UNIVERSIDADE DE FORTALEZA - UNIFOR
CENTRO DE CIÊNCIAS TECNOLÓGICAS
MBA EM CIÊNCIA DE DADOS

GUILHERME TERCEIRO CUNHA

PREDIÇÃO DA VELOCIDADE DO VENTO EM UMA TURBINA EÓLICA A PARTIR DOS DADOS DAS TURBINAS VIZINHAS

Trabalho de conclusão de curso apresentado ao Curso de MBA em Ciência de Dados da Universidade de Fortaleza – UNIFOR como requisito parcial à obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Jorge Luiz Bezerra de Araújo

FORTALEZA

2025

GUILHERME TERCEIRO CUNHA

A Deus.
À minha filha, Liz Terceiro.
À minha esposa, Nágila Terceiro.
Aos meus pais, João Mozart e Gracinda.
Aos meus irmãos, Gustavo e Júlia Maria.
Aos meus tios, Gláucia e João.

RESUMO

Este trabalho desenvolveu um modelo de regressão com o objetivo de prever a velocidade do vento em um aerogerador, utilizando apenas atributos de turbinas vizinhas. A aplicação prática do modelo visa possibilitar a identificação de falhas, a estimativa de geração durante períodos de interrupção de dados e a criação do input de modelos mais completos. Foi seguida uma abordagem estruturada de ciência de dados, compreendendo as etapas de definição do problema, análise exploratória, tratamento de dados ausentes, identificação e tratamento de outliers, normalização/padronização, engenharia e seleção de atributos, além da otimização de hiperparâmetros. Diversos modelos de aprendizado de máquina foram avaliados, como Linear Regression, Random Forest, XGBoost e LightGBM, utilizando o erro médio absoluto (MAE) como métrica de desempenho. Ao longo dos testes, melhorias sucessivas na qualidade das previsões foram obtidas por meio da seleção de atributos, inclusão criteriosa de dados ausentes, criação de variáveis temporais e angulares, e ajuste dos hiperparâmetros utilizando RandomizedSearchCV. O modelo XGBoost apresentou o melhor desempenho, atingindo um MAE de 0,293054, sendo validado também com o Teste de Kolmogorov-Smirnov, que confirmou, com p-valor inferior a 0,05, a superioridade estatística de sua distribuição de erros em relação aos demais modelos. Os resultados evidenciam o potencial de técnicas de aprendizado de máquina na previsão de variáveis meteorológicas e destacam a importância de um processo sistemático de pré-processamento e otimização na construção de modelos mais precisos.

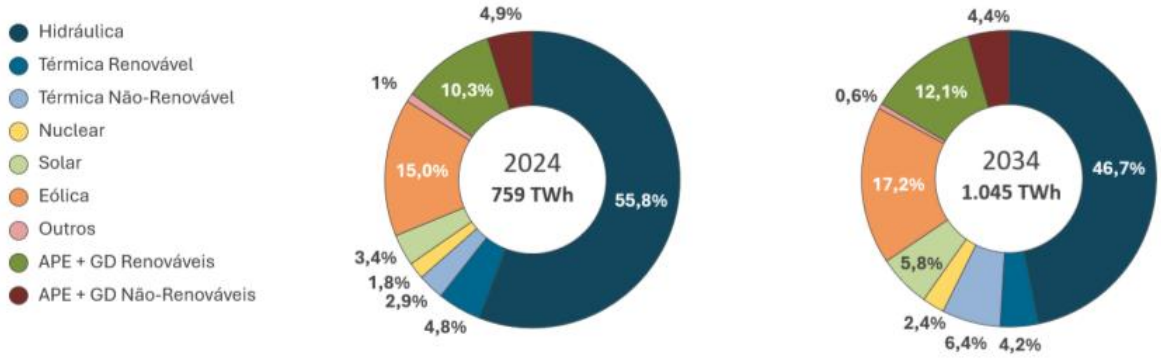
Palavras-chave: Predição de vento, Ciência de Dados, Aprendizado de Máquina, MAE, Aerogeradores.

1. INTRODUÇÃO

O desenvolvimento de um país está associado com o consumo de energia e devido aos acordos ambientais internacionais a busca por novas fontes está caminhando em direção a soluções renováveis. Há uma previsão que a demanda de energia aumente 4% ao ano globalmente até 2027. Esse aumento na demanda surge principalmente da produção industrial, da utilização de ar-condicionado, da eletrificação principalmente no setor de transporte e da expansão de data centers e da Inteligência Artificial (IEA, 2025).

No Brasil, a previsão é que a demanda por energia se expanda 2,1% ao ano até 2034 destacando-se os setores industrial e comercial. O aumento da demanda ocorre em paralelo com a transição energética incentivada principalmente pelos acordos de descarbonização das economias mundiais. O Brasil mantém a predominância de fontes renováveis na sua matriz energética se apoiando em dois recursos com grande potencial em seu território, vento e sol. Na Figura 1, observa-se a perspectiva de crescimento de 2,2% da fonte eólica na sua matriz energética até 2024 (BRASIL, 2025).

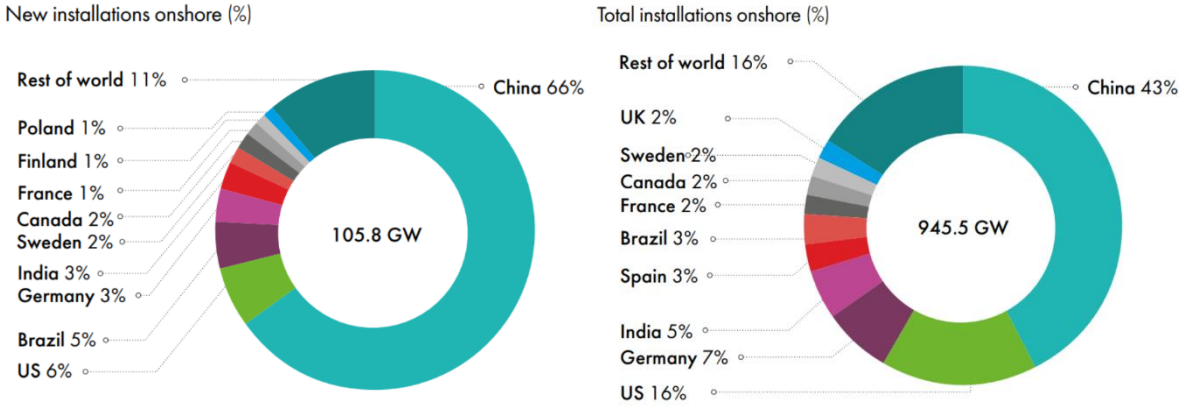
Figura 1: Evolução da geração de energia elétrica no Brasil



Fonte: Brasil, 2025

Observando a classificação mundial de novas instalações de parques eólicos em 2023 (Figura 2), o Brasil, ganhou destaque, ficando em terceiro lugar com mais de 4 GW, atrás apenas da China e dos Estados Unidos. No final do mesmo ano, manteve-se em sexto lugar, com aproximadamente 30 GW instalados, entre as nações com as maiores capacidades de energia eólica instalada. Essas classificações reforçam a força e a capacidade do mercado eólico brasileiro (GWEC, 2024).

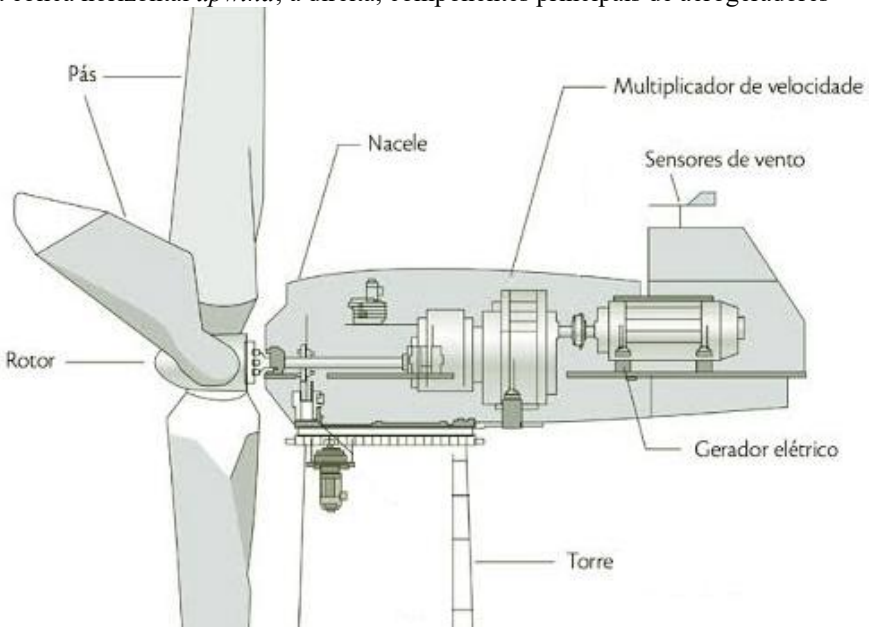
Figura 2: Classificação entre as nações com mais instalações eólicas em 2023 (esquerda) e classificação entre as nações com as maiores capacidade eólicas instalada em 2023.



Fonte: GWEC, 2024

A geração eólica é, fundamentalmente, a conversão da energia cinética do vento em energia elétrica. O equipamento responsável por essa função é a turbina eólica (aerogerador). A Figura 3 (à esquerda) ilustra o modelo horizontal padrão utilizados em parques eólicos (SANTOS, 2022, *apud* PINTO, 2014).

Figura 3: À esquerda, turbina eólica horizontal *upwind*; à direita, componentes principais de aerogeradores



Fonte: WINDBOX, 2020

Existe uma variedade de modelos de turbina no mercado, mas como pode ser visto na Figura 3 (à direita) os componentes principais são:

- Torre: Estrutura em aço ou concreto que sustenta a nacele e o rotor.

- Rotor: Composto pelo cubo e, geralmente, por três pás. As pás possuem perfis aerodinâmicos específicos e otimizados que, com a passagem do vento, promovem a rotação

do rotor e do eixo conectado ao gerador dentro da nacele.

- Nacele: Parte do aerogerador onde a rotação iniciada pelas pás é multiplicada na caixa de transmissão e convertida em energia elétrica no gerador. Abriga diversos equipamentos, como o multiplicador de velocidade (ou caixa de transmissão), gerador elétrico, além de sensores de velocidade e direção do vento (anemômetro e biruta) e outros sistemas de controle.

Normalmente, os aerogeradores operam entre velocidades de vento de 3 m/s a 25 m/s, produzindo uma curva de potência típica, ilustrada na Figura 4. Para proporcionar esse comportamento típico, são necessários controles de *yaw* e *pitch*. O controle de *yaw* está localizado no acoplamento entre torre e nacele, realizando o direcionamento do rotor contra o vento para otimizar a captação de energia pelas pás. Já o controle de *pitch* (ou controle de passo) está localizado no acoplamento entre o rotor e cada pá, controlando o ângulo de ataque possibilitando a concretização da curva de potência típica do aerogerador (SANTOS 2022, *apud* PINTO, 2014).

Segundo Resende (2018), a curva de potência típica (Figura 4) de uma turbina eólica apresenta quatro regiões distintas, conforme a velocidade do vento:

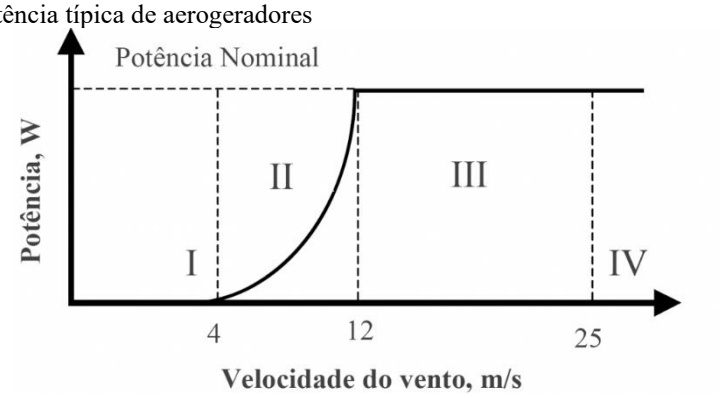
I – Velocidade do vento abaixo do valor mínimo necessário para iniciar a geração de energia.

II – Operação com potência variável, proporcional a velocidade do vento.

III – Operação com potência constante, limitada à sua potência nominal de saída, mesmo havendo variação da velocidade do vento.

IV – Nessa região, a velocidade do vento ultrapassa o valor máximo suportado pelo aerogerador, e o controle de passo (*pitch*) é acionado para evitar danos a estrutura, resultando no desligamento da turbina em seguida.

Figura 4: Curva de potência típica de aerogeradores



Fonte: Resende, 2018

Assim como em outros empreendimento, a otimização da operação e da manutenção da

máquina torna o sistema lucrativo, atraindo cada vez mais investidores para a área. Na energia eólica, essa atenção com a operação se intensifica, pois o investimento inicial é composto por aproximadamente 70% de ativos físicos que serão remunerados ao longo da vida útil da usina. Promover a utilização de sistemas de monitoramento, controle e gestão de ativos, alinhados a boas práticas de manutenção, garante maior eficiência, qualidade e redução de custos (COSTA *et al*., 2016).

O monitoramento é realizando com o sistema de supervisão do parque, SCADA (*Supervisory Control and Data Aquisition*), que centraliza todas as informações geradas por sensores espalhados pela usina eólica e, principalmente, instalados nos aerogeradores. A análise dessas informações torna-se primordial para a tomada de decisão (SANTOS, 2022). Além da otimização da operação e manutenção dos aerogeradores, a análise desses dados pode identificar desvios e comportamentos inesperados das turbinas, possibilitando discussões com o fabricante que podem resultar em garantias ou multas (BARROS, 2019).

O avanço tecnológico tem impulsionado o monitoramento e a otimização da operação e manutenção dos parques, principalmente pela utilização da inteligência artificial. Estudos já avaliaram a aplicação da inteligência artificial na predição de velocidade do vento e da geração de energia, na avaliação de riscos, na detecção e diagnósticos de falhas, bem como na geração de indicadores e clusterização para o aprimoramento do planejamento da manutenção (SOUZA *et al*., 2019).

Neste trabalho, buscou-se desenvolver um modelo de regressão de predição de velocidade de vento, seguindo boas práticas de MLOps aprendidas principalmente na cadeira de Machine Learning e no livro do Géron (2023).

2. OBJETIVO GERAL

Desenvolver um modelo de regressão para predição de velocidade do vento em um determinado aerogerador, avaliando apenas os dados das turbinas vizinhas. O modelo permitirá a verificação de falhas, a comparação entre sensores, a estimativa de geração durante interrupções e a criação do *input* de modelos mais completos de determinação de desalinhamento de *yaw* estático como, conforme comentado por Plumley (2024).

2.1 OBJETIVOS ESPECÍFICOS

- **Definição do Problema:** Contextualizar o problema a ser resolvido, descrever a fonte e o formato dos dados e esboçar o tipo de treinamento e métrica de avaliação.
- **Análise Exploratória dos Dados:** Analisar os dados, buscando informações, como tipos de variáveis, quantidade de valores nulos, características das distribuições, identificação de correlações e de *outliers* com o apoio da visualização por redução de dimensionalidade.
- **Pré-Processamento dos Dados:** Etapa na qual serão removidos os valores nulos, serão escolhidas as colunas mais adequadas, serão realizadas as codificações dos dados categóricos e, por fim, será realizada a padronização dos dados.
- **Treinamento dos Modelos:** Definir o método de treinamento, implementar os modelos e ajustar hiperparâmetros.
- **Avaliação dos Modelos e Conclusão:** Utilizar métrica de avaliação adequada para o problema e definir o melhor modelo a ser utilizado.

3. DESENVOLVIMENTO DO MODELO

No presente estudo, seguiu-se uma sequência estruturada composta pelas seguintes etapas: Análise Exploratória de Dados, Pré-Processamento de Dados e Treinamento dos Modelos finalizando com a Visualização dos Resultados e a Conclusão.

3.1 Definição do Problema

Para a realização desse estudo, foram utilizados dados disponibilizados em uma competição na plataforma *Kaggle* (PLUMLEY, 2024). Os dados são referentes ao parque eólico Kelmarsh, localizado no Reino Unido e pertencente a *Cubico Investments*. O parque possui seis turbinas, mas, no *dataset* fornecido, as informações relativas a turbina 1 foram retiradas, pois, como comentado anteriormente, o objetivo do modelo é predizer a velocidade do vento especificamente nesta turbina, avaliando apenas os dados das outras cinco.

No sistema de supervisão do parque, SCADA, várias informações de cada turbina estão disponíveis, como, por exemplo, velocidade e direção do vento, temperatura, ângulo entre a direção do vento e a direção da nacele, potência gerada, rotação do gerador (rpm) e ângulo *pitch* da pá. Essas informações são armazenadas depois da integração realizada a cada 10 minutos, calculando-se média e desvio padrão (este último apenas de algumas características). Assim, cada dia deve conter 144 linhas no *dataset*.

A velocidade do vento na turbina 1 foi definida como variável alvo (*target*), de modo que os treinamentos dos modelos foram realizados de forma supervisionada e em lote, uma vez que não há um sistema em operação em tempo real. Além disso, o problema em questão caracteriza-se como um caso de regressão, pois, mesmo havendo uma feature *Timestamp*, não é possível utilizar a velocidade do vento na turbina 1 como entrada no modelo.

3.2 Análise Exploratória de Dados

Etapa fundamental em Ciência de Dados, a Análise Exploratória de Dados consiste na busca da compreensão da estrutura, dos padrões e das particularidades dos dados. Esse conhecimento proporciona insights que guiam as etapas seguintes, permitindo a identificação de relações entre atributos (*features*), de erros e inconsistências, além de orientar a escolha dos algoritmos de modelagem.

3.2.1 Visão Geral

O conjunto de dados possui 130.608 registros e 53 colunas, sendo 1 coluna temporal, 1 coluna lógica, 50 colunas numéricas e 1 coluna numérica representando a variável alvo (target), conforme Tabela 1.

Tabela 1: Informações do *dataset*

```
<class 'pandas.core.frame.DataFrame'>
Index: 130608 entries, 0 to 195119
Data columns (total 53 columns):
 #   Column                                     Non-Null Count   Dtype
---  ------                                     --------------   -----
 0   Timestamp                                  130608 non-null  datetime64[ns]
 1   Wind speed (m/s)                           129454 non-null  float64
 2   Wind speed (m/s).1                         100652 non-null  float64
 3   Wind speed (m/s).2                         129444 non-null  float64
 4   Wind speed (m/s).3                         129256 non-null  float64
 5   Wind speed (m/s).4                         129366 non-null  float64
 6   Wind speed, Standard deviation (m/s)       129200 non-null  float64
 7   Wind speed, Standard deviation (m/s).1     100398 non-null  float64
 8   Wind speed, Standard deviation (m/s).2     129190 non-null  float64
 9   Wind speed, Standard deviation (m/s).3     129003 non-null  float64
 10  Wind speed, Standard deviation (m/s).4     129112 non-null  float64
 11  Wind direction (°)                         129454 non-null  float64
 12  Wind direction (°).1                       100839 non-null  float64
 13  Wind direction (°).2                       129444 non-null  float64
 14  Wind direction (°).3                       129440 non-null  float64
 15  Wind direction (°).4                       129366 non-null  float64
 16  Nacelle position (°)                       129454 non-null  float64
 17  Nacelle position (°).1                     100839 non-null  float64
 18  Nacelle position (°).2                     129444 non-null  float64
 19  Nacelle position (°).3                     129440 non-null  float64
 20  Nacelle position (°).4                     129366 non-null  float64
 21  Nacelle position, Standard deviation (°)   129200 non-null  float64
 22  Nacelle position, Standard deviation (°).1 100585 non-null  float64
 23  Nacelle position, Standard deviation (°).2 129190 non-null  float64
 24  Nacelle position, Standard deviation (°).3 129186 non-null  float64
 25  Nacelle position, Standard deviation (°).4 129112 non-null  float64
 26  Vane position 1+2 (°)                      129027 non-null  float64
 27  Vane position 1+2 (°).1                    100412 non-null  float64
 28  Vane position 1+2 (°).2                    129015 non-null  float64
 29  Vane position 1+2 (°).3                    129011 non-null  float64
 30  Vane position 1+2 (°).4                    128935 non-null  float64
 31  Power (kW)                                 129454 non-null  float64
 32  Power (kW).1                               100838 non-null  float64
 33  Power (kW).2                               129444 non-null  float64
 34  Power (kW).3                               129439 non-null  float64
 35  Power (kW).4                               129365 non-null  float64
 36  Nacelle ambient temperature (°C)           129026 non-null  float64
 37  Nacelle ambient temperature (°C).1         100412 non-null  float64
 38  Nacelle ambient temperature (°C).2         129015 non-null  float64
 39  Nacelle ambient temperature (°C).3         129011 non-null  float64
 40  Nacelle ambient temperature (°C).4         128935 non-null  float64
 41  Generator RPM (RPM)                        129454 non-null  float64
 42  Generator RPM (RPM).1                      100839 non-null  float64
 43  Generator RPM (RPM).2                      129444 non-null  float64
 44  Generator RPM (RPM).3                      129440 non-null  float64
 45  Generator RPM (RPM).4                      129366 non-null  float64
 46  Blade angle (pitch position) A (°)         129026 non-null  float64
 47  Blade angle (pitch position) A (°).1       100411 non-null  float64
 48  Blade angle (pitch position) A (°).2       129015 non-null  float64
 49  Blade angle (pitch position) A (°).3       129011 non-null  float64
 50  Blade angle (pitch position) A (°).4       128936 non-null  float64
 51  trainging                                  130608 non-null  bool
 52  target_feature                             129413 non-null  float64
dtypes: category(1), datetime64[ns](1), float64(51)
memory usage: 53.8+ MB
```

As varáveis do *dataset* são:

- Temporal: *Timestamp*. Formatada como *datetime64*.

- Lógica: *training*. Formatada como *bool*.

- Numéricas: *Wind speed (m/s)*, *Wind speed, Standard deviation (m/s)*, *Wind direction (°)*, *Nacelle position (°)*, *Nacelle position*, *Standard deviation (°)*, *Vane position 1+2 (°)*, *Power (kW)*, *Nacelle ambient temperature (°C)*, *Generator RPM (RPM)* e *Blade angle (pitch position) A (°)*. Apenas 10 features foram descritas, mas como são 5 turbinas, as features se repetem para cada turbina totalizando 50 colunas numéricas. Formatadas como *float64*.

- Target: *target_feature*. Representa a velocidade do vento na turbina alvo. Formatada como *float64*.

A Tabela 2 relaciona o nome das turbinas e suas características:

Tabela 2: Identificação dos nomes dos atributos para cada turbina

Nome da turbina	Extensão	Exemplo de feature:
Kelmarsh 1	Não se aplica	*target_feature*
Kelmarsh 2	“”	*Wind speed (m/s)*
Kelmarsh 3	“.1”	*Wind speed (m/s).1*
Kelmarsh 4	“.2”	*Wind speed (m/s).2*
Kelmarsh 5	“.3”	*Wind speed (m/s).3*
Kelmarsh 6	“.4”	*Wind speed (m/s).4*

3.2.2 Identificação de valores ausentes

Na coluna alvo, *target_feature*, foram observados 1.195 linhas com valores ausentes. Na feature *Timestamp*, não foram observados valores ausentes. Já nas características numéricas, verificou-se a ausência pontual de informações em aproximadamente 1,28% (1.673) das linhas, além de uma grande lacuna (*gap*) de dados da turbina Kelmarsh 3, correspondendo a aproximadamente 23,12% (30.000) linhas, conforme pode ser verificado na Tabela 3. Na Figura 5, observa-se que as ausências pontuais estão distribuídas ao longo de todo o *dataset*, enquanto o grande *gap* ocorre em 2019, resultando em uma ausência contínua de dados durante quase todo o ano.

Figura 5: Distribuição dos valores ausentes no tempo de medição

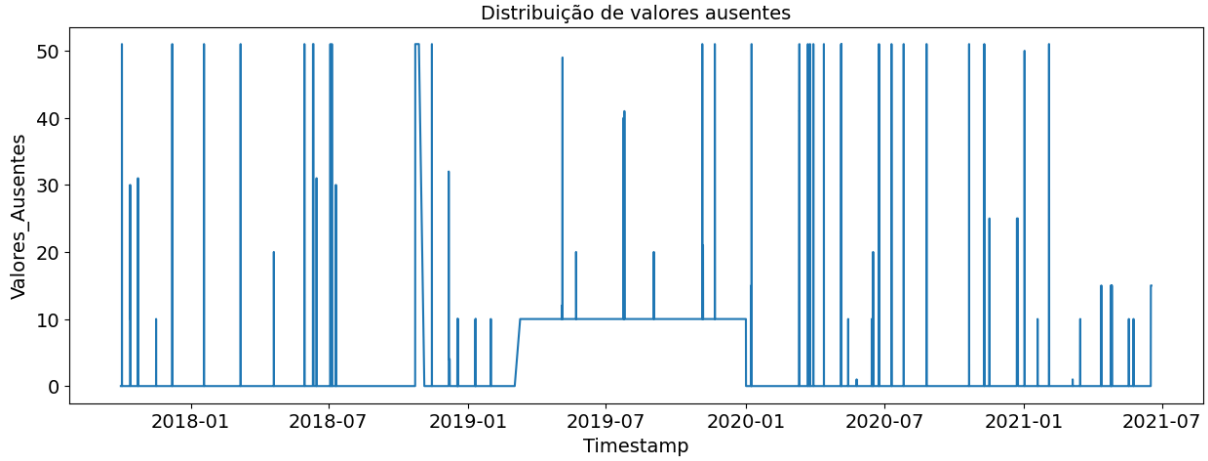


Tabela 3: Contagem e Porcentagem de dados ausentes

	Count	Perc
Wind speed, Standard deviation (m/s).1	30210	23.130283
Blade angle (pitch position) A (°).1	30197	23.120330
Vane position 1+2 (°).1	30196	23.119564
Nacelle ambient temperature (°C).1	30196	23.119564
Nacelle position, Standard deviation (°).1	30023	22.987106
Wind speed (m/s).1	29956	22.935808
Power (kW).1	29770	22.793397
Wind direction (°).1	29769	22.792631
Nacelle position (°).1	29769	22.792631
Generator RPM (RPM).1	29769	22.792631
Vane position 1+2 (°).4	1673	1.280932
Nacelle ambient temperature (°C).4	1673	1.280932
Blade angle (pitch position) A (°).4	1672	1.280167
Wind speed, Standard deviation (m/s).3	1605	1.228868
Vane position 1+2 (°).3	1597	1.222743
Blade angle (pitch position) A (°).3	1597	1.222743
Nacelle ambient temperature (°C).3	1597	1.222743
Nacelle ambient temperature (°C).2	1593	1.219680
Blade angle (pitch position) A (°).2	1593	1.219680
Vane position 1+2 (°).2	1593	1.219680
Nacelle ambient temperature (°C)	1582	1.211258
Blade angle (pitch position) A (°)	1582	1.211258
Vane position 1+2 (°)	1581	1.210492
Nacelle position, Standard deviation (°).4	1496	1.145412
Wind speed, Standard deviation (m/s).4	1496	1.145412
Nacelle position, Standard deviation (°).3	1422	1.088754
Nacelle position, Standard deviation (°).2	1418	1.085692
Wind speed, Standard deviation (m/s).2	1418	1.085692
Nacelle position, Standard deviation (°)	1408	1.078035
Wind speed, Standard deviation (m/s)	1408	1.078035
Wind speed (m/s).3	1352	1.035159
Power (kW).4	1243	0.951703
Nacelle position (°).4	1242	0.950937
Wind speed (m/s).4	1242	0.950937
Generator RPM (RPM).4	1242	0.950937
Wind direction (°).4	1242	0.950937
target_feature	1195	0.914952
Power (kW).3	1169	0.895045
Generator RPM (RPM).3	1168	0.894279
Wind direction (°).3	1168	0.894279
Nacelle position (°).3	1168	0.894279
Wind direction (°).2	1164	0.891216
Generator RPM (RPM).2	1164	0.891216
Power (kW).2	1164	0.891216
Nacelle position (°).2	1164	0.891216
Wind speed (m/s).2	1164	0.891216
Wind direction (°)	1154	0.883560
Generator RPM (RPM)	1154	0.883560
Wind speed (m/s)	1154	0.883560
Nacelle position (°)	1154	0.883560
Power (kW)	1154	0.883560
Training	0	0.000000
Timestamp	0	0.000000

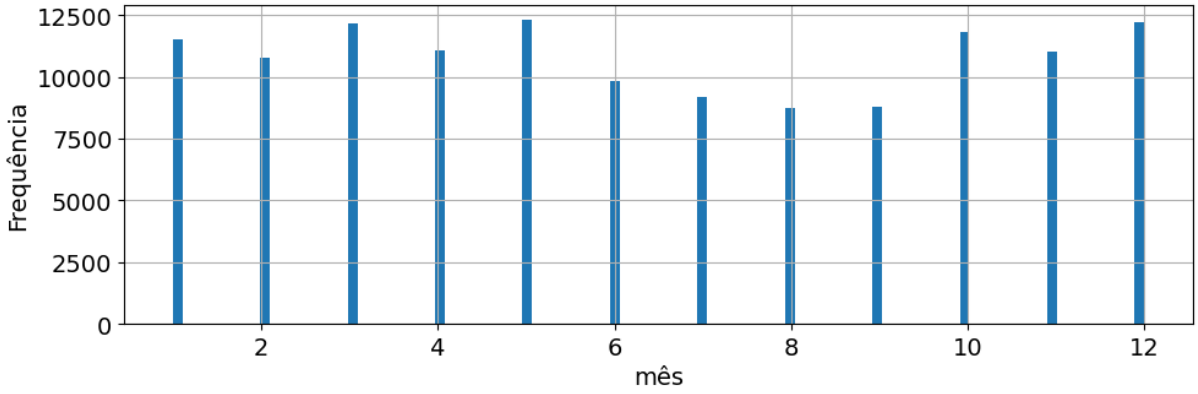
3.2.3 Dados Categóricos e Lógicos

O *dataset* não possui variáveis categóricas. Observando os dados lógicos, só existe um único valor na coluna *training* que é o valor *True.* A coluna foi utilizada pelos criadores da competição no *Kaggle* para separar os dados de treino e teste para o *leaderboard*. Deve-se, portanto, retirar essa coluna na etapa de processamento.

3.2.4 Dados Temporais

A coluna *Timestamp* contém informações sobre ano, mês, dia, hora e minuto. Os registros são realizados a cada 10 minutos, após a integração dos dados por média e desvio padrão (para algumas características). O primeiro registro foi realizado em outubro de 2017 (2017-10-01 00:00:00) e o último registro em junho de 2021 (2021-06-16 23:50:00), totalizando em 1.354 dias de medição. No gráfico de barra na Figura 6, observa-se que os meses 7, 8 e 9 apresentam as frequências mais baixas de registro, possivelmente em função de esses meses não estarem contemplados nem no ano de 2017 nem no de 2021.

Figura 6: Frequência de dados em cada mês do ano do *dataset* inteiro



3.2.5 Dados Numéricos/Quantitativos e Variável *Target*

Nesta seção, foram avaliados o comportamento das distribuições e as medidas de estatística descritiva (média, desvio padrão, quartis, mínimo, máximo, mediana) dos dados numéricos, com o apoio de histogramas e *boxplots*. Devido à quantidade colunas, o foco foi comentar as características e os gráficos considerados mais relevantes para a construção do modelo.

A variável *Wind speed (m/s)* representa a medição da velocidade do vento realizada por anemômetros instalados na nacele de cada aerogerador. Nas cinco turbinas analisadas, o comportamento da velocidade do vento mostrou-se semelhante (Figura 7), não sendo possível identificar outliers apenas com essa avaliação. Em estudos futuros, seria interessante realizar uma análise linha a linha para comparar se as cinco medições estão próximas de um valor comum.

Na Tabela 4, observa-se que médias, medianas, mínimos e máximos diferem entre as cinco turbinas devido a interferência de uma turbina sobre a outra. A turbina Kelmarsh 2 (*Wind speed (m/s)*) registrou a maior velocidade máxima, de 26,40 m/s, possivelmente por ser a primeira a receber o vento e, portanto, não sofre interferência dos outros aerogeradores. Já a turbina Kelmarsh 6 (*Wind speed (m/s).4*) deve ser a última no sentido predominante do vento, apresentando uma velocidade máxima de apenas 23,24 m/s.

Figura 7: Distribuições das variáveis de velocidade do vento nas 5 turbinas

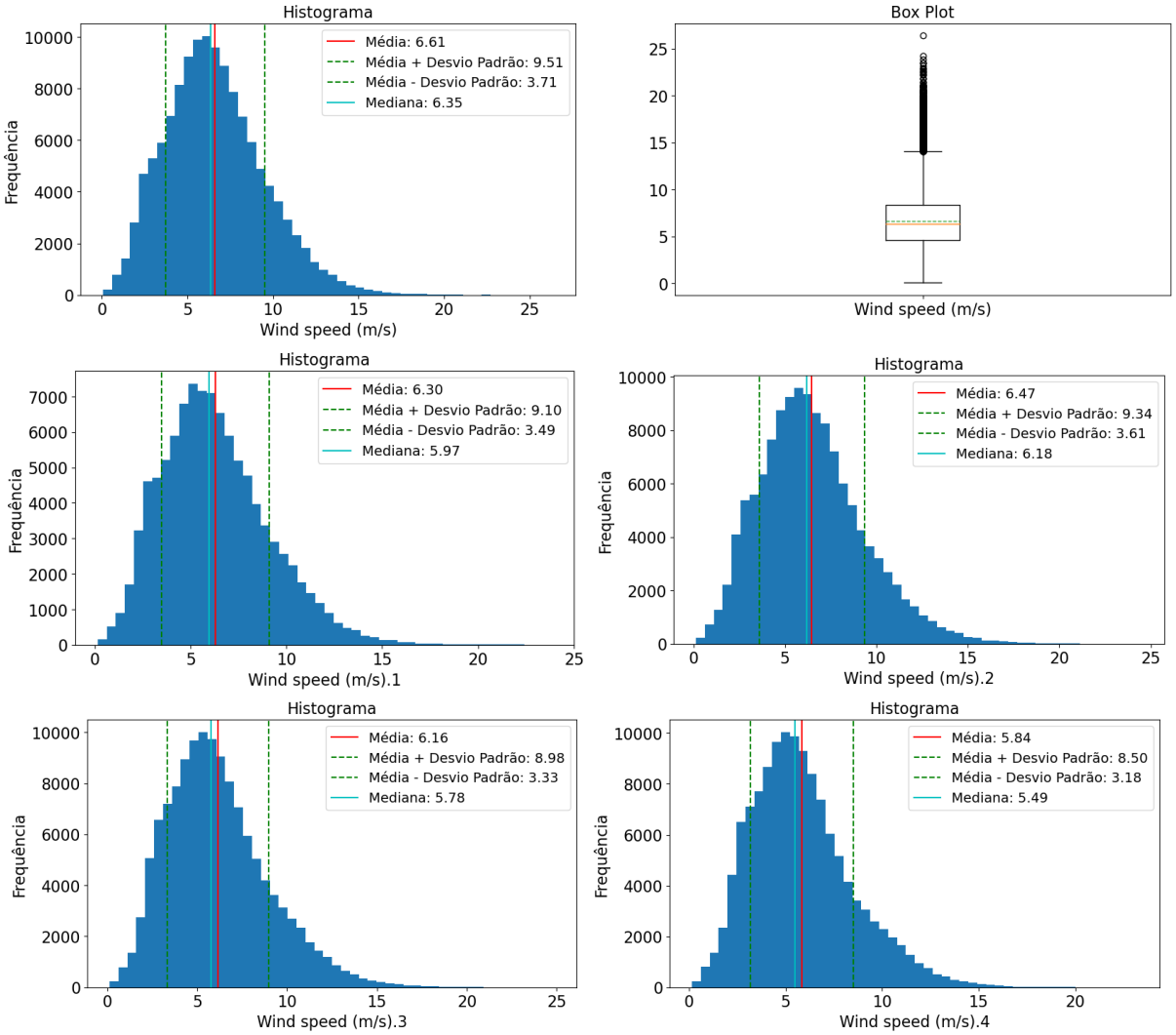


Tabela 4: Medidas estatísticas das variáveis de velocidade do vento nas 5 turbinas

	Wind speed (m/s)	**Wind speed (m/s).1**	**Wind speed (m/s).2**	**Wind speed (m/s).3**	**Wind speed (m/s).4**
count	129382.000000	100616.000000	129376.000000	129193.000000	129302.000000
mean	6.608867	6.296371	6.474510	6.155607	5.839344
std	2.900590	2.802894	2.862391	2.829052	2.658436
min	0.061219	0.153413	0.130069	0.129863	0.141900
25%	4.566530	4.282855	4.455878	4.117782	3.939976
50%	6.346071	5.971344	6.175519	5.782040	5.489077
75%	8.363530	7.972339	8.137524	7.753229	7.303572
max	26.389881	23.826238	24.538780	24.884069	23.241919

A variável alvo, velocidade do vento na turbina Kelmarsh 1, representa a mesma informação (velocidade do vento) coletada nas demais turbinas, confirma-se isso avaliando a distribuição e medidas estatísticas da variável *target* na Figura 8 e na Tabela 5. Com base nos valores, infere-se que essa turbina também sofre interferência de aerogeradores adjacentes. Além disso, a variável alvo apresenta média e mediana próximas (6,33 m/s e 6,04 m/s respectivamente), sugerindo que os dados podem ser aproximados por uma distribuição normal, assim como observado para as demais turbinas.

Figura 8: Distribuição e resumo da variável *target*

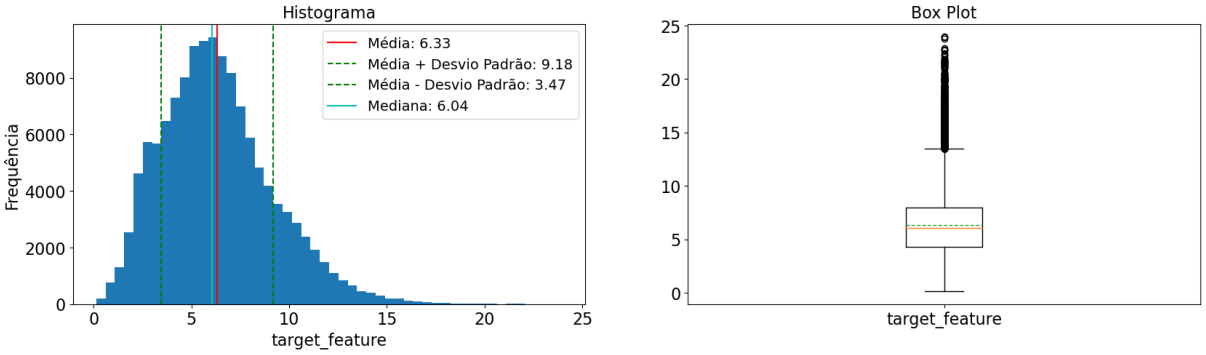


Tabela 5: Medidas estatísticas da variável *target*

	target_feature
count	129413.000000
mean	6.326829
std	2.856556
min	0.134232
25%	4.288056
50%	6.042259
75%	7.970540
max	23.964783

Wind direction (°) representa a direção do vento, *Nacelle position (°)* refere-se à orientação da nacele, e *Vane position 1+2 (°)* corresponde à diferença entre essas duas medidas. Essa diferença é registrada por uma biruta afixada no aerogerador, alinhada com o nacele, enquanto a posição da nacele é conhecida a partir do sistema de controle de rotação em relação ao norte verdadeiro. A direção do vento é, portanto, determinada pela soma da direção da nacele e da diferença registrada pela biruta.

Analisando as Figuras 9 e 10, referentes à turbina Kelmarsh 2, observa-se que a distribuição da direção da nacele é semelhante à distribuição da direção do vento. Esse comportamento é esperado, uma vez que a otimização da geração de energia depende do alinhamento da nacele com a direção do vento. As distribuições de direção são bem características desse tipo de dado, pois observa-se valore de 0 a 360º e uma predominância entre uma faixa específica, neste caso entre 200º e 219º.

Figura 9: Distribuição da direção do vento na turbina Kelmarsh 2 (*Wind direction (°)*)

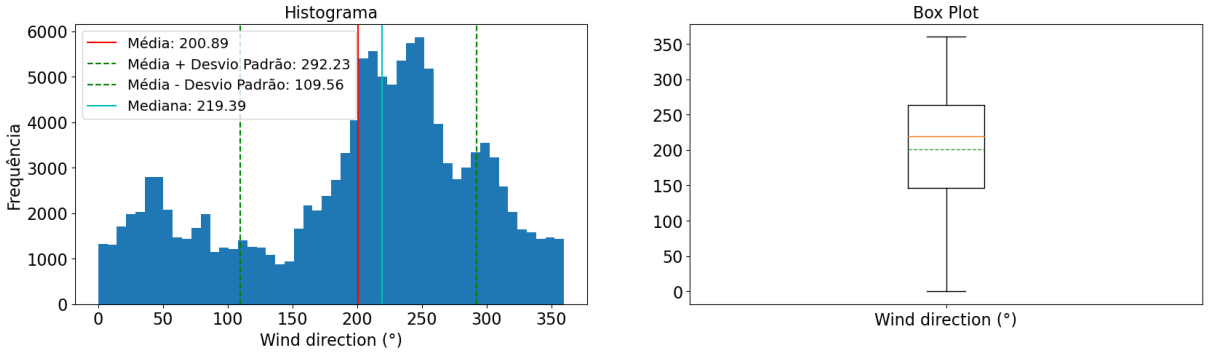
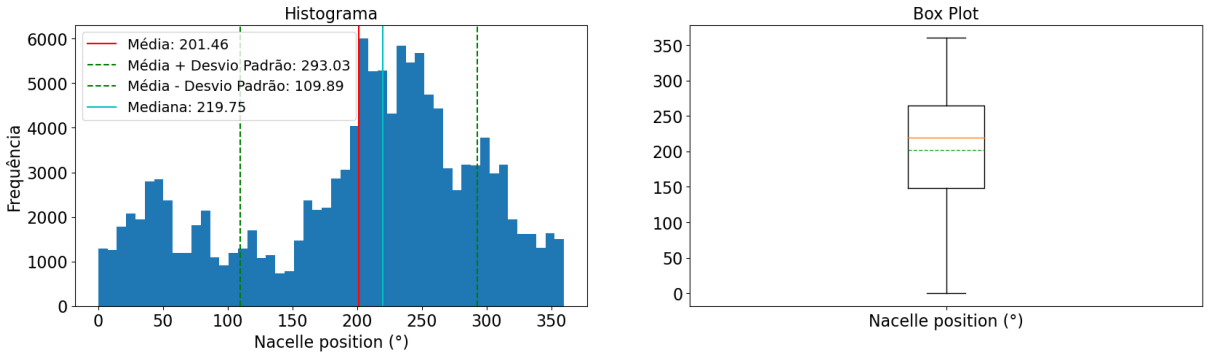


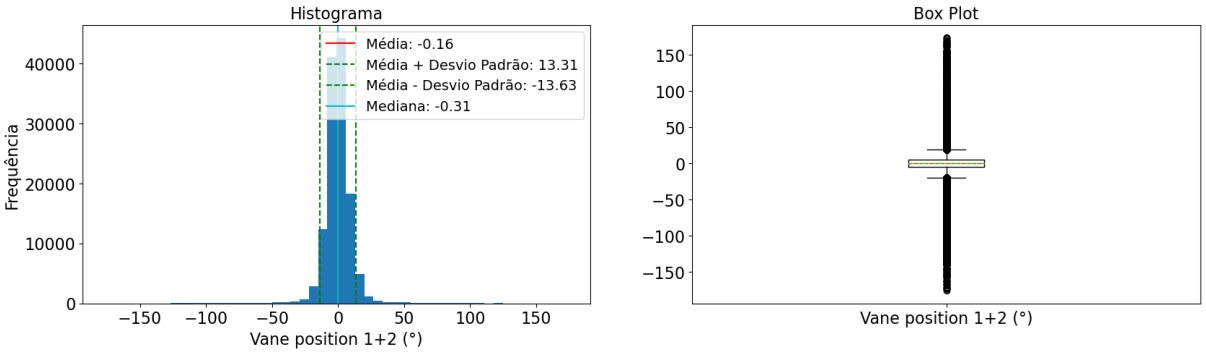
Figura 10: Distribuição da direção da nacele na turbina Kelmarsh 2 (*Nacelle position (°)*)



Em complemento, a distribuição da diferença (*Vane position 1+2 (°)*), apresentada na Figura 11, está centrada em torno de 0°, confirmando que o sistema de orientação da nacele atua de forma eficaz para manter o alinhamento ideal. Apesar de a figura apresentar uma base

relativamente curta, observa-se que o sistema de controle da turbina poderia ser otimizado para reduzir o desvio padrão atual de 13°, o que contribuiria para uma geração de energia ainda mais eficiente.

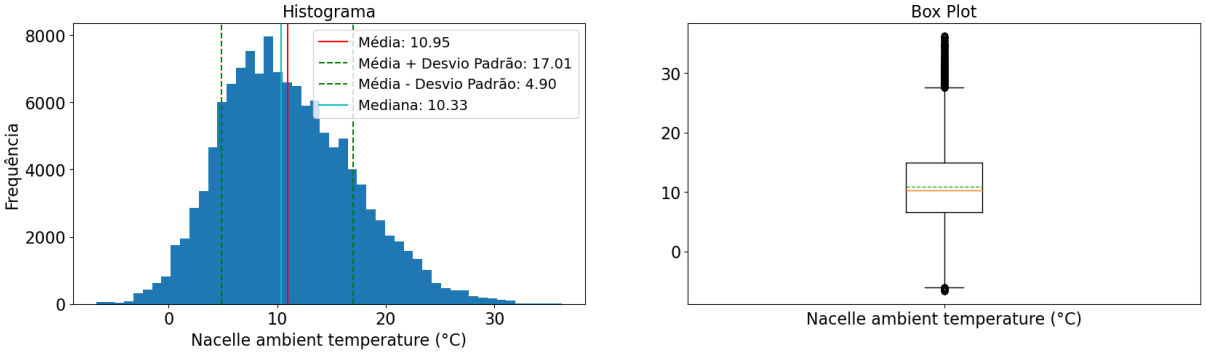
Figura 11: Distribuição da diferença entre direção do vento e direção da nacele na turbina Kelmarsh 2 (*Vane position 1+2 (°)*)



Wind speed, Standard deviation (m/s) representa o desvio padrão integrado a cada 10 minutos da velocidade do vento na nacele, e *Nacelle position, Standard deviation (°)* representa o desvio padrão integrado a cada 10 minutos da direção da nacele. O estudo da distribuição dessas colunas não foi aprofundando neste trabalho.

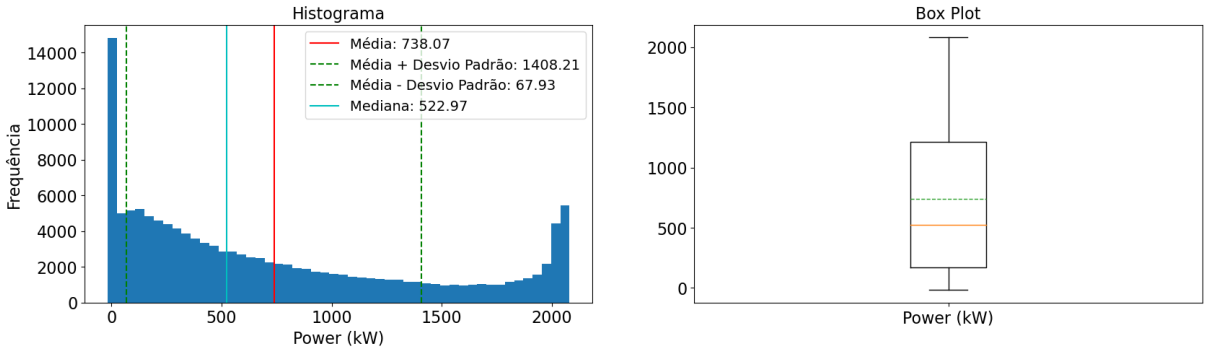
Nacelle ambient temperature (°C) é a temperatura ambiente externa a nacele, medida por um sensor. Como pode ser visto na Figura 12 que ilustra essa característica para a turbina Kelmarsh 2, a distribuição pode ser aproximada por uma distribuição normal. As medidas de estatística descritiva são: média=10,95ºC, mediana=10,33ºC, desvio padrão=6,05 ºC, mínimo=-6,68 ºC e máximo=36,26 ºC.

Figura 12: Distribuição da temperatura ambiente na nacele da turbina Kelmarsh 2 (*Nacelle ambient temperature (°C)*)



Power (kW) representa a potência gerada no aerogerador. Com a distribuição ilustrada na Figura 13, pode-se notar os extremos bem definidos, próximos de 0 kW e 2080 kW. O limite superior, 2080kW, representa a potência máxima que o aerogerador consegue gerar.

Figura 13: Distribuições da potência gerada na turbina Kelmarsh 2 (*Power (kW)*)



Como abordado na introdução deste trabalho, a potência gerada depende da velocidade do vento recebida no rotor, até um limite de segurança, quando o sistema de controle altera o ângulo de ataque das pás para interromper a geração e prevenir danos à turbina.

O aerogerador, além de possuir uma velocidade de corte superior, também possui uma velocidade mínima de entrada em operação, geralmente em torno de 3 m/s. O pico à esquerda observado na distribuição da Figura 13 corresponde aos casos em que a velocidade do vento está abaixo desse limite mínimo de operação. No entanto, também podem existir situações em que a velocidade do vento é superior ao limite mínimo, mas o aerogerador não está gerando energia, o que pode ocorrer por decisão humana — para realização de reparos, por exemplo — ou por outros motivos operacionais.

Tabela 6: Medidas estatísticas das variáveis de potência gerada nas cinco turbinas

	Power (kW)	**Power (kW).1**	**Power (kW).2**	**Power (kW).3**	**Power (kW).4**
count	129382.000000	100803.000000	129376.000000	129377.000000	129302.000000
mean	738.067369	648.686780	678.562457	626.409834	538.522675
std	670.141151	647.051602	650.613433	636.718984	596.767972
min	-17.083929	-16.573292	-18.454697	-17.280342	-16.873207
25%	169.010899	121.708477	142.168906	117.585521	77.207870
50%	522.968280	410.855545	450.805361	383.070744	308.484328
75%	1210.498455	1030.276802	1079.066146	974.694284	791.426328
max	2080.385461	2079.870593	2083.431885	2081.809985	2086.926514

Em alguns casos, quando a máquina está parada, há necessidade de consumo de energia elétrica para manter o funcionamento dos sistemas de controle. Nesses momentos, o valor da potência registrada torna-se negativo, indicando o consumo de energia pela turbina. Na Tabela 6, esses casos podem ser constatados pelos valores mínimos registrados nas cinco turbinas, variando entre -16 kW e -19 kW.

Para a geração de energia, o gerador precisa operar com rotações acima de aproximadamente 1000 rpm. A caixa de transmissão realiza essa conversão, elevando a rotação do rotor — por exemplo, de 20 rpm — para a rotação necessária no gerador (cerca de 1000 rpm) (FOTSO et al., 2021). A variável *Generator RPM (RPM)* representa a rotação do eixo do gerador em rotações por minuto. Assim como a potência gerada, a rotação do gerador depende diretamente da velocidade do vento.

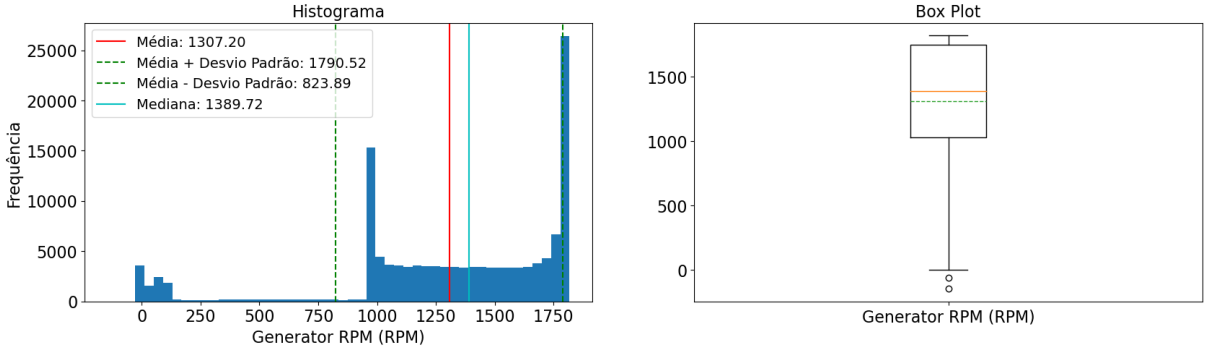
Na distribuição ilustrada na Figura 14, observam-se três picos distintos:

- Valores próximos a 0 rpm (ou até negativos) representam condições de vento abaixo da velocidade mínima necessária ou manobras operacionais humanas.

- Valores próximos a 1000 rpm correspondem à operação normal, onde a caixa de transmissão eleva a rotação do rotor para o nível adequado do gerador.

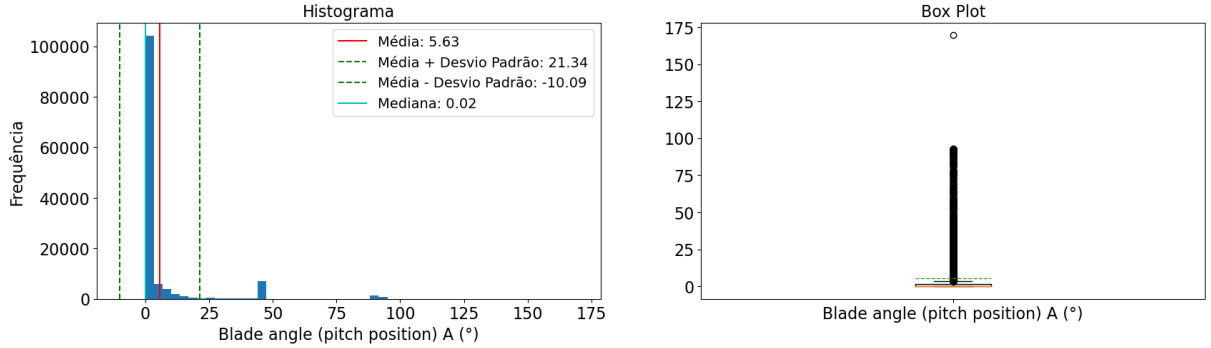
- Valores próximos a 1800 rpm indicam o limite superior de operação, seguido pela atuação dos sistemas de segurança da turbina.

Figura 14: Distribuição da rotação do gerador em rpm da turbina Kelmarsh 2 (*Generator RPM (RPM)*)



A variável *Blade angle (pitch position) A (°)* representa o ângulo de ataque das pás do rotor. Como ilustrado na Figura 15, valores próximos de 0° indicam a configuração para máxima geração. Valores elevados refletem a atuação dos sistemas de controle, que ajustam o ângulo para reduzir a captação de vento em situações de velocidade acima do permitido ou em manobras de operação.

Figura 15: Distribuição do ângulo de ataque da pá da turbina Kelmarsh 2 (*Blade angle (pitch position) A (°)*)

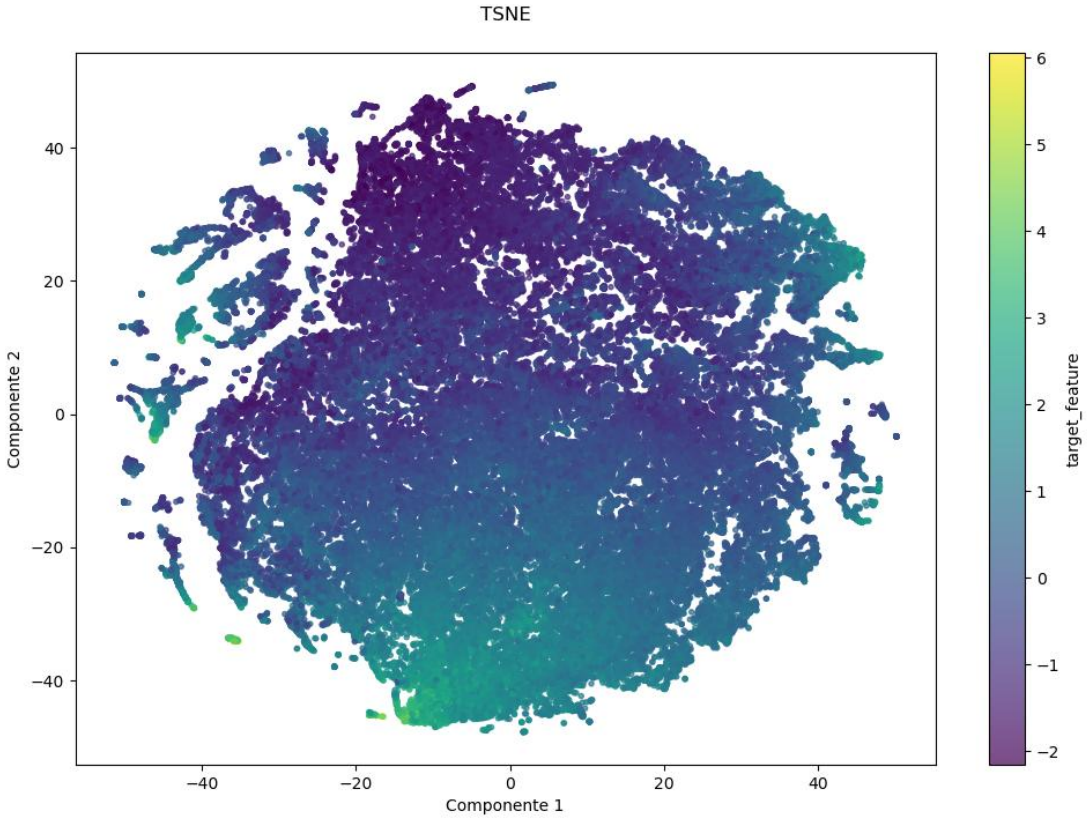


Após a avaliação das variáveis numéricas, é importante ressaltar que *Nacelle position (°)*, *Vane position 1+2 (°), Power (kW)*, *Generator RPM (RPM)* e *Blade angle (pitch position) A (°)* são varáveis passíveis de alteração por ação human. Portanto, a relação dessas variáveis com as variáveis naturais (*Wind speed (m/s), Wind direction (°)* e *Nacelle ambient temperature (°C)*) pode ser impactada, dificultando o aprendizado dos modelos. Na seção 3.2.7 será avaliado essa relação em busca de outlier causado pela ação humana.

3.2.6 Visualização por Redução de Dimensionalidade

Para uma melhor compreensão da estrutura dos dados, foi aplicada a técnica de redução de dimensionalidade t-SNE (*t-Distributed Stochastic Neighbor Embedding*). O t-SNE é uma metodologia não linear que projeta dados de alta dimensão em um espaço bidimensional, preservando principalmente as relações locais entre as amostras. Essa técnica converte as distâncias entre pontos em probabilidades e busca minimizar a diferença entre essas probabilidades nos espaços de alta e baixa dimensão. O t-SNE permite, por meio da visualização, identificar diferentes aglomerados de instâncias, em que elementos de um mesmo grupo apresentam características semelhantes, enquanto elementos de grupos distintos possuem características diferentes. Isso possibilita a identificação visual de padrões distintos e de outliers (GÉRON, 2023).

Figura 16: Gráfico de dispersão de t-SNE



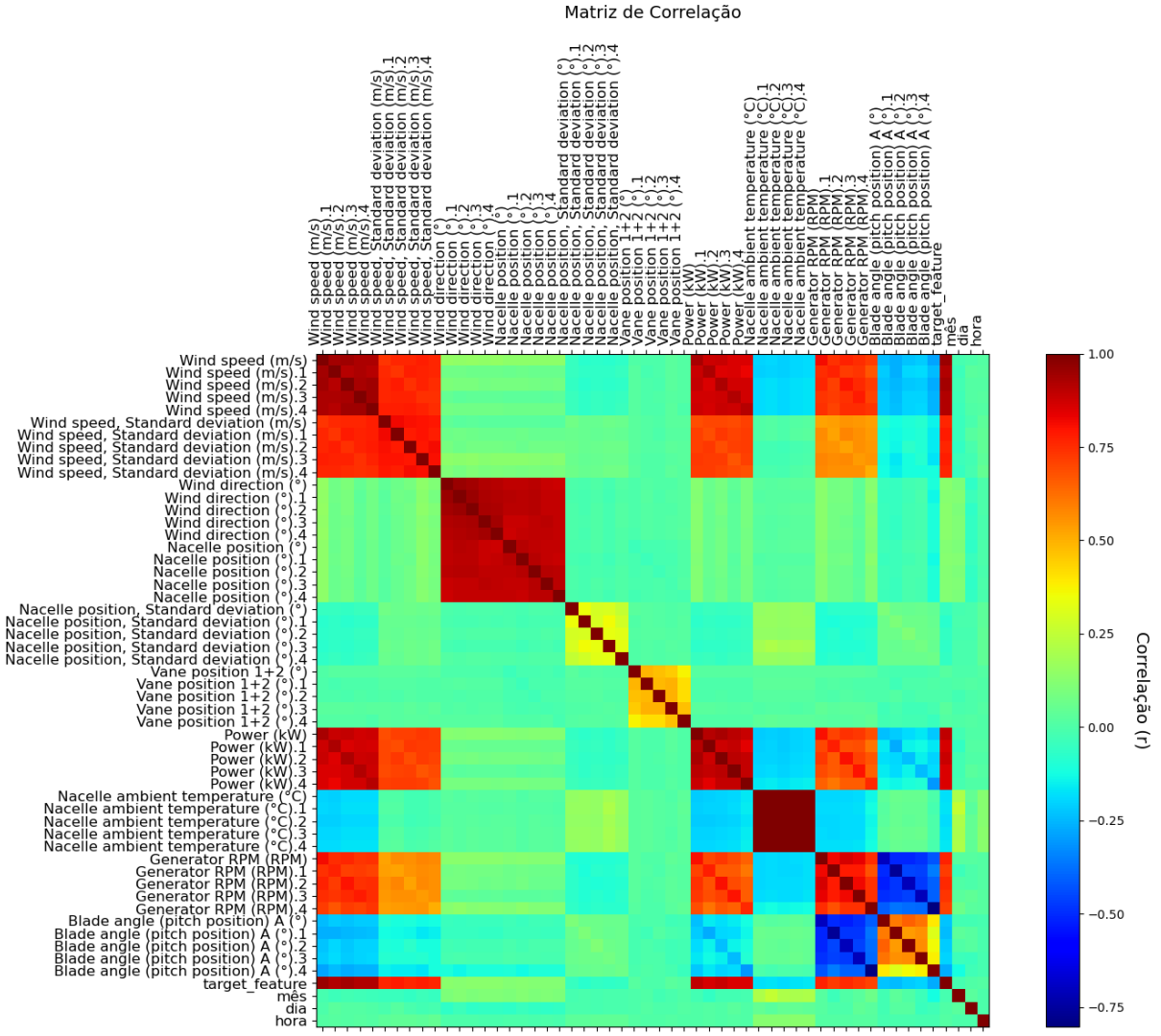
No gráfico de dispersão da Figura 16, o t-SNE permitiu visualizar a distribuição dos dados em relação à variável alvo, revelando agrupamentos e variações locais importantes para a análise exploratória e para o entendimento prévio da complexidade do problema. Observa-se uma grande massa que, mesmo com diferentes valores na variável alvo, permanece conectadas por uma característica semelhante. Ao redor da grande massa, observam-se indícios de *outliers* em pequenos grupos, sem predominância de valores específicos da variável alvo. Além dessas observações, é possível notar a predominância de valores mais baixos da variável alvo, o que corrobora a análise feita no histograma da Figura 8, que apresentou média de 6,32 m/s e valor máximo de 23,96 m/s.

3.2.7 Análise de correlação, relação entre características e verificação de *outliers*

Seguindo com a análise exploratória, foi realizado o cálculo da correlação entre as variáveis, obtendo-se a matriz de correlação conforme Figura 17. Os maiores índices de correlação com a variável alvo, *target_feature*, foram observados nas variáveis *Wind speed (m/s), Wind speed, Standard deviation (m/s), Power (kW)* e *Generator RPM (RPM)*. A variável

Blade angle (pitch position) A (°) apresentou forte correlação inversa com *Generator RPM (RPM)* e correlações inversas fracas com *target_feature*, *Wind speed (m/s)* e *Power (kW)*.

Figura 17: Matriz de correlação

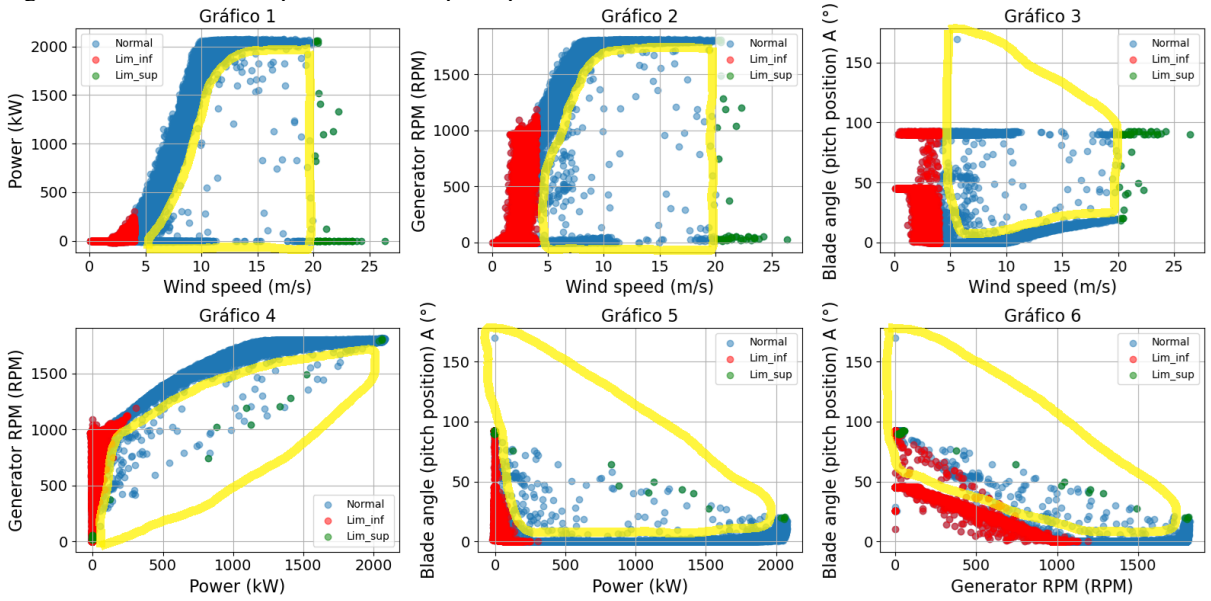


A variável alvo é altamente correlacionada com as velocidades (e seus desvios padrão) por representarem o mesmo fenômeno físico. Para compreender melhor os índices de correlação entre as demais variáveis, foram analisados os gráficos de dispersão ilustrados na Figura 18. Nos gráficos, pontos vermelhos representam registros com velocidade do vento inferior a 4 m/s, enquanto pontos verdes indicam velocidades superiores a 20 m/s.

Observando a relação entre a velocidade do vento e a potência na Figura 18, confirma-se a curva de potência típica de aerogeradores, conforme ilustrado anteriormente na Figura 4. Além disso, é possível visualizar o comportamento esperado das variáveis dentro da faixa de operação normal dos aerogeradores (velocidades do vento entre 4 e 20 m/s). A partir dessa análise, infere-se que valores de potência, rotação e ângulo das pás situados fora da relação

observada dentro dessa faixa, especialmente na região demarcada em amarela do gráfico, indicam outliers possivelmente causados por fatores externos, como manutenções programadas ou serviços operacionais no parque.

Figura 18: Gráficos de dispersão entre as principais variáveis



3.3 Pré-Processamento dos Dados

Nesta etapa, os dados foram tratados considerando as observações realizadas durante a análise exploratória, com o objetivo de tornar o *dataset* mais limpo e consistente para o treinamento dos modelos de aprendizado de máquina. A sequência de fases seguidas foi: tratamento de variáveis, detecção e tratamento de outliers e dados ausentes, normalização e padronização e, por fim, a engenharia e seleção de atributos.

Na seção 3.2.2, foram observados 1.195 linhas com valores ausentes na variável alvo. Por esse motivo, essas linhas foram removidas, visando manter o treinamento supervisionado do modelo. Além disso, a única coluna lógica do *dataset* apresentou o mesmo valor (*True*) em todas as linhas, conforme verificado na seção 3.2.3, e, portanto, foi removida.

3.3.1 Tratamento de Outliers

O tratamento de outliers foi realizado como a primeira etapa, com o objetivo de filtrar o *dataset* e implementar as correções ou ajustes necessários antes de proceder com o tratamento

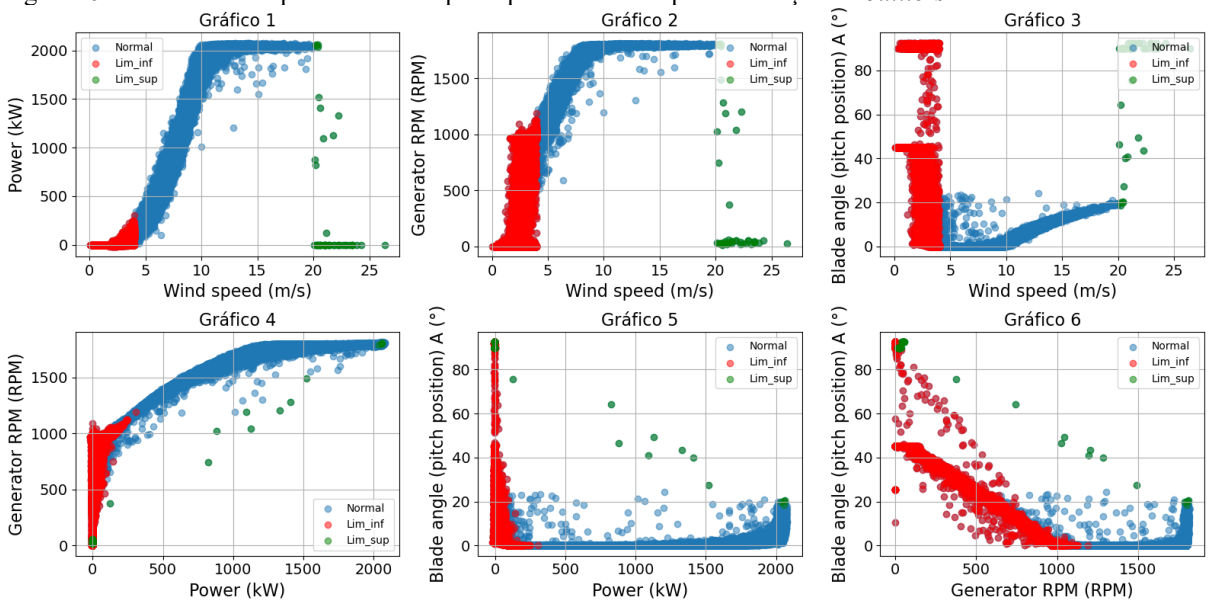
dos dados ausentes. Inicialmente, o *dataset* foi dividido em dois *dataframes*: um contendo apenas linhas sem dados ausentes e outro contendo apenas com linhas com dados ausentes.

No *dataframe* sem dados ausentes, duas abordagens foram aplicadas:

- Alteração dos valores de *Vane position 1+2 (°), Power (kW), Generator RPM (RPM)* e *Blade angle (pitch position) A (°)* situados dentro da região amarela na Figura 18, substituindo-os pela mediana de cada intervale de velocidade de vento, considerando o intervalo de 0,5 m/s. O resultado dessa abordagem pode ser observado na Figura 19. Apesar de vários filtros aplicados, ainda permaneceram alguns pontos na região de outliers.

- Criação de cinco novas colunas, uma para cada turbina, com o objetivo de indicar se, em determinada linha, existe algum outliers (região amarela da Figura 18) nas colunas *Vane position 1+2 (°), Power (kW), Generator RPM (RPM)* e *Blade angle (pitch position) A (°)* de cada turbina.

Figura 19: Gráficos de dispersão entre as principais variáveis após a correção de *outliers*



3.3.2 Tratamento de Dados Ausentes

Na seção 3.2.2, foram identificados dois tipos de ausência de dados no *dataset*: ausências pontuais distribuídas em várias colunas e uma ausência continua ao longo de 2019 nas colunas relacionadas à turbina Kelmarsh 3.

Devido a essa diferença, foram testadas abordagens diferentes para o tratamento de cada tipo de ausência:

- Para a ausência continua nas 10 colunas referentes à turbina Kelmarsh 3, foram testadas duas alternativas: o preenchimento dos dados utilizando a média das respectivas colunas e a imputação por meio de um modelo de regressão linear baseado nas colunas de outras turbinas;

- Para as ausências pontuais nas demais colunas, optou-se pelo preenchimento com a média da respectiva coluna.

3.3.3 Normalização e Padronização

Normalização e padronização são técnicas que permitem ajustar a escala dos atributos garantindo que, no aprendizado de máquina, as características estejam em escalas semelhantes e sejam comparadas de forma justa. Neste trabalho, foi verificado a aplicação de:

- *MinMaxScaler*, que ajusta os dados para o intervalo entre 0 e 1;

- *StandardScaler*, que ajusta os dados para que tenham média zero e desvio padrão igual a 1.

- *RobustScaler*, que realiza o ajuste utilizando a mediana e o intervalo interquartil.

3.3.4 Engenharia e Seleção de Atributos

A engenharia de atributos é a etapa em que se criam variáveis (atributos) a partir de variáveis já existentes e a seleção de atributos é a etapa em que escolhe as variáveis que serão utilizadas no modelo. Elas possuem o objetivo de ganhar performance e de evitar complexidade.

A partir da variável *Timestamp*, foram gerados três novos atributos: mês, dia e hora (inseridas em uma lista nomeada de *list_time*). Como são atributos cíclicos, foram criadas mais 6 colunas para representar, mês, dia e hora utilizando funções seno e cosseno (inseridas em uma lista nomeada de *list_time_mais*).

As equações abaixo foram utilizadas para a criação dessas novas variáveis:

$$newfeature\_\sin = \sin(2 * pi * \frac{feature}{\text{max\_val\_feature}})$$

$$newfeature\_\cos = \cos(2 * pi * \frac{feature}{\text{max\_val\_feature}})$$

Onde *max_val_feature* representa o valor máximo da feature. Para o mês, foi utilizado o valor 12; para a hora, 24; e para 0 dia, 31.

De cada variável que representam um ângulo (*Wind direction (°), Nacelle position (°), Nacelle position, Standard deviation (°), Vane position 1+2 (°)* e *Blade angle (pitch position) A (°)*) foram geradas duas novas variáveis: uma representando o seno e outra representando o cosseno da variável original (inseridas em uma lista nomeada de *list_angle_mais*). As equações abaixo foram utilizadas para essa criação:

$$newfeature\_\sin = \sin(2 * pi * \frac{feature}{360^o})$$

$$newfeature\_\cos = \cos(2 * pi * \frac{feature}{360^o})$$

Inicialmente, o *dataset* possuía 53 atributos e, após as exclusões e criações, passou a contar com 109 atributos. Para a seleção das melhores features, foi utilizado, a princípio, a técnica *SelectKBest* com critério *f_regression*, considerando quantidades de variáveis entre 68 e 94. Além disso, foram criadas duas listas de variáveis:

- Selecionou-se as variáveis com correlação maior que 0,1 (em valor absoluto), inseridas em uma lista nomeada de list_corr_mais_1.

- Selecionou-se as variáveis com correlação maior que 0,2 (em valor absoluto), inseridas em uma lista nomeada de list_corr_mais_2.

3.4 Treinamento dos Modelos

Para a construção dos modelos preditivos, os dados foram divididos em dois conjuntos: 80% para treino e 20% para teste, garantindo que a avaliação final fosse realizada sobre dados não vistos durante o processo de aprendizagem.

Foram utilizados quatro algoritmos de regressão comumente aplicados em problemas de previsão de variáveis contínuas:

- ***Linear Regression***: modelo estatístico simples e interpretável, utilizado como base comparativa.

- ***Random Forest Regressor***: modelo baseado em múltiplas árvores de decisão, robusto a overfitting e eficiente para capturar relações não lineares.

- ***XGBoost Regressor:*** algoritmo de boosting com alta performance, que otimiza o erro residual de modelos anteriores.

- ***LightGBM Regressor***: variante mais leve e rápida do XGBoost, ideal para grandes volumes de dados e com excelente desempenho computacional.

Para avaliar o desempenho dos modelos, foi utilizada a métrica MAE (Mean Absolute Error). Essa métrica calcula a média dos erros absolutos entre os valores previstos e os reais, sendo uma medida direta e interpretável de quão distante, em média, estão as previsões em relação aos valores observados.

A fórmula do MAE é dada por:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|yi - \hat{y}i|)$$

Onde yi é o valor real, $\hat{y}i$ é o valor predito e $\hat{y}i$ é o número total de observações.

O MAE foi escolhido por ser uma métrica intuitiva, robusta a outliers extremos, e por apresentar os erros na mesma unidade da variável alvo, facilitando a interpretação dos resultados no contexto do problema. Além disso, MAE foi escolhido pelo autor da competição do *Kaggle* para avaliar os participantes.

Tabela 7: Configuração e resultado da otimização de hiperparâmetros

# Configurar o RandomizedSearchCV 	
random_search = RandomizedSearchCV(
estimator=model,
param_distributions=param_dist,
n_iter=10,
scoring='neg_mean_squared_error',
cv=3,
verbose=2,
random_state=42,
n_jobs=-1
)
	
	# Range de busca de cada modelo
	# Avaliar 10 combinações aleatórias (100 para XGB e LGBM)
	# Métrica de avaliação
	# Validação cruzada com 3 folds
	
	# Paralelismo

# Melhores parâmetros para Random Forest 	
best_rf = {'n_estimators': 150,
'min_samples_split': 5,
'min_samples_leaf': 2,
'max_features': 'sqrt',
'max_depth': 20,
'bootstrap': False}
	
	# Número de árvores na floresta
	# Número mínimo de amostras necessárias para dividir um nó
	# Número mínimo de amostras que uma folha deve ter
	# Número de atributos considerados na divisão de cada nó
	# Profundidade máxima das árvores
	# Define se as amostras serão selecionadas com reposição

# Melhores parâmetros para XGBoost 	
best_xgb = {'subsample': np.float64(0.6),
'reg_lambda': np.float64(0.0001),
'reg_alpha': np.float64(0.1),
'n_estimators': np.int64(450),
'min_child_weight': np.int64(6),
'max_depth': np.int64(10),
'learning_rate': np.float64(0.042222222222222223),
'gamma': np.float64(0.0),
'colsample_bytree': np.float64(0.7)}
	
	# Fracção de amostras usadas para treinar cada árvore
	# Regularização L2 (ridge)
	# Regularização L1 (sparsidade)
	# Número de árvores (boosting rounds)
	# Peso mínimo para dividir um nó folha
	# Profundidade máxima das árvores
	# Taxa de aprendizado (shrinkage)
	# Redução mínima de perda para uma divisão
	# Fracção de recursos usados para dividir cada nó

# Melhores parâmetros para LightGBM 	
best_lgbm = {'subsample': np.float64(0.7),
'reg_lambda': np.float64(0.001),
'reg_alpha': np.float64(0.1),
'num_leaves': 63,
'n_estimators': np.int64(350),
'max_depth': np.int64(8),
'learning_rate': np.float64(0.10666666666666666),
'colsample_bytree': np.float64(0.5)}
	
	# Proporção de amostras usadas por árvore (evita overfitting)
	# Regularização L2 (controla complexidade)
	# Regularização L1 (sparse features)
	# Número de folhas por árvore (mais folhas = mais complexo)
	# Número de árvores (boosting rounds)
	# Profundidade máxima da árvore (controla complexidade)
	# Taxa de aprendizado (quanto cada árvore contribui)
	# Proporção de colunas usadas por árvore

Após a definição do melhor fluxo de pré-processamento dos dados, foi realizada a otimização de hiperparâmetros dos modelos mais promissores, Random Forest, XGBoost e LightGBM. Esse ajuste fino teve como objetivo maximizar o desempenho preditivo, utilizando técnicas como busca aleatória (*RandomizedSearchCV*) e validação cruzada (*Cross-Validation*). A Tabele 7 apresenta a configuração utilizada no *RandomizedSearchCV* e os melhores parâmetros de cada modelo.

3.5 Visualização dos Resultados

A Tabela 8 apresenta o resultado dos testes realizados. MAE_LR, MAE_RF, MAE_XGB e MAE_LGBM são os erros absolutos médios do modelos *Linear Regression, Random Florest, XGBoost* e *LigthGBM*, respectivamente. KS23, KS24 e KS34 são resultados do Teste de Kolmogorov-Smirnov (KS) entre *Random Florest* (2)*, XGBoost* (3) e *LigthGBM* (4).

Os testes 0, 1, 2 e 3 foram conduzidos no *dataframe*, considerando apenas linhas sem dados ausentes.

- No teste 0, foram avaliadas todas as variáveis numéricas;
- No teste 1, apenas os atributos naturais;
- No teste 2, apenas os atributos artificiais;
- No teste 3, apenas os atributos que apresentaram correlação com a variável target acima de 0,1 (em valor absoluto).

O teste 3 foi o que obteve os menores valores de MAE em todos os modelos avaliados. Devido à proximidade entre os resultados do *Random Florest, XGBoost* e *LigthGBM*, foi decido seguir com os testes apenas com esses dois últimos, além do *Linear Regression*, pois o Random Florest demanda muito tempo para execução. O Random Flores foi utilizado novamente apenas após a definição do melhor processamento e da otimização de hiperparâmetros, nos Testes 30 e 31.

O teste 4 avaliou a correção de outliers, incluindo todas as variáveis numéricas na criação do modelo. O teste 5 avaliou a adição de colunas identificando linhas com outliers para cada turbina, também utilizando todas as variáveis numéricas.

Os testes 4 e 5 aumentaram o erro em comparação com o teste 0 (sem tratamento de outliers). Observa-se também que o tratamento dos outliers apresentou um erro maior que somente a identificação dos outliers em colunas.

Tabela 8: Resultados dos testes

Teste	MAE_LR	MAE_RF	MAE_XGB	MAE_LGBM	KS23	KS24	KS34	Teste
0	0,542013	0,324791	0,325404	0,330957	NÃO	Diferente	NÃO	0
1	0,54348	0,331568	0,333376	0,338022	NÃO	Diferente	NÃO	1
2	0,737846	0,351272	0,356799	0,375844	Diferente	Diferente	Diferente	2
3	0,542751	0,322649	0,324585	0,330722	NÃO	Diferente	NÃO	3
4	0,542618		0,327331	0,333196			NÃO	4
5	0,542029		0,325561	0,331071			NÃO	5
6	0,535106		0,32382	0,333123			Diferente	6
7	0,535106		0,32382	0,333123			Diferente	7
8	0,535106		0,325497	0,332836			Diferente	8
9	0,535106		0,3257	0,332962			Diferente	9
10	0,535106		0,324106	0,333204			Diferente	10
11	0,532797		0,322606	0,331548			Diferente	11
12	0,530051		0,321524	0,328017			NÃO	12
13	0,530432		0,324504	0,328204			NÃO	13
14	0,532236		0,325044	0,328608			NÃO	14
15	0,532671		0,326644	0,330802			NÃO	15
16	0,533538		0,324827	0,32987			Diferente	16
17	0,533558		0,324942	0,329504			Diferente	17
18	0,53244		0,325992	0,328773			NÃO	18
19	0,532283		0,323807	0,328616			NÃO	19
20	0,532236		0,325044	0,328608			NÃO	20
21	0,531887		0,324922	0,328633			NÃO	21
22	0,531848		0,32526	0,328729			NÃO	22
23	0,53178		0,324346	0,328619			NÃO	23
24	0,531776		0,324679	0,328756			NÃO	24
25	0,531732		0,324028	0,328751			NÃO	25
26	0,531743		0,32375	0,328608			NÃO	26
27	0,531726		0,32445	0,328387			NÃO	27
28	0,53172		0,325002	0,328569			NÃO	28
29	0,531603		0,325112	0,32819			NÃO	29
30		0,306478	0,291915	0,30362	Diferente	NÃO	Diferente	30
31		0,308224	0293054	0,305006	Diferente	NÃO	Diferente	31

O teste 6 avaliou o preenchimento de um grande gap de dados por meio de um modelo linear e o preenchimento de dados ausentes pontuais com a média da respectiva coluna, considerando apenas os atributos bem correlacionados para a criação do modelo (assim como no teste 3).

O teste 7 avaliou o preenchimento tanto do grande gap de dados quanto dos dados ausentes pontuais utilizando apenas a média da coluna, também considerando dados bem correlacionados.

Observa-se que os resultados dos testes 6 e 7 foram iguais, e houve uma diminuição do erro com a inclusão e tratamento das linhas com dados ausentes. Por exemplo, no teste 3, o XGBoost apresentou um erro de 0,324585, enquanto no teste 6 o erro caiu para 0,32382. Devido a essa melhoria, o *dataframe* utilizado no teste 6 foi copiado para os testes seguintes.

O teste 8 avaliou a normalização com MinMaxScaler, o teste 9 avaliou a padronização com RobustScaler e o teste 10 avaliou a padronização com StandardScaler. Os três testes apresentaram erros maiores que o teste 6. Com base nesses resultados, optou-se por seguir os testes sem realizar normalização ou padronização.

Os testes 11, 12 e 13 referem-se à engenharia de atributos:

- O teste 11 avaliou a criação das variáveis cíclicas temporais (list_time_mais);

- O teste 12 avaliou tanto as variáveis temporais (list_time_mais) quanto as variáveis cíclicas angulares (list_angle_mais).

- O teste 13 avaliou a retirada das variáveis originais que representam ângulos.

Teste 11 e 12 reduziram o erro em comparação ao teste 6, com destaque para o teste 12. Por conta dessa performance, o *dataframe* utilizado no teste 12 foi copiado para os testes seguintes.

Os testes 14 e 15 são referentes à seleção de atributos realizada manualmente:

- O teste 14 avaliou atributos com correlação acima de 0,1 com a variável target;

- O teste 15 avaliou atributos com correlação acima de 0,2 com a variável target.

Os testes 16 a 29 abordaram a seleção de atributos utilizando a técnica SelectKBest, variando a quantidade de atributos selecionados (68, 70, 72, ..., até 94).

Nenhum dos testes de seleção de atributos conseguiu reduzir o erro em relação ao teste 12.

O teste 30 avaliou a otimização dos hiperparâmetros dos três modelos (*Random Forest, XGBoost e LightGBM*), obtendo os melhores resultados em todos eles. O destaque foi o *XGBoost*, que alcançou o menor erro entre todos os testes, com um valor de 0,291915.

O teste 31 avaliou os modelos realizando o teste com Cross-Validation, e observa-se que os resultados continuam entre os melhores, com *XGBoost* atigindo um MAE de 0,293054.

Além disso, o p-valor de 3,36e-07 menor que 0,05 no Teste de Kolmogorov-Smirnov (KS) indica que a distribuição dos erros do *XGBoost* é estatisticamente diferente das dos outros modelos, consolidando sua escolha como o modelo mais adequado para a tarefa deste estudo.

Para validação desse trabalho, é possível verificar códigos e dados no link https://github.com/guilhermeterceiro/kaggle-predict-the-wind-speed-at-a-wind-turbine.git.

4. CONCLUSÃO

Este trabalho apresentou o desenvolvimento de um modelo de regressão com foco na predição da velocidade do vento em um aerogerador, utilizando exclusivamente atributos de turbinas vizinhas e visando possibilitar a verificação de falhas e a estimativa de geração durante interrupções. Além da aplicação prática, buscou-se também apresentar uma abordagem estruturada de ciência de dados, abrangendo desde a exploração inicial dos dados até a avaliação comparativa de modelos e técnicas de otimização.

Durante o processo, foram realizadas diversas etapas fundamentais: definição do problema, análise exploratória, tratamento de dados ausentes, identificação e tratamento de outliers, normalização/padronização, engenharia e seleção de atributos e otimização de hiperparâmetros. A performance dos modelos foi avaliada utilizando a métrica MAE (*Mean Absolute Error),* que permitiu uma interpretação direta dos erros médios de predição.

Ao longo dos testes realizados, observou-se uma evolução consistente na performance dos modelos a partir das melhorias aplicadas ao longo do processo. Inicialmente, a seleção de atributos com base na correlação com a variável alvo (teste 3) já proporcionou uma boa redução do erro com menor complexidade. Em seguida, a inclusão criteriosa de dados ausentes (teste 8) resultou em melhorias adicionais, ampliando a base de dados de forma eficaz. Posteriormente, a aplicação de técnicas de engenharia de atributos, com a criação de variáveis temporais e angulares (teste 16), levou a uma queda ainda mais significativa no erro. Por fim, a otimização dos hiperparâmetros por meio do RandomizedSearchCV (teste 33) permitiu alcançar os menores valores de MAE entre todos os experimentos, evidenciando o impacto positivo do ajuste fino dos modelos.

O modelo com melhor desempenho foi o *XGBoost*, atingindo um MAE de 0,293054, representando a menor média de erro absoluto entre todos os testes realizados. Este resultado demonstra o potencial de métodos de aprendizado de máquina na previsão de variáveis meteorológicas, especialmente quando combinados com um fluxo de pré-processamento e engenharia de atributos bem planejado. Além disso, o Teste de Kolmogorov-Smirnov (KS) indicou, com p-valor inferior a 0,05, que a distribuição dos erros do *XGBoost* é estatisticamente diferente e superior à dos outros modelos avaliados, reforçando a sua escolha como o modelo mais adequado para a tarefa deste estudo.

REFERÊNCIAS

BARROS, Danilo M. C. Windbox: Eficiência em gestão operacional de parques eólicos. Dissertação - Universidade Federal do Rio Grande do Norte, Natal, 2019.

BRASIL, Ministério das Minas e Energia. Plano decenal de Expansão de energia 2034. Publicado em 09 de abril de 2025.

FOTSO, H. R. F.; KAZÉ, C. V. A.; KENMOÉ, G. D. Real-time rolling bearing power loss in wind turbine gearbox modeling and prediction based on calculations and artificial neural network**.** Tribology International, v. 163, 2021.

GÉRON, Aurélien. *Hands-On Maching Learning with Scikit-Learn, Keras & TensorFlow*. 3. Ed. EUA: O'Reilly Media, 2023.

GWEC, Global Wind Energy Council. Global Wind Report 2024. Publicado em 16 de abril de 2024.

IEA, International Energy Agency. Eletricity 2025: Analysis and forecast to 2027. Publicado em fevereiro de 2025.

PLUMLEY, Charlie. Predict the wind speed at a wind turbine. Kaggle. 2024 Disponível em: <https://kaggle.com/competitions/predict-the-wind-speed-at-a-wind-turbine>. Acesso em 25 de abril de 2025.

RESENDE, Carlos. Fontes de Energia Alternativa: Energia Eólica. 2018. Disponível em: https://www.sharenergy.com.br/fontes-de-energia-alternativa-energia-eolica/. Acesso em 15 de abril de 2025.

WINDBOX. Componentes dos aerogeradores. 2020. Disponível em: <https://windbox.com.br/blog/componentes-dos-aerogeradores/>. Acesso em 15 de abril de 2025.