UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

PREVISÃO DA IRRADIAÇÃO SOLAR UTILIZANDO MÉTODO *ENSEMBLE* PARA SELEÇÃO DE ATRIBUTOS E ALGORITMOS DE APRENDIZADO DE MÁQUINA

EDNA SOFÍA SOLANO MEJÍA

TD 18/23

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2023

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

EDNA SOFÍA SOLANO MEJÍA

PREVISÃO DA IRRADIAÇÃO SOLAR UTILIZANDO MÉTODO *ENSEMBLE* PARA SELEÇÃO DE ATRIBUTOS E ALGORITMOS DE APRENDIZADO DE MÁQUINA

TD 18/23

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2023

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

EDNA SOFÍA SOLANO MEJÍA

PREVISÃO DA IRRADIAÇÃO SOLAR UTILIZANDO MÉTODO *ENSEMBLE* PARA SELEÇÃO DE ATRIBUTOS E ALGORITMOS DE APRENDIZADO DE MÁQUINA

Dissertação submetida à banca examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para a obtenção do Grau de Mestre em Engenharia Elétrica na área de Sistemas de Energia Elétrica

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2023

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará

S684p Solano Mejía, Edna Sofía.
Previsão da irradiação solar utilizando método ensemble para seleção de atributos e algoritmos de aprendizado de máquina / Edna Sofía Solano Mejía. — 2023.
101 f. : il. color.

Orientador(a): Profª. Dra. Carolina de Mattos Affonso
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2023.

1. Seleção de Atributos. 2. Aprendizagem da Máquina. 3. Geração Fotovoltaica. 4. Previsão da Irradiação Solar. 5. Clusterização. I. Título.

CDD 621.31098115

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"PREVISÃO DA IRRADIAÇÃO SOLAR UTILIZANDO MÉTODO ENSEMBLE PARA SELEÇÃO DE ATRIBUTOS E ALGORITMOS DE APRENDIZADO DE MÁQUINA"

AUTORA: **EDNA SOFIA SOLANO MEJÍA**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRA EM ENGENHARIA ELÉTRICA NA ÁREA DE SISTEMAS DE ENERGIA ELÉTRICA.

APROVADA EM: 30/06/2023

BANCA EXAMINADORA:

Prof.ª Dr.ª Carolina de Mattos Affonso
(Orientadora – PPGEE/UFPA)

Prof. Dr. Roberto Célio Limão de Oliveira
(Avaliador Interno – PPGEE/UFPA)

Prof.ª Dr.ª Ahda Pionkoski Grilo Pavani
(Avaliadora Externa – UFABC)

VISTO:

Prof. Dr. Diego Lisboa Cardoso

(Coordenador do PPGEE/ITEC/UFPA)

En reconocimiento por su apoyo incondicional;
Y por nunca dejarme olvidar que soy capaz
de realizar todo objetivo que tenga en mente;
Este trabajo es cariñosamente dedicado a:
mi maravillosa família.

AGRADECIMENTOS

A pesquisa descrita nessa dissertação não seria possível sem a ajuda de inúmeras pessoas, portanto eu gostaria de agradecer.

Agradeço a minha família por seu infinito apoio e amor o tempo todo. Foi seu amor e incentivo ilimitado que me ajudaram a superar todas as dificuldades. Agradeço todos os sacrifícios que eles fizeram por mim.

Expresso minha sincera gratidão a minha orientadora, Professora Carolina Affonso, pela excelência na orientação, por seu contínuo apoio guiando-me na elaboração deste trabalho, ajudando-me a crescer com sucesso.

Aos meus amigos do Laboratório de Smart Grid (LasGRid), que estavam sempre dispostos a me ajudar de todas as maneiras possíveis. Agradeço ao Jorge e à Vivian pelas gentis boas-vindas ao laboratório. Agradeço ao Hugo, João e Amanda pela valiosa e paciente assistência nas disciplinas. E, especialmente, ao Sebastião por compartilhar dúvidas e estar sempre disponível para colaborar com novas perspectivas.

Um agradecimento muito especial à minha querida amiga Melany Gamez, por todo o seu apoio incondicional, pela sua companhia nos momentos bons e difíceis, pelas risadas e lágrimas compartilhadas. Sinto-me profundamente grata por ter vivido essa experiência única ao seu lado.

À CAPES através do custeio da bolsa de estudo sendo um apoio financeiro, ao Programa de Pós-graduação em Engenharia Elétrica da UFPA e aos professores por compartilharem seus conhecimentos que contribuíram enormemente para o meu crescimento acadêmico durante estes anos.

E finalmente, porém não menos importante, a todos aqueles que contribuíram de alguma forma para o meu desenvolvimento pessoal e profissional.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AdaBoost	Boosting Adaptável, do inglês *Adaptive Boosting*
AG	Algoritmo Genético
AM	Aprendizado de Máquina
ANEEL	Agência Nacional de Energia Elétrica
ARIMA	Autoregressivo Integrado de Médias Móveis
ARMA	Autoregressivo de Médias Móveis
CatBoost	Impulso Categórico, do inglês *Categorical Boosting*
CEPEL	Centro de Pesquisas de Energia Elétrica
CH	Índice de validação de Calinski-Harabasz
CNN	Redes Neurais Convolucionais, do inglês *Convolutional Neural Network*
CO_2	Dióxido de Carbono
CRESESB	Centro de Referência para as Energias Solar e Eólica Sérgio de Salvo Brito
CSO	Otimização de Enxame de Frango, do inglês *Chicken Swarm Optimization*
CT-ENERG	Fundo Setorial de Energia
DB	Índice de validação de Davies Bouldin
DBN	Rede de Crenças Profundas, do inglês *Deep Belief Networks*
DM	Diebold – Mariano
DNN	Rede Neural Profunda, do inglês *Deep Neural Networks*
DT	Arvores de Decisão, do inglês *Decision Trees*
ELM	Máquina de Aprendizado Extremo, do inglês *Extreme Learning Machines*
ES	Suavização Exponencial
ET	Arvores Extremamente Aleatórias, do inglês *Extra Trees*
FV	Fotovoltaica
GRU	Unidades Recorrentes Fechadas, do inglês *Gated Recurrent Unit*
GWO	Otimização do Lobo Cinzento, do inglês *Gray Wolf Optimization*
IA	Inteligência Artificial
IDE	Ambiente de Desenvolvimento Integrado, do inglês *Integrated Development Environment*
IEA	Agência Internacional de Energia, do inglês *International Energy Agency*
IM	Informação Mútua
INMET	Instituto Nacional de Meteorologia

IRENA	Agência Internacional de Energia Renovável, do inglês *International Renewable Energy Agency*
kNN	k-Vizinhos mais Próximos, do inglês *k-Nearest Neighbours*
LASSO	Encolhimento e Seleção pelo Menor Valor Absoluto, do inglês *Least Absolute Shrinkage and Selection Operator*
LES	Suavização Exponencial Lineal
LightGBM	Máquina de Impulso Gradiente Leve, do inglês *Light Gradient Boosting Machine*
LpT	Programa Luz para Todos
LSTM	Rede Neural de Memória de Longo Prazo, do inglês *Long Short Term Memory*
MAE	Erro Médio Absoluto, do inglês *Mean Absolute Error*
MAPE	Erro Médio Absoluto Percentual, do inglês *Mean Absolute Percentage Error*
MF	Floresta Mondrian, do inglês *Mondrian Forest*
MLP	Perceptron Multicamadas, do inglês *Multi-Layer Perceptron*
MME	Ministério de Minas e Energia
NWP	Previsões Numéricas do Tempo, do inglês *Numerical Weather Predictions*
PRODEEM	Programa de Desenvolvimento Energético para Estados e Municípios
PSO	Otimização por Enxame de Partículas, do inglês *Particle Swarm Optimization*
R^2	Coeficiente de Determinação, do inglês *Coefficient of Determination*
RBF	Função de base Radial, do inglês *Radial Basis Function*
RF	Floresta Aleatória, do inglês *Random Forest*
RL	Regressão Linear
RMSE	Raiz do Erro Quadrático Médio, do inglês *Root mean square error*
RNA	Redes Neurais Artificias
RP	Regressão Polinomial
RW	Passeio Aleatório, do inglês *Random Walk*
SBS	Seleção Sequencial Regressiva, do inglês *Sequential Backward Selection*
SES	Suavização Exponencial Simples
SFS	Seleção Sequential Progressiva, do inglês *Sequential Forward Selection*
SH	Índice de validação de Silhouette
SOM	Mapas Auto-organizados do inglês *Self-organizing maps*
STC	Condições de Teste Padrão, do inglês *Standard Testing Conditions*
SVR	Regressão de Vetores de Suporte, do inglês *Support Vector Regression*
VMP	Vizinho Mais Próximo

VOA	Votação Média, do inglês *Voting Average*
VOWA	Votação Média Ponderada, do inglês *Weighted Voting Average*
VR	Regressor por Votação, do inglês *Voting Regressor*.
XGBT	Árvores de Aumento de Gradiente Extremo, do inglês *Extreme Gradient Boosting Trees*

RESUMO

A previsão precisa da irradiação solar é essencial para a gestão eficaz de sistemas de energia com geração fotovoltaica significativa. Algoritmos de aprendizado de máquina, que utilizam dados históricos e padrões para fazer previsões, desempenham um papel crucial nessa tarefa. Um aspecto chave é o uso de modelos *ensemble*, que combinam as previsões de vários algoritmos para melhorar a precisão e confiabilidade das previsões. Neste estudo, modelos *ensemble* são utilizados para aprimorar o desempenho das previsões, agregando as previsões de diferentes algoritmos. Além disso, o trabalho propõe um método de seleção de atributos *ensemble*, que envolve identificar os parâmetros de entrada mais relevantes e suas observações passadas relacionadas. Essa abordagem tem como objetivo otimizar os atributos de entrada utilizados pelos algoritmos de aprendizado de máquina, garantindo que apenas as informações mais pertinentes sejam consideradas para previsões precisas de irradiação solar. Ao aproveitar as habilidades de múltiplos algoritmos e selecionar os atributos mais informativos, a abordagem *ensemble* oferece uma estrutura robusta para melhorar a precisão das previsões de irradiação solar.

O desempenho de vários algoritmos de aprendizado de máquina, incluindo modelos ensemble, é comparado para previsão de irradiação solar em dias com diferentes padrões climáticos, utilizando entradas endógenas e exógenas. Os algoritmos considerados são AdaBoost, SVR, RF, XGBT, CatBoost, VOA e VOWA. A seleção de atributos *ensemble* proposta depende dos algoritmos RF, IM e *Relief*. A precisão da previsão é avaliada com base em várias medidas usando um banco de dados real da cidade de Salvador, Brasil. Diferentes previsões climáticas são consideradas: 1 hora, 2 horas, 3 horas, 6 horas, 9 horas e 12 horas com antecedência. Os resultados numéricos mostram que a seleção de atributos *ensemble* proposta melhora a precisão da previsão e que o modelo VOWA selecionado com os algoritmos de melhor desempenho apresenta previsões com maior precisão do que os outros algoritmos em diferentes horizontes de previsão. Esta pesquisa demonstra a eficácia dos modelos *ensemble* e as técnicas de seleção de atributos na melhoria da previsão de irradiância solar, fornecendo insights valiosos para a gestão eficiente de sistemas de energia.

PALAVRAS-CHAVES: Seleção de Atributos; Aprendizagem da Máquina; Geração Fotovoltaica; Previsão da Irradiação Solar, Clusterização, Aprendizagem *Ensemble.*

ABSTRACT

Accurate forecasting of solar irradiance is essential for effective management of power systems with significant photovoltaic generation. Machine learning algorithms, which leverage historical data and patterns to make predictions, play a crucial role in this task. One key aspect is the use of ensemble models that combine the predictions of multiple algorithms to improve forecast accuracy and reliability. In this study, ensemble models are utilized to enhance the forecasting performance by aggregating the predictions of different algorithms. Moreover, the paper proposes an ensemble feature selection method, which involves identifying the most relevant input parameters and their related past observations. This approach aims to optimize the input features used by the machine learning algorithms, ensuring that only the most pertinent information is considered for accurate solar irradiance forecasts. By leveraging the strengths of multiple algorithms and selecting the most informative features, the ensemble approach offers a robust framework for improving the accuracy of solar irradiance predictions.

The performance of several machine learning algorithms, including ensemble models, is compared for solar irradiance forecasting on days with different weather patterns using endogenous and exogenous inputs. The algorithms considered are AdaBoost, SVR, RF, XGBT, CatBoost, VOA, and VOWA. The proposed ensemble feature selection relies on the RF, IM, and Relief algorithms. The forecast accuracy is evaluated based on several metrics using a real database of the city of Salvador, Brazil. Different weather forecasts are considered: 1 hour, 2 hours, 3 hours, 6 hours, 9 hours, and 12 hours in advance. Numerical results show that the proposed ensemble feature selection improves forecast accuracy, and that the VOWA model selected with the best-performing algorithms presents forecasts with higher accuracy than the other algorithms at different forecast time horizons. This research demonstrates the effectiveness of ensemble models and feature selection techniques in enhancing solar irradiance forecasting, providing valuable insights for efficient power system management.

KEYWORDS: Ensemble Feature Selection; Machine Learning; Photovoltaic Generation; Solar Radiation Forecasting, Clustering, Ensemble Learning.

1 INTRODUÇÃO

1.1 MOTIVAÇÃO

O mercado mundial de energia ainda é dominado pelos combustíveis fósseis, principal responsável pela emissão de gases poluentes a atmosfera como o dióxido de carbono (CO_2). Além de seus impactos ambientais negativos, os combustíveis fósseis são recursos naturais esgotáveis, o que vem motivando ações de difusão de fontes de energia mais limpas como as energias renováveis em todo o mundo.

Dentre as fontes renováveis, a energia solar fotovoltaica (FV) tem se destacado, tornando-se cada dia mais popular (ALCAÑIZ et al., 2023). De acordo com dados da Agência Internacional de Energia Renovável, no inglês *International Renewable Energy Agency* (IRENA) a energia fotovoltaica continuou liderando globalmente a expansão da capacidade renovável com um incremento de 137 GW (+19%) em 2021, alcançando um total de 854 GW de capacidade e respondendo por 28% da carteira de geração renovável (IRENA, 2022). No Brasil, a energia fotovoltaica tornou-se a terceira fonte com maior participação na matriz elétrica brasileira, com 10,2%, segundo dados da Agência Nacional de Energia Elétrica (ANEEL) e contabilizados pela Associação Brasileira de Energia Solar Fotovoltaica (ABSOLAR) (ABSOLAR, 2022). A energia fotovoltaica soma 21.349 MW de capacidade instalada no País. Nos últimos 5 anos, o crescimento da geração fotovoltaica é constante, o que permitiu alcançar essa posição destacada apenas 5 anos depois da instalação do primeiro parque. Até 2017, essa fonte não integrava a matriz elétrica.

A energia FV é um recurso de energia limpa, abundante, facilmente acessível e renovável que surgiu como uma solução promissora para reduzir o consumo de combustíveis fósseis e as emissões de CO_2 (GABOITAOLELWE et al., 2023). A energia fotovoltaica pode ser facilmente coletada utilizando painéis fotovoltaicos, seja em pequenas instalações no telhado de residências, ou em grandes fazendas solares, podendo ser transformada em eletricidade e utilizada para fornecer energia elétrica ao proprietário da casa ou integrada à rede elétrica. Como a energia fotovoltaica é ambientalmente favorável, muitos governos estão incentivando sua utilização, considerando que a tecnologia associada a ela está melhorando a um ótimo ritmo, com o aparecimento de novos tipos de células que são candidatas a substituir as tradicionais células de silício cristalino. Estes fatos, juntamente com a necessidade global

de reduzir as emissões de gases de efeito estufa, fazem da energia solar uma das energias renováveis mais promissoras (DEVABHAKTUNI et al., 2013).

Apesar da energia FV ter muitas vantagens sobre as demais fontes tradicionais de energia como carvão e gás natural, a produção de energia fotovoltaica é altamente variável e depende da disponibilidade de irradiação solar e de outros fatores meteorológicos. Adicionalmente, a energia fotovoltaica é uma fonte de energia intermitente, visto que só está disponível durante o dia. A Figura 1.1 mostra a irradiação solar global medida em um dia ensolarado e um dia nublado na cidade de Belém, Pará. Nota-se uma redução considerável nos níveis de irradiação no dia nublado. Com isso, a produção de energia a partir do sistema fotovoltaico também é menor. Além disso, podem ocorrer variações bruscas devido a passagem de nuvens. A geração de energia fotovoltaica flutuante pode levar a uma inversão do fluxo de energia com flutuações de tensão e frequência e um desequilíbrio entre a demanda e a oferta de energia. Com a crescente penetração da energia solar distribuída na rede elétrica, a variabilidade inerente representa um desafio para o funcionamento da rede elétrica e deve ser gerenciada para uma operação confiável da rede.

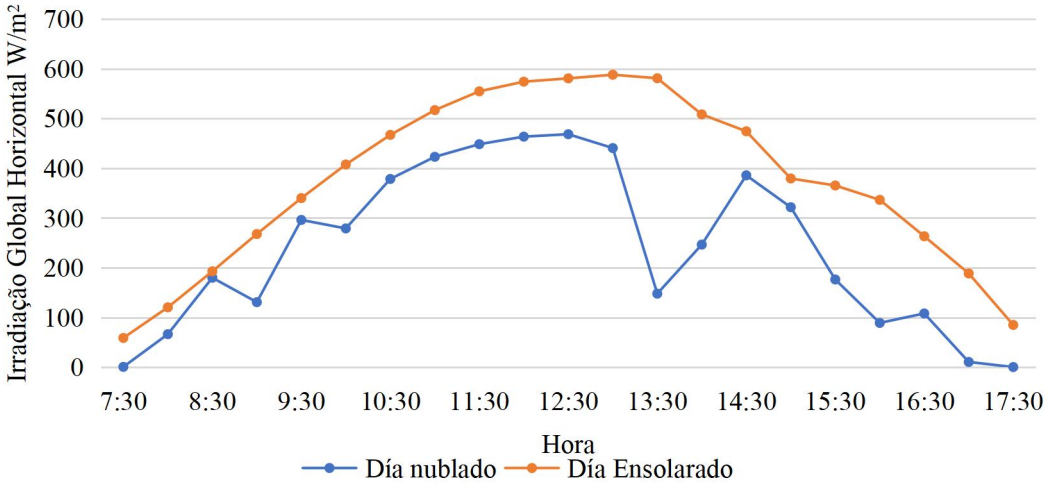


Figura 1.1 – Série cronológica da irradiação global medida na cidade de Belém, Pará .

Fonte: Autor (2023).

A previsão dos níveis de irradiância possibilita a estimação da produção de energia elétrica a partir de fontes fotovoltaicas, sendo essencial para adequada integração da geração FV na rede elétrica. A previsão da geração fotovoltaica em diferentes escalas de tempo é uma

ferramenta poderosa que aproxima a energia fotovoltaica da "despachabilidade" e a torna mais compatível com as operações atuais do sistema de energia (MARTÍN et al., 2010).

Devido à sua capacidade de resolver problemas altamente complexos e aprender relações não lineares entre os dados de entrada e saída, as ferramentas de Inteligência Artificial (IA) se tornaram cada vez mais populares (ALKAHTANI; ALDHYANI; ALSUBARI, 2023; YADAV; CHANDEL, 2014). Os avanços na tecnologia da computação e o potencial dos algoritmos têm ajudado a resolver problemas não apenas na engenharia, assim como em muitas outras áreas, como medicina, finanças e literatura. Por este motivo, as ferramentas da IA estão se tornando cada vez mais utilizadas e podem ser aplicadas a um número cada vez maior de campos ou projetos. A utilização de IA para aumentar a precisão da previsão da irradiação solar pode ajudar a reduzir os problemas de intermitência da geração FV, melhorar a confiabilidade da rede elétrica, manter a qualidade da energia e aumentar a penetração no mercado dos sistemas FV.

1.2 OBJETIVOS

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver um modelo de previsão de irradiação horária global a curto prazo utilizando técnicas de aprendizagem de máquina. Esta metodologia servirá como apoio a previsão da potência gerada pelas usinas fotovoltaicas, minimizando os problemas de intermitência das fontes de energia, permitindo assim uma maior participação da geração de energia fotovoltaica na matriz energética nacional.

1.2.2 Objetivos Específicos

Os objetivos específicos são os seguintes:

1. Comparar o desempenho de diversos algoritmos de aprendizagem de máquina do estado da arte, incluindo um modelo de votação de algoritmos que é pouco utilizado na previsão da irradiação solar;
2. Propor um modelo de seleção de atributos que integre diferentes algoritmos de aprendizagem de máquinas para selecionar as variáveis endógenas e exógenas mais relevantes e seus instantes de atraso;

3. Propor métodos baseados em clusterização para previsão solar que dividem os dias em clusters com características climatológicas semelhantes e, em seguida, construir um modelo separado para cada cluster.
4. Utilizar dados reais de uma cidade do nordeste do Brasil, e analisar a precisão dos algoritmos para diferentes escalas de tempo de previsão através de diversas medidas de desempenho.

1.3 REVISÃO BIBLIOGRÁFICA

Diversos modelos de previsão de irradiação solar têm sido propostos, abrangendo diferentes horizontes de previsão. As previsões de muito curto prazo (de minutos a horas) são úteis para a precificação de eletricidade e monitoramento em tempo real. As previsões de curto prazo (de horas a dias) permitem o compromisso de unidades e programação de despacho. As previsões de médio (de uma semana a um mês) e longo (de um mês a um ano), prazo são importantes para agendar manutenções e planejar a geração e distribuição de eletricidade (DIMD et al., 2022) . Neste trabalho, as metodologias propostas estão direcionadas para as previsões a curto prazo.

Os métodos de previsão de irradiação solar têm sido amplamente abordados na literatura, e podem ser divididos em três categorias: métodos físicos, estatísticos e de IA (INMAN; PEDRO; COIMBRA, 2013; QING; NIU, 2018) . Os modelos físicos fazem previsões baseadas em leis físicas que governam o clima (PELLAND; GALANIS; KALLOS, 2013). Eles usam processamento de imagens de satélite e técnicas de Previsões Numéricas do Tempo, do inglês *Numerical Weather Predictions* (NWP), para prever a intensidade da irradiação solar. Os métodos estatísticos são baseados em séries temporais históricas de dados (COLAK et al., 2015) . Eles são mais simples do que os métodos físicos, porém são frequentemente limitados por suposições sobre normalidade, linearidade, ou dependência de variáveis. Pode-se citar como exemplo os modelos Autoregressivo de Médias Móveis (ARMA), Autorregressivo Integrado de Médias Móveis (ARIMA), bem como Suavização Exponencial (SE). Finalmente, os métodos baseados em IA são orientados por dados e podem aprender relações não lineares entre os dados de entrada e saída (YADAV; CHANDEL, 2014).

Diversos estudos na literatura mostram que os métodos baseados em IA obtém resultados de previsão superiores. Em (KUMAR; AGGARWAL; SHARMA, 2015) , os autores comparam modelos de regressão e Rede Neural Artificial (RNA) para previsão da irradiação solar, e os resultados mostram que a RNA tem desempenho superior. Na referência

(PEDRO; COIMBRA, 2012) , os autores propõem a comparação de diversos modelos tais como o modelo de persistência, ARIMA, k-Vizinhos mais Próximos, do inglês *k-Nearest Neighbours* (kNN), RNA e RNAs otimizadas pelo Algoritmo Genético (AG) sem utilizar entrada exógenas para a previsão da produção de energia solar. Os resultados mostram que os modelos de previsão baseados em RNA têm melhor desempenho do que as outras técnicas de previsão e que melhorias substanciais podem ser alcançadas com a otimização AG dos parâmetros RNA. A referência (DONG et al., 2015) propõe um modelo híbrido para prever a irradiação solar horária baseado em Mapas Auto-organizados do inglês *Self-Organizing Maps* (SOM), Regressão de Vetores de Suporte, do inglês *Support Vector Regression* (SVR) e Otimização por Enxame de Partículas, do inglês *Particle Swarm Optimization* (PSO). Este modelo utiliza apenas a própria serie temporal da irradiação solar horaria. A técnica proposta supera os modelos tradicionais de previsão tais como ARIMA, SE Linear, SE Simples e Passeio Aleatório, do inglês *Random Walk* (RW). Em média, a precisão da previsão do modelo proposto tem 4% menos erro em comparação com a melhor precisão de previsão de todos os outros modelos estatísticos de séries temporais.

Mais recentemente, métodos de Aprendizagem de Máquina (AM), têm sido bastante utilizados para previsão de irradiação solar com excelente desempenho (YADAV; CHANDEL, 2014) . A aprendizagem de máquina é um subcampo da IA capaz de lidar com uma grande quantidade de dados. Os algoritmos de AM comumente encontrados na literatura são: regressão de vetores de suporte, arvores de regressão, floresta aleatória, redes neurais artificiais e aumento de gradiente (VOYANT et al., 2017).

Um modelo de previsão utilizando algoritmos de IA pode ser construído utilizando apenas entradas endógenas, isto é, a própria série temporal de irradiação solar. A referência (RODRÍGUEZ et al., 2022) aplica um modelo de Rede Neural Profunda do inglês *Deep Neural Network* (DNN) para previsão da irradiação solar com um horizonte de tempo de 10 min e utiliza a técnica *wavelet* para o pré-processamento de dados. O modelo de previsão utiliza unicamente observações históricas da irradiação solar, com os resultados mostrando que a abordagem *wavelet* melhora a precisão em 25,59%.

Em (MICHAEL et al., 2022) , os autores propõem um modelo híbrido que combina Rede Neural de Memória de Longo Prazo, do *inglês Long Short Term Memory* (LSTM) e Redes Neurais Convolucionais, do inglês *Convolutional Neural Networks* (CNN) para previsão da irradiação solar de curto prazo. O método utiliza como entrada a série histórica da irradiação solar, e seu desempenho é comparado com outras técnicas de AM como Regressão Linear (RL), SVR e RNA. Os resultados mostram que o modelo híbrido proposto apresenta

desempenho superior quando comparado com os outros algoritmos propostos e os algoritmos padrão CNN e LSTM isolados.

A referência (BOUBAKER et al., 2021) investiga a utilização de diversos modelos de aprendizado profundo para a previsão da irradiação horizontal global tais como LSTM, Bidirecional-LSTM, Unidades Recorrentes Fechadas, do inglês *Gated Recurrent Unit* (GRU), Bidirecional-GRU, CNN de uma dimensão e outras configurações híbridas como CNN-LSTM. Apenas valores históricos da própria irradiação solar foram usados para construir os modelos, enquanto parâmetros climáticos adicionais como temperatura do ar, velocidade do vento, direção do vento, pressão atmosférica e umidade relativa do ar não foram considerados.

Alguns pesquisadores têm se concentrado no desenvolvimento de novos modelos híbridos utilizando apenas entradas endógenas. Entretanto, existem uma alta correlação entre a irradiação solar e alguns parâmetros meteorológicos que conduziu outros autores a utilizar tanto entradas endógenas quanto exógenas para melhorar a precisão da previsão da irradiação solar. As entradas exógenas podem incluir parâmetros meteorológicos tais como temperatura do ar, umidade, velocidade do vento, direção do vento e pressão atmosférica. Desta forma, os modelos de previsão podem ser construídos tanto com entradas endógenas quanto exógenas.

Em (WENTZ et al., 2022), os autores propõem um modelo de previsão da irradiação solar usando LSTM para três horizontes de previsão (1, 15, e 60 minutos). Dois conjuntos de atributos de entrada são considerados, um com sete dados meteorológicos, e outro com um conjunto de dados reduzido com apenas três dados meteorológicos. Os resultados mostram melhor precisão de previsão da abordagem LSTM quando comparada à RNA, e nenhuma diferença significativa entre os dois conjuntos de dados de atributos de entrada. Não obstante, uma metodologia de seleção não foi aplicada adequadamente para selecionar o conjunto reduzido de atributos de entrada.

Em (MASSAOUDI et al., 2021) , os autores propõem um modelo conjunto para previsão de geração fotovoltaica a curto prazo combinando Máquinas de Aprendizado Extremo, do inglês *Extreme Learning Machines* (ELM), Arvores Extremamente Aleatórias, do inglês *Extra Trees* (ET), kNN, Floresta Mondrian, do inglês *Mondrian Forest* (MF) e Rede de Crenças Profundas, do inglês *Deep Belief Networks* (DBN). Vários dados meteorológicos são usados como entrada do modelo de previsão tais como irradiação horizontal global, irradiação horizontal difusa, umidade relativa, direção do vento, tempo de amostragem, temperatura e dados históricos de potência das usinas. Os resultados indicam um desempenho superior do modelo conjunto em relação aos modelos de AM existentes. No entanto, os autores não adotaram uma metodologia de seleção de entradas. Em (JAYALAKSHMI et al.,

2021), os autores propõem um modelo multi-escala para previsão de irradiação solar usando LSTM com Otimização de Enxame de Frango*, do inglês Chicken Swarm Optimization* (CSO) e Otimização do Lobo Cinzento, do inglês *Gray Wolf Optimization* (GWO). Os autores incentivam a adição de atributos exógenos como entradas, porém limitadas à temperatura do ar, altura do sol e velocidade do vento.

Em (MAHMUD et al., 2021), os autores propõem a previsão de geração de energia fotovoltaica baseada em AM, tanto a curto como a longo prazo. Algoritmos como RL, Regressão Polinomial (RP), Arvores de Decisão, do inglês *Decision Trees* (DT), SVR, Floresta Aleatória, do inglês *Random Forest* (RF), LSTM e Perceptron Multicamadas, do inglês *Multi-Layer Perceptron* (MLP). Alguns parâmetros meteorológicos são utilizados no modelo de previsão e são considerados diferentes horizontes de tempo previstos (24 horas, 1 semana e 1 ano). A seleção das entradas é feita de forma intuitiva, e os resultados mostram que o método RF tem um desempenho melhor.

Na referência (RANA; KOPRINSKA; AGELIDIS, 2016), os autores desenvolvem modelos de previsão separados para diferentes tipos de dias os quais são determinados usando algoritmos de clusterização que identificam padrões meteorológicos. Como modelos de previsão, os autores utilizaram conjuntos de RNAs, treinados para prever a saída de energia fotovoltaica para um determinado dia com base nos dados meteorológicos. Os resultados mostraram que sua abordagem do conjunto de RNA supera os outros métodos utilizados para comparação. Entretanto, os autores não empregaram uma metodologia de seleção de atributos.

As pesquisas anteriores comprovaram que a inclusão de entradas exógenas melhora os resultados das previsões. Existem poucos estudos que aplicam uma metodologia de seleção de atributos para a previsão da irradiação solar. A referência (CASTANGIA et al., 2021) investiga a eficácia do uso de insumos exógenos para realizar previsões da irradiação solar a curto prazo com vários modelos de AM. Os autores aplicaram as seguintes técnicas de seleção de atributos: Correlação, Informação Mútua (IM), Seleção Sequencial Progressiva, do inglês *Sequential Forward Selection* (SFS), Seleção Sequencial Regressiva, do inglês *Sequential Backward Selection* (SBS), Encolhimento e Seleção pelo Menor Valor Absoluto, do inglês *Least Absolute Shrinkage and Selection Operator* (LASSO), e RF. Os resultados mostram que as entradas exógenas melhoram significativamente o desempenho das previsões. Em (TAO et al., 2021), os autores propõem um modelo híbrido baseado em *BiasCompensation*–LSTM para realizar previsões de potência FV com uma seleção de entradas exógenas baseada em SFS e comparam com a seleção baseada no algoritmo da Máquina de Impulso Gradiente Leve, do inglês *Light Gradient Boosting Machine* (LightGBM). Os resultados experimentais

mostram que o novo método de seleção de entradas pode melhorar a precisão da previsão. A referência (LONG; ZHANG; SU, 2014) compara o desempenho de RNAs, SVR, kNN e RL para prever a saída de energia FV horária com 1 a 3 dias de antecedência usando tanto a saída de energia anterior quanto os dados meteorológicos. Eles aplicaram um algoritmo de seleção de atributos baseado em clusters e validação cruzada para encontrar um subconjunto de atributos importantes. Durante este procedimento, eles incluíram observações passadas, porém apenas da variável endógena e se limitaram a 3 observações passadas.

Entretanto, a maioria das referências citadas seleciona as entradas exógenas através de algoritmos limitados como o coeficiente de correlação de Pearson, que identifica apenas relações lineares entre atributos, ou experimentando intuitivamente diferentes combinações de atributos de entrada e selecionando a que gera menor erro de previsão. Outra questão importante refere-se à seleção dos instantes de atraso (observações passadas), que também tem grande influência na eficácia da previsão, e que usualmente não é realizada (SURAKHI et al., 2021). Deve-se empregar uma metodologia de seleção de entradas tanto para os atributos endógenos como exógenos e suas correspondentes observações passadas.

Outros estudos recentes concentraram-se no desenvolvimento de modelos conjuntos, também conhecidos como *ensemble models* em inglês, que são cada vez mais utilizados para melhorar o desempenho dos algoritmos de previsão (GUERMOUI et al., 2020). O principal conceito dos modelos *ensemble* é treinar membros do *ensemble* (aprendizes base) e combinar suas previsões em uma única saída para obter um melhor desempenho do modelo (RAHIMI et al., 2023). Eles podem ser combinados com uma ampla gama de técnicas, tais como *bagging* (agregação por amostragem, em português), *boosting* (impulso, em português), *stacking* (empilhamento, em português) ou *voting* (votação, em português) .

A referência (ABDELLATIF et al., 2022) propõe um *ensemble model* baseado na técnica *stacking* para a previsão de potência FV para o dia seguinte. Os seguintes algoritmos de AM são combinados: RF, Árvores de Aumento de Gradiente Extremo, do inglês *Extreme Gradient Boosting Trees* (XGBT) e Boosting Adaptável, em inglês *Adaptive Boosting* (AdaBoost), finalmente são integrados por meio do algoritmo ET. O modelo utiliza como entrada a serie temporal da potência FV. O desempenho do modelo proposto foi comparado com o desempenho de outros *ensemble models*, e os resultados mostraram que ele apresentou um desempenho superior aos outros.

Os autores em (KUMARI; TOSHNIWAL, 2021) propõem um modelo de previsão de irradiação solar horaria e utilizam a técnica *stacking* para combinar dois algoritmos base, XGBT e DNN através de regressão *ridge*. O conjunto de dados de entrada inclui parâmetros

meteorológicos e índice *clear-sky*, e a seleção de atributos é feita implicitamente usando o conceito de ganho de informação. Os resultados mostram a superioridade do modelo.

Em (PARK et al., 2020) é proposto um modelo de previsão para irradiação solar global baseado no algoritmo LightGBM, que é uma técnica de aprendizagem *ensemble* que integra arvores de decisão através da técnica *boosting*. Os autores utilizaram como entrada a data/hora, dados meteorológicos e o histórico da irradiação solar global sem utilizar alguma técnica de seleção. Os resultados demonstraram que o modelo proposto pode alcançar melhor desempenho preditivo do que outros métodos baseado em árvores e de aprendizado profundo.

Em (LEE et al., 2020), os autores investigam o desempenho de *ensemble models* para a previsão da irradiação solar tais como *boosted trees, bagged trees*, RF e RF generalizado que são baseados nas técnicas de *bagging* e *boosting*. Seis variáveis meteorológicas são selecionadas com base em outros estudos, e o desempenho desses *ensemble models* foi comparado com outros dois métodos de previsão comumente conhecidos: Processo de Regressão Gaussiano e SVR. Os resultados mostraram que os *ensemble models* oferecem um desempenho de previsão superior em comparação com os regressores individuais.

Embora existam vários artigos que investigam a melhoria nas previsões de irradiação solar utilizando *ensemble models* baseados em algoritmos de AM, eles utilizam as técnicas *bagging*, *boosting* e *stacking*, enquanto a técnica *voting* não é utilizada.

Este trabalho procura atender esta lacuna na literatura propondo um método de previsão baseado em modelos de AM que incorporem tanto entradas exógenas quanto endógenas, selecionadas através de um método *ensemble* para selecionar não apenas as variáveis de entrada como também seus instantes de atraso, e depois combinar estes modelos individuais através da técnica de *voting*.

1.4 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação de mestrado está organizada da seguinte forma:

O Capítulo 1 apresenta a problemática e justificativa deste trabalho, bem como o objetivo proposto e suas contribuições, além de uma revisão bibliográfica.

O Capítulo 2 descreve a geração solar fotovoltaica, as tecnologias utilizadas na geração assim como a situação atual no mundo e no Brasil.

O Capítulo 3 apresenta a teoria de aprendizagem de máquina, os conceitos e procedimentos de pré-processamento de dados, técnicas utilizadas para seleção de atributos.

O Capítulo 4 apresenta a metodologia de previsão proposta, incluindo a descrição do banco de dados, método de seleção de atributos proposto, além das medidas utilizadas para avaliação dos algoritmos.

No Capítulo 5 são apresentados os resultados obtidos por meio das simulações computacionais e avaliação dos resultados.

Por fim, no Capítulo 6, as conclusões são apresentadas, além de sugestões para trabalhos futuros.

2 GERAÇÃO SOLAR FOTOVOLTAICA

2.1 ORIGEM DA ENERGIA SOLAR FOTOVOLTAICA

A energia proveniente do sol vem sendo utilizada pelo homem ao longo de toda sua história, suprindo necessidades básicas de aquecimento, iluminação e alimentação. No entanto, o uso do sol para a produção de eletricidade é relativamente recente e pode ser feita basicamente de duas formas: de modo direto através dos sistemas fotovoltaicos que convertem a luz solar em eletricidade, e de modo indireto através de sistemas heliotérmicos que produzem energia pelo aquecimento de um fluido a partir do calor acumulado da luz solar.

Este trabalho tem como foco o aproveitamento da energia solar de forma direta através de sistemas fotovoltaicos, que teve início em 1839 quando Edmond Becquerel descobriu o efeito fotovoltaico observando o aparecimento de uma diferença de potencial nos terminais de uma célula eletroquímica causada pela absorção de luz. Mais tarde, em 1877, Adams e seu aluno Richard Day desenvolveram a primeira célula solar, um filme de selénio depositado num substrato de ferro em que um filme de ouro muito fino servia de contacto frontal. Seguidamente começaram a estudar outros materiais e foi então, em 1940, que Russell Ohl desenvolveu no *Bell Labs* a primeira célula solar de silício, chamada naquela época, dispositivo elétrico sensível à luz. Considera-se que a era moderna da energia solar teve início em 1954 quando Calvin Fuller, Gerald Pearson e Daryl Chapin desenvolveram o processo de dopagem do silício dando passo a primeira célula FV de silício e foi formalmente apresentada na reunião anual da *National Academy of Sciences*, em Washington no dia 25 de Abril de 1954. No ano seguinte a célula de silício teve sua primeira aplicação como fonte de alimentação de uma rede telefónica em *Americus*, na Geórgia. Foi apenas em 1956 com o crescimento da área de eletrônica que deu-se início a produção industrial e comercialização dos dispositivos fotovoltaicos, dando um passo para maior aproveitamento da energia solar

em âmbitos de geração de eletricidade (MARQUES LAMEIRINHAS; TORRES; DE MELO CUNHA, 2022).

2.2 GERAÇÃO SOLAR FOTOVOLTAICA NO MUNDO

Segundo dados da Agencia Internacional de Energia, do inglês *International Energy Agency* (IEA), a geração de energia FV vem apresentando crescimento exponencial, atingindo um aumento recorde mundial de 179 TWh em 2021 com um crescimento de 22%. A Figura 2.1 ilustra a evolução anual da capacidade total instalada da geração fotovoltaica. A energia solar FV mostrou-se resiliente frente às interrupções da Covid-19, aos impasses na cadeia de abastecimento e aos aumentos de preços das matérias-primas experimentados em 2021 (IEA, 2022) . Atualmente, a energia FV é a terceira maior tecnologia de eletricidade renovável depois da energia hidrelétrica e eólica.

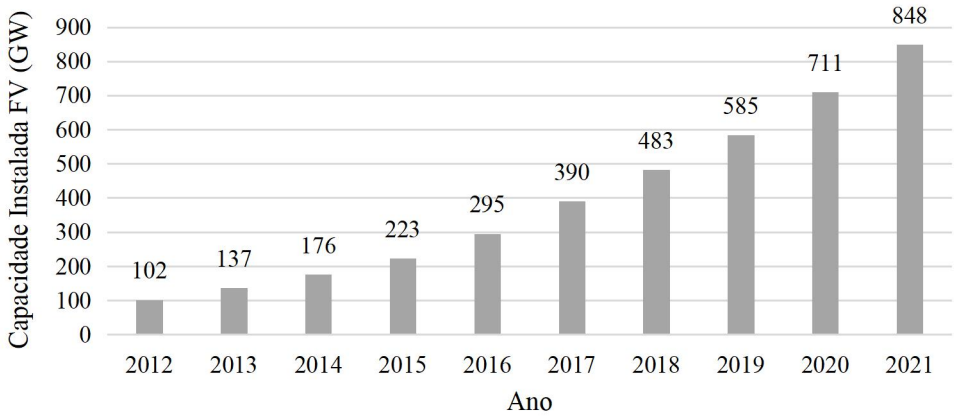


Figura 2.1 – Evolução Global da Capacidade FV Instalada.

Fonte: IRENA (2022)

Em 2021, a China foi responsável por cerca de 38% do crescimento mundial da geração solar FV, e vem liderando o ranking de capacidade instalada de geração fotovoltaica. Em 2021, o país alcançou 306,403 GW, seguida pela União Europeia com 159,535 GW, EUA com 93,713 GW, Japão com 74,191 GW e Índia com 49,342 GW. Na União Europeia, a Alemanha lidera com 58,726 GW, seguida pela Itália com 22,692 GW, França com 14,709 GW, Holanda com 14,249 GW e Espanha com 13,648 GW (IRENA, 2022).

2.3 GERAÇÃO SOLAR FOTOVOLTAICA NO BRASIL

O Brasil está situado quase que totalmente na região limitada pelos Trópicos de Câncer e de Capricórnio. Esta posição geográfica favorece elevados índices de incidência da radiação solar em quase todo o território nacional, inclusive durante o inverno, o que confere ao país condições vantajosas para o aproveitamento energético do recurso solar (TOLMASQUIM, 2016). A Figura 2.2 apresenta o mapa com estimativas de irradiação global horizontal obtidas pelo modelo BRASIL-SR (BUENO PEREIRA et al., 2017). As regiões Nordeste, Centro-Oeste e Sudeste do Brasil apresentam os maiores rendimentos médios anuais. No entanto, considerando os valores de irradiação solar global incidente em qualquer região do território brasileiro que variam entre 4.200 kWh/m^2 e 6.700 kWh/m^2, praticamente todo território brasileiro é elegível ao aproveitamento deste recurso. Como referência, esses níveis são superiores aos estimados na maioria dos países da União Europeia, como Alemanha (900-1.250 kWh/m^2), França (900-1.650kWh/m^2) e Espanha (1.200-1.850 kWh/m^2) (BUENO PEREIRA et al., 2006). Nesse cenário, esta modalidade de geração FV tem participação de destaque na matriz energética nacional.

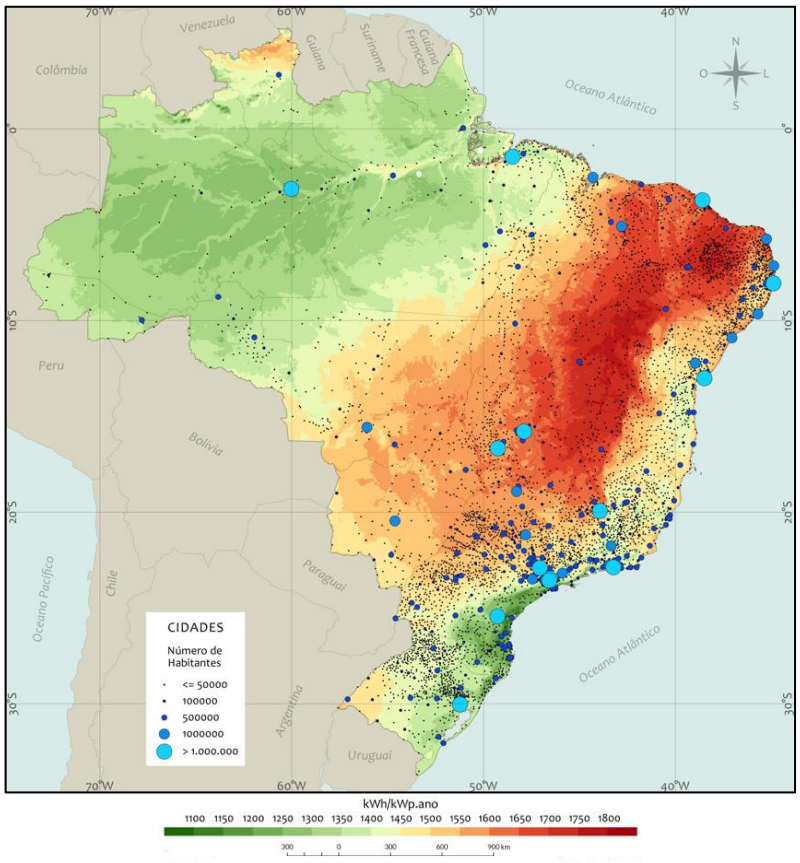


Figura 2.2 – Potencial de geração solar FV-Rendimento energético anual.

Fonte: PEREIRA (2017).

No Brasil, o desenvolvimento de módulos fotovoltaicos deu-se início nos anos 50 no Instituto Nacional de Tecnologia e no Centro Tecnológico de Aeronáutica. Em 1994, o Centro de Referência para as Energias Solar e Eólica Sérgio de Salvo Brito (CRESESB) foi criado por meio de um Convênio entre o Centro de Pesquisas de Energia Elétrica (CEPEL) da Eletrobras e o Ministério de Minas e Energia (MME), com apoio do Ministério da Ciência e Tecnologia. Em 2001 uma iniciativa do governo federal, caracterizada pela criação do Fundo Setorial de Energia (CT-ENERG), resultou em um crescimento das atividade em energia solar FV e na formação de grupos de pesquisa e programas de pós-graduação (PINHO; GALDINO, 2014).

Do ponto de vista do marco regulatório e dos projetos de demonstração, a energia solar começou no Brasil com foco em políticas públicas e projetos-piloto de eletrificação rural em sistemas isolados, com a criação do Programa de Desenvolvimento Energético para Estados e Municípios (PRODEEM) do MME entre 1994 e 2003. Durante este período,

milhares de sistemas fotovoltaicos com armazenamento de energia foram instalados nas partes mais remotas do Brasil, especialmente nas regiões Norte e Nordeste. Em 2003, o PRODEEM tornou-se parte do Programa Luz para Todos (LpT), um programa que priorizava ações para universalizar o acesso à energia. Durante este período, de meados dos anos 2000 até o início de 2013, foram realizados numerosos projetos-piloto com universidades e organizações não governamentais. Até então, a energia solar ainda não era um grande negócio (PINHO; GALDINO, 2014).

Desde 2012, o marco regulatório favorável à energia solar FV deu início a uma nova era para esta tecnologia devido a Resolução Normativa n° 482 publicada pela ANEEL, que estabeleceu condições gerais para o acesso de microgeração e minigeração distribuída aos sistemas de distribuição e o sistema de compensação de energia elétrica. O sistema de compensação de energia permite que unidades consumidoras com geração de até 1 MW de capacidade com base em energia hidráulica, solar, eólica, biomassa ou cogeração qualificada troque energia com a distribuidora local (VIAN et al., 2021). Se ao fim do mês a geração for maior que o consumo, o saldo restante, chamado crédito de energia, pode ser usado para abater o consumo em algum mês subsequente, restando ao consumidor somente o pagamento da tarifa básica. Se o consumo for maior que a geração, o consumidor paga a diferença entre a energia total consumida e a gerada.

A utilização da energia FV vem crescendo no Brasil nos últimos anos. Em janeiro 2018 a energia solar FV atingiu seu primeiro GW de capacidade instalada no Brasil, e um ano mais tarde em janeiro 2019 a geração centralizada solar FV atingiu 2 GW de capacidade instalada acumulada, tornando-se a 7ª maior fonte de energia elétrica do Brasil. Em novembro de 2022, a potência total instalada de geração FV atinge nível superior a 21,3 GW, representando 10,2% da matriz elétrica brasileira, passando a ser a terceira maior fonte de energia no Brasil conforme a Figura 2.3. Desses total, o 69% correspondem a geração centralizada e o restante 31% a geração distribuída (ABSOLAR, 2022).

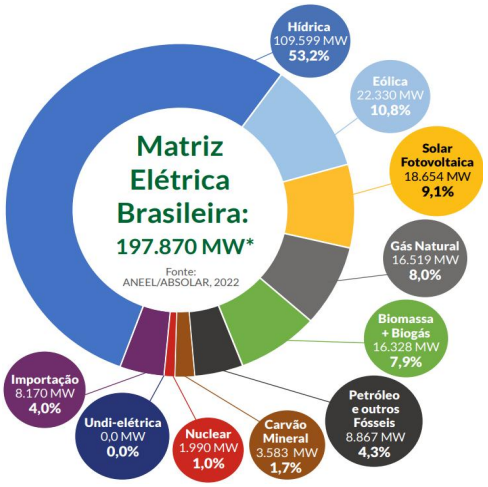


Figura 2.3 – Matriz elétrica brasileira.

Fonte: ABSOLAR (2022)

Atualmente, existem 1.425.747 projetos operacionais de geração distribuída e 99% desses projetos são instalações solares fotovoltaicas (ABSOLAR, 2022) . Sistemas de microgeração e minigeração distribuída solar FV são implantados em residências, comércios, indústrias, propriedades rurais e prédios públicos e os estados de maior destaque são Minas Gerais (14,8%), São Paulo (12,8%) e Rio Grande do Sul (10,9%). São 1.536.897 unidades consumidoras recebendo créditos pelo Sistema de Compensação de Energia Elétrica (ABSOLAR, 2022) . A geração distribuída FV no Brasil está distribuída conforme a Figura 2.4.

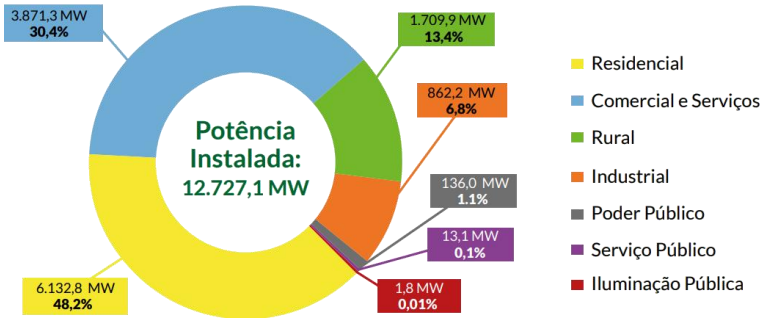


Figura 2.4 – Distribuição da potência instalada FV.

Fonte: ABSOLAR (2022)

Em termos de geração FV centralizada, os estados com maior destaque são Minas Gerais com 1.648,5 MW de potência em operação, Bahia com 1.356,4 MW e Piauí com 1.242,

7 MW. Atualmente no Brasil existem 17603 usinas fotovoltaicas em operação (ANEEL, 2022). A maior usina de energia solar do Brasil e da América do Sul, mostrada na Figura 2.5, está em construção e localiza-se na cidade de São Gonçalo do Gurguéia, Piauí, composta por mais de 2,2 milhões de painéis solares. Atualmente a usina conta com 608 MW em operação e 256 MW em construção (ENEL, 2022).

Figura 2.5 – Usina solar São Gonçalo.

Fonte: ENEL (2022)

2.4 PRINCÍPIO DE FUNCIONAMENTO DOS SISTEMAS FOTOVOLTAICOS

A radiação solar, frequentemente chamada de recurso solar ou combustível da energia FV é a energia radiante, em particular a energia eletromagnética, emitida pelo sol que chega à terra e como ela é do tipo eletromagnética, se propaga na velocidade da luz (PINHO et al., 2008) . Quando a radiação solar entra na atmosfera terrestre, parte da energia incidente é refletida, espalhada ou absorvida pelas moléculas de ar, nuvens e partículas em suspensão. A radiação que não é refletida, espalhada ou absorvida e atravessa diretamente em linha reta desde o disco solar até a superfície terrestre é denominada de radiação direta. A radiação que é espalhada e que chega à superfície da terra é chamada de radiação difusa. A parte da radiação que chega à superfície da terra e é refletida pelo solo é denominada de albedo. A radiação total obtida destas três componentes é chamada de global (WALD, 2021) . Uma grandeza empregada para quantificar a radiação solar é a irradiância, geralmente chamada também de irradiação, expressa na unidade de W/m^2. Trata-se de uma unidade de potência por área.

O recurso solar não pode ser considerado constante devido a sua variação ao longo do dia, ano e localização, o que o torna extremamente variável. Grande parte destas variações se

deve à geografia terrestre, os movimentos astronômicos de rotação e translação e fenômenos climáticos. Por outro lado, além destas variações a irradiação solar varia de acordo com a posição terrestre e o consequente o ângulo de incidência dos raios solares.

A forma mais comum de aproveitar a energia solar é transformá-la diretamente em eletricidade através dos sistemas fotovoltaicos, utilizando painéis que consistem em um arranjo em série de pequenas células fotovoltaicas que convertem a luz solar diretamente em eletricidade (THORPE, 2017). Para compreender como funciona um sistema fotovoltaico, é preciso primeiro entender como funciona uma célula FV.

O principal elemento de um sistema fotovoltaico é a célula FV que, através do efeito fotovoltaico, converte a energia solar em energia elétrica. O efeito fotovoltaico acontece em materiais semicondutores, caracterizados por possuírem condutividade intermediaria entre condutores, na prática são usados principalmente materiais como o silício, que normalmente não conduzem eletricidade, mas sob certas circunstâncias podem fazê-lo.

Uma célula solar é um sanduíche de duas camadas diferentes de silício que foram submetidas ao processo de dopagem. A camada inferior é dopada para que tenha menos elétrons, chamada de tipo-P ou positiva. A camada superior é dopada ao contrário, de modo que tem muitos elétrons, chamada de tipo-N ou negativa (PINHO et al., 2008). Visualiza-se na Figura 2.6 um modelo de uma célula FV.

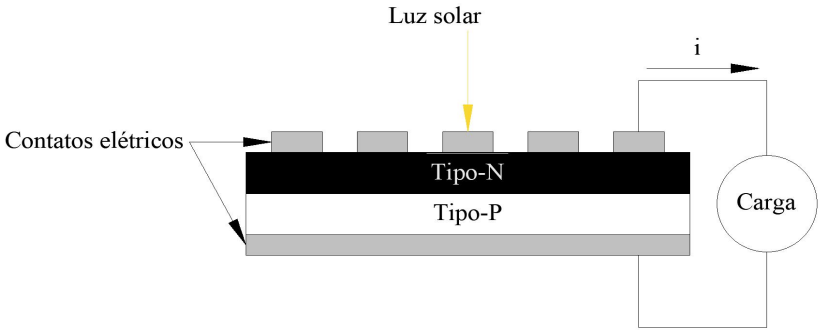


Figura 2.6 – Modelo de uma célula FV.

Fonte: Autor (2023).

Quando a radiação solar, sob a forma de fótons, quando a mesma atinge a célula FV ocorrerá a geração de pares elétron-lacuna. Cada fóton com energia suficiente geralmente libera exatamente um elétron, resultando em um furo livre também. Se isto acontecer suficientemente perto do campo elétrico, ou se o elétron livre e o furo livre entrarem em sua faixa de influência, o campo enviará o elétron para o lado N e o furo para o lado P, gerando

assim, uma corrente através da junção. Este deslocamento de cargas dá origem uma tensão elétrica nos terminais da célula (CRESESB, 2017).

A combinação das células fotovoltaicas em série e/ou paralelo formam um painel ou módulo fotovoltaico, que produz tensão e corrente para o consumo de energia, portanto, a fabricação busca atender os valores de tensão e corrente de projeto, variando a combinação e a quantidade de células (SAMPAIO; GONZÁLEZ, 2017). A corrente elétrica produzida pelos módulos é contínua, sendo necessária a adaptação da energia produzida para os padrões da rede através de um inversor. Na Figura 2.7 é apresentada a configuração de um arranjo solar fotovoltaico.

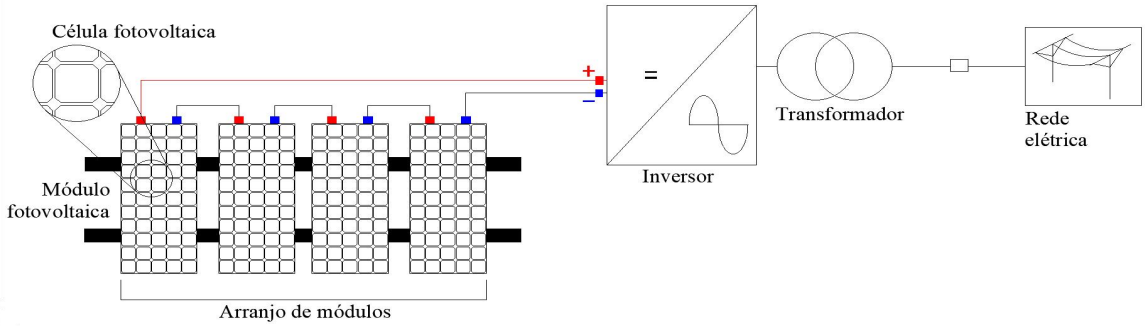


Figura 2.7 – Configuração típica de uma geração solar FV.

Fonte: Autor (2023).

A operação dos módulos FV pode ser descrita pelas curvas I-V (corrente x tensão) e P-V (potência x tensão) mostradas na Figura 2.8. A interconexão de várias células PV em série ou em paralelo para formar painéis solares aumenta a tensão e/ou corrente geral, porém não altera a forma das curvas I-V e P-V. Estas curvas contêm três pontos significativos: ponto A - corrente de curto-circuito; ponto B - potência máxima; ponto C - tensão de circuito aberto.

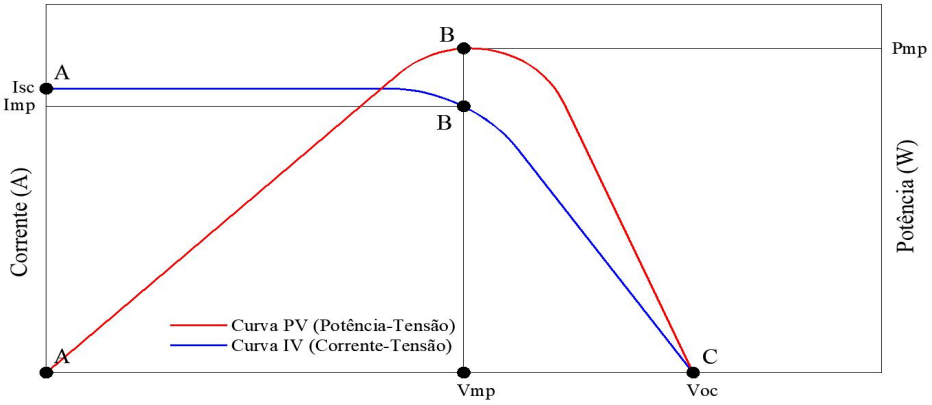


Figura 2.8 – Curvas corrente x tensão (I-V) e potência x tensão (P-V).

Fonte: Autor (2023).

Onde:

Isc = corrente de curto-circuito (em inglês, *short circuit current*);
Voc = tensão de circuito aberto (em inglês, *open circuit voltage*);
Imp = corrente de máxima potência (em inglês, *maximum power current*);
Vmp = tensão de máxima potência (em inglês, *maximum power voltage*);
Pmp = potência no ponto de máxima potência (em inglês, *maximum power point*).

O ponto A de corrente de curto-circuito caracteriza o valor máximo de corrente que o painel pode alcançar, isto ocorre quando os terminais do módulo estão em curto-circuito. Neste ponto, o valor da tensão é zero. O ponto B é o de máxima potência, este é o ponto desejável de operação porque caracteriza a maior geração de energia e é o produto entre Imp e Vmp, podendo ser observadas no "joelho" da curva I-V ou no pico da curva P-V. O ponto C de tensão de circuito aberto caracteriza a tensão máxima do painel, que ocorre quando a corrente é zero, esta tensão é obtida para o painel operando com os terminais abertos (VILLALVA; GAZOLI, 2012).

Normalmente esses dados são encontrados em manuais de fabricantes (tensão e corrente de máxima potência, ponto de máxima potência, tensão de circuito aberto e corrente de curto-circuito, eficiência do módulo, temperatura de operação, tensão máxima do sistema) e são valores levantados em Condições de Teste Padrão, do inglês *Standard Testing Conditions* (STC), com irradiância igual a 1000 W/m² e temperatura do módulo 25 ºC.

A curva I-V e P-V de um módulo pode sofrer alterações devido à influência das variáveis climáticas tais como a irradiação solar e a temperatura de operação. A corrente é sensível a variações nos níveis de irradiação no modulo: quanto mais baixos forem os valores

de irradiação, mais baixos serão os valores de corrente de curto-circuito e consequentemente os valores de potência do modulo. A tensão permanece quase constante. A Figura 2.9 apresenta essas variações nas curvas I-V e P-V.

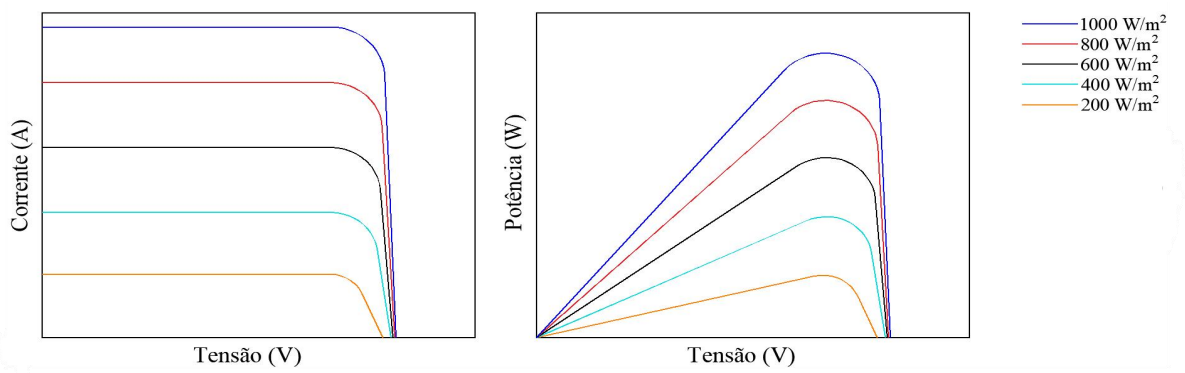


Figura 2.9 – Curvas I-V e P-V para distintos valores de irradiação.

Fonte: Autor (2023).

Desta forma, o sombreamento tem um impacto fortemente prejudicial sobre a geração de energia FV. O sombreamento caracteriza a redução da incidência da radiação solar no sistema, afetando negativamente o nível de geração. Em função disso, pode-se ver como as variáveis climáticas influenciam os níveis de geração de um sistema fotovoltaico. Considerando que o comportamento de tais variáveis é caótico, infere-se que a geração de energia FV também apresenta um comportamento volátil fortemente dependente da irradiação solar do local de instalação.

2.5 CONSIDERAÇÕES FINAIS

A geração de energia solar FV é uma alternativa interessante para a produção de eletricidade e, devido às altas taxas de irradiação solar, conseguiu estabelecer sua posição na matriz energética brasileira ao longo dos anos. Este capítulo apresentou um breve histórico sobre o uso da irradiação solar para a produção de eletricidade, tanto no mundo quanto no Brasil, além de conceitos e princípios de funcionamento das células fotovoltaicas, que são os principais componentes do sistema fotovoltaico. Também foram apresentadas curvas de funcionamento dos sistemas fotovoltaicos em condições padrão e sob diferentes níveis de irradiação solar, mostrando a dependência das condições climáticas e geográficas.

O Brasil é privilegiado geograficamente, com altos índices de irradiação solar, e o cenário atual da energia solar fotovoltaica no país é promissor. No entanto, a característica de intermitência dessa forma de geração, devido à dependência dos níveis de irradiação solar devido a questões climáticas e geográficas, compromete a energia produzida pelas usinas fotovoltaicas, afetando a qualidade da energia injetada na rede e tornando impossível garantir a supressão da demanda.

Realizar previsões de irradiação solar possibilita otimizar o uso de fontes renováveis, como a energia solar fotovoltaica, maximizando sua geração e reduzindo a dependência de fontes não renováveis. Além disso, as previsões permitem melhorar a estabilidade da rede elétrica, equilibrando a oferta e a demanda de energia. Isso contribui para evitar oscilações bruscas e manter a qualidade do fornecimento elétrico. A redução de custos operacionais é outra vantagem, pois as previsões permitem um despacho de energia mais eficiente e a programação do uso de energia em momentos de maior geração solar. Isso reduz custos e evita o acionamento de fontes mais caras durante períodos de menor geração solar. Além disso, as previsões de irradiação solar auxiliam no planejamento e gestão da demanda, otimizando o aproveitamento da energia e contribuindo para a transição energética para uma matriz mais sustentável.

Devido à necessidade de garantir uma operação segura e planejada do sistema de energia, foram desenvolvidos métodos da inteligência artificial, especificamente do campo da aprendizagem de máquinas, a fim de aplicá-los corretamente para fazer previsões de níveis de irradiação solar em diferentes escalas de tempo. No capítulo seguinte, serão apresentados conceitos importantes de aprendizagem de máquinas, tratamento do banco de dados, bem como a apresentação dos algoritmos utilizados nesta dissertação.

3 APRENDIZAGEM DE MÁQUINA

3.1 CONSIDERAÇÕES INICIAIS

O aprendizado de máquinas é um subcampo da IC, que explora o estudo e a construção de algoritmos que podem adquirir conhecimentos a partir de dados e fazer previsões ou decisões. É ciência e a arte de programar computadores para que eles possam aprender com os dados (GÉRON, 2019) . Este processo de aprendizagem pode ser supervisionado ou não supervisionado.

O aprendizado supervisionado utilizada um conjunto de dados rotulados no treinamento dos algoritmos. O aprendizado supervisionado pode ser classificado em dois tipos de problemas: classificação e regressão. Os problemas de classificação utilizam um algoritmo para atribuir dados de teste em categorias específicas com precisão. A regressão usa um algoritmo para entender a relação entre variáveis dependentes e independentes. Os modelos de regressão são úteis para prever valores numéricos com base em diferentes pontos de dados. Exemplos de algoritmos supervisionados são: regressão linear e logística, floresta aleatória, impulso adaptativo, impulso extremo de gradiente, máquina de vetores de suporte e redes neurais.

Por outro lado, o processo de aprendizagem não supervisionado não tem valores de saída conhecidos e o algoritmo tem que descobrir por si só estruturas nos dados. Os modelos de aprendizagem não supervisionados são utilizados para clusterização e associação. Clusterização é uma técnica de mineração de dados para agrupar dados não rotulados com base em suas semelhanças ou diferenças. A associação usa regras diferentes para encontrar relações entre variáveis em um dado conjunto de dados. Exemplos incluem: *K-means*, agrupamento hierárquico e algoritmo Apriori.

Para simplificar, no aprendizado supervisionado o modelo aprende a partir de resultados pré-definidos, utilizando os valores passados da variável alvo para aprender quais devem ser seus resultados de saída e no aprendizado não supervisionado não existem resultados pré-definidos para o modelo utilizar como referência para aprender.

Os algoritmos apresentados para previsão da irradiação solar pertencem ao grupo de aprendizagem supervisionado. Isto porque estão disponíveis dados globais de irradiação solar que podem ser usados como uma saída conhecida e estes algoritmos visam modelar uma função que pode prever uma saída a partir de entradas.

3.2 PREPARAÇÃO DOS DADOS

As bases de dados do mundo real são suscetíveis a dados ruidosos, ausentes ou inconsistentes devido ao seu grande tamanho, sua provável proveniência múltipla e procedimentos de coleta que muitas vezes são mal controlados, resultando em algum nível de erro, tais como valores ausentes e fora da faixa de valores esperados. Estes podem ser erros humanos ou erros nos sistemas de registro. Se os dados não forem previamente cuidadosamente examinados, se pode produzir resultados enganosos, levando a decisões incorretas e análises não confiáveis. Portanto, antes de processar os dados com algoritmos de

aprendizagem de máquina, é importante garantir que eles sejam o mais precisos e consistentes possível.

O pré-processamento é feito principalmente para verificar a qualidade dos dados (GARCÍA; LUENGO; HERRERA, 2015). Há várias técnicas de pré-processamento de dados. A limpeza de dados pode ser aplicada para remover ruídos e corrigir inconsistências nos dados. A integração de dados funde dados de múltiplas fontes em um armazenamento de dados coerente. A redução de dados pode reduzir o tamanho dos dados, agregando, eliminando variáveis redundantes, ou agrupando. As transformações de dados podem ser aplicadas para alterar a escala de dados dentro de uma faixa menor (HAN; KAMBER; PEI, 2011). O resultado esperado das tarefas de pré-processamento de dados é um conjunto final de qualidade, que pode ser considerado correto e útil para os processos posteriores tais como o treinamento de algoritmos de aprendizado de máquina. A qualidade dos dados é a medida da adequação de um conjunto de dados para atender a seu propósito específico. As dimensões mais comuns na qualidade dos dados são completude, atualidade e acurácia, seguidas de consistência e acessibilidade (CICHY; RASS, 2019):

1. Completude: o grau em que os dados são suficientemente amplos, profundos e abrangentes para a tarefa em questão.
2. Acurácia: grau em que os dados são corretos, confiáveis e certificados.
3. Atualidade: o grau em que a idade dos dados é apropriada para a tarefa em questão.
4. Consistência: a medida em que os dados são apresentados no mesmo formato e são compatíveis com os dados anteriores.
5. Acessibilidade: disponibilidade da informação.

As técnicas de tratamento de dados utilizada em uma determinada base variam de um conjunto de dados para outro, não sendo únicas. As técnicas usadas neste trabalho são apresentadas a seguir.

3.2.1 Limpeza de dados

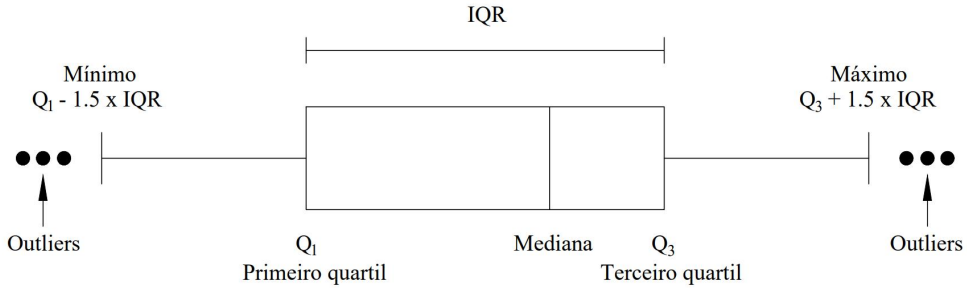
Em quase todas as bases da dados há valores ausentes, ruidosos ou inconsistentes que contribuem para reduzir a qualidade da base de dados. Estes problemas podem ocorrer por vários fatores: formulários de entrada de dados mal projetados, erro humano na entrada de dados, decadência de dados, erros em dispositivos de instrumentação que registram os dados ou dados utilizados para outros fins que não são os originalmente previstos.

A limpeza de dados visa identificar e preencher valores ausentes, identificar anomalias e corrigir inconsistências nos dados (HAN; KAMBER; PEI, 2011). É o processo de correção ou remoção de valores incorretos, corruptos, formatados incorretamente, duplicados ou incompletos dentro de um conjunto de dados.

Os problemas mais comuns são a presença de valores ausentes e anômalos. O valor ausente, é um valor para uma variável que não foi introduzido ou foi perdido no processo de registro (GARCÍA; LUENGO; HERRERA, 2015) . É o problema mais comum e diversos métodos são utilizados para lidar com valores ausentes (PRATAMA et al., 2016):

1. Omissão: é a maneira mais simples, os valores ausentes são completamente ignorados à medida que a análise é realizada. Embora seja um método simples, é arriscado, se a porcentagem de faltas dos dados for grande o suficiente para interromper o resultado da análise.
2. Exclusão: é simplesmente apagar o valor que falta para continuar a análise. Igual que o método anterior, é riscado se a quantidade de informação faltante é elevada e pode ocasionar perda de informação significante.
3. Imputação média/modo/mediana: embora a omissão e a exclusão não ofereçam bons resultados em uma análise com grandes quantidades de faltas, o método de imputação é uma solução para melhores resultados, pois resolve o problema dos valores faltantes e o número de dados esperados continua o mesmo. A proposta deste método é substituir cada um dos valores ausentes pela média, moda ou mediana dos dados observados para aquela variável.
4. Interpolação linear: substitui valores ausentes por interpolação linear, esta técnica é comumente utilizada em bases de séries temporais. O último valor válido antes do valor ausente e o primeiro valor válido após o valor ausente são utilizados para a interpolação.

Já os valores anômalos, conhecidos como *outliers* no inglês, são observações que se desviam tanto de outras observações que levantam suspeitas de que foram gerados por um mecanismo diferente (OSBORNE, 2013) . Técnicas básicas de descrição estatística como *boxplots* e plotagem de dispersão podem ser utilizados para identificar *outliers* (HAN;



KAMBER; PEI, 2011) . Um *boxplot* é um método para representar graficamente grupos de dados numéricos através de seus quartis (SCHWERTMAN; OWENS; ADNAN, 2004) . Ele exibe cinco valores de um conjunto de dados: o valor mínimo, o valor do primeiro quartil, o valor da mediana, o valor do terceiro quartil, e o valor máximo como mostrado na Figura 3.1.

Figura 3.1 – Gráfico *Boxplot.*

Fonte: Autor (2023)

Os dados são ordenados do menor para o maior e divididos em quatro partes iguais. Os seguintes valores podem ser detectados rapidamente:

1. Primeiro quartil (Q_1): é o valor do conjunto que delimita os 25% valores menores, isto é, 25% dos valores são menores do que Q_1 e 75% são maiores do que Q_1.
2. Mediana ou segundo quartil: Divide o conjunto em duas partes iguais.
3. Terceiro quartil (Q_3): é o valor que delimita os 25% valores maiores, isto é, 75% dos valores são menores do que Q_3 e 25% são maiores do que Q_3.
4. Interquartil (IQR): Diferença entre o valor do terceiro quartil e o valor do primeiro quartil.
5. Mínimo e máximo: Estes representam os valores limites e se calculam conforme mostrado na figura acima.

Todos os valores fora dos limites mínimo e máximo são considerados como valores *outliers* segundo a técnica do gráfico *boxplot*. Após identificados esses valores podem ser tratados conforme as técnicas utilizadas para o tratamento dos valores ausentes.

3.2.2 Integração de dados

Antes de proceder à aplicação de algoritmos de AM, é necessário verificar a integração do banco de dados, isto é, se foi realizada uma fusão correta de dados de diferentes fontes. A integração cuidadosa pode ajudar a reduzir e evitar redundâncias e inconsistências no conjunto de dados resultante. Adicionalmente, ajuda a melhorar a fidelidade e a velocidade do processo subsequente a ser aplicado (HAN; KAMBER; PEI, 2011).

Uma questão importante a ser considerada durante a integração de dados é a redundância. Uma variável pode ser redundante se puder ser derivada de outra e pode ocorrer

durante a integração se vier de fontes diferentes. As redundâncias podem ser detectadas através de análise de correlação. Dados duas variáveis, tal análise pode medir o quão fortemente uma variável implica a outra.

Para as variáveis numéricas, a correlação entre *X* e *Y*, pode ser avaliada calculando o coeficiente de correlação também conhecido como coeficiente de Pearson, nomeado em homenagem a seu inventor, Karl Pearson (HAN; KAMBER; PEI, 2011). Avaliar a correlação de Pearson entre duas variáveis pode ajudar a estabelecer a força da relação, assim como determinar se a relação é positiva, negativa ou inexistente (WEAVER et al., 2017) . O coeficiente de Pearson pode ser calculado pela seguinte equação:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\overline{X}\overline{Y}}{n\sigma_X \sigma_Y}, \quad -1 \leq r_{x,y} \leq +1 \tag{3.1}$$

Onde

n é o número de tuplas;

x_i e y_i são os respectivos valores de *X* e *Y*;

$\overline{X}$ e $\overline{Y}$ são os respectivos valores médios de *X* e *Y*;

σ_X e σ_Y são os respectivos desvios padrão de *X* e *Y*.

Se $r_{x,y}$ é maior do que 0, então *X* e *Y* são positivamente correlacionadas, o que significa que os valores de *X* aumentam à medida que os valores de *Y* aumentam. Se $r_{x,y}$ for igual a 0, então *X* e *Y* são independentes e não há correlação entre elas. Se $r_{x,y}$ for negativo, então *X* e *Y* são negativamente correlacionadas, onde os valores de uma variável aumentam conforme os valores da outra diminuem. Quanto mais próximo a -1 ou $+1$ for o coeficiente, mais forte será a correlação e portanto, *X* ou *Y* pode ser removido como uma redundância. Note que a correlação não implica uma causalidade. Isso é, se *X* e *Y* estiverem correlacionadas, esta não implica necessariamente que *X* cause *Y* ou vice-versa.

3.2.3 Transformação de dados

A unidade de medição utilizada pode afetar a análise dos dados. Todos as variáveis devem ser expressas nas mesmas unidades de medida e devem usar uma escala ou faixa comum. O objetivo de normalizar os dados é dar a todos as variáveis o mesmo peso (GARCÍA; LUENGO; HERRERA, 2015). Alguns algoritmos de aprendizagem de máquinas são sensíveis à escala, enquanto outros são praticamente invariáveis. Algoritmos sensíveis são algoritmos baseados na distância entre pontos de dados, como kNN, *K-means* e SVM. Isto

destaca a importância de ter um banco de dados com suas variáveis na mesma escala. As técnicas mais comuns de dimensionamento de variáveis são a normalização e a padronização.

A normalização Mín-Max é o método mais simples e visa escalar todos os valores numéricos v de uma variável numérica *X* para um intervalo comumente especificado [0,1]. Assim, um valor transformado é obtido aplicando a seguinte expressão ao valor *v* para obter o novo valor v' (HAN; KAMBER; PEI, 2011):

$$v' = \frac{v - min_X}{max_X - min_X} \tag{3.2}$$

Onde

v é o valor a transformar;

v' é o novo valor do dado;

min_X é valor mínimo encontrado na variável *X*;

max_X é valor máximo encontrado na variável *X*.

Usando uma normalização para reescalar todos os dados para a mesma faixa de valores evitarão que aquelas variáveis com uma grande diferença $(max_X - min_X)$ dominem sobre as outras, enganando o processo de aprendizagem ao dar mais importância para as primeiras. Esta normalização também é conhecida por acelerar o processo de aprendizagem nas RNAs, ajudando os pesos a convergir mais rapidamente.

3.3 TÉCNICAS DE SELEÇÃO DE ATRIBUTOS

A quantidade de dados de alta dimensão que existe e está disponível publicamente na Internet tem aumentado nos últimos anos. Portanto, os métodos de AM têm dificuldades em lidar com o grande número de colunas de entrada comumente chamadas como atributos, o que representa um desafio interessante porque nem todas as colunas representam valores relevantes e, consequentemente, podem confundir os algoritmos, levando a um mau desempenho dos modelos de aprendizagem da máquina. Por este motivo, a seleção de atributos é uma das técnicas mais frequentes e importantes no pré-processamento de dados, e se tornou um componente indispensável do processo de aprendizagem da máquina.

A seleção de atributos é frequentemente chamada de seleção de variáveis, seleção de características e seleção de subconjuntos de variáveis. É o processo de reduzir as variáveis de entrada para as mais informativas sem alterá-las (ALNUAIMI et al., 2022) . Há três abordagens gerais para a seleção de atributos.

Primeiro, a abordagem *filter* que explora as variáveis gerais dos dados, independentemente do algoritmo de AM. Seguidamente, a abordagem *wrapper* que busca um subconjunto de variáveis ótimo adaptado ao algoritmo de AM específico. E a abordagem *emdebbed* que é feita com um algoritmo de AM específico que realiza a seleção de atributos no processo de treinamento.

3.3.1 Abordagem *filter*

Esta abordagem incorpora uma medida independente para avaliar subconjuntos de variáveis sem envolver um algoritmo de aprendizado (KUMAR, 2014). Eles se concentram na aplicação de medidas estatísticas para atribuir pontuações a cada variável. As variáveis são então ordenados com base na pontuação, resultando na seleção daquelas com a pontuação mais alta e na eliminação das últimas variáveis ordenadas. Estes métodos são geralmente uni variados e podem considerar as variáveis independentemente da variável dependente. Seu diagrama de fluxo é apresentado na Figura 3.2.

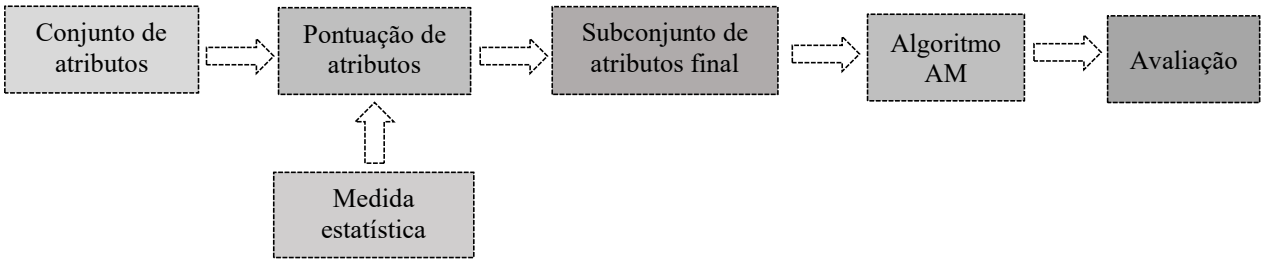


Figura 3.2 – Diagrama de fluxo abordagem *filter*.

Fonte: Autor (2023)

Entre os algoritmos mais conhecidos nesta categoria estão os métodos baseados na teoria da informação e *ReliefF*, os quais foram utilizados neste trabalho.

3.3.1.1 Informação Mútua

IM fornece uma correlação generalizada análoga ao coeficiente de correlação linear, mas é sensível tanto às correlações lineares quanto às não lineares. O conceito principal de informação mútua surge da teoria da informação, proposta em 1948 por Shannon (SHANNON, 1948) . Ela descreve a quantidade de informações compartilhadas entre dois variáveis aleatórias. É uma medida simétrica da relação entre duas variáveis aleatórias e é sempre maior ou igual a zero, onde quanto maior for o valor, maior será a relação entre as

duas variáveis. Se o resultado calculado for zero, então as variáveis são independentes (COVER; THOMAS, 2006).

Considerando duas variáveis aleatórias contínuas $X = [x_1, x_2, ..., x_n]$ e Y $= [y_1, y_2, ..., y_n]$ onde n é o número total de amostras, a informação mútua entre *X* e *Y* é definida pela Equação (3.3).

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.3}$$

onde *H(X)* e *H(Y)* são as entropias de informação de *X* e *Y* que são medidas da incerteza das variáveis aleatórias definidas nas equações (3.4-3.5) respectivamente.

$$H(X) = -\int_x p(x) \log p(x)\, \boldsymbol{dx} \tag{3.4}$$

$$H(Y) = -\int_y p(y) \log p(y)\, \boldsymbol{dy} \tag{3.5}$$

H(X,Y) é a entropia conjunta de *X* e *Y* e está definida como:

$$H(X,Y) = -\int_x \int_y p(x,y) \log p(x,y)\, \boldsymbol{dx\, dy} \tag{3.6}$$

Portanto, para quantificar a quantidade de informação sobre a variável *X* fornecida pela variável *Y* (e vice-versa), que é conhecida como informação mútua, é usada a Equação (3.7).

$$I(X;Y) = \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, \boldsymbol{dx\, dy} \tag{3.7}$$

onde $p(x,y)$ é uma função conjunta de densidade de probabilidade, e $p(x)$ e $p(y)$ são as funções de densidade marginal.

3.3.1.2 Relief

Relief é um algoritmo de seleção de atributos introduzido por (KIRA; RENDELL, 1992) que confere um peso a todos as variáveis do conjunto de dados e podem ser modificados gradualmente. Quanto maior for o peso de uma variável, mais útil será. Os pesos das variáveis são referidos como *W[A]* = peso da variável '*A*'. Dada uma instância escolhida aleatoriamente R_i, *Relief* procura seus dois vizinhos mais próximos: um da mesma classe,

chamado de *hit* mais próximo *H*, e o outro de uma classe diferente, chamado de *miss M* mais próximo. Depois o algoritmo atualiza a estimativa de qualidade *W[A]* para todos as variáveis *A*, dependendo de seus valores para R_i, *M* e *H*. Se as instâncias R_i e *H* têm valores diferentes da variável *A*, então a variável *A* separa duas instâncias com a mesma classe que não é desejável, por isso se diminui a estimativa de qualidade *W[A]*. Por outro lado, se as instâncias R_i e *M* têm valores diferentes da variável *A*, então a variável *A* separa duas instâncias com valores de classe diferentes, o que é desejável, então se aumenta a estimativa de qualidade *W[A]* e o processo é repetido *m* vezes.

Relief tem sido amplamente aplicado por causa de sua simplicidade e eficiência. Mas sua limitação é que só pode lidar com a classificação de duas categorias de dados, e não pode lidar com a situação de falta de dados. Portanto, (ROBNIK-SIKONJA; KONONENKO, 1997) propuseram o algoritmo *RReliefF* que pode lidar com problemas de várias classes com alta eficiência e pode lidar também com problemas de regressão mostrado na Figura 3.3.

Figura 3.3 – Pseudocódigo de *RReliefF*

Algoritmo *RReliefF*

Entrada: para cada instância de treinamento, um vetor de valores de variáveis **x** e o valor previsto $\tau(\boldsymbol{x})$
Saida: o vetor *W* das estimativas das qualidades das variáveis

1. definir todos os $\boldsymbol{N_{dC}}$, $\boldsymbol{N_{dA}[A]}$, $\boldsymbol{N_{dC\&dA}[A]}$, $\boldsymbol{W[A]}$ a 0;
2. **para** i := 1 **a** m **de início**
3. selecionar aleatoriamente a instância $\boldsymbol{R_i}$;
4. selecionar **k** instâncias $\boldsymbol{I_j}$ mais próximas de $\boldsymbol{R_i}$;
5. **para** j := 1 **a** k **de início**
6. $N_{dC} := N_{dC} + |f(R_i) - f(I_i)| \cdot d(i,j)$;
7. **para** A:=1 **a** todos as variáveis **de início**
8. $N_{dA}[A] := N_{dA}[A] + diff(A, R_i, I_j) \cdot d(i,j)$;
9. $N_{dC\&dA}[A] := N_{dC\&dA}[A] + |f(R_i) - f(I_j)| \cdot diff(A, R_i, I_j) \cdot d(i,j)$;
10. **fim;**
11. **fim;**
12. **fim;**
13. **para** A:=1 **a** todas as variáveis **faça**
14. $W[A] := N_{dC\&dA}[A]/N_{dC} - (N_{dA}[A] - N_{dC\&dA}[A])/(m - N_{dC})$;

Fonte: Kira e Rendell (1992)

3.3.2 Abordagem *wrapper*

A abordagem *filter* e *wrapper* são diferenciadas pelos critérios de avaliação. A abordagem *wrapper* utiliza um algoritmo de aprendizagem para a avaliação de subconjuntos de variáveis e seleciona um subconjunto ótimo de que melhor se adapta ao algoritmo

(KUMAR, 2014). Como se mostra na Figura 3.4, o algoritmo de aprendizagem é executado múltiplas vezes e cada vez com um diferente subconjunto de variáveis. O de melhor avaliação é escolhido como o subconjunto final (KOHAVI; JOHN, 1997).

Alguns exemplos comuns de métodos *wrappers* são: seleção sequencial progressiva, seleção sequencial regressiva e seleção sequencial recursiva.

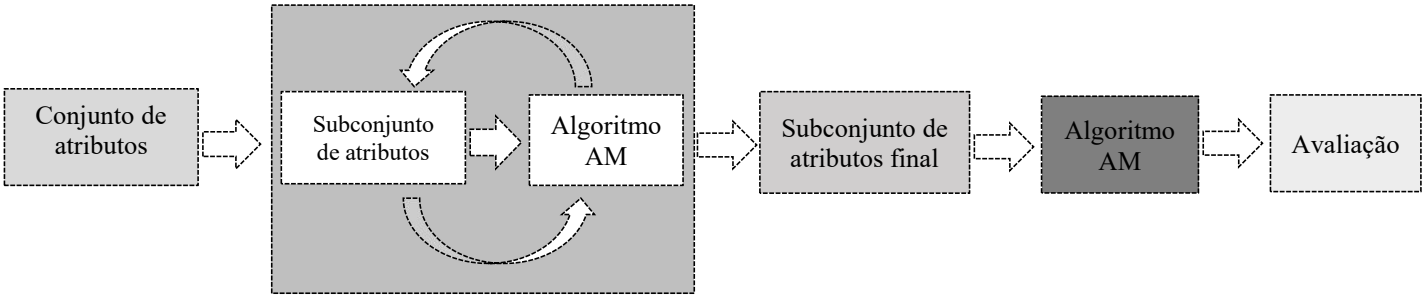


Figura 3.4 – Diagrama de fluxo abordagem wrapper.

Fonte: Autor (2023)

3.3.3 Abordagem *emdebbed*

A abordagem *emdebbed* combina as qualidades dos métodos *filter* e *wrapper*. É implementada por algoritmos que têm seus próprios métodos de seleção de atributos integrados. Baseiam-se no aprendizado sobre qual variável contribui mais para a precisão do modelo à medida que ele está sendo criado (MERA-GAONA et al., 2021). Como se mostra na Figura 3.5, a seleção de atributos é baseada em uma pontuação calculada pelo algoritmo de aprendizagem que é construído usando um conjunto de treinamento e a precisão da previsão avaliada com um conjunto de teste.

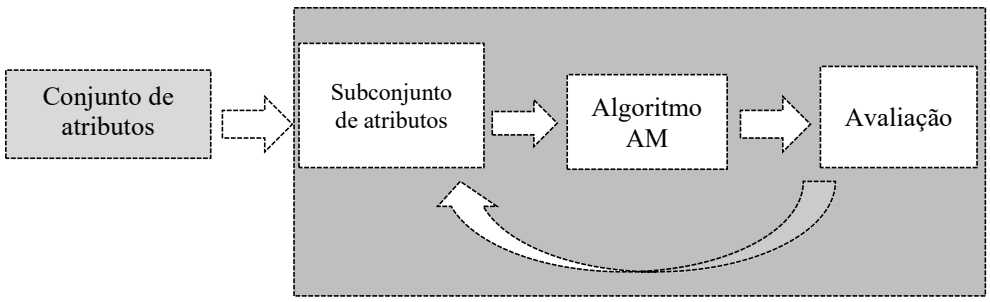


Figura 3.5 – Diagrama de fluxo abordagem embedded.

Fonte: Autor (2023)

Isto então dá o valor de desempenho para cada variável. Como o modelo é construído apenas uma vez para determinar a pontuação dos variáveis, eles têm um custo computacional menor em comparação com os métodos de wrapper. Exemplos são regressão LASSO, regressão *ridge* e *random forest*, o qual foi utilizado neste trabalho.

3.3.3.1 *Random Forest*

O algoritmo RF, é um algoritmo composto por uma coleção de árvores de decisão para classificação ou regressão construída selecionando amostras aleatórias do conjunto de dados e foi introduzido por (BREIMAN, 2001). Em RF, o objetivo é a redução das impurezas a cada divisão de nó. Uma divisão com uma grande diminuição da impureza é considerada importante e, como consequência, as variáveis utilizadas nessa divisão também são consideradas importantes. O valor final da importância para uma variável X_k é calculada pela soma de todas as diminuições em impurezas dos nós onde a variável X_k foi usada para dividir os dados (HJERPE, 2016).

Para classificação, a redução da impureza é tipicamente medida pelo índice de *Gini*, definido na equação 3.10, e para regressão pela soma dos quadrados MSE como se mostra na equação 3.11.

$$Gini(v) = \sum_{k=1}^{K} F_k^v(1 - F_k^v) \tag{3.8}$$

onde F_k^v é a freqüência da classe *k* no nó *v*.

$$MSE(v) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \tag{3.9}$$

onde y_i é o valor de uma instancia, *N* é o número de instancias e a média das instancias.

Após escolhido o critério de divisão do nó denotado por $i(v)$, ele é dividido em nó criança esquerdo e nó criança direito denotados v_L e v_R respectivamente. A diminuição da impureza após a divisão, *d*, é a diferença entre a impureza do nó inicial e a soma ponderada das medidas de impureza dos dois nós crianças e é denotada pela equação 3.12.

$$\Delta i(v) = i(v) - \frac{N_{vL}}{N_v} i(v_L) - \frac{N_{vR}}{N_v} i(v_R) \tag{3.10}$$

onde N_v denota o número de exemplos que alcançam o nó *v*.

A medida de importância para a variável X_k (VI^k) é calculada pela soma de todas as diminuições em impurezas a traves de todas as arvores, *T*, onde a variável X_k é usada para dividir os dados e é denotada pela equação 3.13.

$$VI^k = \frac{1}{n_T} \sum_{n \in S_{Xk}} \Delta i(X_k, v) \tag{3.11}$$

onde n_T é o número de arvores, S_{Xk} é o conjunto de nós divididos por X_k número de variáveis.

3.4 ALGORITMOS DE APRENDIZADO DE MÁQUINA

Nesta seção são apresentados os conceitos teóricos, diagramas e pseudocódigos dos algoritmos supervisionados e não supervisionados propostos para a previsão da irradiação solar.

3.4.1 *Support Vector Regression*

SVR é um algoritmo que pode ser utilizado para classificação ou regressão baseado na teoria do aprendizado estatístico e no princípio da minimização de riscos estruturais, e foi apresentado primeiramente por (CORTES; VAPNIK, 1995).

Para prever uma variável não linear é necessário ter um conjunto de treinamento T:

$$T = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\} \tag{3.12}$$

onde x_m são as entradas e y_m são as saídas. As variáveis de entrada do modelo estão relacionadas à variável-alvo visualizando os dados mapeados em uma função não linear $f(x)$:

$$f(x) = w \cdot \phi(x) + b \tag{3.13}$$

onde w é o vetor pesos, b é uma constante chamada bias e $\phi(x)$ é uma característica espacial de alta dimensão mapeada pelo vetor espacial x. O objetivo é encaixar os dados T, encontrando uma função $f(x)$ que tenha um maior desvio ε em relação aos dados reais objetivos para todos os dados de treinamento T, e ao mesmo tempo é tão pequeno quanto possível. Portanto, a Equação (3.15) se transforma em um problema de otimização como segue:

$$min\ \frac{1}{2}\|w\|^2 \tag{3.14}$$

Caso $f(x)$ não seja viável, podem-se introduzir as variáveis ξ_i, ξ_i^* para lidar com restrições de outra forma impraticáveis do problema de otimização, resultando na seguinte equação:

$$\begin{gathered} min\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ s.a.\ \begin{cases} y_i - \langle w, x_i\rangle - b \le \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \end{cases} \end{gathered} \tag{3.15}$$

A constante C > 0 determina a escolha entre a planicidade de f e a quantidade até a qual os desvios maiores do que ε são tolerados (SMOLA; SCHÖLKOPF, 2004).

3.4.2 *Random Forest*

Árvores de decisão simples estão sujeitas a várias limitações, em particular uma variância elevada. Devido a este fato, (BREIMAN, 2001) propôs no início dos anos noventa uma técnica padrão conhecida como *bagging*, que corresponde a *bootstrap aggregating*, esta técnica consiste em treinar árvores independentes com um subconjunto aleatório do conjunto de treinamento. Isto é, em lugar de construir um único preditor (uma única árvore de decisão), o método gera um conjunto de preditores por *bagging,* técnica de reamostragem estatística que envolve a amostragem aleatória de um conjunto de dados com substituição, sobre a amostra de aprendizagem e depois agrega suas previsões como se mostra na Figura 3.6.

Ao calcular a média de suas previsões, RF alcança uma maior generalização e, portanto, uma precisão superior. Para evitar a correlação das diferentes árvores, RF aumenta a diversidade das árvores ao fazê-las crescer a partir de diferentes subconjuntos de dados de treinamento criados através de bagging por meio de uma nova amostragem aleatória do conjunto de dados original com substituição, ou seja, sem eliminação dos dados selecionados da amostra de entrada para a geração do próximo subconjunto. Assim, alguns dados podem ser usados mais de uma vez no treinamento, enquanto outros podem nunca ser usados.

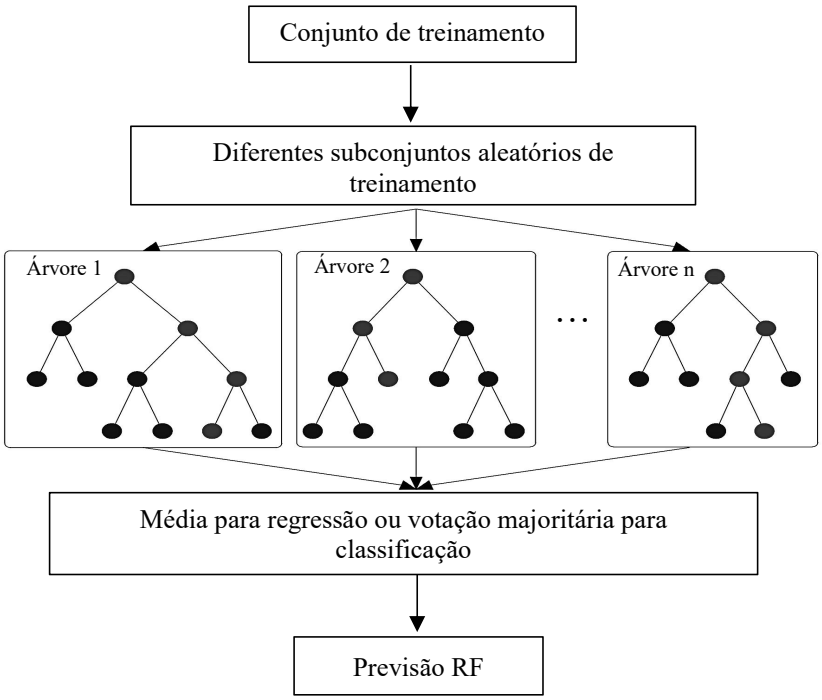


Figura 3.6 – Diagrama *Random Forest* .

Fonte: Autor (2023)

3.4.3 *Adaptive Boosting*

AdaBoost é um método proposto por (FREUND; SCHAPIRE, 1997) . Se baseia na ideia de que um melhor modelo pode ser criado combinando múltiplos modelos individuais considerados como “fracos” adicionados sequencialmente, o que significa que os erros dos modelos anteriores são aprendidos por seus sucessores. A ideia é focar em instâncias que foram previstas erroneamente ao treinar um novo modelo. Isto é obtido aplicando pesos $w_1, w_2, \ldots, w_N$ a cada amostra de treinamento. No início, todos os pesos estão igualmente ajustados e um aprendiz sobre os dados originais é treinado. A cada iteração, os pesos da amostra são atualizados como se mostra na Figura 3.7. Em cada etapa, conforme as previsões corretas são feitas, o peso do exemplo de treinamento é diminuído, e ao contrário, o peso é aumentado se o modelo anterior o previu incorretamente. Em outras palavras, as previsões incorretas aumentam seus pesos para a próxima etapa, enquanto as previsões corretas diminuem seus pesos. Finalmente, as previsões são integradas usando o voto majoritário

ponderado em caso de problemas de classificação ou média ponderada ou mediana em caso de regressão. AdaBoost.R2 é uma versão modificada de AdaBoost adaptada para problemas de regressão apresentada por (DRUCKER, 1997) e utilizado neste trabalho. O método Drucker seguiu o espírito do algoritmo AdaBoost utilizando repetidamente uma árvore de regressão como uma máquina de aprendizagem fraca, seguida do aumento dos pesos dos exemplos mal previstos e da diminuição dos pesos dos exemplos bem previstos. Entretanto, o AdaBoost é um algoritmo com uma função de perda convexa e é sensível a ruídos e *outliers* nos dados, propenso a sobre ajustamento.

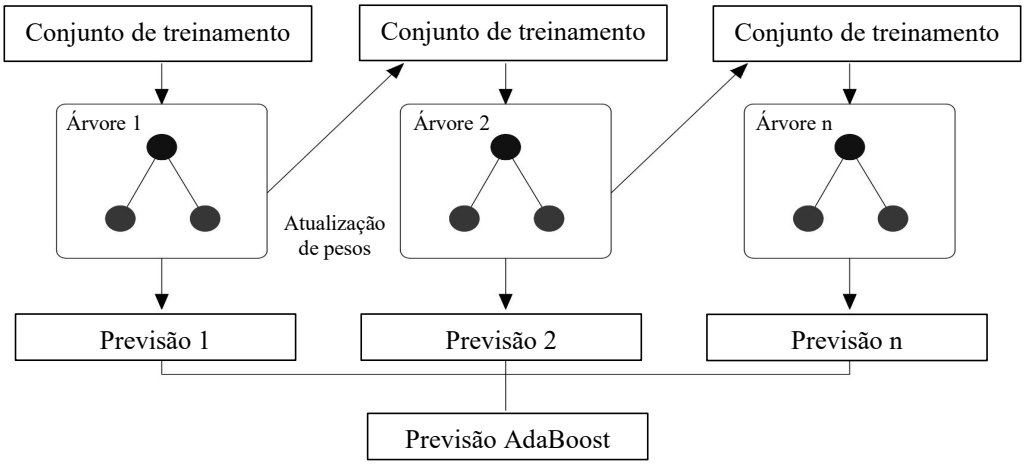


Figura 3.7 – Diagrama AdaBoost.

Fonte: Autor (2023)

...

Atualização
de pesos

3.4.4 *Extreme Gradient Boosting*

XGBT é um algoritmo de aprendizado *ensemble* relativamente novo, proposto por (CHEN; GUESTRIN, 2016) . Similar ao AdaBoost, ele constrói uma série sequencial de aprendizes fracos, neste caso árvores de decisão, e a cada iteração cada aprendiz tenta complementar ao anterior baseando-se nos resíduos das previsões feitas pelo aprendiz anterior e não assina pesos para as instancias dependendo de se foram corretamente previstas ou não. Em adição, o XGBT implementa o algoritmo de aumento de gradiente onde cada árvore de regressão atribui um ponto de dados de entrada para uma de suas folhas contendo uma pontuação contínua. XGBT minimiza uma função objetiva regularizada que combina uma função de perda convexa (baseada na diferença entre os resultados previstos e esperados) e um termo de penalidade de complexidade do modelo. No modelo o treinamento é feito iterativamente, adicionando novas árvores que preveem os resíduos ou erros das árvores anteriores, que são combinados com as anteriores para fazer a previsão final adicionando reforço de gradiente que minimiza a perda ao adicionar novos modelos. As principais vantagens do XGBT são sua velocidade em comparação com outros algoritmos, como o AdaBoost, e seu parâmetro de regularização que reduz com sucesso a variância. E, além do parâmetro de regularização, este algoritmo aproveita uma taxa de aprendizagem e subamostra de variáveis que aumenta ainda mais sua generalização. A Figura 3.8 mostra o fluxograma de XGBT.

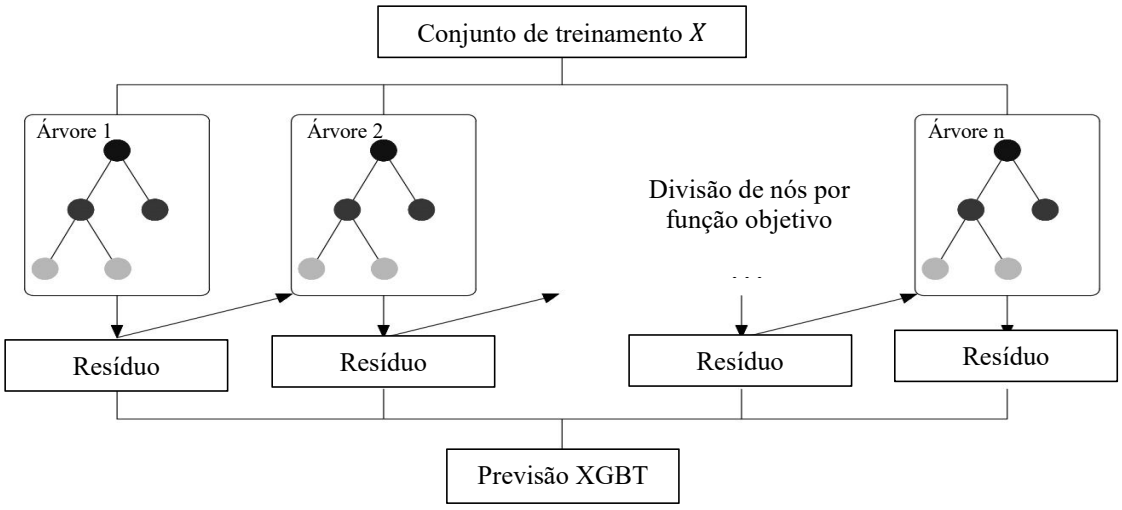


Figura 3.8 – Diagrama Extreme Gradient Boosting .

Fonte: Autor (2023)

3.4.5 *Categorical Boosting*

Impulso Categórico, do inglês *Categorical Boosting* (CatBoost) é uma implementação de código aberto da arvore de decisão de gradiente impulsionada que traz duas inovações: estatística ordenada de objetivos e boosting ordenado. O algoritmo foi proposto por (DOROGUSH; ERSHOV; GULIN, 2018) sendo baseado em árvores de decisão e aumento de gradiente. Ele suporta variáveis numéricas, categóricas e de texto e pode ser utilizado para resolver problemas, tais como regressão, classificação e ranking. Este algoritmo é diferente dos tradicionais algoritmos de aumento de gradiente nos seguintes aspectos:

1. Processamento de variáveis categóricas durante o tempo de treinamento ao em vez de no tempo de pré-processamento. CatBoost permite que todo o conjunto de dados seja utilizado para treinamento. De acordo com os autores, ao implementar a estatística objetiva se consegue de forma mais eficiente lidar com variáveis categóricas com perda mínima de informação.
2. Combinação de variáveis. Todos as variáveis categóricos podem ser combinados como uma nova variável. Ao construir uma nova árvore, CatBoost usa um método ganancioso para considerar combinações. Nenhuma combinação é considerada para a primeira divisão na árvore, mas para a segunda divisão e subsequentes divisões, CatBoost combina todas as combinações pré-definidas com todos as variáveis categóricos no conjunto de dados. Todas as partições selecionadas na árvore são consideradas como uma categoria com dois valores e são utilizadas em combinação.
3. Boosting imparcial com variáveis categóricas. Quando o método é usado para converter variáveis categóricas em valores numéricos, a distribuição será diferente da distribuição original, e o desvio desta distribuição causará o desvio da solução, o que é um problema inevitável para os métodos tradicionais de gradiente boosting. Portanto, os autores desenvolveram um novo método através de análise teórica para superar o desvio do gradiente chamado boosting ordenado.
4. Pontuador rápido. CatBoost utiliza árvores oblíquas como preditores de base, onde o mesmo critério de divisão é utilizado em todo o nível da árvore. Estas árvores são equilibradas e menos propensas ao sobre ajuste ou *overfitting*.

3.4.6 *Voting Regressor*

Regressor por votação, do inglês *Voting Regressor* (VR) é um *ensemble model* proposto por (AN; MENG, 2010) que compreende vários modelos de AM e calcula a média de suas previsões individuais em para formar uma previsão final como se mostra na Figura 3.9. Este método é útil para um conjunto de modelos com bom desempenho para compensar suas fraquezas individuais e construir um modelo único que possa generalizar melhor. No caso de questões de regressão os resultados de todos os modelos são calculados como média para obter uma estimativa final.

Há dois métodos de votação: o Voto Médio Simples, do inglês *Voting Average* (VOA) e o Voto Médio Ponderado, do inglês *Voting Weighted Average* (VOWA). No caso do VOA, os pesos são equivalentes e iguais a 1 e o valor previsto final é obtido tomando o valor médio das previsões resultantes dos modelos de AM individuais:

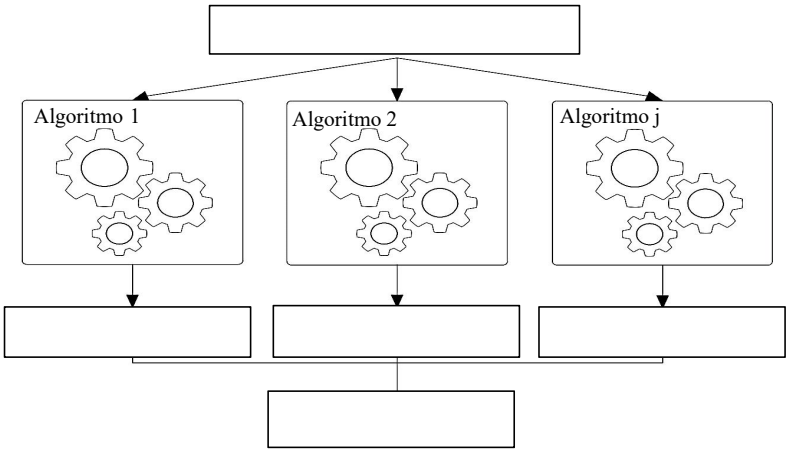
$$\hat{y} = \sum_{j=1}^{m} \hat{y}_j / m \tag{3.16}$$

onde m é o número de algoritmos de AM usados no *ensemble*, $\hat{y}_j$ é o valor previsto do algoritmo j, $\hat{y}$ é o valor previsto final. Uma desvantagem da VOA é que todos os modelos do *ensemble* são aceitos como igualmente eficientes.

VOWA especifica um coeficiente de peso para cada membro do *ensemble*. O peso pode ser um número de ponto flutuante entre 0 e 1 sendo a soma igual a 1, ou um número inteiro de 1 denotando o número de votos concedidos ao membro do *ensemble* correspondente. O valor previsto final é obtido como mostrado na Eq 3.17, onde w_j é o peso do algoritmo *j*.

$$\hat{y} = \frac{\sum_{j=1}^{m} (w_j \hat{y}_j)}{\sum_{j=1}^{m} w_j} \tag{3.17}$$

Figura 3.9 – Diagrama do modelo de votação.



3.4.7 *K-means*

O *K-means* e um algoritmo de aprendizagem não supervisionado e de clusterização, utilizado para particionar dados em k clusters distintos. Ele agrupa dados que compartilham características importantes e parecidas. De modo empírico, uma boa solução para o processo de clusterização e aquela em que os dados do cluster sejam mais semelhantes entre si, do que comparados com outro cluster. *K-means* é um dos mais simples e conhecidos algoritmos de clusterização introduzido por (MACQUEEN, 1967) .

Ao configurar um modelo de clusterização usando o método *K-means*, deve-se especificar um número k que indica o número de centroides que desejado no modelo. O centroide é um ponto representativo de cada cluster. O algoritmo *K-means* atribui cada ponto de dados de entrada a um dos clusters minimizando a soma de quadrados dentro do cluster. No processamento dos dados de treinamento, o algoritmo *K-means* começa com um conjunto inicial de centroides escolhidos aleatoriamente. Os centroides servem como pontos de partida para os clusters e aplicam o algoritmo de Lloyd para refinar seus locais iterativamente.

O algoritmo *K-means* é brevemente apresentado a seguir. Seja $X = \{\vec{x}_j\}$, $j = 1, 2, \ldots, M$ um conjunto de treino composto por M vetores N-dimensionais, com $M \gg k$. O algoritmo *K-means* particiona o espaço vetorial $\mathbb{R}^N$ atribuindo cada vetor de treino a um único cluster através da busca do vizinho mais próximo (VMP). Precisamente, $\vec{x}_j$ pertencerá ao cluster $V(\vec{w}_i)$ se $d(\vec{x}_j, \vec{w}_i) < d(\vec{x}_j, \vec{w}_a), \forall a \neq i$, em que $d(\vec{x}_j, \vec{w}_i)$ denota a distância Euclidiana quadrática entre $\vec{x}_j$ e $\vec{w}_i$. Neste caso, diz-se que $\vec{w}_i$ é o VMP de $\vec{x}_j$. Pode-se associar a busca do VMP a uma função de pertinência, definida por

$$\mu_i(\vec{x}_j) = \begin{cases} 1, & \text{se } \vec{w}_i = \text{VMP}(\vec{x}_j) \\ 0, & \text{caso contrario} \end{cases} \tag{3.18}$$

Dessa forma, a distorção obtida ao se representarem todos os vetores do conjunto de treino pelos respectivos VMPs é dada por

$$J_i = \sum_{i=1}^{k} \sum_{j=1}^{M} \mu_i(\vec{x}_j) d(\vec{x}_j, \vec{w}_i) \tag{3.19}$$

Para minimizar J_i, os vetores $\vec{w}_i$são atualizados como segue:

$$\vec{w}_i = \frac{\sum_{j=1}^{M} \mu_i(\vec{x}_j)\vec{x}_j}{\sum_{j=1}^{M} \mu_i(\vec{x}_j)}, i = 1,2, \ldots k. \tag{3.20}$$

A escolha adequada do número de clusters k no algoritmo *K-means* é uma questão crucial para obter resultados precisos e significativos. Para auxiliar nessa determinação, é comum utilizar três índices bem estabelecidos: Calinski-Harabasz, Silhouette e Davies-Bouldin.

O índice de validação de Silhouette (SH) mostra quão próximo cada ponto de dados está 3, de outros pontos de dados dentro de um cluster e quão bem separados os clusters estão uns dos outros. Em outras palavras, funciona como uma função da distância entre cada ponto dentro e entre os grupos. O valor ideal de k corresponde ao valor máximo de SH que reflete a melhor partição.

O índice de validação Calinski-Harabasz (CH) mostra a qualidade da solução de clusterização com base na soma média dos quadrados entre e dentro dos clusters. Um valor CH mais alto reflete um melhor resultado de agrupação.

O índice de Davies Bouldin (DB) funciona com base na semelhança média entre cada cluster e seu cluster mais parecido. Uma partição de dados bem definida é representada por um baixo valor DB.

Assim, a utilização desses três índices como critérios de avaliação permite uma abordagem mais fundamentada e orientada à obtenção de resultados confiáveis e relevantes na aplicação do algoritmo *K-means*.

4 METODOLOGIA

4.1 INTRODUÇÃO

O objetivo deste trabalho é desenvolver um modelo de previsão de irradiação solar para horizontes de curto prazo, com intervalos de uma hora. Neste sentido, será utilizado um modelo *ensemble* que combina técnicas de seleção de atributos tais como IM, RF e *Relief* e diferentes modelos de AM, como SVR, RF, AdaBoost, XGBT e CatBoost.

O ambiente computacional *Jupyter Notebook* é utilizado para aplicação do modelo, que é um ambiente de programação baseado em células com a possibilidade de escrever e executar código e visualizar o resultado do código logo abaixo. Sua interface flexível permite aos usuários configurar e organizar fluxos de trabalho em ciência de dados, computação científica, jornalismo computacional e AM. Diferentes pacotes *Python* (*pandas, numpy, scikit-learn, seaborn, matplotlib* etc.) podem ser importados para este ambiente.

4.2 MODELO DE PREVISÃO PROPOSTO

O método de previsão proposto combina o uso de um modelo *ensemble* de seleção de atributos com o método de votação com base em diferentes algoritmos de aprendizado de máquina, onde as previsões de irradiação solar são propostas para 1 h, 2 h, 3 h, 6 h, 9 h e 12 h à frente. Este modelo é apresentado na Figura 4.1.

Primeiro, é obtido um banco de dados real e substancial contendo dados de irradiação solar e adicionalmente, informações meteorológicas. Em seguida, o pré-processamento de dados é realizado conforme descrito na seção 3.2 para limpar os dados, identificando os outliers e valores ausentes. A análise de correlação também é realizada para identificar relações entre as variáveis, e em seguida os dados são normalizados para que tenham a mesma escala.

Após isto, os dados são divididos em treinamento, validação e teste, e o algoritmo *K-means* é utilizado para separar os dados em clusters com características meteorológicas semelhantes.

A próxima etapa é selecionar os atributos mais significativos e seus valores de atraso, usando um modelo *ensemble* que combina IM, RF e *Relief*.

Na fase seguinte, os parâmetros dos algoritmos de aprendizado de máquina são otimizados e o desempenho de cada algoritmo é avaliado, para posteriormente construir os modelos de votação por média simples e por média ponderada.

Finalmente, vários indicadores estatísticos são utilizados para quantificar o desempenho dos modelos para diferentes horizontes de previsão. Estas etapas são descritas detalhadamente nas seções seguintes.

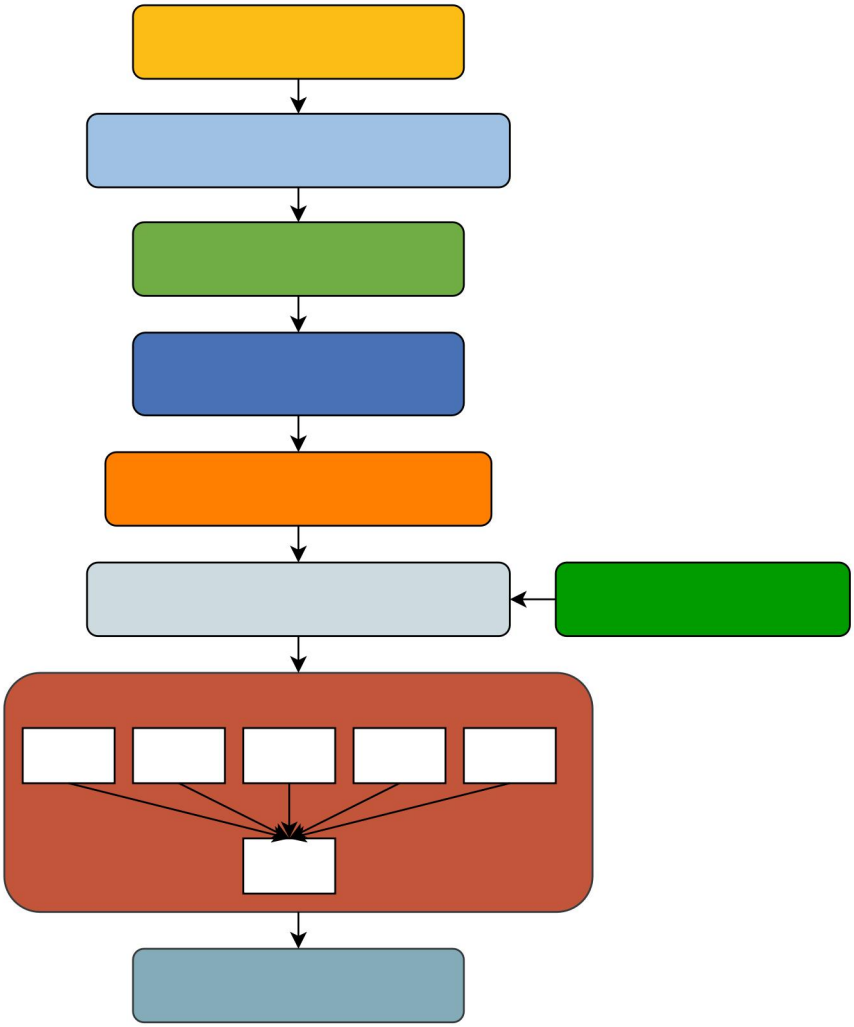


Figura 4.1 – Metodologia de previsão proposta.

Fonte: Autor (2023)

4.3 BANCO DE DADOS

As séries históricas das variáveis meteorológicas utilizadas foram obtidas no portal de dados abertos do Instituto Nacional de Meteorologia (INMET) (INMET, 2022). As estações

meteorológicas do INMET estão equipadas com dispositivos para medir a temperatura (termómetro), vento (anemómetro), precipitação (pluviómetro), pressão atmosférica (barómetro) e irradiação solar (pirómetro), instalados em locais considerados estratégicos e em áreas de interesse do país.

Os dados consultados são da cidade de Salvador, Bahia, localizada na costa brasileira (12°58'28.9992″ S, 38°28'35.9940″ W) a uma altitude de 8 m, como mostrado na Figura 4.2. Salvador tem um clima tropical caracterizado por temperaturas elevadas que vão de 22 °C a 31 °C. A série obtida vai de 01 de janeiro de 2015 e finaliza em 23 de agosto de 2022 com registros de hora em hora, totalizando 67.005 registros.

Na série temporal da irradiação solar, apenas as amostras diurnas são consideradas. Se a série temporal completa for usada, muitos valores observados são zero (período noturno), e os valores previstos também serão zero (ou muito próximos), reduzindo substancialmente o erro da previsão e superestimando o desempenho do modelo de previsão.

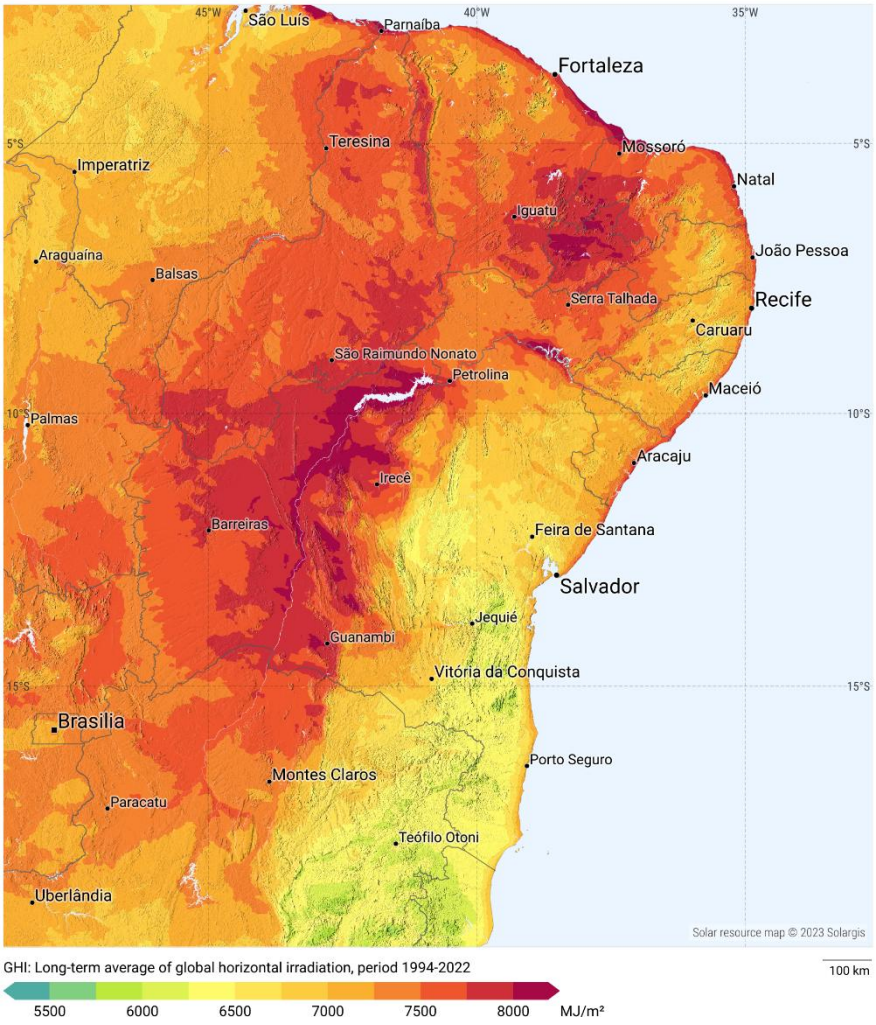


Figura 4.2 – Localização geográfica de Salvador e níveis globais de irradiação horizontal.

Fonte: (SOLARGIS, 2023)

As variáveis do banco de dados são apresentadas na Tabela 4.1, indicando sua respectiva abreviação e a unidade de medição indicada pela estação.

Tabela 4.1 – Base de dados disponível para previsão da irradiação solar.

Variável	Abreviação	Unidade de medida
Hora	H	hora
Irradiação global	R	MJ/m²
Vento rajada máx.	W_g	m/s
Vento velocidade	W_s	m/s
Vento direção	W_d	◦
Temperatura do ar - bulbo seco	T	◦C
Temperatura máxima na hora	T^{max}	◦C
Temperatura mínima na hora	T^{min}	◦C

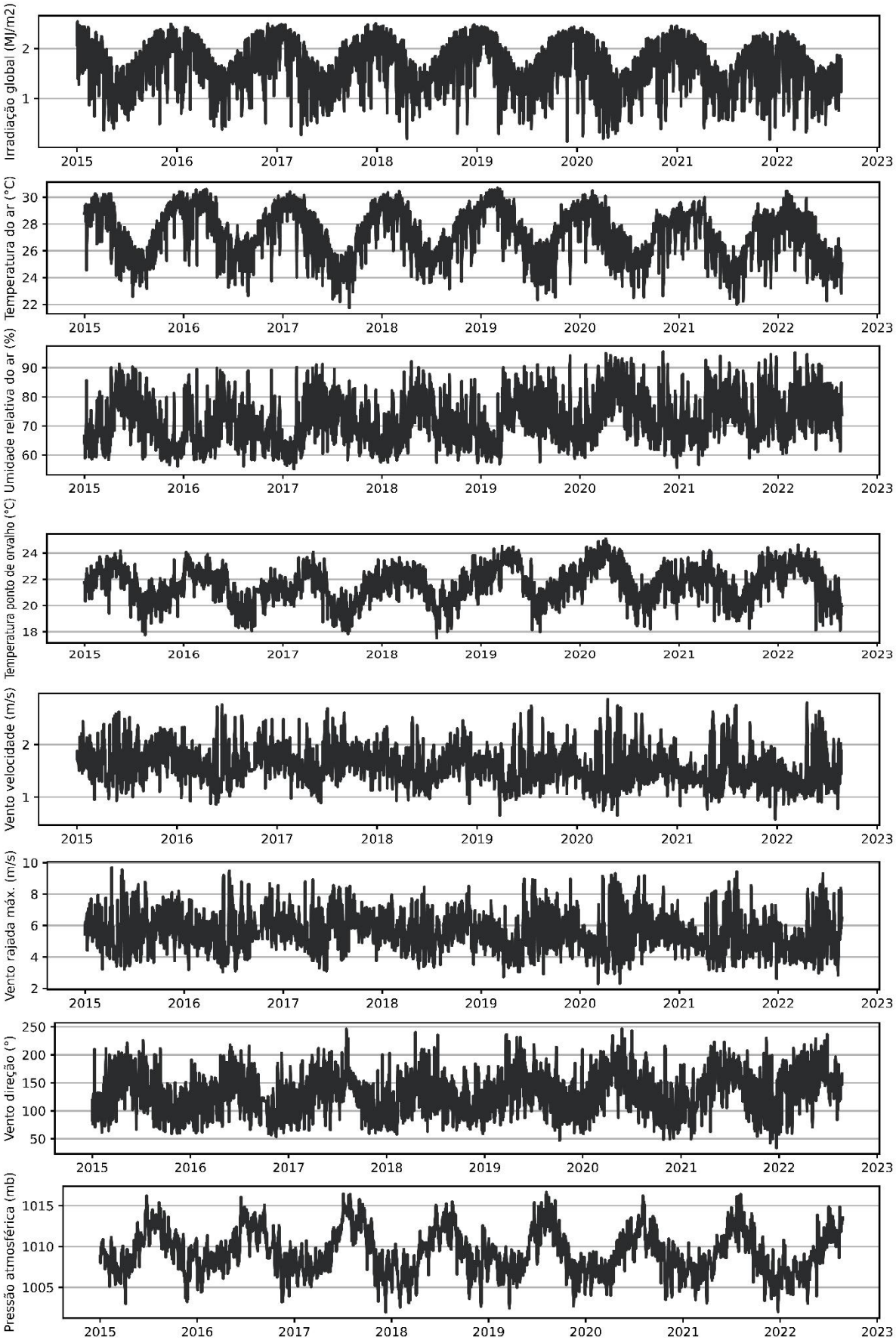
Temperatura do ponto de orvalho	T_d	◦C
Temperatura orvalho máx. na hora	T_d^{max}	◦C
Temperatura orvalho min. na hora	T_d^{min}	◦C
Precipitação total	P	mm
Pressão atmosférica ao nível da estação	A	mb
Pressão atmosférica máx. na hora	A^{max}	mb
Pressão atmosférica min na hora (mb)	A^{min}	mb
Umidade relativa do ar	H	%
Umidade rel. min. na hora	H^{max}	%
Umidade rel. min. na hora	H^{min}	%

A Figura 4.3 apresenta o comportamento das nove principais séries temporais componentes do banco de dados do estudo de 2015 a 2022. As análises estatísticas dos dados diurnos das séries temporais são apresentadas na Tabela 4.2.

Tabela 4.2 – Análises estatística da base de dados.

Variável	Média	Desvio padrão	Min	Max
H	12.00	3.16	7.00	17.00
R	1.64	1.04	0.00	3.57
P	0.25	1.49	0.00	50.40
W_s	1.58	0.54	0.10	3.00
W_g	5.57	1.70	0.60	10.00
W_d	132.55	60.70	1.00	360.00
T	27.29	2.34	20.40	31.70
T^{max}	28.07	2.47	20.80	33.00
T^{min}	26.45	2.31	19.90	30.80
T_d	21.61	1.50	17.40	25.40
T_d^{max}	22.35	1.49	17.90	25.90
T_d^{min}	20.92	1.49	16.80	25.10
H	71.85	10.37	52.00	96.00
H^{max}	75.62	9.88	58.00	97.00
H^{min}	68.68	10.96	47.00	96.00
A	1009.28	2.98	1001.10	1017.10
A^{max}	1009.59	2.94	1001.20	1017.50
A^{min}	1009.09	2.95	1000.70	1017.00

Figura 4.3 – Comportamento series temporais de 2015 a 2022.



Fonte: Autor (2023)

4.4 PRÉ-PROCESSAMENTO DOS DADOS

Os dados foram submetidos a etapa de pré-processamento com procedimentos conforme descrito na seção 3.2. Inicialmente foi feita a detecção e tratamento de valores nulos e *outliers*. Em seguida, aplicou-se a análise de correlação de Pearson entre as séries temporais da base de dados do estudo, conforme ilustra a Figura 4.4. Como esperado, os resultados indicam uma alta correlação linear entre as variáveis e seus valores mínimo e máximo, com valores próximos de -1 ou +1. Assim, as seguintes variáveis são removidas do conjunto de dados: pressão atmosférica horaria máxima e mínima, temperatura horaria máxima e mínima, temperatura do ponto de orvalho horária máxima e mínima, e umidade relativa horaria máxima e mínima.

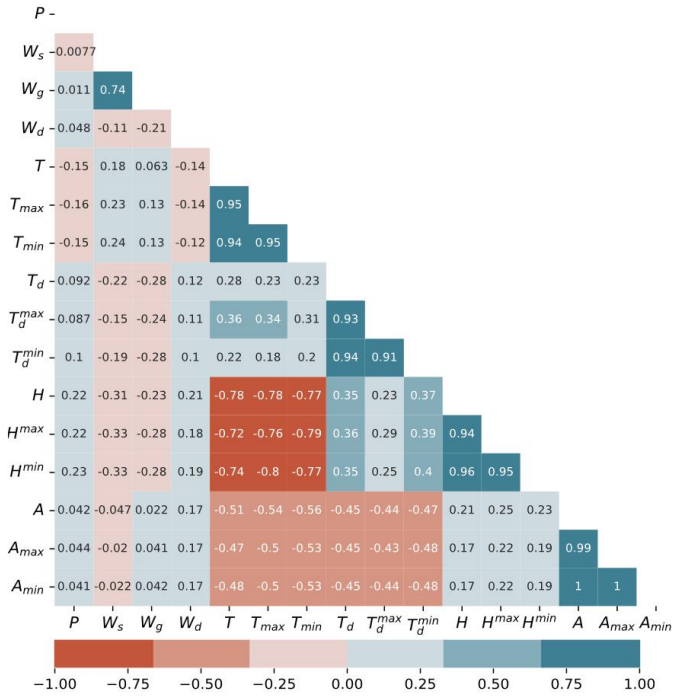


Figura 4.4 – Análises de correlação das series temporais do banco de dados.

Fonte: Autor (2023)

A técnica de normalização Min-Max foi usada nesta dissertação para normalizar a base de dados. Finalmente, o conjunto de dados históricos é dividido em três conjuntos: treinamento, validação e teste. O conjunto de treinamento é utilizado para construir os

modelos aprendizado de máquina. O conjunto de validação é utilizado para ajustar os hiper parâmetros dos modelos. O conjunto de teste é utilizado para estimar o desempenho do modelo em dados não utilizados para treinar o modelo. A divisão aplicada é 70% dos dados para treinamento, 10% para validação e 20% para teste, como mostrado na Figura 4.5.

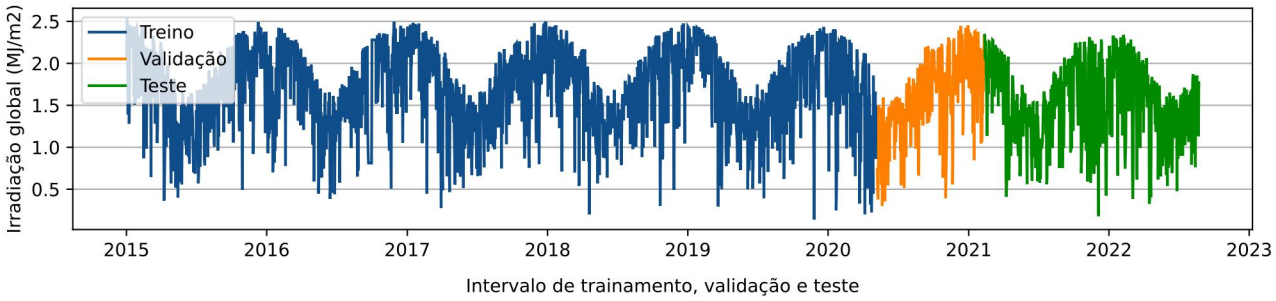


Figura 4.5 – Conjunto de treinamento, validação e teste da série temporal de irradiação solar.

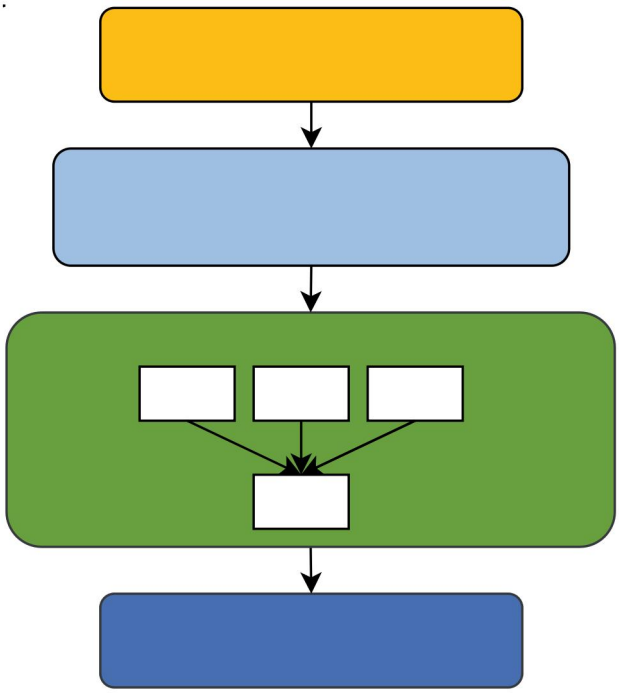
Fonte: Autor (2023)

4.5 CLUSTERIZAÇÃO DOS DADOS

Neste trabalho, o algoritmo de clusterização *K-means* é utilizado para agrupar os dados com padrões meteorológicos semelhantes. Como descrito na seção 3.4.7 o número de clusters é um dos parâmetros de entrada do algoritmo, sendo definido pelo usuário. Para determinar esse valor três índices são utilizados neste trabalho: Calinski-Harabasz, Silhouette e Davies-Bouldin.

Como mostrado na Figura 4.6, o algoritmo *K-means* é primeiramente aplicado aos dados de treinamento variando o número de k clusters de 2 até 10 e calculando os três índices de validação para cada k. O número ótimo de k é selecionado para cada índice separadamente. Finalmente, o número k é escolhido por uma maioria de votos dos três índices.

Figura 4.6 – Metodologia de clusterização proposta.



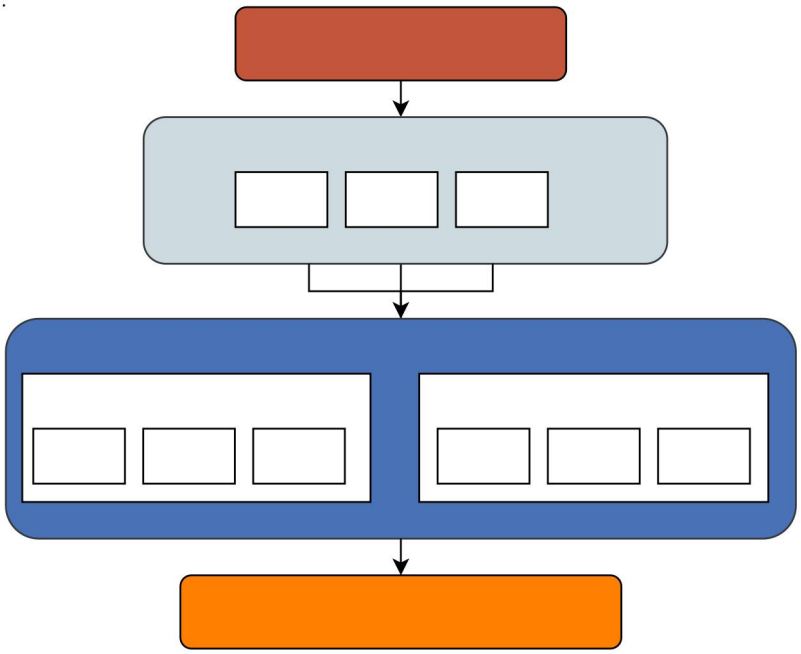
Fonte: Autor (2023)

4.6 MODELO *ENSEMBLE* PARA SELEÇÃO DE ATRIBUTOS

O conjunto de dados utilizado neste trabalho consiste na séries temporal de irradiação solar como variável endógena e outras variáveis meteorológicas como variáveis exógenas. A seleção dos atributos é aplicada para escolher os atrasos mais importantes da variável endógena e as variáveis exógenas mais importantes juntamente com seus atrasos, analisando as relações lineares e não lineares entre elas.

A metodologia de seleção de atributos proposta é apresentada na Figura 4.7. Os algoritmos Informação Mútua, *Random Forest* e *Relief* são utilizados para selecionar os atributos mais significativos. Para cada método, a importância de cada variável é avaliada e seu valor é normalizado. A importância final de cada variável é obtida calculando o valor médio da importância individual de cada variável atribuída pelos três algoritmos de seleção de atributos. Estes valores de importância são posteriormente ordenados do maior ao menor, construindo assim um ranking de importância de cada variável para cada cluster.

Figura 4.7 – Metodologia de seleção de atributos proposta.



Fonte: Autor (2023)

O algoritmo VOA é posteriormente construído combinando os algoritmos: RF, XGBT, SVR, AdaBoost e CatBoost, e seu desempenho é avaliado começando com apenas o atributo mais importante no ranking e, em seguida, inserindo os outros atributos um a um, de acordo com sua importância no ranking, até que o desempenho do algoritmo seja avaliado usando todos os atributos. A medida do erro médio absoluto é usado para avaliar o desempenho, e o conjunto de atributos que produz o menor erro é selecionado como conjunto final.

Como o conjunto de dados consiste em séries temporais multivariadas, a seleção adequada dos atrasos das variáveis (observações passadas) é uma tarefa importante para garantir uma precisão de previsão satisfatória. O próximo passo é selecionar os atrasos das variáveis endógenas e exógenas, aplicando o mesmo modelo de seleção que combina os três algoritmos: Informação Mútua, *Random Forest* e *Relief*. A seleção dos atrasos mais significativos para as variáveis exógenas é feita separadamente da variável endógena, pois elas têm um significado reduzido, mas não menos significativo. Considerando $Xt - k$, um intervalo de *k* horas na variável *X*, o intervalo adotado de atrasos a serem testados é de 1 a 72 para cada variável $(Xt - 1, \ldots, Xt - 72)$, o que é um intervalo suficiente para capturar informações importantes dos valores históricos.

Utilizando o método *ensemble*, um ranking dos atrasos mais importantes para cada variável é construído e sua seleção é realizada empiricamente utilizando o algoritmo VOA

que combina todos os algoritmos e a medida do erro médio absoluto como um instrumento de avaliação de desempenho.

4.7 OTIMIZAÇÃO DE HIPER PARÂMETROS

Os hiper parâmetros são explicitamente definidos pelo usuário para controlar o processo de aprendizagem de qualquer algoritmo de aprendizado de máquina. Estes hiper parâmetros são usados para melhorar a aprendizagem do modelo e seus valores são definidos antes de iniciar o processo de aprendizagem do modelo.

Os hiper parâmetros são importantes porque controlam diretamente o comportamento do algoritmo de treinamento e têm um impacto significativo no desempenho do modelo que está sendo treinado. Uma boa escolha de hiper parâmetros pode aumentar significativamente o desempenho do modelo, por isso, existem técnicas de otimização desenvolvidas com o foco em determinar os melhores valores para os hiper parâmetros de um modelo em uma dada aplicação.

Como o objetivo é encontrar os melhores valores de hiper parâmetros, o método de busca manual pode ser tentado, usando o processo "erro e acerto", porém demoraria muito tempo para construir um único modelo. Das diversas propostas de seleção de hiper parâmetros existentes na literatura, a busca em grade *(*no inglês *GridSearch*) e a validação cruzada do inglês *cross-validation* são as mais utilizadas e são adotadas neste trabalho (AGRAWAL, 2021).

A busca em grade se refere ao processo de busca exaustiva sobre um subconjunto do espaço de trabalho, sendo que a busca é realizada em um espaço formado pelos parâmetros de interesse e seu objetivo é encontrar pontos nos quais a acurácia seja a maior possível. A validação cruzada é uma técnica utilizada para avaliar a capacidade de generalização de um algoritmo quando exposto a um novo conjunto de dados. Sua desvantagem é o tempo de processamento, já que faz uma busca linear no espaço de hiper parâmetros. Outra questão fundamental no método de busca em grade é a delimitação do espaço a ser investigado que deve ser definido pelo usuário.

Os seguintes parâmetros são investigados para cada algoritmo proposto neste trabalho.

4.7.1 Hiper parâmetros do algoritmo SVR

SVR pode ser considerado como um problema de otimização no qual se encontra um hiperplano ideal que separa os dados. Uma função de “kernel” é utilizada no modelo SVR

para mapear o espaço de entrada original para um espaço de dimensão superior. As funções kernel mais utilizadas são a linear, polinomial, função de base radial (RBF do inglês *Radial Basis Function*) e sigmoidal. Os parâmetros de constante de regularização (C) e gama são também fornecidos como entrada para o algoritmo SVR e influenciam o processo de otimização que divide o hiperplano de forma ideal com base nos dados de treinamento. O parâmetro C que controla a quantidade de regularização aplicada aos dados e o parâmetro gama decide até que ponto a influência de um único exemplo de treinamento atinge durante a transformação, o que, por sua vez, afeta a intensidade com que os limites de decisão acabam envolvendo os pontos no espaço de entrada (SHUKLA et al., 2020).

4.7.2 Hiper parâmetros do algoritmo RF

Lembrando que RF é um algoritmo que combina a predição de um conjunto de árvores de decisão, para obter uma única resposta como saída, o número de árvores é um hiper parâmetro do modelo que, em geral, quanto maior é seu valor, melhor a acurácia de predição. Porém, o aumento deste número é benéfico até certo ponto, uma vez que a partir de certa quantidade de árvores, a melhoria na resposta combinada cessa, além do fato de um maior número de árvores consumirem uma maior quantidade de recursos computacionais. Os parâmetros mais relevantes são o número de estimadores (n_estimators) que define a quantidade de árvores gerada na floresta), número máximo de atributos (max_features) consideradas para dividir um nó da árvore, e a profundidade máxima (max_depth) que cada árvore poderá atingir (MOHAPATRA; SHREYA; CHINMAY, 2020).

4.7.3 Hiper parâmetros do algoritmo XGBT

Da mesma forma que o RF, XGBT também é um algoritmo que combina a predição de um conjunto de arvores de decisão, pelo que o número de arvores é um hiper parâmetro presente neste algoritmo e definido a través do parâmetro, número de estimadores (n_estimators).

Adicionalmente, XGBT possui diversos hiper parâmetros, porém, só serão tratados alguns mais importantes que controlam as arvores de regressão e o modelo *ensemble*. O primeiro hiper parâmetro e um dos mais utilizados é a profundidade máxima (max_depth) que informa ao algoritmo a profundidade máxima de cada árvore. Controla o nível de subdivisões que as arvores podem fazer. O segundo parâmetro é o (colsample_bytree) que define a porcentagem de colunas que serão utilizados para construir cada árvore e de acordo com a

documentação, isso ocorre uma vez para toda a árvore construída. O seguinte hiper parâmetro é o (subsample) que basicamente funciona como uma razão das instâncias de treinamento. E um dos hiper parâmetros que controlam o modelo é a taxa de aprendizado (learning_rate) utilizada para controlar o ajuste do modelo. Altos valores garantem a um treinamento mais rápido, porém, pode sobre ajustar o modelo (WADE, 2020).

4.7.4 Hiper parâmetros do algoritmo CatBoost

Um dos destaques e vantagens do algoritmo CatBoost é que, com os parâmetros predefinidos, ele fornece resultados satisfatórios; portanto, o mais importante é definir os parâmetros corretos dependendo do problema que estamos resolvendo como o número máximo de árvores que podem ser construídas por meio do parâmetro iterações (iterations), a taxa de aprendizagem (learning_rate) na qual os pesos do modelo são atualizados após cada amostra de treinamento é analisada, a profundidade das árvores (depth) e o coeficiente de regularização (l2_reg) usado para o cálculo do valor de uma folha (GULIN, 2018).

4.7.5 Hiper parâmetros do algoritmo AdaBoost

Na prática, o AdaBoost é um algoritmo popular por vários motivos, sendo um deles o fato de não exigir a configuração de muitos hiper parâmetros. Apenas dois hiper parâmetros foram otimizados neste trabalho. O número de estimadores (n_estimators) que é o número máximo de aprendizes fracos que serão treinados e a taxa de aprendizagem (learning_rate) que é o peso aplicado a cada classificador em cada iteração que determina em que medida as informações antigas serão substituídas pelas novas.

A Tabela 4.3 apresenta um resumo dos hiper parâmetros a serem otimizados para cada algoritmo investigado neste trabalho.

Tabela 4.3 – Hiper parâmetros utilizados nos algoritmos.

Algoritmo	Hiper parâmetro	Descrição
SVR	C	Constante de regularização
	gamma	Coeficiente do kernel
	kernel	Especifica o tipo de kernel a ser usado no algoritmo.
RF	max_depth	Profundidade das arvores
	n_estimators	Número de estimadores (árvores)
	max_features	Número máximo de atributos para dividir um nó
XGBT	learning_rate	Fator de ponderação da aprendizagem
	max_depth	Profundidade das arvores
	n_estimators	Número de estimadores (árvores)
	subsample	Razão das instâncias de treinamento
	colsample_bytree	Porcentagem de colunas para construir cada árvore

CatBoost	depth	Profundidade das arvores
	l2_reg	Coeficiente de regularização
	learning_rate	Usado para reduzir o gradiente
	iterations	Número máximo de estimadores (árvores)
AdaBoost	learning_rate	Peso aplicado aos regressores em cada iteração
	n_estimators	Número de estimadores (árvores)

4.8 MEDIDAS DE AVALIAÇÃO

A avaliação do resultado é obtida a partir da aplicação de um modelo preditivo, ou a comparação entre os resultados de diferentes modelos. Além disso, deve-se levar em conta a natureza do problema e as características do conjunto de dados em cada caso, pois, a escolha de uma medida inadequada pode levar a falsas conclusões de sucesso para os resultados de um modelo. Felizmente existem medidas amplamente utilizadas na área de ciência de dados, as quais atendem tal finalidade.

O desempenho dos modelos usados neste trabalho é avaliado usando as seguintes medidas: erro médio absoluto, erro médio absoluto percentual, raiz do erro quadrático médio e o coeficiente de determinação. Nas equações seguintes, F_i é o valor previsto, O_i é o valor observado, $\overline{O}_i$ é o valor médio das observações, $\overline{F}$ é o valor médio das previsões e N é o número de amostras.

4.8.1 Erro médio absoluto

O erro médio absoluto, do inglês *Mean Absolute Error* (MAE) é calculado por meio da diferença entre os valores atuais e os preditos:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|F_i - O_i| \tag{4.1}$$

4.8.2 Erro percentual médio absoluto

O erro percentual médio absoluto, do inglês *Mean Absolute Percentage Error* (MAPE) mede o tamanho do erro em termos percentuais:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{F_i - O_i}{O_i}\right| \times 100 \tag{4.2}$$

4.8.3 Raiz do erro quadrático médio

A raiz do erro quadrático médio, do inglês *Root mean square error* (RMSE) penaliza mais os erros maiores, permite avaliar a qualidade de um previsor em relação aos dados:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2} \tag{4.3}$$

4.8.4 Coeficiente de determinação

O coeficiente de determinação, do inglês *Coefficient of Determination* (R^2) também é avaliado. Ele mede a variação nas previsões. Um coeficiente igual a 1 indica que o modelo interpreta perfeitamente os dados observados, enquanto um 0 indica que o modelo não interpreta corretamente dados não vistos.

O coeficiente de determinação é avaliado da seguinte forma:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(F_i - O_i)^2}{\sum_{i=1}^{N}(F_i - \overline{O_i})^2} \quad , \quad \overline{O_i} = \sum_{i=0}^{N-1} F_i \tag{4.4}$$

No caso de MAE, RMSE e MAPE, quanto menores os valores, melhor é a previsão e no caso do R^2 quanto maior o valor melhor é a interpretação dos dados.

4.8.5 Análise estatística

O teste Diebold-Mariano (DM), proposto por (DIEBOLD; MARIANO, 1995) , desempenha um papel fundamental na avaliação e seleção de modelos de previsão, tornando-se uma ferramenta estatística amplamente utilizada na pesquisa acadêmica. Esse teste é aplicado no contexto de previsão de séries temporais e economia, com o objetivo de determinar se existem diferenças estatisticamente significativas na precisão entre os modelos em análise. Para isso, são calculados os erros de previsão dos modelos, que representam as discrepâncias entre os valores previstos e os valores observados.

Definam-se os erros de previsão de dois algoritmos concorrentes como:

$$e_{jt} = \hat{y}_{jt} - y_t \quad , t = 1 \ldots n \tag{4.5}$$

onde $\hat{y}_{jt}$ é o valor previsto pelo j-ésimo algoritmo ($j = 1,2$), y_t é o valor observado, e n é o número de amostras. A função de perda do erro de previsão $g(e_{jt})$ é normalmente adotada como o erro quadrático ou o erro absoluto.

Em seguida, é calculada uma medida estatística conhecida como estatística DM, que compara as diferenças médias nos erros de previsão entre os modelos. O teste DM é baseado na diferença de perda d_t entre as duas previsões concorrentes, definida como:

$$d_t = g(e_{1t}) - g(e_{2t}) \tag{4.6}$$

O teste de Diebold-Mariano formula duas hipóteses: a hipótese nula (H_0) e a hipótese alternativa (H_a). A hipótese nula sustenta que não há diferença estatisticamente significativa na precisão dos modelos de previsão ($H_0 : E(d_t) = 0 \; \forall t$, enquanto a hipótese alternativa afirma que existe uma diferença significativa ($H_a : E(d_t) \neq 0 \; \forall t$).

No teste DM, é estabelecido um nível de significância de $p = 0{,}05$. Em seguida, a decisão de rejeitar ou não a hipótese nula é baseada no valor-p resultante. Se o valor-p for maior que 0,05, não será possível rejeitar a hipótese nula, e as diferenças observadas entre o desempenho dos dois modelos de previsão não são significativas. Caso contrário, se o valor-p for menor que 0,05, a hipótese nula será rejeitada, indicando que as diferenças observadas entre o desempenho dos dois modelos de previsão são significativas.

Em resumo, o teste de Diebold-Mariano possibilita a comparação da precisão entre modelos de previsão por meio da formulação de hipóteses, do cálculo da estatística DM e da avaliação do valor-p em relação ao nível de significância estabelecido. A aceitação ou rejeição da hipótese nula permite determinar se existem diferenças estatisticamente significativas na precisão dos modelos. Essa abordagem contribui significativamente para o avanço da pesquisa em diversos campos, aprimorando a qualidade das previsões.

4.9 CONSIDERAÇÕES FINAIS

Este capítulo apresentou a metodologia a ser seguida para o desenvolvimento da pesquisa, primeiro descrevendo a metodologia em geral e, em seguida, apresentando uma descrição mais detalhada do modelo proposto para clusterização de dados, seleção de atributos, otimização de hiper parâmetros e medidas de avaliação de desempenho dos algoritmos. O capítulo termina com a apresentação do conjunto de dados selecionado para esta pesquisa.

É importante destacar que esta metodologia pode ser adaptada e aplicada a diferentes bases de dados para fazer previsões não só da irradiação solar (proposta neste trabalho), mas também de outras variáveis como velocidade do vento, potência de geração solar, carga energética, entre outras. Esta metodologia será aplicada ao banco de dados meteorológicos da cidade de Salvador, Bahia, descrita na seção 4.3. Os resultados obtidos com a aplicação da metodologia de previsão da irradiação solar proposta são apresentados a seguir.

5 ANÁLISES DOS RESULTADOS

Este capítulo apresenta e discute os resultados obtidos com a aplicação da metodologia proposta na seção anterior, que utiliza diferentes algoritmos de AM para previsão da energia solar sob diferentes horizontes de previsão. Os efeitos do uso da seleção de atributos também são investigados, além do uso de modelos de votação.

Todos os experimentos foram realizados em um notebook com um processador Intel core i5-1035G1 (1,19 GHz) com 8GB de memória RAM e utilizando a linguagem de programação Python 3.8.11. A biblioteca scikit-learn 0.24.0 foi utilizada como suporte para as atividades de aprendizagem de máquina. Além disso, outras bibliotecas de apoio foram utilizadas, como NumPy 1.21.5 para cálculos matemáticos, Pandas 1.4.2 para eventuais

manipulações de dados e Matplotlib 2.0.0, através do módulo pyplot, para geração de gráficos e a plataforma Jupyter Lab foi utilizada como ambiente de desenvolvimento integrado (IDE do inglês *Integrated Development Environment*).

5.1 RESULTADO DA CLUSTERIZAÇÃO DOS DADOS

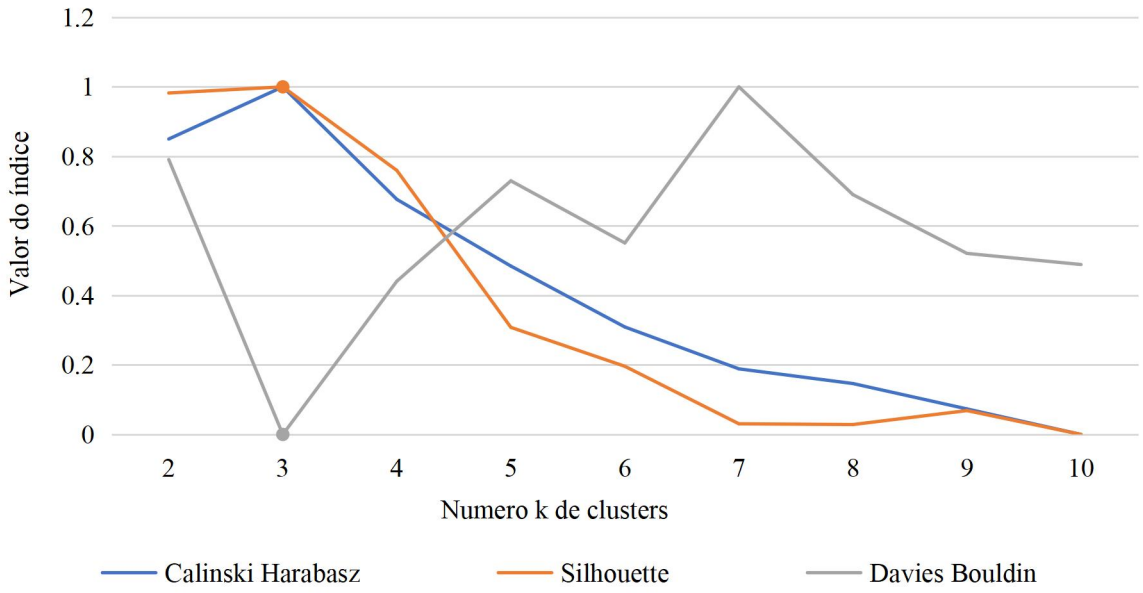
As condições climáticas mudam constantemente, e dependendo desta variação, a irradiação também sofre variações. Portanto, o primeiro passo é aplicar a técnica de clusterização dos dados de acordo com seus padrões meteorológicos, de modo que dentro de um grupo os dados tenham padrões semelhantes entre si e diferentes dos dados dos demais grupos. A Tabela 5.1 apresenta os resultados das três medidas aplicadas ao banco de dados, variando o número k de clusters.

Tabela 5.1 – Resultado dos índices de clusterização dos dados.

Nº de clusters (k)	**Índice Calinski-Harabasz**	**Índice Silhouette**	**Índice Davies Bouldin**
2	866.9865	0.3033	1.3917
3	924.8393	0.3053	1.1702
4	800.0188	0.2786	1.2937
5	726.0164	0.2285	1.3746
6	658.4049	0.2161	1.3245
7	611.9654	0.1978	1.4504
8	595.6785	0.1975	1.3635
9	567.5222	0.2020	1.3161
10	539.2053	0.1944	1.3071

A Figura 5.1 apresenta os valores normalizados das três medidas utilizadas visando a comparação e seleção do número ótimo de k. Pode-se observar que de acordo com todas as técnicas utilizadas o número ideal de k é igual a três. Portanto, a base de dados foi dividida em três clusters de acordo com suas condições meteorológicas.

Figura 5.1 – Resultados de variação do número k de clusters.

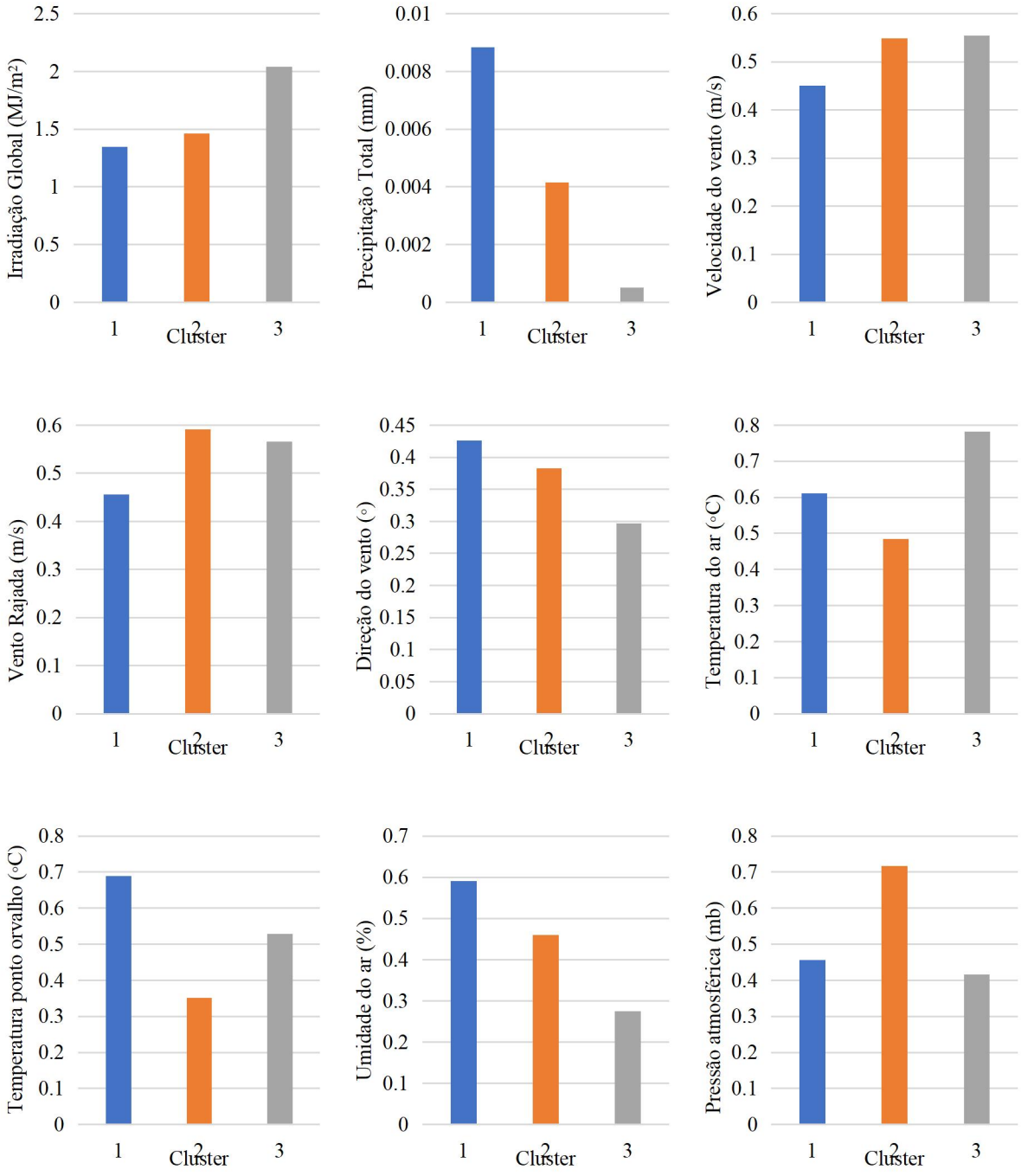


Fonte: Autor (2023)

Uma análise das variáveis de cada cluster é realizada a fim de distinguir as diferenças climatológicas entre os clusters. A Figura 5.2 mostra a média diária dos valores normalizados de cada variável de cada cluster. Nota-se que o cluster 1 contém dados com dias mais chuvosos pois possui nível mais alto de precipitação e nível mais baixo de irradiação solar. Este cluster também possui velocidades de vento mais baixas e um nível de umidade relativa mais alto. Já o cluster 3, ao contrário do cluster 1, contém dados com dias mais ensolarados, pois possui nível mais baixo de precipitação e mais alto de irradiação solar. Este cluster pode ser descrito como o que possui os dias mais calorosos e secos devido à maior velocidade do vento, maior temperatura do ar e menor umidade relativa. Finalmente, o cluster 2 ocupa valores intermediários entre os dois clusters descritos acima. Pode ser descrito como dias frios, não chuvosos, com níveis médios de irradiação solar, umidade relativa e velocidade do vento.

O valor médio diário de irradiação solar no cluster 1 é de 1,34 MJ/m^2, no cluster 2 é de 1,46 MJ/m^2 e no cluster 3 é de 2,03 MJ/m^2. Assim, o cluster 3 tem níveis mais altos de irradiação solar, o que significa que ele tem dados com dias mais ensolarados, em contraste com o cluster 1, que tem níveis mais baixos de irradiação solar.

Figura 5.2 – Média diária dos valores normalizados das variáveis de cada cluster.



Fonte: Autor (2023)

O cluster 1 possui 31,09% dos dados totais, o cluster 2 possui 30,05% e o cluster 3 possui 38,86%. A Figura 5.3 mostra a porcentagem de dias por mês em cada cluster. Nota-se que utilizando a técnica de clusterização, os dados são agrupados de forma diferente da divisão tradicional que segue as estações do ano (verão e inverno, ou período de úmido e seco).

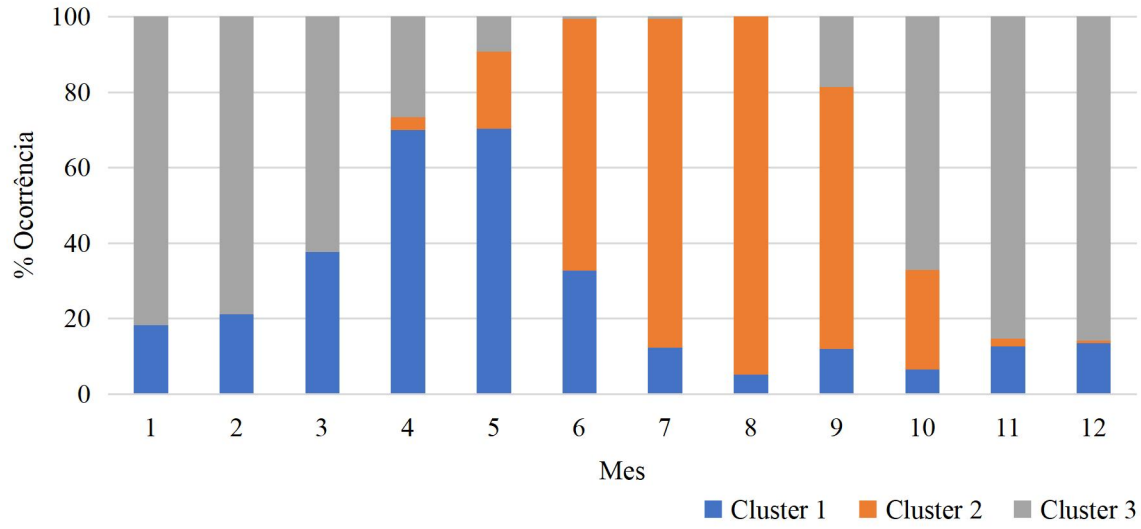


Figura 5.3 – Porcentagem de dias por mês em cada cluster.

Fonte: Autor (2023)

A seção a seguir apresenta os resultados da aplicação da metodologia proposta para a seleção de atributos e atrasos para cada cluster.

5.2 RESULTADO DO MÉTODO *ENSEMBLE* PARA SELEÇÃO DE ATRIBUTOS

Esta seção mostra os resultados da aplicação da metodologia *ensemble* proposta para selecionar os atributos mais significativos de cada cluster. O método proposto integra os seguintes algoritmos: Informação Mútua, *Random Forest* e *Relief*. Para cada método, a importância de cada variável é avaliada e seu valor é normalizado. O ranking final da importância das variáveis é obtido através do cálculo do valor médio dos diferentes métodos utilizados.

A Figura 5.4 mostra o ranking final das variáveis para cada cluster e a Figura 5.5 apresenta o erro médio absoluto utilizando o algoritmo de votação VOA, com as variáveis selecionadas sombreadas em verde. O limiar entre as variáveis selecionadas e as variáveis descartadas foi encontrado empiricamente durante a fase de otimização do modelo utilizando o conjunto de treinamento e validação. Os atributos selecionados variam para cada cluster pois cada um deles possui condições meteorológicas diferentes.

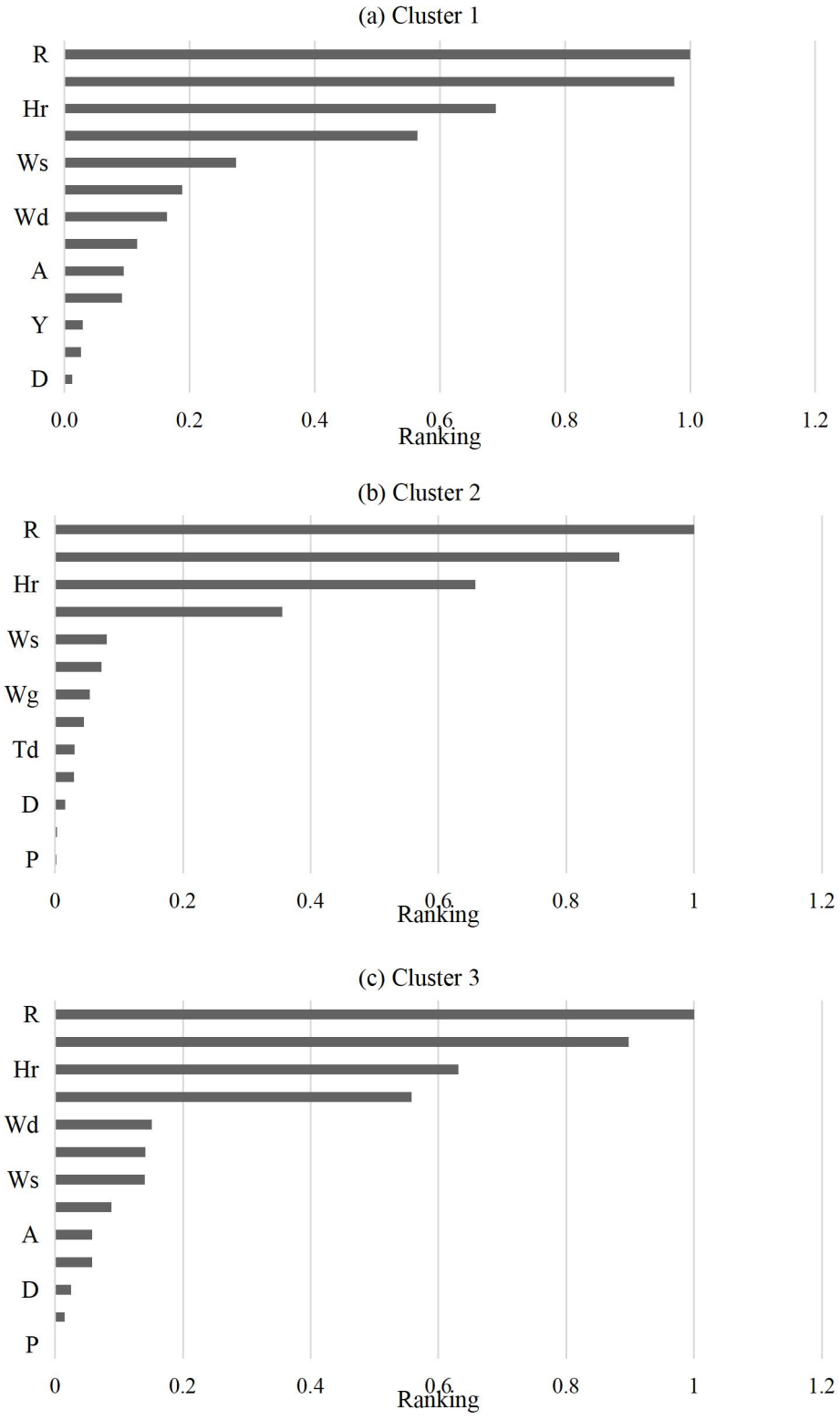
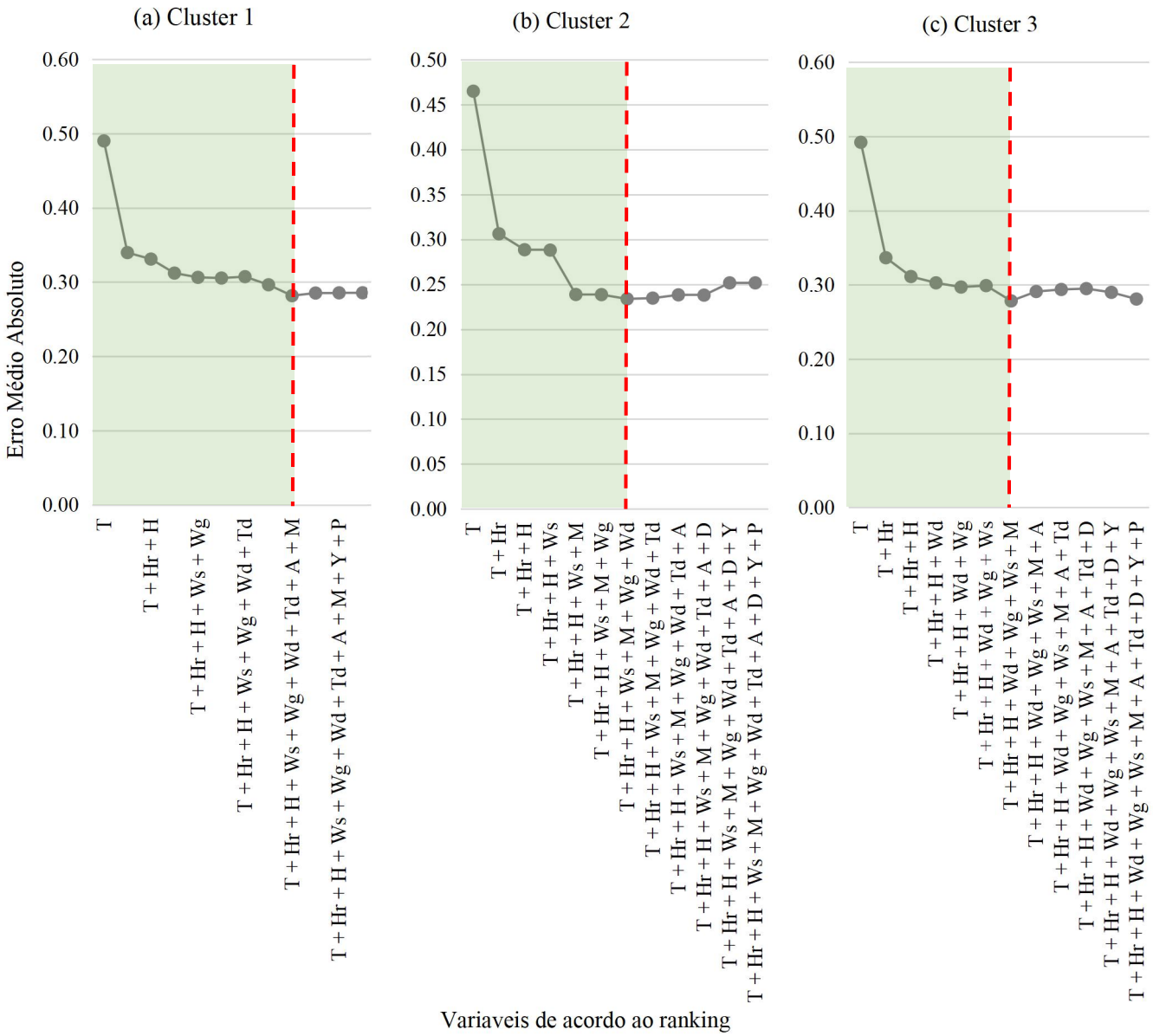


Figura 5.4 – Ranking de importância das variáveis: (a) Cluster 1, (b) Cluster 2 e (c) Cluster 3.

Fonte: Autor (2023)

Figura 5.5 – Desempenho do VOA de acordo com as variáveis de entrada: (a) Cluster 1, (b) Cluster 2 e (c) Cluster 3.



Fonte: Autor (2023)

Os resultados indicam que o cluster 1 tem melhor desempenho quando utiliza como entrada a própria entrada endógena (irradiação global), as variáveis exógenas (temperatura do ar, umidade relativa do ar, velocidade do vento, vento rajada máx., direção do vento, temperatura do ponto do orvalho e pressão atmosférica horaria) e as informações de hora do

dia e mês do ano. Já o cluster 2 tem melhor desempenho quando utiliza como entrada a própria entrada endógena (irradiação global), as variáveis exógenas (temperatura do ar, umidade relativa do ar, velocidade do vento, vento rajada máx. e direção do vento) e as informações de hora do dia e mês do ano. Finalmente o cluster 3 tem melhor desempenho quando utiliza como entrada a própria entrada endógena (irradiação global), as variáveis exógenas (temperatura do ar, umidade relativa do ar, velocidade do vento, vento rajada máx. e direção do vento) e as informações de hora do dia e mês do ano. Os três clusters descartam a utilização da variável exógena precipitação total e as informações de dia do mês e ano. O cluster 2 e 3 descartam adicionalmente a temperatura do ponto de orvalho e a pressão atmosférica.

Após a seleção das variáveis, foi feita a seleção dos atrasos mais significativos de cada variável. Primeiro seleciona-se os atrasos mais significativos da variável endógena e, em seguida, os atrasos mais significativos das variáveis exógenas. Isto é feito separadamente porque as variáveis exógenas são de importância reduzida em relação à variável endógena, porém não mais tênue. Os atrasos foram selecionados e descartados empiricamente utilizando o conjunto de treinamento e validação. O conjunto de entrada final com as variáveis selecionadas e seus atrasos é apresentado na Tabela 5.2.

Tabela 5.2 – Conjunto selecionado de variáveis de entrada e valores de atraso.

Variável	Atrasos
Cluster 1	
Irradiação global	$t-1, t-2, t-23, t-24, t-25, t-48, t-49, t-72$
Temperatura do ar - bulbo seco	$t-1, t-2, t-23, t-24, t-25, t-48, t-49, t-72$
Umidade relativa do ar	$t-1, t-2, t-23, t-24, t-25, t-48, t-49, t-72$
Vento velocidade	$t-1, t-2, t-24, t-25$
Pressão atmosférica ao nível da estação	$t-1, t-2, t-24$
Vento rajada máx.	$t-1, t-24$
Vento direção	$t-1, t-2$
Temperatura do ponto de orvalho	$t-1$
Hora	-
Mes	-
Cluster 2	
Irradiação global	$t-1, t-2, t-23, t-24, t-25, t-48, t-49, t-72$
Temperatura do ar - bulbo seco	$t-1, t-2, t-23, t-24, t-25, t-48, t-49, t-72$
Umidade relativa do ar	$t-1, t-2, t-23, t-24, t-25, t-48, t-49, t-72$
Vento velocidade	$t-1, t-2, t-24, t-48$
Vento rajada máx.	$t-1$
Vento direção	$t-1$
Hora	-
Mes	-
Cluster 3	

Irradiação global	t − 1, t − 2, t − 23, t − 24, t − 25, t − 47, t − 48, t − 72
Temperatura do ar - bulbo seco	t − 1, t − 2, t − 23, t − 24, t − 25, t − 48, t − 49, t − 72
Umidade relativa do ar	t − 1, t − 2, t − 23, t − 24, t − 25, t − 48, t − 49, t − 72
Vento velocidade	t − 1, t − 2, t − 24, t – 25
Vento rajada máx.	t − 1, t – 2
Vento direção	t − 1, t – 2
Hora	-
Mês	-

Para investigar a eficácia do uso do método *ensemble* para seleção de atributos, três casos são analisados:

- Caso 1: O modelo de previsão é treinado utilizando apenas entradas endógenas, que são irradiação solar e suas 10 observações anteriores;
- Caso 2: O modelo de previsão é treinado usando entradas endógenas e exógenas (irradiação solar e outros dados meteorológicos), e suas observações passadas são selecionadas usando o coeficiente de correlação de Pearson;
- Caso 3: O modelo de previsão é treinado usando entradas endógenas e exógenas, selecionados pelo método proposto *ensemble* de seleção de atributos.

O algoritmo utilizado para realizar esta análise é o VOA. Para uma comparação justa, os hiper parâmetros são mantidos iguais nos três casos, os modelos são treinados e testados utilizando o mesmo conjunto de dados de treinamento e validação. Os resultados são apresentados na Tabela 5.3. Os resultados mostram que dentre todos os casos analisados, o caso 3, que utiliza a metodologia proposta neste trabalho, mostra maior precisão de previsão de acordo com todas as medidas de avaliação.

Tabela 5.3 – Desempenho da previsão para diferentes conjuntos de entrada utilizando VOA.

Conjunto de entrada	MAE	RMSE	MAPE	R^2
Cluster 1				
Caso 1: endógenas	0.330	0.459	43.155	0.734
Caso 2: endógenas + exógenas (coeficiente de Pearson)	0.318	0.444	42.572	0.752
Caso 3: endógenas + exógenas (*ensemble*)	**0.316**	**0.441**	**39.599**	**0.756**
Cluster 2				
Caso 1: endógenas	0.257	0.355	37.913	0.864
Caso 2: endógenas + exógenas (coeficiente de Pearson)	0.276	0.363	33.518	0.858
Caso 3: endógenas + exógenas (*ensemble*)	**0.252**	**0.336**	**29.925**	**0.878**
Cluster 3				
Caso 1: endógenas	0.268	0.408	20.840	0.844
Caso 2: endógenas + exógenas (coeficiente de Pearson)	0.260	0.386	19.257	0.860
Caso 3: endógenas + exógenas (*ensemble*)	**0.236**	**0.365**	**16.586**	**0.875**

A Figura 5.6 ilustra a curva de aprendizagem obtida com o VOA para todos os casos. Esta curva mostra como o erro de previsão muda conforme o tamanho do conjunto de treinamento aumenta ou diminui. É possível diagnosticar problemas de viés e variância nos modelos de aprendizagem supervisionada através de sua análise. Há muitas informações a serem extraídas desta curva. Nota-se que quando o tamanho do conjunto de treinamento é baixo o erro do conjunto de treinamento é zero. Este é um comportamento normal, pois o modelo não tem dificuldade de se ajustar perfeitamente a um único dado de entrada, sendo a previsão perfeita. Entretanto, quando testado com o conjunto de validação, o erro tem valor alto, pois o modelo não tem capacidade de generalização para casos que não vistos no treinamento.

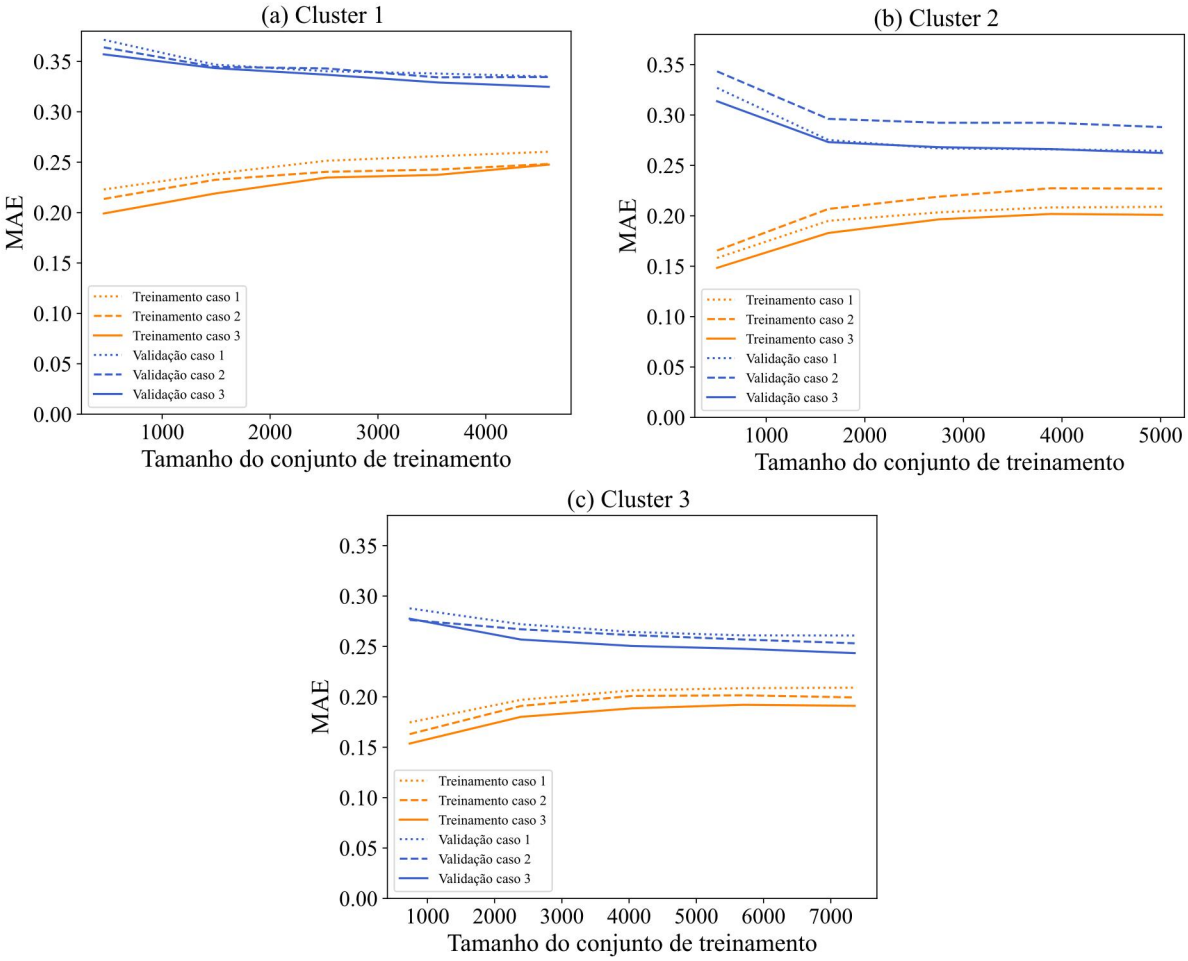


Figura 5.6 – Curvas de aprendizado para o VOA: (a) Cluster 1, (b) Cluster 2 e (c) Cluster 3

Fonte: Autor (2023)

Em todos os casos, à medida que o tamanho do conjunto de treinamento aumenta, o erro de treinamento aumenta e o erro de validação diminui, convergindo para um baixo valor de erro, que é o comportamento desejável. Isto indica um baixo viés, que é a diferença entre a previsão média do modelo e o valor correto que ele está tentando prever. Um modelo com um alto viés é menos sensível aos dados de treinamento e simplifica em excesso o modelo resultando em altos erros nos dados de treinamento e validação.

A estreita diferença entre as curvas de treinamento e validação indica uma variância baixa. A variância refere-se às variações no modelo quando diferentes porções do conjunto de dados de treinamento são utilizadas. Um modelo com alta variância é muito sensível aos dados de treinamento e não generaliza os dados desconhecidos. O caso 3 tem erros de treinamento e validação menores. Isto destaca o impacto positivo da aplicação do método *ensemble* de seleção de atributos proposto e do método de seleção de atrasos, que mantém os atributos e seus respectivos atrasos significativos e descarta aqueles que podem afetar negativamente o processo de aprendizagem.

5.3 OTIMIZAÇÃO DE HIPER PARÂMETROS

Para otimização dos hiper parâmetros foi o utilizado a estratégia da busca em grade, que apresenta uma forma de resolução simples e direta, de fácil implementação e paralelização. No método de busca em grade, cada hiper parâmetro é delimitado em torno de um intervalo particular de busca, no qual acredita-se que seja um potencial local para a varredura. Para cada hiper parâmetro, é estabelecido uma resolução de grade que determina a quantidade de pontos candidatos a serem considerados para cada um destes.

No *scikit-learn*, a busca em grade é implementada através de 'GridSearchCV' onde se avalia cada hiper parâmetro no espaço de configuração previamente definido, e seu desempenho é avaliado por validação cruzada. Após todas as instâncias do espaço de configuração a serem avaliadas, a combinação ideal de hiper parâmetros no espaço de busca definido é obtida com sua pontuação de desempenho. A Tabela 5.4 lista as combinações de hiper parâmetros exploradas de forma a otimizar cada modelo através da técnica ‘GridSearchCV’.

Tabela 5.4 – Espaço de hiper parâmetros explorados para cada algoritmo.

Algoritmo	Hiper parâmetro	Valores investigados
SVR	C	0.1, 1, 10, 100, 1000
	gamma	'auto', 'scale', 0.001, 0.01, 1, 10
	kernel	‘poly’, ‘rbf’, ‘sigmoid’

RF	max_depth	10, 11, 12, 13, 14, 15
	n_estimators	200, 300, 400, 500, 600, 700
	max_features	1, 2, 3
XGBT	learning_rate	0.1, 0.2, 0.3
	max_depth	3, 4, 5, 6, 10
	n_estimators	20, 40, 60, 80, 100, 120, 140, 160
	subsample	0.5, 0.6, 0.7, 0.8, 1
	colsample_bytree	0.2, 0.4, 0.6, 0.8, 1
CatBoost	depth	5, 6, 7, 8, 9, 10
	l2_reg	0.2, 0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
	learning_rate	0.05, 0.1, 0.2, 0.3
	iterations	500, 1000, 1800, 2000, 2200, 2500
AdaBoost	learning_rate	0.1, 0.2, 0.3
	n_estimators	10, 20, 30, 40, 50, 100, 150, 200, 300

A Tabela 5.5 mostra o melhor valor obtido para os hiper parâmetros para cada algoritmo de acordo com a implementação da técnica de busca de grade para cada cluster.

Tabela 5.5 – Resultados da otimização de hiper parâmetros.

Algoritmo	Hiper parâmetro	Cluster 1	Cluster 2	Cluster 3
SVR	C	10	100	10
	gamma	auto	auto	Auto
	kernel	rbf	rbf	Rbf
RF	max_depth	12	11	14
	n_estimators	400	500	600
	max_features	3	3	3
XGBT	learning_rate	0.1	0.1	0.1
	max_depth	4	4	5
	n_estimators	80	120	100
	subsample	0.9	0.6	1.0
	colsample_bytree	0.8	0.8	0.8
CatBoost	depth	6	6	7
	l2_reg	4	4	0.2
	learning_rate	0.05	0.05	0.05
	iterations	2000	2200	1800
AdaBoost	learning_rate	0.1	0.2	0.2
	n_estimators	50	30	40

5.4 DESEMPENHO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Esta seção mostra os resultados dos diferentes modelos de aprendizado de máquina para previsão da irradiação solar. Os modelos utilizam como entrada o mesmo conjunto de dados de treinamento e validação durante o processo de aprendizagem dos algoritmos. Cada cluster utiliza como entrada diferentes atributos e atrasos segundo os resultados da

metodologia *ensemble* de seleção de atributos apresentados na Tabela 5.2 e diferentes valores de hiper parâmetros segundo os resultados da busca em grade apresentados na Tabela 5.5.

A Tabela 5.6 apresenta os resultados das medidas de avaliação para os modelos de AM utilizados. Como pode ser observado, o CatBoost apresenta melhor desempenho para todos as medidas em todos os clusters. Os menores valores de erro são obtidos para o cluster 3, que tem características de dias mais ensolarados, com baixos níveis de precipitação e níveis mais altos de irradiação solar, ao contrário do cluster 1, que apresenta as medidas menos satisfatórias. Todos os algoritmos apresentam coeficiente de correlação satisfatório em torno de 0,8, o que evidencia que eles interpretam os dados corretamente. Em todos os clusters, o desempenho menos satisfatório é apresentado pelo modelo AdaBoost. A Figura 5.7 a seguir compara as quatro medidas de avaliação entre os clusters.

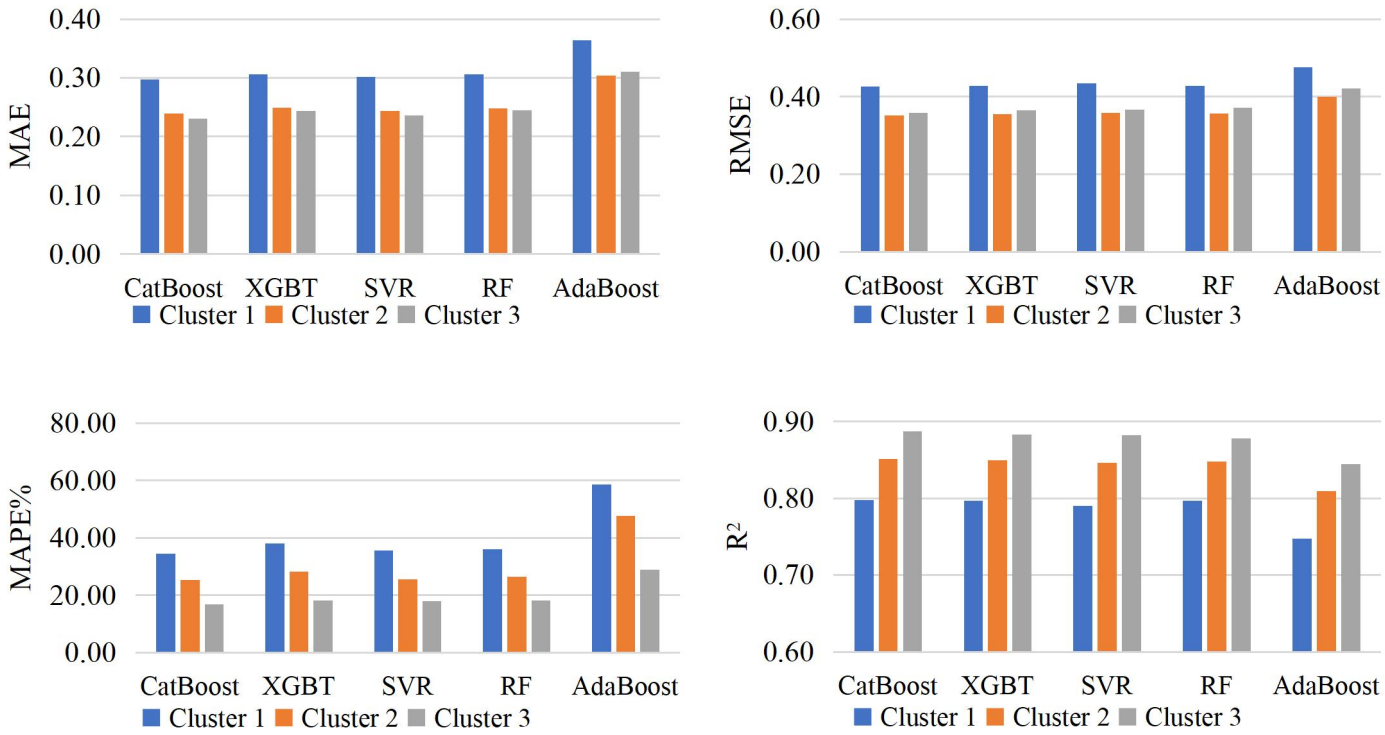
A Tabela 5.7 apresenta algumas medidas estatísticas dos resultados dos modelos de previsão. Observa-se que os valores mínimos de previsão dos modelos estão todos em torno de 0 MJ/m^2, com exceção do AdaBoost, que está em torno de 0,3 MJ/m^2. O modelo com o valor mínimo mais próximo de 0 MJ/m^2 é o SVR. Analisando-se os valores máximos de previsão, o SVR contém os valores mais altos em todos os clusters e o AdaBoost, os valores mais baixos. Todos os valores máximos dos modelos estão próximos de 3,3 MJ/m^2. Os valores medianos de previsão dos modelos apresentam-se bastante semelhantes entre si. Da mesma forma, observa-se que os valores do desvio padrão e da variância situam-se em torno de 0,9 e 0,8, respectivamente, o que indica a dispersão das previsões em torno da média.

Tabela 5.6 – Resultados dos modelos de aprendizado de máquina para previsão da irradiação solar.

	MAE	**RMSE**	**MAPE (%)**	**R^2**
		Cluster 1		
CatBoost	**0.297**	**0.426**	**34.457**	**0.798**
XGBT	0.306	0.427	38.171	0.797
SVR	0.302	0.434	35.617	0.791
RF	0.306	0.427	35.963	0.797
AdaBoost	0.364	0.476	58.716	0.748
		Cluster 2		
CatBoost	**0.239**	**0.352**	**25.366**	**0.852**
XGBT	0.249	0.355	28.328	0.850
RF	0.248	0.356	26.457	0.848
SVR	0.244	0.359	25.465	0.846
AdaBoost	0.304	0.399	47.595	0.809
		Cluster 3		
CatBoost	**0.230**	**0.358**	**16.968**	**0.887**
XGBT	0.243	0.364	18.184	0.883
SVR	0.236	0.366	17.939	0.882
RF	0.245	0.372	18.195	0.878

AdaBoost	0.311	0.420	28.844	0.844

Figura 5.7 – Comparação das medidas de avaliação dos algoritmos.



Fonte: Autor (2023)

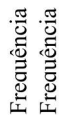
Tabela 5.7 – Análises estatísticas dos resultados dos modelos de aprendizado de máquina para previsão da irradiação solar.

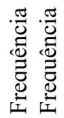
	Min.	Max.	Moda	Mediana	Média	Desvio Padrão	Variância
Cluster 1							
CatBoost	0.038	3.417	0.200	1.278	1.371	0.870	0.757
XGBT	0.075	3.492	0.113	1.289	1.358	0.838	0.702

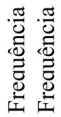
SVR	0.015	3.515	0.183	1.292	1.371	0.874	0.764
RF	0.039	3.306	0.222	1.325	1.370	0.822	0.676
AdaBoost	0.317	2.696	0.317	1.395	1.387	0.727	0.529
Cluster 2							
CatBoost	0.044	3.397	0.210	1.313	1.363	0.873	0.763
XGBT	0.051	3.419	0.147	1.315	1.357	0.849	0.721
RF	0.075	3.394	0.161	1.320	1.361	0.848	0.720
SVR	0.017	3.713	0.180	1.317	1.356	0.877	0.768
AdaBoost	0.308	2.756	0.308	1.335	1.367	0.767	0.589
Cluster 3							
CatBoost	0.111	3.507	0.212	2.103	1.999	1.032	1.066
XGBT	0.123	3.522	0.184	2.050	1.968	1.016	1.032
SVR	0.093	3.595	0.261	2.146	2.005	1.031	1.063
RF	0.133	3.438	0.230	2.053	1.967	1.009	1.019
AdaBoost	0.369	3.171	0.369	2.029	1.970	0.913	0.834

A Figura 5.8 mostra o histograma dos erros absolutos de previsão obtidos com os modelos CatBoost e AdaBoost. No CatBoost, o pico da distribuição de erros encontra-se mais centrado em zero de forma mais acentuada em todos os clusters, indicando a existência de baixos erros na maioria das previsões. No modelo AdaBoost, a distribuição de erros é claramente mais dispersa, indicando erros de previsão maiores.

Figura 5.8 – Histograma dos erros absolutos: (a) CatBoost e (b) AdaBoost.







Fonte: Autor (2023)

Adicionalmente, é importante avaliar o desempenho computacional dos modelos ao lidar com aplicações do mundo real. Os resultados são apresentados na Tabela 5.8. Em todos os clusters, o modelo com o tempo de treinamento mais alto é o RF, com uma média de 62,3 segundos. O modelo com o menor tempo de treinamento é o XGBT, com uma média de 1,5 segundos. A tabela indica que os tempos de treinamento do cluster 3 são ligeiramente mais altos do que os dos outros dois clusters. É possível explicar isso pelo fato de o cluster 3 ter uma porcentagem maior de dados, isto é, há mais dias com tempo ensolarado no banco de dados. Todos os algoritmos possuem uma velocidade de previsão satisfatória, com um tempo computacional médio de 0,4 segundos.

Tabela 5.8 –Tempo de treinamento e previsão dos modelos de aprendizado de máquina para previsão da irradiação solar.

Tempo de treinamento (s)					
	CatBoost	XGBT	SVR	RF	AdaBoost
Cluster 1	15.351	1.391	3.997	41.586	2.445
Cluster 2	15.351	1.391	3.997	41.586	2.445
Cluster 3	23.451	1.973	7.505	103.968	2.459
Tempo de previsão (s)					
	CatBoost	XGBT	SVR	RF	AdaBoost
Cluster 1	0.010	0.013	2.279	0.218	0.036
Cluster 2	0.007	0.015	1.572	0.195	0.021
Cluster 3	0.009	0.008	1.722	0.216	0.019

5.5 DESEMPENHO DOS MODELOS DE VOTAÇÃO

Dois modelos de votação foram testados, sendo estes a votação por média simples (VOA) e a votação por média ponderada (VOWA). Em síntese, foram explorados os seguintes modelos de votação:

- VOA: média simples de CatBoost + XGBT + SVR + RF + AdaBoost
- VOWA: média ponderada de CatBoost + XGBT + SVR + RF + AdaBoost

Em relação ao regressor VOWA, um coeficiente de ponderação deve ser especificado para cada membro do *ensemble*. Neste trabalho, é atribuído um número inteiro que denota o número de votos concedidos ao membro do conjunto correspondente. Os pesos adotados no VOWA são mostrados na Tabela 5.9. O modelo de mais alto desempenho tem o maior número de votos, e o modelo de menor desempenho tem apenas um voto.

Tabela 5.9 – Pesos dos integrantes dos *ensemble* models, VOWA.

	CatBoost	**XGBT**	**SVR**	**RF**	**AdaBoost**
Cluster 1	5	4	3	2	1
Cluster 2	5	4	2	3	1
Cluster 3	5	4	3	2	1

Para ambos os regressores, o mesmo conjunto de dados é utilizado para treinamento, validação e teste, mantendo as mesmas entradas selecionadas para cada cluster. Os resultados das medidas de avaliação são apresentados na Tabela 5.10 e comparados adicionalmente com o modelo de previsão de melhor desempenho, o CatBoost.

Tabela 5.10 – Resultado dos modelos de votação para previsão da irradiação solar.

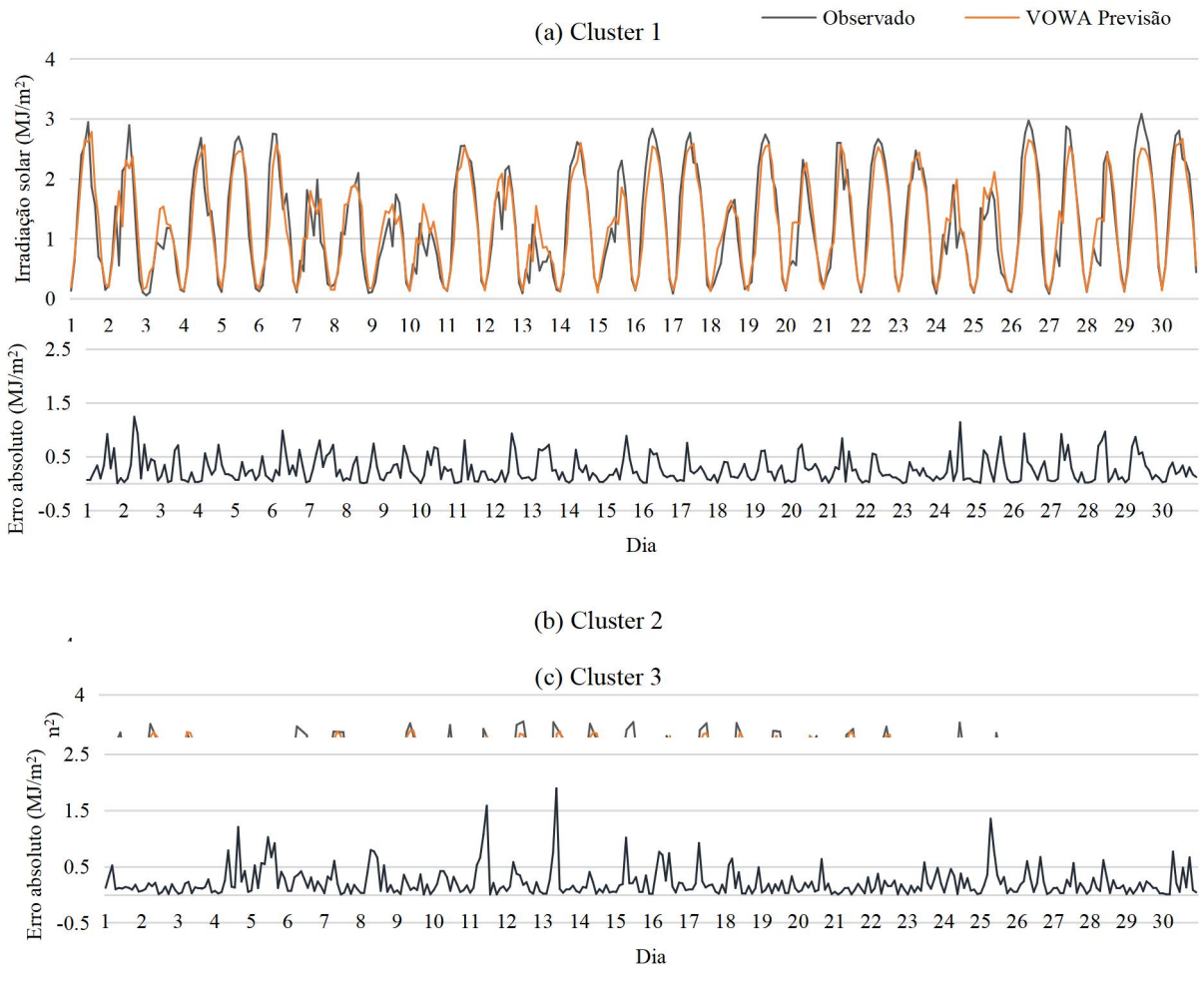
	MAE	**RMSE**	**MAPE (%)**	**R^2**
		Cluster 1		
VOA	0.303	0.423	38.626	0.800
VOWA	**0.296**	**0.421**	**34.059**	**0.802**
CatBoost	0.297	0.426	34.457	0.798
		Cluster 2		
VOA	0.244	0.350	28.423	0.854
VOWA	**0.230**	**0.348**	**25.312**	**0.855**
CatBoost	0.239	0.352	25.326	0.852
		Cluster 3		
VOA	0.241	0.361	18.924	0.886
VOWA	**0.230**	**0.356**	**16.640**	**0.888**
CatBoost	0.230	0.358	16.968	0.887

Uma das desvantagens da votação simples pode ser analisada na tabela de resultados. Na abordagem VOA, todos os modelos são considerados igualmente eficazes. No entanto, essa situação é pouco provável, especialmente se diferentes algoritmos de AM forem utilizados. Nesse caso, o modelo AdaBoost é considerado tão eficiente quanto o modelo CatBoost, uma situação que não é verídica, conforme discutido na seção anterior. A votação do modelo AdaBoost na votação final do *ensemble* é capaz de enviesar os resultados, de modo que o VOA apresenta resultados satisfatórios, porém, não supera o desempenho individual apresentado pelo modelo CatBoost.

A correção desse cenário por meio da votação ponderada, que atribui votos a cada modelo segundo seu desempenho individual, proporciona a melhora dos resultados, conforme ilustrado na tabela acima. O VOWA supera todas as quatro medidas de avaliação propostas nos três clusters diferentes.

A Figura 5.9 mostra a irradiação solar horária observada e prevista, além dos resíduos obtidos utilizando o modelo VOWA. Como os dados são agrupados em clusters com base nas condições meteorológicas de cada dia, eles perdem a continuidade. Por esse motivo, a figura apresenta os últimos 30 dias do conjunto de teste de cada cluster. Os resultados mostram que o modelo de previsão pode acompanhar as variações na irradiação solar.

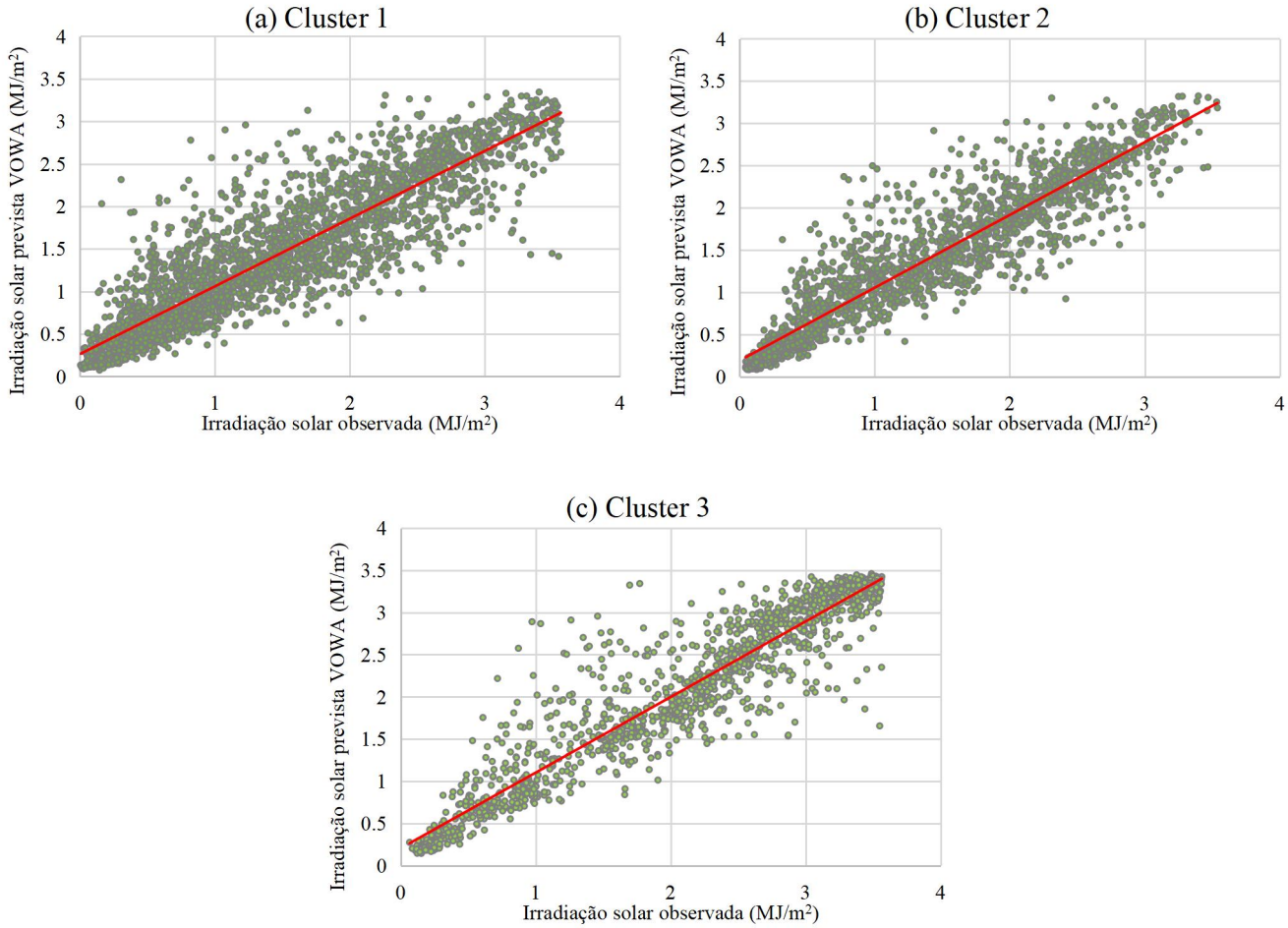
Figura 5.9 – Irradiação solar observada e prevista utilizando o VOWA: (a) Cluster1, (b) Cluster 2 e (c)



Cluster 3.

A Figura 5.10 mostra uma relação linear entre a irradiação solar observada e a prevista pelo modelo VOWA para cada cluster no conjunto de dados de teste. Pode-se ver na figura que os valores previstos estão fortemente correlacionados com os dados de irradiação solar observados para todos os três clusters.

Figura 5.10 – Gráfico de dispersão da irradiação solar utilizando o VOWA: (a) Cluster 1, (b) Cluster 2 e (c) Cluster 3



Fonte: Autor (2023)

5.6 RESULTADOS DA ANÁLISE ESTATÍSTICA

Nesta seção, são discutidos os resultados do teste Diebold-Mariano, incluindo os valores de *p* e as conclusões estatísticas relacionadas à significância das diferenças observadas entre os modelos. Essas análises são fundamentais para a tomada de decisões para a seleção do modelo mais adequado para a previsão no contexto de estudo.

A Tabela 5.11 apresenta os resultados do teste DM, que comparou o desempenho do modelo de votação proposto, VOWA, com os demais algoritmos de previsão em pares. Em todos os casos, os valores de p foram inferiores ao limite de 0,05, o que permite rejeitar a hipótese nula H_0. Isso indica que as diferenças observadas são estatisticamente significativas e que o modelo VOWA proposto é significativamente mais preciso do que os outros modelos.

Com base nesse teste DM, pode-se concluir que o modelo VOWA demonstrou um desempenho superior em relação aos demais algoritmos de previsão analisados. Essa descoberta é de grande importância, pois a significância estatística dos resultados reforça e corrobora a confiabilidade dos estudos anteriores que já indicavam a eficácia do modelo VOWA na obtenção de previsões mais precisas (ERDEBILLI; DEVRIM-İÇTENBAŞ, 2022; NATRAS; SOJA; SCHMIDT, 2022; PHYO; BYUN; PARK, 2022) . Essa consistência de resultados fortalece a confiança na capacidade do modelo VOWA de oferecer melhorias significativas no campo da previsão.

Em resumo, o teste de DM demonstra que o modelo VOWA proposto é significativamente mais preciso do que os outros modelos de previsão que foram comparados.

Tabela 5.11 – Resultados do teste Diebold - Mariano.

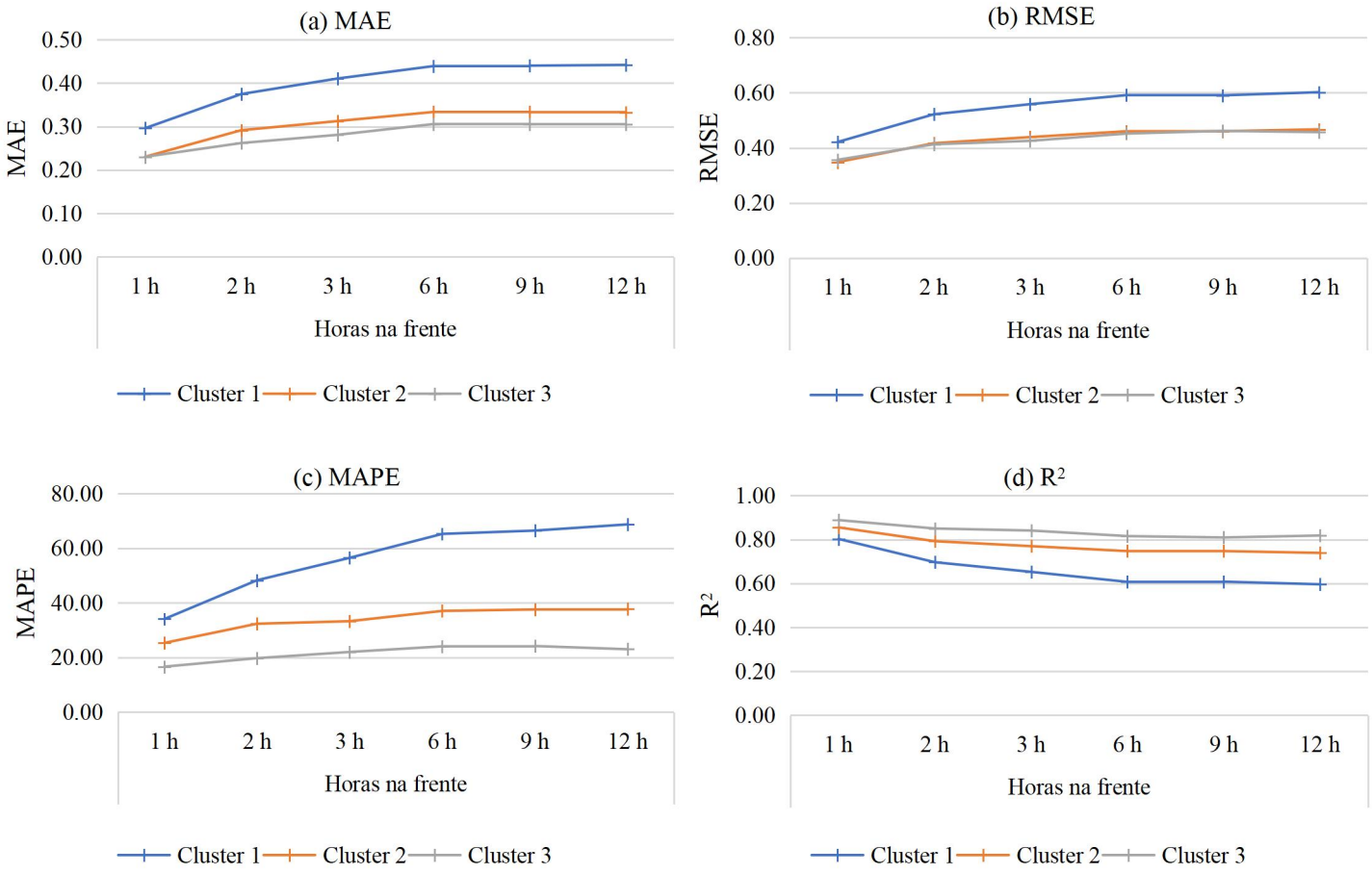
	Valor-*p*		
	Cluster 1	**Cluster 2**	**Cluster 3**
CatBoost - VOWA	0.0001	0.0019	0.0247
XGBT - VOWA	8.86 x 10^(-5)	2.14 x 10^(-5)	0.0017
SVR - VOWA	2.06 x 10^(-6)	0.0006	0.0118
RF - VOWA	0.0009	0.0001	0.0001
AdaBoost - VOWA	2.00 x 10^(-41)	3.31 x 10^(-35)	8.32 x 10^(-26)

5.7 RESULTADOS PARA DIFERENTES HORIZONTES DE PREVISÃO

O modelo proposto VOWA foi avaliado para diferentes horizontes de previsão, de 1 a 12 horas. A Figura 5.11 mostra o resultado. Como esperado, os erros de previsão aumentam com o aumento do horizonte de previsão e o coeficiente de determinação diminui, indicando uma deterioração na qualidade da previsão do algoritmo. É importante destacar na figura que as medidas de avaliação variam mais significativamente para horizontes de previsão de até 3 horas. Depois disso, o desempenho do algoritmo permanece relativamente estável, o que é positivo. Vale ressaltar o fato de que o cluster 1 tem maior erro e maior degradação do desempenho da previsão. Pode-se explicar essa situação pelo fato de que esse cluster contém

dados de dias com alta precipitação e baixa irradiação solar e, portanto, sofre maior intermitência.

Figura 5.11 – Medidas de avaliação para a previsão da irradiação solar em diferentes horizontes utilizando o VOWA: (a) MAE , (b) RMSE, (c) MAPE e (d) R^2



Fonte: Autor (2023)

6 CONCLUSÕES

6.1 CONSIDERAÇÕES SOBRE OS RESULTADOS ALCANÇADOS

A análise deste trabalho propõe uma metodologia inovadora para previsão da irradiação solar, buscando melhorar a precisão dos resultados por meio da utilização de técnicas avançadas de AM. O objetivo é obter previsões mais precisas e confiáveis da irradiação solar, um fator essencial na geração de energia solar e no planejamento de sistemas de energia renovável.

A metodologia proposta consiste em várias etapas bem definidas. Inicialmente, é realizada a etapa de pré-processamento, os dados meteorológicos são limpos e transformados, tratando valores ausentes e discrepantes, normalizando variáveis e removendo variáveis correlacionadas. Em seguida, um algoritmo de clusterização é aplicado para dividir o conjunto de dados em clusters com características semelhantes. A seleção do número de clusters é baseada na comparação de três índices. O resultado são três clusters com diferentes condições meteorológicas, como dias ensolarados, dias com níveis médios de precipitação e irradiação, e dias com alta variabilidade de irradiação solar devido à presença de nuvens e precipitação.

Em seguida, é realizada a seleção de atributos, utilizando uma combinação de algoritmos, como Informação Mútua, *Random Forest* e *Relief*. Além disso, é realizada uma seleção de observações passadas das variáveis. Essas técnicas permitem identificar os atributos mais relevantes para a previsão da irradiação solar, levando em consideração tanto as variáveis meteorológicas endógenas quanto as exógenas, bem como suas observações passadas mais relevantes. Essa seleção de atributos é fundamental para reduzir a dimensionalidade dos dados e melhorar a eficiência dos modelos de AM.

A vantagem do método de seleção de atributos proposto foi validada pela comparação dos resultados obtidos com dois outros casos:

- quando apenas entradas endógenas são utilizadas; e
- quando são utilizadas entradas endógenas e exógenas, selecionadas com o coeficiente de correlação de Pearson.

Os modelos de aprendizado de máquina utilizados nessa metodologia são considerados de última geração e incluem algoritmos como SVR, XGBT, RF, AdaBoost e CatBoost. Cada modelo é treinado com os dados dos diferentes clusters, e seu desempenho é avaliado usando medidas comuns, como MAE, RMSE, MAPE e R^2.

Os resultados obtidos mostraram que o CatBoost obteve o melhor desempenho de previsão em relação aos outros modelos. Ele apresentou medidas de desempenho bastante promissoras, incluindo um MAE de 0,259, um RMSE de 0,379, um MAPE de 26,283% e um R^2 de 0,846. Por outro lado, o AdaBoost teve o desempenho mais fraco, evidenciando a importância de selecionar o modelo de AM mais adequado para cada problema específico.

Uma etapa crucial do trabalho foi a implementação do modelo de votação, que combina as previsões dos diferentes modelos em uma única previsão final. Esse modelo de votação foi testado com duas abordagens diferentes: votação simples e votação ponderada. Os resultados mostraram que a votação ponderada média obteve um desempenho de previsão superior em comparação com a votação simples e os modelos de aprendizado de máquina individuais. Isso reforça a importância de considerar a contribuição relativa de cada modelo para a previsão final.

Por fim, a metodologia proposta apresentou resultados promissores em termos de desempenho de previsão e velocidade de execução computacional. Todos os modelos de aprendizado de máquina investigados mostraram-se adequados para aplicações práticas, com um tempo médio de previsão de apenas 0,42 s. Além disso, a estabilidade das medidas de erro a partir de 3 horas de previsão indica que a metodologia é capaz de fornecer resultados confiáveis em horizontes de previsão mais longos.

Em suma, este trabalho de pesquisa propõe uma metodologia abrangente e inovadora para a previsão solar, incorporando técnicas avançadas de clusterização, seleção de atributos e *ensemble models*. Os resultados obtidos destacam a importância dessas abordagens para melhorar a precisão e confiabilidade das previsões de irradiação solar, contribuindo assim para o avanço da energia solar e sistemas de energia renovável.

Os principais resultados e conclusões estão resumidos a seguir:

- A metodologia de clusterização proposta separa os dados em três tipos de dias com condições meteorológicas claramente diferentes: dias de céu claro e ensolarados, dias com níveis médios de precipitação e irradiação, e dias com níveis mais altos de precipitação, com maior intermitência de irradiação solar.
- A seleção de atributos proposta superou os outros dois casos analisados: um utilizando apenas variáveis endógenas como entradas e outro utilizando variáveis endógenas e exógenas selecionadas com base no coeficiente de correlação de Pearson.

- Todos os modelos de AM investigados apresentaram um desempenho de previsão aceitável. Entre todos os algoritmos, o VOWA obteve o melhor desempenho de previsão, superando os demais modelos nos três clusters.
- Todos os algoritmos apresentaram uma velocidade de previsão adequada para aplicações do mundo real, com um tempo médio de previsão computacional de 0,42 s.
- À medida que o horizonte de previsão aumenta, as medidas de erro se deterioram rapidamente; no entanto, a partir de 3 horas, elas se estabilizam, o que é um aspecto positivo.

6.2 SUGESTÕES PARA TRABALHOS FUTUROS

Os resultados obtidos neste trabalho são considerados interessantes e promissores, demonstrando avanços significativos na área de previsão solar. No entanto, é importante ressaltar que existem questões cruciais que ainda requerem uma análise mais aprofundada.

Um dos aspectos a ser investigado mais profundamente é o número máximo de observações passadas adotado no método de seleção de atributos. Neste estudo, o valor foi escolhido empiricamente, mas é necessário realizar uma investigação mais detalhada para determinar a seleção ideal desse parâmetro. A escolha adequada do número de observações passadas é crucial para obter previsões mais precisas e confiáveis da irradiação solar.

Além disso, a seleção dos pesos adequados no método de votação por média ponderada é uma tarefa desafiadora. Apesar dos resultados promissores obtidos com a votação ponderada, é necessário explorar algoritmos de otimização que possam aprimorar ainda mais a seleção desses pesos. Isso permitirá melhorar o desempenho e a precisão das previsões da irradiação solar.

Outro aspecto importante a considerar é a aplicabilidade da metodologia proposta em diferentes conjuntos de dados provenientes de locais com condições meteorológicas distintas do Brasil. A generalização da metodologia para outras regiões possibilitaria verificar sua eficácia e adaptabilidade em diferentes contextos climáticos, contribuindo para a sua validação e ampliação de aplicabilidade.

Outrossim, a metodologia proposta neste estudo pode ser estendida para lidar com outros problemas de previsão, como previsão da velocidade do vento e a previsão de carga em sistemas de energia. Explorar essas áreas adicionais de aplicação poderia evidenciar a

versatilidade e utilidade da metodologia em diversos cenários relacionados à energia renovável.

Para avançar nessa linha de pesquisa, é fundamental realizar estudos futuros que explorem a otimização dos parâmetros, o aprimoramento da seleção de pesos dos algoritmos na votação ponderada e a aplicação da metodologia em diferentes conjuntos de dados e problemas de previsão. Esses estudos contribuiriam para um maior entendimento das limitações e possíveis melhorias da metodologia proposta, fortalecendo assim o conhecimento científico na área de previsão solar e sistemas de energia renovável.

REFERÊNCIAS BIBLIOGRÁFICAS

ABDELLATIF, A. et al. Forecasting Photovoltaic Power Generation with a Stacking Ensemble Model. **Sustainability**, v. 14, n. 17, p. 11083, 5 set. 2022.

ABSOLAR. **Energia Solar Fotovoltaica no Brasil**. São Paulo: [s.n.]. Disponível em: <https://www.absolar.org.br/mercado/infografico/>. Acesso em: 31 ago. 2022.

AGRAWAL, T. **Hyperparameter Optimization in Machine Learning**. Berkeley, CA: Apress, 2021.

ALCAÑIZ, A. et al. Trends and gaps in photovoltaic power forecasting with machine learning. **Energy Reports**, v. 9, p. 447–471, dez. 2023.

ALKAHTANI, H.; ALDHYANI, T. H. H.; ALSUBARI, S. N. Application of Artificial Intelligence Model Solar Radiation Prediction for Renewable Energy Systems. **Sustainability**, v. 15, n. 8, p. 6973, 21 abr. 2023.

ALNUAIMI, N. et al. Streaming feature selection algorithms for big data: A survey. **Applied Computing and Informatics**, v. 18, n. 1/2, p. 113–135, 1 mar. 2022.

AN, K.; MENG, J. **Voting-Averaged Combination Method for Regressor Ensemble**. International Conference on Intelligent Computing. **Anais**...2010.

ANEEL. **Sistema de Informações de Geração**. Disponível em: <https://app.powerbi.com/view?r=eyJrIjoiNjc4OGYyYjQtYWM2ZC00YjllLWJlYmEtYzdkNTQ1MTc1NjM2IiwidCI6IjQwZDZmOWI4LWVjYTctNDZhMi05MmQ0LWVhNGU5YzAxNzBlMSIsImMiOjR9>. Acesso em: 16 nov. 2022.

BOUBAKER, S. et al. Deep Neural Networks for Predicting Solar Radiation at Hail Region, Saudi Arabia. **IEEE Access**, v. 9, p. 36719–36729, 2021.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BUENO PEREIRA, E. et al. **Atlas Brasileiro de Energia Solar**. 1. ed. São Jose dos Campos: INPE, 2006.

BUENO PEREIRA, E. et al. **Atlas Brasileiro de Energia Solar**. 2. ed. São José dos Campos: INPE, 2017.

CASTANGIA, M. et al. A compound of feature selection techniques to improve solar radiation forecasting. **Expert Systems with Applications**, v. 178, p. 114979, set. 2021.

CHEN, T.; GUESTRIN, C. **XGBoost**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **Anais**...New York, NY, USA: ACM, 13 ago. 2016.

CICHY, C.; RASS, S. An Overview of Data Quality Frameworks. **IEEE Access**, v. 7, p. 24634–24648, 2019.

COLAK, I. et al. **Multi-period Prediction of Solar Radiation Using ARMA and ARIMA Models**. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). **Anais**...IEEE, dez. 2015.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, set. 1995.

COVER, T. M.; THOMAS, J. A. **Elements of Information Theory**. 2. ed. Hoboken: Wiley-Interscience, 2006.

CRESESB. **Tutorial de Energia Solar Fotovoltaica**. Disponível em: <http://www.cresesb.cepel.br/index.php?section=com_content&lang=pt&catid=4>. Acesso em: 9 nov. 2022.

DEVABHAKTUNI, V. et al. Solar energy: Trends and enabling technologies. **Renewable and Sustainable Energy Reviews**, v. 19, p. 555–564, mar. 2013.

DIEBOLD, F.; MARIANO, R. Comparing Predictive Accuracy. **Journal of Business & Economic Statistics**, v. 13, n. 3, p. 253–263, 1995.

DIMD, B. D. et al. A Review of Machine Learning-Based Photovoltaic Output Power Forecasting: Nordic Context. **IEEE Access**, v. 10, p. 26404–26425, 2022.

DONG, Z. et al. A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance. **Energy**, v. 82, p. 570–577, mar. 2015.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. CatBoost: gradient boosting with categorical features support. **ArXiv**, 24 out. 2018.

DRUCKER, H. **Improving Regressors using Boosting Techniques**. International Conference on Machine Learning. **Anais**...1997.

ENEL. **Parque solar São Gonçalo**. Disponível em: <https://www.enelgreenpower.com/pt/nossos-projetos/highlights/parque-solar-sao-goncalo>. Acesso em: 16 nov. 2022.

ERDEBILLI, B.; DEVRIM-İÇTENBAŞ, B. Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey. **Mathematics**, v. 10, n. 14, p. 2466, 15 jul. 2022.

FREUND, Y.; SCHAPIRE, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, v. 55, n. 1, p. 119–139, ago. 1997.

GABOITAOLELWE, J. et al. Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison. **IEEE Access**, v. 11, p. 40820–40845, 2023.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Data Preprocessing in Data Mining**. Cham: Springer International Publishing, 2015. v. 72

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems**. 2. ed. Sebastopol: O'Reilly Media, 2019.

GUERMOUI, M. et al. A comprehensive review of hybrid models for solar radiation forecasting. **Journal of Cleaner Production**, v. 258, p. 120357, jun. 2020.

GULIN, A. **CatBoost**. Disponível em: <https://catboost.ai/en/docs/>. Acesso em: 1 maio. 2023.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining. Concepts and Techniques**. 3. ed. Waltham: Elsevier, 2011.

HJERPE, A. **Computing Random Forests Variable Importance Measures (VIM) on Mixed Continuous and Categorical Data**. Stockholm: KTH Royal Institute of Techonology, 2016.

IEA. **Solar PV**. Paris: [s.n.]. Disponível em: <https://www.iea.org/reports/solar-pv>. Acesso em: 2 mar. 2023.

INMET. **Instituto Nacional de Meteorología**. Disponível em: <https://bdmep.inmet.gov.br/>. Acesso em: 15 abr. 2023.

IRENA. **International Renewable Energy AgencyStatistics Time Series**. [s.l: s.n.]. Disponível em: <https://www.irena.org/Data/View-data-by-topic/Capacity-and-Generation/Statistics-Time-Series>. Acesso em: 28 set. 2022.

JAYALAKSHMI, N. Y. et al. Novel Multi-Time Scale Deep Learning Algorithm for Solar Irradiance Forecasting. **Energies**, v. 14, n. 9, p. 2404, 23 abr. 2021.

KIRA, K.; RENDELL, L. A. A Practical Approach to Feature Selection. Em: **Machine Learning Proceedings 1992**. [s.l.] Elsevier, 1992. p. 249–256.

KOHAVI, R.; JOHN, G. H. Wrappers for feature subset selection. **Artificial Intelligence**, v. 97, n. 1–2, p. 273–324, dez. 1997.

KUMAR, R.; AGGARWAL, R. K.; SHARMA, J. D. Comparison of regression and artificial neural network models for estimation of global solar radiations. **Renewable and Sustainable Energy Reviews**, v. 52, p. 1294–1299, dez. 2015.

KUMAR, V. Feature Selection: A literature Review. **The Smart Computing Review**, v. 4, n. 3, 30 jun. 2014.

KUMARI, P.; TOSHNIWAL, D. Extreme gradient boosting and deep neural network based ensemble learning approach to forecast hourly solar irradiance. **Journal of Cleaner Production**, v. 279, p. 123285, jan. 2021.

LEE, J. et al. Reliable solar irradiance prediction using ensemble learning-based models: A comparative study. **Energy Conversion and Management**, v. 208, p. 112582, mar. 2020.

LONG, H.; ZHANG, Z.; SU, Y. Analysis of daily solar power prediction with data-driven approaches. **Applied Energy**, v. 126, p. 29–37, ago. 2014.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. 1967.

MAHMUD, K. et al. Machine Learning Based PV Power Generation Forecasting in Alice Springs. **IEEE Access**, v. 9, p. 46117–46128, 2021.

MARQUES LAMEIRINHAS, R. A.; TORRES, J. P. N.; DE MELO CUNHA, J. P. A Photovoltaic Technology Review: History, Fundamentals and Applications. **Energies**, v. 15, n. 5, p. 1823, 1 mar. 2022.

MARTÍN, L. et al. Prediction of global solar irradiance based on time series analysis: Application to solar thermal power plants energy production planning. **Solar Energy**, v. 84, n. 10, p. 1772–1781, out. 2010.

MASSAOUDI, M. et al. Enhanced Deep Belief Network Based on Ensemble Learning and Tree-Structured of Parzen Estimators: An Optimal Photovoltaic Power Forecasting Method. **IEEE Access**, v. 9, p. 150330–150344, 2021.

MERA-GAONA, M. et al. Framework for the Ensemble of Feature Selection Methods. **Applied Sciences**, v. 11, n. 17, p. 8122, 1 set. 2021.

MICHAEL, N. E. et al. Short-Term Solar Power Predicting Model Based on Multi-Step CNN Stacked LSTM Technique. **Energies**, v. 15, n. 6, p. 2150, 15 mar. 2022.

MOHAPATRA, N.; SHREYA, K.; CHINMAY, A. Optimization of the Random Forest Algorithm. Em: Singapore: Springer, 2020. p. 201–208.

NATRAS, R.; SOJA, B.; SCHMIDT, M. Ensemble Machine Learning of Random Forest, AdaBoost and XGBoost for Vertical Total Electron Content Forecasting. **Remote Sensing**, v. 14, n. 15, p. 3547, 24 jul. 2022.

OSBORNE, J. W. **Best Practices in Data Cleaning**. 1. ed. Los Angeles: SAGE, 2013.

PARK, J. et al. Multistep-Ahead Solar Radiation Forecasting Scheme Based on the Light Gradient Boosting Machine: A Case Study of Jeju Island. **Remote Sensing**, v. 12, n. 14, p. 2271, 15 jul. 2020.

PEDRO, H. T. C.; COIMBRA, C. F. M. Assessment of forecasting techniques for solar power production with no exogenous inputs. **Solar Energy**, v. 86, n. 7, p. 2017–2028, jul. 2012.

PELLAND, S.; GALANIS, G.; KALLOS, G. Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. **Progress in Photovoltaics: Research and Applications**, v. 21, n. 3, p. 284–296, maio 2013.

PHYO, P.-P.; BYUN, Y.-C.; PARK, N. Short-Term Energy Forecasting Using Machine-Learning-Based Ensemble Voting Regression. **Symmetry**, v. 14, n. 1, p. 160, 14 jan. 2022.

PINHO, J. et al. **Sistemas Híbridos**. 1. ed. Brasilia: Ministério de Minas e Energia, 2008.

PINHO, J.; GALDINO, M. A. **Manual de engenharia para sistemas fotovoltaicos**. Rio de Janeiro: CEPEL - CRESESB, 2014.

PRATAMA, I. et al. **A review of missing values handling methods on time-series data**. 2016 International Conference on Information Technology Systems and Innovation (ICITSI). **Anais**...IEEE, out. 2016.

QING, X.; NIU, Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM. **Energy**, v. 148, p. 461–468, abr. 2018.

RAHIMI, N. et al. A Comprehensive Review on Ensemble Solar Power Forecasting Algorithms. **Journal of Electrical Engineering & Technology**, v. 18, n. 2, p. 719–733, 12 mar. 2023.

RANA, M.; KOPRINSKA, I.; AGELIDIS, V. G. **Solar power forecasting using weather type clustering and ensembles of neural networks**. 2016 International Joint Conference on Neural Networks (IJCNN). **Anais**...IEEE, jul. 2016.

ROBNIK-SIKONJA, M.; KONONENKO, I. **An adaptation of Relief for attribute estimation in regression**. (D. H. Fisher, Ed.)Proceedings of the Fourteenth International Conference on Machine Learning. **Anais**...San Francisco: Morgan Kaufmann Publishers Inc., jul. 1997.

RODRÍGUEZ, F. et al. Forecasting intra-hour solar photovoltaic energy by assembling wavelet based time-frequency analysis with deep learning neural networks. **International Journal of Electrical Power & Energy Systems**, v. 137, p. 107777, maio 2022.

SAMPAIO, P. G. V.; GONZÁLEZ, M. O. A. Photovoltaic solar energy: Conceptual framework. **Renewable and Sustainable Energy Reviews**, v. 74, p. 590–601, jul. 2017.

SCHWERTMAN, N. C.; OWENS, M. A.; ADNAN, R. A simple more general boxplot method for identifying outliers. **Computational Statistics & Data Analysis**, v. 47, n. 1, p. 165–174, ago. 2004.

SHANNON, C. E. A Mathematical Theory of Communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, jul. 1948.

SHUKLA, R. K. et al. (EDS.). **Social Networking and Computational Intelligence**. Singapore: Springer Singapore, 2020. v. 100

SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. **Statistics and Computing**, v. 14, n. 3, p. 199–222, ago. 2004.

SOLARGIS. **Solar Resource Maps of Brazil**. Disponível em: <https://solargis.com/maps-and-gis-data/download/brazil>. Acesso em: 15 abr. 2023.

SURAKHI, O. et al. Time-Lag Selection for Time-Series Forecasting Using Neural Network and Heuristic Algorithm. **Electronics**, v. 10, n. 20, p. 2518, 15 out. 2021.

TAO, C. et al. Short-Term Forecasting of Photovoltaic Power Generation Based on Feature Selection and Bias Compensation–LSTM Network. **Energies**, v. 14, n. 11, p. 3086, 26 maio 2021.

THORPE, D. **Solar Energy Pocket Reference**. 2. ed. New York: Routledge, 2017.

TOLMASQUIM, M. T. **Energia Renovável: Hidráulica, Biomassa, Eólica, Solar, Oceânica**. Rio de Janeiro: Empresa de Pesquisa Energética, 2016.

VIAN, Â. et al. **Energia Solar Fundamentos Tecnologia e Aplicações**. 1. ed. São Paulo: Blucher, 2021.

VILLALVA, M. G.; GAZOLI, J. R. **Energia Solar Fotovoltaica – Conceitos e Aplicações**. 1. ed. São Paulo: Érica, 2012.

VOYANT, C. et al. Machine learning methods for solar radiation forecasting: A review. **Renewable Energy**, v. 105, p. 569–582, maio 2017.

WADE, C. **Hands-On Gradient Boosting with XGBoost and scikit-learn**. 1. ed. Birmingham: Packt, 2020.

WALD, L. **Fundamentals of Solar Radiation**. Boca Raton: CRC Press, 2021.

WEAVER, K. F. et al. **An Introduction to Statistical Analysis in Research**. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2017.

WENTZ, V. H. et al. Solar Irradiance Forecasting to Short-Term PV Power: Accuracy Comparison of ANN and LSTM Models. **Energies**, v. 15, n. 7, p. 2457, 27 mar. 2022.

YADAV, A. K.; CHANDEL, S. S. Solar radiation prediction using Artificial Neural Network techniques: A review. **Renewable and Sustainable Energy Reviews**, v. 33, p. 772–781, maio 2014.