PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO – PUC RJ

DEPARTAMENTO DE ENGENHARIA QUÍMICA E MATERIAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ESTRUTURAL

EQM2118 – MODELAGEM MATEMÁTICA COM APLICAÇÃO DE INTELIGÊNCIA ARTIFICAL

VINÍCIUS DOS SANTOS MOTA

RELATÓRIO

MODELAGEM PREDITIVA DA GERAÇÃO DE ENERGIA SOLAR E EÓLICA NO NORDESTE BRASILEIRO

RIO DE JANEIRO - RJ

OUTUBRO DE 2025

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

VINÍCIUS DOS SANTOS MOTA

MODELAGEM PREDITIVA DA GERAÇÃO DE ENERGIA SOLAR E EÓLICA NO NORDESTE BRASILEIRO

TRABALHO 1

Trabalho apresentado como requisito parcial para obtenção da nota na disciplina de Modelagem Matemática com Aplicação de Inteligência Artificial, do programa de Pós-Graduação em Engenharia, do Departamento de Engenharia Química e de Materiais da PUC-Rio.

Prof. Dr. Brunno Ferreira dos Santos

RIO DE JANEIRO - RJ

OUTUBRO DE 2025

Sumário

Lista de figuras

1. INTRODUÇÃO

Nos últimos anos, o emprego de técnicas de machine learning (ML) vem ampliando significativamente a forma de modelar e prever fenômenos com alta variabilidade, como a geração de energia a partir de fontes renováveis. Em particular, a produção de energia solar e eólica assume papel central devido à crescente demanda por matrizes energéticas sustentáveis. No Brasil, e especialmente na região Nordeste, as condições climáticas — com boa incidência solar e regimes de vento relativamente favoráveis — tornam possível explorar essa produção renovável em escala relevante.

A literatura recente destaca que o uso de variáveis meteorológicas como radiação solar, velocidade e direção do vento, umidade relativa do ar e pressão atmosférica, combinado a algoritmos de ML, tem propiciado melhorias expressivas na previsão de geração eólica e solar. Por exemplo, no contexto brasileiro, estudos de irradiação solar via ML demonstraram que modelos como SVM, ANN e outras variantes alcançaram níveis satisfatórios de desempenho ao predizer radiação ou energia solar (Viscondi & Alves-Souza, 2021) . Em outra linha, para previsão conjunta de recursos solar e eólico, ensaios com ensemble baseados em regressão penalizada (ridge) mostraram que a combinação de modelos pode reduzir de modo significativo os erros de previsão, inclusive no Brasil (Carneiro et al., 2022) .

Diante desse cenário, o presente estudo propõe desenvolver uma análise preditiva da geração energética mensal — tanto solar quanto eólica — em sete estados do Nordeste brasileiro (Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí e Rio Grande do Norte), no período entre 2017 e 2024. Os dados contemplam médias mensais de geração (em MW) por estado e fonte de energia, bem como variáveis meteorológicas agregadas (originais em dados horários) fornecidas pelo Instituto Nacional de Meteorologia (INMET). O objetivo desta primeira etapa consiste em construir, limpar e integrar os bancos de dados, realizar análise exploratória, aplicar técnica de redução de dimensionalidade e seleção de variáveis (PCA e LASSO) com vistas à formação de conjuntos de dados finais robustos para modelagem preditiva.

Com isso, pretende-se oferecer uma base de dados consistente e adequada para alimentar modelos de ML em etapas subsequentes, bem como investigar quais variáveis meteorológicas se mostram mais relevantes para a previsão da geração de energia

renovável, contribuindo assim para o planejamento energético em regiões com elevado potencial.

2. METODOLOGIA

Esta seção descreve os procedimentos adotados para a construção do dataset, desde a aquisição e pré-processamento dos dados até a aplicação das técnicas de análise exploratória, redução de dimensionalidade e seleção de variáveis. O foco está na preparação de um conjunto de dados adequado para tarefas de previsão de geração de energia solar e eólica, a partir de variáveis climáticas.

2.1. Banco de Dados

Dois conjuntos principais de dados foram utilizados neste estudo:

- Dados de Geração de Energia Renovável:

Foram coletadas séries temporais mensais de geração de energia solar e eólica, medidas em megawatts médios (MWmed), para sete estados da região Nordeste do Brasil: Bahia, Ceará, Maranhão, Paraíba, Pernambuco, Piauí e Rio Grande do Norte. As informações cobrem o período de 2017 a dezembro de 2024 e foram organizadas por estado e tipo de fonte (solar ou eólica). Esses dados foram coletados no site do ONS(Operador Nacional do Sistema Elétrico), que é o órgão encarregado de coordenar e controlar a operação das instalações de geração e transmissão de energia elétrica do Sistema Interligado Nacional (SIN), bem como de planejar a operação dos sistemas isolados do país, atuando sob a supervisão e regulação da Agência Nacional de Energia Elétrica (Aneel). A obtenção desses dados permite uma análise segmentada e temporalmente consistente da produção energética renovável.

- Dados Meteorológicos:

As variáveis climáticas foram obtidas por meio do Instituto Nacional de Meteorologia (INMET), que disponibiliza registros horários por estação meteorológica, disponíveis desde 2000. As variáveis coletadas incluem

- PRECIPITAÇÃO TOTAL, HORÁRIA (mm): Quantidade de chuva registrada por hora.
- PRESSÃO ATMOSFÉRICA AO NÍVEL DA ESTAÇÃO, HORÁRIA (mB): Valor da pressão atmosférica na altura da estação meteorológica.

- PRESSÃO ATMOSFÉRICA MÁXIMA NA HORA ANTERIOR (mB): Valor máximo registrado de pressão na hora anterior.
- PRESSÃO ATMOSFÉRICA MÍNIMA NA HORA ANTERIOR (mB): Valor mínimo registrado de pressão na hora anterior.
- RADIAÇÃO GLOBAL (kJ/m²): Quantidade de energia solar incidente por metro quadrado.
- TEMPERATURA DO AR - BULBO SECO, HORÁRIA (°C): Temperatura do ar ambiente medida com termômetro seco.
- TEMPERATURA DO PONTO DE ORVALHO (°C): Temperatura em que o ar se torna saturado (umidade relativa de 100%) e o vapor d'água começa a condensar.
- TEMPERATURA MÁXIMA NA HORA ANTERIOR (°C): Maior temperatura registrada na hora anterior.
- TEMPERATURA MÍNIMA NA HORA ANTERIOR (°C): Menor temperatura registrada na hora anterior.
- TEMPERATURA DO PONTO DE ORVALHO MÁXIMO NA HORA ANTERIOR (°C): Maior temperatura de orvalho registrada na hora anterior.
- TEMPERATURA DO PONTO DE ORVALHO MÍNIMO NA HORA ANTERIOR (°C): Menor temperatura de orvalho registrada na hora anterior.
- UMIDADE RELATIVA MÁXIMA NA HORA ANTERIOR (%): Valor máximo de umidade relativa do ar registrado na hora anterior.
- UMIDADE RELATIVA MÍNIMA NA HORA ANTERIOR (%): Valor mínimo de umidade relativa na hora anterior.
- UMIDADE RELATIVA DO AR, HORÁRIA (%): Umidade relativa do ar registrada na hora.
- VENTO - DIREÇÃO HORÁRIA (°): Direção média do vento em graus em relação ao norte verdadeiro.
- VENTO - RAJADA MÁXIMA (m/s): Maior velocidade instantânea do vento registrada na última hora.
- VENTO - VELOCIDADE HORÁRIA (m/s): Velocidade média do vento ao longo da hora.

2.2. Tratamento Inicial dos Dados

2.2.1. Escolhas Metodológicas Iniciais

Durante o processo de consolidação dos dados, observou-se que os registros de geração de energia solar e eólica estavam disponíveis de forma consistente apenas a partir de 2017. Embora os dados meteorológicos do INMET estivessem disponíveis desde anos anteriores (com início variando por estação), a falta de geração energética nesse período inviabilizava a realização de análises conjuntas.

Dessa forma, para garantir a compatibilidade temporal entre as bases, foi definido que a análise consideraria o intervalo entre janeiro de 2017 e dezembro de 2024. Essa janela oferece um período suficiente para observar padrões sazonais e interanuais, com qualidade e completude adequadas nas duas fontes de dados.

2.2.2. Estrutura dos dados metodológicas

Os dados meteorológicos coletados junto ao Instituto Nacional de Meteorologia (INMET) são fornecidos com frequência horária e organizados por estação meteorológica. No entanto, como a unidade de análise dos dados energéticos é o estado (e não a estação) e sua frequência é mensal, foi necessário realizar um processo de agregação cuidadoso para garantir a compatibilidade entre as bases.

Para isso, adotou-se uma estratégia de agregação mensal por estado, a partir das observações horárias de todas as estações meteorológicas localizadas nos estados mencionados. Assim, dentro de cada mês e estado, todos os registros horários disponíveis foram combinados para formar um conjunto de variáveis climáticas consolidadas que representam as condições meteorológicas médias daquele local e período.

Além de garantir compatibilidade temporal e espacial com os dados de geração de energia, essa escolha também contribui para reduzir a sensibilidade a falhas isoladas de estações específicas, diluindo eventuais lacunas ou ruídos individuais.

Ademais, para cada variável meteorológica original (como temperatura, radiação, umidade, etc.) foram aplicadas funções estatísticas de agregação, resultando em cinco variáveis derivadas para cada dimensão climática:

- Média mensal (mean): Representa o valor médio da variável ao longo do mês em todas as estações do estado.
- Desvio padrão (std): Mede a variabilidade intra-mensal das observações, importante para identificar instabilidade climática.
- Valor mínimo (min): Captura extremos inferiores que podem afetar negativamente a produção energética (ex: baixas radiações).
- Valor máximo (max): Registra extremos superiores que podem sinalizar picos de produção (ex: ventos intensos).
- Total de registros válidos (total_records): Indica a densidade e qualidade dos dados disponíveis naquele mês, servindo também como variável auxiliar no controle da robustez estatística.

Esse conjunto de variáveis permite modelar não apenas condições médias, mas também flutuações e extremos climáticos, que são frequentemente determinantes na produção de energia solar e eólica. Além disso, variáveis como n_stations (número de estações com dados disponíveis no mês) e os identificadores temporais year e month foram mantidos para controle e análise posterior.

2.3. Análise Exploratória de Dados

2.3.1. Geração de Energia ao Longo do Tempo

A análise exploratória teve início com a verificação da disponibilidade e completude dos dados de geração de energia e das variáveis climáticas. Observou-se que os dados de geração de energia solar e eólica não estavam uniformemente distribuídos entre os estados e ao longo do tempo. No caso da energia solar, constatou-se a ausência total de registros para o estado do Maranhão (MA). Além disso, alguns estados, como Ceará (CE) e Paraíba (PB), apresentaram início de registros consideravelmente mais tardios que os demais. O CE, por exemplo, só possui dados a partir de novembro de 2018, o que motivou o primeiro filtro temporal aplicado aos dados solares: foram mantidos apenas os registros a partir de novembro de 2018, garantindo uma base comum entre todos os estados participantes da análise.

Para a energia eólica, a situação é semelhante. Embora todos os estados analisados possuam dados desde 2017, a Paraíba (PB) apresenta registros apenas a partir de julho de

2021. Para garantir comparabilidade entre os estados, os dados eólicos foram filtrados para incluir apenas observações a partir de julho de 2021. Esse critério resultou em duas bases finais de geração energética, uma para a fonte solar e outra para a fonte eólica , com recortes temporais distintos, mas mais homogêneos internamente.

A Figura 1 apresenta a evolução mensal da geração de energia solar nos estados selecionados, já com o filtro temporal aplicado. Nota-se uma trajetória ascendente e relativamente contínua na maioria dos estados, especialmente em Bahia (BA), Piauí (PI) e Ceará (CE), que demonstram forte crescimento nos últimos anos. A ausência de grandes oscilações sazonais sugere que a radiação solar na região é relativamente estável, característica esperada para o clima predominante no Nordeste brasileiro.

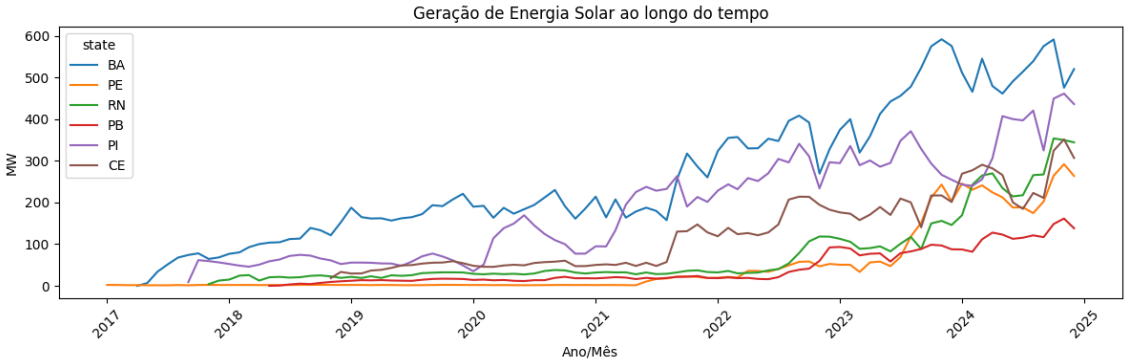

Figura 1. Geração de Energia Solar Ao Longo do Tempo (2017-2024)

Por outro lado, a Figura 2 mostra a série histórica da geração de energia eólica. A produção apresenta padrões sazonais bem definidos, com variações mensais expressivas ao longo dos anos. Os estados do Rio Grande do Norte (RN) e da Bahia (BA) lideram em volume de geração, com picos regulares associados às condições favoráveis de vento. A variabilidade elevada, mesmo entre anos consecutivos, reforça a necessidade de modelos preditivos que considerem a sazonalidade e outros padrões climáticos complexos.

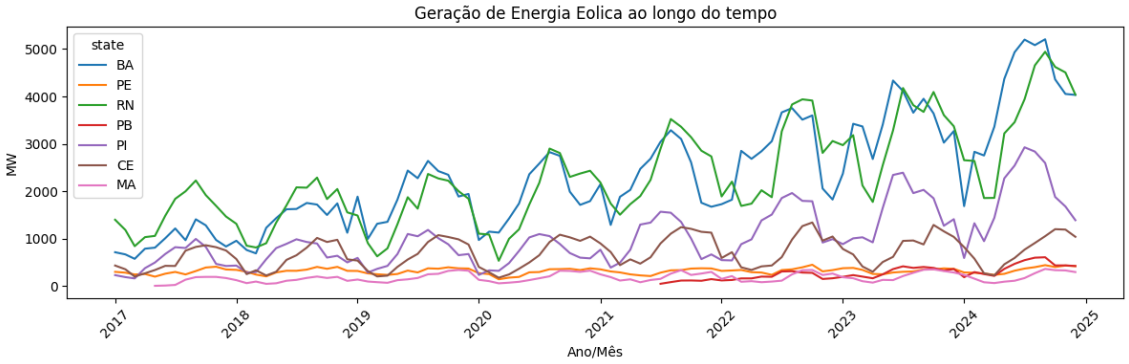

Figura 2. Geração de Energia Eólica Ao Longo do Tempo (2017-2024)

2.3.2. Distribuição das Variáveis Climáticas

Além da produção de energia, as variáveis climáticas também foram exploradas para entender seus comportamentos estatísticos antes da modelagem. A Figura 3 ilustra a distribuição da variável precipitação média, revelando um perfil fortemente assimétrico à direita, com predominância de registros próximos a zero, reflexo das características semiáridas de parte da região analisada.

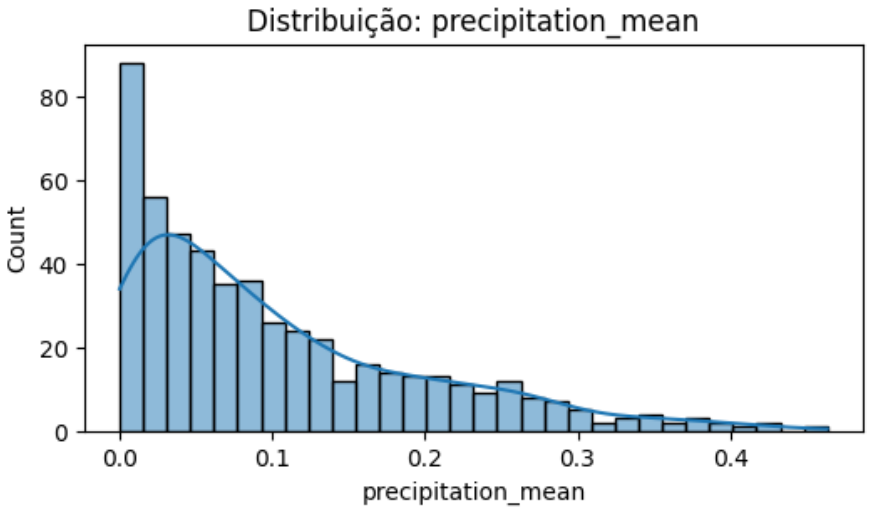

Figura 3.Distribuição da variável precipitação média

A distribuição da temperatura média, por sua vez, representada na Figura 4, é aproximadamente simétrica e concentra-se em torno de 26 °C a 28 °C, com comportamento semelhante ao de uma distribuição normal, coerente com o clima tropical da região.

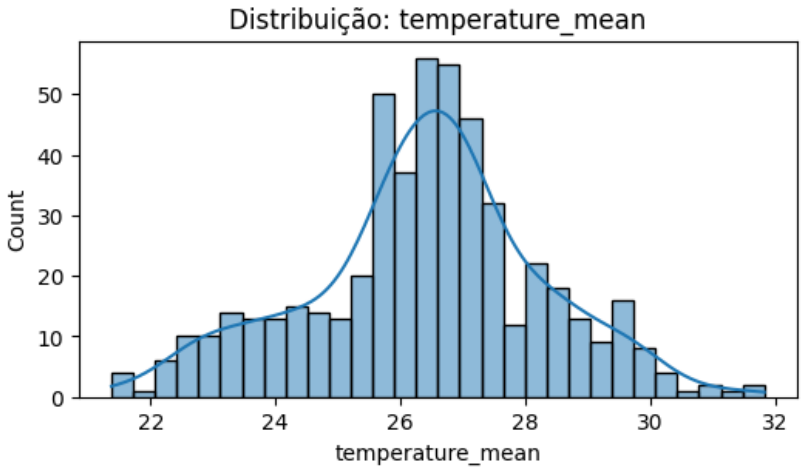

Figura 4.Distribuição da temperatura média

Na Figura 5, observa-se a distribuição da umidade relativa média, com leve assimetria à esquerda. A maior parte dos registros se encontra na faixa entre 60% e 75%, indicando um regime de umidade moderada, porém variável ao longo do ano.

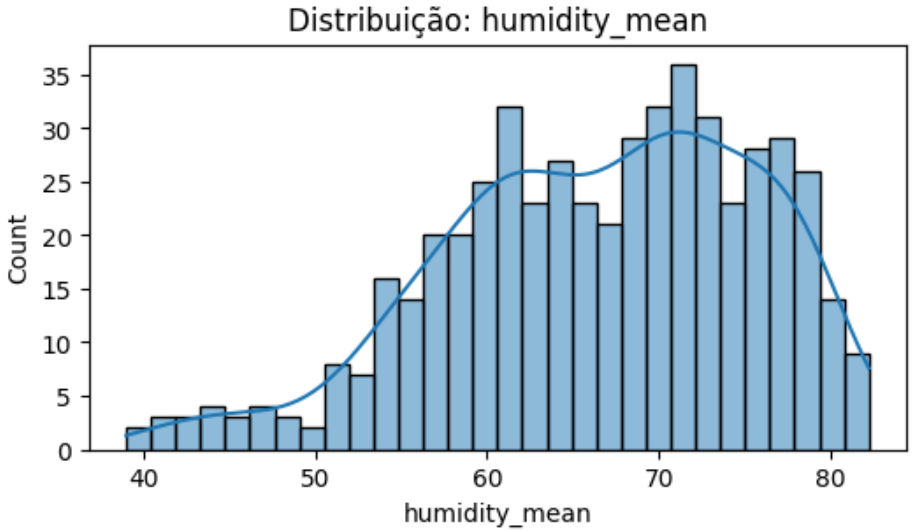

Figura 5.Distribuição da umidade relativa média

A radiação solar média, representada na Figura 6, apresenta distribuição mais simétrica e centrada entre 1300 e 1700 KJ/m², reforçando a alta e constante disponibilidade solar da região Nordeste. A distribuição da velocidade média do vento (Figura 7) mostra assimetria à direita, com maior concentração de observações entre 1,5 m/s e 3,0 m/s — intervalo considerado viável para geração de energia eólica em parques instalados em áreas elevadas ou litorâneas.

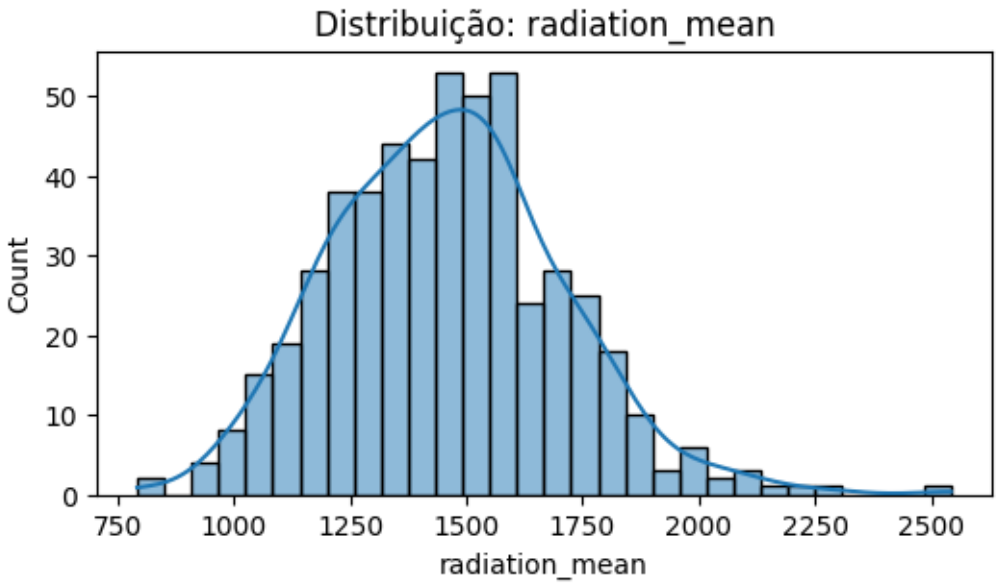

Figura 6.Radiação solar média

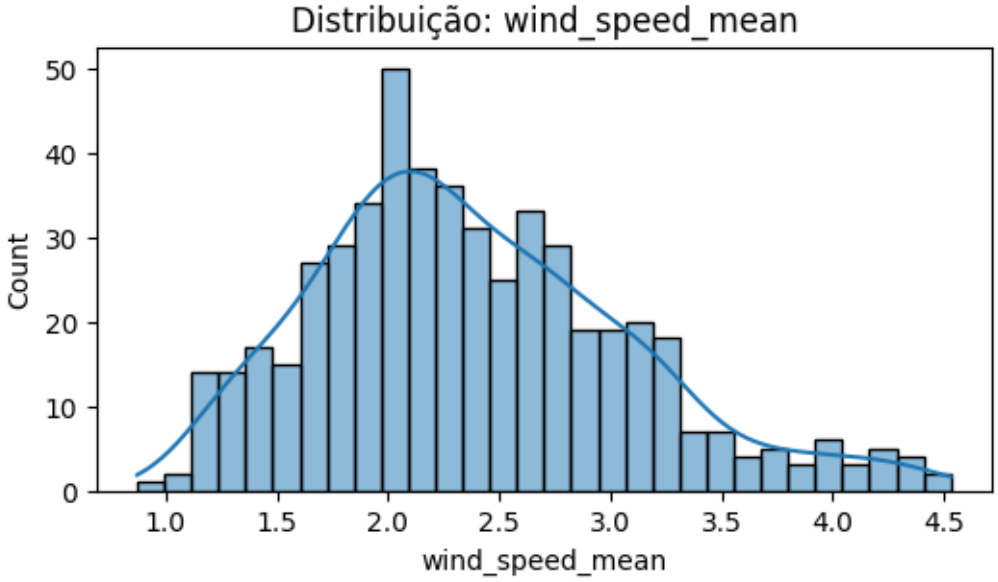

Figura 7.Distribuição da velocidade média do vento

Essas análises iniciais são fundamentais para orientar a seleção de variáveis e a escolha dos métodos de pré-processamento e modelagem, além de fornecer insights sobre a variabilidade espacial e temporal dos fenômenos envolvidos.

2.4. Redução de Dimensionalidade e Seleção de Variáveis

Com os dados climáticos devidamente tratados e compatibilizados temporalmente com os dados de geração de energia, foram criados dois conjuntos distintos para análise:

df_solar_filtrado (444 linhas) e df_eolica_filtrado (294 linhas), correspondentes, respectivamente, aos recortes temporais a partir de novembro de 2018 (solar) e julho de 2021 (eólica). A partir desses dataframes, iniciaram-se as etapas de pré-processamento avançado e seleção de variáveis.

2.4.1. Análise de Correlação

A compreensão das relações entre variáveis climáticas e a geração de energia é um passo essencial para a construção de modelos preditivos robustos. Para isso, foi conduzida uma análise de correlação utilizando o coeficiente de Spearman, por ser uma métrica não paramétrica capaz de capturar relações monotônicas, mesmo em contextos onde a linearidade não está presente. A seguir, são discutidos os resultados encontrados separadamente para os conjuntos de dados relacionados à geração solar e eólica.

No caso da geração de energia solar, observam-se correlações positivas expressivas com variáveis relacionadas ao tempo e à estrutura da base de dados. A variável year, por exemplo, apresentou uma correlação de 0,628 com a variável de geração, indicando uma tendência de crescimento da capacidade instalada e da produção ao longo dos anos. Variáveis como n_stations ($\rho = 0.589$) e total_records ($\rho = 0.590$) também exibiram correlação semelhante, refletindo que a expansão e o aumento da densidade de estações meteorológicas por estado estão associadas ao maior registro de geração solar — o que pode refletir, em parte, o aumento da cobertura de dados ao longo do tempo.

No que diz respeito às variáveis meteorológicas em si, algumas características esperadas foram confirmadas. A radiação global, por exemplo, embora tenha apresentado uma correlação modesta com a geração média (radiation_mean = –0.188), demonstrou correlação positiva com a variabilidade e com os extremos, como no caso de radiation_std ($\rho = 0.237$) e radiation_max ($\rho = 0.213$). Esses resultados indicam que, embora o valor médio de radiação possa não capturar todas as nuances da geração, a presença de picos de irradiação pode estar associada a maiores níveis de produção de energia solar. Por outro lado, a variável wind_speed_mean apresentou uma correlação negativa (-0.249), assim como wind_speed_min (-0.356), possivelmente porque dias com maior incidência de vento coincidem com condições de maior nebulosidade, o que reduz a radiação disponível.

Além disso, as variáveis de precipitação e umidade tendem a mostrar correlações negativas com a geração solar, como precipitation_min ($\rho = -0.221$) e humidity_min ($\rho = -0.177$), o que está em consonância com o entendimento físico de que ambientes mais úmidos e chuvosos reduzem a eficiência da captação solar. A temperatura máxima (temperature_max) teve uma correlação leve, mas positiva ($\rho = 0.229$), o que pode indicar que dias mais quentes (e, portanto, mais ensolarados) estão associados a maior geração, embora temperaturas muito elevadas possam reduzir a eficiência de conversão dos painéis.

Para a geração de energia eólica, o comportamento das correlações foi bastante distinto. Como esperado, as variáveis relacionadas à intensidade dos ventos foram as mais relevantes. A velocidade máxima do vento (wind_speed_max) apresentou a maior correlação positiva com a geração ($\rho = 0.574$), seguida por wind_speed_mean ($\rho = 0.483$), evidenciando a importância direta da força do vento no desempenho energético das turbinas. Outras estatísticas descritivas, como o desvio padrão e a velocidade mínima, também mostraram correlação positiva, embora mais moderada.

Em contrapartida, as variáveis relacionadas à direção do vento apresentaram correlações negativas significativas, como no caso de wind_direction_std ($\rho = -0.411$) e wind_direction_mean ($\rho = -0.357$). Esses resultados sugerem que a variabilidade direcional dos ventos pode comprometer a eficiência das turbinas, especialmente em casos em que as direções predominantes são instáveis e exigem constantes ajustes nos ângulos das pás. A umidade relativa também apareceu como um fator inversamente correlacionado, com destaque para humidity_min ($\rho = -0.512$), o que pode estar relacionado à influência da densidade do ar ou às condições atmosféricas que acompanham diferentes regimes de vento.

Por fim, a variável pressure_max apresentou uma correlação de 0.390 com a geração eólica, sendo uma das únicas variáveis de pressão com impacto positivo significativo. As demais variáveis climáticas apresentaram correlações mais fracas, o que é esperado, dado o protagonismo dos ventos no processo de conversão energética em parques eólicos.

Esses achados forneceram indícios importantes para as etapas subsequentes de redução de dimensionalidade e seleção de variáveis, possibilitando maior foco nas

características mais diretamente relacionadas à geração de energia. Além disso, confirmam tendências já descritas na literatura, como a relação inversa entre precipitação e geração solar, e a importância da constância direcional dos ventos para a performance eólica.

2.4.2. Detecção de Outliers

Após a análise de correlação, uma etapa essencial no pré-processamento dos dados foi a identificação e remoção de outliers. A presença de observações discrepantes pode comprometer a performance e a estabilidade de modelos de aprendizado de máquina, especialmente naqueles mais sensíveis a valores extremos. Para este estudo, optou-se pelo uso do método Elliptic Envelope, uma abordagem robusta baseada na suposição de distribuição gaussiana multivariada, que estima a fronteira elíptica que melhor engloba o conjunto de dados.

A detecção foi aplicada separadamente para os conjuntos de dados solar e eólico. No caso da geração solar, o método identificou 45 observações como outliers. Com a remoção dessas amostras, o conjunto final de dados passou a conter 399 registros distribuídos ao longo de 38 variáveis. Para o conjunto de dados da geração eólica, foram identificados 30 outliers, resultando em um conjunto limpo com 264 observações, também com 38 variáveis.

A remoção dos outliers foi adotada não apenas com o intuito de melhorar a acurácia dos modelos preditivos, mas também para garantir maior consistência estatística nas etapas posteriores de redução de dimensionalidade e seleção de atributos. Essa etapa mostrou-se particularmente importante dada a diversidade de condições meteorológicas observadas entre os estados e ao longo dos anos, o que naturalmente aumenta a heterogeneidade dos dados. A eliminação das amostras extremas permitiu preservar a estrutura geral dos dados sem comprometer a representatividade das variáveis mais relevantes.

2.4.3. Análise de Redução de Dimensionalidade e Seleção de Variáveis

A etapa de redução de dimensionalidade teve como principal objetivo mitigar a alta colinearidade entre as variáveis meteorológicas e reduzir a complexidade dos modelos preditivos, preservando o máximo de informação possível. Para isso, foram

aplicadas duas abordagens complementares: a Análise de Componentes Principais (PCA) e a seleção de variáveis por regressão LASSO, ambas avaliadas separadamente para os conjuntos de dados solar e eólico.

A aplicação do PCA permitiu identificar os componentes principais que explicam a maior parte da variância dos dados climáticos. No conjunto solar, os resultados mostraram que os 9 primeiros componentes explicam 84,5% da variância, enquanto 12 componentes já são suficientes para atingir mais de 90% da variabilidade total. Essa alta variância acumulada com poucos componentes evidencia forte redundância entre as variáveis, o que justifica o uso da técnica para simplificação dos dados. A Tabela abaixo apresenta a variância explicada por componente, dos 20 primeiros componentes:

component	eigenvalue	% of variance	% of variance (cumulative)
0	12.463	31.96%	31.96%
1	7.185	18.42%	50.38%
2	3.431	8.80%	59.18%
3	2.249	5.77%	64.94%
4	2.088	5.35%	70.30%
5	1.692	4.34%	74.63%
6	1.563	4.01%	78.64%
7	1.306	3.35%	81.99%
8	0.984	2.52%	84.51%
9	0.924	2.37%	86.88%
10	0.751	1.92%	88.81%
11	0.682	1.75%	90.56%
12	0.526	1.35%	91.91%
13	0.454	1.16%	93.07%
14	0.373	0.96%	94.03%
15	0.337	0.86%	94.89%
16	0.297	0.76%	95.65%
17	0.279	0.71%	96.36%
18	0.219	0.56%	96.93%
19	0.195	0.50%	97.43%
20	0.160	0.41%	97.84%

Figura 8. Variância acumulada dos 20 primeiros componentes - Energia Solar

No conjunto eólico, vide figura 9, o comportamento foi semelhante. Os 12 primeiros componentes também explicaram mais de 90% da variância, com destaque para os 9 primeiros que juntos explicaram quase 85% do total. Isso reforça a viabilidade do PCA como forma de condensar as informações sem grande perda de conteúdo estatístico. Essa transformação também permitiu observar, por meio da matriz de correlações dos componentes com as variáveis originais, a contribuição relativa de diferentes variáveis (como precipitação média, velocidade do vento, temperatura mínima e radiação solar) na formação dos componentes mais relevantes, oferecendo base para interpretar tendências gerais nos dados meteorológicos.

component	eigenvalue	% of variance	% of variance (cumulative)
0	12.164	31.19%	31.19%
1	7.476	19.17%	50.36%
2	3.550	9.10%	59.46%
3	2.547	6.53%	65.99%
4	2.043	5.24%	71.23%
5	1.769	4.53%	75.77%
6	1.330	3.41%	79.18%
7	1.216	3.12%	82.30%
8	1.005	2.58%	84.87%
9	0.887	2.28%	87.15%
10	0.710	1.82%	88.97%
11	0.683	1.75%	90.72%
12	0.552	1.42%	92.13%
13	0.492	1.26%	93.40%
14	0.404	1.04%	94.43%
15	0.369	0.95%	95.38%
16	0.272	0.70%	96.07%
17	0.229	0.59%	96.66%
18	0.196	0.50%	97.16%
19	0.187	0.48%	97.64%
20	0.151	0.39%	98.03%

Figura 9.Variância acumulada dos 20 primeiros componentes - Energia Eólica

A técnica LASSO foi aplicada com diferentes valores do parâmetro de regularização alpha, variando de 0.001 a 1.0, com o objetivo de avaliar a robustez das variáveis selecionadas sob diferentes níveis de penalização. Essa abordagem busca identificar o subconjunto de variáveis com maior poder explicativo sobre a geração de energia, forçando coeficientes irrelevantes a zero.

No conjunto solar, os seguintes resultados foram obtidos:

- Com $\alpha = 0.001$, todas as variáveis foram selecionadas.
- Aumentando para $\alpha = 0.1$, o conjunto reduziu para 34 variáveis.
- Para $\alpha = 1.0$, o modelo selecionou 24 variáveis.

No geral, variáveis como year, month, n_stations, precipitation_min, pressure_max, temperature_min, humidity_std, wind_speed_mean e radiation_max mantiveram-se recorrentes entre diferentes alphas, indicando relevância consistente.

De maneira análoga, no conjunto eólico, observou-se uma estabilidade ainda maior na seleção. Por exemplo:

- Com $\alpha = 0.001$, foram selecionadas todas as variáveis.
- Com $\alpha = 0.1$, o mesmo número foi mantido.
- Apenas com $\alpha = 1.0$ houve uma leve redução para 34 variáveis.

Entre as variáveis mais frequentes na seleção destacam-se wind_speed_max, humidity_min, precipitation_mean, pressure_std e radiation_std, corroborando os resultados obtidos anteriormente na análise de correlação de Spearman.

A utilização de múltiplos valores de penalização no LASSO foi fundamental para testar a robustez das variáveis sob diferentes condições de parcimônia. Essa abordagem permitiu observar como algumas variáveis são eliminadas precocemente (indicando baixa contribuição marginal), enquanto outras persistem mesmo sob alta regularização.

2.4.4. Consolidação dos Conjuntos de Dados para Modelagem

Concluídas as etapas de pré-processamento, análise exploratória, detecção de outliers, redução de dimensionalidade e seleção de variáveis, foi possível estruturar os

conjuntos de dados finais que servirão de base para o desenvolvimento e avaliação dos modelos de previsão de geração energética. Para garantir uma abordagem metodológica robusta e comparável, foram considerados dois critérios distintos para formação dos datasets: componentes principais (PCA) e variáveis selecionadas via LASSO. Adicionalmente, cada critério foi aplicado separadamente para os dados de geração solar e geração eólica.

Como resultado, foram definidos quatro dataframes distintos:

- df_solar_pca – Base de dados para geração solar composta pelos componentes principais que explicam aproximadamente 92% da variância dos dados climáticos.
- df_solar_lasso – Base de dados para geração solar composta apenas pelas variáveis climáticas selecionadas via regressão LASSO com base na relevância preditiva.
- df_eolica_pca – Base de dados para geração eólica reduzida por meio de PCA, preservando uma alta proporção da variância original dos dados.
- df_eolica_lasso – Base de dados para geração eólica contendo as variáveis mais relevantes segundo os critérios do LASSO, considerando diferentes níveis de regularização.

A partir dessa estrutura, a próxima etapa do trabalho consistirá na aplicação e comparação de algoritmos de aprendizado de máquina para prever a geração energética a partir das variáveis climáticas, avaliando o desempenho de cada abordagem em função da estrutura dos dados utilizados. Esta multiplicidade de bases permite testar não apenas diferentes modelos, mas também analisar o impacto direto da seleção ou transformação de variáveis na capacidade preditiva dos algoritmos.

3. RESULTADOS E DISCUSSÕES

Em breve

4. REFERÊNCIAS

[1] T. Carneiro, P. A. Rocha, P. Carvalho, and L. M. Fernández-Ramírez, “Ridge regression ensemble of machine learning models applied to solar and wind forecasting in Brazil and Spain,” Applied Energy, vol. 309, 2022. DOI: 10.1016/j.apenergy.2022.118936.

[2] M. Paula, C. Marilaine, J. N. Fidalgo, and W. Casaca, “Predicting long-term wind speed in wind farms of northeast Brazil: A comparative analysis through machine learning models,” IEEE Latin America Transactions, vol. 18, no. 12, pp. 2011–2018, 2020. DOI: 10.1109/TLA.2020.9398643.

[3] D. Lima, L. Deon, and F. Lima, “Ensemble learning models for wind power forecasting: A case study about Brazil,” SSRN Electronic Journal, 2024. DOI: 10.2139/ssrn.4709832.

[4] M. AlShafeey and C. Csáki, “Evaluating neural network and linear regression photovoltaic power forecasting models based on different input methods,” Energy Reports, 2021. DOI: 10.1016/j.egyr.2021.10.125.

[5] Jebli et al., “Prediction of solar energy guided by Pearson correlation using machine learning,” Energy, 2021. DOI: 10.1016/j.energy.2021.120109.

[6] J. Fan et al., “Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China,” Energy Conversion and Management, vol. 171, pp. 1376–1391, 2018. DOI: 10.1016/j.enconman.2018.02.087.

[7] G. F. Viscondi and S. N. Alves-Souza, “Solar irradiance prediction with machine learning algorithms: A Brazilian case study on photovoltaic electricity generation,” *Energies*, vol. 14, no. 18, 2021. DOI: 10.3390/en14185657.

ANEXO 1 – Repositório

Google Colab

https://colab.research.google.com/drive/1YxLbjOgLg1EUbCW7Yty8Rd9hsPiJmpHx#scrollTo=zj6uZj0LtDDA