UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA DE ENERGIAS RENOVÁVEIS

VICTOR EMANUEL MARTINS DA SILVA

INFLUÊNCIA DO TRATAMENTO DE DADOS NO DESEMPENHO DE MODELOS DE PREVISÃO DA GERAÇÃO EÓLICA USANDO INFORMAÇÕES DE RESTRIÇÃO DE POTÊNCIA PELO OPERADOR NACIONAL DO SISTEMA ELÉTRICO

FORTALEZA-CE
2023

VICTOR EMANUEL MARTINS DA SILVA

INFLUÊNCIA DO TRATAMENTO DE DADOS NO DESEMPENHO DE MODELOS DE PREVISÃO DA GERAÇÃO EÓLICA USANDO INFORMAÇÕES DE RESTRIÇÃO DE POTÊNCIA PELO OPERADOR NACIONAL DO SISTEMA ELÉTRICO

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação em Engenharia de Energias Renováveis da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Engenheiro de Energia Renovável. Área de concentração: Energias Renováveis e Sistemas de Energia Elétrica.

Orientadora: Profª. Ph.D. Ruth Pastôra Saraiva Leão.
Coorientadora: Profª. Dra. Raquel Cristina Filiagi Gregory.

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S584i Silva, Victor Emanuel Martins da.
Influência do tratamento de dados no desempenho de modelos de previsão da geração eólica usando informações de restrição de potência pelo Operador Nacional do Sistema Elétrico / Victor Emanuel Martins da Silva. – 2023.
43 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia de Energias Renováveis, Fortaleza, 2023.
Orientação: Profa. Dra. Ruth Pastôra Saraiva Leão.
Coorientação: Profa. Dra. Raquel Cristina Filiagi Gregory.

1. Previsão da geração eólica. 2. Pré-Processamento. 3. Aprendizado de máquina. 4. Energia renovável. 5. Restrição operativa. I. Título.

CDD 621.042

VICTOR EMANUEL MARTINS DA SILVA

INFLUÊNCIA DO TRATAMENTO DE DADOS NO DESEMPENHO DE MODELOS DE PREVISÃO DA GERAÇÃO EÓLICA USANDO INFORMAÇÕES DE RESTRIÇÃO DE POTÊNCIA PELO OPERADOR NACIONAL DO SISTEMA ELÉTRICO

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação em Engenharia de Energias Renováveis da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Engenheiro de Energia Renovável. Área de concentração: Energias Renováveis e Sistemas de Energia Elétrica.

Orientador: Profª. Ph.D. Ruth Pastôra Saraiva Leão.
Coorientador: Profª. Dra. Raquel Cristina Filiagi Gregory

Aprovada em: 18/12/2023.

BANCA EXAMINADORA

_____
Profª. Ph.D. Ruth Pastôra Saraiva Leão (Orientadora)
Universidade Federal do Ceará (UFC)

_____
Profª. Dra. Raquel Cristina Filiagi Gregory (Coorientadora)
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Erick Costa Bezerra
Sidia Instituto de Ciência e Tecnologia

_____
Me. André Wagner de Barros Silva
Universidade Federal do Cariri (UFCA)

A Deus.

A minha família.

AGRADECIMENTOS

A minha orientadora, Profª. Ph.D. Ruth Pastôra Saraiva Leão e a minha coorientadora Profª. Raquel Cristina Filiagi Gregory, pelos debates e orientação dada durante os acompanhamentos desse trabalho. Aos meus pais, Wanessa Cunha e Francisco Martins, a minha tia Maria Waleska e à minha parceira Brena Késia, por estarem sempre me acompanhando de perto e me dando suporte em tudo o que preciso. Aos meus amigos Lucas Cordeiro, João Victor e João Rodrigues, por todo o apoio dado ao longo do meu trabalho. À banca examinadora, pelas contribuições dadas a este trabalho, bem como pela disponibilidade para avaliá-lo.

“Essencialmente, todos os modelos estão
errados, mas alguns são úteis.”
(George Box).

RESUMO

A transição para fontes de energia renovável destaca a necessidade de previsões precisas, especialmente para a geração eólica, em que sua natureza intermitente desafia a confiabilidade do sistema elétrico. Em resposta a esse fator, a restrição da produção da energia renovável, aplicada pelo Operador Nacional do Sistema Elétrico (ONS), tem crescido, e se insere como mais uma variável da geração eólica. Métodos de previsão acurados podem contribuir para redução dos custos de operação e aumento da confiabilidade do sistema elétrico. Para isso, diversas etapas estão envolvidas tais como pré-processamento, modelagem preditiva, validação cruzada e análise de resultados. Neste trabalho foi investigado o efeito desse conjunto de práticas na previsão da geração eólica despachada com ênfase no pré-processamento, considerando cenários de restrição com base nos dados extraídos do ONS para o complexo eólico "SERRA DE SANTANA 3", no Rio Grande do Norte, de 108 MWp de capacidade instalada. Após a normalização pela capacidade instalada, foram aplicadas técnicas de aprendizado supervisionado e os algoritmos *Least Absolute Shrinkage Operator* (LASSO), Ridge, *Light Gradient Boosting Machine* (LGBM) e *Support Vector Regressor* (SVR) com otimização bayesiana para previsão de 1 hora. Para treinar esses algoritmos foram aplicadas diferentes técnicas de modelagem de dados incluindo a programação da restrição como variável externa. Para fins de comparação de desempenho, foram usados os modelos de referência usando apenas os dados de geração de energia e também o método de persistência. Com base nos resultados, pode-se concluir que a restrição de geração por ser uma propriedade esparsa teve pouca importância para os modelos frente aos demais atributos, e o método SVR, usando as técnicas de pré-processamento, teve o maior destaque em termos de ganho sobre os modelos de referências. A análise do desempenho dos métodos foi baseada nas métricas *Normalized Mean Absolute Error* (NMAE), *Normalized Root Mean Squared Error* (NRSME) e R2 em termos percentuais, resultando no SVR com o pré-processamento proposto como o melhor método com os índices NMAE 5,30%, e 7,42% e R2 90,70%. A principal contribuição do trabalho foi a avaliação da previsibilidade da geração futura utilizando diferentes métodos de aprendizado de máquina sob diferentes contextos e considerando o aumento das restrições pelo ONS.

Palavras-chave: previsão da geração eólica; pré-processamento; aprendizado de máquina; energia renovável; restrição operativa.

ABSTRACT

The transition to renewable energy sources highlights the need for accurate predictions, especially for wind power generation, where its intermittent nature challenges the reliability of the electrical system. In response to this factor, the restriction of renewable energy production, applied by the National Electric System Operator (ONS), has grown and is introduced as another variable in wind power generation. Accurate forecasting methods can contribute to reduce operating costs and increase the reliability of the electrical system. For this, various steps are involved, such as pre-processing, predictive modeling, cross-validation, and results analysis. In this study, the effect of this set of practices on wind power forecasting was investigated, with emphasis on pre-processing, considering restriction scenarios based on data extracted from the ONS for the "SERRA DE SANTANA 3" wind complex in Rio Grande do Norte, with an installed capacity of 108 MWp. After normalization by installed capacity, supervised learning techniques and algorithms such as Least Absolute Shrinkage Operator (LASSO), Ridge, Light Gradient Boosting Machine (LGBM), and Support Vector Regressor (SVR) with Bayesian optimization were applied for 1-hour ahead prediction. Different data modeling techniques, including wind power curtailment as an external variable, were applied to train these algorithms. For performance comparison, reference models using only energy generation data and the persistence method were used. Based on the results, it can be concluded that the generation restriction, being a sparse property, had little importance for the models compared to other attributes. The SVR method, using pre-processing techniques, stood out the most in terms of improvement over reference models. The performance analysis of the methods was based on the metrics Normalized Mean Absolute Error (NMAE), NRSME, and R2 in percentage terms, resulting in SVR with the proposed pre-processing as the best method with NMAE of 5.30%, NRMSE of 7.42%, and R2 of 90.70%. The main contribution of the study was the evaluation of the wind power forecasting using different machine learning methods under different contexts and considering the growth of wind power curtailment imposed by the ONS.

Keywords: Wind power forecasting; Regression; Data preprocessing; Machine learning; Renewable energy; Wind power curtailment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
CE	Complexo Eólico
DIQ	Distância Interquartil
EFB	Agrupamento de Recursos Exclusivos (*Exclusive Feature Bundling*)
GAM	Modelo Aditivo Generalizado (*Generalized Additive Model*)
GBM	Máquina de Gradiente de Impulso (*Grandient Boosting Machine*)
GEFCom	Competição global de previsão da energia (*Global Energy Forecasting Competition*)
GOSS	Amostragem unilateral de gradiente (*Gradient-Based One Side Sampling*)
LASSO	Operador de redução e seleção mínima absoluta (*Least Absolute Shrinkage and Selection Operator)*
LGBM	Máquina de Gradiente de Impulso Leve (*Light Grandient Boosting Machine*)
MAE	Erro Médio Absoluto (*Mean Absolute Error)*
MDP	Método Desvio Padrão
MLPR	Regressor Perceptron de Multicamada (*Multilayer Perceptron Regressor)*
NMAE	Erro Médio Absoluto Normalizado (*Normalized Mean absolute Error*)
NREL	Laboratório Nacional de Energia Renovável (*National Renewable Energy Laboratory*)
NRMSE	Raíz Quadrada do Erro Médio Quadrático Normalizada (*Normalized Root Mean Squared Error*)
NWP	Previsão Numérica do Clima (*Numerical Weather Prediction*)
ONS	Operador Nacional do Sistema Elétrico
PCA	Análise de Componentes Principais (*Principal Component Analysis)*
PM	Método de Persistência (*Persistence Method*)
PROEÓLICA	Programa Emergencial de Energia Eólica
PROINFA	Programa de Incentivo às Fontes Alternativas de Energia Elétrica
R2	Índice de Determinação
SIN	Sistema Integrado Nacional
SVM	Máquinas de vetores de suporte (*Support Vector Machines)*
SVR	Regressor de vetores de suporte (*Support Vector Regressor*)

LISTA DE SÍMBOLOS

%	Porcentagem
Δ	Delta
$\Sigma_{t=1}^{N}$	Somatório do primeiro ao enésimo termo
$\widehat{y}$	Valor estimado pela regressão
y	Valor observado

SUMÁRIO

1 INTRODUÇÃO

Tendo em vista o avanço da participação da geração de energia eólica na matriz elétrica brasileira, formas de tornar essa geração mais previsível ganham destaque dada a sua natureza intermitente, o que representa um desafio para o setor elétrico (Kisvari, Lin, Liu, 2020). A incerteza causada pela intermitência do vento pode ser abordada por meio de técnicas de previsão, que permitem reduzir o erro sobre a geração futura das unidades geradoras e, assim, permitir que o operador consiga manter o sistema equilibrado em escala.

Apesar dos avanços na previsão da geração renovável variável, não raro os operadores do sistema elétrico precisam fazer ajustes e reduzir o despacho da geração eólica a fim de preservar a estabilidade da rede (Kisvari, Lin, Liu, 2020). As restrições da energia eólica causam prejuízos econômicos, financeiros e afetam o fator de capacidade dos empreendimentos envolvidos.

A redução da produção de energia das usinas eólicas pode ser motivada por diferentes razões, como indisponibilidade em instalações externas às usinas eólicas, confiabilidade elétrica dos equipamentos pertencentes às instalações externas às usinas eólicas ou por razão energética.

Com vistas a fomentar a pesquisa na área de geração renovável, o Operador Nacional do Sistema Elétrico (ONS) no Brasil torna público os dados, em resolução horária, de potência gerada e as restrições em usinas eólicas conectadas à rede básica. O livre acesso à base de dados do ONS torna possível investigar sobre modelos de previsão eólica em usinas que sofreram restrição operativa.

Metodologias de aprendizado de máquina operam como uma função de transformação, convertendo vários parâmetros de entrada em uma saída, por meio de um modelo treinado com dados pré-processados, para que possam treinar o algoritmo e por fim ser empregado. Assim como o vento, que é o insumo para a geração eólica pelas turbinas, os dados são os insumos para o modelo de aprendizado supervisionado e o enriquecimento dos dados pode impactar no desempenho final. A vista disso, são empregadas técnicas de remoção de erros, anomalias, ruído, seleção de atributos e redução de dimensionalidade (Zhou *et al*, 2023).

Após o pré-processamento, a modelagem pode ser feita por métodos mais tradicionais, como regressão linear e suas variações com o termo de penalização, que são muitas vezes preferíveis pela sua interpretabilidade, permitindo o mapeamento direto entre os parâmetros de entrada e seu resultado. Por outro lado, métodos chamados de "caixa-preta" ganham destaque quanto às interações não lineares, visto que são relevantes para o problema de previsão. Diante disso, alguns algoritmos como as redes neurais, as SVM (*Support Vector Machines*) e os métodos *ensemble* como o LGBM (*Light Gradient Boosting Machine*), podem trazer melhorias nos resultados de previsão e, consequentemente, nas decisões e operações baseadas nessas previsões (Liao *et al*, 2023).

No presente estudo, é analisada a aplicação de técnicas de pré-processamento, sendo considerados para realização dos testes os algoritmos: SVM, LGBM, *Least Absolute Shrinkage Operator* (LASSO) e Ridge para avaliação da metodologia proposta acerca da problemática de previsão de geração eólica.

1.1 Motivação

Com o avanço das energias renováveis na matriz elétrica brasileira e o aumento das restrições sobre a capacidade de geração de usinas eólicas, a previsão de geração de complexos eólicos ganha ainda mais relevância. Por outro lado, a crescente limitação na geração eólica pode ser analisada como uma variável extra do sistema e, consequentemente, podendo afetar a precisão de técnicas tradicionais de previsão quando não considerada. Na presente pesquisa, são investigadas metodologias de previsão de geração eólica de curto prazo com uma janela de previsão de 1 hora à frente, usando aprendizado de máquina e técnicas de pré-processamento de dados, considerando a programação de restrição.

1.2 Objetivos

O presente trabalho tem como objetivo geral avaliar as metodologias de pré-processamento para aprendizado de máquina considerando cenários com *constrained-off* para energia eólica 1h à frente e destacar a abordagem com melhor performance. E quanto aos objetivo específicos:

- Destacar as práticas de enriquecimento dos dados para treinamento dos

modelos;

- Propor uma abordagem robusta e simplificada para previsão da geração eólica;
- Avaliar performance dos algoritmos clássicos (LASSO, Ridge, LGBM e SVR) em diferentes cenários de restrição pelo ONS com a implementação da técnica de pré-processamento.

1.3 Organização do Trabalho

Além do capítulo introdutório, este trabalho está organizado em cinco capítulos adicionais: O capítulo 2 contém a revisão bibliográfica, o capítulo 3 resume as metodologias presentes na pesquisa. O capítulo 4 contempla a metodologia aplicada na pesquisa, podendo-se ver nas seções: 4.1) análise descritiva do complexo eólico, 4.2) metodologias de pré-processamento utilizadas e 4.3) otimização de hiperparâmetros. No capítulo 5 são apresentados os resultados da metodologia aplicada e sua discussão e, finalmente, o capítulo 6 traz a conclusão do trabalho.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo tem o objetivo de abordar artigos, pesquisas e trabalhos acadêmicos de maior relevância utilizados para o desenvolvimento deste estudo. A princípio, é feita uma contextualização para o cenário de geração eólica no Brasil dando ênfase ao entendimento do *constrained-off* na geração eólica. Ao final do capítulo, é discutido o papel da modelagem preditiva para geração eólica, destacando aspectos de pré-processamento.

2.1 Restrições Operacionais em Energia Renovável

A crise energética de 2001 provocou uma mudança significativa no cenário do setor energético brasileiro. A escassez de água impactou negativamente a geração de energia hidrelétrica, levando o setor elétrico a buscar alternativas sustentáveis para a diversificação da matriz energética, destacando o potencial de fontes renováveis. Nesse contexto, surgiu o PROEÓLICA (Programa Emergencial de Energia Eólica) e sua expansão para o PROINFA (Programa de Incentivo às Fontes Alternativas de Energia Elétrica) como forma de diversificar a matriz elétrica nacional, até então, fortemente dependente das hidrelétricas (ABEEÓLICA, 2022).

Esse incentivo logo foi acompanhado pelo crescimento de novas fontes de energia, com destaque para a fonte eólica. No entanto, as características do SIN (Sistema Interligado Nacional), combinadas a presença significativa de energia intermitente, demandam medidas de restrição de operação por *constrained-off* pelos motivos demonstrados na Tabela 1, resultando na redução da produção pelos geradores. Esse tipo de procedimento é consolidado para usinas eolioelétricas pela Resolução 1030 da ANEEL (Agência Nacional de Energia Elétrica) (2022a, p.10), como:

> "Art. 13. Para efeitos deste Título, eventos de restrição de operação por Constrained-off são definidos como a redução da produção de energia por usinas eolioelétricas despachadas centralizadamente ou usinas/conjuntos de usinas eolioelétricas considerados na programação, decorrente de comando do ONS, que tenham sido originados externamente às instalações das respectivas usinas."

Nessa perspectiva, o ONS realiza o controle em tempo real da geração de energia elétrica visando otimizar e garantir o fornecimento de eletricidade. Com o objetivo de garantir a confiabilidade do SIN e evitar sobrecargas, o ONS pode impor restrições de geração para usinas eólicas. Essa prática é essencial para manter a confiabilidade do sistema elétrico nacional e é registrada no sistema de dados juntamente com o motivo de cada intervenção de acordo com a tabela Tabela 1.

Tabela 1: Motivos dos eventos de restrição de operação por *Constrained-off*.

Razão de indisponibilidade externa	Motivados por indisponibilidades em instalações externas às respectivas usinas, sejam nas instalações de transmissão ou no âmbito de distribuição.
Razão de atendimento a requisitos de confiabilidade	Motivados por razões de confiabilidade elétrica dos equipamentos pertencentes às instalações externas e que não tenham origem em indisponibilidade dos respectivos equipamentos, ou seja, situações como o atingimento do limite de linhas de transmissão ou de requisitos de estabilidade dinâmica, são passíveis desta classificação.
Razão energética	Motivados pela impossibilidade de alocação de geração de energia na carga

Fonte: (ANEEL, 2022a, p.10)

2.2 Modelagem Preditiva para Energia Eólica

Os algoritmos de aprendizado de máquina são ferramentas importantes para a construção de modelos preditivos, sendo utilizados para reconhecer, extrair padrões, fazer associações e construir modelos de aprendizagem a partir da observação de grande volume de dados (Novaes, 2022).

A previsão de energia eólica envolve o pré-processamento de dados e modelagem. Enquanto a maioria dos estudos anteriores enfatiza a criação de diversos modelos de previsão, pouca ênfase foi dada a métodos inovadores de pré-processamento de dados, o que é contrastado por outras áreas mais comerciais como *marketing* e *e-commerce*. (Zhou *et al*, 2023).

Uma análise comparativa utilizando diversos modelos é apresentada em Liao *et al* (2023), para previsão da geração eólica em diferentes horizontes. O autor faz uso de dois conjuntos de dados univariados (contando com o histórico apenas de geração). No trabalho a metodologia proposta implementa uma abordagem simplificada, pelo uso de modelo aditivo generalizado (*Generalized Additive Model*) como proposta para previsão de curto prazo e supera, em perfomance, vários modelos de referência, obtendo resultados próximos ao de redes neurais, como LSTM (*Long Term Short Term Memory*) e MLP (*Multilayer Perceptron*), porém, o tempo de treino é otimizado. O modelo proposto consegue atingir RMSE de 8.8% e de 12.7% e NMAE de 5% e 8.3% para previsão de 1h em diferentes conjuntos de dados.

Além da performance do *machine learning* clássico ser próxima a de redes neurais na previsão de curto prazo e o tempo de treino ser significativamente maior, também é necessário realizar a otimização dos hiperparâmetros (estrutura da rede, funções de ativação, condições de parada e regularização) o que soma mais complexidade ao processo (Hong *et al*, 2020). Com isso, é possível destacar a modelagem de dados como forma menos custosa, computacionalmente, para avaliar melhorias de performance uma vez que clássicos *machine*

learning para previsão de carga, geração eólica e solar podem beneficiar da modelagem e inclusão de atributos físicos do processo de geração (Hong *et al*, 2020).

Smyl e Hua (2019), utilizaram várias técnicas de pré-processamento para enriquecer a análise univariada para previsão da série temporal de carga na GEFCom 2017, envolvendo criação de novos atributos, seleção de variáveis preditoras e minimização do ruído. Além disso, é dada ênfase que o processo foi realizado com dados extraídos da própria série temporal de estudo e que a inclusão de variáveis preditoras externas, poderiam trazer ganhos em relação a análise univariada.

Nessa perspectiva, vários aspectos podem ser considerados, como as intervenções de manutenção e a restrição operativa. Além disso, previsões de curta duração, abrangendo de minutos a poucas horas também agregam valor e eventos mais recentes se tornam mais relevantes que dados de *Numerical Weather Prediction* (NWP) (Petropoulos *et al*, 2022).

Ademais, séries temporais podem ser decompostas em sub-séries chamadas componentes. Cada componente pode representar características da série como frequência, tendência e sazonalidade que pode ser treinada e prevista de forma separada e por fim integrada em um resultado final baseado na ideia de "dividir e conquistar" (Ribeiro, 2021). Shen *et al* (2018) aplica para previsão da geração eólica pré-processamento por *Empirical Mode Decomposition* para estacionarizar a série temporal em componentes mais estacionárias e o algoritmo *Random Forest* para a previsão se seu valor 1h à frente conseguindo reduzir o NMAE de 11,23 para 7,86%.

Com isso, é possível destacar que técnicas de pré-processamento constituem uma etapa fundamental e que podem expandir os limites dos modelos de aprendizado de máquina em geral, tanto pelo incremento de performance do modelo, quanto pelo potencial de maior aproveitamento de atributos como a restrição operativa. Neste trabalho, é dada ênfase ao conjunto de técnicas de pré-processamento para a previsão da geração eolioelétrica, considerando a dinamicidade dos cenários de restrição operativa que envolvem o complexo eólico.

3 MÉTODOS APLICADOS EM APRENDIZADO SUPERVISIONADO

Na sequência é abordada a metodologia proposta de pré-processamento de dados que consiste na detecção de dados anômalos (*outliers*), transformação dos dados e redução de dimensionalidade.

3.1 Pré-processamento dos dados

Zhou *et al* (2023), sumariza as técnicas de pré-processamento aplicadas a previsão da geração eólica em cinco categorias: decomposição da série, seleção de atributos, extração de atributos, remoção de ruído e tratamento de anomalias.

A etapa de pré-processamento dos dados é aplicada por meio de transformações de dados sobre conjunto de dados, de modo a prepará-los para utilização nos algoritmos. Diante disso, todas as etapas são antecedentes a implementação dos algoritmos de previsão e são essenciais para o correto funcionamento dos mesmos. As seções a seguir detalham a transformação, tratamento de dados anômalos e redução de dimensionalidade.

3.1.1 Transformação

O enriquecimento dos dados pode ser obtido pelo processo de transformação, também conhecido como *Feature Engineering*. Nesse processo os dados brutos passam por transformações para destacar atributos que podem destacar o desempenho do modelo preditivo (Kisvari, Lin, Liu, 2020).

Primeiramente, o efeito de tendência das séries temporais precisa ser representado em processos não estacionários para evitar resultados espúrios (Redl *et al*, 2009). A transformação por diferenciação permite reduzir o efeito de tendência, para isso pode ser aplicada sobre a série temporal a diferença entre valores consecutivos (primeira derivada).

Além disso, outros processos podem ser aplicados a esse tipo de série, como a decomposição de sinais aplicada, que pode ser utilizado na previsão de rampas em complexos eólicos (Qiu *et al*, 2017). Abaixo é definida a fórmula da diferenciação:

$$\Delta Y = Y_t - Y_{t-1}$$

Em que:

- ΔY: Variação da variável alvo.
- Y_t: Valor da observação atual.

- Y_{t-1}: Valor da observação anterior.

Além disso, várias pesquisas sobre seleção de dados e processamento, combinam informações diferentes para melhorar a perfomance dos modelos preditivos, sendo temperatura, potência e velocidade do vento os atributos mais utilizados (Hanifi *et al*, 2020). Ainda sobre o enriquecimento dos dados, atributos estatísticos podem ser incluídos aos dados de treino para aprendizado supervisionado, como: valores passados (*lags*), estatísticas móveis como média, desvio padrão, soma e derivada (Ribeiro, 2021). Smyl e Hua (2019), buscavam enriquecer os dados de uma série univariada para previsão de demanda de carga na GEFCom 2017 a partir da criação de atributos baseados na própria série de forma direta e indireta.

3.1.2 Tratamento de anomalias

A segunda etapa de tratamento realizada trata-se da substituição de anomalias (*outliers*), que são dados que desviam extremamente do comportamento melhor definido dos demais presentes no conjunto de dados, podendo ou não ser útil removê-los ou mesmo ajustá-los. Nessa perspectiva, a maioria das referências categoriza anomalias como dados que se destacam de forma significativa dos demais (Zou *et al*, 2020).

Para mensurar esse destaque, métodos estatísticos podem ser aplicados por meio de indicadores estatísticos em janelas móveis, que consistem na determinação de limites que podem variar no tempo (Zou *et al*, 2020). Nessa perspectiva, é são destacados o método Desvio Padrão (MDP), que assume normalidade na distribuição dos dados, e o método baseado na Distância Interquartil (DIQ), ambos métodos podem ser aplicados em janelas móveis determinam limites superior e inferior dinamicamente (SEO, 2006).

Os dois métodos mencionados, foram aplicados à série temporal da variação do fator de capacidade e após isso foi aplicada a técnica de winsorização, que consiste no ajuste das anomalias pela redução do valor anômalo a um valor limite. Outrossim, tais limites podem ser determinados por parâmetros estatísticos, o que reduz o efeito dos outliers sobre o conjunto de dados e a perda de informação (Kumar *et al*, 2023).

3.1.3 Redução de Dimensionalidade

As técnicas de redução de dimensionalidade permitem a diminuição do número de variáveis consideradas para o treinamento do modelo para torná-lo mais eficiente, além de retirar variáveis que podem ter impacto negativo nas previsões.

3.1.3.1 Seleção de Atributos

A seleção de atributos é uma etapa essencial para aplicação de aprendizado de máquina, permitindo obter um equilíbrio entre número de dimensões e a acuracidade dos resultados. Nessa linha, com a diminuição da redundância, os resultados apresentam menos ruídos, menos tempo de processamento e mais interpretabilidade (Zhao *et al*, 2019).

Além disso, ao lidar com dados reais, é comum enfrentar o problema de *concept shit* (mudança de conceito), alterações nas relações entre a variável preditora e seu alvo ao longo do tempo em que a variável preditora pode assumir mais ou menos relevância nas previsões futuras, o que reflete em variações de performance do em relação ao tempo (Sebastián *et al*, 2023). Neste trabalho é dada ênfase ao método Boruta-Shap, o qual é um método eficiente para selecionar atributos em conjuntos de dados ruidosos, evitando o overfitting e contribuindo para melhoria da performance de *machine learning*.

O funcionamento do método é dado pela remoção de atributos classificados como irrelevantes, para avaliar isso, é feito um teste de hipóteses a partir dos seguintes resultados. O algoritmo cria parâmetros fantasmas a partir dos parâmetros originais, mas reordenados de forma aleatória, perdendo assim a relação dos dados originais. Com isso, o modelo é treinado, testado, e é obtida a importância dos atributos da base de dados original e dos atributos fantasma (Kursa *et al*, 2010). Por fim, os atributos que explicam a variação dos dados de forma mais efetiva do que o melhor dos atributos criados artificialmente para avaliação são marcados como relevantes a cada iteração. Ao final do experimento, é possível obter uma distribuição binomial para cada atributo e classificar sua relevância.

Como resultado, é possível interpretar a distribuição binomial gerada pelo algoritmo em 3 categorias com base no seu nível de significância (Mazzanti, 2020):

1. Área de recusa: atributos dentro dessa faixa são considerados ruídos e podem ser descartados;
2. Área de irresolução: atributos dentro dessa faixa podem ou não ser considerados, o algoritmo não os descarta;
3. Área de aceite: atributos dentro dessa faixa tem alto poder preditivo.

3.1.3.2 Extração de Atributos

A análise de componentes principais (PCA) é uma técnica de redução de dimensionalidade que pode ser aplicada em processamento de sinais e tratamento de fatores como multicolinearidade, ruído e a "Maldição da alta dimensionalidade" citada por Pirolla

(2012), por meio da condensação de vários atributos em um número mais reduzido e preservando a variância dos atributos originais por meio de aproximações lineares.

O método funciona por meio da transformação linear nos dados, projetando-os em um novo sistema de coordenadas, permitindo a projeção do das dimensões originais em um espaço dimensional reduzido. A primeira componente principal captura a maior variação nos dados, a segunda componente principal (ortogonal à primeira) captura a segunda maior variação, e assim por diante (Zha *et al*, 2022).

3.2 Algoritmos de Aprendizado Supervisionado

Nessa seção são discutidos os algoritmos utilizados para o aprendizado supervisionado usados no capítulo seguinte.

3.2.1 Regressão Linear com Penalização

Os métodos de regressão linear tornam a previsão da variável alvo fácil de interpretar, embora seu poder de previsão seja mais limitado por não considerar as relações não lineares entre o vetor de entrada e a variável alvo (Liao *et al*, 2023). Além disso, esse método é sensível a fatores como a colinearidade e a "Maldição da dimensionalidade" citada por Pirolla (2012), o que pode resultar em *overfitting*. Nesse contexto, o fator de penalização contribui para trazer robustez ao algoritmo. Abaixo, seguem as técnicas de regressão linear utilizadas.

3.2.1.1 Ridge

O algoritmo Ridge é uma técnica de regressão linear que incorpora um termo de penalização L2 à função de custo da regressão linear, emergindo como uma ferramenta eficaz para previsão em conjuntos de dados de alta dimensionalidade e colinearidade. Essa abordagem busca evitar o *overfitting*, controlando a magnitude dos coeficientes, otimizando a função de custo para minimizar a soma dos quadrados dos coeficientes penalizada pela norma L2. Além disso, a regressão Ridge se baseia na estabilização da variância do estimador, aplicando penalização quadrática (L2) no vetor de entrada e seus coeficientes penalizados são reduzidos tendendo a zero mas nunca são zerados (Masini *et al*, 2021).

3.2.1.2 Least Absolute Shrinkage and Selection Operator (LASSO)

Um dos pontos mais fortes da regularização, é a redução do ruído presente no vetor de entrada por meio da penalização. Nesse contexto, o algoritmo LASSO é uma das formas mais populares de regularização por ter penalização baseada na seleção de atributos pelo termo de regularização L1, e com isso pode ser aplicado em conjuntos de dados esparsos e de alta dimensionalidade (Masini *et al*, 2021). Ao fazer esse tipo de penalização, essa técnica induz alguns coeficientes a atingirem exatamente zero, simplificando o modelo e garantindo sua estabilidade.

A principal diferença entre as regressões de LASSO e de Ridge reside na forma de penalizar os atributos do conjunto de dados, pela remoção de seu peso ou mesmo sua redução para zero, no caso do LASSO, o que o torna mais interpretável (Ribeiro, 2021).

3.2.2 Métodos não lineares

Para melhorar a precisão dos resultados, metodologias de aprendizado supervisionado mais recentes enfatizam o uso de algoritmos mais complexos que têm maior poder de extrair relações não lineares do conjunto de dados como perceptron multicamadas (Multilayer Perceptron Regressor - MLPR), Máquina de Gradiente de Impulso (Grandient Boosting Machine - GBM) e Máquina de Vetor de Suporte (Support Vector Machines - SVM). Em contraste com práticas tradicionais, essas técnicas por meio da extração de relações mais complexas conseguem atingir melhor performance, mas são reconhecidas como "caixa-preta", pois sua interpretabilidade é perdida no processo (Liao *et al*, 2023).

3.2.2.1 Support Vector Regressor (SVR)

Os SVMs representam técnicas de aprendizado de máquina que adotam uma arquitetura menos complexa, proporcionando uma abordagem direta ao converter problemas não lineares em problemas de otimização convexa. Essas máquinas de aprendizado são aplicadas em tarefas de classificação, regressão, o último também conhecido como Regressor de Vetor de Suporte (Support Vector Regressor - SVR) (Bezerra, 2022).

Além disso, quanto ao seu funcionamento, a essência dos algoritmos baseados em vetor de suporte consiste em determinar pontos próximos ao hiperplano (vetores de suporte) que maximizam a margem entre duas classes de pontos (pontos superiores e inferiores à variável-alvo), obtida a partir da diferença entre o valor-alvo e um limiar (Ribeiro, 2021).

3.2.2.2 Light Gradient Boosting Machine (LGBM)

O LGBM é um algoritmo eficiente de aprendizado de máquina, que faz parte do conjunto de métodos *ensemble*. Além disso, destaca-se por sua capacidade de lidar com conjuntos de dados grandes, através da abordagem de *histogram-based learning* que utiliza a amostragem unilateral baseada em gradiente (Gradient-Based One Side Sampling - GOSS) e agrupamento de recursos exclusivos (Exclusive Feature Bundling - EFB), suporta GPU (*Graphical Processing Unit*) e variáveis categóricas (Novaes, 2022). Sua estratégia de crescimento por folhas reduz o tempo de treinamento, enquanto técnicas como regularização, amostragem e parada antecipada reduzem o risco de *overfitting* (Wang *et al*, 2019).

3.3 Indicadores de Performance

Para avaliação dos resultados, fez-se uso de métricas de desempenho erro médio absoluto normalizado (NMAE), raíz do erro médio quadrático normalizada (NRMSE) e o índice de determinação (R2). Além disso, é destacado que durante a análise de resultados, as métricas de erro (NMAE, NMSE) devem ser minimizadas a zero, pois refletem o erro geral, já o R2 pode refletir o acompanhamento entre o valor real e o valor previsto, podendo variar entre -1 e 1, sendo 1 o seu valor ideal (Chen *et al*, 2022).

Na sequência, são definidos os índices MAE, RMSE, NMAE, NRMSE e R2:

$$MAE = \frac{1}{n}\Sigma_{t=1}^{N}|y_i - \widehat{y}_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{t=1}^{N}(y_i - \widehat{y}_i)^2} \quad (3)$$

$$NRMSE = \frac{RMSE}{max(y) - min(y)} \quad (4)$$

$$NMAE = \frac{MAE}{max(y) - min(y)} \quad (5)$$

$$R^2 = 1 - \Sigma_{i=1}^{n}\frac{(y_i - \widehat{y}_i)^2}{(y_i - \widehat{y}_i)^2} \quad (6)$$

Em que, n representa o número de observações na série temporal, $\widehat{y}_i$ a i-ésima observação prevista e y_i a i-ésima observação real.

4 METODOLOGIA

Quanto aos aspectos metodológicos, a presente pesquisa pode ser caracterizada como quantitativa, segundo os pressupostos teóricos de Gil (2006), o qual assume que as investigações quantitativas pressupõem a mensuração de tudo, isto é, a geração de informações mediante números para classificação e análise.

Nesse viés, para a averiguação da metodologia de pré-processamento, o experimento foi realizado com a série temporal do complexo eólico de “SERRA SANTANA 3” sob o critério de ter o maior *uptime* (tempo em operação), o qual localiza-se no Rio Grande do Norte e é composto por 5 usinas, totalizando 108 MW de capacidade instalada.

Foram utilizados modelos de referência para avaliar sua performance em diferentes cenários de restrição da capacidade do complexo supracitado para previsão de 1h à frente, sendo consideradas as métricas NMAE, NRMSE e R2.

No desenvolvimento desta pesquisa, segue-se uma metodologia rigorosa para analisar e aprimorar modelos de previsão de geração de energia eólica em condições de restrição. O processo metodológico adotado é resumido abaixo:

1. Análise descritiva do conjunto de dados;
2. Pré-processamento;
    2.1. Transformação;
    2.2. Remoção de anomalias;
    2.3. Redução de dimensionalidade;
3. Otimização de hiperparâmetros para todos os algoritmos;
4. Implementação dos algoritmos;
5. Avaliação dos indicadores de performance.

Para obtenção dos hiperparâmetros do modelo, a validação cruzada é aplicada sobre os conjuntos de treino que abrange de 12 de outubro de 2021 a 30 de setembro de 2022. Além disso, o período de teste foi definido como os 12 meses posteriores, de 1 de outubro de 2022 a 1 de outubro de 2023. Além disso, todas as execuções foram realizadas no ambiente virtual Google Colab (versão gratuita) com a linguagem Python 3.10.

4.1 Análise Descritiva

Os dados disponíveis para este trabalho são oriundos de uma base de dados real,

de uso do ONS, os mesmos foram coletados com valores médios com amostras de frequência de 30 minutos. Com o intuito da realização deste estudo, os dados foram sumarizados para obter a média do valor de geração real e de geração limitada por hora para o complexo de estudo. Além disso, a informação da razão de restrição e origem são coletadas para cada registro, visto que a série histórica abrange o período de 10 de outubro de 2021 a 01 de outubro de 2023, consistindo em dados de potência elétrica gerada, valor de geração limitada.

Na Tabela 2, há a transposição dos parâmetros para os principais indicadores das amostras nesse período. No que se refere a descrição do cenário de restrição, para a obtenção de um breve conhecimento sobre os padrões encontrados na usina, são destacadas suas principais condições. Para isso são considerados como tempo em operação a razão de registros com geração maior que 2% dentre todos os registros e a capacidade restringida corresponde a redução imposta pela ONS normalizada pela capacidade do complexo eólico.

Tabela 2 – Principais origens de restrição da geração do complexo.

Origem restrição	Tempo em operação médio	Fator de capacidade médio	Capacidade restringida média
Local	99,5%	52,39%	44,35%
Sem restrição	98,8%	49,39%	0,16%
Sistêmica	98,7%	52,31%	42,03%

Fonte: o próprio autor.

A partir da Tabela 2, é possível inferir que apesar de existirem restrições durante o período total considerado, o impacto na redução da capacidade do complexo é baixo, menos de 2%. Além disso, também é observado que mesmo sem restrição registrada, 0.16% de restrição são observados, fator que pode ser justificado devido aos valores serem obtidos em média horária, pois no intervalo destacado, podem ter transições entre período livre de restrição e período com restrição.

Ademais, ambos os tipos de restrição têm efeito equivalente, reduzindo em média a capacidade do parque em 40%, visto que quando esse tipo de restrição ocorre, o fator de capacidade está maior que no cenário sem restrição. Em visão macro, esse indicador pode passar despercebido.

Na Tabela 3, pode-se inferir que esses valores não são uniformes e que possuem tendência de crescimento em períodos mais recentes, o que pode afetar a previsibilidade do complexo em curto prazo.

Tabela 3 – Principais indicadores de operação para o complexo SERRA DE SANTANA 3, em negrito são destacados os melhores resultados por métrica e trimestre.

Ano	Tempo operante médio	Fator de capacidade médio	Geração média (MW)	Redução da capacidade média	Menor valor limitado (MW)	Capacidade disponível média (MW)
2023.3	99,5%	58,02%	62,66	**6,61%**	0	100,87
2023.2	96,7%	40,94%	44,21	**2,60%**	20,88	105,19
2023.1	99,5%	43,54%	47,03	**0,63%**	40,46	107,31
2022.4	99,4%	51,64%	55,77	0,62%	10,58	107,32
2022.3	100,0%	66,07%	71,35	0,40%	43	107,56
2022.2	97,3%	38,77%	41,87	0,11%	72,79	107,87
2022.1	98,4%	39,32%	42,47	0,00%	108	108
2021.4	99,9%	57,02%	61,58	1,35%	3,26	106,53
Resumo	98,8%	49,48%	53,44	1,55%	0	106,32

Fonte: o próprio autor.

Além disso, com a Tabela 3, é possível observar a sazonalidade anual do fator de capacidade, destacando sua alta estação no terceiro trimestre do ano.

Desse modo, como principais responsáveis pela redução da capacidade de operação, é possível presenciar a "Razão Energética" que também apresenta-se associada a redução do fator de capacidade médio para esse complexo. Além disso, a Tabela 4 sumariza as principais razões de restrição e seus impactos na restrição da capacidade média. É destacado em negrito que sobre esse complexo eólico a capacidade de geração média atinge

44%, a maior redução dentre as categorias listadas, e nesse período o gerador está em plena operação o que pode ser interpretado como não aproveitamento da energia renovável disponível.

Tabela 4 – Principais razões de restrição da capacidade, complexo SERRA DE SANTANA 3.

Razão restrição	Tempo operante médio	Fator de capacidade médio	Redução da capacidade média
Indisponibilidade elétrica	99,7%	58,17%	37,51%
Razão energética	100,0%	**51,28%**	**44,35%**
Requisitos de confiabilidade	97,0%	54,57%	40,29%

Fonte: o próprio autor.

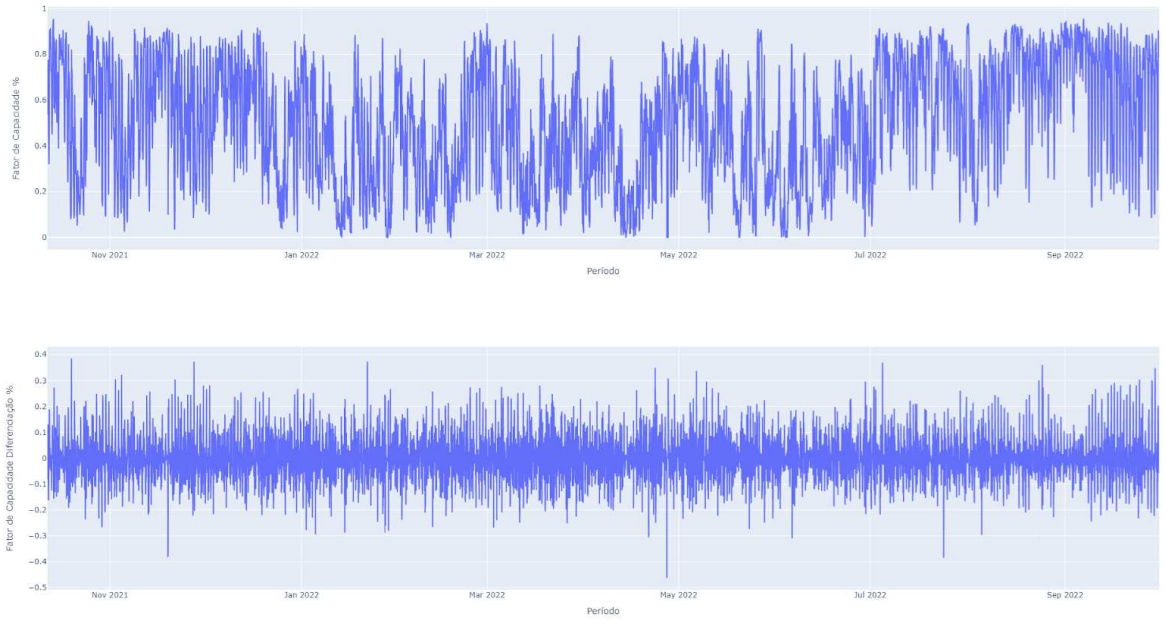
4.2 Pré-processamento

Na seção anterior, foi possível observar as características da geração eólica no complexo estudado. Além disso, na sequência é abordada a metodologia proposta de pré-processamento aplicada aos dados de treino, sendo constituída por transformação, winsorização de anomalias e redução de dimensionalidade.

4.2.1 Transformação

Como pontuado por Redl (2009), a diferenciação é necessária para estacionarizar a variável objetivo. Nesse sentido, a diferenciação da variável de alvo (geração) foi aplicada, o processo pode ser revertido para a forma original por meio da integração ao o valor predecessor. O resultado da transformação pode ser visto de acordo com a Figura 1.

Figura 1- Série temporal do fator de capacidade antes e após a diferenciação primeira, respectivamente



Fonte: o próprio autor.

Além disso, antes do processamento do modelo de *machine learning*, novos atributos podem ser extraídos para enriquecer os dados usados nas previsões. Nesse estudo, informações climatológicas não estão disponíveis, o que exige ainda mais dos dados históricos. Com isso, novos atributos foram gerados automaticamente por meio do pacote *feature-engine.* Com isso, foi possível transformar a série temporal no formato aplicável para aprendizado supervisionado por meio do histórico das últimas 72 horas a partir dos seguintes dados:

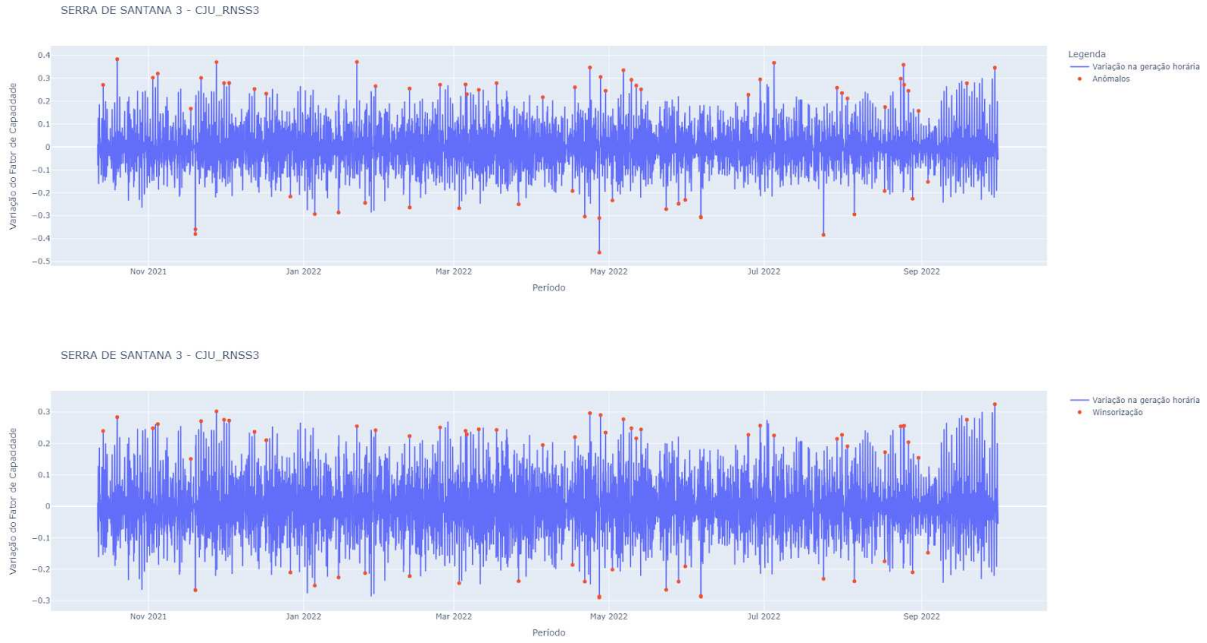
1. Valor da geração do complexo;
2. Hora do dia;
3. Valor da capacidade limitada do CE (variável exógena).

4.2.2 Tratamento de anomalias

Os *outliers* foram ajustados por winsorização para evitar perda de informação. No que se refere ao método por desvio padrão (MDP), foram identificados 66 pontos de alta variação na geração, enquanto que o método pela distância interquartil (MDIQ) detectou 507 pontos. Para a geração eólica, a distribuição não é gaussiana, por esse motivo foi dada sequência pelo método interquartil.

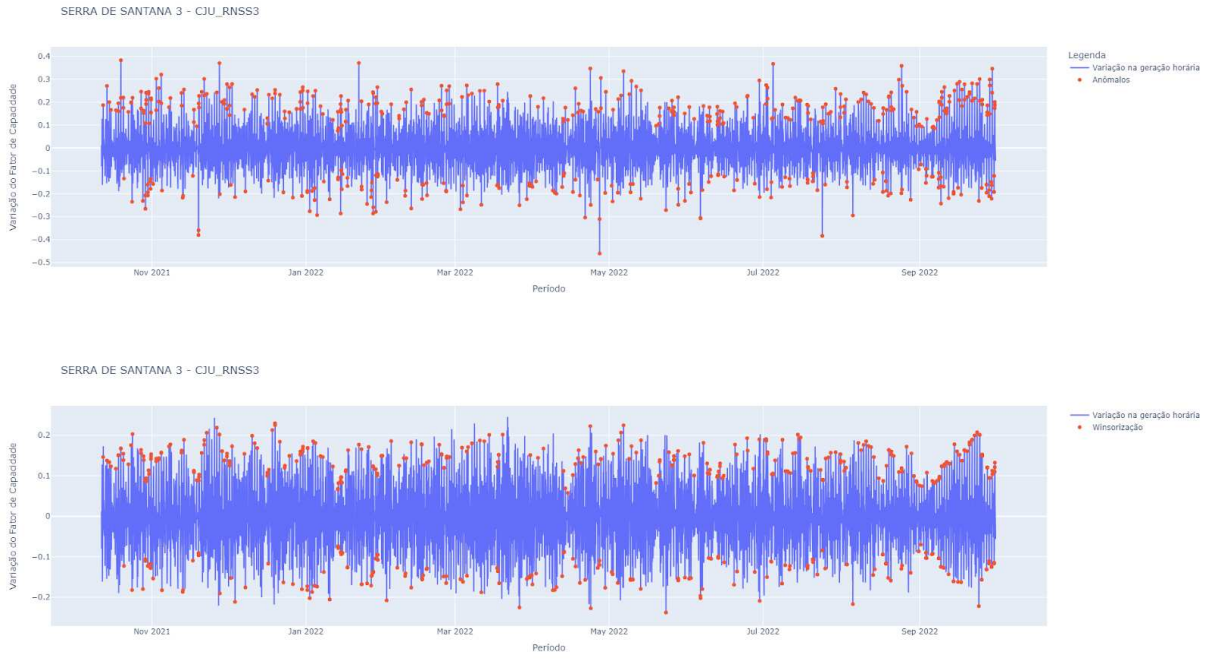
A Figura 2 destaca os registros identificados como anomalias pelos métodos citados e a série após a substituição dos mesmos pelos valores limites de cada método.

Figura 2 - Destaca a identificação das anomalias para os dados de treino no período de outubro de 2021 a outubro de 2022 pelo Método Desvio Padrão, seguida pela substituição desses valores.



Fonte: o próprio autor.

Figura 3 - Destaca a identificação das anomalias para os dados de treino no período de outubro de 2021 a outubro de 2022 pelo Distância Interquartil, seguida pela substituição.

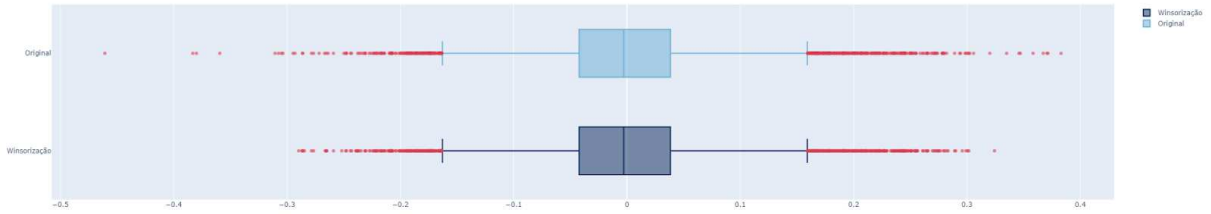


Fonte: o próprio autor.

Observa-se que os pontos identificados como anomalias pelo MDP (Figura 2) também o foram identificados pelo MDIQ (Figura 3), o que é reafirmado pelo seu maior número de pontos marcados. Além disso, ambos os métodos reduziram a amplitude da série

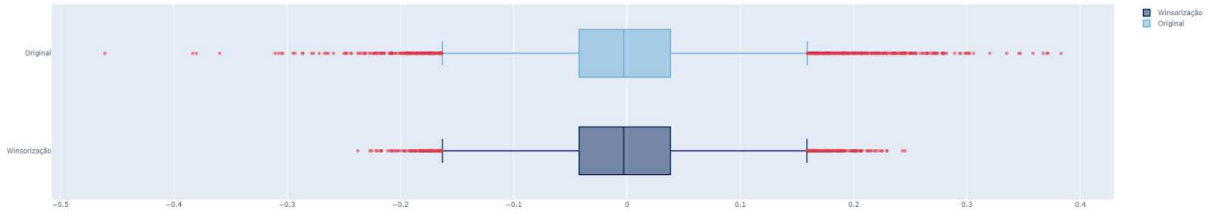
de forma expressiva. As Figura 4 e 5 apresentam o gráfico de caixa (Kumar *et al*, 2023).

Figura 4 - Demonstrando a identificação das anomalias para os dados de treino no período de outubro de 2021 a outubro de 2022 pelo método estatístico desvio padrão



Fonte: o próprio autor.

Figura 5 - Demonstrando a identificação das anomalias para os dados de treino no período de outubro de 2021 a outubro de 2022 pelo método estatístico DIQ



Fonte: o próprio autor

Tabela 5 - Estatística descritiva dos dados antes e após o tratamento dos anômalos sobre os dados de treino e validação para a derivada do Fator de Capacidade.

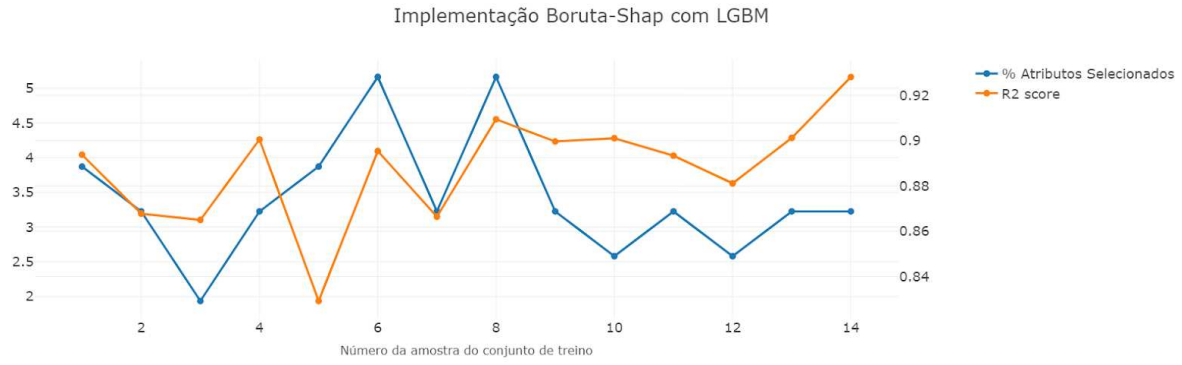
Indicadores	Série original	Série tratada DP	Série tratada DIQ
Registros	8495	8495	8495
Média	0,000	0,000	-0,001
Desvio padrão	0,079	0,078	0,072
Mínimo	-0,461	-0,290	-0,238
25%	-0,042	-0,042	-0,042
50%	-0,003	-0,003	-0,003
75%	0,038	0,038	0,038
Máximo	0,383	0,324	0,245

Fonte: o próprio autor

4.2.4 Redução de dimensionalidade

Como pontuado, o método Boruta-Shap é um algoritmo eficiente para selecionar atributos em conjuntos de dados ruidosos e com intermitência, fator relevante para o presente trabalho. Nessa etapa, o algoritmo é aplicado em conjunto com o regressor LGBM para a extração dos atributos e os marcando como preditores para geração futura com o auxílio da biblioteca shap-hypetune. A Figura 6 destaca os resultados obtidos:

Figura 6 - Seleção de atributos pelo método Boruta-Shap com os dados de treino.



Fonte: o próprio autor

Com os atributos indicados pelo algoritmo, foram considerados os atributos que foram marcados como bons preditores pelo menos 1 vez considerando todas as amostras testadas. A partir do exposto, é possível concluir que o método implementado é bastante efetivo na redução de ruído, eliminando pelo menos 95% dos dados gerados da etapa anterior o ,que deve contribuir para a performance dos algoritmos a serem treinados em termos de estabilidade e velocidade.

4.3 Otimização de hiperparâmetros

Um modelo de aprendizado supervisionado treinado recebe um vetor de entrada e retorna uma saída. Esse resultado é gerado pelos parâmetros do modelo obtidos durante o treino e controlados pelos hiperparâmetros. No entanto, os hiperparâmetros não são obtidos no processo de treinamento do algoritmo e cada modelo possui diferentes configurações dessas configurações e sua definição precisa pode realmente trazer boa performance ao algoritmo (Cheng *et al*, 2019). Nessa perspectiva, para encontrar a melhor combinação de parâmetros foi utilizada a otimização bayesiana por meio da biblioteca Optuna sobre o

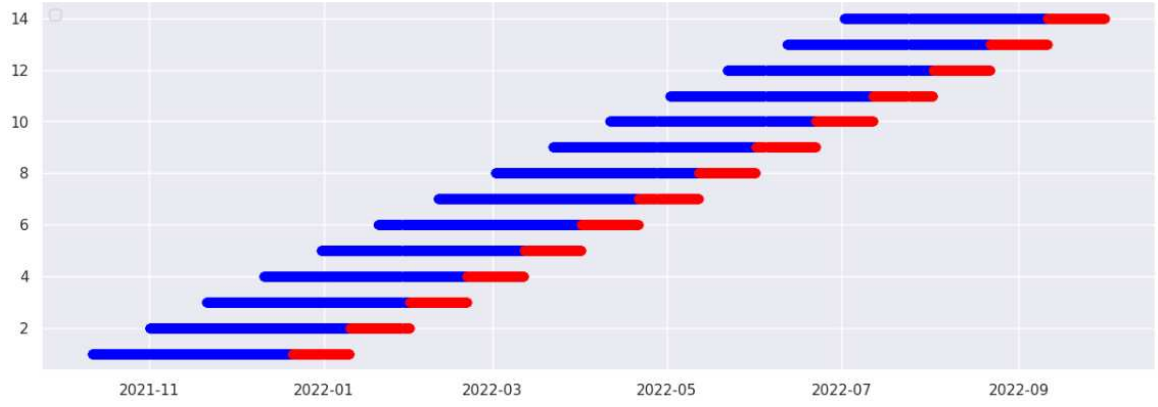
conjunto destinado a treino para todos os modelos para se obter uma comparação justa e assim poder julgar a viabilidade da técnica de pré-processamento em relação aos modelos usados como referência.

Além disso, para tornar as amostras da validação mais representativas das amostras de teste, foi implementada a validação cruzada por janela deslizante. Os dados são sequencialmente deslocados no tempo para que se possa percorrer todo o período disponível para o treino (Petropoulos *et al*, 2022).

Divisão dos dados:

- Inicialmente, os dados foram divididos em conjuntos de treino e teste;
- Para a validação cruzada, utilizou-se uma janela de 365 dias, sendo usadas janelas deslizantes para treino e teste resultando em 14 amostras de 70, 17, 3 (totalizando 90 dias) para treino, teste e espaçamento entre as amostras de treino e teste (*gap*);
- Para o método com pré-processamento, os atributos já foram selecionados pelo método Boruta-Shap;
- Os dados usados para treino passam por winsorização, transformação e redução de dimensionalidade, já os dados de teste passam por transformação e redução;
- No período de 2021 a 2022, 12 meses foram disponibilizados para treino e validação seguida de ajuste nos hiperparâmetros do modelo, os demais 12 meses de outubro de 2022 a outubro de 2023 são reservados para os testes do próximo capítulo.

Figura 7 - Divisão baseada em janela deslizante utilizada na validação cruzada, dados de treino e teste em azul e vermelho, respectivamente.

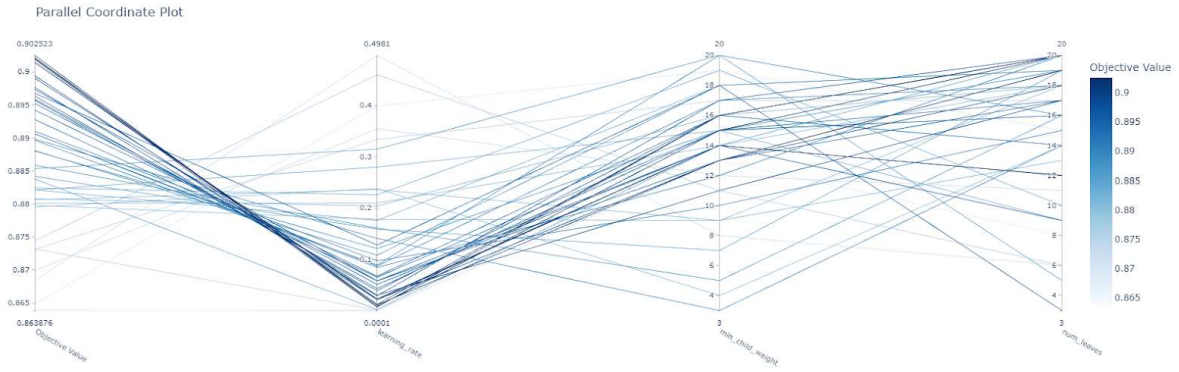


Fonte: o próprio autor

Após o fim do processo, é possível visualizar, pelo gráfico de coordenadas paralelas, como se distribuiu o ajuste dos parâmetros no espaço selecionado para o modelo, utilizando o regressor LGBM, ilustrado na Figura 8.

Como resultado da validação cruzada, são obtidos os hiperparâmetros ajustados e o tempo médio de processamento (considerando treino e teste), o objetivo da otimização era maximizar a mediana do R2 considerando todas as amostras da validação cruzada. A Tabela 6 apresenta a seleção dos hiperparâmetros para os modelos Ridge, LASSO, LGBM e SVR sem o pré-processamento e também para Ridge, LASSO, LGBM e SVR com o pré-processamento e resultado médio obtido pela validação cruzada (marcados com asterisco).

Figura 8 - Resultado da otimização de hiperparâmetros do algoritmo LGBM, índice R2.



Fonte: o próprio autor

Tabela 6 - Hiperparâmetros obtidos após a otimização bayesiana obtidos em 50 iterações.

Método	Hiperparâmetros ajustados	R2	Tempo médio (s)	Desvio padrão (s)
LASSO	alpha': 0.001	87,6271	0,0116	0,0038
LASSO*	alpha': 0.002	90,2455	0,0148	0,0027
LGBM	learning_rate': 0.071, 'num_leaves': 3, 'min_child_weight': 17	87,4534	0,1707	0,0272
LGBM*	learning_rate': 0.008, 'num_leaves': 20, 'min_child_weight': 15	90,1853	0,1567	0,0383
Ridge	alpha': 0.292	88,1268	0,0145	0,0071
Ridge*	alpha': 19.343	90,3096	0,0089	0,0009
SVR	C': 30.800	67,5809	0,1317	0,0262
SVR*	C': 0.884	90,2508	0,1899	0,0033

Fonte: o próprio autor

Com base na Tabela 6, é visível melhor performance para a aplicação do pré-processamento. Quanto aos tempos de processamento, não houve variação expressiva.

5 RESULTADOS

Como linha de base de comparação dos métodos, foi utilizado um modelo de previsão de persistência (PM), no qual a geração da hora h+1 é prevista como igual à da hora anterior (h). Adiciona-se a esse, os algoritmos SVR, LGBM, Ridge e Lasso sem a técnica de pré-processamento proposta.

Nessa perspectiva, os hiperparâmetros para os modelos de referência sem a metodologia de pré-processamento e para os modelos com a metodologia de pré-processamento foram testados com os parâmetros obtidos no capítulo anterior.

Dessa forma, foram usadas amostras de teste de diferentes períodos do ano, utilizadas para representar os diferentes cenários de restrição operativa:

1. Amostra de teste 2022.T4 : 2022-10-04 - 2022-12-31
2. Amostra de teste 2023.T1 : 2023-01-01 - 2023-03-31
3. Amostra de teste 2023.T2 : 2023-04-01 - 2023-06-30
4. Amostra de teste 2023.T3 : 2023-07-01 - 2023-09-30

Figura 9 – Evolução do fator de capacidade e da capacidade restringida no período de teste.

Fonte: o próprio autor

Dos resultados vistos na Tabela 7, para os métodos sem o pré-processamento aplicado, observa-se que todos os métodos superam o método de persistência, com exceção do SVR e do LGBM. Também, os algoritmos lineares LASSO e Ridge com fator de penalização conseguiram manter estabilidade na previsão em praticamente todos os períodos, com destaque para o método LASSO. O resultado destacado pode ser associado ao fator de penalização que esse algoritmo possui, o que garante maior robustez.

Por outro lado, para o pré-processamento aplicado, é possível concluir que todos os modelos tiveram ganho em performance em relação aos modelos iniciais. Desse modo, há o destaque para o SVR que conseguiu gerenciar melhor sua performance após a aplicação proposta. Os métodos de Ridge, LGBM e LASSO também apresentaram evolução, o que reforça o efeito da metodologia para diferentes algoritmos.

Tabela 7 - Resultados para os modelos com a aplicação da metodologia proposta em base percentual, previsão 1h a frente, em negrito são destacados os modelos com a técnica de pré-processamento proposta e em sublinhado o melhor valor obtido. Valores em base percentual.

	Período de teste 2022.4			Período de teste 2023.1			Período de teste 2023.2			Período de teste 2023.3		
Método	NMAE	NRMSE	R2	NMAE	NRMSE	R2	NMAE	NRMSE	R2	NMAE	NRMSE	R2
LASSO	5,3	7,2	92,12	5,89	8,3	87,51	6,38	8,52	88,83	5,81	7,9	88,6
LASSO*	4,98	6,7	93,2	5,46	7,83	88,86	5,98	8,15	89,77	5,38	7,47	89,79
LGBM	5,3	7,12	92,31	6,58	8,84	85,81	6,9	8,98	87,59	5,93	7,99	88,31
LGBM*	4,99	6,84	92,9	5,14	7,59	89,54	5,87	8,08	89,95	5,39	7,6	89,45
PM	6,29	8,74	88,39	5,37	8,27	87,59	6,25	8,68	88,39	6,09	8,66	86,26
Ridge	5,1	6,94	92,68	5,77	8,07	88,17	6,17	8,32	89,34	5,63	7,63	89,35
Ridge*	4,97	6,69	93,2	5,43	7,79	88,98	5,89	8,09	89,91	5,39	7,51	89,69
SVR	8,15	10,23	84,12	16,01	18,62	37,02	18,07	22,63	21,12	9,28	11,56	75,57
SVR*	4,85	6,6	93,4	5,21	7,65	89,39	5,88	8,03	90,06	5,25	7,42	89,94

Fonte: o próprio autor

Ademais, é observado que a performance dos modelos lineares é bastante próxima à observada nos modelo não lineares, o que pode ser visto em intervalos de previsões de curto prazo nos resultados de Liao *et al* (2023). A Tabela 8, destaca os resultados em relação ao PM, é possível observar que, em relação a NMAE, NRMSE e R2, os algoritmos tendem a performar melhor que o PM, e que os maiores destaques são encontrados na versão do algoritmo com a técnica de pré-processamento (LGBM e SVR). Esse efeito pode ser explicado pela redução do efeito do ruído atribuído a seleção de atributos e pela winsorização.

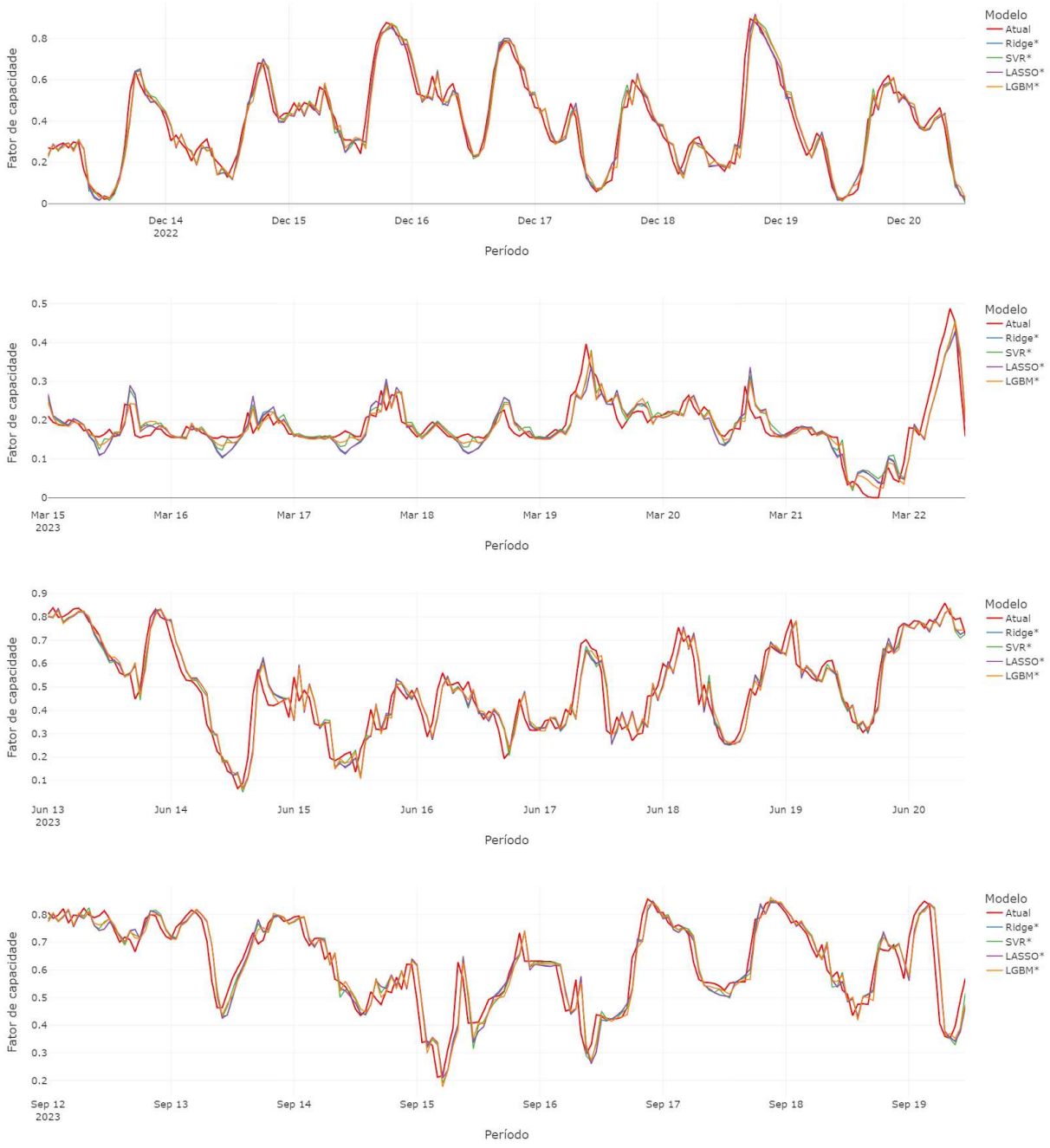
Tabela 8 – Variação absoluta dos resultados relativos ao modelo PM para os modelos aplicados. horizonte de previsão de 1h, complexo SERRA DE SANTANA 3, em negrito são destacados os modelos com a técnica de pré-processamento e em sublinhado o melhor valor obtido. Valores em base percentual.

	Período de teste 2022.4			Período de teste 2023.1			Período de teste 2023.2			Período de teste 2023.3		
Método	NMAE	NRMSE	R2	NMAE	NRMSE	R2	NMAE	NRMSE	R2	NMAE	NRMSE	R2
LASSO	-0,99	-1,54	3,73	0,52	0,03	-0,08	0,13	-0,16	0,44	-0,28	-0,76	2,34
LASSO*	-1,31	-2,04	4,81	0,09	-0,44	1,27	-0,27	-0,53	1,38	-0,71	-1,19	3,53
LGBM	-0,82	-1,38	3,38	1,1	0,52	-1,61	0,78	0,33	-0,91	0,03	-0,5	1,55
LGBM*	-1,04	-1,52	3,7	-0,28	-0,64	1,85	-0,39	-0,55	1,44	-0,61	-0,84	2,55
PM	0	0	0	0	0	0	0	0	0	0	0	0
Ridge	-1,19	-1,8	4,29	0,4	-0,2	0,58	-0,08	-0,36	0,95	-0,46	-1,03	3,09
Ridge*	-1,32	-2,05	4,81	0,06	-0,48	1,39	-0,36	-0,59	1,52	-0,7	-1,15	3,43
SVR	1,86	1,49	-4,27	10,64	10,35	-50,57	11,82	13,95	-67,27	3,19	2,9	-10,69
SVR*	-1,44	-2,14	5,01	-0,16	-0,62	1,8	-0,37	-0,65	1,67	-0,84	-1,24	3,68

Fonte: o próprio autor

Em complemento aos resultados das tabelas 7 e 8, a Figura 10 ilustra os resultados obtidos para as das primeiras 180h de cada amostra de teste com valores reais e previstos pelos modelos com o pré-processamento aplicado.

Figura 10 – Dados reais e previstos para o fator de capacidade horário do complexo SERRA DE SANTANA 3 para os modelos de referência por trimestre, resultado das primeiras 180h de cada período de teste utilizando a metodologia proposta.

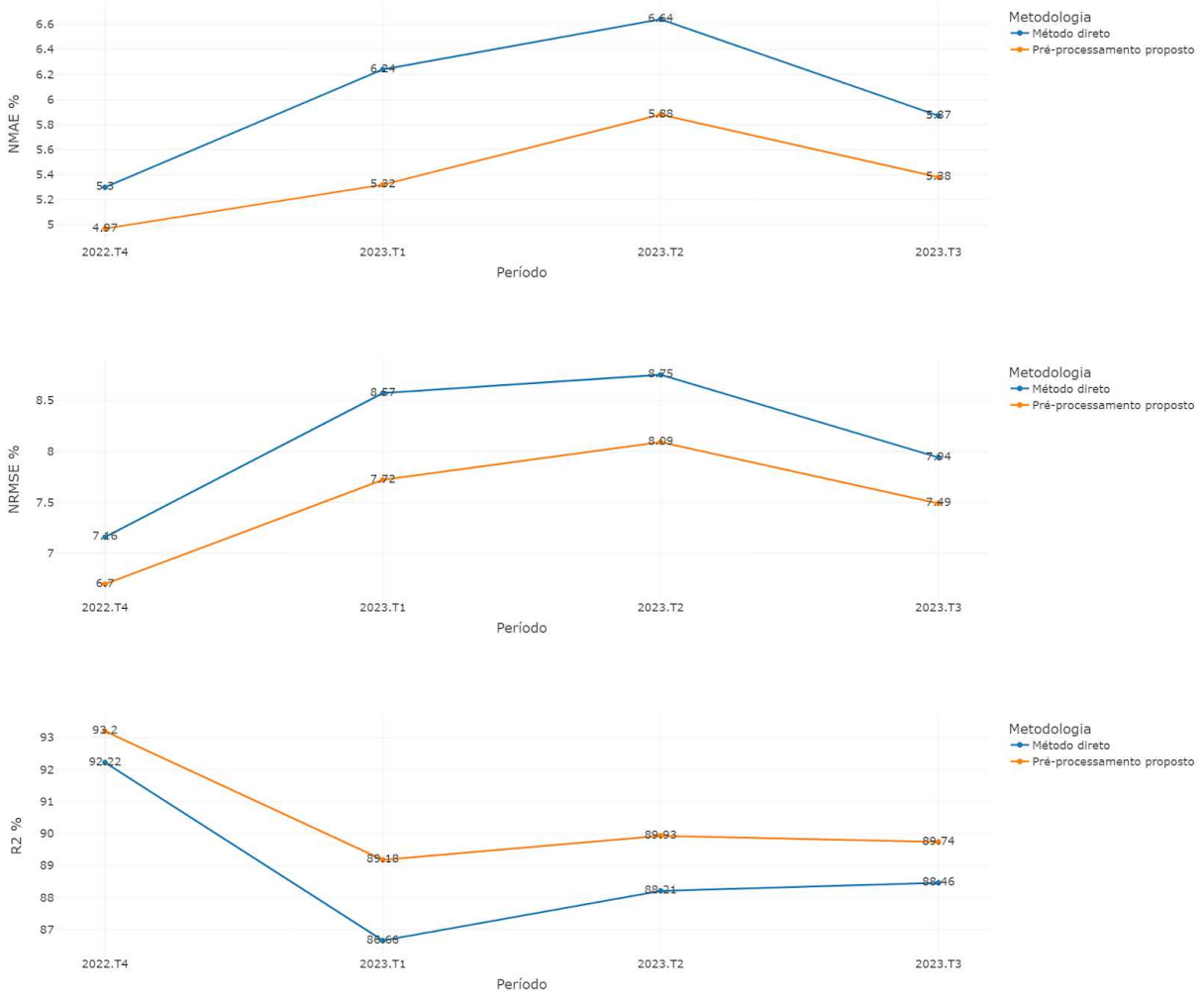


Fonte: o próprio autor.

Considerando a Figura 10, é possível ver que existem períodos com maior distanciamento entre o valor real e os valores previstos para os modelos, podendo-se ver o melhor acompanhamento geral para a amostra de 2022 (2022.T4), o que é confirmado pela Tabela 7 em valores absolutos.

A partir da Figura 11 pode-se concluir que os modelos com o pré-processamento aplicado apresentam maior performance nos diferentes cenários testados em termos de NMAE, NRMSE e R2. Desta maneira, a implementação deste método é bastante satisfatória do ponto de vista de métricas de avaliação e, principalmente, estabilidade dos resultados.

Figura 11 – Alteração da mediana das métricas de performance para os algoritmos que foram submetidos à técnica proposta.

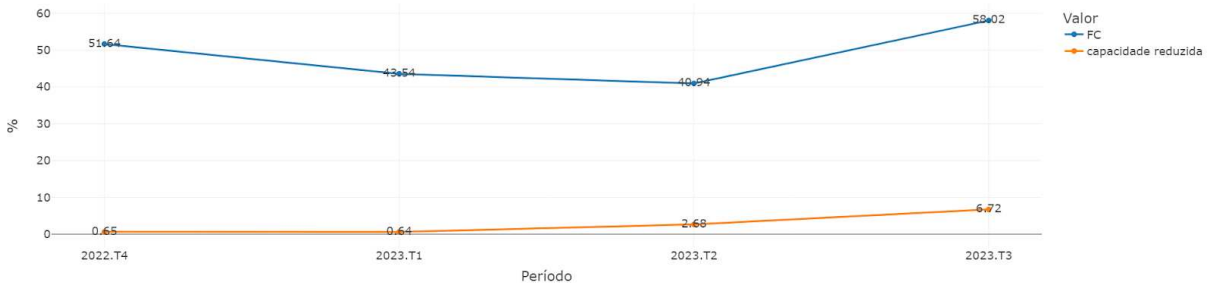


Fonte: o próprio autor.

A partir do exposto na Figura 12, é notável a tendência de crescimento da restrição operativa sobre o complexo no período de teste. Tomando como referência a Figura 11 a

metodologia de pré-processamento aplicada conseguiu não só ter baixa oscilação nos índices de performance como também manter níveis de precisão abaixo da abordagem direta, sem o pré-processamento proposto.

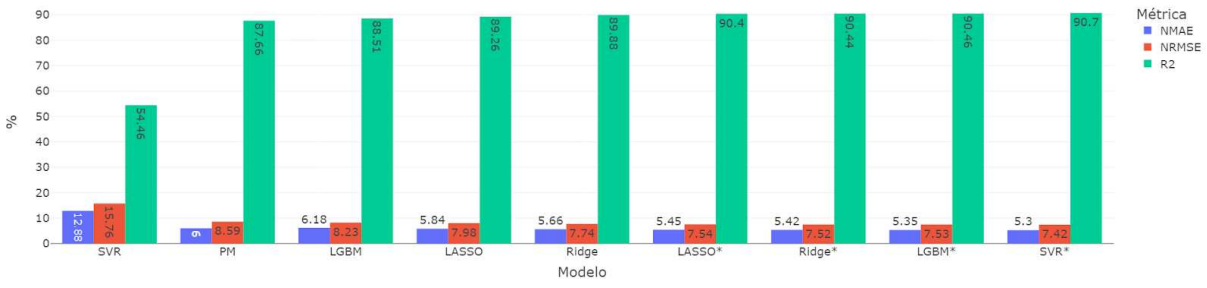
Figura 12 – Evolução da capacidade frustrada e fator de capacidade do complexo no período de teste.



Fonte: o próprio autor.

A Figura 13 resume a média das métricas de avaliação de todos os oito modelos de previsão implementados neste trabalho e do método PM, posicionados da esquerda para a direita em valor crescente de R2. A partir da Figura 13, percebe-se que os modelos com a aplicação do pré-processamento obtiveram a melhor performance, ocupando todas as posições à direita da ilustração.

Figura 13 – Resultados dos modelos considerados no trabalho, horizonte de previsão de 1h, em asterisco são destacados modelos com a metodologia proposta, resultados ordenados pelo R2 em ordem ascendente.



Fonte: o próprio autor.

A partir do exposto nos resultados acima e enfatizado na Figura 13, é notório que o emprego de técnicas de pré-processamento acarreta em ganho de performance e de estabilidade dos resultados, uma vez que o incremento de performance se repetiu em todos os algoritmos aplicados e em todos cenários de restrição operativa representados pelas amostras de teste. Além disso, as variações nas métricas de avaliação pelo uso da técnica de pré-processamento aplicada foram expressivas para os algoritmos SVR e LGBM, o que ressalta sua robustez da metodologia para previsão de curta duração em meio à crescente da restrição operacional aplicada pelo ONS.

6 CONCLUSÃO

Alterações nas condições de previsão podem afetar a estabilidade das previsões e sua acuracidade, o que impacta diretamente na confiabilidade do sistema elétrico. Dito isso, para enfrentar esse desafio, a presente pesquisa propõe uma metodologia de pré-processamento que pode ser aplicada antes do treinamento de modelos, permitindo assim sua aplicação tanto em métodos mais simples, como Ridge e LASSO como em algoritmos mais complexos como o LGBM e o SVR. As seguintes conclusões puderam ser obtidas com base no experimento realizado considerando diferentes períodos do ano e diferentes intensidades na restrição operacional:

1. A técnica de pré-processamento dos dados trouxe ganho em performance e estabilidade que se repetiu em cada modelo e em cada cenário testado, o que reforça sua aplicabilidade;
2. Foi visto que os algoritmos LASSO e Ridge tiveram performance próxima aos métodos LGBM e SVR e ainda tempo de processamento bastante reduzido, mesmo sem a aplicação do pré-processamento;
3. As abordagens por SVR e LGBM se mostraram bastante sensíveis à metodologia proposta, e destacam o pré-processamento como fator limitante aos resultados do modelo;
4. Foi vista durante a validação cruzada redução do tempo de processamento total (treino e teste) os algoritmos Ridge e LGBM e aumento para SVR e LASSO com a aplicação do pré-processamento, sendo os algoritmos LGBM e SVR os que demandaram mais tempo.

Para trabalhos futuros, pode-se estender a avaliação para métodos proeminentes na literatura e que não foram contemplados neste trabalho, como abordagem com a técnica "*escape seasonality*" aplicada por Smyl e Hua (2019) e também a decomposição das séries utilizando decomposições como *Variational Mode Decomposition* e *Seasonal and Trend decomposition using Loess* para reduzir os efeitos de ruído e sazonalidade. Ademais, como apontado por Hong (2020), mais atributos podem contribuir para a previsão de curto prazo como previsões climáticas de curto prazo e as intervenções operacionais aplicadas sobre os geradores. Além disso, pesquisas futuras podem abordar horizontes maiores para investigar a estabilidade do modelo em outros intervalos da previsão de curta duração indo além da previsão de 1h à frente aplicada neste estudo.

REFERÊNCIAS

ABEEOLICA. O Setor: Desenvolvimento da eólica no Brasil. In: Desenvolvimento da eólica no Brasil. [S. l.], 2022. Disponível em: https://abeeolica.org.br/energiaeolica/o-setor/. Acessado em 18 de dezembro de 2023.

ANEEL: **RESOLUÇÃO NORMATIVA** Nº 927, DE 23 DE MARÇO DE 2021. 2021. Disponível em: http://www2.aneel.gov.br/cedoc/ren2021927.html. Acessado em 11 de dezembro de 2023.

ANEEL: **RESOLUÇÃO NORMATIVA** Nº 1.030, DE 26 DE JULHO DE 2022. 2022. Disponível em: http://www2.aneel.gov.br/cedoc/ren20221030.pdf. Acessado em 11 de dezembro de 2023.

BEZERRA, E.C. Abordagem auto-adaptativa baseada no conceito de expectativa de vida aplicada aos métodos Particle Swarm Optimization e máquinas kernel para previsão da velocidade do vento e geração eólica. 2022. 135 f. **Tese** (Doutorado em Engenharia Elétrica) - Centro de Tecnologia, Universidade Federal do Ceará, Fortaleza, 2022

CHENG, H.; DING, X.; ZHOU, W.; e DING, R. (2019). A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange. International Journal of Electrical Power & Energy Systems, 110, 653-666.

CHEN, Y.; HU, X. e ZHANG, E. L. (2022). A review of ultra-short-term forecasting of wind power based on data decomposition-forecasting technology combination model. **Energy Reports**, 8, 14200-14219. DOI: https://doi.org/10.1016/j.egyr.2022.10.342

GIL, Antônio C. Métodos e técnicas de pesquisa social. 5. ed. São Paulo: **Atlas**, 2006.

HANIFI, S.; LIU, X.; LIN, Z. e LOTFIAN, S. A Critical Review of Wind Power Forecasting Methods—Past, Present and Future. **Energies**. 2020; 13(15):3764. https://doi.org/10.3390/en13153764

HONG, T.; PINSON, P.; WANG, Y.; WERON, R.; YANG, D. e ZAREIPOUR, H. (2020). **Energy Forecasting**: A Review and Outlook. IEEE Open Access Journal of Power and Energy, 7, 376-388. https://doi.org/10.1109/OAJPE.2020.3029979

KISVARI, LIN Z. e LIU, X. Wind Power Forecasting – A Data-driven Method along with Gated Recurrent Neural Network. **Renewable Energy**, DOI: https://doi.org/10.1016/j.renene.2020.10.119.

KUMAR DASH CH, Sanjeev; KUMAR BEHERA, Ajit; DEHURI, Satchidananda; GHOSH, Ashish. An outliers detection and elimination framework in classification task of data mining. **Decision Analytics Journal**, 2023.

KURSA, M. e RUDNICKI, W. (2010). Feature Selection with the Boruta Package. **Journal of Statistical Software**, 36(11), 1–13.

LIAO, Wenlong *et al*. Explainable Modeling for Wind Power Forecasting: A Glass-Box Approach with Exceptional Accuracy. **E 23Rd Power Systems Computation Conference**, Paris, France, jul. 2023. DOI: https://doi.org/10.48550/arXiv.2310.18629

MAZZANTI, S. Boruta. **Explained Exactly How You Wished Someone Explained to You** (2020) Disponível em: https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-to-me-4489d70e154a, Acessado em 11 de dezembro de 2023

MINKA, Thomas P. Automatic choice of dimensionality for PCA. In: **Advances in neural information processing systems**, pages 598–604, 2001.

NOVAES, Jonas Arruda Neto. Modelo preditivo de capacidade de pagamento para prospecção PF: atraindo e fidelizando clientes no cenário de Open Finance. 2022. 128 f., il. **Dissertação** (Mestrado Profissional em Economia) — Universidade de Brasília, Brasília, 2022

PETROPOULOS, *et al* (2022). Forecasting: theory and practice. **International Journal of Forecasting**, 38(3), 705–871.

PIROLLA, Francisco Rocha. Redução de dimensionalidade usando agrupamento e discretização ponderada para a recuperação de imagens por conteúdo. **Tese** (Doutorado em Ciência da Computação) - Universidade Federal de São Carlos, São Carlos, 2012.

QIU, X.; REN, Y.; SUGANTHAN, P. N. e AMARATUNGA, G. A. J. (2017). Short-term wind power ramp forecasting with empirical mode decomposition based ensemble learning techniques. 2017 IEEE **Symposium Series on Computational Intelligence** (SSCI). doi:10.1109/ssci.2017.8285421

REDL, C.; HAAS, R.; HUBER, C. e BÖHM B. (2009). Price formation in electricity forward markets and the relevance of systematic forecast errors. **Energy Economics**, 31(3), 356–364. doi:10.1016/j.eneco.2008.12.001

RIBEIRO, Matheus Henrique Dal Molin. Time series forecasting based on ensemble learning methods applied to agribusiness, epidemiology, energy demand, and renewable energy. 2021. **Tese** (Doutorado em Engenharia de Produção e Sistemas) - Pontifícia Universidade Católica do Paraná, Curitiba, 2021.

RICARDO, P. Masini; MARCELO, C. Medeiros e EDUARDO F. Mendes. (2021). **Machine Learning Advances for Time Series Forecasting.** DOI:https://doi.org/10.48550/arXiv.2012.12802

SEBASTIÁN, C.; CARLOS, E. e GONZÁLEZ-GUILLÉN. (2023). **A feature selection method based on Shapley values robust to concept shift in regression.** DOI: https://doi.org/10.48550/arXiv.2304.14774

SEO, Songwon. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. **Dissertação** (Master of Science) — University of Pittsburgh, Pennsylvania, (2006).

SHEN, W.; JIANG, N. e LI, N. (2018). An EMD-RF Based Short-term Wind Power Forecasting Method. In 2018 IEEE **7th Data Driven Control and Learning Systems Conference** (DDCLS) (pp. 283-288).

SMYL, S. N. e HUA, G. (2019). Machine learning methods for GEFCom2017 probabilistic load forecasting. **International Journal of Forecasting**, 35(4), 1424-1431.

WANG, R.; LIU, Y.; YE, X.; TANG, Q.; GOU, J.; HUANG, M. e WEN, Y. (2019). Power System Transient Stability Assessment Based on Bayesian Optimized LightGBM. 2019 IEEE **3rd Conference on Energy Internet and Energy System Integration (EI2)**. doi:10.1109/ei247390.2019.906202

ZHA, W.; LIU, J.; LI, Y. e LIANG, Y. "Ultra-short-term Power Forecast Method for the Wind Farm Based on Feature Selection and Temporal Convolution Network." **ISA Transactions** 129 (2022): 405-14. Web.

ZHOU, Y.; MA, L.; NI, W. e YU, C. Data Enrichment as a Method of Data Preprocessing to Enhance Short-Term Wind Power Forecasting. **Energies**, [S.l.], v. 16, n. 5, p. 2094, 2023. DOI: https://doi.org/10.3390/en16052094.

ZOU, M. e DJOKIC, S. Z. A Review of Approaches for the Detection and Treatment of Outliers in Processing Wind Turbine and Wind Farm Measurements. **Energies**, [S.l.], v. 13, n. 16, p. 4228, 2020. DOI: https://doi.org/10.3390/en13164228