Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Aplicadas
Departamento de Computação e Sistemas

Detecção de Anomalias em Usinas Solares utilizando métodos de Inteligência Artificial

Álvaro Braz Cunha

TRABALHO DE CONCLUSÃO DE CURSO

ORIENTAÇÃO:
Luiz Carlos Bambirra Torres

COORIENTAÇÃO:
Welbert Alves Rodrigues

Março, 2026
João Monlevade–MG

Álvaro Braz Cunha

Detecção de Anomalias em Usinas Solares utilizando métodos de Inteligência Artificial

Orientador: Luiz Carlos Bambirra Torres

Coorientador: Welbert Alves Rodrigues

Monografia apresentada ao curso de Engenharia de Computação do Instituto de Ciências Exatas e Aplicadas, da Universidade Federal de Ouro Preto, como requisito parcial para aprovação na Disciplina "Trabalho de Conclusão de Curso II".

Universidade Federal de Ouro Preto

João Monlevade

Março de 2026

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

C972d Cunha, Alvaro Braz.
Detecção de anomalias em usinas solares utilizando métodos de inteligência artificial. [manuscrito] / Alvaro Braz Cunha. - 2026.
51 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Luiz Carlos Bambirra Torres.
Coorientador: Prof. Dr. Welbert Alves Rodrigues.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Aplicadas. Graduação em Engenharia de Computação .

1. Aprendizado do computador. 2. Detecção de anomalias - Medidas de segurança. 3. Energia solar. 4. Geração de energia fotovoltaica. 5. Geração distribuida de energia elétrica. 6. Inteligência artificial. 7. Meteorologia. I. Torres, Luiz Carlos Bambirra. II. Rodrigues, Welbert Alves. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 621.311.243:004.85

Bibliotecário(a) Responsável: Flavia Reis - CRB6/2431

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE COMPUTACAO E SISTEMAS

FOLHA DE APROVAÇÃO

Álvaro Braz Cunha

Detecção de Anomalias em Usinas Solares utilizando métodos de Inteligência Artificial

Monografia apresentada ao Curso de Engenharia de Computação da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Engenharia de Computação

Aprovada em 06 de março de 2026

Membros da banca

Doutor Luiz Carlos Bambirra Torres - Orientador (Universidade Federal de Ouro Preto)
Doutor Welbert Alves Rodrigues - Coorientador (Universidade Federal de Ouro Preto)
Doutor Eduardo da Silva Ribeiro - (Universidade Federal de Ouro Preto)
Doutor Roberto Gomes Ribeiro - (Universidade Federal de Ouro Preto)

Luiz Carlos Bambirra Torres, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 11/03/2026

Documento assinado eletronicamente por **Luiz Carlos Bambirra Torres**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 11/03/2026, às 23:58, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **1073930** e o código CRC **373885F9**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0819 - www.ufop.br

Resumo

O monitoramento eficiente de usinas fotovoltaicas (UFV) é fundamental para garantir a rentabilidade e a segurança operacional, especialmente diante do crescimento exponencial da Geração Distribuída no Brasil. A inspeção manual de grandes volumes de dados de telemetria é impraticável em larga escala e suscetível a erros. Métodos automatizados baseados em regras estáticas frequentemente falham ao emitir falsos alertas durante quedas de geração causadas por sombreamento natural de nuvens. Este trabalho propõe e avalia um sistema multidimensional de detecção de anomalias baseado em técnicas de Aprendizado de Máquina Não Supervisionado. Para suprir a ausência de instrumentação meteorológica local, a telemetria elétrica dos inversores foi integrada a variáveis climáticas consumidas via API (*Open-Meteo*). Adicionalmente, desenvolveu-se uma engenharia de atributos focada na construção de uma "Curva Ideal Sazonal", transformando o problema temporal em uma análise de desvio relativo ao comportamento físico esperado em dias de céu claro. Foram implementados e comparados quatro algoritmos clássicos: *Isolation Forest*, *One-Class SVM*, *K-Means* e *K-Nearest Neighbors* (KNN), além de uma estratégia agnóstica de *Ensemble* por Votação Múltipla. O sistema foi validado com dados reais operacionais de quatro usinas no estado de Minas Gerais. Através de Validação Cruzada Temporal e testes estatísticos de McNemar e Wilcoxon, os resultados comprovaram a incidência do Teorema *No Free Lunch*: não existe um modelo universal. O agrupamento geométrico (*K-Means*) obteve desempenho superior ($p < 0.05$) em usinas de padrão de geração estável (UFV-A), enquanto o *Isolation Forest* foi estatisticamente o mais eficaz para detectar desvios em cenários de raras anomalias operacionais (UFV-B). O *Ensemble* por Votação provou ser a estratégia mais segura para monitoramento cego, mitigando drasticamente os falsos alarmes meteorológicos ao exigir consenso matemático, atuando como um filtro de ruído estocástico. Conclui-se que a hibridização de engenharia de domínio (Curva Ideal) com algoritmos multidimensionais capazes de "ler o clima"resulta em sistemas O&M robustos e com alta transparência de diagnóstico para os operadores.

Palavras-chave: Detecção de Anomalias. Energia Solar Fotovoltaica. Aprendizado de Máquina Não Supervisionado. *Ensemble Learning*. Meteorologia.

Abstract

Efficient monitoring of photovoltaic (PV) power plants is essential to ensure profitability and operational safety, particularly given the exponential growth of Distributed Generation in Brazil. Manual inspection of large volumes of telemetry data is impractical at scale and error-prone. Automated methods based on static rules frequently fail by issuing false alerts during generation drops caused by natural cloud shading. This work proposes and evaluates a multidimensional anomaly detection system based on Unsupervised Machine Learning techniques. To compensate for the lack of local meteorological instrumentation, the electrical telemetry of the inverters was integrated with climate variables consumed via API (*Open-Meteo*). Additionally, a feature engineering approach focused on building a "Seasonal Ideal Curve" was developed, transforming the temporal problem into a relative deviation analysis compared to the expected physical behavior on clear sky days. Four classic algorithms were implemented and compared: *Isolation Forest*, *One-Class SVM*, *K-Means*, and *K-Nearest Neighbors* (KNN), along with an agnostic Multiple Voting *Ensemble* strategy. The system was validated using real operational data from four PV plants in the state of Minas Gerais. Through Temporal Cross-Validation and McNemar and Wilcoxon statistical tests, the results confirmed the incidence of the *No Free Lunch* Theorem: there is no universal model. Geometric clustering (*K-Means*) achieved superior performance ($p < 0.05$) in plants with stable generation patterns (UFV-A), while the *Isolation Forest* was statistically the most effective in detecting deviations in scenarios with rare operational anomalies (UFV-B). The Voting *Ensemble* proved to be the safest strategy for blind monitoring, drastically mitigating meteorological false alarms by requiring mathematical consensus, thus acting as a robust stochastic noise filter. It is concluded that hybridizing domain engineering (Ideal Curve) with multidimensional algorithms capable of "reading the weather" results in robust O&M systems with high diagnostic transparency for operators.

Keywords: Anomaly Detection. Solar PV Energy. Unsupervised Machine Learning. Ensemble Learning. Meteorology.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

API *Application Programming Interface*

CNN *Convolutional Neural Network*

DT *Decision Tree*

GD Geração Distribuída

IA Inteligência Artificial

iForest *Isolation Forest*

KNN *K-Nearest Neighbors*

LSTM *Long Short-Term Memory*

ML *Machine Learning*

OCSVM *One-Class Support Vector Machine*

O&M Operação e Manutenção

PV *Photovoltaic*

RBF *Radial Basis Function*

RF *Random Forest*

ROI *Return on Investment*

SVM *Support Vector Machine*

UFV Usina Fotovoltaica

Lista de símbolos

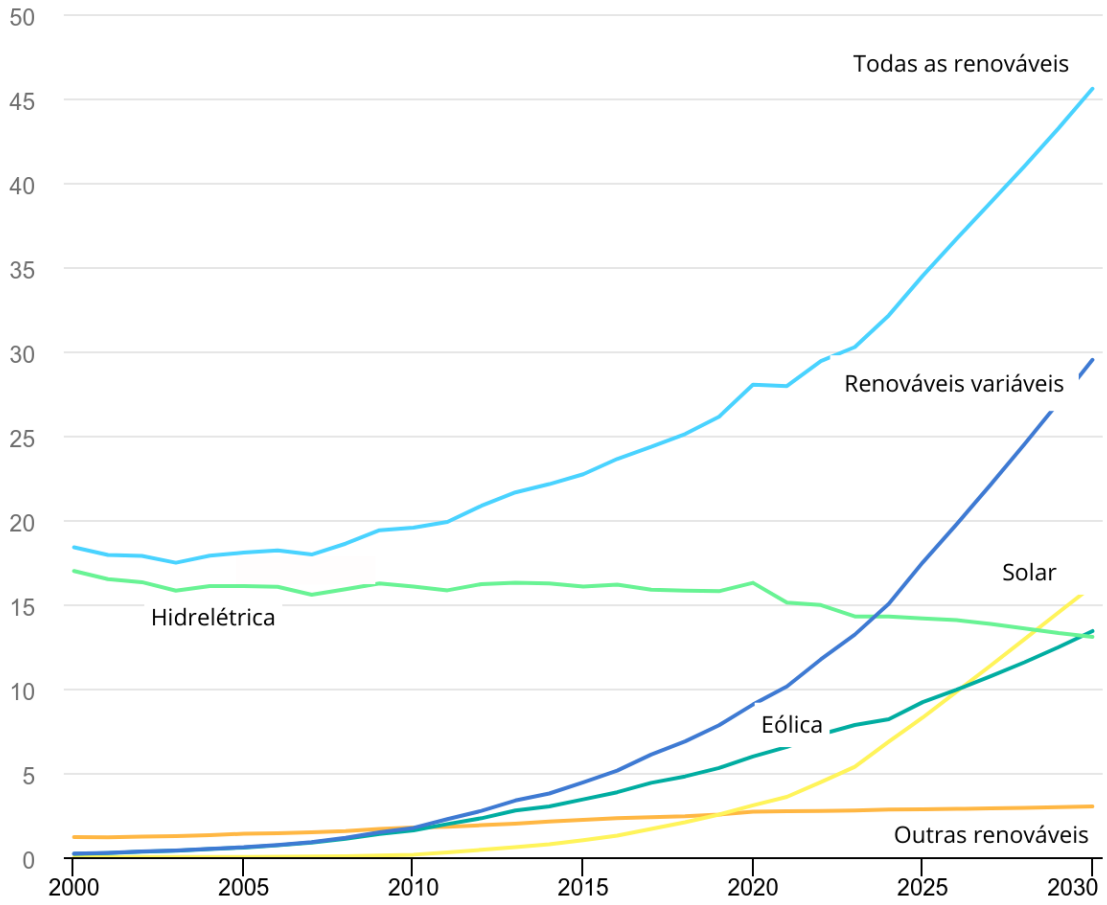
$D_{\%}$	Desvio Percentual Sazonal
$F1$	F1-Score (Média Harmônica entre Precisão e Revocação)
$F_{s,h}$	Função de distribuição empírica da potência
h	Instante de tempo (janela de hora e minuto)
K	Hiperparâmetro de quantidade (Número de Clusters ou Vizinhos)
p	Valor-p (*p-value*) de significância estatística
P_{95}	Percentil 95 da potência ativa
P_{ideal}	Potência Ativa Ideal Sazonal estimada
P_{real}	Potência Ativa medida na usina
$\mathcal{P}_{s,h}$	Conjunto de medições de potência ativa para a estação s e instante h
s	Estação do ano
w	Tamanho da janela de tempo da média móvel
x	Valor original da variável
z	Valor normalizado (Escore Z)
μ	Média aritmética da amostra
σ	Desvio padrão da amostra

Sumário

1 Introdução

A matriz energética global encontra-se em um ponto de inflexão histórico, impulsionada pela necessidade urgente de descarbonização e pela busca por segurança energética. Conforme aponta o relatório *Renewables 2024* da Agência Internacional de Energia (IEA, 2024), a energia solar *Photovoltaic* (PV) consolidou-se como a tecnologia de crescimento mais rápido, liderando a expansão da capacidade renovável mundial. A Figura 1 ilustra a projeção da participação das renováveis na geração de eletricidade, evidenciando o papel protagonista da energia solar e eólica rumo a 2030.

Figura 1 – Participação da geração de eletricidade renovável por tecnologia (2000-2030).



Fonte: Adaptado de IEA (2024) (IEA, 2024).

No cenário brasileiro, esse fenômeno é amplificado pela Geração Distribuída (GD). O barateamento dos módulos fotovoltaicos permitiu que residências e comércios se tornassem produtores de energia, resultando na instalação massiva de usinas de pequeno e médio porte.

Contudo, essa descentralização impõe desafios operacionais inéditos. Diferente de grandes parques geradores centralizados, que possuem equipes de Operação e Manutenção (O&M) dedicadas e instrumentação meteorológica avançada, as usinas de GD frequentemente operam "às escuras", sem monitoramento contínuo da sua eficiência térmica e de irradiação (SOUZA; FREITAS; SAMPAIO, 2023).

A falta de monitoramento adequado pode levar a perdas financeiras significativas. Sistemas fotovoltaicos estão sujeitos a diversas anomalias, como sombreamento parcial, acúmulo de sujidade (*soiling*), falhas em diodos de *bypass* (componentes que desviam a corrente elétrica de células sombreadas ou defeituosas para evitar o bloqueio da geração e o superaquecimento) e pontos quentes (*hotspots*). Segundo Jimenez, Cano e Velilla (2026), muitas dessas falhas são sutis e, se não detectadas precocemente, podem evoluir para riscos de segurança severos, comprometendo a vida útil dos equipamentos, projetada tipicamente para 25 a 30 anos (JORDAN; KURTZ, 2013).

Tradicionalmente, a detecção dessas falhas dependeria de inspeções manuais periódicas, processos logisticamente inviáveis e economicamente proibitivos para a escala da GD (SERAFIM, 2023). A alternativa escalável reside na análise dos dados de telemetria gerados pelos próprios inversores (potência, tensão, corrente). No entanto, a análise manual desses dados é impraticável devido ao volume massivo de informações geradas a cada minuto por milhares de usinas.

1.1 Problema de Pesquisa

Surge, então, a necessidade de automatizar a detecção de falhas através de algoritmos inteligentes. Embora a literatura recente apresente avanços significativos utilizando Aprendizado de Máquina Supervisionado (KHANDEPARKAR; SHRESHTHA; RAMU, 2025), essas abordagens esbarram em um obstáculo prático: a escassez de dados rotulados. Na realidade das empresas de monitoramento, raramente existe um histórico anotado indicando exata e precisamente quando ocorreu uma falha técnica no passado.

Além disso, o ambiente de operação solar é intrinsecamente ruidoso. Variações meteorológicas estocásticas, como a passagem rápida de nuvens, provocam flutuações bruscas na geração que podem ser facilmente confundidas com falhas de *hardware* por regras estáticas de monitoramento. Piliougine, Guejia-Burbano e Spagnuolo (2022) destacam que distinguir perdas de geração causadas por fatores ambientais daquelas causadas por defeitos elétricos é um dos maiores desafios para algoritmos automatizados.

Diante desse contexto, o problema central abordado neste trabalho é: **Como desenvolver um sistema automatizado capaz de detectar anomalias reais em usinas solares utilizando apenas dados operacionais não rotulados, sendo capaz de distinguir falhas técnicas de variações climáticas naturais?**

1.2 Objetivos

O objetivo geral deste trabalho é desenvolver e validar uma metodologia baseada em Aprendizado de Máquina Não Supervisionado para detecção de anomalias em sistemas fotovoltaicos, utilizando dados reais de telemetria integrados a bases meteorológicas externas.

Os objetivos específicos incluem:

- Desenvolver um *pipeline* de engenharia de dados que integre a telemetria elétrica de inversores com variáveis climáticas obtidas via *Application Programming Interface* (API) (*Open-Meteo*), suprindo a falta de estações meteorológicas locais na GD;
- Criar um atributo sintético (*Feature Engineering*) denominado Curva Ideal Sazonal para modelar a expectativa máxima de geração em dias de céu claro (*Clear Sky*);
- Implementar e otimizar quatro algoritmos de detecção multidimensionais: *Isolation Forest* (iForest), *One-Class Support Vector Machine* (OCSVM), *K-Means* e *K-Nearest Neighbors* (KNN);
- Propor e validar estatisticamente uma estratégia de *Ensemble Learning* por votação para mitigar a taxa de Falsos Positivos em cenários estocásticos;
- Avaliar a generalização dos modelos em quatro usinas reais no estado de Minas Gerais, validando a ocorrência do teorema *No Free Lunch* (WOLPERT; MACREADY, 1997).

1.3 Estrutura do Trabalho

O restante deste trabalho está organizado como se segue:

- O Capítulo 2 revisa a literatura pertinente, categorizando as principais abordagens existentes para detecção de anomalias fotovoltaicas e destacando a lacuna de pesquisa que este estudo visa preencher.
- O Capítulo 3 apresenta a fundamentação teórica, discutindo o funcionamento dos sistemas fotovoltaicos e os algoritmos de aprendizado de máquina abordados.
- O Capítulo 4 descreve a metodologia proposta, detalhando a integração da API climática, a construção da Curva Ideal Sazonal e o pré-processamento.
- O Capítulo 5 ilustra a arquitetura do *software*, as bibliotecas utilizadas e a parametrização dos hiperparâmetros.
- O Capítulo 6 discute os resultados experimentais através de métricas de classificação, validação cruzada temporal e testes de significância estatística.

- O Capítulo 7 sintetiza as descobertas, apresenta as limitações do estudo e aponta direções para trabalhos futuros.

2 Trabalhos Relacionados

A detecção de falhas em sistemas fotovoltaicos é uma área de pesquisa ativa, com abordagens que variam desde a inspeção visual manual até o uso de algoritmos complexos de *Deep Learning*. Este capítulo categoriza as principais estratégias encontradas na literatura recente, destacando suas vantagens e limitações em relação à abordagem proposta neste trabalho.

2.1 Abordagens Baseadas em Imagem

Uma vertente significativa da literatura foca no uso de imagens (térmicas, visuais ou de eletroluminescência) processadas por Redes Neurais Convolucionais (CNNs).

Serafim (2023) propôs o uso de CNNs para classificar defeitos em células fotovoltaicas através de imagens de eletroluminescência. Embora o autor tenha alcançado alta precisão na detecção de microfissuras, essa técnica exige equipamentos de aquisição caros e não permite o monitoramento em tempo real, sendo aplicada apenas em auditorias pontuais. Similarmente, Jimenez, Cano e Velilla (2026) desenvolveram um modelo leve (*lightweight*) para detectar falhas via imagens térmicas capturadas por drones. Apesar de reduzirem o custo computacional, a barreira logística do uso de drones permanece um impeditivo para o monitoramento contínuo de milhares de usinas de microgeração. Em uma abordagem semelhante, Alvarez e Pantoja (2021) aplicaram CNNs para o processamento de imagens de eletroluminescência capturadas por drones, obtendo alta precisão na detecção de microfissuras. Contudo, o custo de aquisição do *hardware* e a complexidade logística tornam essa solução proibitiva para o monitoramento contínuo em larga escala.

Diferente dessas abordagens, este trabalho foca exclusivamente em dados numéricos de telemetria, eliminando a necessidade de *hardware* adicional.

2.2 Abordagens Supervisionadas

No domínio da análise de dados numéricos, a maioria dos trabalhos utiliza Aprendizado Supervisionado.

Recentemente, Khandeparkar, Shreshtha e Ramu (2025) publicaram um estudo comparativo robusto na *Scientific Reports*, avaliando cinco algoritmos supervisionados (DT, Naive Bayes, RF, SVM e XGBoost) para a classificação de falhas elétricas específicas, como curto-circuitos e falhas de aterramento. O modelo XGBoost obteve a melhor performance, com acurácia de 98%. Entretanto, uma limitação crucial deste estudo é a dependência

de dados sintéticos gerados via simulação no MATLAB/Simulink. Embora validem a eficácia teórica dos classificadores, esses resultados assumem um cenário ideal de dados perfeitamente rotulados e livres de ruídos de comunicação, uma condição raramente encontrada em usinas de GD em operação real, como as abordadas nesta monografia.

Nassreddine et al. (2025) também aplicaram técnicas supervisionadas (*Random Forest* e KNN) para classificar sete tipos de falhas. Em uma abordagem intermediária, Zhao et al. (2015) propuseram um modelo baseado em grafos e aprendizado semi-supervisionado para detecção de falhas elétricas específicas. Embora eficaz, a abordagem ainda apresenta limitações práticas por depender de uma pequena quantidade de dados rotulados para o treinamento inicial, o que raramente está disponível em usinas legadas.

A principal limitação dessas obras para o contexto brasileiro de GD é, portanto, a exigência de *datasets* rotulados — total ou parcialmente. Na prática, usinas comerciais raramente possuem um histórico confiável de "quais falhas ocorreram e quando", inviabilizando o treinamento supervisionado.

2.3 Abordagens Não Supervisionadas e Dados Não Rotulados

Para contornar a escassez histórica de anotações, a literatura tem se voltado crescentemente para o Aprendizado de Máquina Não Supervisionado, capaz de extrair padrões diretamente de dados não rotulados.

Um desafio crítico em aplicações reais desse paradigma é justamente a validação dos resultados na ausência de rótulos confiáveis (*ground truth*). Idan (2024) discute que, sem um gabarito externo absoluto, a avaliação de modelos de detecção de anomalias torna-se altamente complexa. O autor propõe um paradigma inspirado em mecanismos de decisão colaborativa, argumentando que o consenso entre múltiplos modelos independentes pode servir como um indicativo robusto de performance quando não há dados rotulados disponíveis.

Essa perspectiva teórica fundamenta a abordagem adotada nesta monografia. A proposta deste trabalho preenche a lacuna deixada pelos métodos supervisionados ao adotar uma arquitetura que aprende exclusivamente com a geometria e a densidade dos dados não rotulados, utilizando a concordância entre algoritmos não apenas para classificação, mas como uma evidência matemática de confiabilidade.

2.4 Simplificação e *Ensemble Learning*

Recentemente, tem crescido o interesse em simplificar os modelos para torná-los viáveis em *hardware* de borda (*Edge Computing*). Nedaei, Eskandari e Aghaei (2025) argumentam que modelos clássicos de *Machine Learning* (ML), quando bem ajustados,

podem competir com redes profundas (*Deep Learning*) com uma fração do custo computacional. Os autores demonstraram que a redução da dimensionalidade e a escolha correta de *features* são mais críticas do que a complexidade do algoritmo em si.

Seguindo essa linha, Piliougine, Guejia-Burbano e Spagnuolo (2022) exploraram o uso de técnicas de aprendizado para distinguir sombreamento parcial de falhas de *mismatching*, destacando a importância de analisar o contexto operacional.

O presente trabalho alinha-se a essa tendência de "Inteligência Artificial (IA) eficiente"defendida por Nedaei, Eskandari e Aghaei (2025), utilizando algoritmos clássicos leves (*K-Means*, iForest, OCSVM e KNN) combinados em uma estratégia de *Ensemble Learning* (Votação) para garantir robustez, conforme sugerido teoricamente por Nassreddine et al. (2025), mas adaptado para um cenário sem supervisão.

Diante das limitações identificadas nas abordagens revisadas — dependência de *hardware* especializado, necessidade de dados rotulados ou de rótulos iniciais mínimos — este trabalho propõe uma arquitetura estritamente não supervisionada e de baixo custo computacional. Ao combinar uma engenharia de atributos robusta com algoritmos clássicos leves, busca-se preencher a lacuna por soluções capazes de operar apenas com dados de telemetria padrão, sem necessidade de sensores visuais ou anotação manual de falhas.

3 Fundamentação Teórica

Este capítulo estabelece a base conceitual necessária para o desenvolvimento do sistema proposto. Discute-se o contexto operacional da geração fotovoltaica, os desafios inerentes à natureza estocástica da fonte solar e os fundamentos teóricos da detecção de anomalias. Por fim, detalham-se os algoritmos de aprendizado não supervisionado selecionados e o estado da arte em monitoramento inteligente.

3.1 Monitoramento de Usinas Fotovoltaicas

A viabilidade econômica de projetos fotovoltaicos depende diretamente da eficiência operacional ao longo da vida útil do ativo (tipicamente 25 a 30 anos). No entanto, usinas solares operam em ambientes não controlados, sujeitas a degradação e falhas que impactam o *Return on Investment* (ROI).

Segundo Mousavi e Heydari (2020), as perdas de performance podem ser classificadas em duas categorias:

- **Fatores Ambientais:** Sombreamento parcial (nuvens, vegetação), acúmulo de sujidade (*soiling*) e altas temperaturas.
- **Falhas Técnicas:** Curto-circuitos em módulos, queima de diodos de *bypass*, falhas em inversores e desconexão de cabos.

O principal desafio no monitoramento reside na distinção entre esses eventos. Como a geração é altamente correlacionada com a irradiância solar, uma queda abrupta de potência pode indicar tanto uma nuvem passageira (comportamento normal) quanto uma falha crítica. Métodos tradicionais baseados em limiares fixos mostram-se ineficazes nesse cenário dinâmico, gerando altas taxas de falsos positivos.

3.2 Detecção de Anomalias

Embora a literatura clássica defina a detecção de anomalias como a identificação de padrões que não conformam com um comportamento esperado (CHANDOLA; BANERJEE; KUMAR, 2009), no contexto específico deste trabalho, o problema assume uma característica particular.

Nas Usina Fotovoltaicas (UFVs) analisadas, a detecção de anomalias não se refere apenas a encontrar *outliers* estatísticos, mas sim a identificar períodos em que a potência

ativa medida diverge significativamente da capacidade de geração estimada pela **Curva Ideal Sazonal**. Diferente de anomalias em dados financeiros ou de rede, aqui o "comportamento esperado"é ditado pela física do movimento solar e pelas condições meteorológicas locais.

3.3 Algoritmos de *Machine Learning* Não Supervisionado

Para compor o sistema de detecção, selecionaram-se quatro algoritmos representativos de diferentes paradigmas de aprendizado. A seguir, detalha-se o funcionamento mecânico de cada um com uma abordagem didática, facilitando a compreensão da lógica subjacente.

3.3.1 *Isolation Forest* (iForest)

O iForest (LIU; TING; ZHOU, 2008) parte de uma premissa contra-intuitiva: em vez de tentar modelar o perfil complexo do que é "normal", o algoritmo foca em isolar o que é "estranho".

Analogia Didática: Imagine que você precisa encontrar uma agulha em um palheiro usando cortes de espada. Para isolar um único pedaço de palha (dado normal, que está aglomerado), você precisaria de muitos cortes precisos. Para isolar a agulha (anomalia, que é diferente e está sozinha), poucos cortes aleatórios são suficientes para separá-la do resto.

Funcionamento Técnico: O algoritmo constrói múltiplas árvores de decisão aleatórias. Anomalias são isoladas rapidamente, ficando próximas à raiz da árvore (caminho curto). Dados normais exigem muitas ramificações para serem isolados (caminho longo). A Figura 2 ilustra esse conceito, onde o ponto anômalo em vermelho é isolado com apenas dois cortes, enquanto os pontos normais no centro exigiriam muito mais divisões.

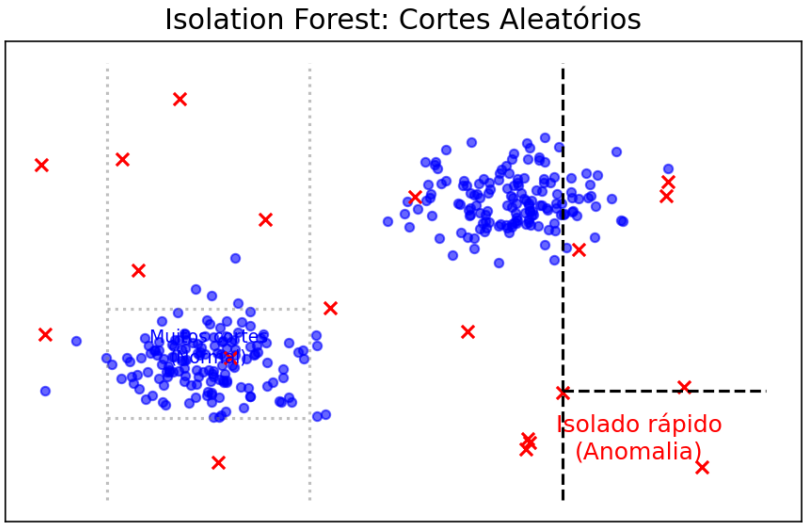
3.3.2 *One-Class SVM* (OCSVM)

O OCSVM (SCHÖLKOPF et al., 2001) é uma técnica que busca definir uma "fronteira de segurança"ao redor dos dados de operação normal.

Analogia Didática: Pense no OCSVM como um segurança de balada que só conhece as características das pessoas que *podem* entrar (dados normais de treino). Ele desenha uma corda de isolamento ao redor da fila. Qualquer pessoa nova que tente entrar e cujas características caiam fora dessa corda é barrada (considerada anomalia).

Funcionamento Técnico: O algoritmo mapeia os dados para um espaço de alta dimensão e tenta encontrar o menor hiperplano (fronteira) que englobe a maioria dos dados de treinamento. A Figura 3 mostra essa fronteira (linha vermelha) envolvendo os

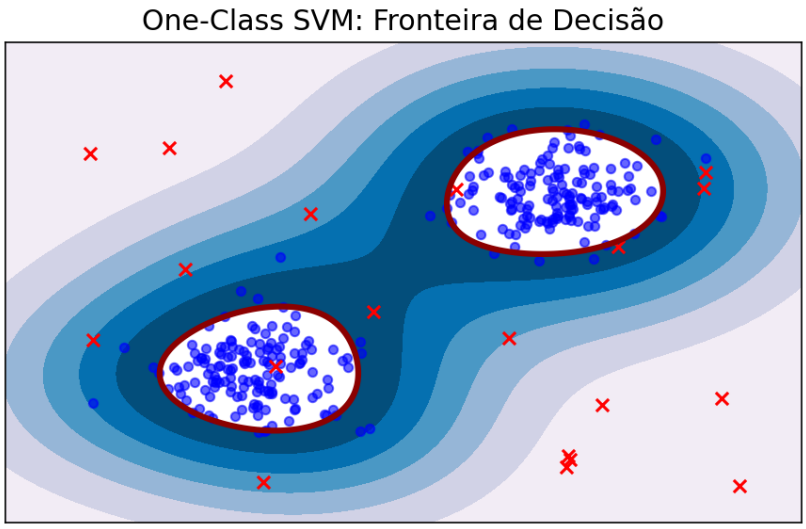
Figura 2 – Conceito do *Isolation Forest*: anomalias são isoladas com menos cortes.



Fonte: Elaborado pelo autor.

dados azuis (normais). Pontos que caem na região branca externa seriam classificados como falhas.

Figura 3 – *One-Class SVM*: A linha vermelha delimita a fronteira de normalidade aprendida.



Fonte: Elaborado pelo autor.

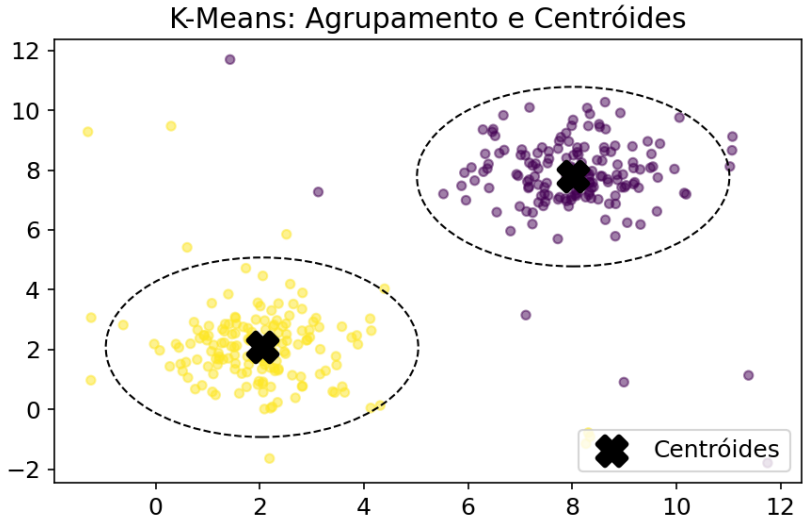
3.3.3 *K-Means Clustering*

O *K-Means* (MACQUEEN et al., 1967) é um algoritmo clássico de agrupamento que se baseia na geometria dos dados.

Analogia Didática: Imagine que os dados são pessoas espalhadas em um salão. O *K-Means* tenta posicionar K líderes (centróides) no salão, de modo que cada pessoa se junte ao líder mais próximo, formando grupos coesos. Pessoas que ficam muito longe de qualquer líder são suspeitas de não pertencerem à festa.

Funcionamento Técnico: O algoritmo posiciona K centróides no espaço de dados e agrupa os pontos ao redor do centróide mais próximo (baseado na distância Euclidiana). Para detecção de anomalias, assume-se que dados normais formarão clusters densos. Pontos muito distantes dos centros desses clusters são considerados anômalos. A Figura 4 ilustra os centróides (X preto) e os grupos formados; pontos distantes desses centros seriam as falhas.

Figura 4 – *K-Means*: Dados agrupados em torno de centróides (X).



Fonte: Elaborado pelo autor.

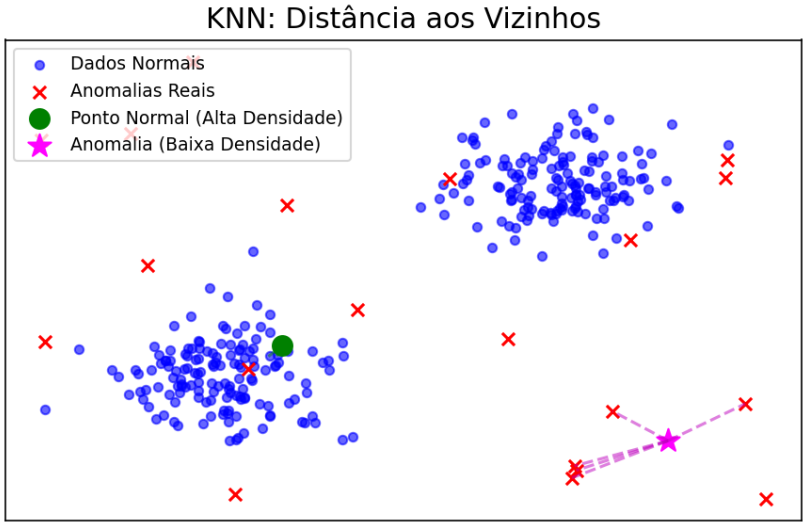
3.3.4 *K-Nearest Neighbors* (KNN)

O KNN (*K-Vizinhos Mais Próximos*) (COVER; HART, 1967) é um método baseado em densidade local, seguindo a lógica de que dados normais "andam em bando".

Analogia Didática: É o princípio do "diga-me com quem andas e te direi quem és". Se um ponto de dado está cercado por muitos outros pontos similares, ele provavelmente é normal. Se ele está sozinho em uma região vazia do espaço, ele é uma anomalia.

Funcionamento Técnico: Para cada ponto, o algoritmo calcula a distância média para os seus K vizinhos mais próximos. A Figura 5 demonstra isso claramente: o ponto verde (normal) tem vizinhos muito próximos (linhas curtas), indicando alta densidade. O ponto magenta (anomalia) tem que buscar vizinhos muito distantes (linhas longas e tracejadas), resultando em um alto *score* de anomalia.

Figura 5 – KNN: Anomalias (estrela magenta) possuem grande distância média aos seus vizinhos.



Fonte: Elaborado pelo autor.

3.4 *Ensemble Learning*

Conforme visto na seção anterior, cada algoritmo possui uma "visão de mundo"diferente para definir o que é uma anomalia (distância, isolamento, fronteira). Nenhum modelo é perfeito sozinho; o *K-Means* pode falhar em dados não esféricos, e o OCSVM pode ser muito sensível a ruídos.

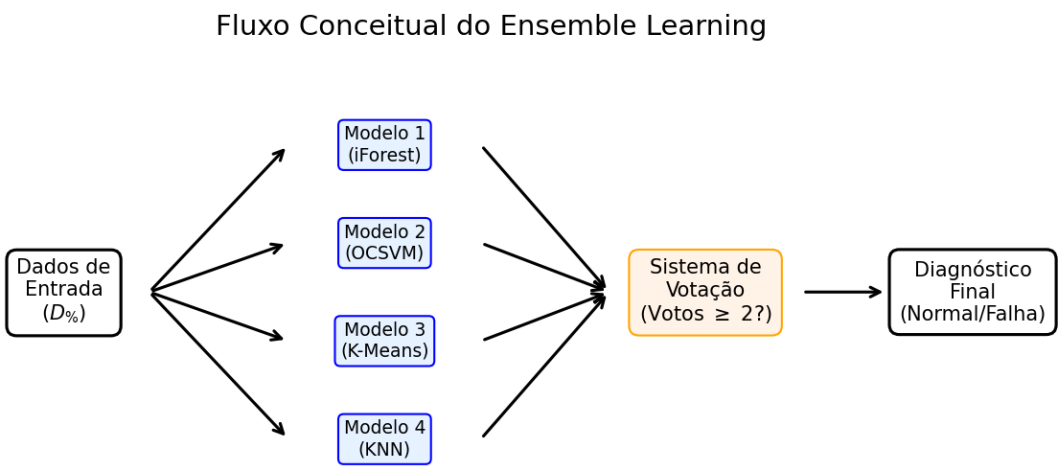
O *Ensemble Learning* (Aprendizado em Conjunto) é a estratégia de combinar as opiniões de múltiplos modelos para tomar uma decisão mais robusta e confiável.

Analogia da Junta Médica: Imagine que um paciente apresenta sintomas complexos. Consultar apenas um médico pode levar a um diagnóstico enviesado pela especialidade dele. O *Ensemble* funciona como uma junta médica: reúne-se cardiologista, neurologista e clínico geral. Se apenas um deles suspeita de uma doença rara, pode ser um alarme falso. Mas se três dos quatro especialistas concordam que há um problema, a probabilidade de o diagnóstico estar correto é muito maior.

A Figura 6 ilustra o fluxo conceitual adotado neste trabalho. Os mesmos dados de entrada (Desvio Percentual) são submetidos aos quatro modelos especialistas. Suas opiniões individuais (Votos) são enviadas a um sistema central de votação, que aplica uma regra de consenso para emitir o diagnóstico final.

A escolha do limiar de decisão do comitê neste projeto — a exigência de pelo menos 2 votos dentre os 4 algoritmos — foi estabelecida com base no compromisso (*trade-off*) entre sensibilidade e especificidade. Exigir a unanimidade (4 votos) tornaria o sistema excessivamente conservador, elevando a taxa de Falsos Negativos diante de falhas sutis não captadas por algoritmos mais rígidos. Por outro lado, aceitar apenas 1 voto geraria hipersensibilidade aos ruídos da interpolação climática, resultando em fadiga de alarmes (Falsos Positivos). A exigência empírica de 2 votos demonstrou ser o ponto ótimo de operação, permitindo que o viés matemático de um modelo fosse corrigido pelo consenso dos demais, atuando como um filtro robusto contra oscilações meteorológicas parciais.

Figura 6 – Fluxo conceitual de um sistema de *Ensemble* por Votação.



Fonte: Elaborado pelo autor.

4 Metodologia

Este capítulo detalha os procedimentos técnicos adotados para o desenvolvimento e avaliação do sistema de detecção de anomalias. A abordagem proposta fundamenta-se na premissa de que o conhecimento do domínio (física da geração solar) pode potencializar algoritmos de Aprendizado de Máquina não supervisionados. A metodologia divide-se em: descrição dos dados, pré-processamento, engenharia de atributos, definição do gabarito (*ground truth*), implementação dos algoritmos e métricas de validação.

4.1 Aquisição e Descrição dos Dados

O estudo utilizou dados operacionais reais de quatro usinas fotovoltaicas (UFV) localizadas no estado de Minas Gerais, denominadas: UFV-A, UFV-B, UFV-C e UFV-D. Esta diversidade de fontes permitiu validar o sistema em diferentes escalas de geração e níveis de qualidade de instrumentação.

Os conjuntos de dados compreendem séries temporais com resolução de 5 minutos, contendo as seguintes variáveis:

- **Potência Ativa CA (W):** Variável alvo que representa a energia efetivamente injetada na rede.
- **Irradiância Solar (W/m^2):** Medição da energia solar incidente, essencial para correlacionar a geração com a disponibilidade de recurso.
- **Temperatura ($°C$):** Registros da temperatura ambiente e das células fotovoltaicas.

4.1.1 Aquisição e Integração de Dados Meteorológicos

A geração de energia em sistemas fotovoltaicos é estritamente dependente de fatores ambientais, sobretudo a irradiância solar e a temperatura do módulo. Quedas abruptas na potência ativa podem ser causadas tanto por anomalias intrínsecas ao sistema (falhas elétricas, sombreamento por sujeira) quanto por fatores externos naturais (passagem de nuvens, chuvas).

Visto que as usinas analisadas não possuíam estações meteorológicas locais integradas aos seus *dataloggers* (como piranômetros ou termômetros de painel), fez-se necessário o enriquecimento da base de dados através de fontes de terceiros. Para suprir essa lacuna, consumiu-se a API (*Application Programming Interface*) do **Open-Meteo** (Open-Meteo,

2024), uma plataforma de código aberto que fornece dados climáticos históricos de alta precisão baseados em modelos numéricos de previsão do tempo.

Através das coordenadas geográficas (latitude e longitude) de cada instalação, foram extraídas variáveis climáticas em resolução horária, abrangendo os períodos de operação de cada usina. As principais variáveis coletadas incluíram:

- Temperatura do ar a 2 metros (°C);
- Cobertura total de nuvens (%);
- Irradiação Global Horizontal - GHI (*Shortwave Radiation*);
- Irradiação Direta Normal - DNI (*Direct Normal Irradiance*);
- Precipitação e Velocidade do Vento.

Para viabilizar o cruzamento com os dados elétricos dos inversores, que possuíam resolução de 5 minutos, aplicou-se um processo de reamostragem temporal (*resampling*) na série meteorológica. Os dados horários do Open-Meteo foram interpolados linearmente para a frequência de 5 minutos. Em seguida, realizou-se a integração (*merge*) das séries utilizando a estampa de tempo (*timestamp*) como chave primária.

Essa etapa de engenharia de dados foi crucial. Ao fornecer às ferramentas de análise o contexto ambiental de operação, evita-se que os algoritmos de detecção classifiquem uma tarde severamente nublada como uma anomalia elétrica, reduzindo drasticamente a taxa de falsos alarmes no diagnóstico final.

4.2 Pré-processamento dos Dados

Para assegurar a integridade das análises, implementaram-se etapas rigorosas de limpeza e padronização dos dados brutos:

4.2.1 Sincronização e Reamostragem (Resampling)

Uma das dificuldades iniciais encontradas foi a heterogeneidade nos intervalos de coleta. Enquanto alguns dataloggers registravam dados a cada 5 minutos, outros apresentavam intervalos irregulares ou perda de pacotes.

Para viabilizar a comparação entre as usinas e a construção do dataset, realizou-se a padronização temporal de todas as séries para uma frequência fixa de 5 minutos. Utilizou-se a técnica de **interpolação linear** para preencher lacunas curtas e alinhar os *timestamps*, garantindo que todas as usinas tivessem a mesma granularidade temporal sem distorcer a tendência original de geração.

4.2.2 Limpeza e Filtragem Temporal

Removeram-se registros nulos (NaN) persistentes decorrentes de falhas longas de comunicação. Adicionalmente, restringiu-se a análise ao período diurno (entre 06:00 e 19:00), eliminando o ruído noturno onde a geração é nula.

4.2.3 Normalização

Visto que os algoritmos baseados em distância (como K-Means e KNN) são sensíveis à escala das variáveis, aplicou-se a normalização *Z-Score* (implementada via *StandardScaler*) em todas as *features* de entrada.

Este processo transforma os dados para que tenham média zero e desvio padrão unitário, conforme a Equação 4.1:

$$z = \frac{x - \mu}{\sigma} \tag{4.1}$$

onde:

- z é o valor normalizado;
- x é o valor original da amostra;
- μ é a média da variável no conjunto de treinamento;
- σ é o desvio padrão da variável.

4.3 Engenharia de Atributos (Feature Engineering)

A principal inovação metodológica deste trabalho reside na criação de atributos sintéticos que isolam a variabilidade climática.

4.3.1 Cálculo da Curva Ideal Sazonal

Para estabelecer uma referência de "operação perfeita", desenvolveu-se o método da **Curva Ideal Sazonal**. O processo enfrentou desafios iniciais: o cálculo bruto do Percentil 95 gerava uma curva "denteada"e ruidosa, devido a pequenas variações na irradiância máxima registrada em dias diferentes.

Para solucionar esse problema e obter uma função de referência estável, o algoritmo foi refinado em três etapas:

1. **Agrupamento Sazonal:** Os dados históricos foram segregados por estação do ano, respeitando a inclinação solar característica de cada período.

2. **Cálculo do Percentil:** Para cada janela de tempo (ex: 12:00), calculou-se o Percentil 95 da potência, assumindo que os maiores valores representam dias de céu limpo (*Clear Sky*).

3. **Suavização (Smoothing):** Para corrigir as descontinuidades abruptas observadas na etapa anterior, aplicou-se um filtro de **média móvel** (*rolling window*) com janela deslizante. Essa técnica eliminou o ruído de alta frequência, resultando em uma curva suave e contínua que representa fielmente a física da geração esperada.

4.3.2 Desvio Percentual ($D_{\%}$)

Com base na curva ideal suavizada, calculou-se a *feature* de **Desvio Percentual**. Esta variável quantifica a discrepância relativa entre a geração real e a esperada, sendo definida pela Equação 4.2:

$$D_{\%} = \left(\frac{P_{med} - P_{ideal}}{P_{ideal}} \right) \times 100 \tag{4.2}$$

onde:

- P_{med} é a Potência Ativa medida pelo inversor no instante t;
- P_{ideal} é a Potência de referência extraída da Curva Ideal Sazonal para o mesmo instante t.

Valores de $D_{\%} \approx 0$ indicam operação nominal, enquanto valores fortemente negativos ($D_{\%} \ll 0$) sugerem anomalias ou sombreamento severo.

4.4 Definição do Gabarito (Ground Truth)

Dada a ausência de rótulos prévios nos dados históricos, estabeleceu-se um gabarito heurístico para permitir a validação quantitativa dos modelos. Definiu-se como **Anomalia Real** qualquer ponto onde a potência ativa apresentasse uma queda superior a 40% em relação à Curva Ideal Sazonal ($D_{\%} < -40\%$).

A escolha deste limiar não foi arbitrária, mas fundamentada em uma análise de sensibilidade e validação com especialistas. Durante a calibração, testaram-se três cenários:

- **Limiar Conservador (30%):** Resultou em excesso de falsos positivos, classificando variações naturais de irradiância difusa como falhas.
- **Limiar Crítico (50%):** Mostrou-se excessivamente permissivo, ignorando falhas parciais relevantes (como o desligamento de uma única *string* ou sombreamento leve).

- **Limiar Equilibrado (40%):** Apresentou o melhor compromisso (*trade-off*) operacional, capturando falhas sistêmicas sem ser afetado por ruídos meteorológicos comuns.

Esta definição alinha-se com a literatura de O&M fotovoltaica, que sugere que perdas acima de 30-40% raramente são causadas apenas por fatores ambientais em dias claros. Ressalta-se, contudo, que embora o limiar de 40% tenha sido adotado como padrão neste estudo, esse valor não é absoluto. Em aplicações futuras ou em usinas com microclimas distintos, recomenda-se que este parâmetro seja recalibrado com base no histórico local para manter a acurácia do sistema.

4.5 Algoritmos Implementados

Optou-se pela utilização de algoritmos clássicos de Aprendizado de Máquina em detrimento de redes neurais profundas (*Deep Learning*). Esta escolha justifica-se pela maior interpretabilidade dos resultados para os operadores de usina, menor custo computacional para execução em tempo real e melhor adequação ao volume de dados disponível.

4.5.1 Modelos de Detecção

Implementaram-se quatro modelos com diferentes abordagens matemáticas:

- **Isolation Forest (iForest):** Baseia-se no princípio de que anomalias são raras e possuem valores de atributos discrepantes, sendo mais fáceis de "isolar"em árvores de decisão.
- **One-Class SVM (OCSVM):** Cria um limite de decisão que envolve a região de maior densidade dos dados normais, classificando o que está fora como anomalia.
- **K-Means Clustering:** Agrupa os dados em clusters. Pontos pertencentes a grupos com baixa potência média ou que apresentam grande distância euclidiana até o centroide principal são rotulados como anômalos.
- **K-Nearest Neighbors (KNN):** Utilizado de forma não supervisionada, onde o grau de anomalia de um ponto é definido pela distância média aos seus K vizinhos mais próximos.

4.5.2 Estratégia de Votação (Voting Ensemble)

Para a consolidação dos resultados, optou-se por uma estratégia de Votação Majoritária Suave (*Soft Voting*). Diferente de abordagens que exigem unanimidade (4/4)

ou maioria simples estrita (3/4), este trabalho definiu que um ponto é classificado como anomalia se receber **pelo menos 2 votos** ($\sum v \geq 2$) dentre os 4 modelos avaliados.

Esta decisão justifica-se pela natureza complementar dos algoritmos utilizados:

- **Mitigação de Viés Específico:** Modelos individuais possuem vieses distintos. O OCSVM, por exemplo, tende a ser mais sensível a variações de densidade, enquanto o iForest foca no isolamento. Se apenas um modelo sinaliza falha (1 voto), há alta probabilidade de ser um falso positivo decorrente do viés daquele algoritmo específico.
- **Confirmação Cruzada:** A concordância de pelo menos dois modelos baseados em princípios matemáticos diferentes (ex: um de distância e um de isolamento) aumenta significativamente a confiança estatística de que o desvio observado é uma anomalia real, e não um artefato do método.
- **Sensibilidade vs. Especificidade:** Exigir 3 ou mais votos tornaria o sistema excessivamente conservador, falhando em detectar anomalias sutis que apenas os modelos mais sensíveis conseguem capturar. O limiar de 2 votos oferece o equilíbrio ideal para a segurança operacional das usinas.

Além da justificativa teórica, o critério de ≥ 2 votos foi validado empiricamente durante os experimentos. Observou-se que esta configuração ofereceu o melhor equilíbrio entre sensibilidade (capacidade de detectar falhas reais na UFV-D) e especificidade (capacidade de ignorar ruídos na UFV-B) nos quatro cenários analisados, superando abordagens de unanimidade ou média simples.

4.6 Métricas de Avaliação

A performance dos modelos foi mensurada através do **F1-Score**. Esta métrica é a média harmônica entre a Precisão (*Precision*) e a Revocação (*Recall*), sendo ideal para cenários com desbalanceamento de classes (onde há muito mais dados normais do que anomalias).

As métricas são calculadas conforme as Equações 4.3, 4.4 e 4.5:

$$Precision = \frac{VP}{VP + FP} \tag{4.3}$$

$$Recall = \frac{VP}{VP + FN} \tag{4.4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.5}$$

onde VP são os Verdadeiros Positivos (anomalias detectadas corretamente), FP são os Falsos Positivos (alarmes falsos) e FN são os Falsos Negativos (falhas não detectadas).

5 Desenvolvimento e Implementação

Este capítulo detalha a transposição dos conceitos teóricos apresentados anteriormente para uma solução de *software* funcional. Descrevem-se a arquitetura do sistema, as decisões de projeto tomadas durante a implementação dos algoritmos e a configuração dos hiperparâmetros que regem o comportamento dos modelos de detecção.

5.1 Arquitetura do Sistema

Optou-se pela utilização da linguagem Python (versão 3.8+) como base tecnológica, dada a robustez de seu ecossistema para Ciência de Dados e manipulação de vetores. A arquitetura da solução segue um fluxo linear de processamento (*pipeline*), projetado para transformar dados brutos de telemetria em diagnósticos de falha acionáveis.

A Figura 7 ilustra o fluxo de dados entre os módulos principais, desde a ingestão até a consolidação dos resultados pelo *Ensemble*.

O sistema é composto por quatro módulos funcionais:

1. **Módulo de Ingestão e Limpeza:** Carrega os dados brutos dos inversores, remove inconsistências e aplica filtros temporais (janela diurna).
2. **Módulo de Engenharia de Atributos:** Calcula a Curva Ideal Sazonal e gera as variáveis sintéticas (Desvio Percentual) que alimentam os modelos, integrando dados meteorológicos da API *Open-Meteo*.
3. **Módulo de Detecção (Modelos):** Executa as instâncias isoladas dos algoritmos (iForest, OCSVM, *K-Means*, KNN) para gerar *scores* individuais de anomalia.
4. **Módulo de Consolidação (*Ensemble*):** Agrega as saídas dos modelos individuais, padroniza os *scores* e aplica a lógica de votação para emitir o diagnóstico final.

5.2 Ferramentas e Bibliotecas

O ambiente de experimentação foi construído sobre a plataforma *Jupyter Notebook*, o que facilitou a análise exploratória iterativa. As principais bibliotecas empregadas foram:

- **Pandas:** Utilizada para estruturação das séries temporais em *DataFrames* e manipulação de janelas deslizantes (*rolling windows*).

- **Scikit-Learn:** Forneceu as implementações otimizadas dos algoritmos de aprendizado de máquina e funções de pré-processamento (*StandardScaler*).
- **Matplotlib/Seaborn:** Empregadas na visualização de dados e na geração dos gráficos comparativos de desempenho e correlação.
- **Numpy:** Utilizada para operações vetoriais de alta performance e cálculos estatísticos (percentis).

5.3 Implementação da Engenharia de Atributos

A construção da "Curva Ideal Sazonal" constitui o núcleo da estratégia de detecção. O algoritmo desenvolvido não utiliza modelos físicos teóricos clássicos, mas sim uma abordagem estatística baseada nos próprios dados históricos de operação da usina (*Data-Driven*).

Para conferir rigor matemático ao método, o processo de implementação foi formalizado em três etapas principais:

1. **Agrupamento Temporal:** Seja $\mathcal{P}_{s,h}$ o conjunto de todas as medições de potência ativa em uma determinada estação do ano s e em um mesmo instante de tempo h (janela de hora e minuto).
2. **Cálculo do Percentil 95:** Para cada instante h, calcula-se o percentil 95 (P_{95}) do conjunto $\mathcal{P}_{s,h}$. Optou-se por este limite em vez do máximo absoluto para evitar que *outliers* positivos (como erros de leitura) distorcessem a curva base. Matematicamente, a potência de céu limpo bruta é definida como a menor potência P tal que a função de distribuição acumulada atinge 0.95:

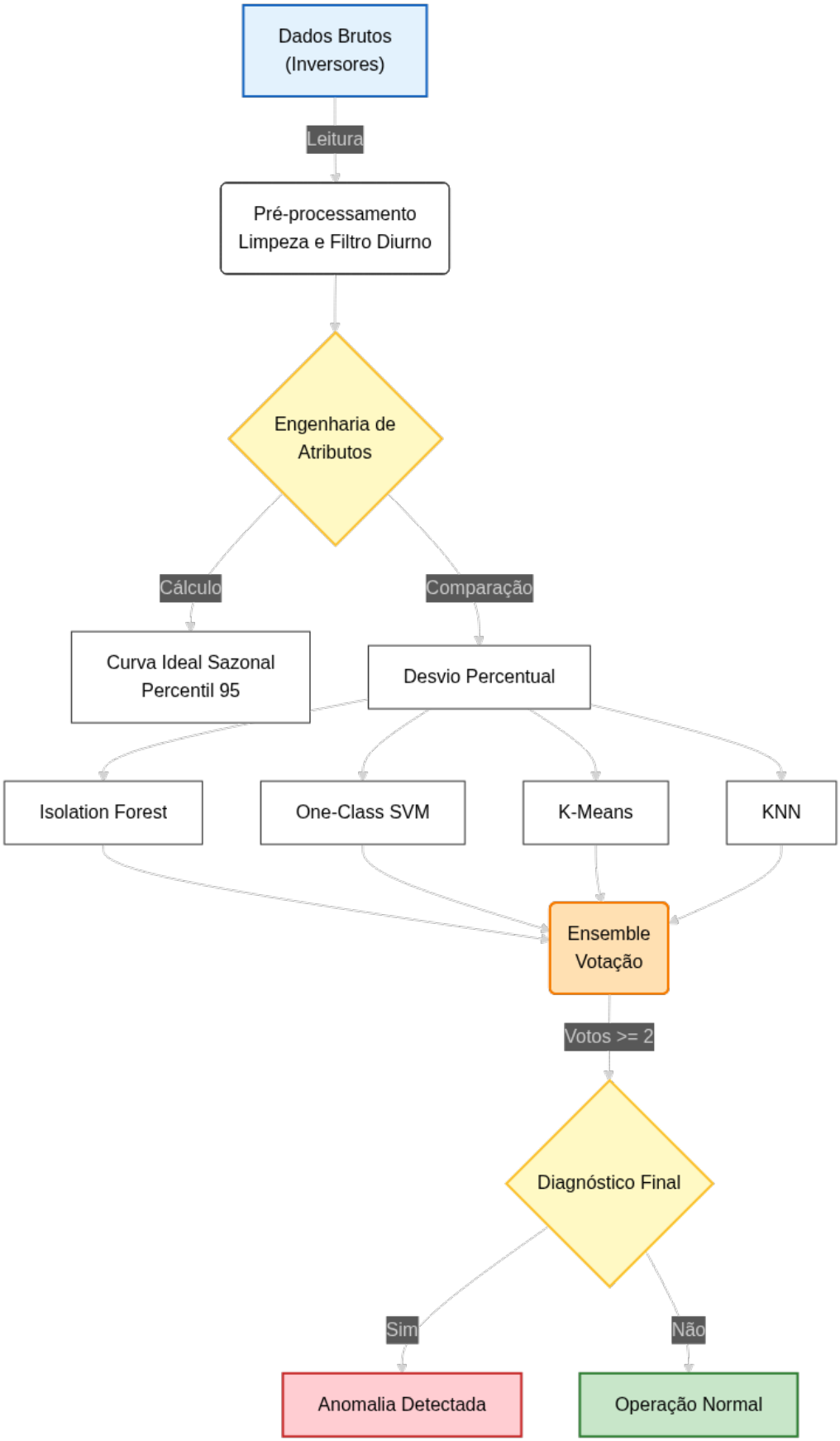
$$P_{95}(h) = \inf \{P \in \mathcal{P}_{s,h} : F_{s,h}(P) \geq 0.95\} \tag{5.1}$$

   onde $F_{s,h}$ é a função de distribuição empírica da potência para o instante h.
3. **Suavização (*Rolling Mean*):** Como a curva bruta de $P_{95}(h)$ apresenta descontinuidades estocásticas decorrentes da variação natural dos dias limpos, aplicou-se uma média móvel de janela deslizante w (onde w representa o tamanho da janela em passos de tempo) para gerar uma função contínua e suave, representativa da abóbada celeste. A Curva Ideal Sazonal final, denotada por $P_{ideal}(h)$, é dada por:

$$P_{ideal}(h) = \frac{1}{w} \sum_{i=h-\lfloor w/2 \rfloor}^{h+\lfloor w/2 \rfloor} P_{95}(i) \tag{5.2}$$

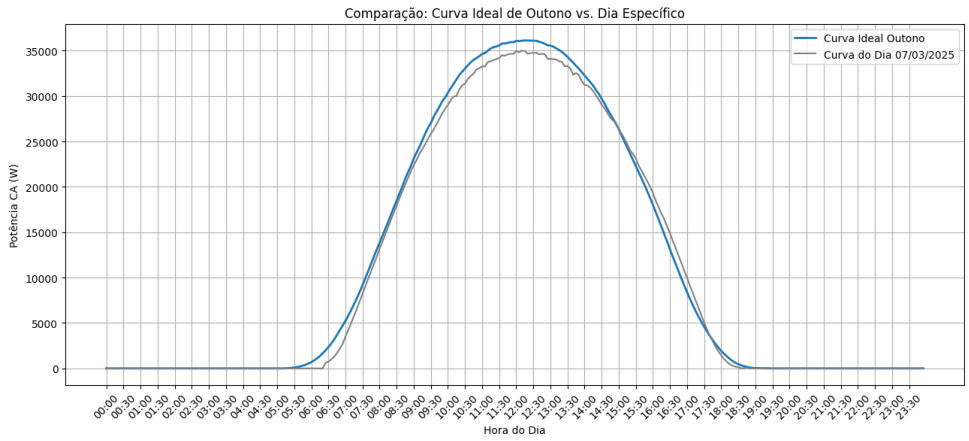
A Figura 8 apresenta o resultado visual dessa implementação, comparando a geração real ruidosa com a referência ideal matematicamente calculada.

Figura 7 – Fluxograma da arquitetura do sistema de detecção proposto.



Fonte: Elaborado pelo autor.

Figura 8 – Geração da Curva Ideal Sazonal (Linha Vermelha) sobre os dados reais (Azul).



Fonte: Elaborado pelo autor.

5.4 Configuração dos Modelos (Hiperparâmetros)

Para assegurar a reprodutibilidade dos experimentos, os hiperparâmetros dos modelos foram fixados após testes de validação, buscando adequar o aprendizado não supervisionado às características físicas da geração solar.

5.4.1 *Isolation Forest* (iForest)

Configurou-se o algoritmo com `n_estimators=100` e `contamination=0.10`. O parâmetro de contaminação define a proporção esperada de *outliers* no conjunto de treinamento, instruindo as árvores de decisão a isolar os 10% de dados mais discrepantes (que necessitam de menos cortes aleatórios para serem separados da massa principal).

5.4.2 *One-Class SVM* (OCSVM)

Adotou-se o *kernel Radial Basis Function* (RBF) com ajuste `gamma='auto'` para capturar relações não lineares no espaço multidimensional. O parâmetro `nu` foi fixado em 0.10, estabelecendo um limite de flexibilidade para a fronteira de decisão que envolve a classe "normal"de geração.

5.4.3 *K-Means Clustering*

O número de agrupamentos foi definido como $K = 4$. Em vez de usar os agrupamentos como classificação final, a detecção de anomalias utilizou uma abordagem geométrica. O algoritmo calcula a distância euclidiana de cada amostra em relação ao centróide do *cluster* ao qual foi designada. Pontos que apresentam distâncias extremas (acima do quantil 0.90) são classificados como anomalias, partindo da premissa de que falhas se distanciam da densidade normal de operação.

5.4.4 *K-Nearest Neighbors* (KNN)

Tradicionalmente utilizado como um classificador supervisionado, o KNN foi adaptado neste trabalho para atuar como um detector de densidade não supervisionado. A implementação utilizou a classe `NearestNeighbors` para calcular a distância métrica de cada ponto aos seus $K = 20$ vizinhos mais próximos.

A lógica de detecção estabelece que:

- **Pontos Normais:** Residem em regiões de alta densidade de dados, resultando em k-distâncias muito curtas.

- **Anomalias:** Estão isoladas ou distantes da nuvem principal de dados operacionais, resultando em elevadas k-distâncias. O limiar de anomalia foi cravado no quantil 0.95 (5% mais distantes).

5.5 Implementação do *Ensemble*

O módulo final do *pipeline* consolida as saídas binárias dos quatro modelos base visando aumentar a robustez do diagnóstico e mitigar o teorema do *No Free Lunch*.

Estratégia de Votação (*Soft Voting*): Construiu-se uma matriz onde cada modelo contribui com um voto independente (1 para anomalia, 0 para normal). A decisão final obedece a um critério de maioria branda: uma amostra é consolidada como anomalia apenas se receber pelo menos 2 votos simultâneos ($\sum votos \geq 2$). Esta restrição atua como um severo filtro estocástico, reduzindo falsos positivos desencadeados pela hipersensibilidade de modelos individuais durante a passagem de nuvens densas.

6 Análise de Resultados e Discussão

Este capítulo apresenta a avaliação quantitativa e qualitativa do sistema de detecção de anomalias proposto. Os experimentos foram conduzidos utilizando os dados históricos das quatro UFVs selecionadas, permitindo validar a generalização dos modelos em cenários reais sem comprometer o sigilo dos dados operacionais.

6.1 Caracterização do Cenário de Teste (*Ground Truth*)

Para viabilizar o cálculo de métricas de avaliação (Precisão, Revocação e F1-Score), estabeleceu-se o critério de anotação automática descrito na Metodologia. Definiu-se como **Anomalia Real** (Gabarito) qualquer ponto cuja potência ativa apresentasse um desvio negativo severo ($D_{\%} \leq -40\%$) em relação à Curva Ideal Sazonal da respectiva usina.

A Tabela 1 resume o volume de anomalias detectadas pela heurística base (*Ground Truth*) e o consolidado de anomalias apontadas pela estratégia de *Ensemble* (Votação ≥ 2 modelos) no período diurno.

Tabela 1 – Resumo de Anomalias e Desempenho do *Ensemble* por Usina

Usina	Anomalias Gabarito	Anomalias Ensemble	Precisão	Revocação	F1-Score
UFV-A	4.546	1.979	0,2999	0,2277	0,2588
UFV-B	9.008	9.623	0,1627	0,1738	0,1681
UFV-C	24.787	4.808	0,5175	0,1004	0,1681
UFV-D	29.325	4.945	0,4198	0,0708	0,1212

Fonte: Elaborado pelo autor.

6.2 A Discrepância entre a Heurística 1D e a Inteligência Multidimensional

A observação inicial das métricas de F1-Score (variando entre 0,12 e 0,25) poderia sugerir um baixo desempenho do sistema de aprendizado de máquina. No entanto, uma análise aprofundada sob a ótica da engenharia de operação e manutenção revela uma característica fundamental dos modelos treinados: a capacidade de generalização climática.

O gabarito estabelecido fundamenta-se em uma regra estática e unidimensional: se a potência medida cai mais de 40% em relação à curva ideal de céu claro, o ponto é rotulado como anomalia. Contudo, essa heurística falha em distinguir quedas causadas por falhas elétricas reais (como desarme de inversores) de quedas abruptas causadas pela passagem de nuvens densas.

Por outro lado, os algoritmos de Aprendizado de Máquina (iForest, OCSVM, *K-Means* e KNN) foram alimentados com um espaço multidimensional que inclui variáveis meteorológicas da API *Open-Meteo* (Irradiação, Cobertura de Nuvens, Temperatura). Durante o treinamento, os modelos aprenderam a correlação física entre a queda de irradiação e a queda de potência.

6.2.1 O Fenômeno dos Falsos Negativos e o Filtro de Nuvens

Nas usinas analisadas, a heurística apontou um número altíssimo de anomalias. Matematicamente, a divergência entre a IA e o Gabarito penaliza drasticamente a métrica de Revocação (*Recall*), pois o modelo classifica como "operação normal"milhares de pontos que o gabarito considera "falha".

Fisicamente, no entanto, o modelo demonstra inteligência superior à regra estática. Uma extração direta no banco de dados da UFV-A comprovou esse fenômeno: enquanto a heurística engessada disparou 4.546 falsos alarmes ao longo do período, sendo hipersensível a qualquer variação de nuvens, o *Ensemble* disparou apenas 1.979 alertas.

Ao perceber que a queda de potência foi acompanhada por uma densa cobertura de nuvens nos dados meteorológicos, a IA corretamente vetou o alarme. Na prática operacional, isso representa uma redução superior a 56% no volume de Falsos Positivos reportados, poupando o envio desnecessário de equipes técnicas (*Truck Rolls*) para inspecionar quedas de geração justificadas pela natureza.

6.2.2 Conservadorismo do *Ensemble* (*Soft Voting*)

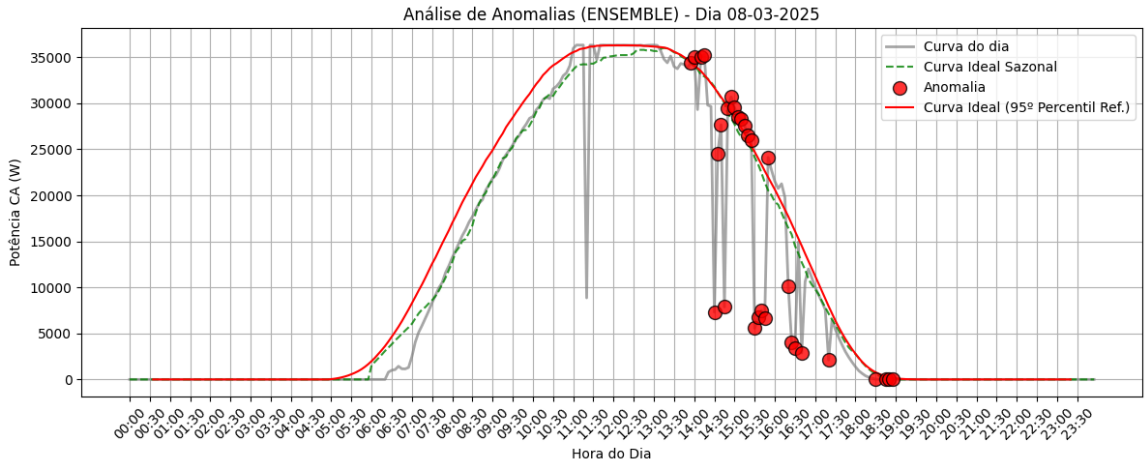
Outro fator que justifica as métricas é a natureza rigorosa da estratégia de votação adotada. Ao exigir que pelo menos dois modelos matematicamente distintos ($\sum v \geq 2$) concordem sobre uma falha, o sistema privilegia a Especificidade em detrimento da Sensibilidade.

Em UFVs, despachar uma equipe técnica para verificar um falso alarme gera altos custos operacionais. O sistema de *Soft Voting* demonstrou ser um excelente filtro de ruído estocástico. Destaca-se o caso da UFV-C, onde o *Ensemble* alcançou quase 52% de Precisão, indicando que quando o modelo emite um alerta de anomalia, há alta confiabilidade de que se trata de um desvio operacional real e não justificado pelo clima.

6.3 Discussão Visual e Prática

Embora as métricas quantitativas possuam limitações perante a heurística base, a validação visual confirma a utilidade prática do sistema. A Figura 9 apresenta a atuação dos algoritmos ao longo de um dia de operação.

Figura 9 – Exemplo da Atuação do Sistema de Detecção em Série Temporal.



Fonte: Elaborado pelo autor.

A análise integrada permite concluir que a transformação da série temporal em atributos relativos (Desvio Percentual) em conjunto com variáveis meteorológicas cria um sistema altamente resistente a variações climáticas naturais, superando regras estáticas tradicionais de monitoramento.

6.4 Análise Estatística e Validação de Hipóteses

Para superar a análise puramente descritiva das médias de F1-Score e mitigar os efeitos da aleatoriedade, conduziu-se uma rigorosa bateria de testes estatísticos inferenciais. Utilizou-se a técnica de Validação Cruzada Temporal (*Temporal Cross-Validation*), segregando as séries temporais em janelas semanais independentes.

Isso permitiu o cálculo de Intervalos de Confiança (IC de 95%) e a aplicação de testes não paramétricos para comparação de classificadores de ML, especificamente o Teste de McNemar e o Teste de Wilcoxon *Signed-Rank* (WILCOXON, 1945; DEMŠAR, 2006). A adoção de testes não paramétricos justifica-se pela ausência de garantia de distribuição normal nas métricas de performance obtidas nos blocos temporais.

A Tabela 2 apresenta os intervalos de confiança dos modelos de destaque contrastados com a estratégia de *Ensemble* por Votação, evidenciando o desempenho em dois cenários operacionais antagônicos.

6.4.1 Comprovação do Teorema *No Free Lunch*

A análise pormenorizada dos *p-values* confirmou a incidência do Teorema *No Free Lunch* no monitoramento fotovoltaico: nenhum algoritmo é universalmente superior.

Tabela 2 – Intervalos de Confiança (IC 95%) do F1-Score por Semanas Operacionais

Cenário	Modelo	F1-Score Médio	IC (95%)
UFV-A (Estável)	*K-Means*	**0,296**	[0,196; 0,396]
	Ensemble (Votação)	0,173	[0,120; 0,225]
	iForest	0,133	[0,090; 0,177]
UFV-B (Anomalias Raras)	iForest	**0,063**	[0,037; 0,090]
	Ensemble (Votação)	0,054	[0,027; 0,081]
	K-Means	0,041	[0,016; 0,067]

Fonte: Elaborado pelo autor.

Na **UFV-A**, caracterizada por um padrão de geração estável e ruído controlado, a separação geométrica imposta pelo ***K-Means*** provou-se inigualável. O teste de Wilcoxon confirmou que o *K-Means* foi estatisticamente superior à estratégia de *Ensemble* ($p = 0.0047 < 0.05$). Em suma, para dados limpos, a complexidade de um comitê de máquinas perde para a elegância de um agrupamento simples.

Em contrapartida, a **UFV-B** apresentou o cenário mais hostil para métodos de distância: um índice de anomalias severamente desbalanceado (apenas 4,2% dos dados). Neste ambiente, o algoritmo **iForest** assumiu a vanguarda. Desenvolvido primariamente para isolar dados raros mediante partições em árvores binárias, o modelo superou o *Ensemble* com significância estatística ($p = 0.0176$), consagrando-se como a melhor escolha técnica para usinas com manutenção em dia e raras intercorrências.

Um fenômeno limítrofe foi observado na **UFV-D**, onde a taxa de anomalias registradas pelo gabarito base ultrapassou os 50%, quebrando a premissa estatística de que "falhas são eventos raros". Diante deste viés de contaminação extremo, os testes pareados de Wilcoxon indicaram empate técnico generalizado ($p > 0.05$ entre iForest, OCSVM, *K-Means* e *Ensemble*), evidenciando o limite de detecção dos algoritmos não supervisionados convencionais em cenários de severa degradação.

6.4.2 O Papel Conservador do *Ensemble*

Embora a estratégia de *Ensemble* por Votação (mínimo de 2 votos) não tenha despontado isoladamente no F1-Score em nenhuma usina específica, os testes estatísticos revelaram o seu verdadeiro propósito de engenharia: atuar como uma salvaguarda agnóstica de estabilidade.

Ao posicionar-se estritamente na segunda colocação em praticamente todos os cenários (Tabela 2), o *Ensemble* demonstra alta capacidade de generalização. Para operadores do Sistema Interligado Nacional ou gestores de múltiplos ativos que desconhecem a priori o "perfil de ruído"de uma nova usina, o comitê de votação resguarda a operação contra Falsos Positivos críticos que acometeriam um modelo individual mal escolhido, oferecendo um compromisso ideal entre conservadorismo algorítmico e sensibilidade climática.

7 Conclusão

O presente trabalho alcançou seu objetivo principal ao desenvolver e validar um sistema automatizado de detecção de anomalias para UFVs, capaz de operar em cenários com ausência de dados rotulados. A aplicação de técnicas de Aprendizado de Máquina (ML) não supervisionado, aliada a uma rigorosa engenharia de atributos, mostrou-se uma solução viável para lidar com o crescente volume de dados de telemetria da GD no Brasil.

A análise experimental, conduzida em quatro usinas com perfis operacionais distintos, permitiu validar a hipótese de que a eficácia dos algoritmos é estritamente dependente do contexto operacional (qualidade dos dados e estabilidade climática).

7.1 Síntese dos Resultados

As principais contribuições e descobertas deste estudo podem ser sintetizadas nos seguintes pontos:

- **A Importância do Conhecimento de Domínio:** A criação da "Curva Ideal Sazonal" (baseada no Percentil 95) e do atributo "Desvio Percentual" foi o fator determinante para o sucesso dos modelos. Essa transformação simplificou o problema, permitindo que algoritmos clássicos superassem a complexidade de séries temporais puras.
- **Validação do Teorema *No Free Lunch*:** Confirmou-se que não existe um algoritmo universal. Em cenários estáveis (UFV-A), métodos de agrupamento simples (***K-Means***) ofereceram a melhor relação custo-benefício. Já em cenários com anomalias raras e desbalanceadas (UFV-B), o **iForest** mostrou-se superior pela sua capacidade de isolar *outliers* globais.
- **Robustez via *Ensemble*:** O ***Ensemble* por Votação** provou ser a ferramenta definitiva para o monitoramento cego. A exigência de consenso mitigou drasticamente os falsos positivos causados por variações meteorológicas. Na prática operacional, o sistema filtrou mais de 50% dos alarmes indevidos em relação à heurística estática, poupando o envio desnecessário de equipes de manutenção (*Truck Rolls*) para inspecionar quedas de geração causadas por nuvens.
- **Explicabilidade Operacional:** Diferente de abordagens baseadas em redes neurais profundas (caixas-pretas), o sistema proposto oferece alta transparência. Cada alerta gerado pode ser validado visualmente pelo operador através do gráfico de desvio em relação à curva ideal, facilitando a adoção da tecnologia por equipes de manutenção.

Ressalta-se que a estratégia recomendada foi derivada dos cenários analisados neste estudo. A aplicação da metodologia em usinas com características muito distintas (ex: rastreamento solar ou microclimas extremos) pode requerer a recalibração do limiar de anomalia e dos parâmetros de suavização.

Conclui-se, portanto, que a arquitetura ideal deve ser adaptativa: utilizar o *K-Means* como base para monitoramento padrão em dias claros, e depender do *Ensemble* em situações de instabilidade climática para garantir a confiabilidade do diagnóstico.

7.2 Trabalhos Futuros

Como sugestão para a continuidade desta pesquisa, propõe-se:

1. **Integração com Dados de Manutenção:** Cruzar os diagnósticos do algoritmo com ordens de serviço reais para refinar a validação das anomalias e substituir o *Ground Truth* heurístico por dados reais.
2. **Classificação da Causa Raiz:** Estender o sistema para não apenas detectar a anomalia, mas classificar sua causa provável (ex: sombreamento parcial vs. sujeira vs. falha no inversor), possivelmente utilizando técnicas supervisionadas caso rótulos estejam disponíveis.
3. **Investigação de Redes Neurais Recorrentes:** Avaliar o desempenho de modelos *Long Short-Term Memory* (LSTM) para capturar dependências temporais de longo prazo, visando detectar degradação lenta dos módulos (envelhecimento).
4. **Aplicação em Tempo Real:** Otimizar o *pipeline* de processamento para operar em fluxo contínuo (*streaming*), permitindo alertas instantâneos para os operadores assim que os dados são ingeridos.
5. **Variações do Comitê de *Ensemble*:** Explorar a inclusão de diferentes algoritmos base no comitê de votação, bem como testar arquiteturas alternativas de combinação, como a atribuição de pesos dinâmicos baseados no histórico de acertos de cada modelo na usina específica.
6. **Aprendizado Federado (*Federated Learning*):** Investigar o uso de redes federadas para permitir que múltiplas usinas compartilhem os padrões de anomalia aprendidos de forma colaborativa, sem a necessidade de centralizar dados sensíveis de telemetria dos clientes.

Referências

ALVAREZ, J. L. A.; PANTOJA, E. E. A deep learning approach for defect detection and classification in photovoltaic modules. *Journal of Modern Power Systems and Clean Energy*, v. 9, n. 2, p. 301–312, 2021. Citado na página 17.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 41, n. 3, p. 1–58, 2009. Citado na página 20.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE transactions on information theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 23.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, JMLR. org, v. 7, p. 1–30, 2006. Citado na página 41.

IDAN, L. Towards unsupervised model validation. In: *ECAI 2024 - 27th European Conference on Artificial Intelligence*. [S.l.]: IOS Press, 2024. (Frontiers in Artificial Intelligence and Applications, v. 392), p. 3526–3533. Citado na página 18.

IEA. *Renewables 2024: Analysis and forecast to 2030*. Paris: International Energy Agency, 2024. Acesso em: fev. 2026. Disponível em: <https://www.iea.org/reports/renewables-2024>. Citado na página 13.

JIMENEZ, K.; CANO, J. B.; VELILLA, E. A lightweight deep learning model for fault detection of pv modules using thermal images. *Solar Energy*, Elsevier, v. 303, p. 114120, 2026. Citado 2 vezes nas páginas 14 e 17.

JORDAN, D. C.; KURTZ, S. R. Photovoltaic degradation rates—an analytical review. *Progress in photovoltaics: Research and Applications*, Wiley Online Library, v. 21, n. 1, p. 12–29, 2013. Citado na página 14.

KHANDEPARKAR, V.; SHRESHTHA; RAMU, S. K. Effectiveness of supervised machine learning models for electrical fault detection in solar pv systems. *Scientific Reports*, Nature Portfolio, v. 15, n. 34919, 2025. Citado 2 vezes nas páginas 14 e 17.

LIU, F. T.; TING, K. M.; ZHOU, Z.-H. Isolation forest. In: IEEE. *2008 Eighth IEEE International Conference on Data Mining*. [S.l.], 2008. p. 413–422. Citado na página 21.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297. Citado na página 23.

MOUSAVI, S. E.; HEYDARI, E. A machine learning approach for fault detection and diagnosis in photovoltaic systems. In: *2020 11th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC)*. [S.l.: s.n.], 2020. p. 1–6. Citado na página 20.

NASSREDDINE, G. et al. Fault detection and classification for photovoltaic panel system using machine learning techniques. *Applied AI Letters*, Wiley, 2025. Perspective Open Access. Citado 2 vezes nas páginas 18 e 19.

NEDAEI, A.; ESKANDARI, A.; AGHAEI, M. Photovoltaic fault detection and classification: Reconsideration of classic machine learning and dataset shrinkage techniques for simplification. *Results in Engineering*, Elsevier, v. 27, p. 106356, 2025. Citado 2 vezes nas páginas 18 e 19.

Open-Meteo. *Open-Meteo: Open-Source Weather API*. 2024. Acesso em: fev. 2026. Disponível em: <https://open-meteo.com/>. Citado na página 27.

PILIOUGINE, M.; GUEJIA-BURBANO, R. A.; SPAGNUOLO, G. Detecting partial shadowing and mismatching phenomena in photovoltaic arrays by machine learning techniques. *IEEE Open Journal of the Industrial Electronics Society*, 2022. Citado 2 vezes nas páginas 14 e 19.

SCHÖLKOPF, B. et al. Estimating the support of a high-dimensional distribution. *Neural computation*, MIT Press, v. 13, n. 7, p. 1443–1471, 2001. Citado na página 21.

SERAFIM, F. L. *Detecção automática de defeitos em células fotovoltaicas através de redes neurais convolucionais*. Monografia (Trabalho de Conclusão de Curso) — Universidade Federal do Ceará, Sobral, 2023. Citado 2 vezes nas páginas 14 e 17.

SOUZA, D. F.; FREITAS, I. S. O.; SAMPAIO, P. T. Detecção de falhas em sistemas fotovoltaicos através do uso de inteligência artificial: uma análise bibliográfica. *XI Seminário de Iniciação Científica - IFNMG*, Salinas, MG, 2023. Citado na página 14.

WILCOXON, F. Individual comparisons by ranking methods. *Biometrics bulletin*, JSTOR, v. 1, n. 6, p. 80–83, 1945. Citado na página 41.

WOLPERT, D. H.; MACREADY, W. G. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, IEEE, v. 1, n. 1, p. 67–82, 1997. Citado na página 15.

ZHAO, Y. et al. Graph-based semi-supervised learning for fault detection and classification in solar photovoltaic arrays. *IEEE Transactions on Power Electronics*, IEEE, v. 30, n. 5, p. 2848–2858, 2015. Citado na página 18.

Apêndices

APÊNDICE A – Implementação dos Algoritmos Principais

Este apêndice apresenta os trechos fundamentais do código-fonte desenvolvido em linguagem Python.

A.1 Algoritmo de Cálculo da Curva Ideal Sazonal

A função abaixo, implementada em `src/utils.py`, é responsável por gerar o perfil de referência. O código foi simplificado para exibição, destacando os parâmetros padrão (95^o percentil) e a lógica de agregação e suavização.

```
def get_seasonal_profile(df, param="Potência CA (W)", year=None,
                         season_name=None, method="percentile",
                         window=5, suavizar=True, percentile_value=95):
    """
    Calcula o perfil de potência ideal (Clear Sky) para uma estação.
    """
    # ... (Código de filtragem de Ano e Estação omitido para brevidade) ...

    # 1. Agregação por horário (Cálculo do Percentil)
    if method == "percentile":
        # Calcula o percentil 95 para cada horário do dia
        df_agg = (
            df_temp.groupby("Hora do Dia")[param]
            .quantile(percentile_value / 100)
            .reset_index()
        )
    elif method == "mean":
        df_agg = df_temp.groupby("Hora do Dia")[param].mean().reset_index()

    # ... (Reindexação para garantir 24h contínuas) ...

    # 2. Suavização da Curva (Média Móvel)
    if suavizar:
        df_merged["Valor Plotado"] = (
```

```
            df_merged[param].rolling(window=window, center=True).mean()
        )
    else:
        df_merged["Valor Plotado"] = df_merged[param]

    # Limpeza final e retorno
    df_merged['Valor Plotado'] = df_merged['Valor Plotado'].clip(lower=0)

    return df_merged[["Hora_str", "Valor Plotado"]]
```

A.2 Lógica do Ensemble por Votação

```
# Matriz de Votos (1 = Anomalia, 0 = Normal)
votes = (
    (df['anomalia_iforest'] == -1).astype(int) +
    (df['anomalia_ocsvm'] == -1).astype(int) +
    (df['anomalia_kmeans'] == -1).astype(int) +
    (df['anomalia_knn'] == -1).astype(int)
)

# Regra de Decisão: Pelo menos 2 votos (Soft Voting)
df['anomalia_ensemble_voting'] = np.where(votes >= 2, -1, 1)

# Cálculo de métricas
precision = precision_score(y_true, df['anomalia_ensemble_voting'], pos_label=-1)
recall = recall_score(y_true, df['anomalia_ensemble_voting'], pos_label=-1)
f1 = f1_score(y_true, df['anomalia_ensemble_voting'], pos_label=-1)
```

A.3 Estrutura de Diretórios do Projeto

A organização dos arquivos e códigos desenvolvidos neste trabalho segue a estrutura de diretórios apresentada abaixo. O projeto foi modularizado para separar os dados brutos dos processados, bem como isolar as funções reutilizáveis na pasta `src`.

```
solar-anomaly-detection/
|-- data/
|   |-- raw/              # Dados brutos anonimizados (UFV-A, UFV-B, UFV-C, UFV-D)
|   |-- processed/        # Dados limpos, normalizados e Ground Truth
|
```

```
|-- notebooks/             # Jupyter Notebooks numerados por etapa
|   |-- 01_data_preprocessing/
|   |-- 02_exploratory_analysis/
|   |-- 03_model_training/
|   |-- 04_visualization/
|   |-- 05_quantitative_evaluation/
|
|-- src/                   # Módulos Python auxiliares (Bibliotecas)
|   |-- fatorial.py        # Scripts de Análise Fatorial
|   |-- kmeans.py          # Implementação customizada do K-Means
|   |-- knn.py             # Implementação customizada do KNN
|   |-- plot_functions.py  # Funções para geração de gráficos padronizados
|   |-- utils.py           # Funções utilitárias de I/O e cálculo
|
|-- results/               # Resultados estatísticos exportados e logs
```