses hiperparâmetros é essencial para garantir que o modelo atinja seu potencial máximo em termos de acurácia e capacidade de generalização (KUNAPULI, 2023).

O método usado para otimização do hiperparâmetro foi de força bruta. O método de força bruta para otimização de hiperparâmetros é uma técnica simples e direta que consiste em testar todas as combinações possíveis de valores para os

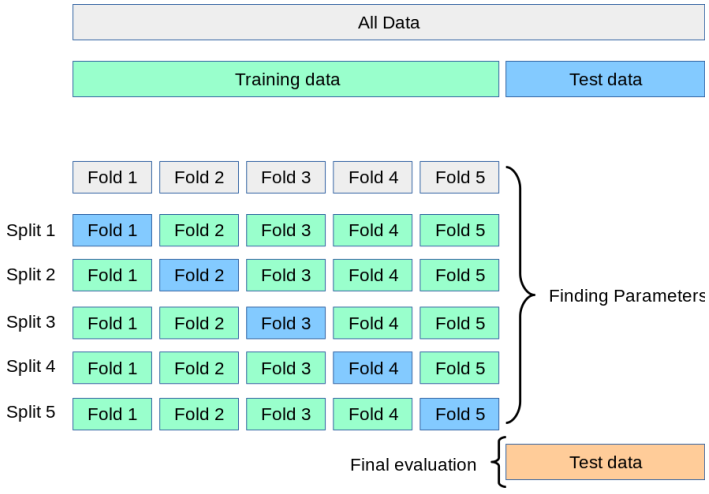
hiperparâmetros. Esta técnica é eficiente para conjuntos de dados pequenos, mas pode ser inviável para conjuntos de dados grandes. Para sua implementação é necessário definir um espaço de busca para os hiperparâmetros. Este espaço deve incluir valores que sejam prováveis de resultar em um bom desempenho do modelo. Por exemplo, se o conjunto de dados é pequeno, é provável que um número pequeno de árvores resulte em um bom desempenho.

2.5. Validação Cruzada

A validação cruzada (*Cross-Validation*) é usada para avaliar o desempenho de um modelo e mitigar vieses na divisão dos dados, ao dividi-los para treinamento e teste. Assim seu objetivo é obter uma estimativa mais confiável do desempenho do modelo e avaliá-lo em diferentes subconjuntos de dados (CUNHA, 2019).

No modelo de validação *K-Fold Cross-Validation* os dados são divididos em k partes. O modelo é treinado em $k - 1$ partes. Esse processo é repetido k vezes, cada vez usando um conjunto diferente para teste. A métrica do desempenho é dada pela média das métricas obtidas em cada interação. Podemos observar na Figura 7 uma ilustração do funcionamento da *K-Fold Cross-Validation.*

Figura 7 – Ilustração validação cruzada.



Fonte: Scikit Learn (2023).

Pode-se observar a divisão em k subconjuntos e, em cada interação, um subconjunto é selecionado como um conjunto de validação, enquanto o restante é combinado como conjunto de treinamento e o processo é repetido k vezes. Então, o modelo terá classificado todos os dados disponíveis, assim, os resultados de cada interação compõem o desempenho final do modelo.

2.6. Medidas de Desempenho

As medidas de desempenho são indicadores de regressão, são usados para medir a qualidade de um modelo de regressão e seu ajuste aos dados verdadeiros (HARRISON, 2020).

2.6.1. Root Mean Square Error

RMSE (*Root Mean Square Error*), ou Raíz do Erro Quadrático Médio, é uma métrica usada para avaliar o desempenho de modelos de regressão. Essa métrica fornece uma medida da diferença entre os valores observados e os valores previstos pelo modelo. A Equação (6) expressa o valor do RMSE.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}, \tag{6}$$

Em que n é a quantidade de amostras; Y é o valor real; $\hat{Y}$ é o valor previsto pelo modelo.

O RMSE calcula a raiz quadrada da média dos quadrados das diferenças entre os valores observados e os valores previstos. Essa métrica penaliza mais fortemente erros maiores, tornando-se sensível a valores discrepantes.

Assim, podemos interpretar o valor do RMSE da seguinte forma, quanto menor o valor, melhor o modelo está em fazer previsões assertivas (IZBICKI; SANTOS, 2020).

2.6.1. Coeficiente de Determinação

O Coeficiente de Determinação, denotado por R^2, é uma métrica estatística, usada para avaliar a qualidade de um modelo de regressão. Ele fornece uma medida da proporção da variabilidade na variável dependente, que pode ser explicada pelo modelo. Assim, calculamos o R^2 conforme a Equação (7).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}, \tag{7}$$

Em que n é a quantidade de amostras; Y é o valor real; $\hat{Y}$ é o valor previsto pelo modelo; $\bar{Y}$ é o valor médio dos valores reais.

O valor de R^2 varia de 0 a 1, 1 indica que o modelo explica toda a variabilidade na variável dependente, 0 indica que o modelo não explica nada de variabilidade dependente, $0 < R^2 < 1$ indica a proporção de variabilidade explicada pelo modelo.

2.6.1. MAPE

O MAPE (*Mean Absolute Percentage Error*), ou Erro Percentual Médio Absoluto, é uma métrica comumente utilizada para avaliar a precisão de modelos de previsão. O MAPE mede a média percentual das diferenças absolutas entre os valores observados e os valores previstos.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{(Y_i - \hat{Y}_i)}{Y_i} \times 100, \quad (8)$$

Em que n é o número de amostras; Y é o valor real; $\hat{Y}$ é o valor previsto pelo modelo.

O MAPE expressa os erros em porcentagem, tornando-o uma métrica intuitiva. Quanto menor o valor do MAPE, melhor é o desempenho do modelo – indicando uma menor média percentual dos erros.

3. MATERIAL E MÉTODOS

Este capítulo aborda a metodologia utilizada para obtenção dos resultados do trabalho. São apresentadas informações sobre o parque eólico analisado, o pré-processamento dos dados, a construção do modelo preditivo e a obtenção das métricas de avaliação. Para melhor entendimento do leitor, o fluxograma representado na Figura 8, representa as etapas deste processo.

Figura 8 – Fluxograma das etapas da metodologia utilizada no trabalho.

Fonte: Autoria Própria (2023).

Nos subtópicos seguintes, são apresentadas informações detalhadas para cada etapa da metodologia.

3.1. Descrição do Parque Eólico Analisado

Os dados usados como base para o estudo foram obtidos do parque eólico real, localizado no Nordeste do Brasil, na Microrregião do Agreste meridional a 254 km do Recife. Este parque possuí 141 MW de capacidade instalada, distribuída em um conjunto de 75 aerogeradores, dos quais 11 estão no conjunto estudado. O conjunto estudado possui uma Torre de Medição Anemométrica (TMA) situada em Latitude 8°54'16.07"S, Longitude 36°43'39.45"O, identificada como AMA VII, com três anemômetros de copo, dois *windvane,* um termo-higrômetro, e um barômetro. Esses instrumentos possuem certificados de calibração reconhecidos por instituições internacionais.

A capacidade do conjunto escolhido é de 25,3 MW. Os aerogeradores possuem 2,3 MW de potência nominal, produzidos pela General Electric, plataforma 2.X. Esta

turbina tem uma altura de *hub* de 80,0 m e diâmetro de rotor de 116 m. Na Figura 9 apresenta uma das turbinas do conjunto.

Figura 9 – Aerogerador Plataforma 2.X.

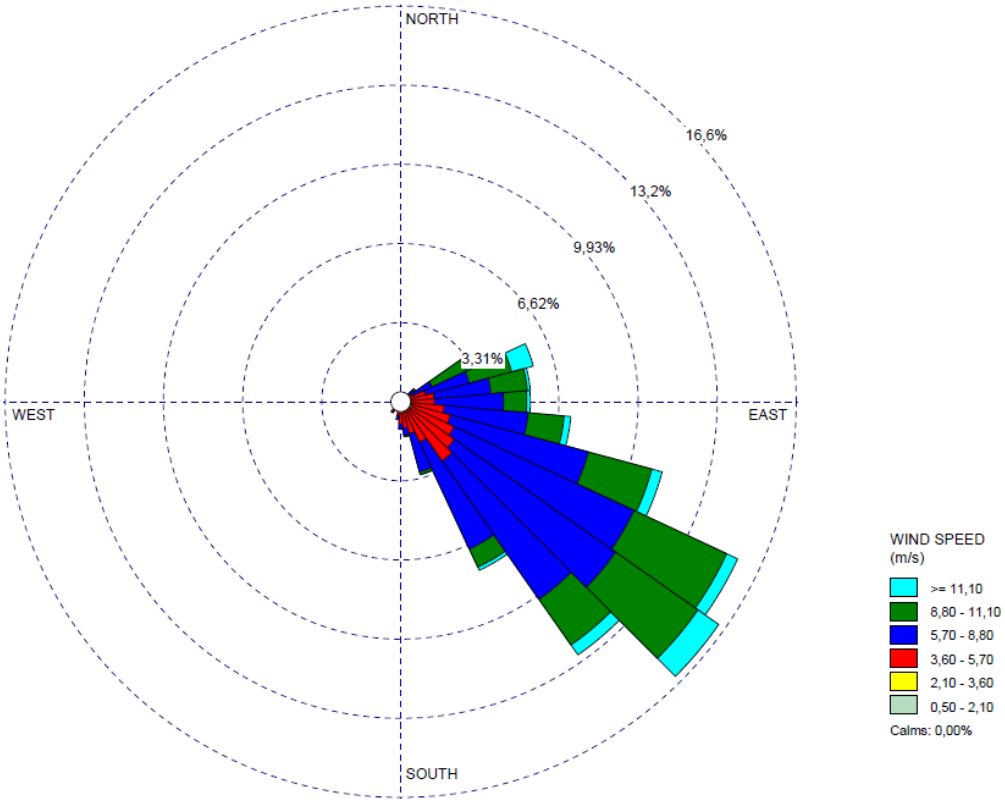
Fonte: Autoria Própria.

A distribuição das turbinas e da TMA pode ser observada na Figura 10. O complexo possui direção predominante de Leste-Sudeste (LSE), conforme mostrado na rosa dos ventos na Figura 11, esta direção predominante de vento dá-se pela influência dos ventos alísios.

Figura 10 – Distribuição do conjunto de aerogeradores.

Fonte: Adaptado de Google Earth (2023).

Figura 11 – Rosa dos Ventos.

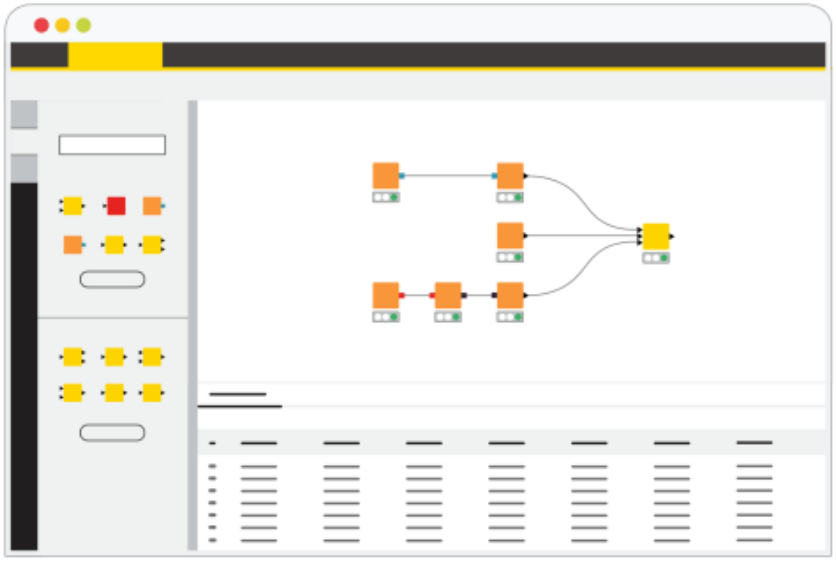


Fonte: Autoria Própria.

3.2. Ambiente Computacional

Como recurso computacional para execução da metodologia foi utiliza a plataforma KNIME (*Konstanz Information Miner*), que é uma plataforma de análise de dados de código aberto e uma ferramenta de análise preditiva. Ela fornece uma interface gráfica amigável que permite aos usuários criarem fluxo de trabalhos analíticos, integrando diversas etapas de pré-processamento de dados, análise estatística, mineração de dados e visualização. Na Figura 12 apresenta uma ilustração do ambiente computacional do KNIME.

Figura 12 – Ilustração ambiente KNIME.



Fonte: (KNIME, 2023).

O KNIME oferece uma interface gráfica intuitiva e baseada em *point-and-click*, assim facilitando a criação dos fluxos de trabalho. Uma das vantagens da metodologia é ser *codeless*, assim, mesmo sem conhecimento profundo em programação pode usá-lo para desenvolvimento.

O KNIME oferece integração com várias ferramentas e extensões, permitindo a incorporação de algoritmos de aprendizado de máquina populares, como Scikit-Learn, TensorFlow e outros algoritmo de análise de dados.

Sua função neste trabalho vai desde o tratamento de dados, passando pelo treinamento do modelo até a avaliação dos resultados, o que justifica sua escolha, minimizando todas as interações em um único ambiente. Essas análises demandam

um esforço computacional. Com isso, foi utilizada um *hardware* cuja especificações estão presentes na Tabela 1.

Tabela 1 – Especificações do *hardware*.

Componente	Descrição
CPU	• AMD Ryzen™ 7-5800H • Clock de até 4,4GHz • 8 Núcleos de processamento
GPU	• NVIDIA® GeForce® RTX 3050 4GB
Memória RAM	• 16 GB • 3,2 GHz
Armazenamento	• 1 TB tipo SSD.

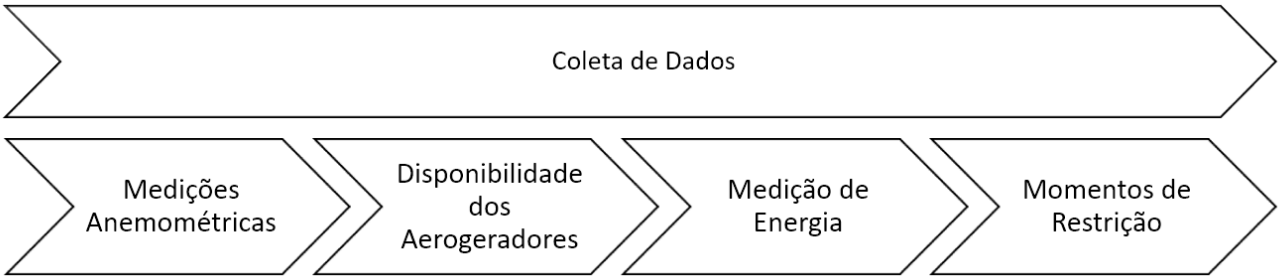
Fonte: Autoria Própria.

3.3. Pré-processamento

3.3.1. Coleta dos Dados

Para considerar o intervalo de um ano completo, o período dos dados escolhido foi do ano de 2022, contemplando de 00:00h 01/01/2022 até 23:50 31/12/2023. As subseções abaixo descrevem o processo de obtenção dos dados analisados para construção do modelo, conforme Figura 13.

Figura 13 – Fluxograma de coleta de dados.



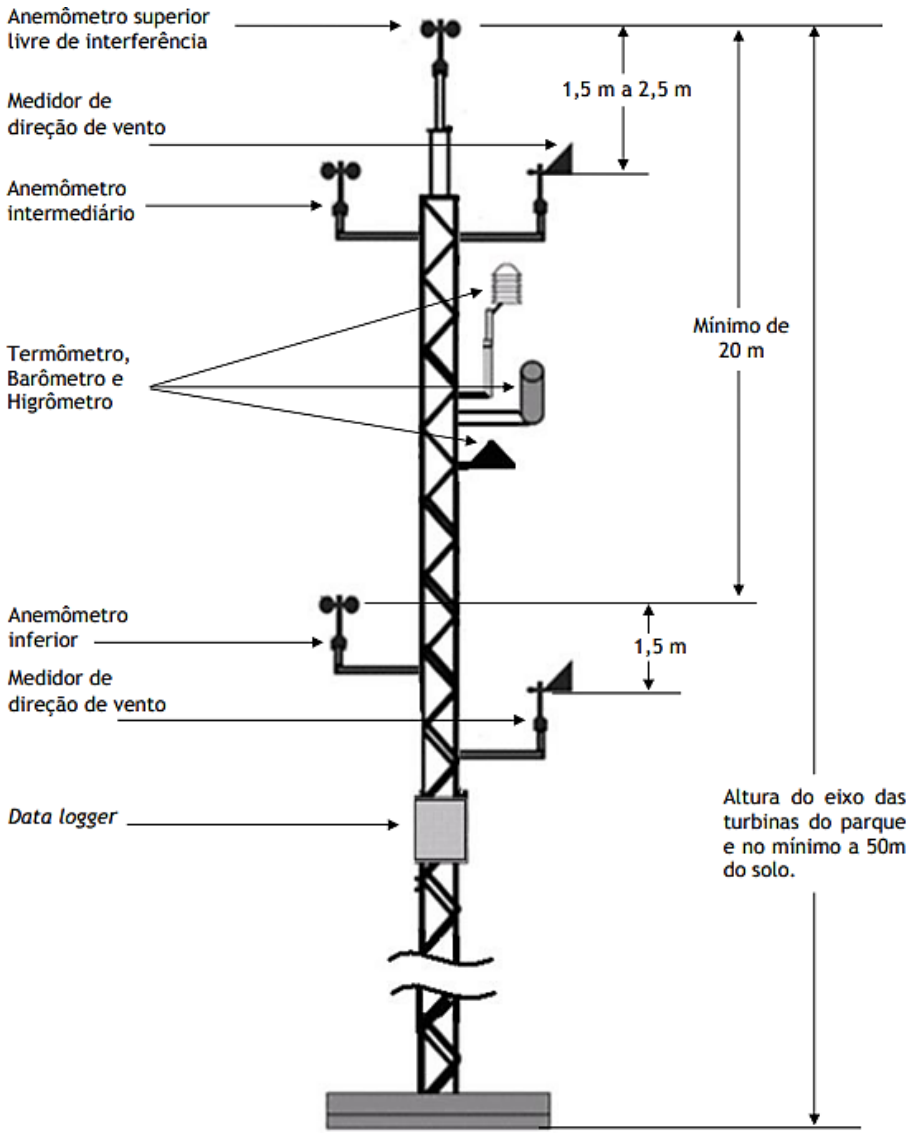
Fonte: Autoria Própria.

3.3.1.1. Medições Anemométricas

Conforme comentado, os dados ambientais foram originados da torre de medição anemométrica (TMA) proveniente da TMA VII, a qual está representada na Figura 14. Estes dados foram obtidos em intervalos que representam as médias de 10 minutos, com intervalo de integração de um segundo. O *data logger* responsável pelo armazenamento dos dados é o EOL Zenith, de fabricação da Kintech

Engineering, representado na Figura 15. A relação dos sensores, suas alturas e variáveis medidas estão descritas na Tabela 2.

Figura 14 – Ilustração Torre de Medição Anemométrica.



Fonte: EPE (2014).

Figura 15 – *Data Logger* EOL Zenith.

Fonte: Kintech Engineering (2023).

Tabela 2 – Relação dos sensores.

Equipamento	Altura (m)	Marca/Modelo	Medição
Anemômetro de copo 1	80	Thies Clima/First Class Advanced	Velocidade do Vento (m/s)
Anemômetro de copo 2	78	Thies Clima/First Class Advanced	
Anemômetro de copo 3	22	Thies Clima/First Class Advanced	
Windvane 1	78	Thies Clima/Thies TMR K360V	Direção do Vento (Graus)
Windvane 2	21	Thies Clima/Thies TMR K360V	
Termo-higrômetro	78	Galtec/KPC 1/5	Temperatura (ºC) e Umidade (%)
Barômetro	11	Setra/Setra 276 800-1100	Pressão (hPa)

Fonte: Autoria Própria.

Os dados foram coletados do *data logger* através do software Atlas, da própria Kintech Engineering. Após coleta, os dados foram salvos na extensão “.wnd”, que pode ser facilmente aberta com um leitor de texto padrão. Após coleta, os arquivos são agrupados em intervalo de meses, para compilá-los em um único arquivo na extensão “.csv” foi utilizado o código apresentado no APÊNDICE A.

3.3.1.2. Dados de Geração

A geração proveniente deste grupo de turbinas escoa através de uma Rede de Média Tensão (RMT) para subestação elevadora. A energia transmitida por este circuito da RMT é medida por um Sistema de Medição de Faturamento (SMF). O medidor responsável por essa medição é o ION8650c, fabricado pela Schneider

Electric, representado na Figura 16. Este medidor armazena as médias de 5 minutos, com intervalo de integração de 1.024 amostras por ciclo.

Figura 16 – Medidor ION8650c.

Fonte: Schneider Electric (2023).

Os dados de geração são coletados diariamente pela Câmara de Comercialização de Energia Elétrica (CCEE) e compilados em médias, no intervalo de uma hora. Assim, os dados de geração foram obtidos através do modulo de Sistema de Coleta de Dados de Energia (SCDE) da CCEE, em intervalo de uma hora.

3.3.1.3. Dados de Disponibilidade

A energia gerada também depende do número de aerogeradores disponíveis, assim os dados de disponibilidade também foram considerados para construção da base de dados.

Estes dados foram obtidos através do WindSCADA, que nada mais é que o supervisório que permite o gerenciamento dos aerogeradores. A obtenção desses dados foi feita através das medições registradas pelo *Wind Farm Management System* (WFMS) que gerencia e integraliza várias variáveis do conjunto de turbinas, dentre elas está a informação do número de turbinas disponíveis para geração.

3.3.1.4. Momentos de Restrição de Energia

Os momentos de restrição de energia são instantes importantes pois auxiliam a remover da base de dados conjunta os momentos em que as variáveis do ambiente pouco têm influência na produção de energia do parque, já que a potência máxima está limitada a um valor definido pelo ONS.

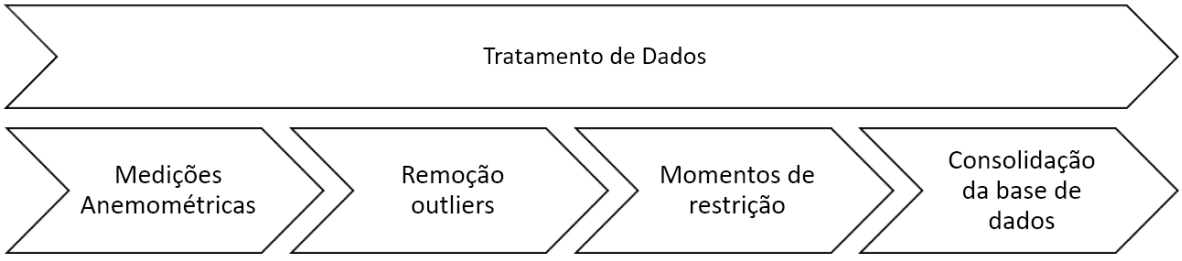
A base de dados com esses momentos é disponibilizada através do Sistema de Apuração de Geração (SAGER).

3.3.2. Tratamento dos Dados

Após a conclusão da fase de coleta de dados, inicia-se uma etapa crucial no fluxo de trabalho, que é o processo de tratamento de dados. Essa fase representa um conjunto de ações e métodos elaborados com o propósito de aprimorar tanto a confiabilidade quanto a qualidade dos dados previamente coletados.

Nesse estágio, diversos procedimentos são implementados para assegurar que as informações obtidas sejam precisas, consistentes e livres de possíveis distorções ou anomalias. Isso envolve a identificação e correção de erros, a padronização de formatos, a normalização de valores e, quando necessário, a exclusão de dados inconsistentes. Um esboço do processo pode ser encontrado na Figura 17.

Figura 17 – Fluxo do tratamento de dados.



Fonte: Autoria Própria.

3.3.2.1. Medições Anemométricas

O tratamento prévio das medições foi realizado mediante a aplicação de um pré-filtro, que engloba os limiares estabelecidos EPE para cada variável. Esse pré-filtro desempenha um papel fundamental na preparação dos dados, uma vez que estabelece critérios específicos de aceitação para as medições, conforme definido pelos padrões e normas estabelecidos pela EPE.

Ao empregar esse pré-filtro, é possível realizar uma triagem dos dados, removendo potenciais valores discrepantes ou imprecisos que possam comprometer a qualidade e confiabilidade das medições. Os limiares estabelecidos pela EPE servem como referência para determinar a validade e consistência das observações,

contribuindo para a obtenção de uma base de dados mais consistente e alinhada com os requisitos e padrões estabelecidos pela entidade.

Dessa forma, o pré-filtro representa uma etapa essencial no processo de tratamento dos dados, assegurando que apenas as medições que atendam aos critérios estabelecidos pela EPE sejam incluídas na análise subsequente. Isso promove a integridade e confiabilidade dos dados, sendo crucial para garantir resultados robustos e precisos na construção do modelo. Os limiares podem ser encontrados na Tabela 3.

Tabela 3 – Limiares EPE.

Medição	Limiar Mínimo	Limiar Máximo
Pressão atmosférica (hPA)	800	1.060
Temperatura (ºC)	-15	50
Umidade do ar (%)	0	110
Direção do Vento	0	360
Velocidade média do vento (m/s)	0	50
Velocidade máxima do vento (m/s)	0	70

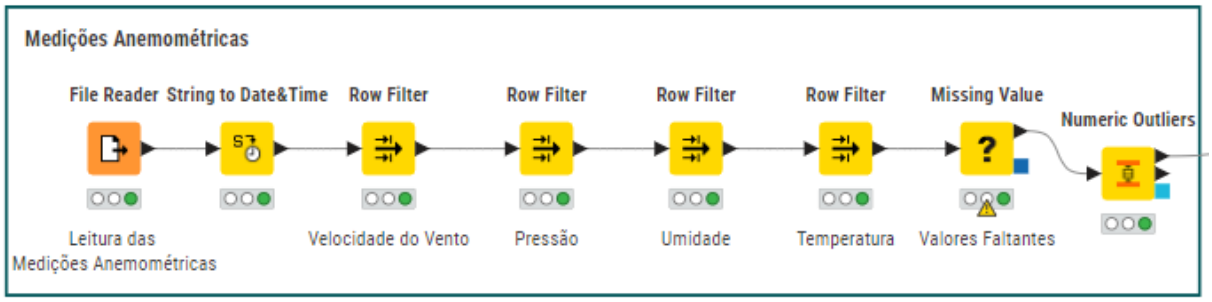
Fonte: Adaptado (EMPRESA DE PESQUISA ENERGÉTICA, 2014).

Após aplicação destes filtros, podem surgir lacunas deixadas por valores foras destes padrões. Para solução dessas lacunas, optou-se pela aplicação do método de média móvel, considerando valores uma hora antes e uma hora depois, efetuando o preenchimento dessas lacunas de forma temporalmente coerente.

Posteriormente, reconhecendo a possibilidade da introdução de valores atípicos durante ou após o processo de preenchimento, propôs-se a aplicação de um filtro de *outlier*. Essa abordagem visa não apenas a identificação e remoção de outliers originais, mas também a detecção de possíveis valores discrepantes que possam surgir como resultado do tratamento de lacunas. A aplicação desses passos visa garantir que os dados anemométricos, utilizados nas análises subsequentes, sejam, não apenas, conformes às normativas estabelecidas pela EPE, mas também representativos, completos e livres de distorções que possam comprometer a sua validade.

O fluxograma do processo implementado no KNIME pode ser observado na Figura 18.

Figura 18 – Fluxograma Tratamento Dados Anemométricos.



Fonte: Autoria Própria.

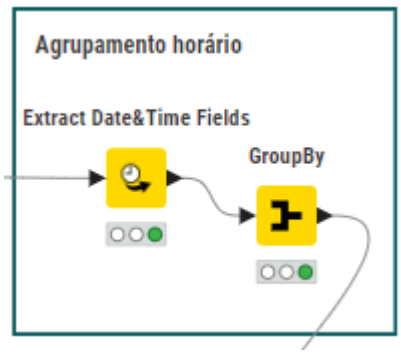
3.3.2.2. Disponibilidade e Geração

É fundamental observar que as medições de disponibilidade e energia produzida contam com uma confiabilidade já garantida tanto pelo equipamento utilizado quanto pelos processos regulares de correção e realizados mensalmente, bem como as aplicações de medidores principais e de retaguarda. Essa abordagem reforça a qualidade dos dados utilizados nas análises, proporcionando uma base sólida para interpretações precisas e decisões informadas.

3.3.3. Consolidação da Base de Dados

Antes de consolidar todas as bases de dados, é importante notar que as medições anemométricas foram registradas em intervalos que contemplam a média de 10 minutos, enquanto as de geração e disponibilidade foram computadas em médias de uma hora. Assim para transformar as médias de 10 minutos em médias horárias, usou-se o fluxograma no KNIME conforme representado na Figura 19.

Figura 19 – Agrupamento por hora.



Fonte: Autoria Própria.

Então, para agrupar as médias em intervalos de uma hora, primeiro extrai-se do campo onde estão as informações de data e hora, criando colunas de ano, mês dia e hora, então agrupa-se esses valores do intervalo de uma hora, calculando as médias das linhas correspondentes agregadas.

Após os passos descritos, pode-se prosseguir para a etapa de agrupamento de dados, unindo os dados de geração, medições anemométricas, disponibilidade e restrição de energia em uma única base, agrupando-os pela coluna horária. Após a consolidação das bases, aplica-se um filtro para retirar horários em que houve restrição da geração. Assim, conclui-se a etapa de pré-processamento dos dados.

3.4. Seleção de Variáveis

O Processo de seleção de variáveis dá-se pela escolha dos valores mais correlatos com a variável *target,* que é a energia gerada em MWh, utilizou-se como métrica de avaliação o coeficiente de relação de Spearman. A escolha de variáveis baseada no coeficiente de relação de Spearman oferece uma abordagem que vai além das correlações lineares tradicionais, possibilitando a detecção de associações mais complexas e não lineares entre as variáveis explicativas e a variável *target*, assim, foram escolhidas as variáveis que apresentam $|\rho| \geq 0{,}7$.

Ao adotar essa estratégia, busca-se garantir que as variáveis selecionadas para análise estejam intrinsecamente ligadas à variável *target*, proporcionando, assim, uma base sólida para a construção do modelo.

3.5. Treinamento do Modelo

Antes do processo de treinamento do modelo é necessário realizar a normalização dos dados. Foi utilizada a normalização unitária conforme a Equação (4).

Após a definição das variáveis que comporão o modelo, visando alcançar o *target*, é realizada a remoção das colunas com variáveis que não serão úteis para o treinamento, seja por apresentar baixa correlação, seja porque são variáveis auxiliares utilizadas para cálculos intermediários.

Para o processo de otimização dos hiperparâmetros, o método de *brute force* é utilizado para determinar os hiperparâmetro mais adequados para o algoritmo *Random Forest.* O funcionamento deste método é discutido na subseção 2.4.2 e os parâmetros utilizados no processo são apresentados na Tabela 4.

Tabela 4 – Parâmetros utilizados no método de Força Bruta.

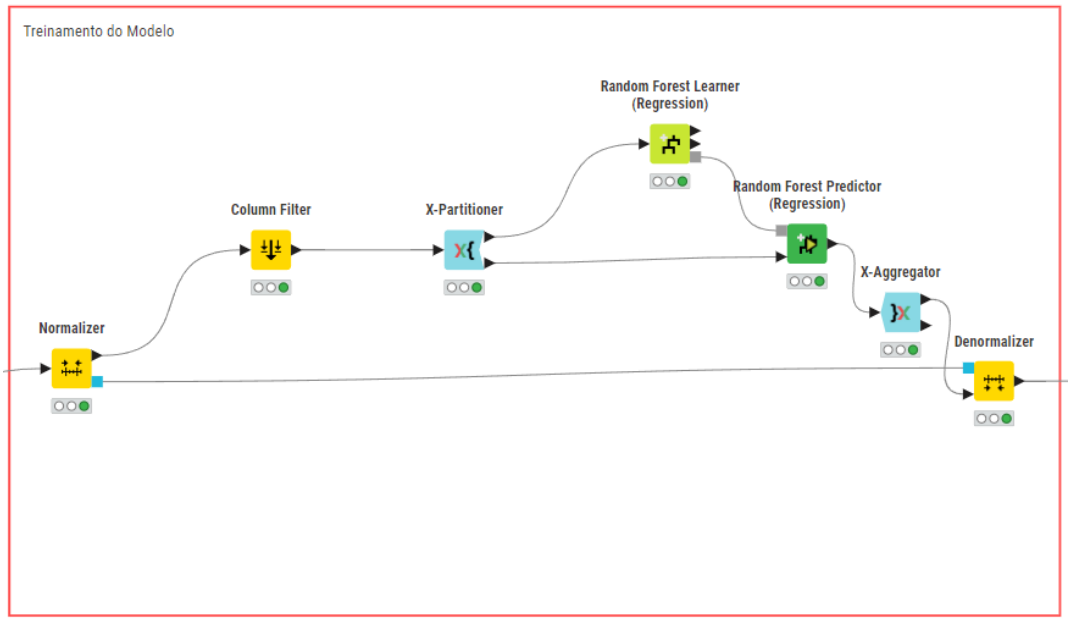
Parâmetros	Limite inferior	Limite superior	Tamanho do passo
Número de níveis das árvores	2	20	1
Número de árvores	100	500	1

Fonte: Autoria Própria.

Os limites inferiores e superiores apresentados na Tabela 4. Foram definidos conforme as recomendações de Kunapuli.

Por fim, após a otimização do hiperparâmetro, define-se o conjunto de dados para treinamento e o conjunto para teste, que serão aplicados na etapa de validação cruzada. Na validação cruzada o conjunto de dados é dividido na proporção de 80% para treinamento e 20% para teste. Após treinamento e previsão, os valores são desnormalizados para que o resultado possa ser utilizado no cálculo das métricas de desempenho do modelo. Na Figura 20 pode-se observar o fluxograma de treinamento conforme.

Figura 20 – Fluxograma de treinamento do modelo.



Fonte: Autoria Própria.

3.6. Avaliação do Modelo

A avaliação dos resultados do modelo foi realizada por meio da análise das métricas de erro definidas: o coeficiente de determinação (R^2), o MAPE e o RMSE.

Essas métricas desempenham o papel de mensuração da acurácia do modelo, proporcionando *insights* sobre sua capacidade de explicar a variação nos dados, bem como sua exatidão em termos de previsão. O coeficiente de determinação avalia a proporção da variabilidade da variável dependente que é explicada pelo modelo, enquanto o MAPE e o RMSE oferecem informações sobre a acurácia e a qualidade das previsões, respectivamente.

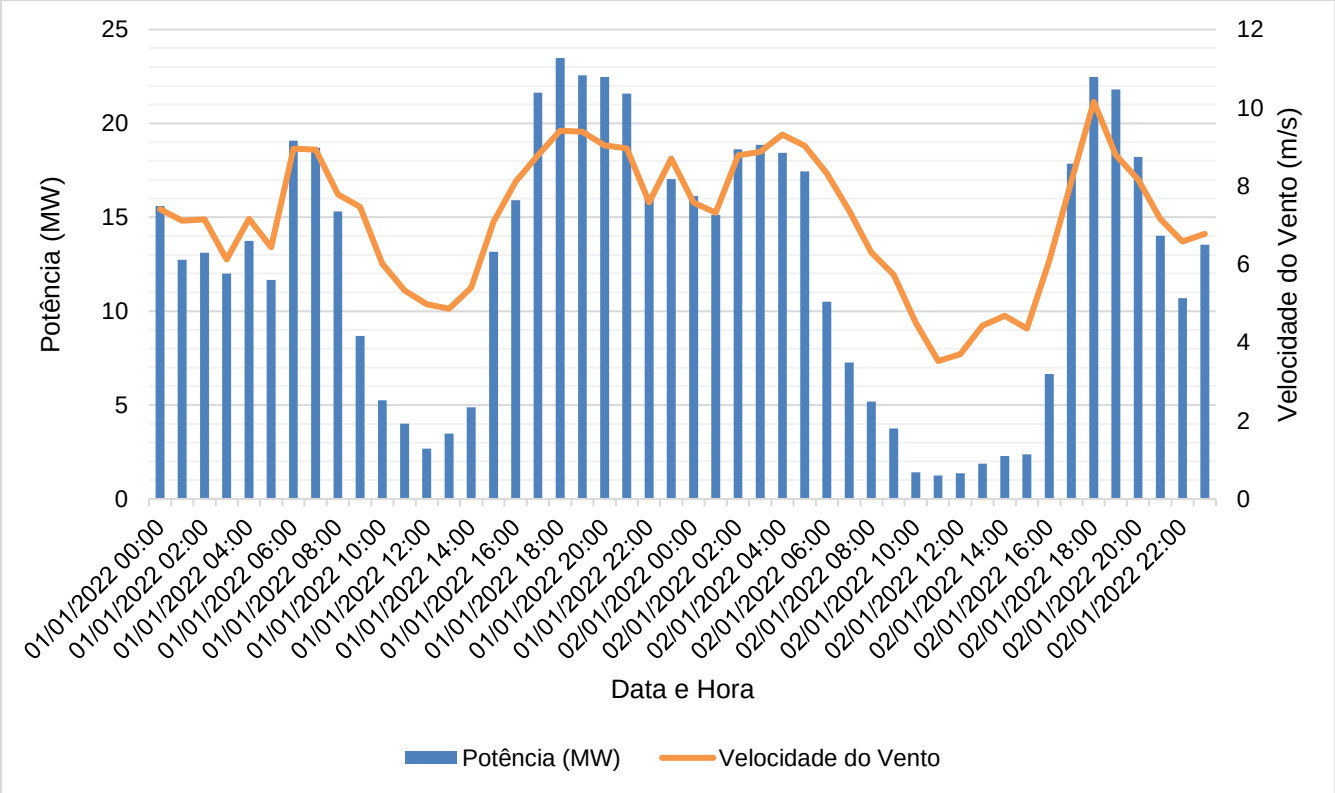
4. RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados obtidos a partir do processo de tratamento de dados, passando pela seleção de variáveis, treinamento e avaliação do modelo, e finaliza apresentando a geração frustrada estimada para o parque eólico em análise.

4.1. Análise Preliminar dos Dados

Inicialmente, foi realizada uma análise preliminar dos dados obtidos após a etapa de pré-processamento, com objetivo de descobrir e analisar informações relevantes nos dados, como aquela apresentada no gráfico da Figura 21, que correlaciona a potência gerada no parque com a velocidade do vento do horário.

Figura 21 – Velocidade do vento e potência



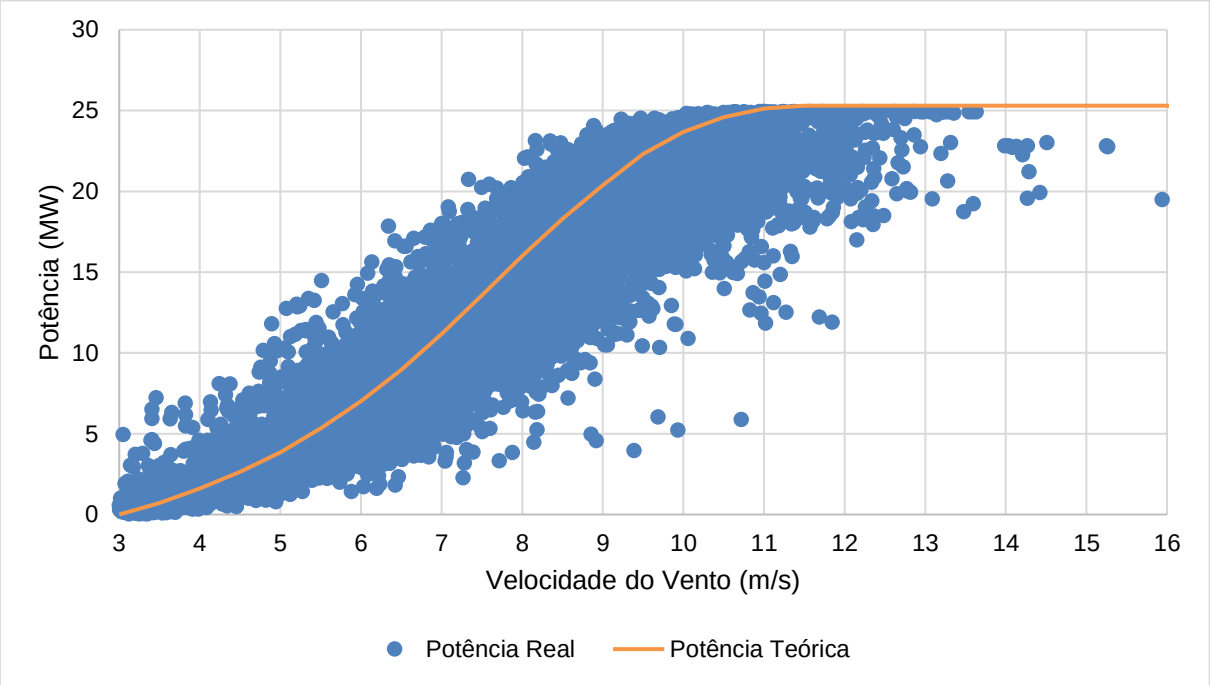
Fonte: Autoria Própria.

A análise da Figura 21 demonstra o comportamento da a velocidade do vento e a potência do parque no intervalo de um dia. A partir da visualização, é possível

observar que os dados de energia gerada e velocidade do vento foram agrupados de forma correta. Esse comportamento pode-se ser explicado pela Equação(1), sendo a potência diretamente proporcional ao cubo da velocidade.

A partir dos dados obtidos também foi possível determinar a curva de potência do parque eólico, conforme gráfico apresentado na Figura 22.

Figura 22 – Curva Potência x Vento.



Fonte: Autoria Própria.

Pode-se notar uma correlação entre a curva de potência real e a curva de potência teórica na Figura 22. Ao ajustar realizar um ajuste nas curvas, obtém-se o valor de R^2=0,842 para potência real e R^2=0,958, estes valores indicam que sua dispersão está bem agrupada e com poucas variações, o que indica uma boa qualidade nos dados. A variação dispersão mais próxima à curva ideal pode se dá pela variação de temperatura ou pressão, conforme as Equações (2) e (1), presentes na subseção 2.1.2. Já os pontos mais afastados do agrupamento, podem ser explicados pela variação da disponibilidade dos aerogeradores.

Após a análise preliminar dos dados, foi realizada a seleção das variáveis mais correlacionadas com a variável *target* no estudo, e os resultados desta etapa são apresentados na seção seguinte.

4.1.1. Seleção de Variáveis

Na seleção de variáveis, inicialmente, verificou-se a correlação das variáveis entre si, e suas correlações com a variável *target*, a qual foi nomeada de MED_G, que representa a geração bruta do conjunto, pela grande quantidade de variáveis candidatas a matriz de correlação com a variável *target* é apresentada no APÊNDICE C. O resultado da correlação em termos do coeficiente Spearman, para as variáveis selecionadas, pode ser visualizado na Tabela 5. Estes valores foram obtidos conforme procedimento descrito na subseção 2.3.5. e na Equação (4).

Tabela 5 – Correlação obtida pelo coeficiente de relação Spearman para variável MED_G.

Variável	ρ
Velocidade do Vento (80m)	0,977
Direção do Vento	-0,221
Pressão	-0,732
Umidade	0,636
Temperatura	-0,639
Disponibilidade	0,830

Fonte: Autoria Própria.

Pode-se observar uma forte correlação entre as variáveis escolhidas e a energia gerada, assim, implicando em uma possível explicação da variável *target,* através das variáveis correlatas, conforme discutido na subseção 2.1.1.

A partir da análise do resultado na Tabela 5 é possível notar que a influência das variáveis anemométricas na variável *target*, o que é explicado na subseção 2.1.2, como também a influência da disponibilidade de turbinas. Das variáveis apresentadas no APÊNDICE C que não estão na Tabela 5, destaca-se a baixa correção com a variável *target*, exceto pela Direção do Vento, representada pelo valor do Windvane em graus. Esta variável foi escolhida pela influência da direção do vento na energia gerada, por ser apresentada em graus sua variação de forma não linear. Ao avaliar o desempenho do modelo sem essa variável, percebe-se uma queda no seu desempenho.

4.2. Treinamento e Avaliação do Modelo

Conforme subseções 2.4 e 2.6, foi definido os hiperparâmetros ótimos do algoritmo escolhido com aplicação do método de Força Bruta*,* esse método foi

escolhido devido a facilidade de implementado e pelo seu custo computacional ser suportado pelo *hardware*, a métrica de desempenho usada para seleção dos valores ótimo foi RMSE. Os resultados da otimização são apresentados na Tabela 6.

Tabela 6 – Hiperparâmetros ótimos para o algoritmo.

Hiperparâmetro	Valor ótimo
Número de árvores	100
Número de níveis das árvores	5

Fonte: Autoria Própria.

Ao aplicar os hiperparâmetros ótimos e o conjunto de variáveis selecionadas na etapa anterior, foi realizado o treinamento do modelo. Então pode-se prosseguir para a avaliação de desemprenho a partido do processo de validação cruzada *k-fold,* conforme comentado na subseção 2.6. O resultado desse processo é exposto na Tabela 7.

Tabela 7 – Resultado da validação cruzada para o modelo.

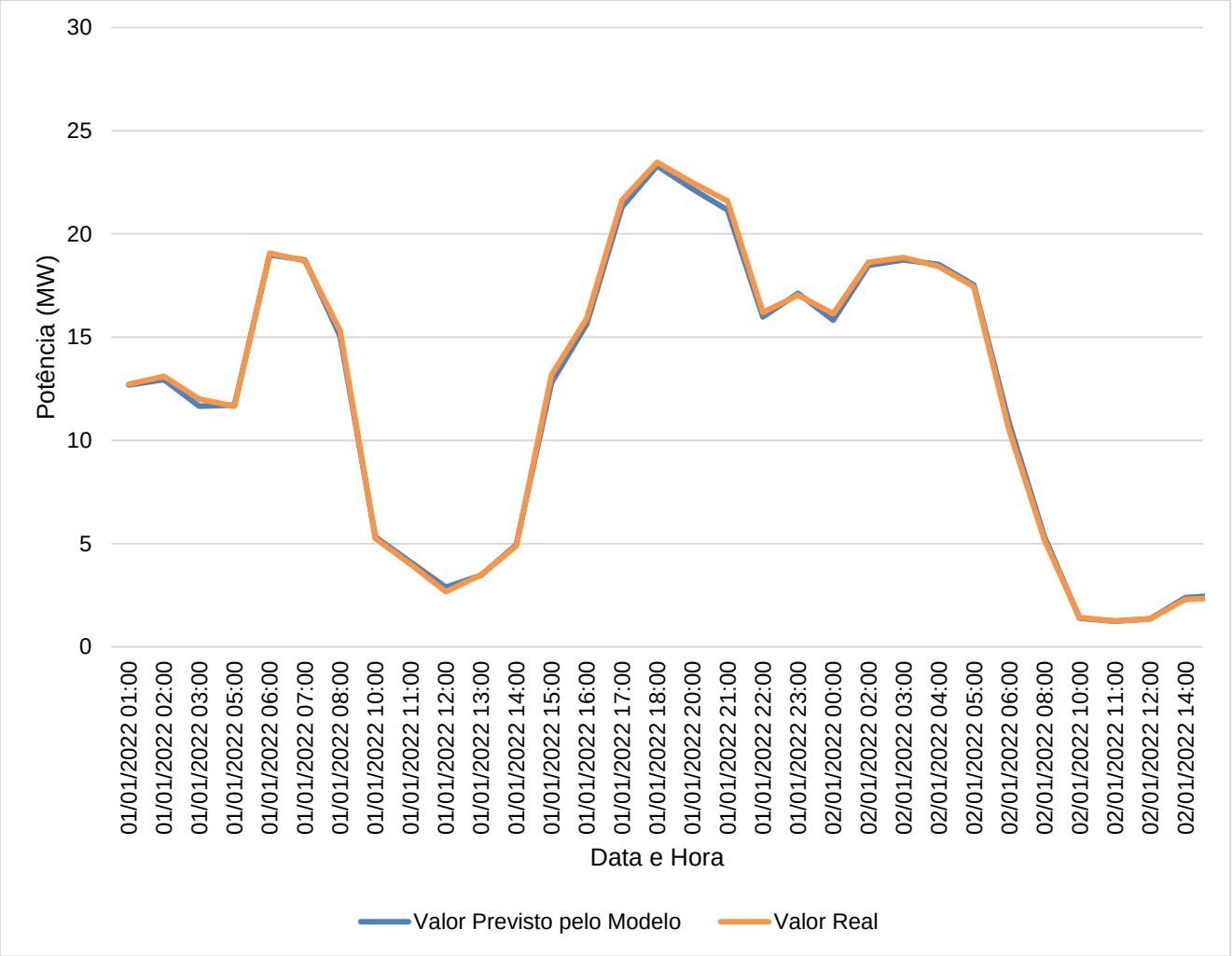
Indicador	Valor
R^2	0,983
RMSE	0,354 MWh
MAPE	3,8 %

Fonte: Autoria Própria.

A partir dos valores da Tabela 7, é possível notar que o coeficiente de determinação (R^2) demonstra uma excelente adaptação do modelo aos dados verdadeiros, sugerindo que aproximadamente 98,3% da variabilidade nos dados é explicada pelo modelo. Um RMSE de 0,354 MWh indica que, em média, as previsões do modelo estão a cerca de 0,354 MWh dos valores reais, média dos valores reais é de 11,5 MWh. Quanto menor o RMSE, melhor o modelo está em termos de precisão. O MAPE de 3,8% indica que, em média, as previsões do modelo apresentam um desvio de 3,8% em relação aos valores reais.

Os resultados das métricas de erro demonstram o excelente desempenho do modelo para a tarefa de predição da energia gerada no parque eólico. Uma ilustração visual da exatidão do modelo pode ser observada na Figura 23, de modo a complementar a análise dos resultados.

Figura 23 – Curva de valores previstos e valores reais.



Fonte: Autoria Própria.

A análise da Figura 23 demonstra mais vez que o modelo de Random Forest está desempenhando muito bem na tarefa para a qual foi treinado, proporcionando previsões exatas e explicando a maioria da variabilidade nos dados.

4.3. Determinação da Geração Frustrada

Uma vez que a validade do modelo já foi demonstra com os resultados apresentados anteriormente, foi possível utilizá-lo para prever a geração nos períodos correspondentes as restrições de geração estabelecidas pelo ONS. Assim, é possível determinar a geração frustrada no período subtraindo dos valores previstos os valores de geração efetivamente registrados. A Tabela 8 apresenta os resultados obtidos para todos os períodos de restrição registrados no ano de 2022 para o parque eólico em questão.

Tabela 8 – Resultado dos valores estimados de geração frustrada.

Data	Geração Prevista (MWh)	Geração Realizada (MWh)	Geração Frustrada (MWh)	Geração Frustrada (%)
03/04/2022	427,76	294,26	133,50	31,21%
01/05/2022	405,09	333,93	71,16	17,57%
07/08/2022	244,02	243,76	0,26	0,11%
04/09/2022	423,2	417,53	5,67	1,34%
09/10/2022	338,72	337,63	1,09	0,32%
25/12/2022	404,47	387,82	16,65	4,12%
Total	2.243,26	2.014,93	228,73	10,20%

Fonte: Autoria Própria.

A partir da análise dos dados apresentados na Tabela 8 constata-se uma frustração na geração de cerca de 228,37 MWh, o que representa 10,20% para geração prevista no período 2.243,26. O dia com maior perda foi 03/04/2022, onde observa-se uma perda percentual de 31,21%, representando quase 1/3 da energia prevista para o dia. As horas em restrição contabilizaram um total de 32 horas.

Considerando o valor de venda de energia de R$ 212,37/MWh, estima-se que essa energia perdida impacta num valor monetário de, aproximadamente, R$ 48.575,39.

Até a presente data, no ano de 2023 estimam-se um total de 477 horas em restrição em todos os 6 parques presentes no complexo em questão. Isso representa um aumento de 1490% em relação à 2022. Estima-se o impacto monetário da ordem de milhões de reais.

5. CONCLUSÕES

A criação do modelo proposta foi desenvolvida e validada na pesquisa, contribuindo para a criação de uma metodologia de cálculo de perdas de energia. O resultado obtido pode auxiliar os produtores de energia eólica a mensurarem suas perdas no intuito de ter um método comparativo ao do ONS. Assim, possibilitando aos produtores a possibilidade de arguição junto ao ONS quanto aos valores a serem restituídos.

Os resultados obtidos na avaliação do modelo de Random Forest são promissores, refletindo um desempenho excepcional na tarefa de previsão. O coeficiente de determinação elevado, alcançando 0,983, evidencia que o modelo é capaz de explicar a variabilidade nos dados, indicando um ajuste robusto e consistente. A baixa magnitude da Raíz do Erro Quadrático Médio, que atinge 0,354 MWh, denota uma precisão nas previsões em termos absolutos. Isso sugere que, em média, as estimativas do modelo estão extremamente próximas dos valores reais, ressaltando a eficácia do algoritmo Random Forest na modelagem. Além disso, o Erro Percentual Absoluto Médio de 4,5% confirma a precisão do modelo em termos relativos, indicando uma discrepância média de apenas 4,5% entre as previsões e os valores reais. Essa baixa taxa de erro percentual destaca a confiabilidade e acurácia do modelo na projeção das variáveis alvo.

Diante desses resultados, podemos concluir que o modelo de Random Forest demonstra um desempenho excelente na tarefa de previsão, oferecendo resultados confiáveis e precisos. Essas métricas robustas fortalecem a confiança na capacidade do modelo.

Algumas sugestões para trabalhos futuros:

- Avaliar custos e retornos financeiros considerando a implantação de uma usina de hidrogênio verde junto a uma termoelétrica para o escoamento da energia, limitada para a geração de hidrogênio, e seu uso nos momentos de baixa geração.
- Realizar método comparativo ao imposto pelo ONS para previsão de energia perdida, avaliando a precisão do modelo de cálculo de perdas impostos pelo ONS.

- Explorar o desempenho do modelo de previsão de energia em conjunto com um modelo de previsão de vento, com o intuito de viabilizar a venda de energia no mercado de curto prazo.

REFERÊNCIAS

ABEEÓLICA. **ABBEólica**, 2023. Disponivel em: https://abeeolica.org.br/. Acesso em: 18 nov. 2023.

AGÊNCIA INFRA. iNFRA. **AGENCIAINFRA.COM**, 2023. Disponivel em: https://www.agenciainfra.com/blog/eolicas-acumulam-prejuizos-de-r-75-milhoes-por-restricoes-impostas-desde-o-apagao-diz-associacao/. Acesso em: 30 nov. 2023.

ANEEL. Sistema de Informações de Geração da ANEEL SIGA, 2023. Disponivel em: https://app.powerbi.com/view?r=eyJrIjoiNjc4OGYyYjQtYWM2ZC00YjllLWJlYmEtYzdkNTQ1MTc1NjM2IiwidCI6IjQwZDZmOWI4LWVjYTctNDZhMi05MmQ0LWVhNGU5YzAxNzBlMSIsImMiOjR9. Acesso em: 16 nov. 2023.

BETZ, Albert. Das Maximum der theoretisch möglichen Ausnützung des Windes durch Windmotoren. *In:* **Zeitschrift für das gesamte Turbinenwesen**. German: [s.n.], v. 26, 1920. p. 307-309.

BURTON, Tony *et al.* **Wind Energy Handbook**. 2ª. ed. [S.l.]: John Wiley & Sons, 2011.

CBBE. CENTRO BRASILEIRO DE ENERGIA EÓLICA, 2000. Disponivel em: www.eolica.com.br.

CUNHA, João Paulo Zanola. **Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos**. Universidade de São Paulo. São Paulo. 2019.

EMPRESA DE PESQUISA ENERGÉTICA. **Leilões de Energia: Instruções para as medições anemométricas e climatológicas em parques eólicos**. EPE. Rio de Janeiro, p. 17. 2014. (NOTA TÉCNICA DEA 08/14).

ESCOVEDO, Tatiana ; KOSHIYAMA, Adriano. **Introdução a Data Science:** Algoritmos de Machine Learning e métodos de análise. [S.l.]: Casa do Códgio, 2020.

FERREIRA, Artur Gonçalves. **Meteorologia prática**. São Paulo: Oficina de Textos, 2006.

GRUS, Joel. **Data science do zero**. Rio de Janeiro: Alta Editora e Consultoria Eireli, 2016.

GSI. Segurança de Infraestruturas Críticas. **Site do Gabinete de Segurança Institucional (GSI) do governo do Brasil**, 2022. Disponivel em: https://www.gov.br/gsi/pt-br/assuntos/seguranca-de-infraestruturas-criticas-sic. Acesso em: 30 nov. 2023.

GWEC. **Global Wind Report 2022**, 2022. Disponivel em: https://gwec.net/global-wind-report-2022/. Acesso em: 15 nov. 2023.

HARRISON, Matt. **Machine Learning–Guia de referência rápida:** trabalhando com dados estruturados em Python. São Paulo: Novatec Editora, 2020.

HAWKINS, Doulgas. **Identification of outliers**. Londres: Chapman and Hall, 1980.

IBGE. CENTRO DE DOCUMENTAÇÃO E DISSEMINAÇÃO DE INFORMAÇÕES. **Normas de apresentação Tabular**. 3 ed. ed. Rio de Janeiro: IBGE, 1993.

IBM. What is Random Forest? **IBM**, 2023. Disponivel em: https://www.ibm.com/topics/random-forest. Acesso em: 25 nov. 2023.

IZBICKI, Rafael ; SANTOS, Tiago Mendonça dos. **Aprendizado de máquina:** uma abortagem estatística. 1ª. ed. São Carlos: [s.n.], 2020.

KINTECH ENGNEERING. EOL ZENITH. **Kintech Engneering**, 2023. Disponivel em: https://www.kintech-engineering.com/pt-br/catalogue/data-loggers/eol-zenith/#ywtm_2531. Acesso em: 30 nov. 2023.

KNIME. KNIME. **KNIME Analytics Platform Version 5**, 2023. Disponivel em: https://www.knime.com/knime-analytics-platform-version-5. Acesso em: 25 nov. 2023.

KUNAPULI, Gautam. **Ensemble Methods in Machine Learning**. Shelter Island, NY: Manning , 2023.

LEVENTHAL, Barry. An introduction to data mining and other techniques for advanced analytics. **Journal of Direct, Data and Digital Marketing Practice**, 2010. 137-153.

MITCHELL, Tom M. **Machine Learning**. 1ª. ed. [S.l.]: McGraw-Hill Education, 1997.

ONS. **Submódulo 10.6 - Controle de Geração**. ONS. [S.l.], p. 16. 2021.

ONS. **Submódulo 6.5**. ONS. [S.l.], p. 51. 2022.

REIS, Lineu Belico dos. **Geração de energia elétrica**. 2ª. ed. Barueri, SP: Editora Manole, v. 978-85-204-3039-2, 2011.

SCHNEIDER ELECTRIC. M8650C0C0H6E1A0A. **Schneider Electric**. Disponivel em: https://www.se.com/br/pt/product/M8650C0C0H6E1A0A/medidor-ion8650c-32mb-base-1-5a-65120vac-eth/. Acesso em: 30 nov. 2023.

SCIKIT LEARN. Scikit Learn. **3.1. Cross-validation:** evaluating estimator performance, 2023. Disponivel em: https://scikit-learn.org/stable/modules/cross_validation.html. Acesso em: 27 nov. 2023.

SILVA, Amanda Zilli da. **Constrained-off de Usinas Eólicas: análise preliminar da Consulta Pública**. Universidade Federal de Santa Caratina. Araranguá, p. 71. 2023.

SMOLA, Alex ; VISHWANATHAN, S.V.N.. **INTRODUCTION TO MACHINE LEARNING**. 2ª. ed. Cambridge: CAMBRIDGE UNIVERSITY PRESS, 2008.

APÊNDICE

APÊNDICE A – Código usado para compilar os dados.

```
import pandas as pd
import glob

# Lista todos os arquivos CSV no diretório atual
arquivos_csv = glob.glob('*.csv')

# Lista para armazenar os DataFrames de cada arquivo CSV
dataframes = []

# Loop pelos arquivos CSV e lê cada um deles em um DataFrame, ignorando
as primeiras 3 linhas e usando ';' como delimitador
for arquivo_csv in arquivos_csv:
    df = pd.read_csv(arquivo_csv, delimiter=';', skiprows=3)
    dataframes.append(df)

# Concatena os DataFrames em um único DataFrame
resultado = pd.concat(dataframes, ignore_index=True)

# Salva o DataFrame compilado em um novo arquivo CSV
resultado.to_csv('dados_compilados.csv', index=False, sep=';')

print("Arquivos CSV compilados com sucesso em 'dados_compilados.csv'.")
```

APÊNDICE B – Variáveis dos dados anemométricos.

Indicador	Medida
Battery	Tensão da bateria do *Data logger*
Anemômetro 80m	Velocidade média do vento
Anemômetro 80m Min.1	Velocidade mínima do vento
Anemômetro 80m Max.1	Velocidade máxima do vento
Anemômetro 80m TI30.1	Intensidade da turbulência
Anemômetro 78m	Velocidade média do vento
Anemômetro 78m Min	Velocidade mínima do vento
Anemômetro 78m Max	Velocidade máxima do vento
Anemômetro 22m	Velocidade média do vento
Anemômetro 22m -Min	Velocidade mínima do vento
Anemômetro 22m -Max	Velocidade máxima do vento
Windvane1	Direção do Vento
Windvane2	Direção do Vento
Barômetro	Pressão
Higrômetro	Umidade
Termômetro	Temperatura

APÊNDICE C – Variáveis candidatas e suas correlações.

Variável	**Medida**	**MED_G (Target)**
Anemômetro 80m	Velocidade média do vento	0,977
Anemômetro 80m Min.1	Velocidade mínima do vento	0,880
Anemômetro 80m Max.1	Velocidade máxima do vento	0,899
Anemômetro 80m TI30.1	Intensidade da turbulência	-0.453
Anemômetro 78m	Velocidade média do vento	0,926
Anemômetro 78m Min	Velocidade mínima do vento	0,881
Anemômetro 78m Max	Velocidade máxima do vento	0,899
Anemômetro 22m	Velocidade média do vento	0,821
Anemômetro 22m -Min	Velocidade mínima do vento	0,856
Anemômetro 22m -Max	Velocidade máxima do vento	0,789
Windvane1	Direção do Vento	-0,218
Windvane2	Direção do Vento	0,076
Barômetro	Pressão	-0,732
Higrômetro	Umidade	0,636
Termômetro	Temperatura	-0,639
MED_G	Energia Gerada	1
Disponibilidade	Disponibilidade	0,832