UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
ESCOLA DE ENGENHARIA
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E TRANSPORTES

TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO

PREVISÃO DE VELOCIDADE DO VENTO PARA GERAÇÃO DE ENERGIA EÓLICA A PARTIR DE UM MODELO PREDITIVO CAUSAL

ANTONIO BERNARDO SILVA DA CUNHA

Orientador: Michel José Anzanello, PhD

PORTO ALEGRE
ABRIL/2023

PREVISÃO DE VELOCIDADE DO VENTO PARA GERAÇÃO DE ENERGIA EÓLICA A PARTIR DE UM MODELO REGRESSIVO CAUSAL

Antonio Bernardo Silva da Cunha – antoniobscunha@gmail.com

Michel José Anzanello, PhD – anzanello@produção.ufrgs.br

Universidade Federal do Rio Grande do Sul (UFRGS), Escola de Engenharia, Departamento de Engenharia de Produção.
Av. Osvaldo Aranha, 99, 5° andar, 90055-190, Porto Alegre, RS, Brasil.

1. Introdução

As crescentes preocupações relacionadas à evolução da situação climática e a independência energética em relação a combustíveis fósseis têm apoiado o desenvolvimento tecnológico e regulatório direcionado à expansão da matriz energética mundial (LETCHER, 2008). Dentre as fontes renováveis candidatas a participar desta expansão, a energia eólica apresenta implantação rápida e consistente capacidade de geração de energia elétrica (PINSON, 2013). Segundo o *Global Wind Report* (2022), ao fim de 2021 a capacidade cumulativa instalada de energia eólica no mundo era de 837 GW. Essa fonte alternativa de energia continua crescendo de forma acelerada, com expectativa de atingir mais de 3000 GW de capacidade instalada em 2030, sendo o Brasil um dos principais expoentes desta expansão.

A energia eólica no Brasil teve sua implementação impulsionada pelo Programa de Incentivo às Fontes Alternativas de Energia (PROINFA) em 2002, sendo que em 2009 teve início a sua expansão dentro da matriz energética brasileira através de leilões para contratação de energia no ACR (Ambiente de Contratação Regulado) (MIGUEL, 2021). De 2009 a 2021, a potência instalada de geração eólica passou de 1 GW para 21 GW, consistindo em 11% da matriz energética brasileira. Estima-se que o mercado brasileiro possui capacidade de agregar uma taxa de 5 GW por ano (GWEC, 2022).

A entrada de fontes de geração de energia elétrica baseadas em recursos renováveis estocásticos e intermitentes na rede elétrica apresenta desafios na gestão da oferta de energia estável e segura. Devido à estocasticidade intrínseca aos recursos eólicos, denota-se uma variabilidade natural da velocidade do vento que se manifesta em diferentes escalas temporais. No curto prazo, flutuações na velocidade do vento ocorrem aleatoriamente em questão de segundos, minutos, horas e/ou dias, que pode alterar o equilíbrio instantâneo entre oferta e demanda por eletricidade (MIGUEL, 2021).

Com vistas a mitigar a influência de tais variações climáticas, buscam-se gerar políticas públicas que norteiem a expansão do sistema elétrico brasileiro (EPE, 2020). No âmbito do agente gerador, os efeitos da intermitência se manifestam na gestão dos contratos de venda de energia para o cumprimento da entrega de energia no tempo contratado, sob pena de incorrer na compra de energia a preços de mercado. Além de serem potencialmente mais caros que a energia negociada nos contratos em que é parte, tal descumprimento também pode incorrer em multas e outras sanções aplicadas pelas autoridades reguladoras do setor (PINHEIRO, 2020).

Além das ações em políticas públicas, as previsões de geração de energia são ferramentas valiosas para integrar a energia eólica à rede de fornecimento de eletricidade. Em parques eólicos, tais previsões são usadas principalmente para operação da rede de distribuição e transmissão, programação e comercialização de usinas de energia (FOCKEN et al., 2008). Por outro lado, a previsão de geração de energia pode aumentar a eficiência do agente gerador em captar valor no mercado a partir de sua geração, seja prevendo geração excedente ou detectando um déficit de geração. Nesse sentido, diversos autores propuseram abordagens para predição da geração de energia. Lai *et al. (2020)* revisaram modelos de *machine learning* para previsão de geração de energia a partir de fontes renováveis, concluindo que há um uso majoritário de técnicas de inteligência artificial e de modelos híbridos na previsão de geração de energia eólica. Heinermann *et al.* (2016) afirmam que a energia eólica somente pode ser integrada à rede elétrica juntamente com um modelo de previsão que forneça uma previsão confiável e que seja suficientemente eficiente a fim de fornecê-las em tempo hábil, e propõem um modelo *ensemble* que combina diferentes técnicas de *machine learning.* Por fim, Sharifzadeh *et al.* (2019) implementaram um modelo baseado em técnicas de *machine learning* com o intuito de promover a

integração da previsão de geração a partir de fontes renováveis à previsão de demanda de energia elétrica a fim de amenizar os efeitos da intermitência da geração e as incertezas e flutuações da demanda por energia.

O presente trabalho tem como objetivo avaliar a utilização de técnicas regressivas para predição da velocidade do vento no ponto de captação (rotor do aerogerador), velocidade essa que implica diretamente na geração de energia eólica. Para predição da velocidade, serão coletados dados meteorológicos relacionados (velocidade do vento e sua direção em solo, temperatura, umidade e precipitação, dentre outras). A variável dependente será a velocidade do vento na altura do rotor do aerogerador, que influencia diretamente a geração de energia eólica.

Após a coleta dos dados, três técnicas de predição serão ajustadas e testadas em termos de suas capacidades preditivas no que diz respeito à velocidade do vento na altura do rotor: Regressão Linear Múltipla, Regressão *Partial Least Squares* e Redes Neurais. A análise do impacto das variáveis de entrada do modelo sobre a previsão da velocidade do vento na altura do rotor também será executada, permitindo uma melhor compreensão de como as variáveis meteorológicas influenciam a geração de energia eólica indiretamente. Além de permitir a identificação das variáveis mais relevantes para predição do vento, será possível identificar a técnica regressiva mais adequada para prever a velocidade do vento na altura do rotor do aerogerador, contribuindo para uma estimativa mais precisa da geração de energia eólica.

Este artigo, além da introdução, estrutura-se da seguinte forma: a segunda seção corresponde a uma revisão teórica acerca dos tópicos abordados, a terceira seção apresentará a metodologia utilizada, por fim, a quarta seção discutirá os resultados e as conclusões obtidas com o presente trabalho.

2. Referencial Teórico

2.1. Técnicas Multivariadas de Predição

De acordo com Lattin et al. (2011), técnicas multivariadas de predição compreendem o conjunto de procedimentos implementados para investigar dois ou mais grupos de variáveis, simultaneamente, tendo em conta suas associações e relações com o intuito de explicar os resultados observados. Os diferentes métodos

podem ser aplicados em tipos e conjuntos específicos de dados, visando elucidar diferentes análises (BARTHOLOMEW, 2010).

Para além dos modelos preditivos, técnicas de análises multivariadas podem ser empregadas em diferentes áreas e objetivos: redução de dados e simplificação estrutural, ordenamento e agrupamento de variáveis, investigação de dependência entre variáveis e validações por testes de hipóteses (JOHNSON et al, 2007). Nos tópicos seguintes são apresentados os fundamentos das técnicas multivariadas utilizadas neste estudo.

2.1.1. Regressão Linear Múltipla (RLM)

Segundo Draper et al. (1998), a regressão é uma técnica estatística amplamente utilizada para encontrar relações lineares entre observações e variáveis. A Regressão Linear Simples (SLR) estabelece uma relação linear entre uma única variável independente e uma variável dependente. Quando há várias variáveis independentes, a Regressão Linear Múltipla pode ser aplicada para modelar essas relações (MONTGOMERY et al., 2012).

A Regressão Linear Múltipla estende a Regressão Linear Simples para considerar várias variáveis independentes, permitindo a análise de situações mais complexas e realistas (JAMES et al., 2013). O modelo matemático geral da Regressão Linear Múltipla é dado pela equação (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \quad (1)$$

onde Y é a variável dependente, $X_1, X_2, \ldots X_p$ são as variáveis independentes, β_0 é o intercepto, $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes de regressão e ε é o termo de erro (WOOLDRIDGE, 2012).

Os coeficientes de regressão são estimados usando método dos mínimos quadrados ordinários (OLS), que busca minimizar a soma dos quadrados das diferenças entre os valores observados e os valores previstos pelo modelo (HAYASHI, 2000). A solução dos mínimos quadrados ordinários pode ser encontrada usando álgebra matricial, conforme mostrado na equação (2):

$$\beta = (X'X)^{-1}X'Y \quad (2)$$

onde β é o vetor de coeficientes de regressão, X é a matriz de variáveis independentes, Y é o vetor de variáveis dependentes e $(X'X)^{-1}X'$ é a matriz inversa generalizada de Moore-Penrose, que é utilizada para calcular os coeficientes de regressão.

A Regressão Linear Múltipla é uma técnica poderosa e simples para modelar a relação entre uma variável dependente e várias variáveis independentes, permitindo uma análise mais detalhada de fenômenos complexos. No entanto, é importante estar ciente das suposições subjacentes à Regressão Linear Múltipla, como a normalidade dos erros, a independência dos erros e a ausência de multicolinearidade entre as variáveis independentes. Quando essas suposições não são atendidas, outras técnicas de regressão podem ser mais adequadas.

2.1.2. Regressão PLS (*Partial Least Squares Regression*)

Segundo Malinowski (2002), a regressão é uma das formas mais comuns de encontrar relações lineares entre observações e variáveis. O método da Regressão Linear Múltipla (RLM) pode ser implementado quando há muitas variáveis. Entretanto, seu desempenho pode ser reduzido no caso em que o número de variáveis independentes for maior que o número de observações (gerando risco de *overfitting*) (TOBIAS, 1995), ou ainda quando existir forte colinearidade entre as variáveis (GELADI, 1986). Para estes casos, propõe-se a utilização da Regressão PLS (*Partial Least Squares Regression*).

A proposta deste método consiste em considerar no modelo apenas as variáveis independentes que possuem mais peso sobre as variações das variáveis de resposta, nomeando os fatores (várias independentes) de variáveis latentes. Isto é realizado por meio da geração de componentes que irão compor um conjunto chamado de vetores latentes. Este conjunto será composto tanto de variáveis independentes como de variáveis de reposta, e tem por finalidade explicar a covariância entre esses dois grupos distintos (ABDI, 2010).

Morelatto (2010) descreveu o método PLS, baseado no trabalho de Wold *et al.* (2001), através da equação (3),

$$Y = XB + E \tag{3}$$

em que, $Y = (y_1, \dots, y_M)$ é uma matriz $(N \times M)$ de variáveis de resposta, $X = (x_1, \dots, x_N)$ é uma matriz $(N \times K)$ de variáveis preditoras, B é uma matriz $(K \times M)$ dos coeficientes de regressão, e E é a matriz de ruídos para o modelo que tem a mesma dimensão de Y.

A regressão PLS extrai um pequeno número de "novas" variáveis, que são chamadas de fatores ou componentes e denotadas por t_a $(a = 1, \dots, A)$. Os fatores são preditores de Y e descrevem X, desta forma tanto X como Y são, pelo menos em parte, modelados pelas mesmas variáveis latentes. A ideia da PLS é extrair componentes que consigam capturar as variâncias das covariáveis e obter correlações com as variáveis independentes. Isto pode ser atingindo a partir da maximização da covariância entre os fatores de X, t_a e Y.

O número de componentes extraídos de X é menor que o número de covariáveis $(A < K)$ e são ortogonais. Estes são obtidos como combinações lineares das variáveis originais x_k, com os coeficientes, chamados de pesos, w_a $(a = 1, \dots, A)$, dados por

$$T = XW \tag{4}$$

em que $T = (t_1, \dots, t_A)$ é a matriz $(N \times A)$ de fatores e $W = (w_1, \dots, w_A)$ é a matriz $(K \times A)$ de pesos. As matrizes X e Y, por sua vez, são decompostas da seguinte forma:

$$X = TP' + F \tag{5}$$

e

$$Y = UC' + G \tag{6}$$

sendo que T e U são matrizes $(N \times A)$ de fatores de X e Y, respectivamente; P' e C' são matrizes $(A \times K)$ de cargas de X e Y, respectivamente; e F e G são matrizes de erros.

Como citado acima, na decomposição de X as componentes, t_a, são obtidas de maneira que as covariâncias entre elas e as variáveis de resposta da matriz Y sejam maximizadas.

Com a dimensão de X reduzida em A componentes t_a $(A < K)$ pode-se efetuar a regressão de Y sobre T na forma

$$Y = TC' + E. \quad (7)$$

Para conseguir os coeficientes da regressão PLS referentes aos dados originais, basta substituir a igualdade em (2), na equação (5), e obter

$$Y = TC' + E = XWC' + E = XB + R \quad (8)$$

assim, os coeficientes da regressão PLS podem ser escritos como segue na equação (7).

$$B = WC' \quad (9)$$

2.1.3. Redes Neurais Artificiais

As redes neurais artificiais (RNA) são um conjunto de técnicas de aprendizado de máquina inspiradas na estrutura e funcionamento do cérebro humano. Segundo Wang (2003), uma RNA é composta por uma camada de entrada de neurônios, camadas intermediárias ou camadas ocultas de neurônios e uma camada final de neurônios de saída. Cada conexão entre neurônios está associada a um peso, que é ajustado durante o treinamento da rede para otimizar seu desempenho na tarefa em questão.

As camadas intermediárias, são responsáveis por realizar transformações não-lineares nos dados de entrada, permitindo que a rede aprenda a mapear entradas complexas em saídas desejadas. A saída h_I do neurônio i na camada intermediária é calculada a partir da equação:

$$h_i = max\left(0, \sum_{j=1}^{N} V_{ij}x_j + T_i^{hid}\right) \quad (10)$$

onde $\max()$ é a função ReLU, N é o número de neurônios de entrada, V_{ij} o peso, x_j é a entrada no neurônio de entrada e T_i^{hid} é o termo de limiar de neurônios ocultos.

A função ReLU (*Rectified Linear Unit*) é uma função de ativação não-linear que se tornou popular em redes neurais devido a sua simplicidade e eficácia. Ela retorna o valor máximo entre zero e o argumento, o que significa que se o resultado da soma ponderada for negativo, o neurônio não será ativado e retornará zero. Caso contrário, a saída será igual ao valor da soma ponderada. Na construção de uma RNA pode-se utilizar diversas funções de ativação coma a ReLU, Sigmoide, Tangente Hiperbólica, dentre outras.

Durante o treinamento, a rede é exposta a um conjunto de exemplos de entrada e saída desejada, e os pesos das conexões são ajustados para minimizar o erro entre as saídas produzidas pela rede e as saídas desejadas. Esse processo de aprendizado é realizado por meio de algoritmos de otimização, como o Gradiente Descendente, que ajustam os pesos de forma a minimizar a função de perda da rede.

Uma das principais vantagens das RNAs é a sua capacidade de aprender a partir de exemplos, sem a necessidade de programação explicita. Seu potencial para aproximar funções complexas com alta precisão torna-as uma ferramenta valiosa para resolver problemas em diversas áreas.

Uma RNA construída no formato acima é capaz de aproximar qualquer função computável com uma precisão arbitrária. Os números entregues aos neurônios de entrada são variáveis independentes e os números retornados pelos neurônios de saída são as variáveis dependentes da função que está sendo aproximada pela RNA.

2.2 Técnicas multivariadas na previsão de geração de energia

A natureza estocástica da geração eólica é inerente a variabilidade das condições climáticas das quais se vale para gerar energia elétrica. Por conta disto, o comportamento errático do vento pode gerar problemas críticos durante a operação do empreendimento, bem como da rede como um todo. O desenvolvimento de predição de geração de energia eólica permite um dimensionamento da disponibilidade de energia em um determinado momento e que a partir disso possam ser tomadas decisões tanto no âmbito comercial como estrutural.

De Caro *et al.* propuseram um modelo baseado em *machine learning* dinâmico para previsão de geração de energia eólica, que implementa diversas técnicas de seleção de variáveis e obtiveram valores mais baixos de Ero Absoluto Médio (MAE) e Erro Quadrático Médio (MSE) do que métodos baseados em séries temporais. WU *et al.* aumentarem a acurácia de previsões de curtíssimo prazo para geração eólica propondo um modelo combinando *wavelet denoising* e redes neurais LSTM (*Long Short Term Memory).* Bai *et al.* utilizaram um modelo multivariado DFM (*Dynamic Factor Model)* para produzir cenários de previsão de geração de energia eólica para alimentar um modelo dinâmico estocástico para otimizar o fluxo de potência que alimenta baterias.

3. Procedimentos Metodológicos

Esta seção está dividida em duas partes. Inicia com a classificação da pesquisa, seguida pela modelagem e previsão da velocidade do vento através das técnicas elencadas no referencial teórico.

3.1. Classificação de pesquisa

O presente estudo pode ser classificado em relação a quatro categorias: natureza, abordagem, objetivos e procedimentos. Relativo à natureza, este estudo é considerado como pesquisa aplicada, já que os dados utilizados são reais e os resultados advindos poderão implicar em melhorias no empreendimento analisado. A segunda categoria trata da abordagem do estudo, no caso analisado o estudo é baseia-se em análises numéricas para implementação, execução e avaliação dos métodos propostos, considera-se a abordagem quantitativa. O objetivo do estudo é classificado como pesquisa exploratória, sendo que propõe a utilização associada de técnicas de previsão. Em relação ao procedimento, tratando-se de uma avaliação e proposição de metodologia focada em um empreendimento real, aspirando a melhoria de processos, este trabalho caracteriza-se como estudo de caso (GIL, 2010; VERGARA, 2013).

3.2. Etapas do trabalho

O presente trabalho apresenta uma metodologia para previsão através de modelos preditivos causais. Para tal, este estudo é calcado em três etapas operacionais, as quais serão detalhadas nessa seção: (i) Coleta dos dados históricos das variáveis climáticas associadas a cada passo temporal e da variável climática alvo (velocidade do vento à 80 metros de altura) (ii) Preparação dos dados (iii) ajuste dos modelos de previsão (iv) avaliação dos resultados obtidos por cada modelo (Figura 1).

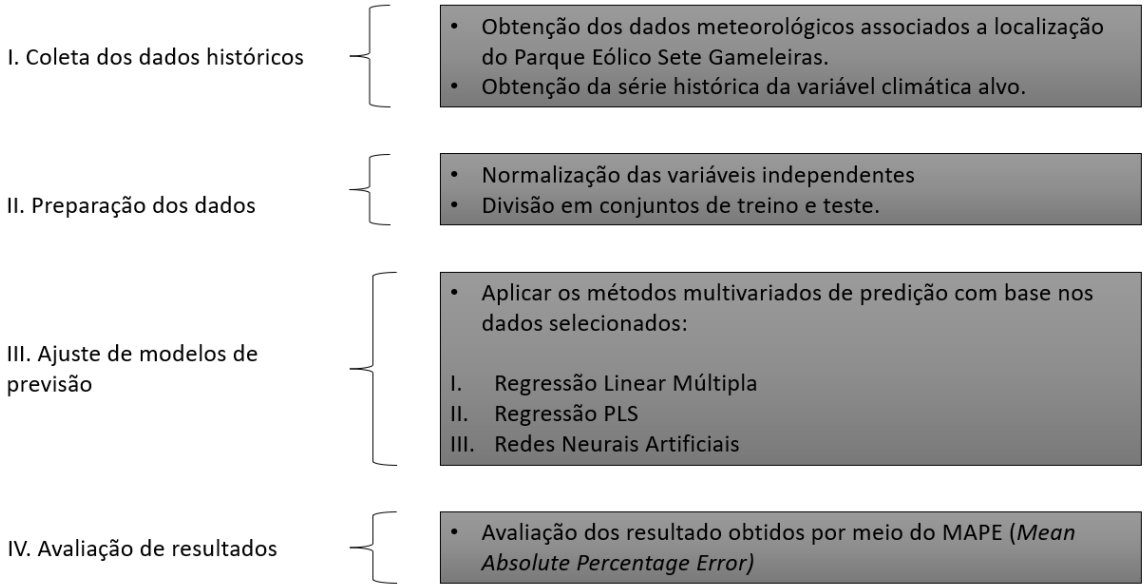

Figura 1 - Etapas do método proposto

Fonte: Elaborado pelo próprio autor

3.2.1. Coleta de dados

Os dados coletados para utilização neste trabalho consistem em dados meteorológicos obtidos através da empresa Visual Crossing, líder em fornecimento de dados meteorológicos. O banco de dados conta com 6428 entradas que descrevem o comportamento climático horário da região em que o parque está instalado. As variáveis independentes que compõem o banco de dados são: temperatura, ponto de orvalho, umidade, precipitação, rajadas de vendo, velocidade do vento ao nível do solo, direção do vento ao nível do solo, pressão atmosférica, nebulosidade e irradiação solar. De acordo com a literatura, as variáveis meteorológicas mais utilizadas para modelagens preditivas baseadas para geração de energia eólica incluem velocidade do vento, direção do vento, temperatura, umidade e pressão (LIU *et al.,* 2015; HUANG *et al.,* 2011).

A variável dependente que consta no banco de dados é a velocidade do vento a 80 metros de altura. Esta variável dependente foi escolhida por conta da especificação do aerogerador utilizado no Parque Eólico Sete Gameleiras: modelo V-100 de 2MW da fabricante Vestas, que tem o rotor localizado a 80 metros de altura. A previsão da velocidade do vento na altura do rotor do aerogerador é de extrema importância para o setor da energia eólica, visto que a energia gerada por um aerogerador é

proporcional à velocidade do vento elevada ao cubo (o que significa que pequenas variações na velocidade do vento podem ter grandes impactos na geração de energia).

3.2.2. Preparação dos dados

A preparação dos dados é fundamental para assegurar que o modelo de análise seja capaz de gerar resultados confiáveis e válidos. Nesta pesquisa, a preparação dos dados envolveu três etapas principais: análise exploratória de dados, normalização das variáveis independentes e divisão do banco de dados em treino e teste.

A análise exploratória de dados é uma técnica importante para compreender a estrutura, a distribuição e as relações existentes nos dados (TUKEY, 1977). Nesta etapa, foi realizada uma análise descritiva das variáveis, para identificar outliers e possíveis anomalias nos dados. Além disso, a correlação entre as variáveis (vide tabela 1) foi analisada para identificar relações lineares entre as variáveis independentes e a variável dependente (HAIR et al., 2014).

	Temperatura	*Ponto de Orvalho*	*Umidade*	*Precipitação*	*Rajadas de Vento*	*Velocidade do Vento*	*Direção do Vento*	*Pressão Atmosférica*	*Nebulosidade*	*Irradiação Solar*	*Veloc. Vento 80 m.*
Temperatura	1										
Ponto de Orvalho	-0,319	1									
Umidade	-0,847	0,742	1								
Precipitação	-0,040	0,122	0,117	1							
Rajadas de Vento	-0,463	-0,162	0,202	-0,080	1						
Velocidade do Vento	0,119	-0,265	-0,254	-0,075	0,323	1					
Direção do Vento	-0,063	0,045	0,084	0,080	0,070	0,023	1				
Pressão Atmosférica	-0,591	-0,125	0,296	-0,062	0,543	0,349	-0,005	1			
Nebulosidade	-0,081	0,408	0,283	0,082	-0,096	0,001	0,022	0,020	1		
Irradiação Solar	0,547	-0,143	-0,462	-0,047	-0,355	0,345	-0,091	-0,034	0,101	1	
Veloc. Vento 80 m.	-0,248	-0,270	-0,011	-0,099	0,904	0,486	0,061	0,531	-0,068	-0,099	1

Tabela 1 - Correlação entre variáveis

Na sequência normalizou-se as variáveis independentes para evitar efeitos de escala por conta da magnitude (e.g. influência excessiva de variáveis com escalas maiores). Utilizou-se a técnica de normalização min-max, que transforma os dados de forma a situá-los no intervalo de 0 a 1, preservando a distribuição dos mesmos e mantendo a relação entre as variáveis (PATRO; SAHU, 2015).

Por fim, a divisão do banco de dados em conjuntos de treino e teste é essencial para avaliar o desempenho do modelo de análise em dados não utilizados na construção do modelo (KELLEHER et al., 2015). Neste estudo, os dados foram divididos em dois conjuntos: 70% para treino e 30% para teste, seguindo a proporção comumente adotada na literatura (TRAIN, 2009). Essa divisão permite a construção e a validação do modelo de análise de forma eficiente, minimizando o risco de sobreajuste (*overfitting*) e garantindo a capacidade de generalização do modelo.

A partir dessas etapas de preparação dos dados, o modelo de análise pôde ser desenvolvido com maior precisão e confiabilidade. A análise exploratória de dados possibilitou uma compreensão mais aprofundada das características dos dados, enquanto a normalização das variáveis independentes assegurou uma análise justa e adequada de suas contribuições ao modelo. Além disso, a divisão do banco de dados em conjuntos de treino e teste permitiu uma avaliação realista do desempenho do modelo, garantindo a generalização dos resultados obtidos e a aplicabilidade do modelo a novos casos.

3.2.3. Seleção das variáveis mais informativas para posterior geração dos modelos

Nesta etapa, aplica-se um algoritmo de seleção de variáveis (*feature selection*) previamente ao ajuste dos modelos aos dados. A seleção de variáveis é uma etapa crucial no processo de aprendizado de máquina, pois permite reduzir a dimensionalidade dos dados, melhorar a performance dos modelos e minimizar o ruído e a multicolinearidade entre as variáveis.

Neste estudo, para os modelos de RLM e RNA, foi utilizada uma abordagem sistemática para selecionar o melhor conjunto de variáveis a serem incluídas no modelo. A técnica empregada é baseada no método *SelectKBest* do pacote *Scikit-Learn*, uma biblioteca de aprendizado de máquina em *Python*. A função de pontuação utilizada na técnica foi o Coeficiente de Correlação de Pearson para classificar as

variáveis independentes com base em suas correlações com a variável dependente. Esta abordagem visa identificar as variáveis que tem maior sincronia com a variável de resposta e reduzir a dimensionalidade dos dados, minimizando assim o risco de sobreajuste e melhorando a generalização do modelo.

O processo é realizado iterativamente para um intervalo de valores de “k” (número de variáveis independentes selecionadas), variando de 1 a 10 (número máximo de variáveis independentes). Em cada iteração, as seguintes etapas são executadas:

1) As “k” melhores variáveis independentes são selecionadas do conjunto de dados de treino com base na pontuação calculada.
2) Um modelo de regressão (RLM/RNA) é construído e treinado a partir dos dados de treino selecionados.
3) O modelo treinado é usado para fazer previsões nos conjuntos de dados de treinamento selecionados (contendo apenas as “k” melhores variáveis independentes)
4) São calculadas métricas de desempenho para as previsões do conjunto de treinamento, incluindo Erro Percentual Absoluto Médio (MAPE), Coeficiente de Determinação (R^2) e Erro Quadrático Médio (MSE).

Após todas as iterações, os resultados são analisados levando em consideração o MAPE obtido em cada rodada para selecionar as variáveis independentes que deverão fazer parte do modelo a ser treinado nas próximas etapas. A iteração que obtiver o menor valor de MAPE será a vencedora e as variáveis constantes no seu conjunto “k” serão selecionadas, fornecendo informações importantes sobre as variáveis independentes que têm maior impacto sobre a variável dependente.

Para a Regressão PLS (que se apoia em fundamentos distintos de RLM e RNA), foi utilizada uma abordagem sistemática para identificar componentes latentes que maximizem a covariância entre as variáveis independentes e a variável dependente. O processo se dá forma iterativa para um intervalo de valores de “n” componentes (número de componentes latentes), variando de 1 a 10. O limite inferior do intervalo, 1, representa o mínimo necessário de variáveis latentes, enquanto o limite superior, 10, é igual ao número de variáveis independentes disponíveis no conjunto de dados. Ao explorar o espaço de soluções possíveis dentro desse intervalo, pode-se identificar o número ideal de componentes latentes que equilibra a

complexidade do modelo e o desempenho na previsão dos dados de teste. Essa abordagem iterativa permite obter um modelo PLS apoiado em número adequado de componentes, que melhora a eficácia e a robustez do modelo. Tal processo é executado como segue:

1) Um modelo de regressão PLS é criado com “n” componentes e é ajustado aos dados de treinamento.
2) O modelo PLS treinado é utilizado para fazer previsões nos conjuntos de dados de treinamento.
3) São calculadas métricas de desempenho para as previsões do conjunto de treinamento, incluindo Erro Percentual Absoluto Médio (MAPE), Coeficiente de Determinação (R^2) e Erro Quadrático Médio (MSE).

Após todas as iterações, os resultados são analisados levando em consideração o MAPE obtido em cada rodada para determinar quais variáveis independentes deverão continuar no modelo para as próximas etapas. A iteração que obtiver o menor valor de MAPE será a vencedora e o número de componentes latentes “n” será selecionado e utilizado nas próximas etapas.

3.2.4. Ajuste dos modelos regressivos aos dados

A partir da definição das variáveis independentes mais informativas e do número de componentes a serem consideradas em cada modelo PLS, executa-se o ajuste final dos modelos aos dados de treino e teste. Para tanto, aplica-se a técnica de validação cruzada *m-fold* aos dados de treinamento, cujo método consiste na divisão do conjunto de dados de treino em *m* subconjuntos mutuamente exclusivos contendo o mesmo número de observações. A partir disto, um subconjunto é reservado para teste (ou seja, validação da capacidade preditiva em dados não considerados na geração do modelo) e os outros *m-1* conjuntos são utilizados para estimar os parâmetros. Este procedimento é realizado *m* vezes alternando de forma circular o conjunto de teste (Figura 2).

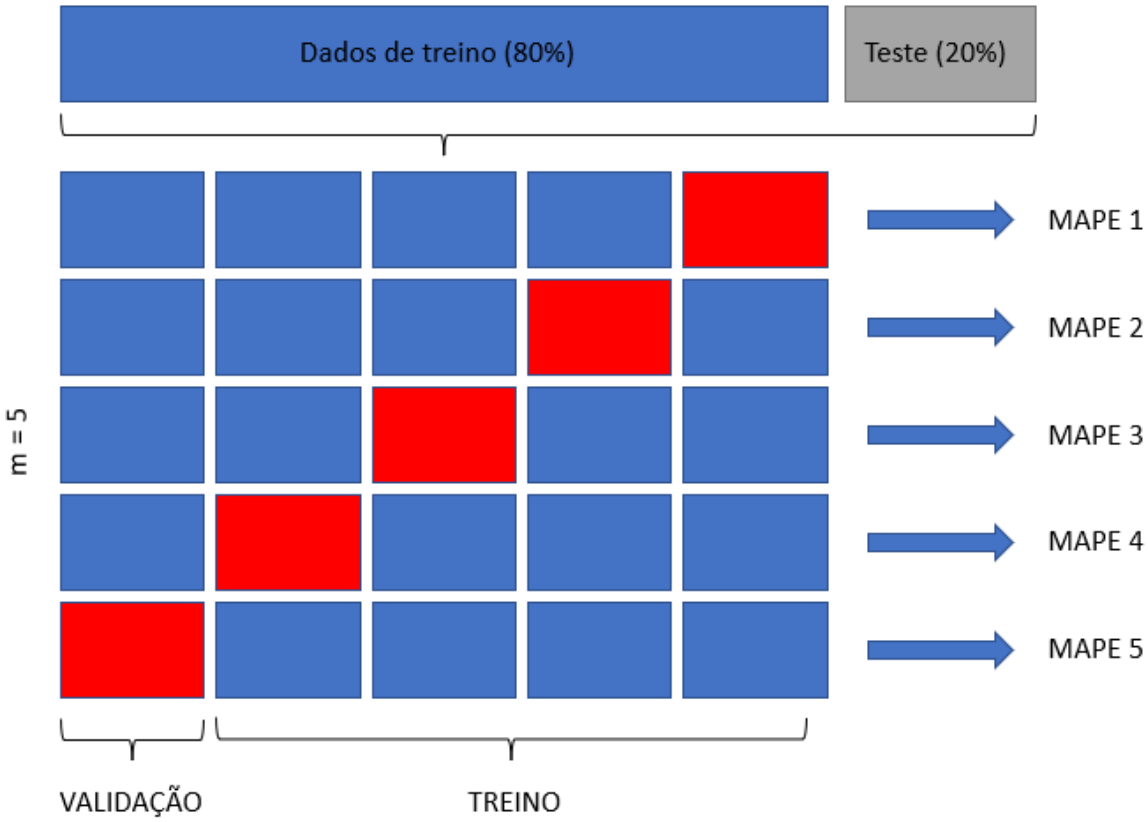

Figura 2 – Exemplo de Validação Cruzada m-Fold (m=5)

Fonte: Elaborado pelo próprio autor

Através do Erro Percentual Absoluto Médio (MAPE), calcula-se a acuracidade e validam-se as previsões a cada porção de teste. O MAPE baseia-se no cálculo dos erros percentuais absolutos de cada previsão, soma-se os erros absolutos individuais e os divide pela geração de cada período, o resultado é a média dos erros. Para cada método utilizado será calculado o MAPE médio a partir das *m* iterações da validação cruzada.

Por fim, após o treinamento dos modelos e da validação cruzada, o modelo treinado é utilizado para fazer previsões nos conjuntos de dados de teste a fim de se obter as métricas finais de avaliação dos modelos. A técnica multivariada responsável pelo menor MAPE é a indicada para geração de previsões de velocidade do vento voltadas à geração de energia.

4. Resultados

Nesta seção de apresentação dos resultados, serão abordadas as análises realizadas no contexto do Parque Eólico Sete Gameleiras, localizado no município de Santo Sé, Bahia. Este empreendimento de energia eólica possui uma capacidade instalada de 30 megawatts (MW) e encontra-se em uma região com elevado potencial eólico e condições favoráveis para a geração de energia a partir desta fonte. A análise do referido parque eólico tem por objetivo avaliar a eficácia de técnicas de previsão da geração de energia eólica, considerando a significativa variabilidade do vento na área. As técnicas empregas foram Regressão Linear Múltipla, Regressão PLS e Rede Neural Artificial.

Nesta seção, serão apresentados os resultados derivados da análise de um conjunto de 6.428 observações, estabelecendo-se a relação entre as variáveis independentes (temperatura, ponto de orvalho, umidade, precipitação, rajadas de vento, velocidade do vento ao nível do solo, direção do vento ao nível do solo, pressão atmosférica, nebulosidade e irradiação solar), e a variável dependente (velocidade do vento a uma altura de 80 metros). A análise busca compreender como as condições meteorológicas influenciam a velocidade do vento naquela altura e, consequentemente, a geração de energia eólica no Parque Eólico Sete Gameleiras. Nos próximos parágrafos, serão detalhados os resultados obtidos por meio deste estudo.

4.1. Seleção de Variáveis

A seguir, será apresentada a análise dos resultados referentes à seleção de variáveis para os três modelos em estudo. O critério adotado para a escolha das variáveis mais relevantes foi o menor valor do Erro Médio Absoluto Percentual (MAPE).

4.1.1. Regressão Linear Múltipla (RLM)

O menor MAPE foi obtido a partir da iteração com k=8, sendo as seguintes as variáveis independentes selecionadas: temperatura, ponto de orvalho, precipitação, rajadas de vento, velocidade do vento ao nível do solo, pressão atmosférica, nebulosidade e irradiação solar. A iteração vencedora apresentou valores de MAPE de 11,421% no conjunto de treino e 11,518% no conjunto de teste. Além disso, o

modelo alcançou um coeficiente de determinação R^2 de 0,896 no treino. Quanto aos valores do Erro Quadrático Médio (MSE), foram obtidos 0,999 m/s para o treino e 1,050 m/s para o teste, respectivamente. Os resultados obtidos durante o procedimento de seleção de variáveis podem ser observados na Tabela 2.

Nº Variáveis Ind.	MAPE treino [%]	MAPE teste [%]	R^2 treino	MSE treino [m/s]	MSE teste [m/s]
1	16,807	16,299	0,814	1,744	1,741
2	16,846	16,373	0,821	1,722	1,720
3	13,631	13,774	0,862	1,326	1,380
4	13,046	12,942	0,868	1,270	1,122
5	11,968	12,186	0,888	1,072	1,055
6	11,497	11,578	0,895	1,010	1,055
7	11,496	11,604	0,895	1,009	1,056
8	11,421	11,518	0,896	0,999	1,050
9	11,427	11,569	0,896	0,997	1,047
10	11,430	11,589	0,889	0,991	1,039

Tabela 2 - Seleção de Variáveis - Regressão Linear Múltipla

A análise dos coeficientes em um modelo de regressão linear múltipla permite estimar o efeito de cada variável independente na variável dependente (ver Tabela 3). Nesta tabela, os coeficientes indicam que a velocidade do vento a 80 metros de altura é influenciada positivamente pelas rajadas de vento, velocidade do vento ao nível do solo e pressão atmosférica, e negativamente pelo aumento do ponto de orvalho e nebulosidade. A temperatura é uma variável que apresenta uma relação positiva moderada com a variável dependente, enquanto a precipitação e irradiação solar possuem uma relação positiva fraca.

Temperatura	Ponto de Orvalho	Precipitação	Rajadas de Vento	Velocidade do Vento (solo)	Pressão Atmosférica	Nebulosidade	Irradiação Solar
2,778	-0,684	0,425	14,724	2,051	1,286	-0,344	1,154

Tabela 3 – Coeficientes do modelo de Regressão Linear Múltipla

4.1.2. Redes Neurais Artificiais (RNA)

O menor MAPE foi obtido a partir da iteração com k=6; as variáveis independentes selecionadas são: temperatura, ponto de orvalho, rajadas de vento, velocidade do vento ao nível do solo, pressão atmosférica e irradiação solar.

A iteração vencedora apresentou valores de MAPE de 9,385% no conjunto de treino e 9,577% no conjunto de teste. Além disso, o modelo alcançou um coeficiente de determinação R^2 de 0,924 no treino. Quanto aos valores do Erro Quadrático Médio

(MSE), foram obtidos 0,733 m/s para o treino e 0,779 m/s para o teste. Os resultados obtidos durante o procedimento de seleção de variáveis podem ser observados na Tabela 4.

Nº Variáveis Ind.	MAPE treino [%]	MAPE teste [%]	R^2 treino	MSE treino [m/s]	MSE teste [m/s]
1	14,597	14,277	0,825	1,676	1,669
2	17,010	16,493	0,820	1,728	1,717
3	13,349	13,453	0,862	1,324	1,365
4	12,424	12,151	0,876	1,189	1,197
5	52,264	49,399	0,000	9,602	9,437
6	9,385	9,577	0,924	0,733	0,779
7	10,817	10,895	0,914	0,828	0,860
8	10,392	10,475	0,918	0,789	0,835
9	9,978	10,119	0,920	0,770	0,815
10	10,609	10,664	0,919	0,777	0,813

Tabela 4 - Seleção de Variáveis - Rede Neural Artificial

A rede neural utilizada no modelo possui três camadas. A primeira camada densa tem 10 neurônios e usa a função de ativação *ReLU*. A segunda camada densa tem 10 neurônios e usa a função de ativação *ReLU*. A última camada tem apenas um neurônio de saída e usa a função de ativação linear. O modelo é treinado com a função de perda de Erro Médio Quadrático (*Mean Squared Error*). Em resumo, a rede neural é uma arquitetura simples de regressão que tenta aprender uma relação não-linear entre as entradas e a saída.

4.1.3. Regressão *Partial Least Squares* (PLS)

Para o modelo PLS, o menor valor de MAPE foi obtido com apenas 1 componente, sendo o MAPE de treino de 61,356% e MAPE de teste de 58,939%. O R^2 de treino foi de 0,731. Os valores de MSE foram 2,580 m/s e 2,549 m/s para treino e teste, respectivamente. Os resultados obtidos durante o procedimento de seleção de variáveis podem ser observados na Tabela 5.

Nº Variáveis Ind.	MAPE treino [%]	MAPE teste [%]	R^2 treino	MSE treino [m/s]	MSE teste [m/s]
1	61,356	58,939	0,731	2,580	2,549
2	62,398	60,083	0,815	11,776	1,790
3	63,183	61,075	0,882	1,135	1,205
4	63,332	61,230	0,894	1,022	1,072
5	63,374	61,242	0,897	0,993	1,044
6	63,376	61,237	0,897	0,992	1,041
7	63,377	61,239	0,897	0,992	1,040
8	63,378	61,238	0,897	0,992	1,040
9	63,378	61,237	0,897	0,992	1,040
10	63,378	61,234	0,897	0,991	1,039

Tabela 5 – Seleção de Componentes – Regressão PLS

Os coeficientes encontrados em um modelo de regressão PLS permitem estimar o efeito de cada variável independente na variável dependente. Ao analisar os valores das variáveis apresentadas na Tabela 6 é possível observar que algumas delas apresentam correlações mais fortes com a variável dependente, como é o caso das rajadas de vento, que possuem um peso positivo forte na predição da velocidade do vento na altura do rotor, assim como a pressão atmosférica, que também possui um peso positivo moderado. Já outras variáveis apresentam correlações mais fracas, como a umidade, precipitação, nebulosidade e irradiação solar, que possuem pesos negativos ou próximos a zero.

Temperatura	Ponto de Orvalho	Umidade	Precipitação	Rajadas de Vento	Velocidade do Vento (solo)	Direção do Vento	Pressão Atmosférica	Nebulosidade	Irradiação Solar
-0,373	-0,413	-0,023	-0,139	1,361	0,740	0,102	0,798	-0,107	-0,143

Tabela 6 – Coeficientes da regressão PLS

Comparando os três modelos, a RNA apresentou os melhores resultados em termos de MAPE e R^2, tanto para os dados de treino quanto para os dados de teste. A RLM também obteve resultados razoáveis, porém inferiores aos da RNA. Por fim, o modelo PLS apresentou os piores resultados, com valores de MAPE mais altos e R^2 mais baixos.

4.2. Ajuste final dos modelos regressivos com base nas variáveis selecionadas

Após realizar a seleção de variáveis para os três modelos (Regressão Linear Múltipla – RLM, Redes Neurais Artificiais – RNA e *Partial Least Squares* – PLS), uma validação cruzada foi realizada utilizando a porção de treinamento do banco de dados,

selecionados na etapa anterior. A validação cruzada foi executada com 10 *folds* para cada modelo, e os resultados (Tabela 6) obtidos são apresentados a seguir.

A validação cruzada para o modelo RLM mostrou um desempenho consistente com os resultados anteriores. O MAPE médio foi de 11,443%, com um desvio padrão de 0,714%, indicando uma boa precisão do modelo. O R^2 médio foi de 0,895, com um desvio padrão de 0,011, demonstrando que o modelo explica cerca de 89,5% da variância nos dados. O MSE médio foi de 1,003 m/s e seu desvio padrão foi de 0,075 m/s, sugerindo que o modelo tem um erro quadrático médio relativamente baixo. Esses resultados indicam que a RLM tem um bom desempenho na previsão dos dados, embora não seja o melhor entre os três modelos analisados.

O modelo RNA apresentou um desempenho ainda melhor na validação cruzada. O MAPE médio foi de 10,328%, com um desvio padrão de 0,810%, o que indica que o modelo é mais preciso do que a RLM. O R^2 médio foi de 0,910 e seu desvio padrão foi de 0,013, mostrando que o modelo explica aproximadamente 91% da variância nos dados. O MSE médio foi de 0,858 m/s e seu desvio padrão foi de 0,113 m/s, revelando que o modelo possui um erro quadrático médio menor do que o da RLM. A RNA demonstrou ser o modelo mais adequado entre os três analisados, com o melhor desempenho em termos de precisão e capacidade explicativa.

O modelo PLS, por outro lado, apresentou o pior desempenho na validação cruzada entre os três modelos. O MAPE médio foi de 61,220%, com um desvio padrão de 13,151%, indicando uma precisão significativamente menor em comparação com a RLM e a RNA. O R^2 médio foi de 0,725 e seu desvio padrão foi de 0,026, o que significa que o modelo explica apenas cerca de 72,5% da variância nos dados. O MSE médio foi de 2,620 m/s e seu desvio padrão foi de 0,233 m/s, mostrando que o modelo tem um erro quadrático médio consideravelmente mais alto do que os outros dois modelos. Portanto, o PLS não é o modelo mais adequado para esta aplicação.

Esses resultados reforçam a conclusão anterior de que a RNA é o modelo mais adequado entre os três para esta aplicação. A validação cruzada também ajuda a confirmar a robustez do modelo escolhido, uma vez que fornece uma estimativa mais confiável do desempenho do modelo em novos dados.

Após a validação cruzada (dados de treino), os modelos foram aplicados na porção de teste dos dados e os resultados obtidos foram:

	RLM	RNA	PLS
MAPE médio - treino[%]	11,443	10,328	61,220
MAPE desvio padrão - treino[%]	0,714	0,810	13,151
R^2 médio - treino	0,895	0,910	0,725
R^2 desvio padrão - treino	0,011	0,013	0,026
MSE médio [m/s] - treino	1,003	0,858	2,620
MSE desvio padrão [m/s] - médio	0,075	0,113	0,233
MAPE - teste [%]	11,518	11,229	58,939
MSE - teste [m/s]	1,050	0,945	2,549

Tabela 7 - Resultados

Analisando os resultados (Tabela 7) obtidos na porção de teste dos dados, a RNA continua sendo o modelo de melhor desempenho entre os três, com o menor MAPE (11,229%) e o maior R^2 (0,899). O modelo RLM também apresenta um bom desempenho, com um MAPE de 11,518% e um R^2 de 0,888, porém ainda inferior à RNA. O modelo PLS, por outro lado, mantém o pior desempenho, com um MAPE significativamente maior (58,939%) e um R^2 menor (0,728).

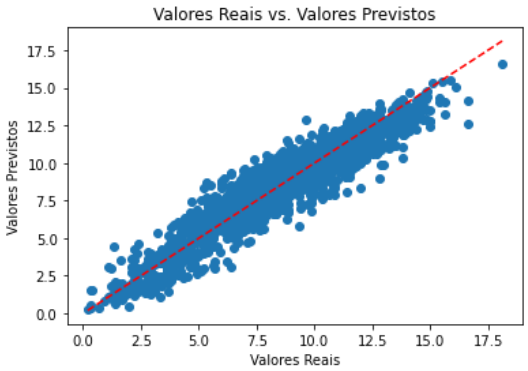

Figura 3 - Regressão Linear Múltipla

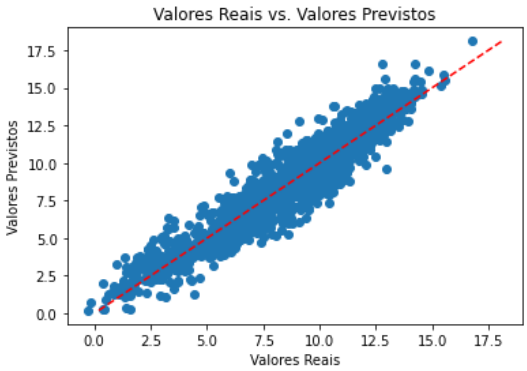

Figura 3 - Rede Neural Artificial

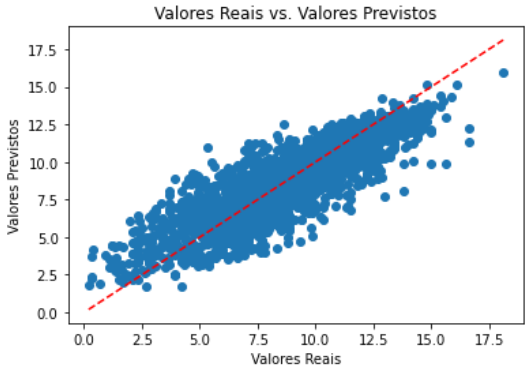

Figura 5 - Regressão PLS

Os gráficos de dispersão (Figuras 3 a 5) contendo os valores reais e os valores previstos permitem visualizar a qualidade do ajuste do modelo. Para todos os modelos é possível verificar que não há *overfitting* e *underfitting* uma vez que os dados se distribuem ao longo da linha central.

Além disso, todos os modelos apresentaram resíduos que se distribuem aleatoriamente (Figuras 6 a 8) e têm um histograma que se aproxima de uma distribuição normal (Figuras 9 a 11). Isso indica que os modelos são capazes de capturar a maior parte das variações nos dados e que os erros são predominantemente aleatórios, sem padrões sistemáticos. A presença de resíduos com distribuição normal é uma das suposições básicas de muitos modelos de regressão, incluindo a RLM, e sugere que os modelos estão bem ajustados aos dados.

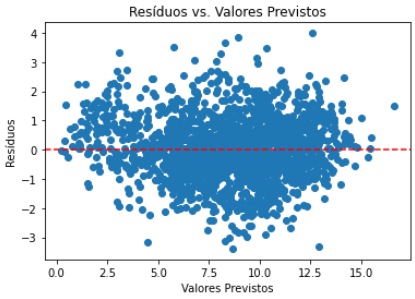

Figura 6 - Regressão Linear Múltipla

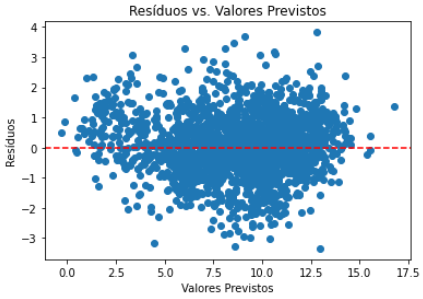

Figura 7 - Rede Neural Artificial

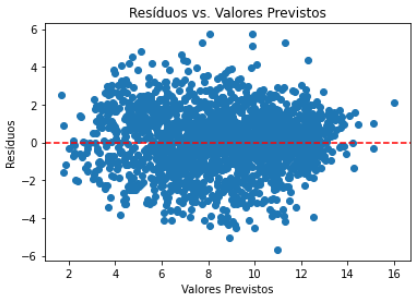

Figura 8 - Regressão PLS

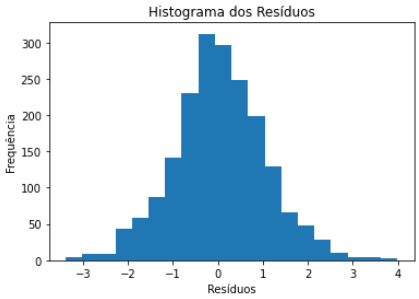

Figura 9 - Regressão Linear Múltipla

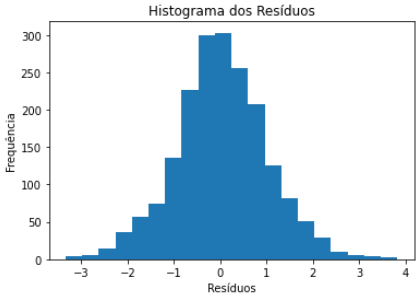

Figura 10 - Rede Neural Artificial

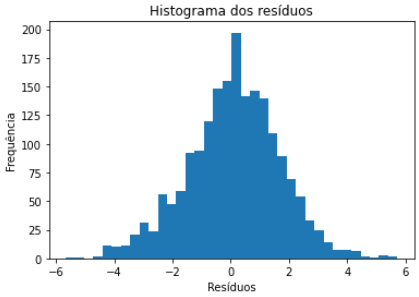

Figura 11 - Regressão PLS

Com base no conjunto de resultados acima descritos, conclui-se que a RNA é o modelo mais adequado entre os três para esta aplicação. Os resultados na porção de teste dos dados corroboram a análise feita com base na validação cruzada e demonstram que a RNA é capaz de fornecer previsões mais precisas e com melhor ajuste aos dados. A análise dos resíduos reforça a adequação dos modelos e fornece mais confiança na aplicabilidade da RNA para este problema específico.

5. Conclusões

Neste trabalho o objetivo principal era selecionar o melhor modelo para prever a velocidade do vento a 80 metros de altura, correspondente à altura do rotor de um aerogerador. Para isso, foram testados e comparados três modelos de aprendizado de máquina: Regressão Linear Múltipla (RLM), Redes Neurais Artificiais (RNA) e *Partial Least Squares* (PLS). A seleção de variáveis e a validação cruzada foram aplicadas para garantir a robustez e a generalização dos modelos. Além disso, os resíduos dos modelos foram analisados para verificar a adequação às suposições básicas e identificar possíveis padrões sistemáticos nos erros.

A partir da análise dos resultados, foi constatado que a RNA é o modelo mais adequado para prever a velocidade do vento a 80 metros de altura entre os três modelos estudados. A RNA apresentou o melhor desempenho tanto na validação cruzada quanto na porção de teste dos dados, com o menor MAPE e o maior R^2. Os resíduos dos modelos apresentaram uma distribuição aleatória e normal, indicando que os modelos estão bem ajustados aos dados e que os erros são predominantemente aleatórios. Portanto, pode-se concluir que a RNA é a abordagem mais promissora para estimar a velocidade do vento a 80 metros de altura e, assim, contribuir para a eficiência na geração de energia eólica. A aplicação bem-sucedida da RNA neste contexto demonstra a utilidade e a versatilidade das técnicas de aprendizado de máquina para resolver problemas complexos e melhorar a eficiência em diversos setores, incluindo a energia renovável.

6. Referências Bibliográficas

LETCHER, Trevor M. (Ed.). Future energy: improved, sustainable, and clean options for our planet. Elsevier, 2020.

PINSON, Pierre. Wind Energy: Forecasting challenges for its operational management. Statistical Science, v. 28, n. 4, p. 564-585, 2013.

COUNCIL, Global Wind Energy. GWEC Global Wind Report 2022. Global Wind Energy Council: Bonn, Germany, 2022.

MIGUEL, José Vítor Pereira. Avaliação da geração de energia elétrica no Brasil em condições de escassez de recursos eólicos e hídricos. Tese de Doutorado. Universidade de São Paulo.

LANGE, Matthias; FOCKEN, Ulrich. New developments in wind energy forecasting. In: 2008 IEEE power and energy society general meeting-conversion and delivery of electrical energy in the 21st century. IEEE, 2008. p. 1-8.

JÓNSSON, Tryggvi; PINSON, Pierre; MADSEN, Henrik. On the market impact of wind energy forecasts. Energy Economics, v. 32, n. 2, p. 313-320, 2010.

MALINOWSKI, E. R. Factor Analysis in Chemistry. New York: John Wiley and Sons, 2002.

BARTHOLOMEW, D. J. The interpretation of Multivariate Data. International Encyclopedia of Education, 3, 12-17, 2010.

LATTIN, J.M.; CARROL, J.D.; GREEN, P. E. Análise de dados multivariados. São Paulo: Cengage Learning, 2011. 475 p.

JOHNSON, R. A.; WICHERN, D. W. Applied Multivariate Statistical Analysis. New Jersey: Pearson Prentice Hall, 2007.

TOBIAS, R. D. An Introduction to Partial Least Squares Regression, SAS Institute Inc., Cary, NC, 1995.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. Anal. Chim. Acta, 185, 1-17, 1986.

ABDI, H., Partial least square regression, projection on latent structure regression, PLS-Regression. Wiley Interdiscip. Rev. Comput. Stat 2, 97-106, 2010.

WANG, Sun-Chong. Artificial neural network. In: Interdisciplinary computing in java programming. Springer, Boston, MA, 2003. p. 81-100.

LAI J-P, Chang Y-M, Chen C-H, Pai P-F. A Survey of Machine Learning Models in Renewable Energy Predictions. Applied Sciences. 2020; 10(17):5975.

HEINERMANN, Justin; KRAMER, Oliver. Machine learning ensembles for wind power prediction. Renewable Energy, Volume 89, 2016, Pages 671-679.

SHARIFZADEH, Mahdi; SIKINIOTI-LOCK, Alexandra; SHAH, Nilay. Machine-learning methods for integrated renewable power generation: A comparative study of

artificial neural networks, support vector regression, and Gaussian Process Regression. Renewable and Sustainable Energy Reviews, Volume 108, 2019.

GIL, A. C. Como elaborar projetos de pesquisa. 5. ed. Atlas, 2010.

VERGARA, S. C. Projetos e Relatórios de Pesquisa em Administração. 12. ed. Atlas, 2013.

DE CARO, F.; DE STEFANI, J.; VACCARO, A.; BONTEMPI, G. DAFT-E: Feature-Based Multivariate and Multi-Step-Ahead Wind Power Forecasting. In: IEEE Transactions on Sustainable Energy, vol. 13, no. 2, pp. 1199-1209, April 2022.

WU, S.; JIA, L.; LIU, Y. Ultra-short-term wind energy prediction based on wavelet denoising and multivariate LSTM. 2021 Power System and Green Energy Conference (PSGEC), 2021, pp. 443-447.

BAI, W.; LEE, D.; LEE, K. Y. A multivariate time series forecast model for wind and storage integrated system operation. 2017 IEEE Power & Energy Society General Meeting, 2017, pp. 1-5, doi: 10.1109/PESGM.2017.8274436.

EPE. Empresa de Pesquisa Energética; Plano Decenal de Expansão de Energia 2031. Eficiência Energética e Recursos Energéticos Distribuídos: Micro e minigeração. Brasília: MME/EPE, 2020.

PINHEIRO, André Fialho; HERVÉ, Márcio. Riscos e Desafios no Desenvolvimento de Parques Eólicos no Brasil. Boletim do Gerenciamento, [S.l.], v. 21, n. 21, p. 65-72, dez. 2020. ISSN 2595-6531.

LIU, Da; WANG, Jilong; WANG, Hui. Short-term wind speed forecasting based on spectral clustering and optimized echo state networks. Renewable Energy, v. 78, p. 599-608, 2015.

HUANG, Chi-Yo et al. Predicting of the short term wind speed by using a real valued genetic algorithm based least squared support vector machine. In: Intelligent Decision Technologies. Springer, Berlin, Heidelberg, 2011. p. 567-575.

DRAPER, N. R.; SMITH, H.; POWNELL, E. Applied Regression Analysis. 3. ed. New York: John Wiley & Sons, 1998.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to Linear Regression Analysis. 5. ed. Hoboken, NJ: Wiley, 2012.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013.

WOOLDRIDGE, J. M. Introductory Econometrics: A Modern Approach. 5. ed. Mason, OH: South-Western Cengage Learning, 2012.

HAYASHI, F. Econometrics. Princeton, NJ: Princeton University Press, 2000

MORELLATO, Saulo Almeida. Modelos de regressão PLS com erros heteroscedásticos. 2010. 60 f. Dissertação (Mestrado em Ciências Exatas e da Terra) - Universidade Federal de São Carlos, São Carlos, 2010.

WOLD, S.; SJOSTRON, M.; LIN J. G. PLS-regression: a basic tool of chemometrics, Chemometrics and Intelligent Laboratory Systems, 58: 109-130. 2001.

TUKEY, John W. et al. Exploratory data analysis. 1977.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. Multivariate Data Analysis. 7. ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2014.

PATRO, SGOPAL; SAHU, Kishore Kumar. Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462, 2015.

KELLEHER, John D.; MAC NAMEE, Brian; D'ARCY, Aoife. Fundamentals of machine learning for predictive data analytics: algorithms. Worked examples, and case studies, 2015.