COORDENAÇÃO DO CURSO SUPERIOR DE BACHARELADO EM ENGENHARIA ELÉTRICA

SÁVIO MURILLO DIAS BASTOS

PROJETO FINAL DE CURSO

APLICAÇÃO DE MODELOS PREDITIVOS DE APRENDIZADO DE MÁQUINA PARA GERAÇÃO DE ENERGIA SOLAR

João Pessoa

Dezembro de 2023

Sávio Murillo Dias Bastos

Aplicação de Modelos Preditivos de Aprendizado de Máquina para Geração de Energia Solar

Projeto Final de Curso submetido à Coordenação do Curso Superior de Bacharelado em Engenharia Elétrica do Instituto Federal da Paraíba como parte dos requisitos necessários para a obtenção do grau de Bacharel em Engenharia Elétrica.

Orientador(a):
Profa. Dra. Diana Moreno Nobre de Souza

João Pessoa
Dezembro de 2023

Dados Internacionais de Catalogação na Publicação – CIP
Biblioteca Nilo Peçanha – IFPB, *campus* João Pessoa

B327a Bastos, Sávio Murillo Dias.
Aplicação de modelos preditivos de aprendizado de máquina para geração de energia solar / Sávio Murillo Dias Bastos. – 2023.
47 f. : il.

TCC (Graduação em Engenharia Elétrica) – Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB / Coordenação de Engenharia Elétrica.
Orientadora : Profa. Dra. Diana Moreno Nobre de Souza.

1. Inteligência artificial – Aprendizado de máquina. 2. Modelos preditivos. 3. Energia solar. I. Título.

CDU 004.8:620.91

Bibliotecária responsável Josinete Nóbrega de Araújo – CRB15/116

Sávio Murillo Dias Bastos

Aplicação de Modelos Preditivos de Aprendizado de Máquina para Geração de Energia Solar

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Superior de Bacharelado em Engenharia Elétrica do Instituto Federal da Paraíba como parte dos requisitos necessários para a obtenção do grau de Bacharel em Engenharia Elétrica.

Trabalho Aprovado em 11 / 12 / 2023 pela banca examinadora:

gov.br Documento assinado digitalmente
DIANA MORENO NOBRE DE SOUZA
Data: 20/12/2023 21:36:52-0300
Verifique em https://validar.iti.gov.br

_____

Prof(a). Dra. Diana Moreno Nobre de Souza
Orientador(a), IFPB

gov.br Documento assinado digitalmente
FRANKLIN MARTINS PEREIRA PAMPLONA
Data: 21/12/2023 13:01:34-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Franklin Martins Pereira Pamplona
Examinador(a), IFPB

gov.br Documento assinado digitalmente
ADEMAR GONCALVES DA COSTA JUNIOR
Data: 21/12/2023 11:50:55-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Ademar Gonçalves da Costa Júnior
Examinador(a), IFPB

João Pessoa

Dezembro de 2023

À minha avó, Maria Eugênia por todo amor, carinho e amparo.
À minha mãe, Iaponira Dias, por todo esforço em prol da minha criação.

AGRADECIMENTOS

Quero expressar minha gratidão, primeiramente a Deus e depois a minha família, à minha companheira, Sayonara, por todo apoio e incentivo nos momentos bons e ruins. Por fim, agradeço ao IFPB por proporcionar oportunidades no decorrer do ensino técnico e da graduação que me ajudaram na busca constante pelo aprendizado.

RESUMO

Este estudo propõe a aplicação de algoritmos de aprendizado de máquina voltados para séries temporais, visando avaliar sua precisão e selecionar o modelo com a melhor acurácia. O conjunto de dados utilizado foi obtido de uma usina de geração solar na Índia, disponibilizado pelo Kaggle. A pesquisa proporciona uma oportunidade única de integrar conceitos da Engenharia Elétrica adquiridos durante a graduação com os princípios relacionados à Ciência de Dados e Inteligência Artificial, enfocando especificamente a análise de Séries Temporais em um contexto de geração solar.

Palavras-chave: aprendizado de máquina; séries temporais; energia fotovoltaica; ciência de dados.

ABSTRACT

This study proposes the application of machine learning algorithms focused on Time Series, aiming to assess their accuracy and select the model with the best accuracy. The dataset used was obtained from a solar power plant in India, made available by Kaggle. The research provides a unique opportunity to integrate concepts from Electrical Engineering acquired during undergraduate studies with principles related to Data Science and Artificial Intelligence, specifically focusing on the analysis of Time Series in the context of solar generation.

Keywords: machine learning; time series; photovoltaics; data science.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADF	Teste de Dickey-Fuller Aumentado
ARIMA	Médias Móveis Integradas AutoRegressivas
C	Ciclo
CA	Corrente Alternada
CC	Corrente Contínua
CNN	Rede Neural Convolucional
GW	Gigawatt (unidade de potência equivalente a um bilhão de watts)
I	Variação Irregular
IEA	Agência Internacional de Energia
IRENA	Agência Internacional de Energias Renováveis
MAE	Erro Absoluto Médio
RMSE	Erro Quadrático Médio
S	Sazonalidade
SVM	Máquina de Vetores de Suporte
T	Tendência
TCC	Trabalho de Conclusão de Curso
TW	Terawatt (unidade de potência equivalente a um trilhão de watts)

SUMÁRIO

1. INTRODUÇÃO

Segundo a Agência Internacional de Energia Renovável (IRENA), no final de 2022, o mundo testemunhou um aumento notável na capacidade global de geração de energia renovável, atingindo 3372 Gigawatts (GW). Esse aumento registrou um crescimento recorde de 295 GW, representando um aumento de 9,6% em comparação ao ano anterior. Além disso, 83% de toda a capacidade energética adicionada ao longo do ano foi proveniente de fontes de energia renovável.

Os dados mais recentes, apresentados nas Estatísticas de Capacidade Renovável de 2023 pela IRENA, confirmam a continuação desse crescimento surpreendente. Isso ocorre mesmo em meio a incertezas globais, reforçando a tendência de declínio na produção de energia a partir de combustíveis fósseis. A crescente demanda por fontes de energia sustentável e limpa coloca a energia solar no centro das atenções como uma solução ecologicamente correta para atender às necessidades energéticas globais (IRENA, 2022). Nesse contexto, o desenvolvimento de modelos preditivos de geração de energia solar se torna fundamental para otimizar o uso dessa fonte de energia renovável.

Este Projeto Final de Curso (PFC) concentra-se na aplicação e análise de modelos preditivos de geração de energia solar utilizando algoritmos de aprendizado de máquina com intuito de selecionar o modelo com melhor assertividade para as peculiaridades de uma usina estudada de geração de energia solar localizada na Índia, composta por 22 inversores.

1.1. DESCRIÇÃO DO PROBLEMA

É notável o crescimento projetado das energias renováveis, conforme indicado pelo relatório da Agência Internacional de Energia (IEA) de 2021. A previsão de que as energias renováveis representarão quase 95% do aumento da capacidade de energia global até 2026 é um passo significativo na transição para uma matriz energética mais sustentável.

O aumento previsto de 60% no uso de fontes renováveis nos próximos cinco anos em comparação com os números de 2020 destaca a rápida adoção e integração dessas tecnologias no setor energético global. Esse crescimento pode ser atribuído, em parte, ao reconhecimento dos benefícios ambientais e econômicos das energias renováveis.

A aspiração de muitos países em zerar as emissões líquidas de carbono até 2050, como afirmam Sales e Uhlig (2021), demonstra um compromisso coletivo em combater as mudanças climáticas. Essa meta ambiciosa exige a implementação de políticas e práticas

sustentáveis, bem como a contínua inovação tecnológica para alcançar um equilíbrio entre o desenvolvimento econômico e a preservação do meio ambiente.

O setor de energia desempenha um papel crucial nessa transição, e a absorção rápida das tecnologias emergentes é fundamental para atingir esses objetivos. A colaboração global e o investimento em pesquisa e desenvolvimento são essenciais para impulsionar ainda mais a eficiência e a acessibilidade das energias renováveis.

Essas tecnologias se tornam cada vez mais presentes na geração e consumo de energia elétrica, visando reduzir impactos causados pelo alto consumo e otimizando o processo de geração de energia.

1.2. OBJETIVOS

1.2.1. Objetivo Geral

Utilizando dois conjuntos de dados disponibilizado pela plataforma *Kaggle,* coletados ao longo de um período de 34 dias de uma usina de energia solar na Índia, composta por 22 inversores que oferecem informações tanto sobre a planta da usina quanto realiza uma monitorização sobre as condições climáticas da região, o propósito central deste trabalho é analisar a aplicabilidade dos algoritmos de aprendizado de máquina com intuito de selecionar o modelo com melhor acurácia na previsão de geração de energia solar na usina em questão.

1.2.2. Objetivos Específicos

Para alcançar o propósito global, é essencial adotar um procedimento que abranja a análise criteriosa de dados, o tratamento adequado, a seleção do algoritmo, a avaliação do modelo empregado e a definição precisa do mesmo. Portanto, torna-se necessário desdobrar esses elementos em objetivos específicos, a fim de estabelecer um processo bem definido.

Os objetivos específicos delineados para este estudo compreendem:

- Análise exploratória e tratamento de dados: realizar tratamento e classificação eficazes nos dados coletados da usina de energia solar na Índia. Este passo é fundamental para garantir a qualidade e a consistência necessárias durante as fases subsequentes da análise;

- Aplicação de Algoritmos de *machine learning*: definir e justificar a escolha dos algoritmos de *machine learning* que serão empregados no desenvolvimento do modelo

preditivo. Esta etapa visa garantir uma abordagem alinhada com os objetivos do estudo e a natureza dos dados disponíveis;

- Treinamento e avaliação de modelos: implementar o treinamento dos modelos preditivos utilizando os conjuntos de dados preparados, seguido por uma avaliação rigorosa do desempenho. Essa análise crítica permitirá a identificação de pontos fortes e áreas de aprimoramento em cada modelo;

- Definição do modelo com melhor desempenho: selecionar o modelo com o melhor desempenho, considerando métricas específicas de avaliação. Esta etapa é crucial para escolher a solução mais eficiente e precisa para a previsão da geração de energia solar na região estudada.

1.3 ESTRUTURA DO TRABALHO

O presente trabalho está organizado em cinco seções, começando pela introdução (seção atual) e seguido por fundamentação teórica, metodologia, resultados e conclusão.

Na seção 2 de fundamentação teórica, serão explorados os conceitos essenciais para a compreensão do trabalho. Essa seção incluirá subseções que elucidam os temas de energia solar, aprendizado de máquina e suas aplicações, além das métricas utilizadas para avaliar o desempenho dos algoritmos.

A seção 3 detalha a metodologia adotada, delineando cada processo de maneira individual. Serão apresentadas a base de estudos utilizada e as etapas de análise exploratória, tratamento de dados, aplicação e validação dos algoritmos.

A seção 4 discute os resultados obtidos, destacando qual algoritmo demonstrou melhor desempenho com base nas métricas de avaliação.

Por fim, a seção 5 traz a conclusão do trabalho, apresentando os desdobramentos dos objetivos propostos e delineando as expectativas futuras relacionadas ao projeto desenvolvido.

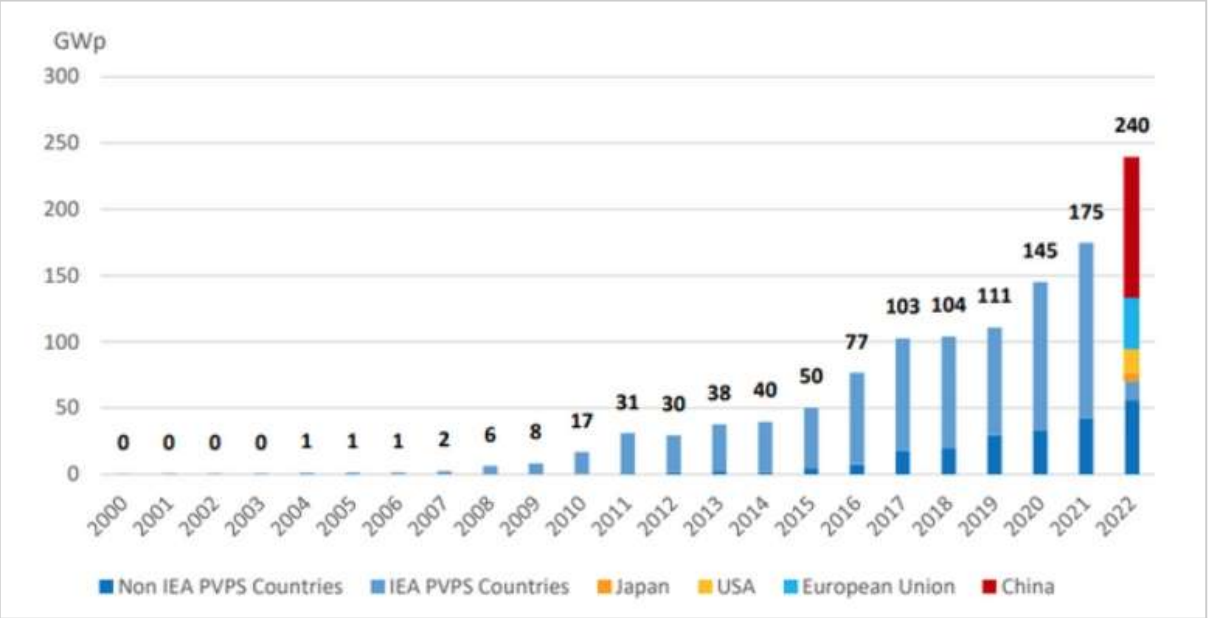
2. FUNDAMENTAÇÃO TEÓRICA

Primeiramente, é fundamental adquirir compreensão acerca dos temas que orientam esta pesquisa. No decorrer deste capítulo, exploraremos o panorama de expansão da energia solar e seus conceitos básicos, discutiremos os fundamentos do aprendizado de máquina, abordaremos sua aplicação e avaliação, e aprofundaremos o entendimento sobre o conceito de séries temporais.

2.1. ENERGIA SOLAR

Em 2022, a capacidade global de energia solar registrou um notável crescimento de 240 GW, alcançando uma marca acumulada impressionante de 1,2 TW, conforme revelado por uma análise da Agência Internacional de Energia (IEA). Os resultados destacam que 23 nações implementaram sistemas com pelo menos 1 gigawatt, enquanto 16 dessas atingiram uma capacidade acumulada superior a 10 GW até o final do ano, conforme pode ser observado na Figura 1.

Figura 1 — Evolução da capacidade global.



Fonte: Portal Solar, 2023.

Esse notável aumento na capacidade global de energia solar em 2022 não apenas reflete uma tendência de progresso significativo, mas também sublinha a crescente importância da transição para fontes renováveis (PORTAL SOLAR, 2023).

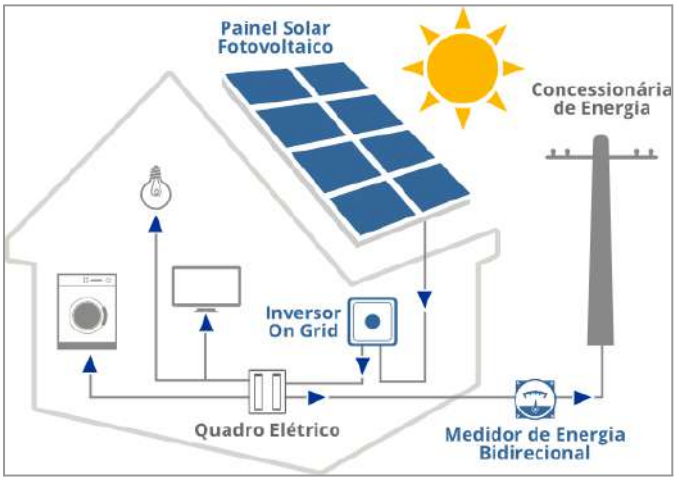
A energia solar é uma fonte alternativa, renovável e sustentável de energia que deriva da radiação eletromagnética (luz e calor) emitida pelo sol. Diversas tecnologias aproveitam

essa energia, incluindo aquecedores solares, painéis fotovoltaicos e usinas heliotérmicas (DUFFIE, J. A.; BECKMAN, W. A.; MCGOWAN, J. G, 1985).

A conversão da energia solar em eletricidade é realizada pela tecnologia fotovoltaica. Os painéis solares, compostos por células fotovoltaicas geralmente feitas de silício, captam a luz solar e a convertem em eletricidade por meio do efeito fotovoltaico. Esse efeito ocorre quando a radiação solar excita os elétrons nos átomos do material semicondutor, gerando corrente elétrica (DUFFIE, J. A.; BECKMAN, W. A.; MCGOWAN, J. G, 1985).

Um sistema solar fotovoltaico básico, apresentado na Figura 2, inclui painéis solares, inversor, controlador de carga e/ou baterias (quando usado em sistemas isolados da rede elétrica). Os painéis geram corrente contínua (CC), convertida em corrente alternada (CA) pelo inversor, tornando-a compatível com dispositivos elétricos convencionais. O controlador de carga regula a carga e descarga das baterias em sistemas isolados. A energia gerada alimenta a casa, e o excedente pode ser devolvido à concessionária de energia, gerando créditos (DUFFIE, J. A.; BECKMAN, W. A.; MCGOWAN, J. G, 1985).

Figura 2 — Usina fotovoltaica residencial.



Fonte: FÓTON ENGENHARIA, 2023.

Existem dois tipos principais de sistemas: residenciais e usinas. As usinas residenciais (Figura 2) são projetadas para atender às necessidades de uma casa, compreendendo painéis solares e inversores. Elas podem suprir toda a demanda residencial, incluindo iluminação, ar-condicionado e eletrodomésticos (SIDERURGIA BRASIL, 2022).

As usinas de grande porte (Figura 3) são projetadas para atender a grandes consumidores, como empresas e indústrias. Com milhares de painéis solares instalados em áreas extensas com exposição solar, essas usinas podem gerar quantidades significativas de

eletricidade, sendo uma opção viável para atender às necessidades de grandes consumidores (SIDERURGIA BRASIL, 2022).

Figura 3 — Usina fotovoltaica de grande porte localizada no Piauí.

Fonte: PORTAL SIDERURGIA, 2022.

2.2. APRENDIZADO DE MÁQUINA

O aprendizado de máquina pode ser definido como um ramo da inteligência artificial que permite que os computadores aprendam com os dados fornecidos, sem a necessidade de serem explicitamente programados para realizar uma tarefa específica. Este campo é comumente dividido em dois paradigmas principais: aprendizado supervisionado e aprendizado não supervisionado. No aprendizado supervisionado, os modelos são treinados com conjuntos de dados que incluem exemplos de entrada e saída correspondentes. Isso permite que o computador associe entradas específicas a saídas desejadas, capacitando-o a fazer previsões precisas para novos conjuntos de dados (SING, 2019).

Em contrapartida, o aprendizado não supervisionado implica no treinamento de modelos utilizando conjuntos de dados desprovidos de saídas conhecidas. Nesse contexto, a ênfase reside na identificação de padrões intrínsecos nos dados, viabilizando a agregação de informações similares ou a previsão de dados ainda não observados. Essas abordagens não apenas se distinguem em termos de metodologia, mas também encontram aplicações específicas em diversas áreas. Para além desses dois paradigmas, outras facetas do aprendizado de máquina merecem destaque, como o aprendizado por reforço, que envolve a capacidade do modelo em realizar ações em um ambiente, recebendo recompensas ou penalidades com base no desempenho dessas ações (SING, 2019).

Desafios inerentes ao aprendizado de máquina incluem questões de *overfitting* e *underfitting*, onde os modelos podem se ajustar demais ou insuficientemente, respectivamente, aos dados de treinamento, e a seleção apropriada de recursos, fundamental para o desempenho ideal. A interpretabilidade dos modelos também se destaca como uma preocupação crescente, especialmente em contextos nos quais compreender as decisões tomadas pelos algoritmos é crucial (SING, 2019).

Na prática, o aprendizado de máquina é amplamente empregado em diversas aplicações. Desde o reconhecimento de imagens, onde o aprendizado supervisionado é utilizado para classificar objetos, até o processamento de linguagem natural, onde técnicas não supervisionadas podem ser utilizadas para detectar anomalias, associar compras realizadas, agrupar tipos de clientes e diversas outras atribuições. À medida que os modelos se tornam mais avançados, é imperativo abordar considerações éticas e de segurança, destacando a importância de um desenvolvimento cuidadoso e responsável dessas tecnologias inovadoras (RODRIGUEZ-GALIANO;SÁNCHEZ-CASTILLO;CHICA-OLMO;CHICA-RIVAS, 2015).

2.2.1. APLICAÇÕES DE ALGORITMOS SUPERVISIONADOS

Os modelos supervisionados desempenham um papel crucial na era da inteligência artificial, oferecendo soluções poderosas para uma variedade de problemas complexos. Ao permitir que algoritmos aprendam a partir de dados rotulados, esses modelos têm a capacidade de generalizar padrões e realizar tarefas específicas com notável precisão. Nas próximas seções, serão apresentados alguns cenários em que a aplicabilidade desses modelos se faz necessária (SING, 2019).

2.2.1.1. Classificação de e-mails

Utilizado para tratar o alto volume de mensagens eletrônicas, modelos supervisionados como Máquinas de Vetores de Suporte (SVM), Redes Neurais e Naive Bayes desempenham um papel vital. Ao aprender padrões a partir de exemplos rotulados, esses algoritmos capacitam sistemas a identificar com precisão e eficiência e-mails como "*spam*" ou "não *spam*", contribuindo para uma experiência mais segura e livre de inconvenientes para os usuários (GÉRON, 2019).

2.2.1.2. Diagnóstico médico

Em um contexto de saúde, a aplicação de modelos supervisionados como árvores de decisão e redes neurais oferece uma abordagem promissora para a classificação de pacientes. A análise de características médicas e histórico clínico permite a identificação eficaz de portadores ou não de uma doença, fornecendo aos profissionais de saúde ferramentas valiosas para tomada de decisões precisas e personalizadas (GÉRON, 2019).

2.2.1.3. Reconhecimento de imagens

No universo visual, modelos supervisionados como Redes Neurais Convolucionais (CNN) e SVM desempenham um papel central no reconhecimento de padrões em imagens. Esses algoritmos capacitam sistemas a identificar objetos, rostos e padrões, sendo essenciais para aplicações como reconhecimento facial e categorização de objetos, tanto em ambientes de segurança quanto em soluções de organização de dados visuais (GÉRON, 2019).

2.2.1.4. Previsões no mercado financeiro

A complexidade do mercado financeiro encontra resposta na aplicação de modelos supervisionados, como séries temporais e regressão linear. Ao analisar padrões históricos e indicadores financeiros, esses algoritmos oferecem ferramentas valiosas para prever o valor futuro de ações (GÉRON, 2019).

Em síntese, a aplicação de modelos supervisionados destaca-se como uma abordagem eficaz em diversos domínios, capacitando sistemas a aprender e realizar tarefas complexas com base em exemplos rotulados.

2.2.2. APLICAÇÕES DE ALGORITMOS NÃO SUPERVISIONADOS

Enquanto os modelos supervisionados destacam-se na aprendizagem a partir de dados rotulados, os modelos não supervisionados abordam desafios complexos de maneira diferente, explorando padrões e estruturas intrínsecas nos dados. Nesta seção, examinaremos a aplicação desses modelos em diversas áreas, como agrupamento de dados, geração de recomendações e detecção de anomalias (SING, 2019).

2.2.2.1. Agrupamento de dados

Os modelos não supervisionados, como o *k-Means* e o *Hierarchical Clustering*, desempenham um papel importante no agrupamento de dados, sendo abordagens comumente utilizadas em diversos nichos de mercado como publicidade e recursos humanos. Esses algoritmos identificam padrões não rotulados, agrupando observações similares e revelando estruturas subjacentes nos conjuntos de dados. Essa abordagem é essencial em áreas como segmentação de clientes, onde a identificação de grupos sem informações prévias pode guiar estratégias de marketing mais eficazes (RASCHKA;PATTERSON;NOLET, 2020).

2.2.2.2. Geração de recomendações

Nos domínios de comércio eletrônico e entretenimento, modelos não supervisionados, incluindo sistemas de recomendação baseados em fatoração de matrizes, têm um impacto significativo. Ao analisar padrões de comportamento de usuários sem depender de rótulos explícitos, esses algoritmos geram recomendações personalizadas, melhorando a experiência do usuário e impulsionando a relevância de produtos e conteúdos (RASCHKA;PATTERSON;NOLET, 2020).

2.2.2.3. Detecção de anomalias

Na segurança e na manutenção de sistemas, modelos não supervisionados como *Isolation Forests* e *One-Class* SVM são cruciais para a detecção de anomalias. Ao aprender o comportamento normal dos dados, esses algoritmos destacam padrões anômalos, alertando para atividades suspeitas ou falhas, contribuindo para a segurança e estabilidade de sistemas complexos (GÉRON, 2019).

Os modelos não supervisionados desempenham um papel crucial na revelação de padrões ocultos e estruturas intrínsecas nos dados, sem depender de rótulos explícitos. Essa abordagem, fundamental para desafios onde a natureza dos dados é complexa ou desconhecida, continua a impulsionar avanços significativos em diversas áreas, desde a organização eficiente de informações até a melhoria da segurança e da experiência do usuário (GÉRON, 2019).

2.3. MODELOS PREDITIVOS

Os modelos preditivos representam uma peça fundamental no arsenal da inteligência artificial, permitindo a antecipação de eventos e comportamentos futuros com base em padrões e dados históricos. Neste contexto, exploraremos diferentes tipos de modelos preditivos, destacando sua aplicação em áreas como finanças, saúde, energia e análise de comportamento do usuário online (RODRIGUEZ-GALIANO;SÁNCHEZ-CASTILLO;CHICA-OLMO;CHICA-RIVAS, 2015).

2.3.1. Regressão

Os modelos de regressão entram em cena quando a variável de saída desejada é contínua. Em finanças, por exemplo, esses modelos podem prever o preço futuro de ações com base em indicadores históricos, oferecendo aos investidores uma ferramenta valiosa para tomada de decisões informadas diante da volatilidade do mercado (GÉRON, 2019).

2.3.2. Classificação

Para variáveis de saída categóricas, os modelos de classificação são essenciais. Na área da saúde, esses modelos podem prever se um paciente tem ou não uma determinada condição médica, utilizando resultados de exames e histórico clínico. Essa abordagem facilita diagnósticos precoces e otimiza o processo de cuidados médicos (GÉRON, 2019).

2.3.3. Séries temporais

O conceito de séries temporais é definido como um conjunto de valores de um fenômeno específico ao longo do tempo, onde essa variação temporal (Y) é dividida em intervalos uniformes (NIELSEN, 2021). Esse tipo de série apresenta componentes que caracterizam seu comportamento, sendo expresso matematicamente pela Equação (1):

$$Y=T+C+S+I \qquad (1)$$

Os parâmetros são definidos da seguinte forma:

- Tendência (T): observa a direção geral de crescimento ou decrescimento de uma variável ao longo do tempo.
- Ciclo (C): analisa padrões que se repetem em curtos intervalos de tempo, indicando ciclos recorrentes.

- Sazonalidade (S): descreve variações específicas em momentos particulares do tempo, como sazonalidades sazonais.

- Variação Irregular (I): refere-se a flutuações imprevisíveis causadas por fatores externos fora de controle.

Quando a temporalidade é crucial, os modelos de séries temporais assumem protagonismo. No setor de energia, por exemplo, esses modelos podem prever a demanda de eletricidade para os próximos meses, incorporando sazonalidades e padrões históricos, possibilitando um planejamento eficiente da produção.

Dentro dos algoritmos mais utilizados nos dia de hoje, dois se destacam, sendo eles o modelo estatístico ARIMA (*Autoregressive Integrated Moving Average*) e o *framework* de aprendizado de máquina desenvolvido pelo *Facebook*, o *Prophet.*

Cada tipo de modelo preditivo aborda contextos específicos e oferece soluções distintas para problemas relacionados à previsão, permitindo uma aplicação versátil em diferentes domínios.

2.3.3.1. ALGORITMO ARIMA

A análise e previsão de séries temporais são elementos fundamentais em diversas disciplinas, fornecendo informações essenciais para a tomada de decisões embasadas. Nesse cenário, o ARIMA se destaca como uma ferramenta estatística essencial, amplamente adotada devido à sua capacidade de modelar padrões temporais complexos. A partir da figura 4, entende-se como estão dispostos os parâmetros no modelo estatístico.

Figura 4 — ARIMA.

Fonte: Autoria própria, 2023.

O ARIMA desmembra uma série temporal em três componentes principais, cada uma contribuindo para uma compreensão abrangente da dinâmica temporal dos dados. A Auto-regressão (AR) captura a dependência linear entre uma observação atual e seus valores passados, representada pela ordem de auto-regressão "p". A Diferenciação (I) é crucial para tornar a série estacionária, removendo tendências e sazonalidades, com a ordem de diferenciação "d"indicando a quantidade de diferenciações necessárias. A Média Móvel (MA) modela a dependência entre uma observação e os erros residuais de um modelo de média móvel aplicado a observações passadas, com a ordem da média móvel "q" definindo o número de erros residuais considerados (NIELSEN, 2021).

A determinação dos parâmetros p, d e q é uma etapa crítica na construção de um modelo ARIMA eficaz, garantindo uma representação precisa da complexidade inerente à série temporal. Para simplificar esse processo, a biblioteca “pmdarima”[1] em *Python* apresenta a função “auto_arima”, que automatiza a escolha de parâmetros por meio de uma abordagem passo a passo. Essa automação reduz a complexidade associada à seleção manual de parâmetros, tornando o ARIMA uma ferramenta acessível e eficiente, especialmente em cenários que envolvem grandes conjuntos de dados temporais (NIELSEN, 2021).

Em síntese, o ARIMA oferece uma base teórica robusta para a modelagem de séries temporais, proporcionando uma abordagem sistemática para compreender e prever padrões

[1] pmdarima: Hyndman RJ, & Athanasopoulos G. (2018). pmdarima: An open source Python package for ARIMA modeling. Journal of Open Source Software, 3(32), 1026. https://doi.org

em dados temporais. A inclusão da função “auto_arima” adiciona uma camada de automação, aprimorando a aplicabilidade prática desse modelo estatístico.

2.3.3.2. ALGORITMO *FACEBOOK PROPHET*

O *Facebook* Prophet é uma *framework* de previsão de séries temporais desenvolvida para simplificar a complexa tarefa de modelagem e previsão, tornando-a acessível até mesmo para usuários com pouca experiência nesse domínio. Criado pela equipe de pesquisa do *Facebook*, o *Prophet* opera com uma abordagem flexível e eficaz para lidar com uma variedade de padrões temporais (BOSCOA, 2022).

O funcionamento do *Prophet* é caracterizado por diversos componentes-chave. Primeiramente, o modelo identifica e modela automaticamente a tendência temporal dos dados, ajustando-se a mudanças significativas ao longo do tempo, permitindo uma representação mais precisa de padrões de crescimento ou declínio não lineares. Além disso, o *Prophet* lida eficientemente com padrões sazonais, como variações diárias, mensais ou anuais, captando automaticamente as sazonalidades e facilitando previsões em séries temporais com ciclos regulares (BOSCOA, 2022).

Outra característica distintiva é a capacidade do *Prophet* de incorporar informações sobre feriados. Essa funcionalidade permite ao modelo considerar eventos específicos que podem impactar as séries temporais, resultando em previsões mais precisas durante períodos de feriados ou eventos especiais. Além disso, o *Prophet* é robusto em relação a dados ausentes *(outliers)*, conseguindo lidar eficientemente com lacunas, uma qualidade valiosa em situações do mundo real, onde dados incompletos são comuns (DATARISK, 2023).

O *Facebook* Prophet se destaca como uma solução robusta para previsões de séries temporais. Sua notável facilidade de uso, exigindo poucos ajustes de parâmetros, aliada à rapidez nas previsões, o torna ideal para análises em tempo real. Além disso, sua flexibilidade permite uma ampla gama de aplicações, desde previsões financeiras até análises científicas. Em resumo, o Prophet é uma ferramenta valiosa e adaptável, consolidando-se como uma escolha confiável em diversas situações práticas (DATARISK, 2023).

2.4. MÉTRICAS DE AVALIAÇÃO

A avaliação de modelos de aprendizado de máquina é uma etapa crucial no desenvolvimento de soluções eficazes e precisas. As métricas desempenham um papel fundamental nesse processo, oferecendo uma maneira quantitativa de medir o desempenho do modelo em relação aos dados observados. Dentre as métricas comumente utilizadas, o MAE

(Erro Absoluto Médio), R² (coeficiente de determinação) e RMSE (Erro Quadrático Médio) desempenham papéis específicos, cada um trazendo pontos valiosos sobre diferentes aspectos do desempenho do modelo (GÉRON, 2019).

2.4.1. MAE

O MAE é uma métrica comumente utilizada para avaliar o desempenho de modelos de previsão ou regressão. Ele quantifica a média absoluta dos erros entre as previsões do modelo e os valores reais. O MAE é calculado da seguinte maneira, conforme exposto pela Equação 2 (CARMO;SILVA, 2023):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |yi - \hat{Y}i| \qquad (2)$$

Onde:

- n é o número total de observações.
- yi é o valor real da observação i.
- Ŷi é a previsão do modelo para a observação de i.

2.4.2. R² SCORE

O R² é uma métrica estatística frequentemente utilizada para avaliar o desempenho de modelos de regressão. Essa métrica fornece uma medida da proporção da variabilidade da variável dependente que é explicada pelas variáveis independentes do modelo. O R² varia de 0 a 1, onde:

- R²=0: O modelo não explica nenhuma variabilidade na variável dependente.
- R²=1: O modelo explica completamente a variabilidade na variável dependente.

O cálculo do R² é feito pela fórmula exibida abaixo pela Equação 3:

$$R2 = 1 - \frac{\left[\sum_{i=1}^{n} (yi - \hat{Y}i)^2\right]}{\sum_{i=1}^{n} (yi - \hat{Y})^2} \qquad (3)$$

- n é o número total de observações.
- yi é o valor real da observação i.
- Ŷi é a previsão do modelo para a observação de i.

- Ŷ é a média dos valores reais de y.

Em termos simples, o R² avalia a adequação do modelo em relação à variabilidade dos dados reais. Um R² mais próximo de 1 indica que o modelo ajusta-se bem aos dados, enquanto um R² próximo de 0 sugere que o modelo não é eficaz na explicação da variabilidade observada. No entanto, o R² também tem limitações, especialmente em modelos complexos ou com sobreajuste *(overfiting)*, e deve ser interpretado juntamente com outras métricas de desempenho (GÉRON, 2019).

2.4.3. RMSE

O RMSE é uma métrica amplamente utilizada para avaliar a precisão de modelos de previsão ou regressão. Ele mede a média da raiz quadrada dos erros quadrados entre as previsões do modelo e os valores reais. O RMSE é calculado pela fórmula abaixo, exposta na Equação 4:

$$RMSE = \sqrt{\frac{1\sum_{i=1}^{2}(Yi-\hat{Y}i)^2}{n}} \qquad (4)$$

Onde:

- n é o número total de observações.
- Yi é o valor real da observação i.
- Ŷi é a previsão do modelo para a observação de i.

O RMSE penaliza erros maiores de maneira mais significativa do que erros menores devido ao processo de elevação do quadrado. A raiz quadrada é então aplicada para retornar à métrica à mesma unidade da variável original, facilitando a interpretação.

Em termos simples, o RMSE mais baixo indica melhor desempenho do modelo, refletindo previsões mais próximas dos valores reais. Essa métrica, embora útil, requer interpretação em conjunto com outras informações e considerações específicas do contexto (GÉRON, 2019).

2.4.4. TESTE DICKEY-FULLER

O Teste de Dickey-Fuller é uma ferramenta estatística utilizada para avaliar a estacionariedade de uma série temporal. Ele é projetado para testar a presença de uma raiz unitária em uma série, o que é crucial para determinar se a série é estacionária ou não.

A hipótese nula do teste é que a série temporal possui uma raiz unitária, o que implica não estacionariedade. Se o teste rejeitar a hipótese nula com um nível de significância estabelecido, isso sugere que a série é estacionária, indicando a ausência de tendências ou padrões sistemáticos.

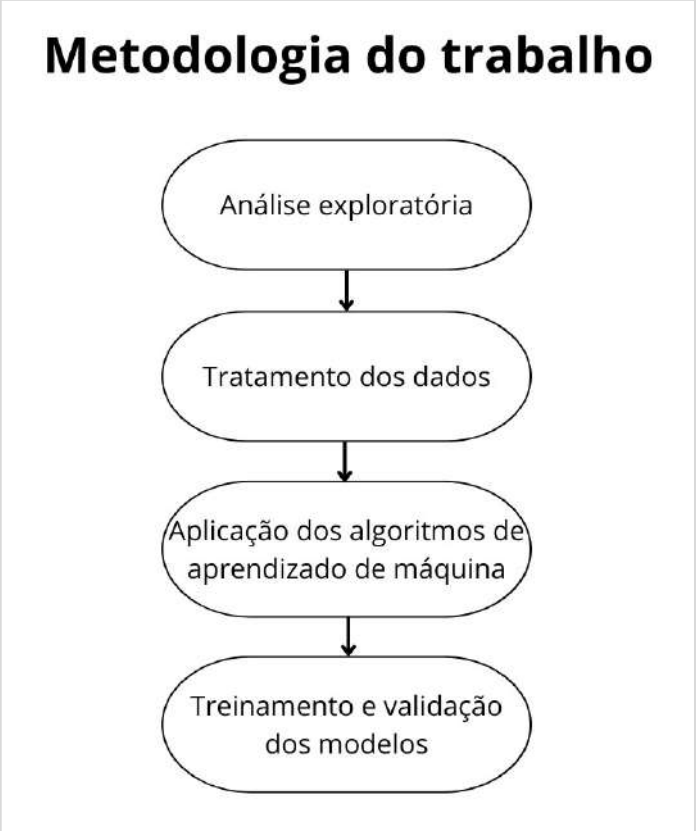
O resultado do teste é geralmente interpretado usando um valor-p (*p-value*). Se o valor-p for menor que um nível de significância crítico (como 0,05), a hipótese nula é rejeitada, indicando estacionariedade. Se o valor-p for maior que o nível de significância, não há evidência suficiente para rejeitar a hipótese nula, sugerindo não estacionariedade.

Em resumo, o Teste de Dickey-Fuller é uma ferramenta valiosa para verificar se uma série temporal é estacionária, uma propriedade fundamental em muitos métodos de análise de séries temporais (NIELSEN, 2021).

3. METODOLOGIA

O projeto em análise segue o fluxograma apresentado na Figura 5, detalhando os processos envolvidos e buscando alcançar êxito na seleção do modelo de previsão.

Figura 5 — Metodologia do trabalho.



Fonte: Autoria própria, 2023.

Este estudo adota uma abordagem metodológica abrangente, iniciando com uma análise exploratória dos dados para compreender sua estrutura e características. Essa análise é essencial para compreender as características do conjunto de dados e, assim, escolher os algoritmos mais adequados.

Em seguida, procede-se ao tratamento dos dados, que inclui etapas de limpeza e normalização, preparando-os para a aplicação de algoritmos de *machine learning*. A escolha criteriosa dos algoritmos é baseada nos *insights* adquiridos na análise exploratória, garantindo que eles sejam adequados ao conjunto de dados específico.

O treinamento e a avaliação do modelo são realizados de forma iterativa, visando otimizar continuamente o desempenho. A combinação da análise exploratória com a aplicação de técnicas de *machine learning* tem como objetivo proporcionar resultados confiáveis e *insights* valiosos para este estudo.

3.1. BASE DE ESTUDO

O presente estudo faz uso de uma base de dados pública hospedada no *Kaggle*[2], intitulada "*Solar Power Generation Data*". Esta base de dados engloba informações referentes à geração de energia e às condições climáticas de duas usinas fotovoltaicas localizadas na Índia, durante um período de 34 dias nos meses de maio a junho. É crucial destacar que a análise se restringe exclusivamente a uma das usinas.

A escolha desse banco de dados específico decorre do conjunto de informações proporcionado pelo monitoramento climático da região, resultando em uma maior eficiência no desenvolvimento do trabalho e otimização do tempo no processo de extração de dados. Considera-se esse estudo como o ponto de partida para futuras aplicações acadêmicas.

A unidade de geração de energia é representada por duas tabelas, uma abrangendo os componentes do sistema de geração e outra abordando as condições climáticas locais a cada intervalo de 15 minutos ao longo do dia. A Tabela 1 fornece uma visão geral do escopo das informações contidas no banco de dados, relacionadas à produção de eletricidade e seus fatores associados.

Tabela 1 — Base de estudo de geração da usina 1

DATA_HORA	POTENCIA_CC	POTENCIA_AC	PRODUCAO_DIARIA	PRODUCAO_TOTAL	INVERSOR
31/05/2020 07:00	1215.12	118.15	73.500	7.328	INVERSOR6
28/05/2020 15:30	8386.42	820.68	7.706	7.264	INVERSOR5
28/05/2020 13:30	7660.85	749.04	6.031	6.287	INVERSOR2

Fonte: Dados da pesquisa, 2023.

Onde os principais parâmetros da tabela 1 incluem:

- **DATA_HORA**: indica a data e a hora em que a informação foi registrada.
- **INVERSOR**: identificador exclusivo para cada inversor.
- **POTENCIA_CC**: representa a geração de corrente contínua.
- **POTENCIA_CA**: reflete a geração de corrente alternada.
- **PRODUCAO_DIARIA**: indica a quantidade de energia gerada diariamente, permitindo uma análise do desempenho ao longo do tempo.
- **PRODUCAO_TOTAL**: representa a geração total acumulada.

[2] Anikannal. (2023). Solar Power Generation Data. Kaggle. Retrieved from https://www.kaggle.com/datasets/anikannal/solar-power-generation-data

A Tabela 2 apresenta os dados referentes ao clima na região da usina, conforme podemos observar abaixo:

Tabela 2 — Base de estudo do clima da usina 1

DATA_HORA	TEMP_AMBIENTE	TEMP_MODULO	IRRADIACAO
31/05/2020 07:00	22.76	23.13	0.082451
28/05/2020 15:30	33.45	49.20	0.572510
28/05/2020 13:30	32.80	49.77	0.513674

Fonte: Dados da pesquisa, 2023.

Onde os principais parâmetros da tabela 2 incluem:

- **TEMP_AMBIENTE:** temperatura no âmbito da usina;
- **TEMP_MODULO:** temperatura referente aos módulos fotovoltaicos;
- **IRRADIACAO:** irradiação sobre a usina.

Após realizar um prévio entendimento acerca das variáveis dispostas nos bancos de dados, foi realizada uma união dos conjuntos com intuito de mesclar informações e obter melhores avaliações sobre a usina.

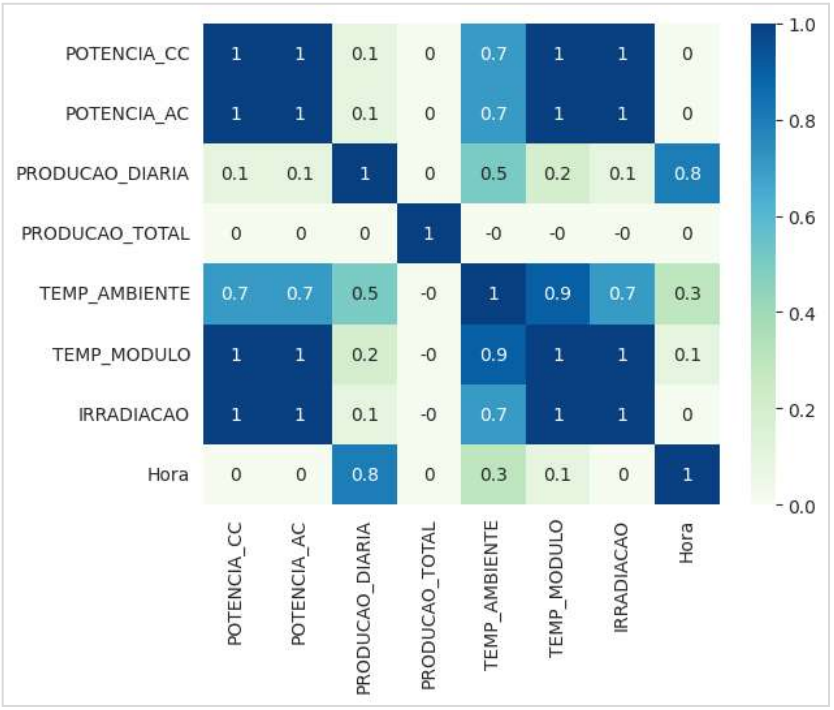
3.2. ANÁLISE EXPLORATÓRIA

A análise em questão concentrou-se em compreender as interações entre as variáveis envolvidas, assim como o desempenho da usina em torno de uma variável alvo relacionada a potência de corrente alternada (CA). O objetivo é extrair informações mais aprofundadas com intuito de orientar o desenvolvimento contínuo do trabalho e sinalizar possíveis modificações, caso necessárias.

O ponto inicial de investigação, conforme ilustrado na Figura 6, consistiu na análise das correlações entre as variáveis em questão, realizado por meio da função .corr() da biblioteca pandas[3] e com o objetivo de discernir as informações pertinentes das que são consideradas irrelevantes.

[3] pandas. (2023). Pandas Documentation. Retrieved from https://pandas.pydata.org/docs/

Figura 6 — Correlação de variáveis.



Fonte: Dados da pesquisa, 2023.

Ao analisar os dados, é evidente a presença de correlações substanciais entre diversas variáveis, alinhando-se com as expectativas prévias. Notavelmente, as potências de corrente contínua (CC) e corrente alternada (CA) apresentam uma correlação extremamente forte, indicando uma relação direta e proporcional entre elas. Esta constatação destaca-se como um ponto crucial para a compreensão do desempenho dos inversores no sistema.

A temperatura do módulo fotovoltaico surge como um fator de grande influência, manifestando uma correlação significativa com tanto a irradiação solar quanto a temperatura ambiente. A inter-relação aponta para a importância de considerar o impacto ambiental nas condições de operação dos módulos, uma vez que variações na temperatura podem influenciar diretamente a eficiência do sistema.

Além disso, a análise revela uma forte correlação entre a produção diária e a hora do dia, indicando uma relação direta e proporcional entre essas variáveis. Esse apontamento sugere que a hora do dia desempenha um papel crucial na determinação da produção diária, ressaltando a relevância de uma gestão temporal eficiente para otimização do desempenho do sistema.

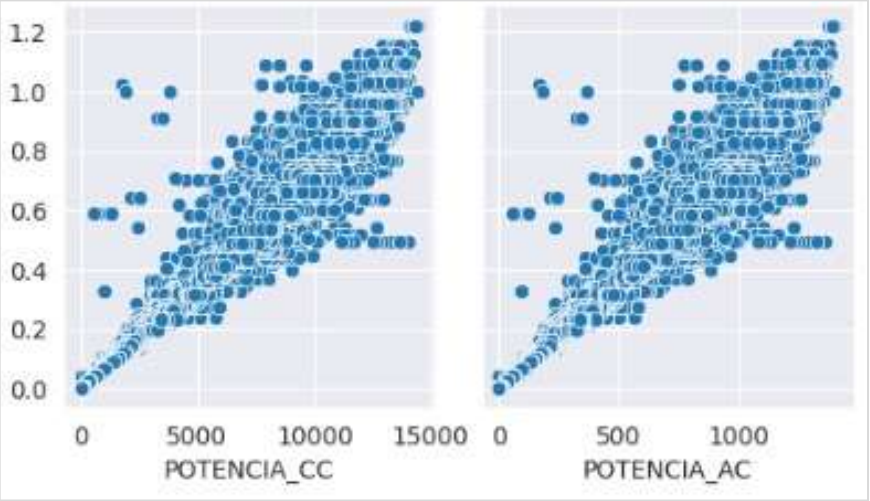
Em resumo, a observação atenta das correlações entre as variáveis fornece informações valiosas sobre o funcionamento do sistema, apontando para áreas específicas que

merecem uma atenção mais detalhada durante a análise e otimização do sistema de energia solar em questão.

Continuando a análise e aprofundando na primeira correlação identificada, que diz respeito à relação entre as potências, conforme ilustrado na Figura 7, notou-se que, embora a relação entre as potências seja proporcional, o crescimento não ocorre em uma escala semelhante. Surpreendentemente, constatou-se que a geração de corrente alternada (CA) representa apenas 10% da geração de corrente contínua (CC). O desequilíbrio encontrado sugere a existência de possíveis problemas nos inversores da usina, podendo estar relacionados a questões de manutenção ou mesmo à vida útil dos equipamentos.

A discrepância observada entre a geração de CC e CA destaca-se como um ponto crítico, exigindo uma investigação mais aprofundada para identificar e corrigir as potenciais falhas nos inversores. Essa análise mais detalhada poderá fornecer informações para melhorar a eficiência do sistema como um todo, contribuindo para um desempenho mais consistente e otimizado ao longo do tempo.

Figura 7 — Geração CC x CA em kW pela irradiação.



Fonte: Dados da pesquisa, 2023.

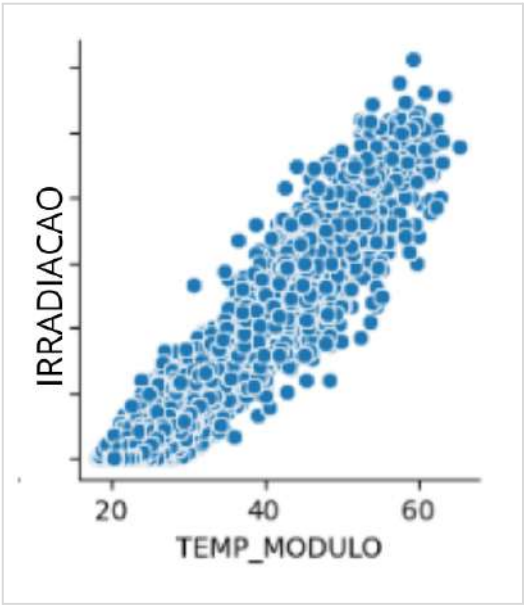
Entretanto, é importante ressaltar que a análise mais aprofundada necessária para identificar as causas do desequilíbrio na geração de corrente contínua (CC) e corrente alternada (CA) não está dentro do escopo do presente projeto. Optamos, portanto, por destacar a possível existência desse problema e concentrar nossos esforços no objetivo principal: a previsão da geração na usina com base nas condições atuais.

Ao focar na previsão da geração, podemos direcionar nossos recursos para otimizar e aprimorar a precisão do modelo, considerando as variáveis relevantes sem, necessariamente,

resolver de imediato as questões relacionadas aos inversores. Este enfoque estratégico nos permitirá atender aos objetivos estabelecidos para este projeto, enquanto a questão dos inversores pode ser tratada posteriormente em uma análise mais específica e direcionada.

No que diz respeito à relação entre a irradiação e a temperatura do módulo, observa-se, conforme evidenciado na Figura 8, a temperatura do módulo no momento é consequência da irradiação.

Figura 8 — Irradiação x temp. módulo

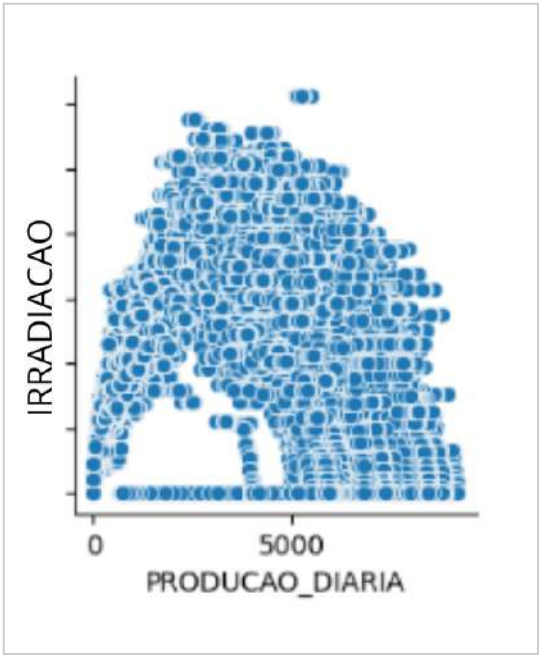


Fonte: Autoria própria, 2023.

Ambas as variáveis, irradiação e temperatura do módulo, desempenham papéis cruciais no monitoramento eficaz da usina. A constatação de que o relacionamento entre essas variáveis não está sendo respeitado sugere a possibilidade da existência de um defeito. Isso ressalta a importância de monitorar de perto esses parâmetros, pois qualquer desvio do comportamento esperado pode indicar problemas operacionais ou falhas nos componentes do sistema.

A seguir, na Figura 9, é possível observar o padrão de comportamento da produção diária em relação à irradiação.

Figura 9 — Irradiação x produção diária (kW)

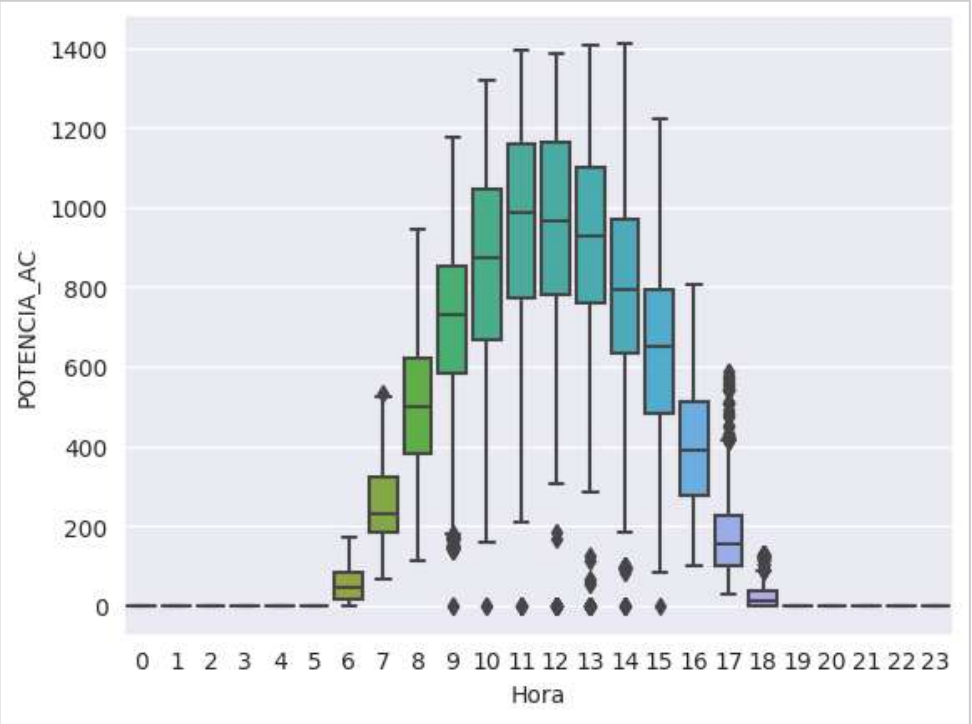


Fonte: Dados da pesquisa, 2023.

O comportamento em questão apresenta uma distribuição normal, evidenciando que a produção está diretamente vinculada aos valores de irradiação. Nota-se que a produção segue as oscilações correspondentes nos níveis de irradiação ao longo do período de monitoramento. Essa relação próxima entre a produção diária e a irradiação destaca a sensibilidade do sistema à variação na intensidade da luz solar, reforçando a importância de considerar essa variável ao realizar previsões ou otimizações no contexto da usina.

Concluindo a análise, foi crucial conduzir a detecção de valores atípicos, conhecidos como *outliers*. *Outliers* são pontos que se destacam, não seguindo o padrão esperado, e podem indicar eventos como desligamentos na usina, problemas nos equipamentos, ou outras irregularidades. A Figura 10 proporciona uma análise da geração de corrente alternada (CA) ao longo do dia na usina, permitindo a identificação de padrões e potenciais *outliers*. Este passo é fundamental para garantir a integridade dos dados e promover uma compreensão mais precisa do comportamento do sistema, contribuindo assim para uma análise mais robusta e confiável.

Figura 10 — Geração CA (kW) x Hora.



Fonte: Dados da pesquisa, 2023.

A figura 10, expõe uma gráfico conhecido como boxplot, um *boxplot* é um gráfico estatístico que oferece uma representação visual da distribuição de um conjunto de dados. Ele inclui uma caixa que abrange o intervalo interquartil com uma linha indicando a mediana. Os "bigodes" se estendem até os valores extremos dentro de um alcance aceitável, e pontos fora desse alcance são considerados *outliers*. O *boxplot* é útil para visualizar a dispersão, simetria e presença de valores atípicos em um conjunto de dados.

Ao analisar a figura 10, destaca-se uma presença significativa de valores zero na geração durante o intervalo de 9:00 às 15:00. Em contextos de geração solar, esse padrão requer uma investigação aprofundada. Como sugerido pela Figura 7, existe uma possível indicação de defeito nos inversores da usina, o que pode explicar esses valores abaixo do esperado. No entanto, é importante ressaltar que essa análise mais detalhada está além do escopo do presente projeto.

Dado que o foco principal reside na assertividade do modelo de previsão, optamos por desconsiderar esses valores e excluí-los do conjunto de dados. Vale mencionar que foram identificados 63 valores zero nesse intervalo de tempo, aproximadamente 0,09% dos dados totais (68.778), uma quantidade considerada irrelevante em relação ao tamanho total do banco de dados disponível. Essa abordagem visa garantir a consistência e confiabilidade dos dados utilizados no desenvolvimento do modelo de previsão.

Em resumo, a análise exploratória possibilitou uma compreensão do estado operacional da usina, identificando as principais variáveis a serem avaliadas e destacando o impacto delas sobre a variável alvo relacionada à potência de corrente alternada (CA). Essa abordagem forneceu grandes informações sobre a inter-relação entre diferentes parâmetros, evidenciando áreas de atenção e, potencialmente, indicando a necessidade de futuras investigações para otimização e manutenção do sistema. O entendimento obtido por meio da análise exploratória servirá como base sólida para o desenvolvimento e aprimoramento do modelo de previsão, contribuindo para a eficiência e confiabilidade da operação da usina.

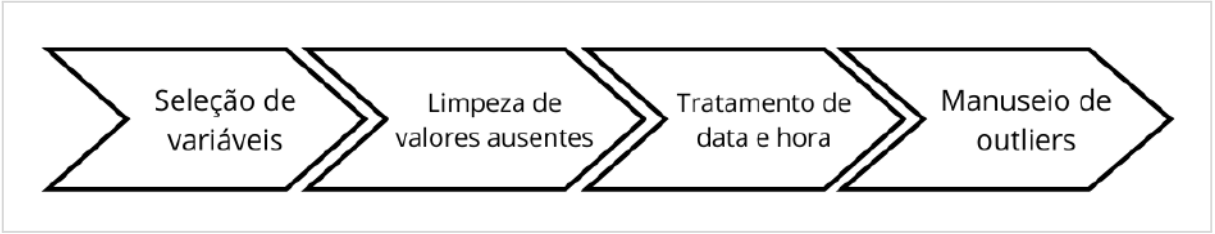
3.3. TRATAMENTO DE DADOS

A etapa de tratamento de dados, muitas vezes denominada pré-processamento de dados, é uma fase crucial no ciclo de vida da análise de dados, desempenhando um papel fundamental na transformação de dados brutos em informações valiosas e prontas para análise. Esta etapa ocorre antes da aplicação de algoritmos ou técnicas analíticas, e seu principal objetivo é garantir que os dados estejam em um estado adequado para serem utilizados de forma eficaz.

Contextualizando melhor, podemos entender essa etapa como um processo de refinamento e organização dos dados brutos coletados durante a fase de aquisição. Durante a coleta, os dados podem apresentar uma variedade de desafios, como inconsistências, valores ausentes, *outliers* e formatos diversos. A etapa de tratamento visa abordar esses desafios, tornando os dados mais confiáveis, consistentes e prontos para análise.

Na Figura 11, é possível visualizar a estrutura e organização adotadas para o processo de pré-processamento de dados no âmbito deste trabalho. A figura oferece uma representação gráfica que ilustra de forma clara como as etapas de seleção de variáveis, limpeza de valores ausentes, tratamento de data e manuseio de *outliers* foram delineadas e integradas, proporcionando uma compreensão visual do fluxo de trabalho adotado para preparar os dados brutos para análises subsequentes.

Figura 11 — Fluxo de tratamento de dados.



Fonte: Autoria própria, 2023.

3.3.1. SELEÇÃO DE VARIÁVEIS

Com base na correlação identificada na Figura 6 durante a análise exploratória, notou-se que algumas variáveis exercem um impacto mais significativo sobre as demais. Essa observação possibilitou a seleção criteriosa das variáveis que apresentam uma relação mais forte, visando otimizar o processamento durante o desenvolvimento do modelo. Essa abordagem visa concentrar os esforços nas características mais relevantes, contribuindo para a eficiência e desempenho aprimorados do modelo em questão. Na Figura 12, podemos observar as variáveis selecionadas.

Figura 12 — Variáveis selecionadas.

DATA_HORA
POTENCIA_CC
POTENCIA_AC
TEMP_MODULO
TEMP_AMBIENTE
PRODUCAO_DIARIA
IRRADIACAO

Fonte: Dados da pesquisa, 2023.

Conforme mencionado anteriormente, a seleção das variáveis foi conduzida com base em suas correlações, dando prioridade àquelas que apresentam uma correlação principal com a variável alvo "POTENCIA_AC". A última será a variável com a qual nos dedicaremos para realizar a previsão. Essa abordagem estratégica visa direcionar o foco analítico para as características que mais influenciam a variável de interesse, fortalecendo a capacidade de previsão do modelo.

3.3.2. LIMPEZA DE VALORES AUSENTES

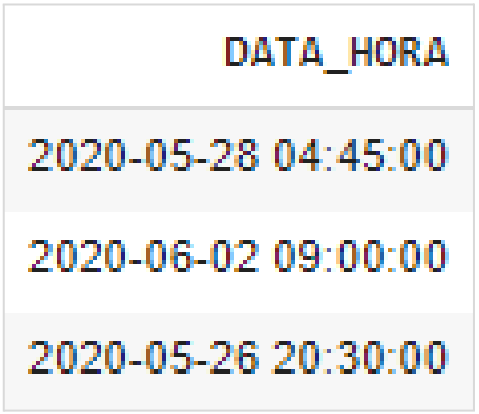
A limpeza de valores ausentes é uma etapa essencial no pré-processamento de dados, crucial para garantir a qualidade das análises. Ao remover ou tratar dados faltantes, evita-se polarização nos resultados, preserva a consistência e otimiza o desempenho de modelos analíticos. Isso facilita a interpretação clara dos resultados, previne erros em análises futuras e conserva recursos computacionais. Em resumo, a limpeza de valores ausentes é fundamental para assegurar a precisão e confiabilidade das informações provenientes dos dados.

Na seção 3.2, examinou-se a possível presença de valores faltantes, mas constatou-se que eles não estavam presentes em nosso conjunto de dados. Com isso, avançaremos para as demais etapas de tratamento. Entretanto, é crucial enfatizar a importância de realizar essa verificação para validar a confiabilidade dos dados manipulados e assegurar a necessidade desta etapa no processo de pré-processamento.

3.3.3. TRATAMENTO DE DATA E HORA

O tratamento das variáveis de data e hora desempenha um papel crucial no processo de pré-processamento de dados, sendo essencial para a preparação adequada dos dados destinados a análises subsequentes. Um aspecto fundamental desse tratamento é a padronização de formatos de data, exposto na Figura 13, garantindo consistência e facilitando a interpretação dos dados. Além disso, a extração de informações relevantes a partir de variáveis temporais, como dia da semana ou mês, enriquece a análise e pode ser crucial para a construção de modelos preditivos mais robustos.

Figura 13 — Data e hora.



DATA_HORA
2020-05-28 04:45:00
2020-06-02 09:00:00
2020-05-26 20:30:00

Fonte: Autoria própria

Além disso, o tratamento aprimora a capacidade de indexação e filtragem, tornando mais fácil a seleção de intervalos temporais específicos para análise. A visualização de dados

temporais também é beneficiada, com gráficos mais claros e precisos. Em cenários nos quais o conjunto de dados contém informações de diversas fontes, o tratamento de datas e horas facilita a sincronização e integração coesa desses dados.

A capacidade de realizar cálculos precisos de duração e intervalo entre eventos é outra vantagem significativa do tratamento das variáveis temporais. Em resumo, o tratamento adequado das variáveis de data e hora é essencial para garantir a qualidade, consistência e utilidade dos dados em análises temporais. Essa prática não apenas facilita a interpretação e visualização, mas também prepara os dados de maneira eficaz para serem utilizados em modelos analíticos e preditivos, contribuindo para insights mais precisos e informados.

3.3.4. MANUSEIO DE *OUTLIERS*

Como mencionado durante a análise exploratória, a identificação de *outliers* foi uma observação destacada, conforme ilustrado na Figura 10. Abordar valores atípicos é uma parte crucial do pré-processamento de dados, e é importante ressaltar que a exclusão nem sempre é a abordagem correta. Neste trabalho, a decisão de exclusão foi tomada considerando que esses valores poderiam impactar negativamente o comportamento do algoritmo. Além disso, a análise aprofundada para compreender a razão por trás da baixa geração não estava dentro do escopo do projeto. Cada decisão de tratamento de *outliers* deve ser cuidadosamente ponderada, levando em conta os objetivos específicos do projeto e o impacto potencial na validade e interpretação dos resultados.

Em conclusão, a etapa abrangente de tratamento de dados foi realizada de forma meticulosa, incluindo a limpeza de valores ausentes, o tratamento de *outliers* e a adequação das variáveis de data e hora. Essas ações foram essenciais para garantir a integridade e qualidade do conjunto de dados. Agora, com os dados preparados e refinados, deve-se avançar para as etapas subsequentes relacionadas ao desenvolvimento do modelo. Este processo robusto de pré-processamento estabeleceu uma base sólida, proporcionando dados confiáveis e consistentes que servirão como alicerce para análises mais aprofundadas e a construção efetiva do modelo preditivo.

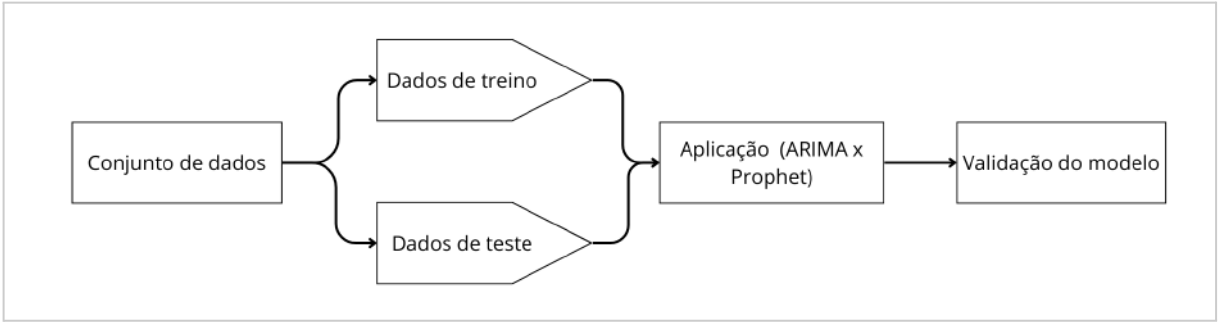
3.4. APLICAÇÃO DO MODELO

Neste estudo, foi explorada a aplicação de dois algoritmos amplamente utilizados, ARIMA e *Facebook Prophet*, na previsão da geração de energia solar. Com o objetivo de simplificar o processamento e facilitar a visualização, restringimos nossos dados de treino e

teste a um intervalo específico, do dia 13 ao dia 16 de junho, dividindo 70% do período para treino e 30% para validação/teste.

A aplicação dos modelos, ilustrada na Figura 14, compreendeu a separação do conjunto de dados em conjuntos de treinamento e teste, com foco exclusivo no treinamento durante o intervalo de 13 a 16 de junho. Para avaliar o desempenho de cada modelo, restringimos o período de teste ao dia 16 de junho. A delimitação desses intervalos tem como único propósito otimizar a capacidade de processamento dos algoritmos, preservando o objetivo principal de identificar o desempenho superior entre os modelos aplicados.

Figura 14 — Aplicação dos algoritmos.



Fonte: Autoria própria, 2023.

Ainda durante a fase de aplicação dos modelos, realizou-se um teste de estacionariedade. Em séries temporais estacionárias, média, variância e covariância permanecem constantes ao longo do tempo, possibilitando previsões confiáveis com base na suposição de que o comportamento futuro será semelhante ao passado. Se a série temporal não for estacionária, podendo apresentar tendências ou sazonalidade, é necessário transformá-la para atingir a estacionariedade antes de aplicar modelos de previsão.

Para constatar se a série é estacionária ou não, foi utilizado o teste estatístico de Dickey-Fuller, a fim de observar variáveis como termo ADF, *p value* e valores críticos, conforme observa-se na Tabela 3.

Tabela 3 — Teste ADF.

ADF	-3,821		
P Value	0,0027		
Valores críticos	**1%**	**5%**	**10%**
	-3,455	-2,872	-2,572

Fonte: Dados da pesquisa, 2023.

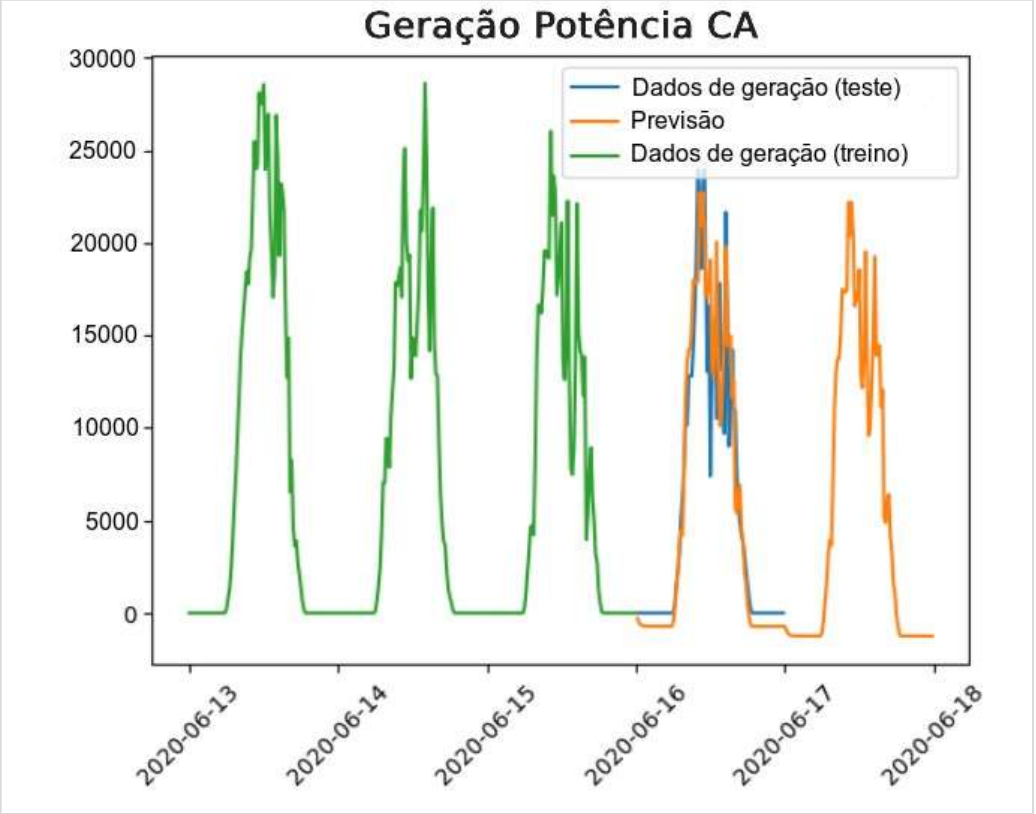
Em nossa análise, o teste ADF exibe um valor expressivamente negativo, fornecendo uma evidência robusta contra a existência de uma raiz unitária, o que sugere a não estacionariedade na série temporal. Adicionalmente, o valor de *p* é significativamente pequeno, inferior ao nível de significância convencional de 0,05. Consolidando ainda mais a rejeição da hipótese nula de raiz unitária, os valores críticos (1%, 5% e 10%) estão abaixo do valor de ADF, corroborando a presença de estacionariedade.

Após a constatação da estacionariedade da série e da implementação de ambos os algoritmos, avança-se para a etapa de validação do modelo. Nessa fase, será realizada a aplicação de métricas estatísticas como a principal ferramenta de avaliação e comparação entre os modelos utilizados.

3.4.1. Aplicação do ARIMA

Ao empregar o modelo ARIMA, o qual o resultado obtido está apresentado na Figura 16, é possível analisar tanto o padrão da previsão gerada pelo ARIMA quanto os dados reais já existentes em nosso conjunto de dados

Figura 15 — Geração CA.



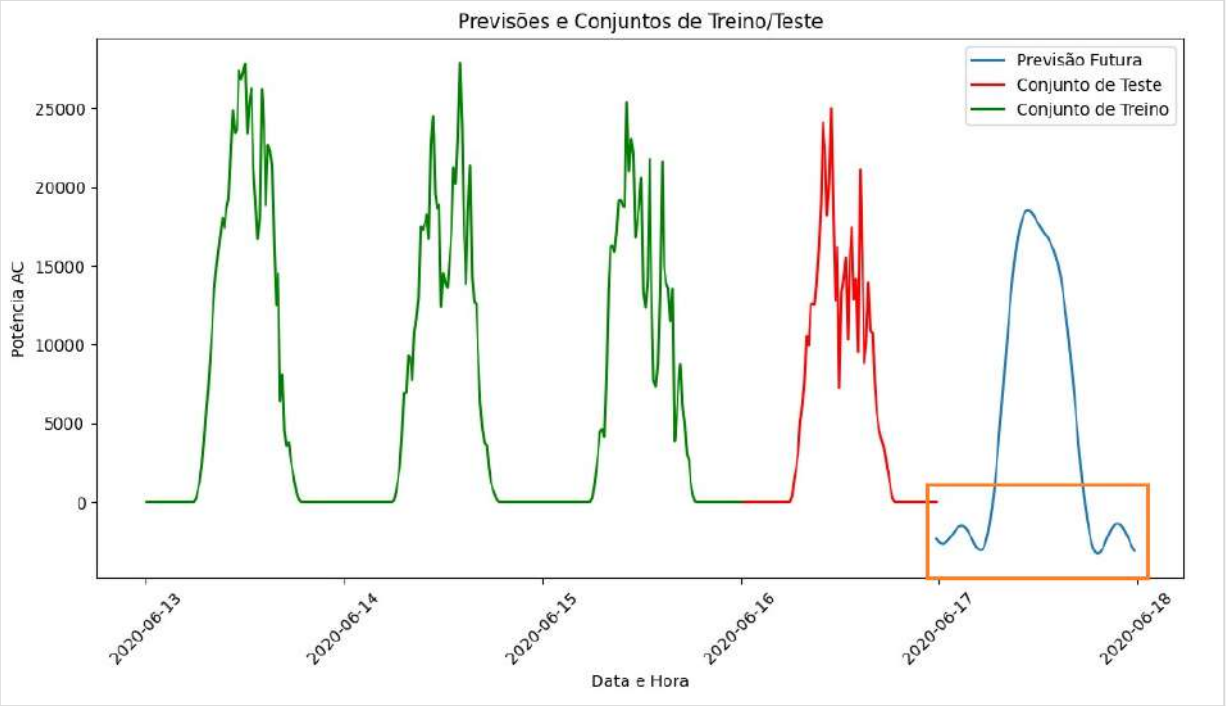
Fonte: Dados da pesquisa, 2023.

Como resultado da aplicação do modelo ARIMA, nota-se que a curva de previsão acompanha significativamente os dados de teste, indicando um ajuste sólido ao modelo. No entanto, é importante ressaltar que a presença de alguns valores negativos pode ter um impacto na integridade geral do modelo.

3.4.2. Aplicação do *Facebook Prophet*

Conforme mencionado anteriormente, ao treinar e aplicar o modelo do *Facebook Prophet*, é necessário dividir o conjunto de dados em conjuntos de treino e teste. Na Figura 17 abaixo, apresentamos a representação visual da execução das previsões, destacando a comparação entre esses dois conjuntos previamente mencionados.

Figura 16 — Previsões Potência CA (kW) pelo *Prophet*



Fonte: Dados da pesquisa, 2023.

A previsão, ainda que tenha acertado boa parte do comportamento das curvas de treino e teste, abrange uma área que apresenta potencial negativo (destacada em laranja na figura). A possível razão para esse cenário pode ser associada ao sobreajuste do modelo. O sobreajuste ocorre quando um modelo se adapta de maneira excessiva aos dados de treinamento, comprometendo sua capacidade de generalização para novos dados. Nessas circunstâncias, o modelo pode gerar previsões impraticáveis, manifestando-se, por exemplo, em valores negativos.

4. RESULTADOS DO MODELO

Para validar os modelos, utilizaremos três métricas estatísticas essenciais: R² *Score*, MAE e RMSE. Essas métricas desempenharão um papel crucial na avaliação da precisão de ambos os modelos.

A avaliação entre cada métrica varia de acordo com suas características particulares, vejamos:

- R² *Score*: Quanto maior, melhor. O R² *Score* varia de 0 a 1, onde 1 indica um ajuste perfeito do modelo aos dados e 0 indica que o modelo não explica a variabilidade dos dados.
- MAE (Erro Absoluto Médio): Quanto menor, melhor. O MAE é a média das diferenças absolutas entre as previsões do modelo e os valores reais. Um MAE mais baixo indica que o modelo tem menor erro médio.
- RMSE (Erro Quadrático Médio): Quanto menor, melhor. O RMSE é a raiz quadrada da média dos erros quadrados entre as previsões e os valores reais. Assim como o MAE, um RMSE mais baixo indica um modelo mais preciso.

Na tabela 4, estão expostos os resultados de cada métrica, vejamos abaixo:

Tabela 4 — Métricas de avaliação.

	ARIMA	Prophet
R² Score	0.894025	0.897808
MAE	15.882,18	20.236,22
RMSE	25.400,80	27.049,77

Fonte: Dados da pesquisa, 2023.

Ao analisar a validação dos algoritmos, observa-se que o *Facebook Prophet* supera o modelo ARIMA no que diz respeito ao R² *Score*, embora ambos tenham se aproximado significativamente do valor 1. No entanto, em relação às outras métricas (MAE e RMSE), o ARIMA se destacou de maneira substancial, revelando uma maior precisão em comparação ao seu concorrente.

5. CONSIDERAÇÕES FINAIS

O estudo dedicado à aplicação de modelos preditivos de aprendizado de máquina em séries temporais acompanha os avanços tecnológicos no setor de geração solar obtidos nos últimos anos. Inicialmente, foi proposta a análise da viabilidade desses modelos em um cenário específico, a usina localizada na Índia. Ao seguir a metodologia estabelecida para análise, tratamento de dados e implementação do modelo, pode-se examinar o desempenho do algoritmo de aprendizado de máquina *Facebook Prophet* em comparação com o modelo estatístico ARIMA.

Ao concluir a análise e implementação, observamos que o modelo ARIMA superou o *Prophet* em termos de desempenho, embora ambos tenham apresentado previsões que se distanciaram da realidade. Vale ressaltar a escolha da Potência CA como variável alvo, pois, ao visar o desempenho da usina, identificamos potenciais defeitos nos equipamentos durante a análise exploratória.

Como oportunidade de aprimoramento, sugerimos a implementação e teste de outros algoritmos baseado em outras formas de aprendizado de máquina além das séries temporais, bem como a adaptação do algoritmo a diferentes contextos, tanto em áreas residenciais quanto em usinas de grande porte, como abordado neste trabalho. Essas melhorias são de grande relevância para o mercado solar atual, pois, além dos cálculos de demanda realizados, prever a geração com base nas condições locais valorizará o produto e diminuirá a necessidade de futuras manutenções corretivas nas usinas.

REFERÊNCIAS

BOSCOA, Vini. Prophet: prevendo o futuro em séries temporais. Vini Boscoa's Blog, 2022. Disponívelem:https://www.viniboscoa.dev/blog/prophet-prevendo-o-futuro-em-series-temporais. Acesso em: 29 de nov. 2023.

Carmo, C. R. S.; Silva, J. R. M. (2023). Aprendizado de máquina e prestação de serviços de armazenamento de dados: métricas para análise e validação de algoritmos preditivos. Revista GeTeC, 17(1), 1-10. Acesso em: 29 de nov. 2023.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. Disponível em: https://api.semanticscholar.org/CorpusID:4650265. Acesso em: 25 de nov. 2023.

DATARISK. Ciência de dados e energias renováveis: Facebook Prophet para a previsão de geração de energia solar. Data Risk, 2023. Disponível em: https://www.datarisk.io/ciencia-de-dados-e-energias-renovaveis-facebook-prophet-para-a-previsao-de-geracao-de-energia-solar/. Acesso em: 29 de nov. 2023.

DUFFIE, J. A.; BECKMAN, W. A.; MCGOWAN, J. G. Solar Engineering of Thermal Processes. American Journal of Physics, v. 53, p. 382-382, 1985. Disponível em: https://api.semanticscholar.org/CorpusID:129469958. Acesso em: 25 de nov. 2023.

ENGIE. Avanço das tecnologias suportam a transição energética. Disponível em: https://www.alemdaenergia.engie.com.br/avanco-das-tecnologias-suportam-a-transicao-energetica/. Acesso em: 23 de nov. 2023.

GÉRON, Aurélien. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn e TensorFlow. Alta Books, 2019. ISBN 9788550803814.

IRENA. Crescimento recorde em energias renováveis alcançado apesar da crise energética. Disponívelem:https://www.irena.org/News/pressreleases/2023/Mar/Record-Growth-in-Renewables-Achieved-Despite-Energy-Crisis-PT. Acesso em: 23 de nov. 2023.

LEWIS, Nathan S. Oportunidades de pesquisa para avançar a utilização da energia solar. Science, v. 351, n. 6271, p. aad1920, 2016. DOI: 10.1126/science.aad1920. Disponível em: https://www.science.org/doi/full/10.1126/science.aad1920. Acesso em: 23 de nov. 2023.

NIELSEN, Aileen. Análise prática de séries temporais: predição com estatística e aprendizado de máquina. Edição em português. O'Reilly, 2021.

PORTAL SOLAR. Energia Solar Cresce 240 GW no mundo em 2022 e atinge 1,2 TW. Disponívelem:https://www.portalsolar.com.br/noticias/mercado/internacional/energia-solar-cresce-240-gw-no-mundo-em-2022-e-atinge-1-2-tw. Acesso em: 25 de nov. 2023.

RASCHKA, S.; PATTERSON, Joshua; NOLET, Corey J. Machine Learning in Python: Main developments and technology trends in data science, machine learning, and artificial intelligence. ArXiv, v. abs/2002.04803, 2020. Disponível em: https://api.semanticscholar.org/CorpusID:211082718. Acesso em: 27 de nov. 2023.

RODRIGUEZ-GALIANO, Victor Francisco; SÁNCHEZ-CASTILLO, Manuel; CHICA-OLMO, Mario; CHICA-RIVAS, Mario. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, v. 71, p. 804-818, 2015. Disponível em: https://api.semanticscholar.org/CorpusID:129774903. Acesso em: 27 de nov. 2023.

Siderurgia Brasil. Usinas solares de grande porte. Disponível em: https://siderurgiabrasil.com.br/2022/06/01/usinas-solares-de-grande-porte/. Acesso em: 25 de nov. 2023.

SINGH, Ajit. Machine Learning With Python. 2019. Disponível em: https://api.semanticscholar.org/CorpusID:67183937. Acesso em: 25 de nov. 2023.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus João Pessoa
Av. Primeiro de Maio, 720, Jaguaribe, CEP 58015-435, Joao Pessoa (PB)
CNPJ: 10.783.898/0002-56 - Telefone: (83) 3612.1200

Documento Digitalizado Ostensivo (Público)

Tcc Sávio Murillo Dias Bastos

Assunto:	Tcc Sávio Murillo Dias Bastos
Assinado por:	Savio Murillo
Tipo do Documento:	Projeto
Situação:	Finalizado
Nível de Acesso:	Ostensivo (Público)
Tipo do Conferência:	Cópia Simples

Documento assinado eletronicamente por:

- **Sávio Murillo Dias Bastos**, **ALUNO (20191610041) DE BACHARELADO EM ENGENHARIA ELÉTRICA - JOÃO PESSOA**, em 22/12/2023 16:55:13.

Este documento foi armazenado no SUAP em 22/12/2023. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1035486
Código de Autenticação: 6b479a98bb