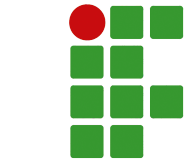

Instituto Federal de Goiás
Câmpus Formosa
Análise e Desenvolvimento de Sistemas
http://www.ifg.edu.br/formosa

APLICAÇÃO DE ESTRATÉGIAS DE APRENDIZADO DE MÁQUINA AO PROBLEMA DE PREDIÇÃO DE ENERGIA SOLAR FOTOVOLTAICA

RUAN ROCHA DE SOUZA FEITOSA

Trabalho de Conclusão de Curso

FORMOSA
2025

INSTITUTO FEDERAL
Goiás

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
SISTEMA INTEGRADO DE BIBLIOTECAS

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

[ ] Tese	[ ] Artigo Científico
[ ] Dissertação	[ ] Capítulo de Livro
[ ] Monografia – Especialização	[ ] Livro
[X] TCC - Graduação	[ ] Trabalho Apresentado em Evento
[ ] Produto Técnico e Educacional - Tipo: _____	

Nome Completo do Autor: Ruan Rocha de Souza Feitosa
Matrícula:20231070130096
Título do Trabalho: Aplicação de estratégias de aprendizado de máquina ao problema de predição de energia solar fotovoltaica

Autorização - Marque uma das opções

1. ( ) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. (X) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data 27/01/2027 (Embargo);
3. ( ) Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

( ) O documento está sujeito a registro de patente.
(X) O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
( ) Outra justificativa: _____

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

i. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
ii. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
iii. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Formosa, 29/01/2026.
Local Data

gov.br
Documento assinado digitalmente
RUAN ROCHA DE SOUZA FEITOSA
Data: 30/01/2026 15:14:44-0300
Verifique em https://validar.iti.gov.br

_____
Assinatura do Autor e/ou Detentor dos Direitos Autorais

Instituto Federal de Goiás
Câmpus Formosa
Análise e Desenvolvimento de Sistemas
http://www.ifg.edu.br/formosa

APLICAÇÃO DE ESTRATÉGIAS DE APRENDIZADO DE MÁQUINA AO PROBLEMA DE PREDIÇÃO DE ENERGIA SOLAR FOTOVOLTAICA

Ruan Rocha de Souza Feitosa

Trabalho de Conclusão de Curso apresentado ao Departamento de Áreas Acadêmicas do Instituto Federal de Goiás campus Formosa, como requisito parcial para obtenção do grau de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador: Prof. Me. Mario Teixeira Lemes

FORMOSA
2025

Dados Internacionais de Catalogação na Publicação

F311a

Feitosa, Ruan Rocha de Souza.

Aplicação de estratégias de aprendizado de máquina ao problema de predição de energia solar fotovoltaica. / Ruan Rocha de Souza Feitosa - 2025.

51 f. : il.

Trabalho de conclusão de curso - Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Formosa, 2025.

Orientador: Prof. Me. Mario Teixeira Lemes

1. Aprendizado do computador. 2. Energia solar- Predição. 3. Modelagem hídrica. I. Lemes, Mario Teixeira. II. Título. III. Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Formosa.

CDD: 333.793

Bibliotecária: Jéssica Fleury Nunes CRB1-DF no 3277

Formulário de Metadados para Disponibilização da Produção Técnico-Científica no ReDi IFG

TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO (TCC)
MONOGRAFIA DE ESPECIALIZAÇÃO

* Preenchimento Obrigatório

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

	TCC		Monografia de Especialização

Informe o título do documento. NÃO DIGITAR EM CAIXA ALTA!	
Título: *Aplicação de Estratégias de Aprendizado de Máquina ao Problema de Predição de Energia Solar Fotovoltaica	
Informe o título alternativo. Recomenda-se preencher com a tradução do título para o inglês, para maior visibilidade do documento.	
Título Alternativo:* Application of machine learning strategies to the problem of photovoltaic solar energy prediction	
Permissão de acesso ao documento*	Acesso aberto ( ) Acesso restrito (X) Embargo ( )
Se o documento for de acesso restrito ou embargo, informe o motivo:	( ) O documento está sujeito a registro de patente. (X) O documento pode vir a ser publicado como livro, capítulo de livro ou artigo. ( ) Outra justificativa: _____
Caso haja restrição de acesso, indicar data para que o documento possa ser disponibilizado no ReDI.	
Data para disponibilização no ReDI:	27 /01/2027
Informe a data da defesa	
Data da defesa*:	13/01/2026

AUTOR(ES)

1	Informe o nome do(s) autores(s), conforme o formato de citação.	
	Último Nome + "Jr", Ex. Silva Último Nome: *	Feitosa
	Primeiro(s) nome(s), ex. João Primeiro Nome: *	Ruan Rocha de Souza
	URL do Currículo Lattes:	http://lattes.cnpq.br/7562748275488362

ORIENTADOR

Informe o nome do orientador, conforme o formato de citação.	
Último Nome + "Jr", Ex. Silva Último Nome: *	Lemes

Primeiro(s) nome(s), ex. João Primeiro Nome: *	Mario Teixeira
URL do Currículo Lattes: *	http://lattes.cnpq.br/4918126641251231

MEMBROS DA BANCA

1	Informe o nome do(s) membro(s) da banca, conforme o formato de citação.	
	Último Nome + "Jr", Ex. Silva Último Nome: *	Lemes
	Primeiro(s) nome(s), ex. João Primeiro Nome: *	Mário Teixeira
	URL do Currículo Lattes: *	http://lattes.cnpq.br/4918126641251231
2	Último Nome + "Jr", Ex. Silva Último Nome: *	Neto
	Primeiro(s) nome(s), ex. João Primeiro Nome: *	Afrânio Furtado de Oliveira
	URL do Currículo Lattes: *	http://lattes.cnpq.br/1169788371765122
3	Último Nome + "Jr", Ex. Silva Último Nome: *	Paiva
	Primeiro(s) nome(s), ex. João Primeiro Nome: *	João Ricardo Braga de
	URL do Currículo Lattes: *	http://lattes.cnpq.br/9175757340129330

DESCRIÇÃO DO TRABALHO

Informe as palavras-chave do documento descrito. Sugere-se também o uso de termos em inglês. Caso o idioma original seja inglês optar por outro idioma.	
Palavras-Chave: *	Aprendizado de máquina; Energia solar fotovoltaica; Predição; Modelagem híbrida.
Selecione a grande área, área do conhecimento e subárea correspondente, de acordo com tabela do CNPq.	
Áreas de conhecimento de acordo com tabela do CNPq: *	Grande área: Ciências Exatas e da Terra / Área: Ciência da Computação / Subárea: Metodologia e Técnicas da Computação
Resumo do documento. Preencha o campo de acordo com o idioma do documento.	
Resumo: *	A predição da geração de energia solar é essencial para garantir a estabilidade das redes elétricas, otimizar o planejamento energético e ampliar a integração de fontes renováveis de forma segura e eficiente. Este trabalho apresenta estudo sobre a utilização de algoritmos de aprendizado de máquina na predição de energia solar fotovoltaica. Foram analisados os algoritmos Árvore de Decisão (AD), Floresta Aleatória (FA) e Perceptron Multicamada (PM), bem como suas combinações em modelos híbridos duplos e em uma configuração híbrida tripla. A pesquisa explorou dados meteológicos públicos de geração de energia solar de três locais distintos, avaliado por meio das métricas de Erro Médio Absoluto (EMA), Raiz do Erro Quadrático Médio (REQM) e Coeficiente de Determinação (R^2). Os resultados indicam que os modelos híbridos superaram os individuais, com

	destaque para a modelagem híbrida composta pela Floresta Aleatória e Perceptron Multicamada, que apresentou reduções de até 12,5% no EMA, melhorias de aproximadamente 6,6% no REQM e valores de R^2 superiores a 0,90 em todos os locais. Embora o modelo triplo tenha alcançado desempenho próximo em alguns cenários, não conseguiu superar o FA + PM de forma consistente. Os resultados revelam o potencial das abordagens híbridas para aumentar a confiabilidade na predição de potência gerada por sistemas fotovoltaicos, contribuindo positivamente para o planejamento e a operação de sistemas energéticos sustentáveis.

Abstract do documento. Preencha com o resumo em outro idioma.

Abstract: *	The prediction of solar energy generation is essential to ensure grid stability, optimize energy planning, and enable the safe and efficient integration of renewable sources. This work presents a study on the use of machine learning algorithms for photovoltaic solar energy prediction. The algorithms Decision Tree (DT), Random Forest (RF), and Multilayer Perceptron (MLP) were analyzed, as well as their combinations in dual hybrid models and a triple hybrid configuration. The research utilized public meteorological and solar generation datasets from three distinct locations, evaluated using the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) metrics. The results indicate that the hybrid models outperformed the individual ones, with emphasis on the hybrid model composed of Random Forest and Multilayer Perceptron (Model 6), which achieved reductions of up to 12.5% in MAE, improvements of approximately 6.6% in RMSE, and R^2 values above 0.90 across all sites. Although the triple hybrid model achieved comparable performance in some scenarios, it did not consistently surpass Model 6. The results demonstrate the potential of hybrid approaches to enhance the reliability of photovoltaic power generation prediction, contributing positively to the planning and operation of sustainable energy systems.

Referência bibliográfica do documento (como o documento deve ser citado). Use as normas de acordo com a área, por exemplo: ABNT, APA, Vancouver.

Citação: *	FEITOSA, R. R. S. Aplicação de estratégias de aprendizado de máquina ao problema de predição de energia solar fotovoltaica. Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) – Departamento de Áreas Acadêmicas, Instituto Federal de Goiás, Formosa Goiás, 2026.

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Na presente data realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso intitulado **Aplicação de estratégias de aprendizado de máquina ao problema de predição de energia solar fotovoltaica**, sob orientação de Mario Teixeira Lemes , apresentado pelo aluno **Ruan Rocha de Souza Feitosa (20231070130096)** do Curso **Superior de Tecnologia em Análise e Desenvolvimento de Sistemas (Câmpus Formosa)**. Os trabalhos foram iniciados às **19:00** do dia **13/01/2026** pelo Professor presidente da banca examinadora, constituída pelos seguintes membros:

- **Mario Teixeira Lemes** (Presidente)
- **Joao Ricardo Braga de Paiva** (Examinador Interno)
- **Afranio Furtado de Oliveira Neto** (Examinador Interno)

A banca examinadora, tendo terminado a apresentação do conteúdo do Trabalho de Conclusão de Curso, passou à arguição do candidato. Em seguida, os examinadores reuniram-se para avaliação e deram o parecer final sobre o trabalho apresentado pelo aluno, tendo sido atribuído o seguinte resultado:

[X] Aprovado [ ] Reprovado Nota : 9,3

Observação / Apreciações:

Proclamados os resultados pelo presidente da banca examinadora, foram encerrados os trabalhos e, para constar, eu **Mario Teixeira Lemes** lavrei a presente ata que assino juntamente com os demais membros da banca examinadora.

Documento assinado eletronicamente por:

- **Mario Teixeira Lemes**, em **27/01/2026 16:24:20** com chave **c755c920fbb511f0b46b005056a537a4**.
- **Joao Ricardo Braga de Paiva**, em **13/01/2026 21:55:35** com chave **bc3a0380f0e311f0b95b005056a537a4**.
- **Afranio Furtado de Oliveira Neto**, em **14/01/2026 11:00:44** com chave **6b3ab524f15111f0938f005056a537a4**.

Este documento foi emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QrCode ao lado ou acesse https://suap.ifg.edu.br/comum/autenticar_documento/ e informe os dados a seguir.

Tipo de Documento: Ata de Projeto Final

Data da Emissão: 27/01/2026

Código de Autenticação: 1d8d6e

Dedico este trabalho a todos que caminharam comigo e àqueles que, mesmo ausentes, continuam presentes em minha trajetória.

Agradecimentos

Gostaria de expressar minha mais profunda gratidão. Em primeiro lugar, aos meus pais, pelo apoio incondicional, incentivo e compreensão em cada etapa da minha trajetória. Agradeço igualmente aos meus tios, Lizandra e Watila, com um reconhecimento especial à tia Lizandra, que foi fundamental ao me apresentar o curso e ao Instituto Federal de Goiás - *Câmpus* Formosa. Meu agradecimento também se estende aos meus avós, pelo constante apoio, carinho e compreensão em todos os momentos.

Sou imensamente grato aos meus amigos, em especial à Patrícia Lima, cujo incentivo foi decisivo para que eu tentasse ingressar no curso, acreditando em mim desde o início. Registro também meu apreço a todos os professores do curso pelo conhecimento compartilhado, e dedico um agradecimento especial ao meu orientador, professor Mario Lemes, por acreditar em mim durante um período particularmente difícil, oferecendo paciência, orientação e apoio essenciais para a conclusão deste trabalho. Agradeço também à minha turma, cuja convivência, companheirismo e trocas frutíferas tornaram a jornada mais leve.

Por fim, direciono um agradecimento geral a todos que, de maneira direta ou indireta, contribuíram para a realização deste trabalho e para a minha formação acadêmica e pessoal. A cada um, meu sincero obrigado.

Parecemos estar levando tudo muito a sério, como se nos recusássemos a morrer antes de salvar o mundo de toda a sua feiura.

—ADAM SILVERA

Resumo

A predição da geração de energia solar é essencial para garantir a estabilidade das redes elétricas, otimizar o planejamento energético e ampliar a integração de fontes renováveis de forma segura e eficiente. Este trabalho apresenta estudo sobre a utilização de algoritmos de aprendizado de máquina na predição de energia solar fotovoltaica. Foram analisados os algoritmos Árvore de Decisão (AD), Floresta Aleatória (FA) e Perceptron Multicamada (PM), bem como suas combinações em modelos híbridos duplos e em uma configuração híbrida tripla. A pesquisa explorou dados meteorológicos públicos de geração de energia solar de três locais distintos, avaliados por meio das métricas de Erro Médio Absoluto (EMA), Raiz do Erro Quadrático Médio (REQM) e Coeficiente de Determinação (R^2). Os resultados indicam que os modelos híbridos superaram os individuais, com destaque para a modelagem híbrida composta pela Floresta Aleatória e Perceptron Multicamada, que apresentou reduções de até 12,5% no EMA, melhorias de aproximadamente 6,6% no REQM e valores de R^2 superiores a 0,90 em todos os locais. Embora o modelo triplo tenha alcançado desempenho próximo em alguns cenários, não conseguiu superar o FA + PM de forma consistente. Os resultados revelam o potencial das abordagens híbridas para aumentar a confiabilidade na predição de potência gerada por sistemas fotovoltaicos, contribuindo positivamente para o planejamento e a operação de sistemas energéticos sustentáveis.

Palavras-chave: Aprendizado de máquina; Energia solar fotovoltaica; Predição; Modelagem híbrida.

Abstract

The prediction of solar energy generation is essential to ensure grid stability, optimize energy planning, and enable the safe and efficient integration of renewable sources. This work presents a study on the use of machine learning algorithms for photovoltaic solar energy prediction. The algorithms Decision Tree (DT), Random Forest (RF), and Multilayer Perceptron (MLP) were analyzed, as well as their combinations in dual hybrid models and a triple hybrid configuration. The research utilized public meteorological and solar generation datasets from three distinct locations, evaluated using the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2) metrics. The results indicate that the hybrid models outperformed the individual ones, with emphasis on the hybrid model composed of Random Forest and Multilayer Perceptron (Model 6), which achieved reductions of up to 12.5% in MAE, improvements of approximately 6.6% in RMSE, and R^2 values above 0.90 across all sites. Although the triple hybrid model achieved comparable performance in some scenarios, it did not consistently surpass Model 6. The results demonstrate the potential of hybrid approaches to enhance the reliability of photovoltaic power generation prediction, contributing positively to the planning and operation of sustainable energy systems.

Keywords: Machine learning; Photovoltaic solar energy; Forecasting; Hybrid modeling.

Lista de Figuras

Lista de Tabelas

Lista de Acrônimos

Sumário

1

Introdução

Com o aumento da demanda energética mundial, a eletricidade tornou-se indispensável para suprir as necessidades de uma sociedade moderna e globalizada. Nesse cenário, a integração de fontes renováveis às redes elétricas surge como alternativa promissora do ponto de vista energético. No entanto, essa integração apresenta alta complexidade devido à natureza variável e imprevisível dessas fontes (Impram; Nese; Oral, 2020; Lara-Fanego et al., 2012; Gao; Wang; Shen, 2020; Espinar et al., 2010). Essa irregularidade resulta em desafios como a dificuldade de monitorar o balanço entre entrada e saída de energia, flutuações de tensão, perda de qualidade e instabilidade no fornecimento (Anderson; Leach, 2004; Moreno-Munoz et al., 2008).

Nesse contexto, entre as fontes de energia renovável, a energia solar se destaca por seu potencial de contribuição para um futuro sustentável (Lorenz et al., 2009). Os sistemas fotovoltaicos constituem uma das principais soluções para mitigar impactos das mudanças climáticas e promover práticas ambientalmente responsáveis (Victoria et al., 2021).

Entretanto, geração de energia a partir da luz solar é fortemente influenciada por variáveis meteorológicas incertas e incontroláveis, como temperatura do ar, cobertura de nuvens, irradiação difusa, direta, extraterrestre e em superfície horizontal. Tais fatores impactam não apenas o desempenho energético, mas também a viabilidade econômica e a confiabilidade operacional dos sistemas fotovoltaicos (Malvoni; De Giorgi; Congedo, 2017; Pierro et al., 2022).

Diante desses desafios, para garantir integração eficiente das fontes renováveis à rede elétrica, é essencial obter estimativas precisas de geração (Yang et al., 2021). A predição da energia solar fotovoltaica consiste em estimar a quantidade de energia que será produzida em um determinado intervalo de tempo. Essa capacidade de predição possibilita aos operadores agir de forma proativa diante de interrupções ou variações no fornecimento (Gaboitaolelwe et al., 2023).

Nesse sentido, modelos baseados em aprendizado de máquina têm demonstrado resultados promissores na predição de energias renováveis (Das et al., 2018). Por meio de algoritmos capazes de identificar padrões e relações nos dados sem necessidade de programação explícita, esses modelos constroem predições mais precisas, contribuindo para a gestão otimizada das redes elétricas (Leva et al., 2017).

Além disso, para aprimorar o desempenho preditivo, é fundamental detectar e eliminar

outliers nos conjuntos de dados utilizados no treinamento dos modelos. Considerando esse cenário, o objetivo geral deste trabalho é implementar e avaliar os algoritmos de aprendizagem de máquina Árvore de Decisão, Floresta Aleatória e Perceptron Multicamada para predição de energia solar fotovoltaica a partir de dados meteorológicos públicos, considerando o impacto da remoção de anomalias sobre a precisão dos modelos preditivos.

Os objetivos específicos incluem: (i) implementar os algoritmos de aprendizagem de máquina Árvore de Decisão (AD), Floresta Aleatória (FA) e Perceptron Multicamada (PM), individualmente e em combinação, utilizando dados meteorológicos públicos, (ii) analisar o impacto da remoção de anomalias com a técnica de Floresta de Isolamento, e (iii) avaliar o desempenho preditivo dos modelos individuais e híbridos por meio das métricas de Erro Absoluto Médio (EAM), Raiz do Erro Quadrático Médio (REQM) e Coeficiente de Determinação (R^2).

Esse trabalho está organizado em 5 (cinco) capítulos. No Capítulo 1, é apresentada a introdução ao tema, na qual se discute a relevância da predição da geração de energia solar, a importância das técnicas de aprendizado de máquina e as principais contribuições propostas neste estudo. O Capítulo 2 aborda os conceitos fundamentais, além de discutir trabalhos relacionados. Os materiais e métodos são descritos no Capítulo 3, enquanto o Capítulo 4 apresenta o processamento dos dados, a implementação e a avaliação dos algoritmos segundo as métricas adotadas. Por fim, o Capítulo 5 expõe a conclusão e apresenta recomendações para trabalhos futuros.

2

Referencial Teórico

Este capítulo tem como objetivo abordar conceitos fundamentais de energia solar fotovoltaica, incluindo o processo de conversão de energia, fatores de eficiência e diferentes tipos de sistemas fotovoltaicos existentes. Em seguida, são explorados os princípios da Inteligência Artificial (IA), com foco em aprendizado de máquina, através dos paradigmas de aprendizado supervisionado e não supervisionado. De forma complementar, é abordado o processo de treinamento de modelos e os tipos de modelos físicos, bem como os algoritmos Árvore de Decisão, Floresta Aleatória e Perceptron Multicamada. O Capítulo também discute desafios existentes, como *overfitting* e variância, e apresenta as principais métricas utilizadas para a avaliação de desempenho dos modelos de predição de séries temporais. Finalmente, é realizada uma análise de trabalhos correlatos, contextualizando a pesquisa na área de predição de sistemas fotovoltaicos.

2.1 Energia Solar Fotovoltaica

Energia solar fotovoltaica é definida como a energia obtida a partir da luz emitida pelo Sol (Bayod-Rújula, 2019). Uma das formas mais eficientes de converter essa energia em eletricidade é através do uso de células fotovoltaicas, dispositivos semicondutores projetados para transformar diretamente a luz solar em energia elétrica. Esse processo ocorre por meio do "efeito fotovoltaico", fenômeno no qual a radiação solar estimula elétrons do material semicondutor, gerando corrente elétrica. Além do uso direto do efeito fotovoltaico, sistemas de energia solar podem ser divididos em duas categorias solar térmica e solar elétrica. A energia solar térmica utiliza o calor do sol diretamente, sendo amplamente empregada para aquecer água em residências e piscinas. A energia solar elétrica converte a luz solar em eletricidade, também por meio do efeito fotovoltaico, utilizando células solares (Singh, 2013).

Quanto à configuração, sistemas fotovoltaicos podem ser configurados de diversas formas. Há sistemas autônomos, que funcionam de forma independente da rede elétrica, sistemas para veículos solares, e sistemas conectados à rede elétrica, que injetam a energia gerada no sistema de distribuição (Singh, 2013). Em locais onde o acesso à rede elétrica convencional é inviável, sistemas autônomos são utilizados para suprir as necessidades energéticas em localizações remotas (Gayen; Chatterjee; Roy, 2024). Esses sistemas, por não estarem conectados à rede

elétrica, apresentam diversidade elevada em tamanho e possibilidades de uso (Singh, 2013).

Em cenários conectados à rede, a energia gerada pelas células solares é convertida por meio de inversores e integrada ao sistema de distribuição. Essa tecnologia tem demonstrado sua necessidade, especialmente em situações de emergência, fornecendo energia quando o serviço da concessionária é interrompido (Singh, 2013). Além disso, vem sendo amplamente utilizada a produção em larga escala visando a diminuição de emissão de gases do efeito estufa e o bem estar ambiental (Gayen; Chatterjee; Roy, 2024) (Victoria et al., 2021).

2.2 Inteligência Artificial

A Inteligência Artificial (IA) é uma área da Ciência da Computação que permite o desenvolvimento de sistemas que podem realizar tarefas de forma inteligente, simulando os processos cognitivos humanos (Duan; Da Xu, 2012). Suas técnicas apresentam diversas vantagens, incluindo a capacidade de generalizar informações, lidar com múltiplas variáveis simultaneamente, integrar conhecimentos físicos em modelos e identificar padrões valiosos a partir de grandes volumes de dados (Das et al., 2018).

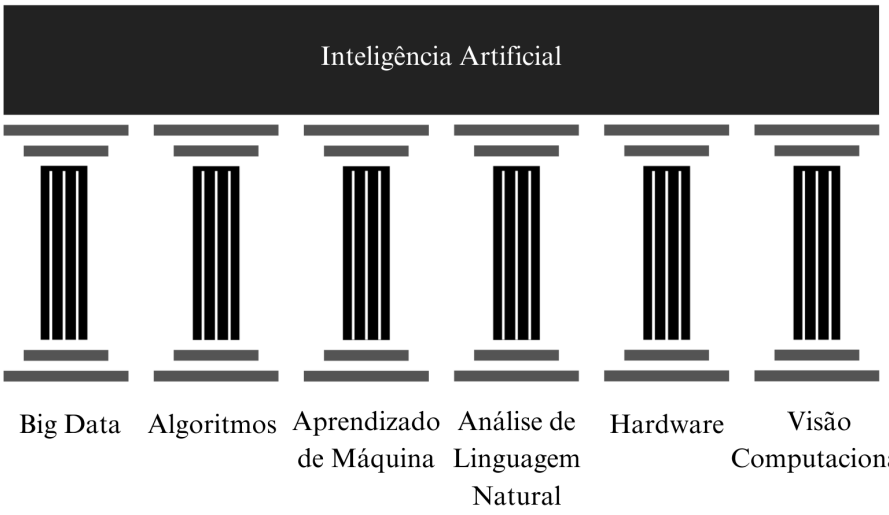

Figura 2.1: Pilares da paradigma de Inteligência Artificial

A Figura 2.1, apresenta os principais pilares que sustentam os sistemas de IA contemporâneos. O *Big Data* fornece grandes volumes de informações, essenciais para o treinamento de modelos preditivos, enquanto os algoritmos processam esses dados e definem regras que orientam o comportamento das aplicações. O Aprendizado de Máquina permite que os sistemas melhorem seu desempenho com base em padrões extraídos dos dados e o Processamento de Linguagem Natural torna possível a interpretação e geração da linguagem humana, viabilizando interfaces intuitivas entre homem e máquina. Além disso, a infraestrutura é sustentada pelo Hardware, que oferece o poder computacional necessário para aplicações da IA, e a Visão Computacional permite que sistemas interpretem imagens e vídeos, sendo aplicada em reconhecimento facial, análise de ambientes e automação (Zhang; Lu, 2021).

2.2.1 Aprendizado de Máquina

No âmbito da Inteligência Artificial, o aprendizado de máquina permite construção de sistemas que podem aprender por meio de dados. Com isso, computadores podem adquirir a capacidade de realizar tarefas sem a necessidade de serem explicitamente programados (Ray, 2019). Por meio dessa abordagem, modelos de aprendizado de máquina são amplamente utilizados para identificar padrões entre entrada e saída. Essa característica permite sua aplicabilidade em gama variada de problemas, por exemplo, o reconhecimento de padrões, resolução de impasses em classificação e desafios de predição (Voyant et al., 2017). Uma característica dos modelos de aprendizagem de máquina é a dependência de uma base de dados significante, pois esses modelos dependem diretamente da qualidade dos dados utilizados, o que torna fundamental a escolha e o preparo adequado (Gaboitaolelwe et al., 2023).

Nesse contexto, dados de treinamento consistem de conjunto de exemplos utilizados para ensinar o modelo. Cada exemplo é formado por um par, onde há um objeto de entrada (*input*) e o valor de saída desejado (*output*). Esse conjunto de padrões permite que o modelo aprenda a identificar relações entre dados de entrada e suas respectivas saídas, ajustando seus parâmetros para realizar predições ou classificações com base no aprendizado adquirido (Voyant et al., 2017). Para que essas predições sejam precisas, é fundamental um conjunto de dados de boa qualidade, com atributos apropriados que contribuam para a identificação de padrões relevantes (Gupta et al., 2021).

O processo de treinamento, este envolve a aplicação de modelos de aprendizado de máquina sobre um conjunto de dados previamente dividido em duas partes, que podem variar em tamanho dependendo do objetivo e base de dados. Aproximadamente 70% dos dados disponíveis são utilizados para treinar o modelo, permitindo aprendizado de padrões e ajuste de parâmetros. Os 30% restantes são reservados para a aplicação real e avaliação, sendo essas as etapas em que se avalia o resultado e desempenho do modelo (Das et al., 2018). Dessa forma, o processo de treinamento pode ser feito seguindo diferentes paradigmas de aprendizado, que se diferenciam principalmente pela forma como os dados de entrada e saída são apresentados ao modelo. Entre esses paradigmas, temos o aprendizado supervisionado e não supervisionado, nos quais o treinamento é realizado com base em dados não rotulados ou previamente rotulados, respectivamente.

2.2.1.1 Aprendizado Supervisionado e Não Supervisionado

No aprendizado supervisionado, o computador recebe entradas de exemplo e os resultados esperados fornecidos por um guia (ou professor). A partir disso, a máquina aprende com parâmetros definidos, como regras que relacionam dados de entrada e saída. Os dados são rotulados para treinar o modelo e permitir predição com base no conhecimento aprendido (Inman; Pedro; Coimbra, 2013). Esse tipo de método necessita de orientação durante o treinamento, e com os dados fornecidos, o modelo pode classificar informações em categorias específicas ou

realizar predição numéricas (Gaboitaolelwe et al., 2023).

De acordo com a Figura 2.2, os dados de entrada são inseridos no treinamento do modelo, representados pelas formas geométricas. Essas formas são rotuladas com uma saída esperada de organização, permitindo que a máquina aprenda com parâmetros definidos e estabeleça regras para relacionar entradas e saídas. Esse processo de treinamento com dados rotulados capacita o modelo a realizar tarefas com base no conhecimento adquirido.

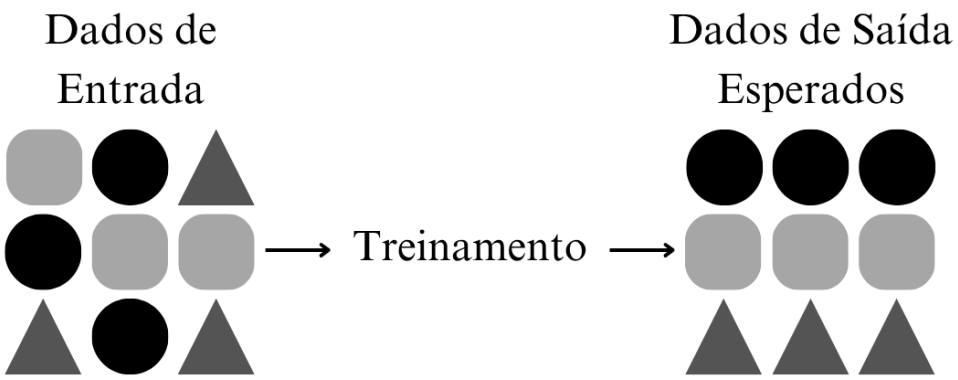

Figura 2.2: Fase de treinamento de aprendizado de máquina supervisionado.

Na fase de aplicação, apresentada na Figura 2.3, pode ser observada a aplicação de um modelo já treinado. Esse modelo, ao ser alimentado com figuras geométricas, as organiza com base no conhecimento adquirido durante o treinamento, aplicando regras aprendidas para classificar e agrupar as formas corretamente de acordo com dados rotulados.

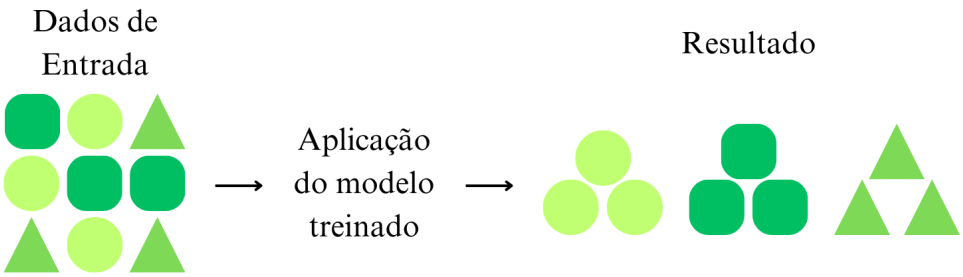

Figura 2.3: Fase de aplicação do aprendizado de máquina supervisionado.

O aprendizado não supervisionado utiliza dados não rotulados para ser treinado (Nguyen; Müsgens, 2022). Isso exige que o modelo identifique padrões ocultos ou agrupamentos de forma autônoma, sem informações prévias sobre saídas esperadas. Os algoritmos dessa abordagem geram modelos capazes de analisar, agrupar e categorizar dados, além de detectar anomalias (Gaboitaolelwe et al., 2023).

Conforme apresentado na Figura 2.4, formas geométricas não rotuladas, ou seja, aquelas que não possuem dados de entrada, saída ou uma organização específica, são inseridas em um modelo de aprendizado não supervisionado. Esse modelo busca identificar padrões e semelhanças entre os dados, aprendendo com eles sem a necessidade de rótulos pré-definidos.

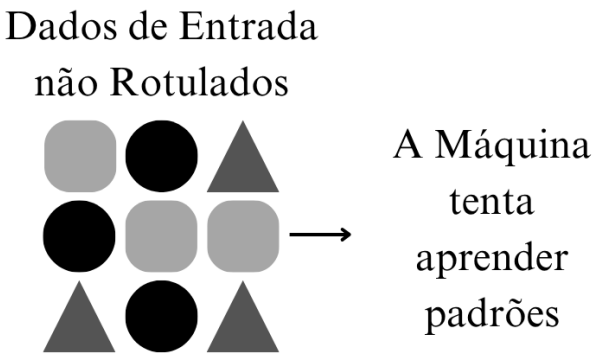

Figura 2.4: Fase de treinamento de aprendizado de máquina não supervisionado.

Na Figura 2.5, é apresentada a aplicação do modelo que foi previamente treinado de maneira não supervisionada. O modelo analisa as formas geométricas, identifica padrões e as agrupa com base em semelhanças em comum.

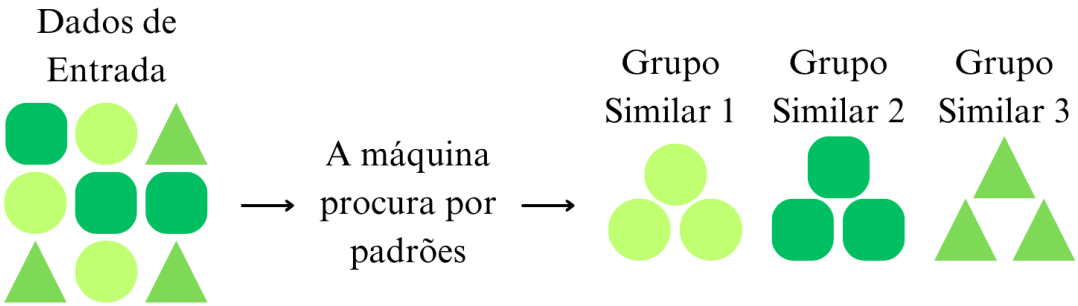

Figura 2.5: Fase de aplicação do aprendizado de máquina não supervisionado.

2.2.1.2 Modelos Físicos, Baseados em Dados e Modelagem Híbrida

Estimativas de produção de energia solar podem ser feitas com várias metodologias, as principais são modelos físicos e modelos de aprendizado de máquina. Sendo que a escolha do modelo a ser seguido depende de quanto tempo se deseja prever a saída de potência (Voyant et al., 2017). Os modelos físicos são baseados no comportamento do sistema de produção solar, levando em consideração propriedades físicas básicas, como eficiência dos painéis, ângulo de incidência da irradiação solar e o sombreamento dos painéis. Além disso, esses modelos incorporam variáveis ambientais, como temperatura, cobertura de nuvens e umidade (Gaboitaolelwe et al., 2023). Também são consideradas características específicas do sistema, como detalhes das instalações, configurações elétricas, localização geográfica e aspectos técnicos (Dobos, 2014) (Al-Dahidi et al., 2024).

Em contraste com os modelos físicos, modelos baseados em dados dependem de dados históricos para fazer inferências, identificar padrões e estabelecer relações que podem gerar predições. Eles se baseiam em estatística e algoritmos que aprendem com padrões e relações nos dados. Esses modelos podem ser estatísticos ou de aprendizado de máquina (Al-Dahidi et al., 2024).

Os modelos estatísticos analisam dados históricos e, a partir disso, fazem estimativas futuras. São utilizadas técnicas estatísticas e algoritmos matemáticos para ajustar dados históricos e identificar relações entre diferentes variáveis. Diferentemente dos modelos de aprendizado de máquina, modelos estatísticos não fazem parte dessa categoria, pois se fundamentam exclusivamente em métodos estatísticos.

Segundo Gaboitaolelwe et al. (2023), a principal diferença entre modelos estatísticos e modelos de aprendizado de máquina, é a necessidade de intervenção humana. Modelos estatísticos demandam atenção na seleção e planejamento de características. Por outro lado, o aprendizado de máquina pode ser aplicado com dados brutos, sem necessidade de pré-processamento.

Buscando superar as limitações das abordagens individuais, sistemas híbridos, combinam diferentes tipos de aprendizado de máquina para fazer predições, que podem ser modelos físicos ou baseado em dados (Nguyen; Müsgens, 2022). O objetivo é equilibrar vantagens e mitigar limitações de cada abordagem, embora isso possa acarretar custos computacionais elevados (Voyant et al., 2017)(Gaboitaolelwe et al., 2023). Para Al-Dahidi et al. (2024), comparados aos métodos tradicionais aplicados isoladamente, modelos híbridos oferecem resultados mais precisos sem comprometer a confiabilidade das predições.

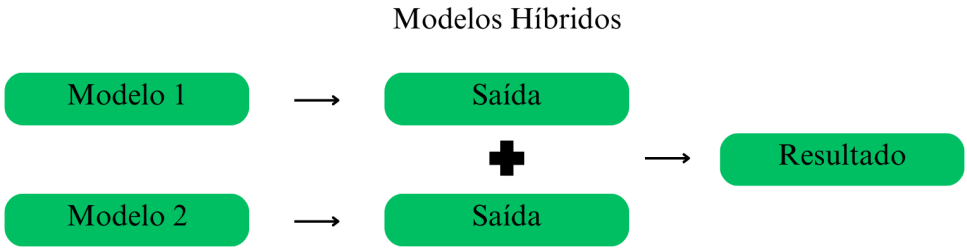

Figura 2.6: Representação da modelagem híbrida: as saídas de dois (ou mais) modelos distintos são combinadas para geração do resultado.

Conforme ilustrado na Figura 2.6, é apresentada a arquitetura de um modelo híbrido, na qual:

- O Modelo 1 de ML gera uma primeira predição;
- O Modelo 2, com abordagem distinta, produz uma segunda saída;
- Os resultados são armazenados e combinados, gerando uma saída.

Dentro do contexto de modelagem híbrida, o Regressor por Votação, constitui método que combina múltiplos modelos de regressão independentes com objetivo de produzir uma estimativa. A abordagem é baseada no princípio de que diferentes modelos podem capturar diferentes aspectos da estrutura dos dados, assim, ao integrar suas predições, obtem-se resultado com menor variância e maior capacidade de generalização. Cada regressor produz uma predição individual $fm(x)$ e com isso o resultado é calculado pela média das predições individuais, expresso como:

$$\hat{y}(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) \tag{2.1}$$

onde:

- $\hat{y}(x)$ representa o valor estimado da variável resposta para a amostra de entrada x;
- M indica o número total de modelos de regressão;
- m é o índice que identifica cada modelo individual, com $m = 1, 2, \ldots, M$;
- $f_m(x)$ denota a predição produzida pelo modelo de regressão para a entrada x;
- $\frac{1}{M}$ é a normalização da soma, resultando em uma média aritmética das predições.

2.2.1.3 Árvore de Decisão

A Árvore de Decisão é um modelo de aprendizado supervisionado que mapeia hierarquicamente um domínio de dados em um conjunto de respostas. Nesse sentido, o modelo divide os dados recursivamente em subdomínios, garantindo que cada divisão maximize o ganho de informação em relação ao nó anterior. O objetivo do algoritmo de otimização é encontrar a melhor divisão possível. Além disso, na estrutura da árvore, cada nó interno representa uma pergunta sobre uma característica dos dados, cada ramo corresponde a uma possível resposta e cada nó folha indica uma decisão final ou classe de saída (Suthaharan; Suthaharan, 2016).

A Figura 2.7 ilustra um exemplo de aplicação da Árvore de Decisão. Inicialmente, considera-se na raiz o conjunto de dados $X = 3,5,6,8,9,10$, onde cada elemento possui um rótulo de classe correspondente $R = 1,1,0,1,1,0$. O primeiro critério de divisão utiliza a média dos valores de $X(6,8)$, separando os dados em dois subconjuntos: $C0 = 3,5,6$ (elementos menores ou iguais a 6,8) e $K0 = 8,9,10$ (elementos maiores que 6,8). No entanto, como $C0$ ainda contém rótulos mistos (classes 1 e 0), aplica-se uma nova divisão usando sua média (4,6), resultando no nó folha $C1 = 3$ (classe 1 pura) e no nó interno $C2 = 5,6$. Em seguida, este subconjunto é dividido pela média 5,5, gerando os nós folha finais $C3 = 5$ (classe 1 pura) e $C4 = 6$ (classe 0 pura). Desse modo, o processo recursivo de divisão garante que todos os nós terminais alcancem pureza máxima em sua classificação.

No ramo direito da divisão inicial (valores maiores que 6,8) isola o subconjunto K0 = 8, 9, 10, onde rótulos correspondentes são R = 1, 1, 0. Visto que K0 também apresenta classes mistas, calcula a média de seus elementos (9) para estabelecer um novo critério de separação. A aplicação divide os dados em dois nós folhas definitivos: o K1 = 8, 9 (valores menores ou iguais a 9), que resulta em uma classificação pura da classe 1, e o K2 = 10 (valor maior que 9), que isola a classe 0. Atingindo também homogeneidade total também neste lado da árvore, completando a estruturação do modelo.

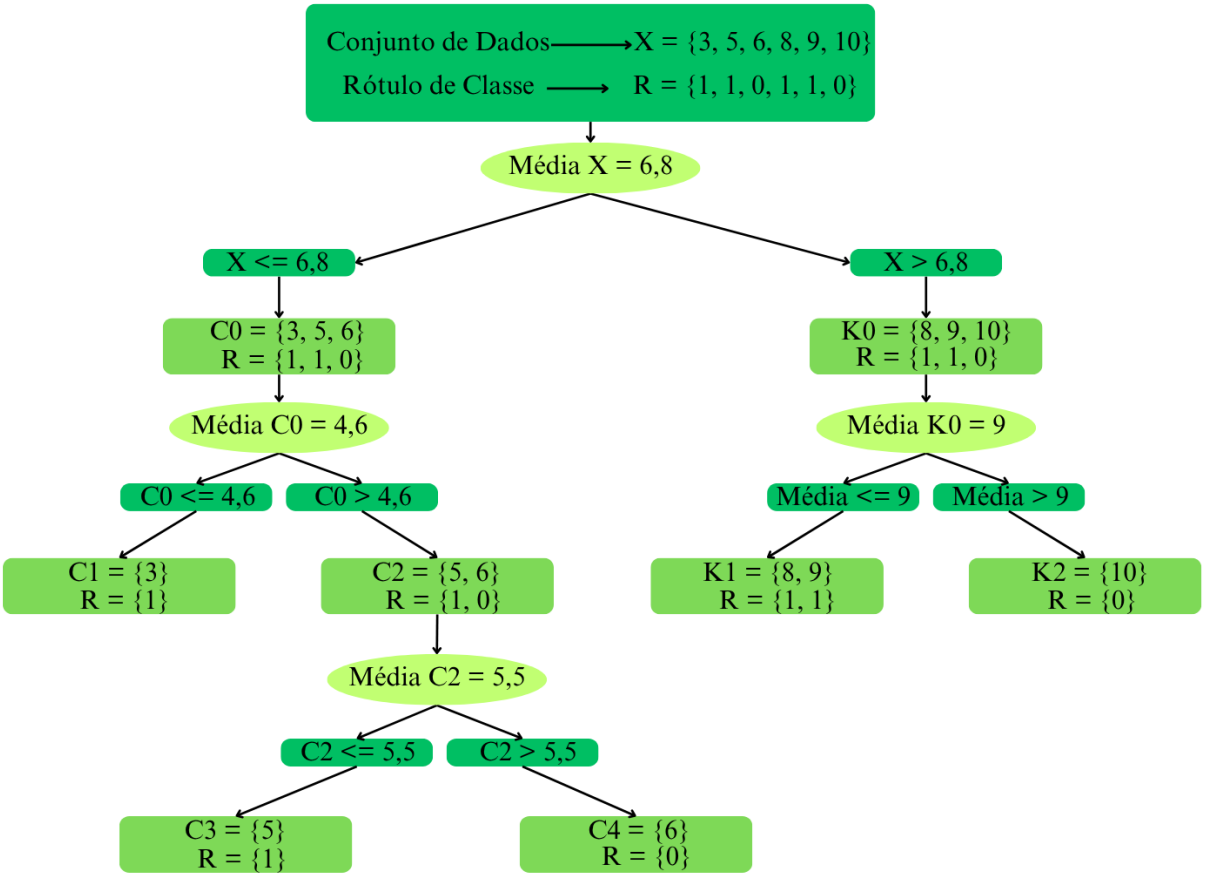

Figura 2.7: Exemplo de aplicação do algoritmo Árvore de Decisão

No contexto da regressão, os modelos de árvores de decisão são conhecidos pela simplicidade e eficiência para lidar com grande número de possibilidades, para isso usamos algoritmos de divisão e conquista, que faz uma divisão de dados em conjuntos menores. Contudo, é importante destacar a dificuldade desse modelo em lidar com decisões em níveis mais baixos da árvore. Formalmente, a estrutura lógica do modelo pode ser expressa pela Equação 2.2.

$$m(x) = \sum_{i=1}^{l} k_i \cdot I(x \in D_i) \tag{2.2}$$

Onde:

- $m(x)$ Representa o valor predito para a entrada x;
- k_i Valor constante associado à folha, retorna a predição do nó;
- D_i Representação de nó da arvore;
- $I(x \in D_i)$ Retorna 1 se x estiver em D_i e 0 se não estiver presente.

Nesse sentido, o algoritmo de Partição Recursiva (PR) constrói a árvore de decisão através de um processo recursivo de divisão do conjunto de treinamento em subconjuntos cada vez menores (Suthaharan; Suthaharan, 2016). O método opera sobre pares de dados (x, y), onde

x representa as características de entrada e y os valores alvo. Além disso, um aspecto importante do algoritmo é o critério de parada, implementado através de um nó de teste t. Para que uma amostra x seja direcionada a um determinado ramo da árvore, ela deve satisfazer a condição ótima de divisão s^* associada ao nó. Esse teste de decisão é necessário para o particionamento hierárquico dos dados. Portanto, em um conjunto de dados D podemos ter duas opções: uma que satisfaz o critério (Equação 2.3) e outra que não satisfaz (Equação 2.4). Essas condições são usadas para melhor definição de divisão para um determinado nó.

$$D = \{(x_i, y_i) : x_i \text{ satisfaz } s^*\} \tag{2.3}$$

$$D = \{(x_i, y_i) : x_i \text{ não satisfaz } s^*\} \tag{2.4}$$

Para a construção de um modelo de regressão, é fundamental determinar os parâmetros que minimizam o critério Mínimos Quadrados (MQ). O método dos mínimos quadrados é um critério estatístico amplamente utilizado para ajustar modelos de regressão. Sua principal finalidade é minimizar a soma dos quadrados das diferenças entre os valores observados e os valores previstos pelo modelo. A fórmula dos quadrados mínimos pode ser expressa pela Equação 2.5.

$$MQ = \frac{1}{n}\sum_{i=1}^{n}(y_i - r(\beta, x_i))^2 \tag{2.5}$$

Onde:

- MQ: Representa os mínimos quadrados;
- n: Quantidade de dados de amostra;
- y_i: Rótulo de treinamento;
- x_i: Amostra para treinamento;
- $r(\beta, x_i)$: Representa a predição feita pelo modelo de regressão para a entrada x.

Com o objetivo de minimizar o valor esperado do erro quadrático, utiliza-se K_l que é dado pela Equação 2.6:

$$K_l = \frac{1}{n_l}\sum_{D_l} y_i \tag{2.6}$$

Onde:

- K_l: Constante;
- y_i: Dado rotular;

- n_l: Número de exemplos em uma folha;
- D_l Conjunto de exemplos que caem na folha l.

Para avaliar os testes que otimizam a precisão da árvore, podemos quantificar o erro associado a cada nó t por meio da Equação 2.7:

$$Err(t) = \frac{1}{n_t} \sum_{D_t} (y_i - k_t)^2 \tag{2.7}$$

Onde:

- $Err(t)$: Representa o erro do nó analisado;
- n_t: Quantidade de dados da interação;
- y_i: Dado rotular;
- k_i: Erro quadrático minimizado dado pela Equação 2.6.
- D_t representa o conjunto de dados presente no nó t

A fim de realizar uma divisão binária, ou seja, particionamento do espaço de características em dois subconjuntos distintos, para isso, implementamos a regra de divisão criteriosa. Esta regra tem como objetivo principal minimizar o erro de predição da árvore resultante desse particionamento. Formalmente, definimos o erro da divisão s através da Equação 2.8:

$$Err(s,t) = \frac{n_{t_l}}{n_t} \cdot Err(t_l) + \frac{n_{t_R}}{n_t} \cdot Err(t_R) \tag{2.8}$$

Onde:

- $Err(s,t)$: Erro de divisão;
- t: O nó atual antes da divisão;
- s: Erro de uma divisão candidata;
- t_l: Nó à esquerda depois da divisão que contém os dados que satisfazem a condição;
- t_r: Nó à direita que contém os dados que não satisfazem a condição;
- n_t: número de exemplos em t;
- n_{t_l}: Número de exemplos de nós esquerdos;
- n_{t_R}: Número de exemplos de nós direitos;
- $Err(t_l)$ o erro calculado do nó esquerdo;

- $Err(t_r)$ o erro calculado do nó direito.

Dessa forma, pode-se determinar a melhor divisão para um nó t dado um conjunto s de possíveis divisões. Esse critério orienta a escolha das divisões nos nós internos da árvore de regressão. A cada iteração do algoritmo de partição recursiva, são testadas todas as divisões possíveis das variáveis, conforme Equação 2.9.

$$\Delta Err(s,t) = Err(t) - Err(s,t) \tag{2.9}$$

Onde:

- $\Delta Err(s,t)$: Melhor divisão;
- s: Erro de uma divisão candidata;
- t: O nó atual antes da divisão;
- $Err(t)$: Erro médio de um nó expresso pela Equação 2.7;
- $Err(s,t)$: Erro de divisão expresso pela Equação 2.8.

Esse critério orienta a escolha das divisões nos nós internos da árvore de regressão. A cada iteração do algoritmo de partição recursiva, são testadas todas as divisões possíveis das variáveis. Além disso também devemos definir uma profundidade máxima, que vamos testando ao decorrer de testes do modelo.

2.2.1.4 Floresta Aleatória

A Floresta Aleatória é um método de aprendizado de máquina supervisionado que se destaca tanto em tarefas de classificação quanto de regressão (Al-Dahidi et al., 2024). Nesse contexto, essa técnica utiliza um conjunto de árvores de decisão para produzir predições mais precisas e estáveis do que modelos baseados em uma única árvore (Gaboitaolelwe et al., 2023) (Al-Dahidi et al., 2024).

O princípio da Floresta Aleatória consiste na combinação de múltiplas árvores de regressão, onde cada árvore é treinada em um subconjunto aleatório dos dados de treinamento. Além disso, a cada divisão de um nó da árvore, apenas um subconjunto aleatório de características é considerado, o que aumenta a diversidade entre as árvores e reduz a correlação entre elas. Como resultado, a predição final do modelo é obtida pela média das predições individuais de todas as árvores, resultando em um modelo generalizado e menos suscetível a *overfitting* (Al-Dahidi et al., 2024).

O *overfitting* é considerado um desafio em aprendizado de máquina. Esse fenômeno ocorre quando um modelo se ajusta excessivamente aos dados de treinamento, a ponto de perder a habilidade de se aplicar a dados novos, resultando em um desempenho inferior nos testes (Ying,

2019). Em vez de identificar padrões úteis, o modelo acaba "decorando"os dados, incluindo variações e elementos irrelevantes, que não são representativos do comportamento real dos dados.

Esse problema pode ser visto como uma falta de equilíbrio entre a capacidade do modelo de se ajustar bem aos dados observados e sua habilidade de fazer predições eficazes com dados inéditos (Ying, 2019). Quando o *overfitting* ocorre, o modelo pode ter um desempenho excepcional nos dados de treinamento, mas não consegue generalizar corretamente para o conjunto de teste, o que leva a uma queda de precisão em novas situações.

As Árvores de Decisão individuais tendem a se ajustar excessivamente aos dados de treinamento, capturando ruídos e particularidades. Formalmente, a Equação 2.10 representa o funcionamento da Floresta Aleatória.

$$RF = \frac{1}{n}\sum_{i=1}^{n} AD \tag{2.10}$$

Onde:

- RF: Representa a Floresta Aleatória;
- n: Número de elementos da iteração;
- AD: Representa a predição realizada pelo algoritmo.

2.2.1.5 Perceptron Multicamada

O Perceptron Multicamada é uma rede neural composta por várias camadas de neurônios interconectados. Esses neurônios utilizam funções de ativação não lineares, permitindo que a rede reconheça e interprete padrões complexos presentes nos dados. De modo geral, a rede é estruturada em três camadas: (i) entrada, onde são inseridas as variáveis independentes, (ii) ocultas, responsáveis pelos cálculos e transformações dos dados, e (iii) saída, que é responsável por gerar as predições (Al-Dahidi et al., 2024).

Matematicamente, o funcionamento do Perceptron Multicamada é apresentado na Equação 2.11.

$$\hat{y} = f_a\left[\sum_{i=1}^{n} w_i\left(\sum_{j=1}^{m} w_j x_i + b_1\right) + b_2\right] \tag{2.11}$$

Na Equação 2.11, f_a representa a função de ativação do neurônio oculto, denominada *Leaky Rectified Linear Unit* (Leaky ReLU) (Al-Dahidi et al., 2024). Além disso, os parâmetros m e n correspondem, respectivamente, ao número de neurônios nas camadas ocultas e de saída. Os pesos das conexões são indicados por w_i e w_j, enquanto os vieses das camadas de entrada e saída são representados por b_1 e b_2.

O processo de aprendizado da rede é realizado por meio do método de retropropagação do erro, que ajusta os pesos e vieses com base nas informações mais recentes, buscando minimizar o erro na camada de saída. Com isso, a atualização dos pesos é descrita pela Equação 2.12.

$$w^* = w - \alpha \frac{\partial e}{\partial w} \tag{2.12}$$

Onde:

- $\hat{y}$ é a predição do modelo;
- w^* é o novo peso atualizado;
- w representa o peso anterior;
- α é a taxa de aprendizado;
- $\frac{\partial e}{\partial w}$ indica a variação do erro em função do peso.

Por fim, a função de erro utilizada para avaliar o desempenho do modelo é expressa pela Equação 2.13.

$$e = \frac{1}{2} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{2.13}$$

Na Equação Equação 2.13, os parâmetros são definidos da seguinte forma:

- e: erro médio quadrático do modelo;
- $\hat{y}_i$: valor previsto pela rede;
- y_i: valor real esperado para a amostra;
- n: número total de amostras;
- $(\hat{y}_i - y_i)^2$: erro ao quadrado entre o valor previsto e o valor real;
- $\frac{1}{2}$: fator utilizado para simplificar o cálculo da derivada na etapa de otimização.

2.2.2 Variância e Tratamento de Anomalias

Variância é uma medida que indica o quanto as predições de um modelo mudam quando ele é treinado em diferentes conjuntos de dados. Modelos com alta variância são muito sensíveis a pequenas flutuações nos dados de treinamento, o que significa que podem capturar até mesmo o ruído presente nesses dados (Voyant et al., 2017) (Al-Dahidi et al., 2024).

Essa sensibilidade excessiva é uma característica de modelos complexos e está diretamente relacionada ao *overfitting*. A alta variância está associada ao comportamento instável do modelo diante de dados diferentes. (Ying, 2019).

Anomalias são padrões em dados que não tem comportamento e estrutura normal (Liu; Ting; Zhou, 2008). Podem ser induzidas nos dados por uma variedade de razões, como quebra

de sistema ou erros de medição em sensores no geral, como meteorológicos e medidores de saída de energia (Chandola; Banerjee; Kumar, 2009).

Adicionalmente, em cenários onde os dados envolvem múltiplas variáveis dependentes do tempo, as técnicas de detecção de *outliers* devem considerar correlações entre essas variáveis para identificar anomalias temporalmente alinhadas (Blázquez-García et al., 2021). Esses métodos não apenas capturam desvios pontuais em uma única dimensão, mas também padrões atípicos que afetam simultaneamente várias variáveis, indicando falhas ou eventos de maior impacto.

2.2.2.1 Floresta de Isolamento

A Floresta de Isolamento é um método de detecção de anomalias baseado em partições. Esse algoritmo aproveita duas propriedades fundamentais das anomalias: São instâncias raras (minoria) e possuem valores distintos das instâncias normais (Liu; Ting; Zhou, 2008). O método constrói um conjunto de árvores de isolamento para um conjunto de dados, identificando as discrepâncias com base no comprimento médio do caminho percorrido até o isolamento do elemento. Nesse contexto, quanto menor o caminho, mais fácil é isolar a instância, indicando um valor aleatório (*outlier)*.

Na Figura 2.8, observa-se o processo de partição utilizado para isolar pontos específicos. Por exemplo, o ponto (a) exigiu seis partições para ser isolado, enquanto o ponto (b) necessitou de apenas uma. Dessa forma, percebe-se que (b), por ser isolado rapidamente, é uma anomalia, enquanto (a), com maior dificuldade de isolamento, não apresenta características de uma anomalia.

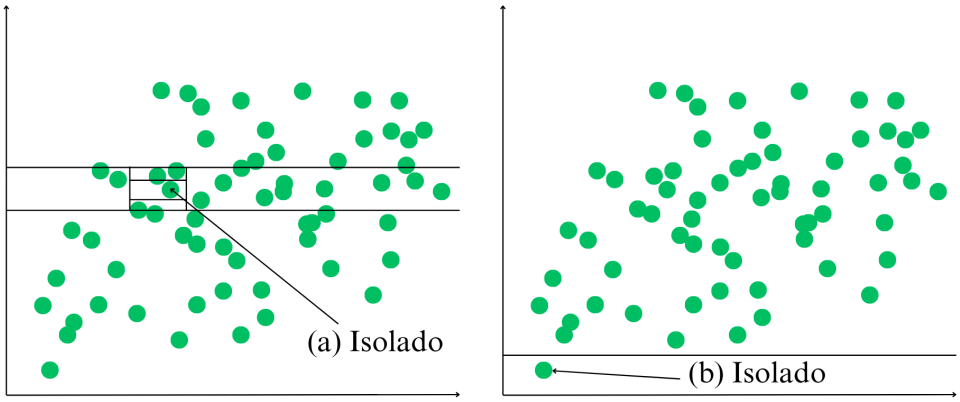

Figura 2.8: Processo de partição dos dados para encontrar discrepâncias.

O comprimento do caminho até o nó folha de uma instância é determinado pelo número de divisões necessárias para isolá-la. A partir desse valor, é possível calcular a pontuação de anomalia. Para isso, inicialmente, primeiro definimos alguns valores: $h(x)$ representa o número de divisões necessárias até que a instância alcance um nó folha. Esse valor depende da quantidade de dados disponíveis, representada por n. No entanto, como a profundidade da árvore cresce com o número de elementos, essa característica pode afetar a comparação entre instâncias (Liu;

Ting; Zhou, 2008). Assim, para contornar esse problema, utilizamos um valor normalizado com base na amostra n, expressa pela Equação 2.14:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n} \tag{2.14}$$

Onde:

- $H(n-1)$: representa o número harmônico estimado por $\ln(i) + 0,5772156649$;
- $c(n)$ corresponde ao comprimento médio normalizado esperado de $h(x)$ para uma instância ;
- n: representa a quantidade de dados.

Com base nesse valor normalizado, pode-se calcular a pontuação de anomalia, conforme apresentada pela Equação 2.15:

$$s(x,n) = 2\frac{E(h(x))}{c(n)} \tag{2.15}$$

Onde:

- $s(x,n)$: Pontuação de anomalia para uma instância x em um conjunto de dados com n instâncias;
- $E(h(x))$ corresponde à média de $h(x)$ da instância x ao passar pelas árvores de isolamento;
- $c(n)$ corresponde ao comprimento médio normalizado esperado de $h(x)$ para uma instância .

O valor de $s(x)$ indica o grau de anomalia da instância x, considerando a média do comprimento do caminho $E(h(x))$ percorrido nas árvores de isolamento (Liu; Ting; Zhou, 2008). Quando $E(h(x))$ é pequeno, $s(x)$ se aproxima de 1, o que indica um forte indício de anomalia. (Liu; Ting; Zhou, 2008). Por outro lado, se $E(h(x))$ é aproximadamente igual a $c(n)$, então $s(x) = 0.5$, o que significa um comportamento comum (Liu; Ting; Zhou, 2008). Por fim, quando $E(h(x))$ é grande, $s(x)$ se aproxima de 0, indicando que a instância é normal (Liu; Ting; Zhou, 2008).

2.2.3 Métricas de Desempenho

A qualidade ou precisão das predições de saída fotovoltaica pode ser avaliada por meio da diferença entre os valores reais e os valores preditos. Essa diferença é representada por métricas de erro, que ajudam a quantificar o desempenho dos modelos de predição, fornecendo uma medida clara sobre o quão próximos ou distantes as predições estão dos resultados observados

(Nguyen; Müsgens, 2022). Com isso, para a avaliação dos modelos de aprendizado é necessário a aplicação de métricas fundamentais, como EAM, REQM e R², uma vez que cada uma delas oferece uma perspectiva única sobre a eficácia dos modelos (Al-Dahidi et al., 2024).

2.2.3.1 Erro Absoluto Médio (EAM)

EAM fornece medida direta da magnitude média dos erros, calculando a média das diferenças absolutas entre os valores previstos e observados. Nesse contexto a métrica é essencial para avaliar a precisão geral dos modelos, independentemente da direção do erro (Abumohsen et al., 2024)(Al-Dahidi et al., 2024). Sua representação matemática é apresentada pela Equação 2.16:

$$EAM = \frac{1}{n}\sum_{i=1}^{n}(|y_i - \hat{y}|) \tag{2.16}$$

Onde:

- EAM: Erro Absoluto Médio;
- n: Número de observações;
- y_i: Valor real da observação;
- $\hat{y}_i$: Valor predito ou estimado.

2.2.3.2 Raiz do Erro Quadrático Médio (REQM)

REQM corresponde à raiz quadrada da média das diferenças quadráticas entre os valores preditos e os valores reais. Além disso ele representa o desvio padrão dos erros e é amplamente utilizado para identificar desvios em predições, sendo uma métrica útil para comparar o desempenho de modelos aplicados a diferentes conjuntos de dados (Abumohsen et al., 2024)(Al-Dahidi et al., 2024). Para calcular o REQM, utiliza-se a fórmula expressa pela Equação 2.17:

$$REQM = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{2.17}$$

Onde:

- REQM: Raiz do Erro Quadrático Médio
- n: número de observações
- y_i valor real da observação
- $\hat{y}_i$ valor predito ou estimado

2.2.3.3 Coeficiente de Determinação (R^2)

R^2 avalia a proporção da variância da variável dependente que é explicada pelas variáveis independentes utilizadas no modelo. Nesse sentido, essa métrica é particularmente importante para compreender o quão bem o modelo se ajusta aos dados e sua capacidade de capturar as relações existentes entre as variáveis (Abumohsen et al., 2024)(Al-Dahidi et al., 2024). Sua expressão matemática é expressa pela Equação 2.18:

$$R^2 = 1 - \frac{\sum_{i=i}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.18}$$

Onde:

- n: número de observações;
- y_i: valor real da observação;
- $\hat{y}_i$: valor predito ou estimado;
- $\bar{y}_i$: média dos valores reais.

2.3 Trabalhos Relacionados

Ledmaoui et al. (2023) realizaram estudo comparativo para medir a precisão da predição de energia solar com base em dados coletados de uma usina que se encontra em Benguerir, Marrocos. Os autores utilizaram como variáveis a produção de energia, a irradiância e a temperatura ambiente. No entanto, o uso reduzido de parâmetros pode ser insuficiente para um treinamento robusto de modelos de aprendizado de máquina. Na avaliação proposta, foram testados seis algoritmos de aprendizado de máquina: Regressão por Vetores de Suporte (RVS), Redes Neurais Artificiais (RNA), Árvore de Decisão, Floresta Aleatória, Modelo Aditivo Generalizado (MAG), e *Extreme Gradient Boosting* (XGBOOST). O desempenho dos modelos foi aferido pelas métricas: Erro Quadrático Médio (EQM), EAM, Erro Absoluto Médio Escalonado (EAME) e R^2. Os autores concluíram que as Redes Neurais obtiveram melhor desempenho entre os modelos testados, demonstrando alta eficiência e precisão na predição de demandas energéticas. Contudo, a remoção e tratamento de *outliers* não são devidamente apresentados e também inexiste o uso de abordagens híbridas.

Em Al-Dahidi et al. (2024) é executado uma avaliação entre modelos de aprendizado de máquina e seu impacto na predição de energia fotovoltaica. Os algoritmos testados foram: Modelo Linear Robusto (MLR), Árvore de Decisão, Floresta Aleatória, RVS e Perceptron Multicamada. Os autores utilizam quatro variáveis climáticas de análise: velocidade do vento, umidade relativa, temperatura ambiente e irradiação solar. Além disso, os autores implementaram o algoritmo de otimização de chimpanzés para seleção de hiperparâmetros. Para fins de avaliação foram utilizadas as métricas EQM, EAM e R^2. Diferente de Ledmaoui et al. (2023), Al-Dahidi

et al. (2024) apresentaram análise exploratória completa, aplicando técnicas de normalização e utilizando de variáveis climáticas relevantes. Em relação aos resultados apontados evidenciou-se que a radiação solar foi a variável mais influente na geração de energia. Em termos de desempenho, o Perceptron Multicamada se destacou com o melhor resultado. Entretanto, cabe ressaltar algumas limitações: (1) não foram exploradas abordagens híbridas, (2) não há uso de técnicas de remoção de anomalias, e (3) os hiperparâmetros utilizados no modelo MLR não foram disponibilizados.

No estudo de Abumohsen et al. (2024), o objetivo foi desenvolver modelos com alta precisão para predizer a geração de energia solar. Para isso, foram aplicadas algumas técnicas, incluindo aprendizado de máquina, aprendizado profundo e modelos híbridos. Entre os modelos utilizados estão: *Bi-directional LSTM* (BI-LSTM), *Gated Recurrent Units* (GRU), Redes Neurais Recorrentes (RNR), Floresta Aleatória, Máquina de Vetores de Suporte (MVS), BI-LSTM e Rede Neural Convolucional (RNC). O modelo híbrido RNC-BI-LSTM-FA, apresentou os melhores resultados em termos de precisão. Os dados utilizados no treinamento foram coletados entre 03 de junho de 2022 e 31 de julho de 2023, fornecidos pela *Tubas Electricity Company*, localizada na Palestina. As variáveis consideradas incluíram: potência de saída, radiação solar, temperatura, umidade, velocidade do vento e pressão atmosférica. Para avaliação dos modelos, foram adotadas as métricas: EQM, EAM e R^2. Na análise comparativa, os modelos de aprendizado de máquina demonstraram que a Floresta Aleatória superou a MVS em termos de precisão. Entre os modelos de aprendizado profundo, como *Long Short-Term Memory* (LSTM), BI-LSTM, RNR e GRU, o BI-LSTM se destacou com melhor desempenho. Em relação aos modelos híbridos, foi analisada a combinação de LSTM-FA, bem como um segundo modelo composto por RNC-LSTM-FA. Ao comparar modelos individuais e híbridos, é evidente que o modelo RNC-LSTM-FA apresentou melhores resultados, confirmando a hipótese de que abordagens híbridas são mais eficazes do que modelos isolados (Voyant et al., 2017). Apesar dos resultados obtidos, alguns pontos devem ser destacados: (1) não foram disponibilizados os hiperparâmetros utilizados, o que limita a replicabilidade dos experimentos, e (2) não há qualquer menção à aplicação de técnicas de controle e remoção de anomalias.

No estudo de Amiri et al. (2024), foram implementados diversos modelos de aprendizado de máquina, incluindo FA, RVS, PM, Regressão Linear (RL), Aprimoramento por Gradiente (AG), K-Vizinhos Mais Próximos (KNN), Regressão *Ridge* (RR), Lasso Regressor (LASSO), Regressão Polinomial (RP) e *Extreme Gradient Boosting* (XGBoost). Observa-se, que não foi explorada nenhuma modelagem híbrida, e os hiperparâmetros utilizados não foram detalhados pelos autores. Embora a aplicação isolada da Floresta Aleatória tenha apresentado desempenho consistente, quando adotamos abordagens híbridas – em nossas próprias experimentações, verificou-se uma tendência de redução tanto no EAM quanto no REQM. Considerando as diferenças entre as bases de dados utilizadas, o modelo desenvolvido neste trabalho obteve valores de EAM e REQM superiores aos registrados para o modelo isolado de Amiri et al. (2024). Nesse sentido, a adoção de uma abordagem híbrida, caso fosse implementada por Amiri

et al. (2024), provavelmente resultaria em melhorias no desempenho. Como aspecto positivo do trabalho de Amiri et al., destaca-se a implementação de solução para detecção e tratamento de anomalias. A Floresta Aleatória se destacou como o modelo com melhor desempenho geral entre os avaliados.

Tabela 2.1: Resultados de Amiri et al. (2024)

Métrica	**RP**	**FA**	**RVS**	**MLP**	**AG**	**LR**	**KNN**	**RR**	**LASSO**	**XGBoost**
RMSE	26.57	21.02	27.12	25.46	23.15	27.96	25.25	27.96	28.01	24.06
MAE	9.79	7.40	7.63	9.24	7.94	10.50	7.79	10.50	10.49	7.63
R^2	0.93	0.96	0.93	0.94	0.95	0.92	0.94	0.92	0.92	0.94

Por fim, observa-se que nenhum dos trabalhos utilizou ou disponibilizou o conjunto de dados em suas avaliações, o que representa uma limitação significativa, uma vez que o compartilhamento dessa informação contribui com a possibilidade de novas interpretações interdisciplinares, a preservação da integridade dos dados a longo prazo, otimização de recursos e a transparência científica (Tenopir et al., 2011).

3
Metodologia

Este capítulo descreve a metodologia adotada para o desenvolvimento e avaliação dos modelos preditivos de geração de energia solar fotovoltaica. São detalhados os materiais utilizados, incluindo a base de dados pública e as ferramentas computacionais, bem como os métodos empregados nas etapas de processamento de dados, tratamento de anomalias, configuração dos modelos e avaliação de desempenho.

3.1 Descrição

Conforme ilustra a Figura 3.1, foram desenvolvidos sete modelos para analisar o desempenho das técnicas de aprendizado de máquina empregadas. Nos Modelos 1, 2 e 3, cada técnica, Árvore de Decisão (AD), Floresta Aleatória (FA) e Perceptron Multicamada (PM), foi aplicada de forma independente, possibilitando avaliar o potencial preditivo individual de cada abordagem. Já os Modelos 4, 5 e 6 investigaram combinações híbridas, correspondentes às configurações AD e FA, AD e PM, e FA e PM, com o objetivo de verificar efeitos sinérgicos entre as técnicas. Por fim, o Modelo 7 reuniu simultaneamente as três abordagens (AD, FA e PM), configurando o cenário mais abrangente e integrador da metodologia proposta.

3.2 Processamento de Dados e Tratamento de Anomalias

Para o treinamento e validação dos modelos foi utilizada uma base de dados pública disponível no site AI on Demand[1]. Os registros foram coletados em intervalos horários, abrangendo três localidades distintas, no período de 22 de novembro de 2022 a 2 de novembro de 2023. A base de dados contém, variáveis relacionadas às condições meteorológicas e aos níveis de energia elétrica gerada, conforme descrito a seguir:

- Temperatura do ar em graus Celsius (°C);
- Quantidade de cobertura de nuvens no céu, expressa em percentual (%);

[1] AI on Demand

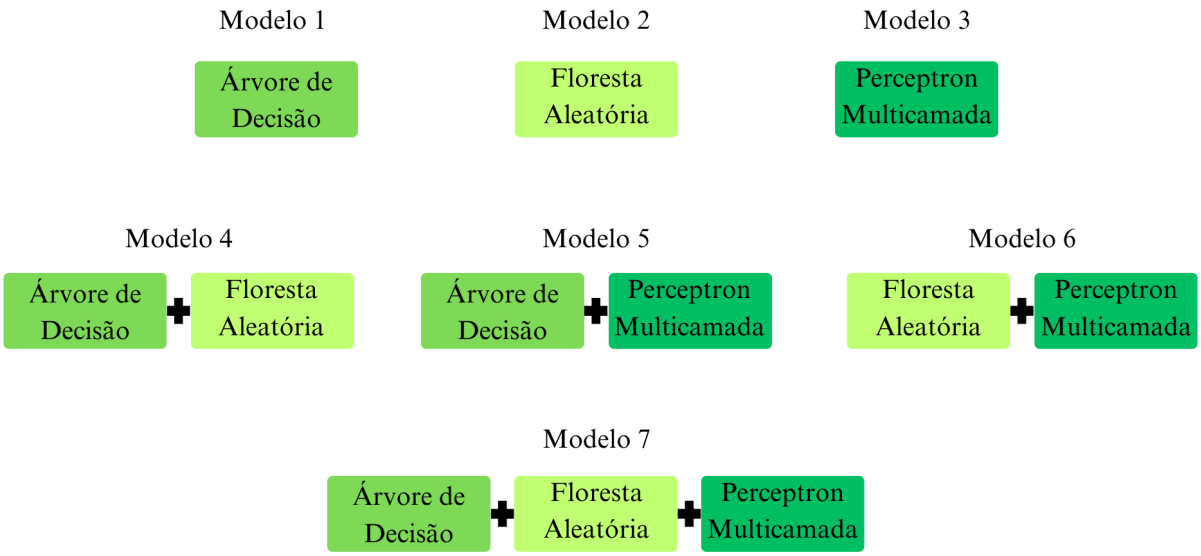

Figura 3.1: Esquema dos modelos de aprendizado de máquina.

- Irradiância Difusa Horizontal: Radiação solar recebida em superfície horizontal proveniente de todo o céu, incluindo luz direta e difusa. Medida em quilowatts por metro quadrado (kW/m²);
- Irradiância Direta Normal: Representa a radiação solar recebida diretamente do sol em superfície perpendicular aos raios solares. Medida em quilowatts por metro quadrado (kW/m²);
- Irradiância Extraterrestre Horizontal: Quantidade de radiação solar que seria recebida na superfície da Terra se não houvesse atmosfera. Medida em quilowatts por metro quadrado (kW/m²);
- Irradiância Global Horizontal: Radiação solar total recebida em uma superfície horizontal, incluindo componentes diretos e difusos. É medida em quilowatts por metro quadrado (kW/m²);
- Produção – Local 1/Local 2/Local 3: Representa a produção de eletricidade dos painéis solares no Local 1, Local 2 e Local 3, em quilowatt-horas (kWh);

O processamento de dados é etapa fundamental para garantir a qualidade do conjunto de treinamento. Como discutido no Capítulo 2, essas informações devem estar em condições ideais para evitar perdas na precisão dos resultados de predição. Técnicas de pré-processamento, como a remoção de valores ausentes por meio da função *Dropna* da biblioteca Pandas, foram aplicadas à base de dados. Posteriormente, os dados foram filtrados de modo a incluir apenas os registros correspondentes ao período de incidência significativa de radiação solar, entre 06:00 e 18:00. Em seguida, utilizando a função *Train Test Split* da biblioteca *Scikit-learn*, o conjunto de dados foi dividido em 75% para treinamento e 25% para teste. Além disso, todas variáveis

foram avaliadas tanto individualmente quanto em conjunto, a fim de verificar o desempenho do modelo em diferentes combinações. Para auxiliar compreensão de relações entre variáveis e embasar a escolha das características para os modelos, foi realizada análise que incluiu a geração de um mapa de calor de correlação entre todas variáveis numéricas do conjunto de dados a fim de auxiliar nos testes das combinações.

Para o tratamento de anomalias, utilizou-se a técnica de Floresta de Isolamento. Inicialmente, foi aplicado o algoritmo Árvore de Decisão aos dados brutos. Em seguida, o procedimento foi repetido, aplicando-se a Árvore de Decisão aos dados previamente tratados com Floresta de Isolamento, a fim de verificar o efeito positivo do tratamento. Por fim, o desempenho entre os dois cenários foi comparado, através das métricas de desempenho EAM, REQM e R^2.

Para a Floresta de Isolamento, a taxa de contaminação foi testada individualmente em cada modelo, com objetivo de identificar o valor ideal para a detecção de *outliers*. Este parâmetro determina a porcentagem de observações do conjunto de dados que serão consideradas potenciais anomalias. No Modelo 1, foi utilizada a taxa de contaminação de 0,10. No Modelo 2, a taxa foi de 0,04. Para o Modelo 3, foi aplicada a taxa de 0,05. No Modelo 4, novamente foi utilizada a taxa de 0,10. Já nos Modelos 5, 6 e 7, a configuração adotada foi de 0,075. Em todos os modelos, foram utilizadas 100 árvores, considerando todos os atributos em cada divisão e até 256 amostras por árvore.

3.3 Ajuste de Hiperparâmetros

Nos experimentos realizados, foram adotados os hiperparâmetros apresentados na Tabela 3.1.Esses hiperparâmetros foram definidos com objetivo de otimizar a performance de cada modelo na tarefa de predição.

Para a Árvore de Decisão, o critério de divisão utilizado foi o Erro Quadrático, o que significa que a escolha das melhores divisões nos nós da árvore foi feita minimizando a soma dos quadrados dos erros nos valores preditos. A profundidade máxima da árvore foi limitada a 6 níveis, evitando que o modelo se tornasse excessivamente complexo e se ajustasse aos dados de treinamento. Não foram impostas restrições para o número máximo de atributos ou folhas. O tamanho mínimo de amostras necessárias para formar uma folha foi definido como 1, enquanto o mínimo de amostras exigidas para realizar uma divisão foi 2.

No caso da Floresta Aleatória, foi adotada estratégia de amostragem com reposição, ou seja, cada árvore da floresta foi treinada em uma amostra aleatória dos dados com repetições. O critério de divisão utilizado dentro das árvores foi o Erro Absoluto, que orienta a formação dos nós minimizando a soma das distâncias absolutas entre as previsões e os valores reais. A profundidade máxima das árvores foi limitada a 7 e o número máximo de atributos considerados por divisão foi definido como 1.0, indicando que todas as variáveis disponíveis puderam ser consideradas. Assim como na Árvore de Decisão, não houve limitação no número de folhas de amostras. O número mínimo de amostras por folha foi 1, com 2 como mínimo para permitir uma

Tabela 3.1: Ajuste de hiperparâmetros dos modelos, detalhando valores definidos para cada algoritmo.

Modelo	**Hiperparâmetro**	**Valor**
Árvore de Decisão	Critério de Divisão	Erro Quadrático
	Profundidade Máxima	6
	Número Máximo de Atributos	Nenhum
	Número Máximo de Folhas	Nenhum
	Número Mínimo de Amostras por Folha	1
	Número Mínimo de Amostras para Divisão	2
	Método de Divisão	Melhor
Floresta Aleatória	Amostragem com Reposição	Verdadeiro
	Critério de Divisão	Erro Absoluto
	Profundidade Máxima	7
	Número Máximo de Atributos	1.0
	Número Máximo de Folhas	Nenhum
	Número Máximo de Amostras	Nenhum
	Número Mínimo de Amostras por Folha	1
	Número Mínimo de Amostras para Divisão	2
	Número de Estimadores	100
Rede Neural	Função de Ativação	ReLU
	Regularização alpha	0,0001
	Tamanho de Lote	Auto
	Tamanhos das Camadas Ocultas	(100, 50)
	Taxa de Aprendizado	Constante
	Taxa Inicial de Aprendizagem	0,001
	Máximo de Funções Internas	15000
	Número Máximo de Iterações	2000
	Número de Iterações sem Mudança	10
	Solver	Adam
	Tolerância	0,0001
	Fração de Validação	0,1

divisão. A floresta consistiu 100 estimadores, ou seja, 100 árvores de decisão independentes.

Para a Rede Neural, adotou-se a função de ativação ReLU, amplamente utilizada por permitir convergência eficiente em redes profundas. A regularização foi controlada pela variável alpha com valor 0,0001, ajudando a evitar o *overfitting*. O tamanho do lote de treinamento foi definido automaticamente, enquanto a arquitetura da rede incluiu duas camadas ocultas com 100 e 50 neurônios, respectivamente. A taxa de aprendizado foi mantida constante, com um valor inicial de 0,001, e o solver escolhido para otimização foi o Adam, método eficiente para grandes espaços de parâmetros. O algoritmo foi configurado para realizar até 2000 iterações, com máximo de 15000 funções internas. O critério de tolerância para parada foi 0,0001, e o treinamento pararia se não houvesse melhoria após 10 iterações consecutivas.

3.4 Ferramentas Utilizadas

Para o desenvolvimento e a implementação dos experimentos, são utilizadas as principais tecnologias descritas a seguir:

- **Ambiente de desenvolvimento:** o ambiente de programação utilizado foi o *Google Colab*[2], plataforma colaborativa que oferece recursos para desenvolvimento de código Python, além de integração com diversos serviços de armazenamento e execução em nuvem.
- **Manipulação e análise de dados:** para tratamento e manipulação dos dados, foram utilizadas as bibliotecas *Pandas*[3] (versão 2.2.2) e *NumPy*[4] (versão 2.0.2), amplamente utilizadas para análise de dados estruturados e computação numérica.
- **Visualização:** para criação de gráficos e visualizações dos dados, foram empregadas as bibliotecas *Matplotlib*[5] (versão 3.10.0) e *Seaborn*[6] (versão 0.13.2), que permitem a construção de visualizações customizadas.
- **Aprendizado de máquina e pré-processamento:** o pré-processamento de dados e a aplicação dos algoritmos de aprendizado de máquina foram realizados utilizando a biblioteca *Scikit-learn*[7] (versão 1.6.1), que oferece gama de ferramentas para tarefas de modelagem preditiva e avaliação de desempenho.
- **Versionamento de código:** para controle de versão do código-fonte e colaboração no desenvolvimento, foi utilizado o *Git*[8], sistema distribuído para gerenciamento de versões.
- **Armazenamento e disponibilização:** o código final e os recursos utilizados foram armazenados e disponibilizados através da plataforma *Github*[9].

[2]https://colab.google/
[3]https://pandas.pydata.org
[4]https://numpy.org
[5]https://matplotlib.org
[6]https://seaborn.pydata.org
[7]https://scikit-learn.org/stable/
[8]https://git-scm.com
[9]https://github.com/Ruanrochafeitosa/paper-RJM-2026

4

Resultados e Discussão

Este capítulo apresenta os resultado da aplicação e avaliação dos diferentes modelos de aprendizado de máquina definidos no Capítulo 3.

4.1 Correlação de Variáveis

O mapa de correlação entre variáveis numéricas exibido na Figura 4.1 evidencia relações de intensidade distintas. Em particular, observa-se que variáveis associadas à radiação solar, como Irradiância Direta Normal, Irradiância Horizontal Extraterrestre e Irradiância Global Horizontal, apresentam correlações muito altas (valores próximos a 0,9 ou superiores). Este padrão sugere que essas variáveis capturam conceitos físicos sobre radiação solar que coexistem e variam em conjunto no mesmo ambiente.

No que se refere à relação entre variáveis climáticas e produção, há correlações positivas moderadas a altas entre as variáveis de radiância (DNI, EBH, GHI) e as variáveis de produção. Este fato indica que condições de maior radiação solar tendem a associar-se a níveis maiores de produção, o que é coerente com a hipótese física de que a geração depende da intensidade da radiação disponível. Em contrapartida, a variável *CloudOpacity*, relacionada a opacidade de nuvens, apresenta correlações negativas com a produção, o que sugere que maior cobertura ou densidade de nuvens pode atenuar a produção elétrica, ainda que esse efeito não seja tão intenso quanto as influências diretas da radiação.

4.2 Tratamento de Anomalias

Conforme ilustra a Figura 4.2, os resultados da Árvore de Decisão indicam que o EAM apresentou uma queda expressiva em todos os três locais analisados após a aplicação da Floresta de Isolamento, com reduções de 15,30%, 7,41% e 8,83%, respectivamente. De modo semelhante, o REQM apresentou diminuições de 19,67% no Local 1, 5,72% no Local 2 e 9,07% no Local 3 após o tratamento das anomalias.

Ainda com base nos resultados exibidos na Figura 4.2, observou-se também avanço no R^2 de 6,76% no Local 1. Por outro lado, no Local 2 houve uma redução de 2,25%, enquanto

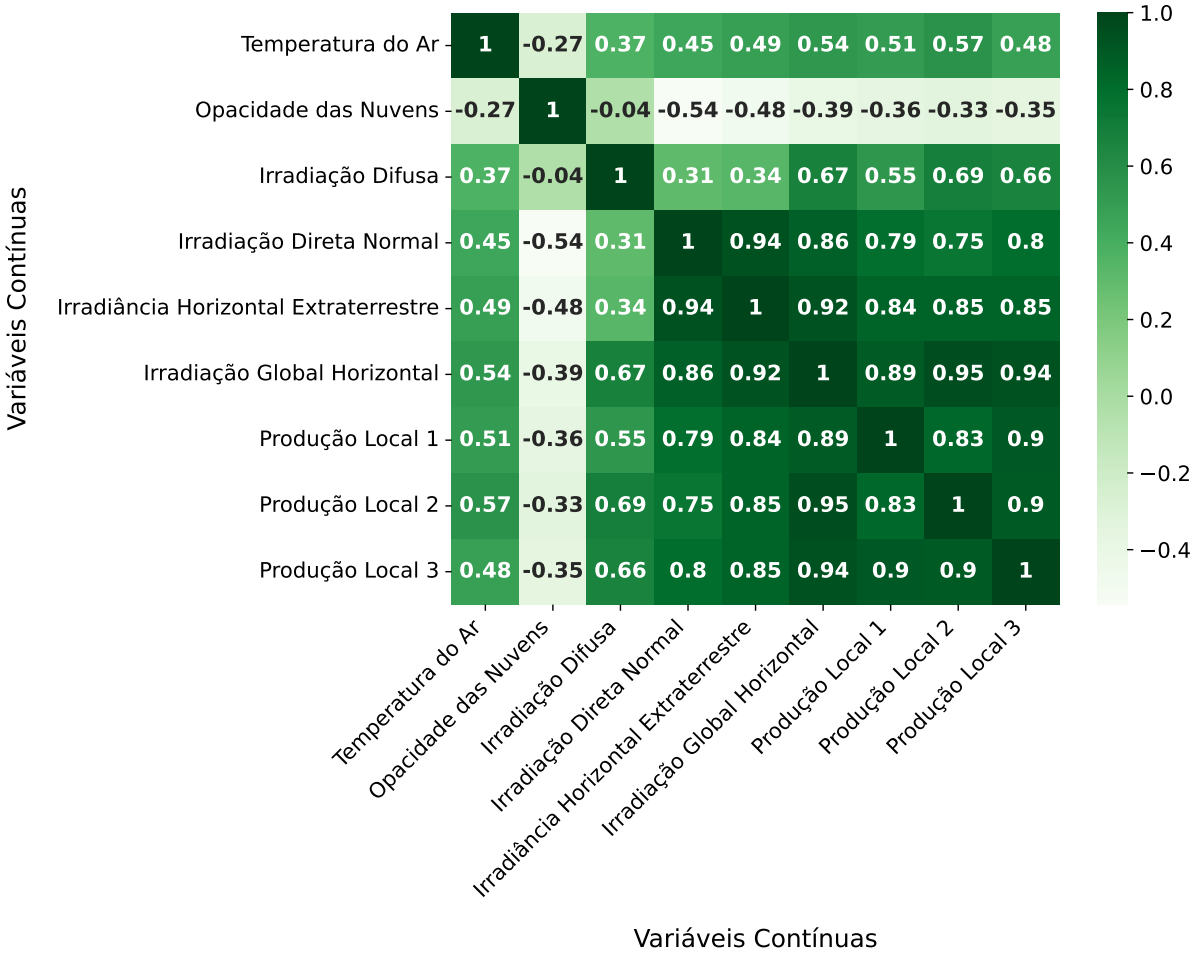

Figura 4.1: Correlação das variáveis numéricas da base de dados.

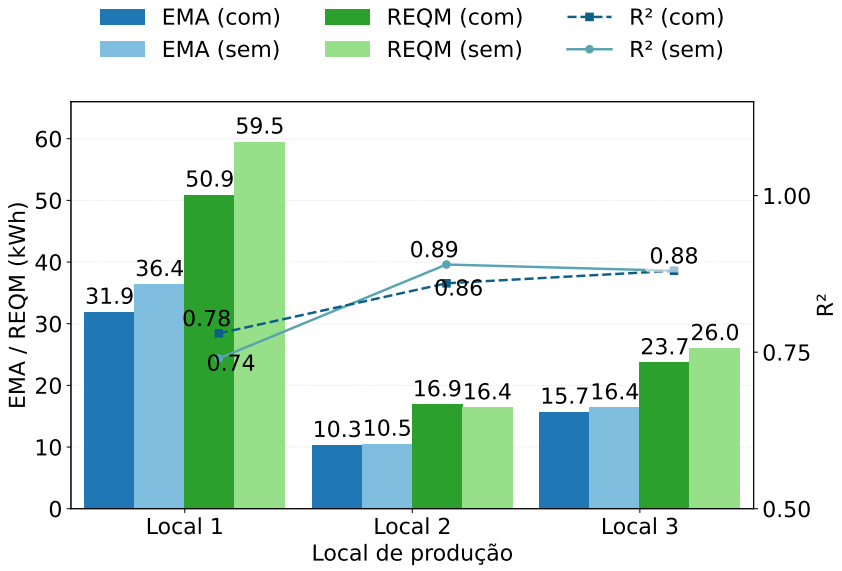

Figura 4.2: Comparação do desempenho da Árvore de Decisão com e sem remoção de anomalias.

no Local 3 não foram registradas mudanças. Esses resultados sugerem que os dados do Local 1 continham quantidade considerável de anomalias que comprometiam o ajuste do modelo

preditivo, enquanto, nos demais locais, a remoção de anomalias pode ter eliminado informações relevantes para as previsões.

No segundo experimento, foi aplicada a Floresta Aleatória com e sem remoção de anomalias. O comportamento observado foi similar ao teste anterior. O EAM diminuiu 12,9% no Local 1, 2,5% no Local 2 e 7,9% no Local 3, como evidencia a Figura 4.3. De forma complementar, o REQM foi reduzido em 15,3% no Local 1, 4,2% no Local 2 e 5,0% no Local 3. Quanto ao R^2, houve discreto decréscimo de 1,1% no Local 3, nenhuma variação no Local 2 e uma melhora de 3,8% no Local 1.

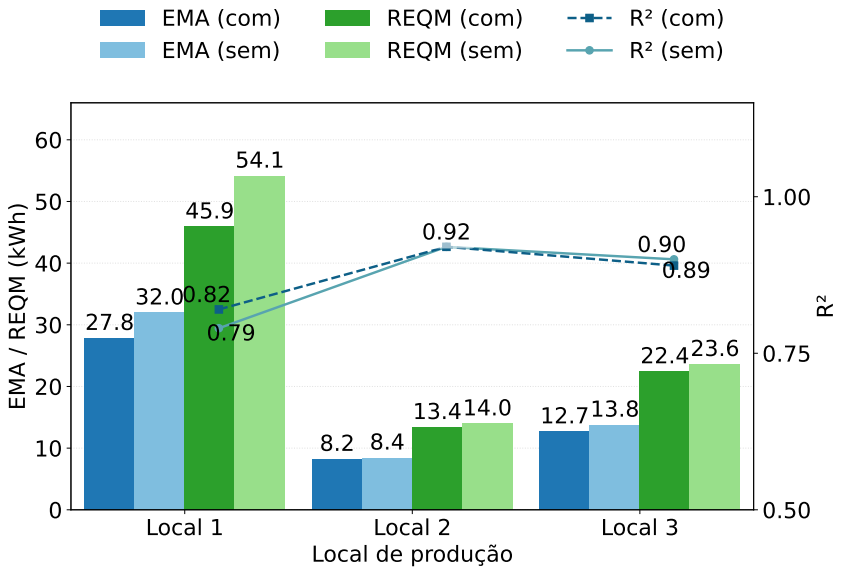

Figura 4.3: Comparação do desempenho da Floresta Aleatória com e sem remoção de anomalias.

No terceiro experimento, foi aplicado o Perceptron Multicamada (MLP) com e sem remoção de anomalias. Conforme mostra a Figura 4.4, o EAM apresentou redução em todos os locais: no Local 1, houve queda de 9,0%; no Local 3, de 4,3%; no Local 2 ocorreu um aumento de 2,1%. O REQM apresentou comportamento similar: no Local 1, reduziu-se 10,3%; no Local 3, 9,1%; enquanto no Local 2 observou-se diminuição de 3,4%.

Em relação ao R^2, conforme exposto na Figura 4.4, constatou-se aumento de 1,3% no Local 1, piora de 2,2% no Local 2 e nenhuma variação (0%) no Local 3. Esses resultados sugerem que, para o Perceptron Multicamada, em alguns casos a remoção de anomalias pode eliminar dados relevantes ao ajuste do modelo, especialmente no Local 2, prejudicando tanto o EAM e o REQM quanto o valor de R^2.

4.3 Modelos Individuais

A Tabela 4.1 sintetiza os resultados obtidos nos testes individuais por local de produção. O Local 2 apresenta desempenho superior, evidenciado pelos menores valores de EAM e REQM, bem como pelo valor mais elevado R^2, em comparação aos demais locais. No âmbito dos

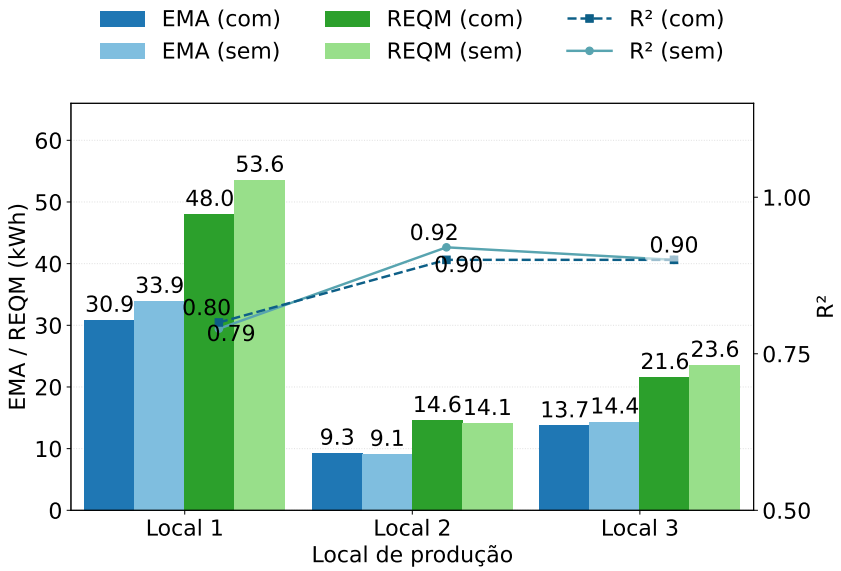

Figura 4.4: Comparação do desempenho do Perceptron Multicamada com e sem remoção de anomalias.

modelos isolados, a técnica FA demonstrou maior eficácia, superando AD e PM. É possível que, dado o caráter público da base de dados utilizada, a mesma contenha anomalias que AD não conseguiu detectar ou eliminar de forma adequada.

Tabela 4.1: Desempenho dos Modelos 1, 2 e 3 segundo EMA, REQM e R², por local de produção.

Modelo	EMA (kWh)			REQM (kWh)			R²		
	Local 1	Local 2	Local 3	Local 1	Local 2	Local 3	Local 1	Local 2	Local 3
Modelo 1	38.58	12.64	18.60	59.62	18.95	27.93	0.74	0.86	0.86
Modelo 2	31.96	8.43	13.77	54.15	13.97	23.60	0.79	0.92	0.90
Modelo 3	33.93	9.11	14.36	53.57	14.12	23.55	0.79	0.92	0.90

Especificamente, o Modelo 1 (Árvore de Decisão) apresentou EAM de 38,58 kWh (Local 1), 12,64 kWh (Local 2) e 18,60 kWh (Local 3); REQM de 59,62 kWh, 18,95 kWh e 27,93 kWh, respectivamente; e R² de 0,74, 0,86 e 0,86, nessa mesma ordem.

No Modelo 2 (Floresta Aleatória), observou-se redução do EAM para 31,96 kWh (queda de 17,1 % em relação ao Modelo 1) no Local 1, 8,43 kWh (–33,3%) no Local 2 e 13,77 kWh (–25,9%) no Local 3; o REQM diminuiu para 54,15 kWh (–9,2%), 13,97 kWh (–26,3%) e 23,60 kWh (–15,6%); enquanto o R² permaneceu em 0,79 no Local 1, mas subiu para 0,92 no Local 2 e 0,90 no Local 3, indicando maior consistência do modelo em ambos os locais.

Já o Modelo 3 (PM) registrou EAM de 33,93 kWh (–10,0%) no Local 1, 9,11 kWh (–27,9%) no Local 2 e 14,36 kWh (–22,8%) no Local 3; REQM de 53,57 kWh (–10,2%), 14,12 kWh (–25,5%) e 23,55 kWh (–15,8%); e R² de 0,79 (Local 1), 0,92 (Local 2) e 0,90 (Local 3), confirmando desempenho similar ao do Modelo 2.

A Figura 4.5, Figura 4.6 e Figura 4.7 ilustram comparativamente os valores reais e preditos por meio de gráficos de linha, com a linha sólida azul representando os valores reais e a linha amarela tracejada os valores estimados. A análise dos resultados, considerando o

Modelo 1, Modelo 2 e Modelo 3 respectivamente, permite avaliar a proximidade entre as curvas e, consequentemente, a capacidade preditiva de cada modelo ao replicar o comportamento dos dados originais.

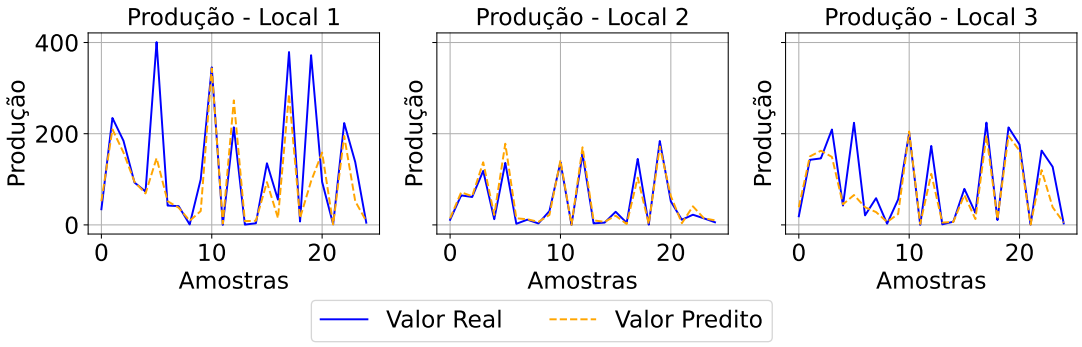

Figura 4.5: Comparativo entre valores observados e estimados pelo modelo Árvores de Decisão (AD)

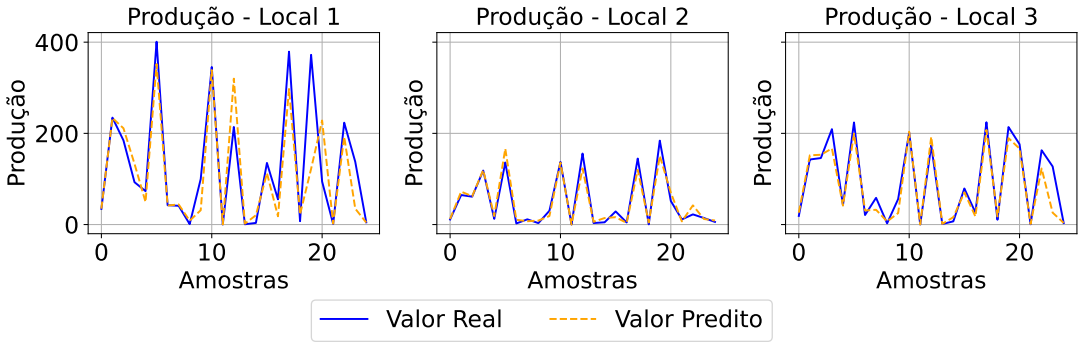

Figura 4.6: Comparativo entre valores reais e estimados pelo modelo Floresta Aleatória (FA)

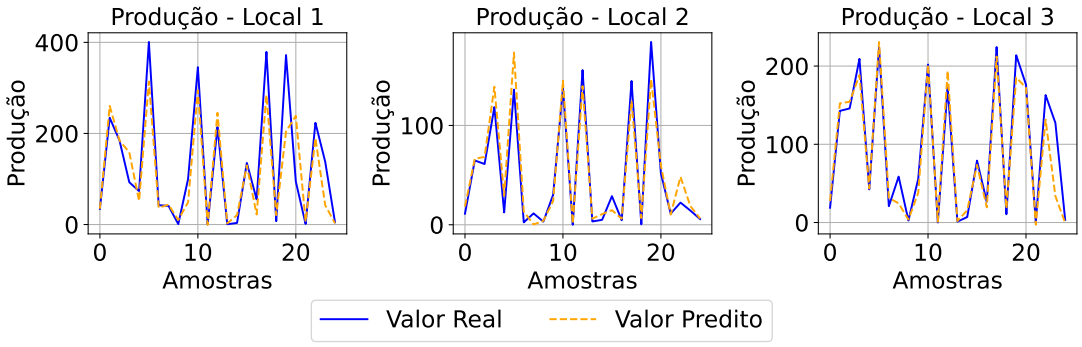

Figura 4.7: Comparativo entre valores reais e estimados pelo modelo Perceptron Multicamada (PM)

4.4 Modelos Híbridos Duplos

A Tabela 4.2 apresenta o desempenho aferido dos Modelos 4, 5 e 6 segundo EMA, REQM e R^2, por local de produção. Dentre os híbridos avaliados, o Modelo 6 apresentou o melhor desempenho em todos os locais de produção.

Tabela 4.2: Desempenho dos Modelos 4 a 6 segundo EMA, REQM e R², por local de produção.

Modelo	EMA (kWh)			REQM (kWh)			R²		
	Local 1	Local 2	Local 3	Local 1	Local 2	Local 3	Local 1	Local 2	Local 3
Modelo 4	29.25	7.93	13.61	46.75	13.61	21.93	0.80	0.90	0.89
Modelo 5	30.71	8.04	13.10	46.81	13.00	19.69	0.81	0.91	0.92
Modelo 6	28.85	7.26	11.46	46.01	12.15	18.40	0.81	0.92	0.93

Para o EAM, o Modelo 4 obteve valores de 29,25 kWh (Local 1), 7,93 kWh (Local 2) e 13,61 kWh (Local 3). O Modelo 5 apresentou um aumento de aproximadamente 4,9% no Local 1, 1,4% no Local 2, e redução de cerca de 3,8% no Local 3. A Figura 4.8 e a Figura 4.9 exibem o comportamento dos valores observados e estimados com a aplicação dos Modelos 4 e 5, respectivamente.

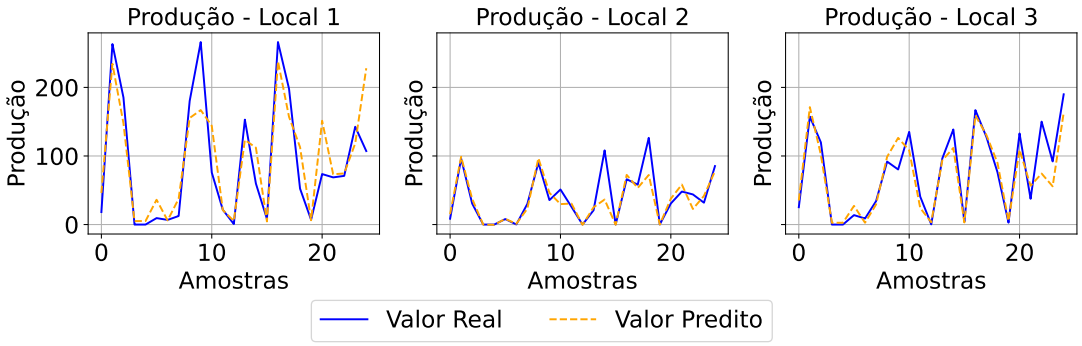

Figura 4.8: Comparativo entre os valores observados e os estimados pelo modelo híbrido - Árvores de Decisão + Floresta Aleatória.

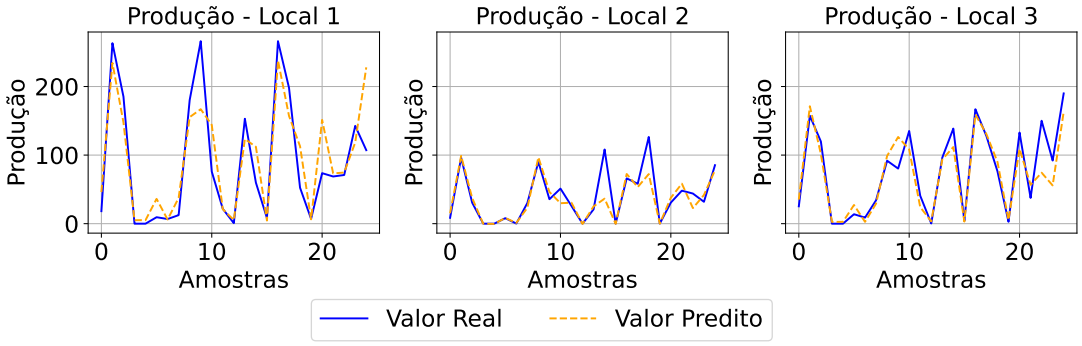

Figura 4.9: Comparativo entre os valores observados e os estimados pelo modelo híbrido - Árvores de Decisão + Perceptron Multicamada.

O Modelo 6 reduziu o EMA em relação ao Modelo 5: cerca de 6,1% no Local 1, 9,7% no Local 2 e 12,5% no Local 3. Em relação ao REQM, o Modelo 5 provocou pequenas variações em comparação ao Modelo 4 (aproximadamente +0,1% no Local 1; redução de cerca de 4,5% no Local 2; redução de aproximadamente 6,2% no Local 3). O Modelo 6 apresentou melhorias adicionais com reduções de REQM de cerca de 1,7% (Local 1), 6,5% (Local 2) e 6,6% (Local 3).

No que se refere ao coeficiente de determinação (R^2), houve evolução consistente dos modelos 4 para 5 e, em seguida, para o modelo 6: comparado ao Modelo 4, o Modelo 5 melhorou em cerca de 1,3% (Local 1), 1,1% (Local 2) e 3,4% (Local 3); o Modelo 6 manteve ou ultrapassou esses valores, com ganhos de 0,6%, 1,1% e 1,1%, respectivamente. Esses resultados confirmam que o Modelo 6 é superior aos outros modelos híbridos avaliados nos três locais de produção. A Figura 4.10 exibe o comparativo entre valores observados e estimados pelo modelo híbrido duplo composto pelos algoritmos Floresta Aleatória e Perceptron Multicamada.

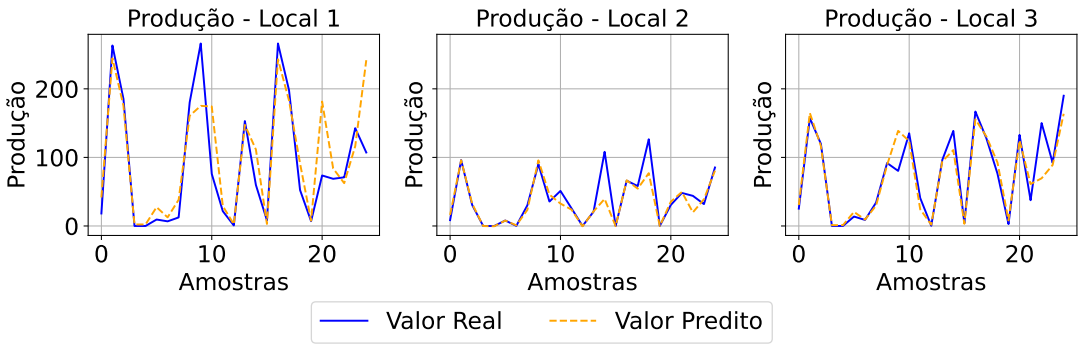

Figura 4.10: Comparativo entre os valores observados e os estimados pelo modelo híbrido - Floresta Aleatória (FA) + Perceptron Multicamada (PM)

4.5 Modelo Híbrido Triplo

Após avaliação dos modelos híbridos duplos, analisa-se o Modelo 7, que combina Árvore de Decisão, Floresta Aleatória e Perceptron Multicamada. No que se refere ao EAM, o Modelo 7 alcançou 28,94 kWh no Local 1 com aumento de 0,31% em relação ao Modelo 6, 7,36 kWh no Local 2 onde teve um aumento de 1,37% e 11,89 kWh no Local 3 com aumento de 3,75%. Já o REQM foi de 47,50 kWh no Local 1 com elevação de 3,23%, 12,49 kWh no Local 2 aumento de 2,79% e 19,24 kWh no Local 3 crescimento de 4,56%. Por fim, o R^2 registrou 0,80 no Local 1 com uma redução de 1,23% e 0,92 no Local 2 que se manteve estável, enquanto no Local 3 houve uma pequena redução de 1,07%, atingindo 0,92. A Tabela 4.3 exibe o desempenho do Modelo 7 segundo EMA, REQM e R^2, por local de produção, enquanto a Figura 4.11 exibe o comparativo entre valores observados e estimados pelo modelo híbrido triplo composto pelos algoritmos Árvore de Decisão, Floresta Aleatória e Perceptron Multicamada.

Como pode ser observado na Tabela 4.3, o EAM foi de 28,59 kWh no Local 1, representando redução de 0,9% em relação ao Modelo 6, 7,24 kWh no Local 2, com queda de 0,3%, e 11,69 kWh no Local 3, com aumento de 2%. O REQM atingiu 47,16 kWh no Local

Tabela 4.3: Desempenho do Modelo 7 segundo EMA, REQM e R^2, por local de produção.

Modelo	EMA (kWh)			REQM (kWh)			R^2		
	Local 1	Local 2	Local 3	Local 1	Local 2	Local 3	Local 1	Local 2	Local 3
Modelo 7	28.94	7.36	11.89	47.50	12.49	19.24	0.80	0.92	0.92

1, apresentando aumento de 2,5% em comparação ao Modelo 6, 12,36 kWh no Local 2, com incremento de 1,7%, e 18,94 kWh no Local 3, registrando aumento de 3,0%. Em relação ao R^2, observou-se 0,81 no Local 1, mantendo o mesmo valor do Modelo 6, 0,92 no Local 2, também estável, e 0,92 no Local 3, com leve redução de 1,1%. Esses resultados indicam que o Modelo 7 apresenta desempenho inferior em relação ao Modelo 6.

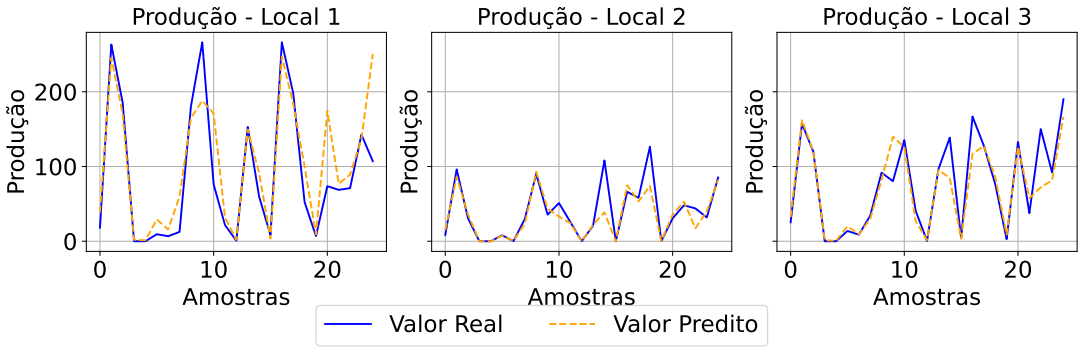

Figura 4.11: Comparativo entre os valores observados e os estimados pelo modelo híbrido - Árvores de Decisão + Floresta Aleatória + Perceptron Multicamada.

Os resultados indicam que o Modelo 7 apresentou desempenho inferior ao Modelo 6, com pequenas melhorias pontuais no EAM, mas aumento no REQM e perda de ajuste no R^2. Adicionalmente, nota-se diferença significativa entre os locais de produção. O Local 2 se mostrou como o de melhor desempenho em todos os modelos avaliados, alcançando maiores valores de R^2 e os menores erros, tanto no EAM quanto no REQM. Esse desempenho superior pode ser justificado devido as variações nas medições dos sensores decorrentes das características da localização.

Na comparação entre os modelos, observa-se que, quanto ao EAM, o Modelo 7 obteve os menores valores nos Locais 1 e 2, enquanto no Local 3 o melhor resultado foi obtido pelo Modelo 6. Já em relação ao REQM, o Modelo 6 apresentou os menores erros em todos os locais, consolidando-se como o mais consistente nesse critério. Considerando o R^2, o Modelo 6 também demonstrou desempenho superior frente aos demais.

A análise individual dos algoritmos indica que a Floresta Aleatória apresentou menor erro absoluto, mantendo valores de R^2 semelhantes aos outros modelos, o que a torna o algoritmo com melhor desempenho dentre as técnicas analisadas. Além disso, observa-se correlação entre o desempenho dos modelos individuais e seus equivalentes híbridos: algoritmos mais preditivos, quando combinados, tendem a gerar resultados superiores. Entretanto, nota-se que os modelos que incorporaram a Árvore de Decisão como preditor, em especial os Modelos 4, 5 e 7, apresentaram desempenho ligeiramente inferior, ainda que a diferença em relação aos demais modelos seja pequena.

5
Conclusão

Este trabalho investigou a aplicação de diferentes estratégias de aprendizado de máquina na predição de energia solar fotovoltaica, considerando modelos individuais, através dos algoritmos de Árvore de Decisão, Floresta Aleatória e Perceptron Multicamada, e abordagens híbridas resultantes das combinações em duplas e tripla. Os experimentos de análise preditiva dos modelos foram realizados com dados reais de três locais distintos de produção, avaliados por meio das métricas de EAM, REQM e R^2.

A remoção de anomalias possui potencial significativo de aprimorar o desempenho preditivo de modelos, reduzindo erros medidos por EAM e REQM. Todavia, esse ganho não é uniforme e depende do modelo e conjunto de dados. Enquanto modelos baseados em árvore responderam, de forma favorável à eliminação de observações atípicas, apresentando reduções expressivas nos erros e, em alguns casos, elevação do R^2, o algoritmo Perceptron Multicamada demonstrou maior sensibilidade à supressão de dados extremos, chegando a apresentar piora do R^2 ou aumento do erro em determinados locais. Esse comportamento evidencia que, embora a remoção de anomalias seja estratégia relevante, sua aplicação deve ser feita com cautela e avaliada localmente, especialmente em modelos mais complexos, para evitar exclusão de informação relevante ao ajuste e à generalização preditiva.

Observou-se que os modelos híbridos superaram os individuais, destacando-se o composto pela combinação entre Floresta Aleatória e Perceptron Multicamada, tornando-se a solução mais consistente e equilibrada para predição de energia solar fotovoltaica no contexto analisado. Esse modelo apresentou reduções significativas nos erros das predições e valores de R^2 superiores a 0,90 em todos os locais de produção. Embora o Modelo, que combinou simultaneamente as três técnicas, tenha mostrado bom desempenho em alguns cenários, ele não conseguiu superar o FA + PM de forma consistente, sobretudo em termos de REQM e estabilidade dos erros.

A análise individual reforçou a importância da estratégia de Floresta Aleatória, que se destacou pelo menor erro absoluto e pela manutenção de altos valores de R^2, confirmando-se como o algoritmo mais preditivo. Consequentemente, verificou-se que modelos mais precisos individualmente tendem potencializar positivamente os resultados quando integrados em arquiteturas híbridas, embora a inclusão da Árvore de Decisão em algumas combinações tenha reduzido

ligeiramente o desempenho.

Os resultados experimentais comprovam a relevância do uso de abordagens híbridas de aprendizado de máquina para a predição da geração solar fotovoltaica, promovendo confiabilidade no planejamento e operação de sistemas energéticos sustentáveis. Como direções para trabalhos futuros, recomenda-se a investigação de arquiteturas alternativas, tais como o uso de Redes Neurais Recorrentes (RNNs), e adoção de diferentes métodos de otimização e ajuste de hiperparâmetros. Além disso, para fortalecer a robustez e ampliar a capacidade de generalização, é desejável o uso desses modelos em bases de dados mais extensas e diversificadas.

Referências

Abumohsen, M. et al. Hybrid machine learning model combining of CNN-LSTM-RF for time series forecasting of Solar Power Generation. **e-Prime-Advances in Electrical Engineering, Electronics and Energy**, [S.l.], v.9, p.100636, 2024.

Al-Dahidi, S. et al. Enhancing solar photovoltaic energy production prediction using diverse machine learning models tuned with the chimp optimization algorithm. **Scientific Reports**, [S.l.], v.14, n.1, p.18583, August 10 2024.

Amiri, A. F. et al. Improving photovoltaic power prediction: insights through computational modeling and feature selection. **Energies**, [S.l.], v.17, n.13, p.3078, 2024.

Anderson, D.; Leach, M. Harvesting and redistributing renewable energy: on the role of gas and electricity grids to overcome intermittency through the generation and storage of hydrogen. **Energy policy**, [S.l.], v.32, n.14, p.1603–1614, 2004.

Bayod-Rújula, A. A. Solar photovoltaics (PV). In: **Solar hydrogen production**. [S.l.]: Elsevier, 2019. p.237–295.

Blázquez-García, A. et al. A review on outlier/anomaly detection in time series data. **ACM computing surveys (CSUR)**, [S.l.], v.54, n.3, p.1–33, 2021.

Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: a survey. **ACM computing surveys (CSUR)**, [S.l.], v.41, n.3, p.1–58, 2009.

Das, U. K. et al. Forecasting of photovoltaic power generation and model optimization: a review. **Renewable and Sustainable Energy Reviews**, [S.l.], v.81, p.912–928, 2018.

Dobos, A. **PVWatts version 5 manual-technical report NREL**. [S.l.]: TP-6A20, 2014.

Duan, L.; Da Xu, L. Business intelligence for enterprise systems: a survey. **IEEE Transactions on Industrial Informatics**, [S.l.], v.8, n.3, p.679–687, 2012.

Espinar, B. et al. Photovoltaic Forecasting: a state of the art. In: EUROPEAN PV-HYBRID AND MINI-GRID CONFERENCE, 5. **Anais...** [S.l.: s.n.], 2010. p.Pages–250.

Gaboitaolelwe, J. et al. Machine learning based solar photovoltaic power forecasting: a review and comparison. **IEEE Access**, [S.l.], v.11, p.40820–40845, 2023.

Gao, J.; Wang, H.; Shen, H. Smartly handling renewable energy instability in supporting a cloud datacenter. In: IEEE INTERNATIONAL PARALLEL AND DISTRIBUTED PROCESSING SYMPOSIUM (IPDPS), 2020. **Anais...** [S.l.: s.n.], 2020. p.769–778.

Gayen, D.; Chatterjee, R.; Roy, S. A review on environmental impacts of renewable energy for sustainable development. **International Journal of Environmental Science and Technology**, [S.l.], v.21, n.5, p.5285–5310, 2024.

Gupta, A. et al. Solar energy prediction using decision tree regressor. In: INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND CONTROL SYSTEMS (ICICCS), 2021. **Anais...** [S.l.: s.n.], 2021. p.489–495.

Impram, S.; Nese, S. V.; Oral, B. Challenges of renewable energy penetration on power system flexibility: a survey. **Energy Strategy Reviews**, [S.l.], v.31, p.100539, 2020.

Inman, R. H.; Pedro, H. T.; Coimbra, C. F. Solar forecasting methods for renewable energy integration. **Progress in energy and combustion science**, [S.l.], v.39, n.6, p.535–576, 2013.

Lara-Fanego, V. et al. Evaluation of the WRF model solar irradiance forecasts in Andalusia (southern Spain). **Solar Energy**, [S.l.], v.86, n.8, p.2200–2217, 2012.

Ledmaoui, Y. et al. Forecasting solar energy production: a comparative study of machine learning algorithms. **Energy Reports**, [S.l.], v.10, p.1004–1012, 2023.

Leva, S. et al. Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. **Mathematics and computers in simulation**, [S.l.], v.131, p.88–100, 2017.

Liu, F. T.; Ting, K. M.; Zhou, Z.-H. Isolation forest. In: 2008 EIGHTH IEEE INTERNATIONAL CONFERENCE ON DATA MINING. **Anais...** [S.l.: s.n.], 2008. p.413–422.

Lorenz, E. et al. Benchmarking of different approaches to forecast solar irradiance. In: EUROPEAN PHOTOVOLTAIC SOLAR ENERGY CONFERENCE, 24. **Anais...** [S.l.: s.n.], 2009. p.21–25.

Malvoni, M.; De Giorgi, M. G.; Congedo, P. M. Forecasting of PV Power Generation using weather input data-preprocessing techniques. **Energy Procedia**, [S.l.], v.126, p.651–658, 2017.

Moreno-Munoz, A. et al. Very short term of solar radiation. In: IEEE PHOTOVOLTAIC SPECIALISTS CONFERENCE, 2008. **Anais...** [S.l.: s.n.], 2008. p.1–5.

Nguyen, T. N.; Müsgens, F. What drives the accuracy of PV output forecasts? **Applied Energy**, [S.l.], v.323, p.119603, 2022.

Pierro, M. et al. Progress in regional PV power forecasting: a sensitivity analysis on the italian case study. **Renewable Energy**, [S.l.], v.189, p.983–996, 2022.

Ray, S. A Quick Review of Machine Learning Algorithms. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, BIG DATA, CLOUD AND PARALLEL COMPUTING (COMITCON), 2019. **Anais...** [S.l.: s.n.], 2019. p.35–39.

Singh, G. K. Solar power generation by PV (photovoltaic) technology: a review. **Energy**, [S.l.], v.53, p.1–13, 2013.

Suthaharan, S.; Suthaharan, S. Decision tree learning. **Machine learning models and algorithms for big data classification: thinking with examples for effective learning**, [S.l.], p.237–269, 2016.

Tenopir, C. et al. Data sharing by scientists: practices and perceptions. **PloS one**, [S.l.], v.6, n.6, p.e21101, 2011.

Victoria, M. et al. Solar photovoltaics is ready to power a sustainable future. **Joule**, [S.l.], v.5, n.5, p.1041–1056, 2021.

Voyant, C. et al. Machine learning methods for solar radiation forecasting: a review. **Renewable energy**, [S.l.], v.105, p.569–582, 2017.

Yang, D. et al. Operational solar forecasting for grid integration: standards, challenges, and outlook. **Solar Energy**, [S.l.], v.224, p.930–937, 2021.

Ying, X. An overview of overfitting and its solutions. In: JOURNAL OF PHYSICS: CONFERENCE SERIES. **Anais...** [S.l.: s.n.], 2019. v.1168, p.022022.

Zhang, C.; Lu, Y. Study on artificial intelligence: the state of the art and future prospects. **Journal of Industrial Information Integration**, [S.l.], v.23, p.100224, 2021.