UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE TECNOLOGIA
CURSO DE ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Eduardo Goulart da Silva

CLASSIFICAÇÃO DO TIPO DE SOMBREAMENTO E DE SUJIDADE EM MÓDULOS FOTOVOLTAICOS UTILIZANDO LGBM E CURVAS I-V

Santa Maria, RS
2023

Eduardo Goulart da Silva

CLASSIFICAÇÃO DO TIPO DE SOMBREAMENTO E DE SUJIDADE EM MÓDULOS FOTOVOLTAICOS UTILIZANDL LGBM E CURVAS I-V

Monografia apresentada ao Curso em Engenharia de Controle e Automação na Universidade Federal de Santa Maria (UFSM), como requisito parcial para obtenção do grau de Bacharel em Engenharia de Controle e Automação.

Orientador: Prof. Dr. Daniel Fernando Tello Gamarra

Santa Maria, RS
2023

RESUMO

CLASSIFICAÇÃO DO TIPO DE SOMBREAMENTO E DE SUJIDADE EM MÓDULOS FOTOVOLTAICOS UTILIZANDO LGBM E CURVAS I-V

AUTOR: EDUARDO GOULART DA SILVA
ORIENTADOR: DANIEL FERNANDO TELLO GAMARRA

Com a crescente demanda global por energia elétrica, torna-se fundamental investir na produção de energia renovável e limpa. O Brasil, com seu vasto potencial geográfico, tem a oportunidade de diversificar sua matriz energética com fontes renováveis, trazendo benefícios econômicos, sociais e ambientais. Investir em produção de energia renovável é essencial para construir uma matriz sustentável capaz de atender às necessidades de consumo de maneira eficiente. Reduzir os custos com operação e manutenção da energia solar é crucial para garantir a viabilidade econômica de projetos fotovoltaicos. A gestão eficiente desses custos pode ser a diferença entre uma operação bem-sucedida e uma instalação pouco rentável, especialmente considerando o aumento da capacidade instalada de sistemas solares. Este trabalho propõe a estruturação de um conjunto de dados utilizando curvas I-V de um arranjo fotovoltaico em diferentes condições de sombreamento e sujeira, tendo em vista aplicar um algoritmo de aprendizagem de máquina que seja capaz de detectar essas condições. O conjunto de dados utilizado contém medições de corrente, tensão, temperatura do módulo e irradiação no momento da amostragem. O modelo utilizado para avaliação e comparação é o *Light Gradient-Boosting Machine*, que atingiu uma acurácia de 99,1% na avaliação da sujidade e de 96,5% na avaliação de sombreamento. Os resultados de outros modelos também foram comparados, chegando a acurácias de 100% na avaliação da sujidade.

Palavras chaves: Aprendizagem de máquina, energia solar, detecção de sujidade, detecção de sombreamento.

ABSTRACT

CLASSIFICATION OF SHADING AND DIRTINESS IN PHOTOVOLTAIC MODULES USING LGBM AND I-V CURVES

AUTHOR: EDUARDO GOULART DA SILVA

ADVISOR: DANIEL FERNANDO TELLO GAMARRA

With the growing global demand for electricity, it becomes essential to invest in clean and renewable energy production. Brazil, with its vast grographic potential, has the opportunity to diversify its energy matrix with renewable sources, bringing economic, social, and environmental benefits. Investing in renewable energy production is crucial for creating a sustainable matrix that can efficiently meet consumption needs. Reducing costs for solar energy operation and maintenance is critical to ensure the economic viability of photovoltaic projects. Efficient management of these costs can make the difference between a successful project and an unprofitable installation, especially considering the increase in the installed solar system capacity. This work proposes the structuring of a dataset using I-V curves from a photovoltaic array under different shading and dirtiness conditions, aiming for reducing maintenance costs. The dataset used contains measurements of current, voltage, module temperature and irradiation at the time of sampling. The model used for evaluation and comparison is the Light Gradient-Boosting Machine, which achieved an accuracy of 99.1% in dirtiness evaluation and 96.5% in shading evaluation. Results from other models were also compared, reaching an accuracy of 100% in dirtiness evaluation.

Keywords: Machine learning, solar energy, dirtiness detection, shading detection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GW	Gigawatts
LCOE	Custo nivelado da eletricidade
FV	Fotovoltaica
O&M	Operação e manutenção
LGBM	*Light Gradient-Boosting Machine*
SFI	Sistema isolado
SFVCR	Sistema conectado à rede elétrica
I-V	Corrente e tensão
V_{OC}	Tensão de circuito aberto
I_{SC}	Corrente de curto-circuito
MPP	Ponto de máxima potência
VMP	Tensão no ponto de máxima potência
IMP	Corrente no ponto de máxima potência
P-V	Potência e tensão
P_{max}	Potência máxima
GOSS	*Gradient-based one-side sampling*
EFB	*Exclusive feature bundling*
GAF	*Gramian angular difference field*
SVM	Máquina de Vetores de Suporte
ANN	Redes neurais artificiais
KNN	K-vizinhos Próximos
CNN	Redes neurais convolucionais
GAF	*Gramian Angular Field*
LSTM	*Long Short Term Memory*

SUMÁRIO

1 INTRODUÇÃO

A expansão acelerada no consumo de energia elétrica no mundo e a utilização cada vez maior de energia não renovável se apresentam como um dos grandes desafios energéticos que a humanidade enfrenta para seu futuro. Tendo em vista que o desenvolvimento da sociedade exige cada vez mais de um crescimento na produção de energia, técnicas e soluções criativas são demandadas para que os custos na produção de energia provindas de fontes renováveis sejam viabilizados como medida de atenuação da geração poluente. Visando essa problemática, estudos que indicam possíveis alternativas que não causem danos irreparáveis ao meio ambiente são de grande relevância.

As energias renováveis possuem fontes cujo reabastecimento ocorre de forma mais rápida do que seu consumo – o sol, o vento e a água são exemplos dessas fontes naturais. A energia renovável faz o uso de tecnologias para gerar energia elétrica, calor ou energia mecânica a partir dessas fontes. Atualmente, é conveniente utilizar carvão e petróleo para o abastecimento elétrico e de combustível. O grande problema reside no fato de que o consumo destes insumos é mais rápido do que o seu reabastecimento, ficando limitada a produção de energia a partir dessas fontes (NREL, 2001).

Uma das tendências mais notáveis no mercado de energia renovável é a crescente competitividade das energias eólica e solar. O custo dos painéis solares caiu drasticamente nos últimos anos, tornando-se uma das fontes de energia renovável mais econômicas. Do mesmo modo, o custo da energia eólica também reduziu, se tornando uma opção cada vez mais viável para atender às necessidades de energia. Em 2020, houve um aumento de 256 gigawatts (GW) de capacidade de geração renovável, representando um acréscimo de 30% em relação ao último recorde de crescimento anual de produção. Desde 2010, o custo nivelado de energia (LCOE) retraiu 85%, enquanto as usinas eólicas *onshore* (usinas com instalações em terra) retraiu 56% (REN21, 2021). O LCOE representa uma medida de custo relativo da energia produzida por diferentes fontes de geração de energia.

De acordo com as projeções publicadas pela Empresa de Pesquisa Energética para 2023, as estimativas de consumo, no Brasil, estarão situadas em aproximadamente 650 TWh/ano (EPE, 2017). Dessa forma, o crescimento expressivo na oferta de energia será exigido. Para suprir esta demanda, serão realizados fortes investimentos em usinas hídricas, mas que podem não suportar todo o crescimento de energia exigida. Por isso, uma das alternativas para auxiliar no suprimento seria o uso de fontes renováveis de forma distribuída, como é o exemplo da

energia fotovoltaica (FV) por ser uma maneira não poluente, silenciosa, eficiente e não prejudicial ao meio ambiente (RÜTHER, 2004) de geração elétrica.

Durante o ano de 2021, a matriz energética brasileira apresentou predominância da geração de energia hidráulica (56,8%), enquanto a energia solar representou 2,47% e a eólica com 10,6% (EPE, 2022). Em 2022, foram superados 16 GW de potência instalada no Brasila partir de instalações fotovoltaicas, somando 5 GW de usinas de grande porte e 11 GW dos sistemas de geração própria de energia elétrica instalados por consumidores em telhados, fachadas, etc (EXAME, 2022). Tendo em vista o potencial geográfico brasileiro para a produção de energia FV, o país ainda se encontra da 13ª posição no ranking mundial de energia solar, atrás de países que possuem extensão territorial menor.

1.1 MOTIVAÇÃO

A geração solar pode ser considerada como uma alternativa para a geração de energia elétrica, devido aos seus benefícios econômicos e sociais. No Brasil, em particular, a adoção da energia solar pode reduzir significativamente a pressão sobre a geração hidrelétrica, que atualmente representa 56,8% da matriz energética do país, de acordo com a (EPE, 2022). A forte concentração de geração a partir de hidrelétricas pode ser problemática quando fatores imprevisíveis, como períodos de estiagem, afetam a produção. Em relatório da (ONS, 2021) mostra que os reservatórios do eixo Sudeste e Centro-Oeste estiveram em níveis muito abaixo de suas médias históricas nos últimos 4 anos. Com a redução na geração das hidrelétricas, a alternativa é recorrer à ativação das termelétricas cuja produção é mais cara.

A diversificação na matriz energética brasileira será benéfica, uma vez que os riscos de pressão inflacionária serão mitigados devido a uma menor concentração de geração de energia a partir de hidrelétricas. Além disso, a energia solar possui vantagens em quesitos ambientais, tais como sua geração sustentável e livre de emissões de gases de efeito estufa. Outro aspecto positivo é a capacidade de implementação da geração distribuída, possibilitando a instalação de painéis solares em edifícios e áreas ruais. Somado a isso, o setor solar tem contribuído para a geração de empregos em suas diferentes etapas, desde a fabricação até a instalação e manutenção dos sistemas solares.

1.2 JUSTIFICATIVA

A etapa de operação e manutenção (O&M) é um aspecto crucial na avaliação da viabilidade econômica de projetos de energia solar. O processo de O&M busca mitigar os riscos potenciais, melhorar a confiabilidade de longo prazo, o LCOE, os preços dos contratos de compra e venda de energia, além de impactar positivamente o retorno sobre o investimento (TRETER, 2022). Geralmente, os custos de O&M da energia solar são baixos em comparação a outros formas de geração, uma vez que é exigida uma manutenção periódica de limpeza e inspeção elétrica. No entanto, a fase de O&M pode se estender por 20 a 35 anos, tendo um impacto significativo nos custos totais do projeto. As atividades que buscam reduzir os custos derivados de O&M estão se tornando cada vez mais importantes para a viabilidade dos projetos fotovoltaicos.

Desenvolver maneiras de reduzir os custos com O&M irão impactar diretamente na viabilidade econômica dos projetos, que influenciará para um maior crescimento da adoção dessa geração no país. Analisar a eficiência da produção das plantas fotovoltaicas pode ser um caminho para desenvolver projetos que buscam otimizar o tempo de manutenção.

Em instalações fotovoltaicas, vários parâmetros são responsáveis por afetar a geração. O principal deles é a radiação solar, que depende diretamente da orientação geográfica, bem como a inclinação da instalação. A temperatura dos módulos, o sombreamento e o estado de limpeza também influenciam a performance do sistema gerador fotovoltaico (RÜTHER, 2004)

O sombreamento é considerado uma questão crítica para a eficiência dos módulos. Uma sombra sobre a célula do painel, como é o exemplo de uma sombra de antena, reduz acentuadamente o rendimento do sistema. Isso ocorre, pois, a célula sobre a qual incidir a menor quantidade de radiação é que irá determinar a corrente de operação do conjunto conectado em série, afetando a potência.

A sujeira é outro responsável direto por limitar a produção de energia fotovoltaica. Ela leva a uma redução direta de radiação solar absorvida pelas células que compõem o módulo. Diante dos impactos negativos da sujidade dos módulos, faz-se necessária a limpeza com uma periodicidade razoável. A frequência com que a sujeira irá acumular no painel dependerá de elementos que compõem ambiente que, considerando a diversidade de fatores, é relevante indicações de período ótimo de limpeza.

Após projetar e ao ser feita a instalação dos módulos, para garantir ótimas operações, é necessário conhecer vários fatores que influenciam a eficiência do sistema de energia, pois esses sistemas são expostos a diversas falhas e defeitos que afetam a energia gerada, como, falhas por

curto-circuito, circuito aberto e mesmo algum dano durante o transporte e instalação dos painéis. O diagnóstico torna-se uma prática fundamental para detectar possíveis defeitos. Através de medições realizadas por um traçador de curvas de corrente e tensão (I-V) proposto em (TRETER, 2022) é proposto neste trabalho criar um conjunto de dados para ser aplicado no modelo *Light Gradient-Boosting Machine*, de modo a detectar sujidade e o grau de sombreamento incidente sobre os painéis solares.

1.3 OBJETIVOS

Tem-se como objetivo geral do trabalho:

- Analisar e avaliar o algoritmo de aprendizagem de máquina *Light Gradient-Boosting Machine* (LGBM) na detecção do tipo de sombreamento e sujeira em módulos fotovoltaicos, a partir de detecção de padrões em curvas I-V. Dessa forma, busca-se encontrar meios de tornar mais eficiente o processo de planejamento da manutenção de plantas solares, visando otimizar a geração de energia.

1.3.1 Objetivos Específicos

- Estruturar um banco de dados com variáveis preditoras a partir de medições de curvas I-V;
- Realizar uma análise exploratória nos dados para o treinamento dos modelos;
- Classificar o tipo de sombreamento incidente sobre módulos fotovoltaicos;
- Classificar a sujidade de módulos fotovoltaicos;
- Avaliar o modelo LGBM para predição de sujidade e sombreamento;
- Comparar o LGBM a outros modelos de aprendizagem de máquina.

1.4 ESTRUTURA DO TRABALHO

A fim de situar o leitor no contexto da pesquisa, o trabalho está organizado da seguinte forma. No capítulo 2, apresenta-se o referencial teórico, onde são abordados os temas essenciais para compreensão do desenvolvimento do trabalho, tais como energia solar, curvas I-V, LGBM, além dos modelos que serão utilizados como comparação: Máquina de Vetores de Suporte (SVM) e *K-nearest neighbors* (KNN). No capítulo 3, é descrita a estruturação das variáveis

preditoras utilizadas nas classificações das falhas, a partir de um conjunto de dados com medições das curvas I-V, além de detalhar as ferramentas e bibliotecas utilizadas ao longo do trabalho. No capitulo 4, são apresentadas as avaliações dos treinamentos do LGBM e a comparação com outros algoritmos, e, aqui, busca-se compreender o processo de aprendizagem do LGBM por meio de análise da importância das variáveis preditoras na classificação. Finalmente, o capítulo 5 traz as conclusões que sumarizam o conteúdo abordado durante o trabalho, indicando sugestões de trabalhos futuros.

2 REFERENCIAL TEÓRICO

2.1 CURVAS I-V

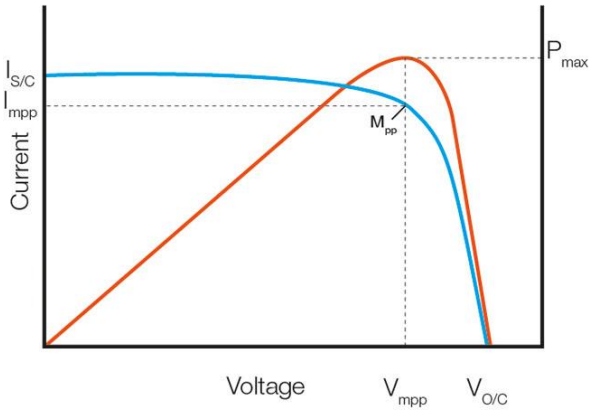
As curvas I-V são utilizadas para medir e caracterizar o desempenho de um sistema FV. De forma prática, elas mostram como a corrente elétrica varia com a tensão em um sistema de energia solar. A curva de potência e tensão (P-V), que mostra como a potência elétrica varia de acordo com a tensão, é utilizada comumente para se obter o rendimento do sistema, ou seja, que podem ser comparadas a potência adquirida com a projetada.

A análise da curva I-V proporciona vantagens significativas em detrimento a outros métodos de monitoramento. Para sua medição, o equipamento normalmente utilizado é o traçador de curva I-V. Através das características de corrente e tensão, os traçadores de curvas funcionam para verificar sinais observáveis de desempenho, defeitos, desgaste e degradação do módulo (WILLOUGHBY; OSINOWO, 2018).

O traçador de curva I-V relaciona a tensão e a corrente na saída de um módulo FV, medindo a corrente de curto-circuito (I_{SC}) e a tensão de circuito aberto (V_{OC}) para estimar a máxima potência do módulo. Com este equipamento, aliado a sensores de temperatura e irradiância, é possível mensurar os principais parâmetros elétricos de um módulo, podendo ser possível estimar quando algum efeito está interagindo no rendimento do sistema.

A Figura 1 traz duas curvas características do sistema FV. Na curva I-V, representada em azul, a I_{SC} indica a máxima corrente que o módulo pode fornecer e a V_{OC} indica a máxima tensão que o módulo pode fornecer. O ponto de máxima potência é o ponto entre a tensão no ponto de máxima potência (VMP) e a corrente no ponto de máxima potência (IMP). O MPP se situará justamente no “joelho” da curva I-V. Isto pode ser verificado na curva de potência e tensão (P-V), representada em vermelho, indicando a potência máxima (P_{max}) no ponto de VMP.

Figura 1 - Curvas I-V e P-V características.



Fonte: (SEAWARD)[1].

O rendimento de um módulo é diretamente afetado por obstáculos que bloqueiam a radiação solar, acúmulo de sujeira, defeitos inerentes do processo de fabricação, degradação natural ou por angulação imprópria. Os efeitos de degradação das células FV afetam diretamente as características de corrente e tensão e, portanto, o sistema de medição dos parâmetros de saída dos painéis FV pode fornecer informações em relação ao seu desempenho que podem ser usadas para detecção de falhas (PAPAGEORGAS, et al. 2015).

2.1.1 Associações de Células e Módulos Fotovoltaicos

É possível obter níveis de tensão e corrente desejados em dispositivos FV por meio da associação em série e/ou paralela. Tais dispositivos podem ser células, módulos ou arranjos FV. Os arranjos FV são compostos por um conjunto de módulos que são conectados eletricamente em série e/ou paralelo para fornecer uma saída única de tensão e corrente (CRESESB, 2014).

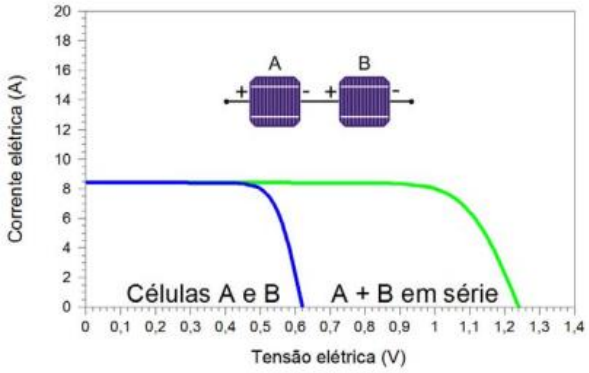
Na conexão em série, cuja representação da curva I-V está descrita na Figura 2, o terminal positivo de um dispositivo FV é conectado ao terminal negativo do outro dispositivo. Para dispositivos idênticos e submetidos à mesma irradiância, quando a ligação é em série, as tensões são somadas e a corrente elétrica não é afetada, como exemplificado nas equações (1) e (2).

$$I = I_1 = I_2 = \cdots = I_n \tag{1}$$

[1] Disponível em: https://www.seaward.com/gb/support/solar/faqs/84179-what-is-solar-pv-i-v-curve-tracing/

$$V = V_1 + V_2 + \cdots + V_n \quad (2)$$

Figura 2 - Soma das tensões na associação série.



Fonte: (CRESESB, 2014).

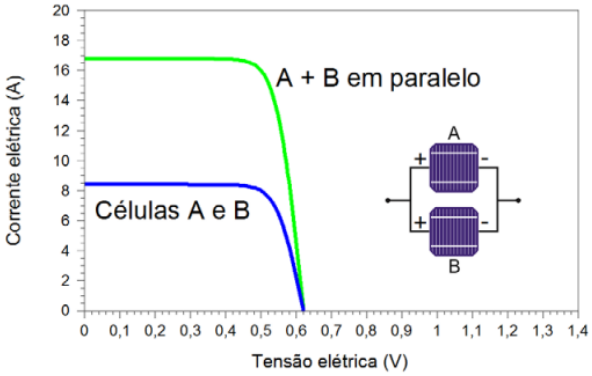
Na associação em paralelo, os terminais positivos dos dispositivos são interligados, assim como os terminais negativos. Aqui, obtém-se a soma das correntes elétricas em células ideias conectadas em paralelo, conforme as equações (3) e (4).

$$I = I_1 + I_2 + \cdots + I_n \quad (3)$$

$$V = V_1 = V_2 = \cdots = V_n \quad (4)$$

A Figura 3 ilustra o resultado da soma das correntes em uma associação paralela.

Figura 3 - Soma das correntes em uma associação paralela.

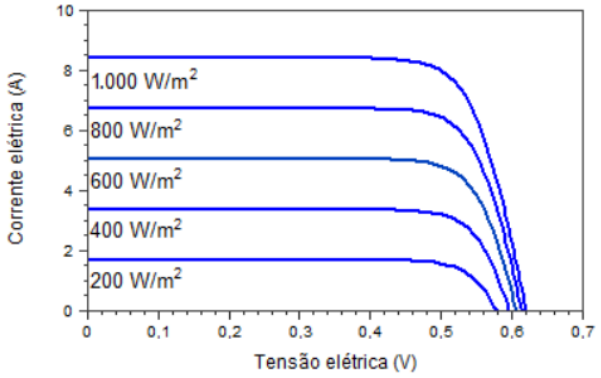


Fonte: (CRESESB, 2014).

2.1.2 Efeitos da Irradiação e da Temperatura nas Curvas I-V

A irradiação solar e a temperatura são dois fatores que influenciam a produção de energia pelos módulos FV. A corrente gerada pelo módulo varia linearmente com a irradiância, porém esta variação acontece de forma logarítmica com a variação de temperatura (CRESESB, 2014). A Figura 4 ilustra as curvas I-V características de acordo com diferentes níveis de irradiância. O MPP é diretamente proporcional ao nível de irradiância incidente sobre o módulo.

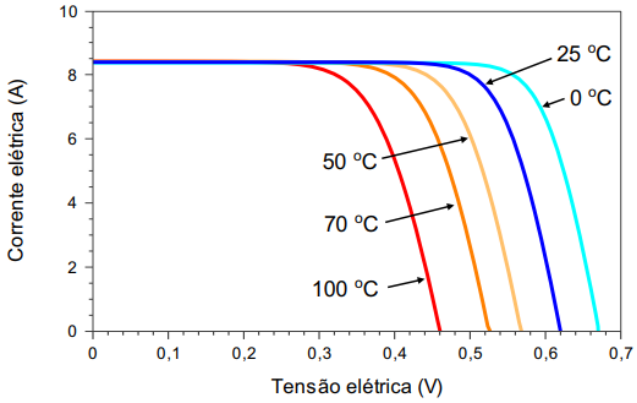
Figura 4 - Curvas características em diferentes níveis de irradiância.



Fonte: (CRESESB, 2014)

De Forma semelhante, as curvas I-V que relacionam diferentes níveis de temperatura estão ilustradas na Figura 5. Em que é possível alcançar maiores MPP com menores temperaturas nos módulos FV. É interessante observar o efeito da redução na V_{OC} derivado do aumento de temperatura.

Figura 5 - Curvas características em diferentes temperaturas.



Fonte: (CRESESB, 2014)

2.2 EFEITOS DA SUJEIRA EM PAINÉIS FOTOVOLTAICOS

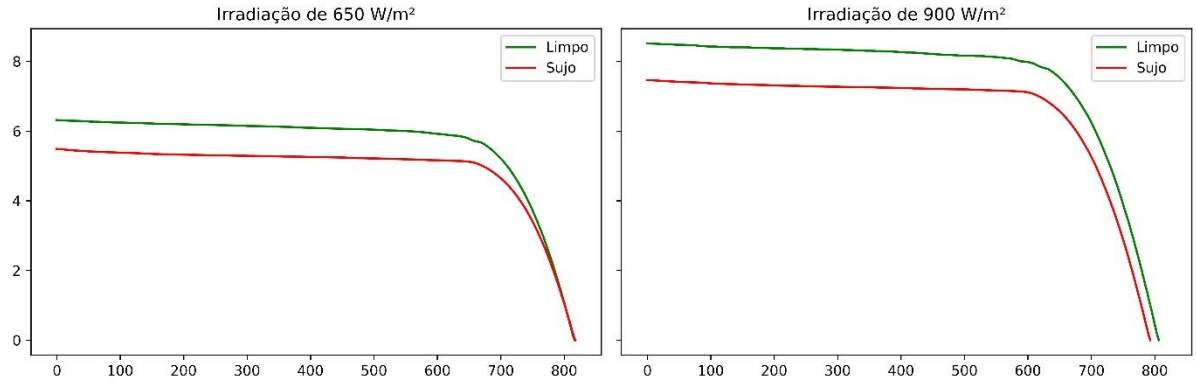
O estudo de (BARBOSA, 2018) indica os impactos negativos da sujeira nos módulos FV. Na pesquisa, foram comparados três módulos, sendo um deles recentemente limpo enquanto os outros possuem um período de um ano sem manutenção. Na Figura 6, é possível visualizar uma comparação entre os módulos sujos e limpos do estudo. Os resultados obtidos indicam uma perda de eficiência de 10,26% nos painéis sem manutenção.

Figura 6 - Comparação entre módulos sujos e limpos.

Fonte: (BARBOSA, 2018)

O acúmulo de partículas na superfície das células provenientes do ambiente externo provoca um bloqueio da radiação solar, impedindo que as células FV sejam induzidas para produzirem energia, reduzindo a eficiência da instalação. Os custos totais de um projeto de sistema FV dependem do impacto do acúmulo de sujeira nos módulos. A Figura 7 demonstra o comportamento das curvas I-V em dois valores diferentes de irradiação: 650 e 900 W/m². Torna-se perceptível a diferença na I_{SC} das curvas de acordo com a sua sujidade, bem como o seu incremento em maiores irradiações.

Figura 7 – Comparação de curvas I-V de módulos sujos e limpos em diferentes irradiações.



Fonte: Autor.

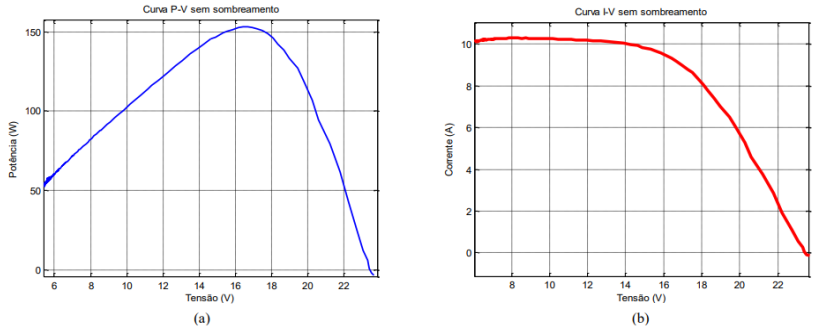
2.3 EFEITOS DO SOMBREAMENTO NA EFICIÊNCIA DO SISTEMA FOTOVOLTAICO

Quando os sistemas FV que são integrados a coberturas ou fachadas de edificações inseridas em meios urbanos tendem a receber sombreamentos parciais de seu entorno, afetando diretamente na produção de energia. Para mitigar os efeitos de sombreamento, os sistemas FV podem ser projetados de forma a minimizá-las, levando em consideração a trajetória solar e o ambiente em que está inserido.

O sombreamento pode possuir um impacto significativo do desempenho geral dos sistemas FV, reduzindo a absorção de radiação solar e, consequentemente, prejudicando a geração de energia. De acordo com a presença de radiação solar, as células FV fornecem corrente elétrica para um inversor FV. Quando uma célula deixa de receber radiação, ela deixa de conduzir corrente e, como existem ligações em série entre os painéis FV, as células associadas também deixarão de conduzir, impedindo que seja fornecida corrente ao circuito (COUTINHO et al., 2016).

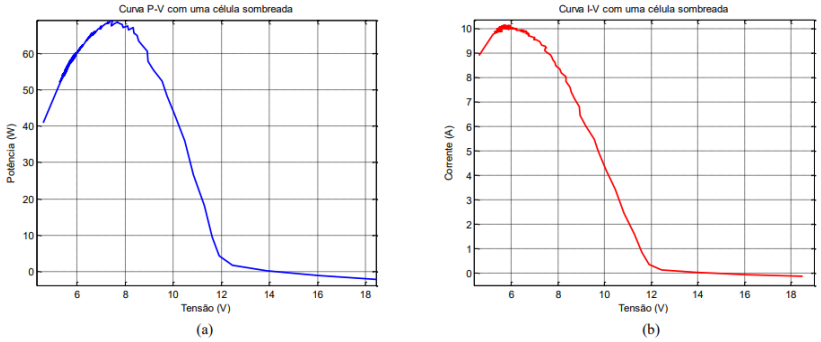
Para evitar problemas mais graves, um diodo é inserido em paralelo com uma ou mais células, de forma a desviar a corrente da célula sombreada. Com o sombreamento de parte de um módulo, as correntes que circulam nas células são reduzidas, o que ocasiona um acréscimo na tensão das células não sombreadas, podendo causar danos devido a sobreaquecimento. Este diodo, denominado de *by-pass*, consegue impedir que todo o módulo seja afetado pelo sombreamento de uma única célula, além de proteger o módulo contra tensões altas que poderiam causar o aparecimento de pontos quentes e, consequentemente, causar uma redução na vida útil do módulo. Os sistemas FV inseridos próximos a regiões que possuem alta densidade urbana, acabam sendo prejudicados pela presença de sombreamento provocada por edificações. Conforme as Figuras 8, 9 e 10 é possível realizar comparações no trabalho de (COUTINHO et al., 2016) da queda de rendimento devido à redução da MPP pelo efeito do sombreamento.

Figura 8 - Curvas P-V e I-V de um módulo sem sombreamento.



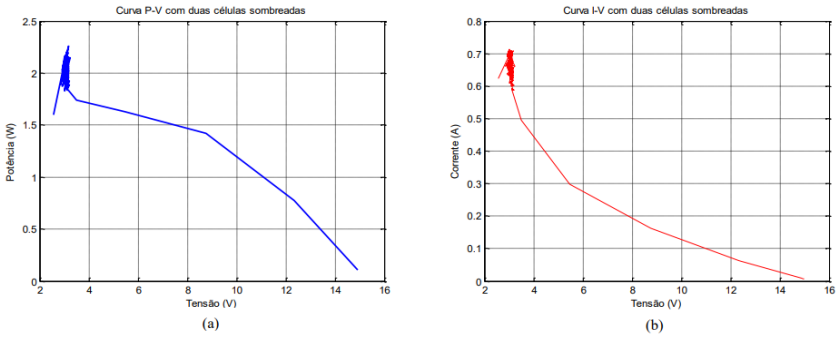
Fonte: (COUTINHO et al., 2016)

Figura 9 - Curvas P-V e I-V de um módulo com uma célula sombreada.



Fonte: (COUTINHO et al., 2016)

Figura 10 - Curvas P-V e I-V de um módulo com duas células sombreadas.



Fonte: (COUTINHO et al., 2016)

2.4 ALGORITMOS BASEADOS EM ÁRVORES DE DECISÃO

A árvore de decisão, consiste em um algoritmo de aprendizado de máquina supervisionado, amplamente utilizada para tomada de decisões e modelagem de problemas complexos. Este algoritmo se baseia em uma representação hierárquica em forma de árvore

cuja composição é representada a partir de nós e folhas (vértices e arestas). Cada nó representa uma decisão ou uma característica do problema em análise (ROKACH; MALMON, 2005).

A construção de uma árvore de decisão envolve a divisão recursiva do conjunto de dados com base em critérios específicos, a fim de otimizar a qualidade das divisões das folhas. Os critérios utilizados nos treinamentos para estimar o melhor modelo são baseados no ganho de informação, como a entropia. O objetivo com isso é classificar as variáveis preditoras que mais fornecem informação para que sejam postas no topo da árvore.

A entropia é uma medida de impureza utilizada nas árvores de decisão para avaliar a homogeneidade dos dados. Quanto menor a entropia, maior a pureza dos dados em relação às classes existentes, em que seu cálculo é baseado na distribuição de frequência das classes nos dados (BATRA; AGRAWAL, 2018). Supondo um conjunto de dados S com C classes preditas, em que p_i representa a proporção dos exemplos que pertencem à classe i, a descrição matemática da entropia pode ser dada através da equação (5).

$$H_{(S)} = \sum_{i=1}^{C} -p_i \log_2 p_i \tag{5}$$

Uma vez calculada a entropia das classes, o ganho de informação pode ser calculado para medir a redução da entropia de cada variável preditora. Os maiores valores de ganho de informação representam as variáveis preditoras que conseguem fornecer maior capacidade preditiva para o modelo. A representação matemática do ganho de informação está descrita na equação (6), em que: v são os possíveis valores da variável preditora; Si é o número de amostras com esse valor; N é o total de amostras; $H_{(Si)}$ é a entropia da variável preditora.

$$G_{(S,A)} = H_{(S)} - \sum_{i=1}^{v} \frac{S_i}{N} . H_{(Si)} \tag{6}$$

Existem alguns fatores que fazem com as árvores de decisão percam espaço para outros algoritmos mais complexos, inclusive derivados das árvores de decisão. Pode-se elencar alguns fatores como:

Tendência a sobre ajuste: As árvores de decisão têm a capacidade de se ajustar muito bem aos dados de treinamento, o que pode resultar em um modelo super ajustado aos padrões específicos desses dados.

Sensibilidade a pequenas variações nos dados: Uma pequena alteração nos dados de treinamento pode resultar em uma árvore de decisão completamente diferente. Isso ocorre, pois, as árvores são construídas com base em divisões determinísticas nos dados.

Dificuldade de lidar com dados de alta dimensionalidade: À medida que o número de variáveis preditoras aumenta, a complexidade das árvores também aumenta. Isso resulta em aumento do custo computacional e de risco de sobre ajuste.

Limitações na representação de relações complexas: As árvores de decisão possuem uma representação binária das relações entre as características dos dados, podendo dificultar a captura de relações complexas e interações entre as variáveis.

O *boosting* de gradiente visa melhorar o desempenho de um modelo preditivo baseado em árvores de decisão que combina vários estimadores fracos em um modelo mais forte. A ideia central dessa técnica é construir um modelo aditivo, em que cada novo estimador é ajustado aos erros residuais (ou gradientes) do modelo anterior (). Esses estimadores são tipicamente árvores de decisão rasas, conhecidas como aprendizes fracos. Os aprendizes fracos, que possuem baixa habilidade de generalização, serão combinados para comporem um modelo preditivo forte. Como objetivo principal, o *boosting* de gradiente busca reduzir uma função de custo iterativamente, utilizando-se de um passo de aprendizagem que ditará a contribuição da saída de cada aprendiz fraco.

Considerando um conjunto de dados $S = \{(x_i, y_i)\}_{i=1}{}^{n}$, em que n é o número total de amostras, e uma função de custo $L(y_i, F_{(x)})$, em que y_i é a saída verdadeira e $F_{(x)}$ é a saída predita, como entradas de um algoritmo de *boosting* de gradiente, consideram-se os seguintes passos para seu treinamento (BENTÉJAC; CSÖRGO; MARTÍNEZ-MUÑOZ, 2019):

Inicialização do modelo: O algoritmo começa com um modelo simples, geralmente um valor constante. Na classificação, a saída predita do primeiro classificador, conforme a equação (7), será um valor γ que minimiza o somatório da função de custo aplicada a todas amostras. Em outras palavras, γ será um valor de predição inicial e constante que resulta no menor erro entre as amostras.

$$F_{0(x)} = argmin_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma) \quad (7)$$

Iteração: Iterativamente, o algoritmo ajusta um novo estimador aos erros residuais do modelo anterior. Os erros residuais são calculados como a diferença entre as saídas reais e as previsões atuais do modelo. Dessa forma, uma nova árvore de decisão, treinada a partir dos resíduos é criada para corrigir as predições anteriores.

Atualização do modelo: O novo estimador é adicionado ao modelo existente com um fator de aprendizado α, que controla a contribuição do estimador para o modelo geral. Geralmente, o valor do passo de aprendizado é um valor pequeno para evitar *sobre ajuste*. A representação matemática da atualização pode ser descrita conforme a equação (8), em que a nova predição será a predição do modelo anterior somada a nova predição considerando uma taxa de aprendizado. Dessa forma, busca-se gerar resíduos menores que serão utilizados para o treinamento do modelo subsequente.

$$F_{(x)} = F_{i-1(x)} + \alpha . F_{i(x)} \quad (8)$$

Atualização dos pesos: Durante o processo de *boosting*, cada exemplo de treinamento é atribuído a um peso que indica a sua importância relativa. Os exemplos mal classificados possuem maiores pesos, o que leva aos estimadores subsequentes maior atenção aos exemplos mais difíceis de serem classificados.

Convergência: O processo iterativo é finalizado quando um critério de parada é alcançado, como um número máximo de estimadores ou uma melhora insuficiente na performance do modelo.

Predição: Para realizar uma previsão em um novo exemplo, os estimadores individuais são ponderados de acordo com seus fatores de aprendizado e combinados para formar o estimador final conforme resumido pela equação (9).

$$F_{(x)} = F_{0(x)} + \sum_{i=1}^{M} \alpha . F_{M(x)} \quad (9)$$

2.5 LGBM

O LGBM, de *Light Gradient-Boosting Machine*, recebe este nome por derivar da técnica de *boosting* de gradiente, enquanto se propõe a otimizar o tempo de treinamento e a acurácia dos modelos. Esse algoritmo, proposto em (KE, 2017), se diferencia, pois, utiliza duas técnicas chamadas *Gradient-based One-Side Sampling* (GOSS) e *Exclusive Feature Bundling* (EFB). A fim de melhorar a precisão das árvores de decisão que utilizam *boosting* de gradiente, o LGBM realiza seu treinamento utilizando apenas algumas variáveis preditoras e amostras. A seleção das amostras e variáveis que irão compor o treinamento são resultados do GOSS e do EFB.

Observou-se no trabalho de (KE, 2017) que as instâncias de dados com diferentes gradientes desempenham papéis diferentes no cálculo de ganho de informação. Em particular, as instâncias com gradientes maiores irão contribuir mais para o ganho de informação. Se uma instância for associada a um gradiente pequeno, o erro de treinamento para essa instância é pequeno e ela já está bem treinada. Assim, o GOSS se propõe a manter todas as instâncias com gradientes grandes e realiza uma amostragem aleatória nas amostras com gradiente pequeno.

Geralmente em aplicações reais, embora existam muitas variáveis preditoras, o conjunto de dados é muito esparso, isto é, contém uma grande quantidade de elementos nulos em sua matriz. Por exemplo, um conjunto esparso é visto em problemas de aprendizado de máquina que possuem muitas variáveis preditoras categóricas, principalmente quando é utilizado a codificação *one-hot*.

Pensando em reduzir a dimensão de variáveis preditoras, o EFB é proposto como diferencial do LGBM para agregar variáveis mutualmente exclusivas. Supondo um conjunto de dados com 3 variáveis: A, B e C. Considerando que A e B são exclusivas, ou seja, raramente assumem valores diferentes de zero simultaneamente. Por outro lado, C não é exclusiva e pode assumir valores diferentes de zero junto com as características A e B. Nesse caso, o EFB irá agrupar as variáveis para formarem apenas dois grupos: [A,B] e [C]. Isso resulta em apenas dois grupos que serão utilizadas para o treinamento do modelo.

Por se tratar de um algoritmo baseado em *boosting* de gradiente, é imposto ao LGBM uma dificuldade em lidar com valores discrepantes, que acabam distorcendo os resultados da predição. Uma vez que esses registros terão um resíduo grande e os aprendizes subsequentes

tentarão realizar predições em cima desse erro, o modelo se adaptará erroneamente para corrigir esses valores que fogem da normalidade.

2.5.1 Parâmetros do Modelo

Alguns parâmetros são importantes para a configuração de um algoritmo LGBM, que objetivam melhorar a acurácia final ou evitar sobre ajuste do treinamento:

Número de árvores (*n_estimators*): quantidade máxima de árvores que podem ser criadas;

Passo de aprendizagem (*learning_rate*): taxa de aprendizagem para que nenhuma árvore possua predominância na capacidade preditiva do sistema. Esse parâmetro é o mais importante para evitar sobreajuste na predição final.

Máxima profundidade (*max_depth*): se refere à máxima profundidade que cada árvore do conglomerado poderá possuir e possuirá impacto direto na prevenção de sobre ajuste em cada aprendiz fraco;

Número de folhas (*n_leaves*): Número máximo de folhas que a árvore poderá ter. Entende-se por folha o nó final de cada árvore que conterá apenas subconjuntos de dados homogêneos. Esse parâmetro também terá influência na prevenção de sobre ajuste em cada aprendiz fraco;

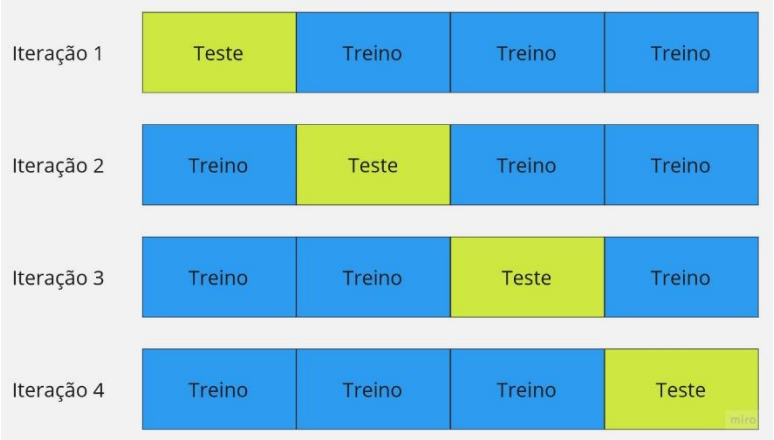
Objetivo: se refere a função de custo que será utilizada no problema. Na classificação binária, a função utilizada é a Entropia Cruzada, enquanto na classificação multiclasse, a função *Softmax* é comumente adotada.

2.5.2 Validação Cruzada

A validação cruzada consiste em ser uma técnica de avaliação que permite analisar a capacidade de generalização do modelo. É amplamente utilizada em problemas onde o objetivo da modelagem é a predição. Seu conceito principal é o particionamento do conjunto de dados de treinamento em subconjuntos mutualmente exclusivos que serão usados para estimar os melhores parâmetros do modelo. Dessa forma, a validação cruzada consiste em dividir o conjunto de dados aleatoriamente em “k” subconjuntos de mesmo tamanho. A cada iteração do

treinamento, um conjunto formado por k-1 subconjuntos serão utilizados no treinamento e o subconjunto restante para a avaliação do treinamento, conforme exemplificado na Figura 11, gerando um resultado (uma acurácia, por exemplo). A acurácia média dos resultados da validação cruzada pode ser empregada para avaliar quão bem o estimador está desempenhando (SHAH, R., 2021) durante o treinamento.

Figura 11 - Exemplo de aplicação da validação cruzada.



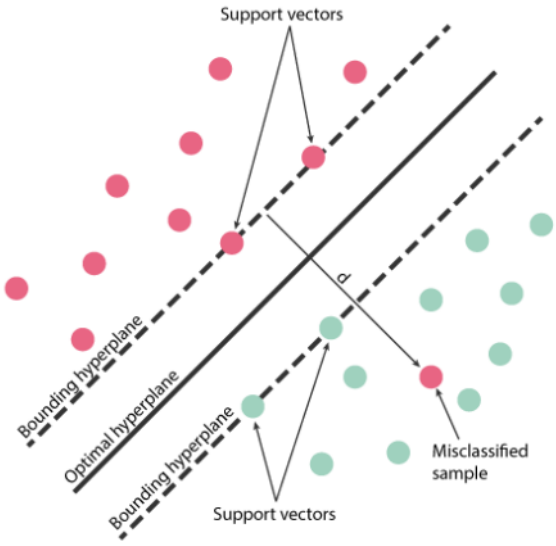
Fonte: Autor.

2.6 MÁQUINA DE VETORES DE SUPORTE

Alguns estudos tem proposto métodos para classificar falhas em dispositivos de geração de energia. O SVM tem se destacado por sua alta capacidade de predição. A aplicação de um modelo SVM é utilizado em (HÜBNER, 2021) para uma classificação multiclasse, utilizando a velocidade estimada de rotação do rotor de aerogeradores como variável de entrada. O SVM foi capaz de predizer níveis de desbalanceamento de massa nas pás dos aerogeradores com uma acurácia de 84,5%. Como entrada do treinamento, foram fornecidas amplitudes de uma série de frequências resultante da aplicação da técnica Densidade Espectral de Potência (PSD) na velocidade do rotor.

O SVM tem como funcionamento a definição da fronteira de separação, chamada de hiperplano, que melhor distingue as classes da classificação. Seu principal objetivo é encontrar o hiperplano de margem máxima, ou seja, que possui distância máxima entre os pontos das classes. Ao maximizar essa margem, é fornecida uma confiança maior para que novos pontos possam ser classificados evitando enviesamento. Os pontos mais próximos do hiperplano são chamados de vetores de suporte. Na Figura 12, está exemplificado um hiperplano linear que melhor consegue classificar as amostras.

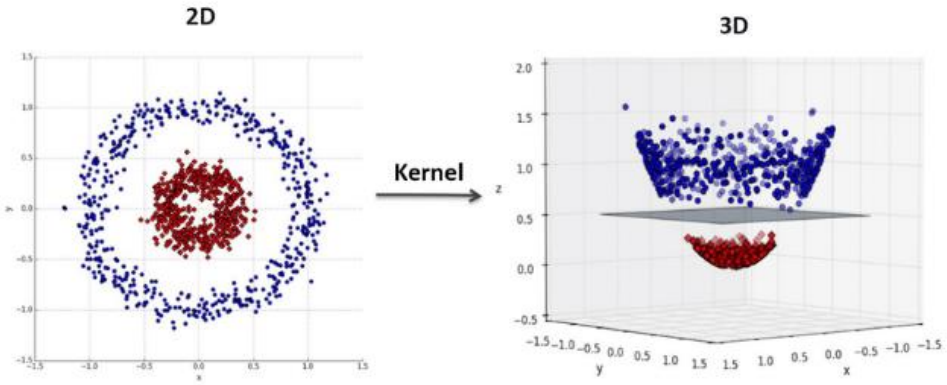
Figura 12 - Exemplo da definição do hiperplano do SVM.



Fonte: (CARDOSO-FERNANDES, 2020).

Apesar do SVM se basear em dados cujas classes são linearmente separáveis, o emprego do modelo não se limita a problemas dessa natureza. Como solução para conjuntos de dados mais complexos, são utilizados *kernels* que são utilizados para transformar os pontos em uma superfície linearmente separável, resultando em um acréscimo de uma dimensão do problema, como ilustrado na Figura 13. Essa técnica se baseia na aplicação de uma função nos dados de entrada.

Figura 13 - Transformação gerada pelo *kernel*.



Fonte: (HACHIMI et al, 2020).

2.7 K-VIZINHOS PRÓXIMOS

O KNN se trata de outro algoritmo que vem sendo utilizado com boas performances em problemas onde séries temporais são envolvidas. No trabalho de (LI et al, 2021) foi utilizada uma metodologia para aproveitar toda a medição da curva I-V no diagnóstico de falhas em módulos FV. A pesquisa explora diversas técnicas de aprendizado de máquina, incluindo redes neurais artificiais (ANN), para classificar oito condições de falha. O estudo tem como foco identificar a combinação mais eficaz de características de entrada e classificadores em termos de precisão e tempo de processamento. Além disso, a pesquisa investiga a robustez dos classificadores em relação a ruídos inseridos no conjunto de dados e busca explorar técnicas de redução de dimensionalidade de variáveis. Os resultados demonstram que o KNN alcança uma acurácia de 97,21%, enquanto outro modelo baseado em ANN e processamento de imagem obteve acurácia de 100%.

O KNN tem como objetivo prever a classe correta para os dados de teste calculando a distância entre os dados de teste e todos os pontos de treinamento. Em seguida, são selecionados os “k” pontos que estão mais próximos da instância de teste. Seu funcionamento pode ser descrito seguindo alguns passos:

Passo 1 – Selecionar o número “k” de vizinhos

São selecionados o número de vizinhos que serão considerados para a classificação de um novo ponto de dados, isto é, o novo ponto será classificado de acordo com a menor distância entre as “k” instâncias mais próximas.

Passo 2 – Calcular a distância euclidiana dos “k” vizinhos

É calculada a distância euclidiana o novo ponto de dados e todas as outras instâncias de treinamento. A distância euclidiana é uma medida que calcula a distância entre dois pontos em um espaço multidimensional e é utilizada para determinar a proximidade entre os pontos.

Passo 3 – Selecionar os “k” vizinhos mais próximos

São selecionados os vizinhos mais próximos de acordo com a distância euclidiana. Com base nisso, é realizada uma contagem das classes dos vizinhos. A classe com maior representação será utilizada para classificar a nova instância.

3 MATERIAIS E MÉTODOS

3.1 OBTENÇÃO DAS CURVAS I-V

O banco de dados utilizado deriva do trabalho de (TRETER, 2022), em que esta seção se destina a destacar os principais componentes e técnicas utilizadas para a medição da curvas I-V do estudo. No trabalho, foi proposto um traçador de curvas I-V que viabilizou a coleta de dados automatizado de curvas I-V de uma usina. Os parâmetros climáticos também são coletados de modo a que condições impostas sejam satisfeitas antes da medição, como uma irradiação mínima. Após a coleta da curva, os dados são exportados para um banco de dados *online*.

No estudo de (TRETER, 2022), a usina utilizada para a coleta das curvas I-V possui 384 módulos FV, abrangendo uma área de aproximadamente 2.000 m². Esses módulos são instalados em estruturas metálicas projetadas com uma inclinação de 30º, direcionados para o norte. A fabricante dos módulos é a Canadian Solar, e as especificações técnicas podem ser encontradas na Tabela 1.

Tabela 1 - Especificações técnicas dos módulos fotovoltaicos utilizados para coleta das curvas.

Parâmetro	Valor
Potência nominal máxima (P_{max})	270 W
Tensão no ponto de máxima potência (V_{mp})	30,8 V
Corrente no ponto de máxima potência (I_{mp})	8,75 A
Tensão de circuito aberto (V_{OC})	37,9 V
Corrente de curto-circuito (I_{SC})	9,32 A
Total de células FV por módulo	60 (6x10)

Fonte: (CANADIAN SOLAR, 2017)

A aquisição das curvas I-V foi realizada por meio de um traçador com carga capacitiva, em que as curvas coletadas foram comparadas com as do traçador comercial PVA1000-S, produzido pela Solmetric Corporation. As curvas obtidas foram semelhantes, resultando em um erro percentual 1,02% da potência extraída pelo módulo em menores irradiações. Reforça-se que a sobreposição das curvas obtidas permitiu validar o desempenho do traçador proposto.

Os capacitores do traçador permitem traçar a curva I-V de forma natural, precisa e com baixa ondulação na tensão e na corrente. O sensor LAH 25-NP, fabricado pela LEM, é o responsável por medir a corrente elétrica. Esse sensor possui precisão nominal de $\pm$0,3% e suporta valores nominais de 8, 12 ou 25 A, dependendo da configuração. Foi utilizada a

configuração de 12 A. A tensão elétrica foi medida a partir do sensor LV 25-P/SP5, produzido pela LEM, com precisão nominal de $\pm$0,8% e capacidade de medir até 1.500 V em corrente contínua.

A medição da irradiância no plano inclinado é uma importante forma de quantificar a entrada de energia no arranjo FV. Geralmente, essa medição é realizada utilizando células de referência ou pirômetros. No estudo foi utilizada uma célula de referência Si-I-420-T da Meteo Control. Essa célula tem uma faixa de medição de 0 a 1.500 W/m², com uma incerteza de $\pm$5 W/m².

A medição de temperatura de operação dos módulos FV pode ser realizada de diferentes maneiras: por contato (com sensor fixado no módulo), sem contato (com uma câmera infravermelha) ou de forma indireta (analisando a tensão de circuito aberto). No estudo foi proposta a medição por contato devido à facilidade de implementação e aos resultados satisfatórios obtidos por esse método. O sensor utilizado foi uma termorresistência PT-100, amplamente utilizado na indústria, que funciona com base na variação da resistência elétrica em função da temperatura.

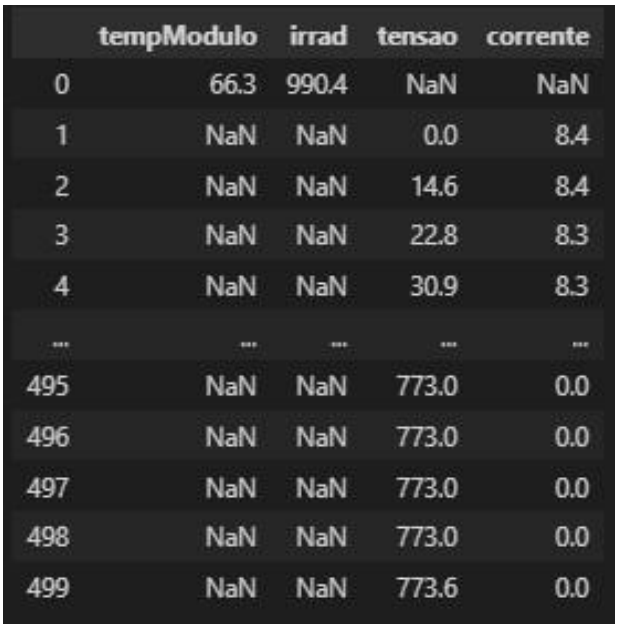
Para demonstrar o procedimento de criação do banco, três passos foram exigidos: extração, transformação e carregamento.

A extração, considerada a primeira etapa, possui o objetivo de coletar os dados das fontes. Neste caso, os valores de irradiação e temperatura são medidos para verificação das condições climáticas exigidas. Por fim, a curva I-V será medida, gerando uma série de corrente e outra de tensão.

Na etapa de transformação, os dados são convertidos, formatados e limpos para que possam ser armazenados. O primeiro passo foi o tratamento de *outliers* que consistem em dados que fogem da normalidade. Estes registros podem ser resultados devido a falhas no circuito de conversão A/D, por exemplo, e são prevenidos de serem populados no banco.

A última etapa, de carregamento, é responsável por estruturar os dados a fim de serem utilizados para análises exploratórias. No caso das medições das curvas I-V, os resultados são armazenados na nuvem por meio de arquivos em formato “.csv”. Assim, cada amostra registrada no banco de dados é definida de acordo com a Figura 14, em que se destaca uma série de corrente e de tensão com 500 pontos de medição cada. Estes pontos de medição de corrente e tensão correspondem a uma curva I-V.

Figura 14 - Formato do registro das medições das curvas I-V.



	tempModulo	irrad	tensao	corrente
0	66.3	990.4	NaN	NaN
1	NaN	NaN	0.0	8.4
2	NaN	NaN	14.6	8.4
3	NaN	NaN	22.8	8.3
4	NaN	NaN	30.9	8.3
...	...	...	...	...
495	NaN	NaN	773.0	0.0
496	NaN	NaN	773.0	0.0
497	NaN	NaN	773.0	0.0
498	NaN	NaN	773.0	0.0
499	NaN	NaN	773.6	0.0

Fonte: Autor.

3.1.1 Criação das Variáveis Preditoras

As variáveis preditoras são elementos usados em análise estatística para explicar o valor de uma variável predita. Em outras palavras, são os fatores que se acredita influenciarem ou terem relação com a variáveis que se tem interesse em estudar ou prever. No trabalho, existe o interesse em analisar o comportamento da predição utilizando a curva I-V completa. Dessa forma, o banco de dados de treinamento foi criado com os 500 pontos de corrente e tensão como variáveis preditoras, além da irradiação e da temperatura do módulo.

Cada medição de curva I-V está disposta em um arquivo ".csv", em que é necessário transformar o formato de dado tabular, a exemplo da Figura 14, para se adequar aos requisitos das bibliotecas *pycaret* e *scikit-learn*, utilizadas no trabalho. Como resultado final do processamento das curvas I-V, tem-se o conjunto de dados de treinamento, ilustrado na Figura 15, em que cada linha representa uma amostra da curva I-V. O conjunto final possui 569 amostras, com 1000 variáveis preditoras (os valores de tensão no curto-ciruito e de corrente no circuito aberto foram removidos, pois serão sempre zero).

Figura 15 - Estrutura final do conjunto de dados.

amostra	temp	irrad	label_sujeira	label_sombra	tensao_1	...	corrente_497	tensao_498	corrente_498	tensao_499	corrente_499
0	57.27	1015.756377	Limpo	Ensaio 1	0.0	...	0.040642	803.358322	0.040642	804.018888	0.0
1	36.37	1040.200000	Limpo	Ensaio 1	0.0	...	0.010000	881.100000	0.010000	882.700000	0.0
2	47.79	1094.686567	Limpo	Ensaio 1	0.0	...	0.029947	824.936811	0.029947	825.817566	0.0
...	...	...	...	...	...	...	...	...	...	...	...
566	66.27	966.200000	Sujo	Ensaio 6	0.0	...	0.050000	771.000000	0.050000	772.100000	0.0
567	66.30	990.385958	Sujo	Ensaio 6	0.0	...	0.047059	772.972287	0.047059	773.632853	0.0
568	66.30	993.500000	Sujo	Ensaio 6	0.0	...	0.050000	773.000000	0.050000	773.600000	0.0

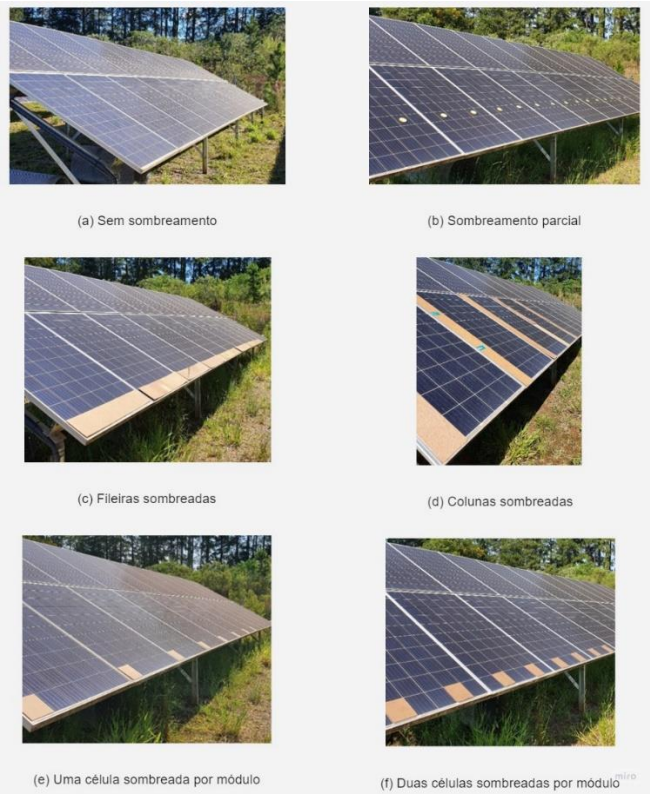
569 rows × 1002 columns

Fonte: Autor.

Dentre as variáveis preditas, as colunas “label_sujeira” e “label_sombra” dispõe para os modelos supervisionados a saída desejada. É importante ressaltar que cada amostra possui uma condição de sombreamento e uma de sujidade. As variáveis preditas tem seus objetivos definidos como:

Sombreamento: Reside a intenção de classificar com acurácia satisfatória o tipo de sombreamento incidente sobre o módulo. Assim, neste caso, os problemas de aprendizagem de máquina terão um caráter multiclasse, isto é, são problemas que possuem o objetivo de classificar três ou mais classes. A Figura 16 exemplifica os tipos de sombreamento das medições realizadas no experimento de (TRETER, 2022).

Figura 16 - Módulos fotovoltaicos com sombreamento.



Fonte: (TRETER, 2022)

Sujidade: Para a predição da sujidade do módulo, é proposta uma classificação binária. Dessa forma, os modelos objetivarão predizer apenas se o módulo está limpo ou não. Assim, a divisão entre as amostras se dará sendo que 297 pertencerão aos módulos limpos e 272 aos módulos sujos.

3.2 TREINAMENTO DOS MODELOS

Nesta seção, serão apresentadas as ferramentas utilizadas na realização do trabalho. Optou-se pela linguagem Python devido a sua ampla variedade de bibliotecas destinadas para análise de dados e treinamento de modelos. As bibliotecas disponíveis permitem realizar análises complexas de dados e implementar algoritmos de aprendizagem de máquina com maior facilidade e eficiência.

Sintetizando os objetivos do trabalho, dois experimentos serão propostos para cada problemática (sujidade e sombreamento). O primeiro experimento conterá os 499 pontos de tensão e de corrente, mais irradiação e temperatura do módulo, totalizando 1000 variáveis preditoras. Já o segundo experimento consiste em reduzir o tamanho de variáveis preditoras

para evitar problemas de multicolinearidade e para melhorar a agilidade no treinamento dos modelos.

3.2.1 Python para Análise de Dados

Atualmente, ferramentas que possam lidar com grandes quantidades de dados com facilidade e rapidez é uma necessidade. Considerando o crescimento das aplicações de aprendizagem de máquina, é importante utilizar as ferramentas que permitem ler grandes quantidades de informações, limpá-las e processá-las para uso. O Python se adapta ao projeto devido as suas bibliotecas e ferramentas. Além de ser uma linguagem flexível e intuitiva, ela contém ferramentas propícias para a análise de dados.

Dentre os motivos principais para a escolha do Python para o projeto está a biblioteca *pandas*. Nela estão contidas funções orientadas a colunas que possibilitam o processamento de dados de forma simples. Dentre as funcionalidades úteis estão aquelas ligadas à leitura e escrita de dados em formato estruturado (ou formato tabular), facilitando a leitura dos arquivos ".csv".

Outras bibliotecas que serão utilizadas são as relacionadas à visualização de dados, como é o exemplo do *seaborn* e do *matplotlib*. Estas bibliotecas são úteis, não só na etapa de avaliação dos modelos, mas também na análise exploratória de dados. Este processo é crítico na etapa precedente ao da limpeza dos dados, pois permite descobrir padrões, detectar anomalias, testar hipóteses e conhecer estatisticamente o conjunto de dados.

Os treinamentos dos estimadores serão realizados utilizando a biblioteca *pycaret*. Esta biblioteca possui código aberto e seu objetivo é automatizar o fluxo de trabalho dos treinamentos dos algoritmos de aprendizagem de máquina. De forma simples, através do *pycaret* se consegue criar e avaliar modelos disponíveis em outras bibliotecas – XGBoost, scikit-learn, LightGBM, CatBoost, entre outras.

3.2.2 Separação dos Dados

Para avaliar o desempenho dos modelos que estão sendo treinados, é necessário realizar alguns procedimentos adicionais. Primeiramente, os dados serão divididos em conjuntos de treinamento e teste. Esse passo é crucial para garantir que as métricas finais possam ser avaliadas de forma independente dos dados de treinamento, aumentando a confiabilidade da avaliação da generalização dos modelos.

No trabalho, dividiu-se o banco de dados em 80% para amostras de treinamento e 20% para amostras de validação. O conjunto de validação será utilizado para avaliar a capacidade do modelo em generalizar novos dados que não foram utilizados no treinamento. Dessa forma, consegue-se simular um comportamento em situações desconhecidas.

3.2.3 Normalização

Alguns algoritmos de aprendizagem de máquina, como a regressão logística e as redes neurais, que utilizam a descida do gradiente como uma técnica de otimização necessitam passar por uma normalização das variáveis preditoras. Nos algoritmos de distância, como no KNN e o SVM, a normalização também afetará os resultados. Estes modelos são mais afetados pela variação da escala das variáveis preditoras, pois seus treinamentos se baseiam no cálculo de distância entre pontos de dados para determinar a semelhança entre eles.

Como as variáveis irradiação e temperatura estão em diferentes escalas, existe uma chance de o modelo considerar um maior peso à variável irradiação (por ter maior magnitude). Este procedimento, portanto, evita que os algoritmos baseados em distância resultem em enviesamento para variáveis de maiores magnitudes.

A padronização será a técnica aplicada de normalização, em que consiste em subtrair a média de cada valor da distribuição e, depois, divide-se pelo desvio padrão. Como resultado da aplicação da padronização, cada variável predita será transformada de modo a possuírem média igual a zero e desvio padrão igual a 1. Destaca-se que a aplicação do escalonamento não interfere na distribuição das variáveis e será utilizada para comparar outros algoritmos, pois os modelos baseados em árvore são imunes ao escalonamento. A equação (10) descreve o cálculo da padronização, em que μ é a média da variável e σ é seu desvio padrão.

$$z = \frac{x - \mu}{\sigma} \tag{14}$$

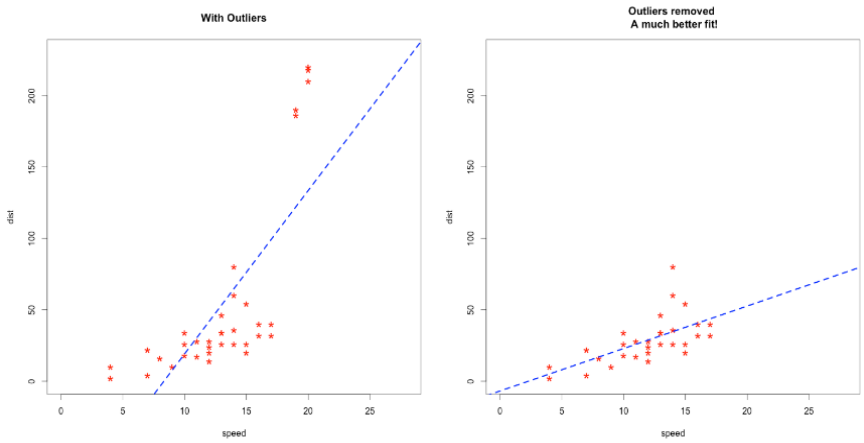
3.2.4 Tratamento de Outliers

Um *outlier* consiste em uma observação que se diferenciam drasticamente em uma amostra de uma população. São considerados valores que fogem da normalidade de observação que poderão causar anomalias nos padrões do conjunto de dados. Antes de lidar com os *outliers*

é necessário saber a sua origem. Existem três classificações para os *outliers*: erro de medição experimental; variação natural e de amostragem.

A presença de *outliers* pode impactar negativamente a análise exploratória, pois, devido a alterações na média e no desvio padrão do conjunto de dados, podem apresentar distribuições inconformes com outras amostras da população. Consequentemente, os *outliers* influenciam na redução da normalidade do conjunto de dados. Alguns algoritmos são severamente afetados pela presença de *outlier*, a Figura 17 descreve as diferenças no treinamento de uma regressão linear.

Figura 17 - Exemplo de remoção de *outliers* em uma regressão linear.



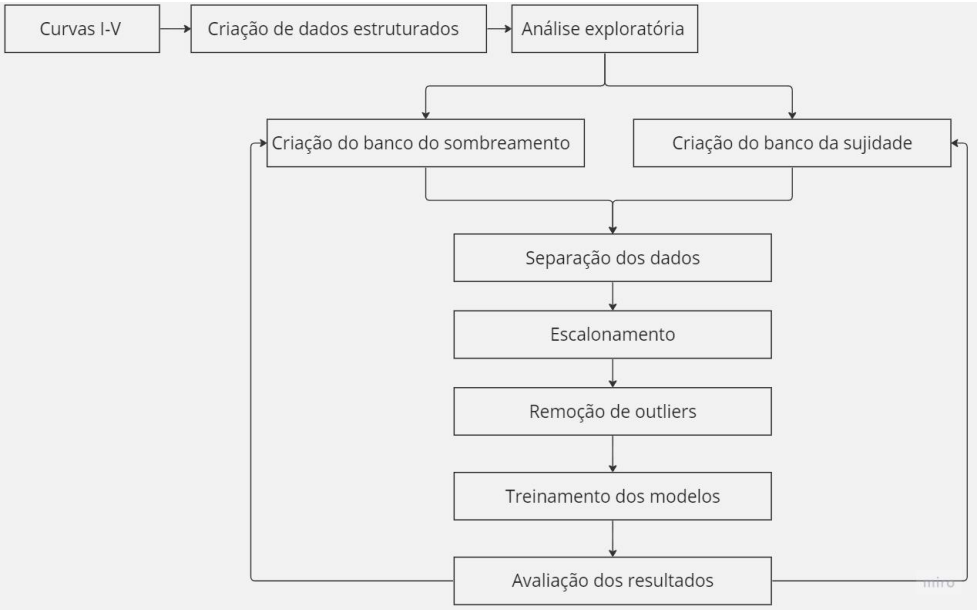
Fonte: (R-STATISTICS)[2]

3.2.5 Processamento de Dados

A Figura 18 sintetiza toda etapa de processamento de dados até a avaliação dos resultados modelo. De acordo com os resultados obtidos, é possível alterar parâmetros de configuração dos algoritmos e manipular variáveis preditoras no objetivo de alcançar um maior desempenho. Assim, o processo se torna cíclico até se encontrar resultados satisfatórios para os treinamentos.

[2] Disponível em: http://r-statistics.co/Outlier-Treatment-With-R.html

Figura 18 - Fluxo de processamento de dados para ambas problemáticas.



Fonte: Autor.

4 RESULTADOS E DISCUSSÃO

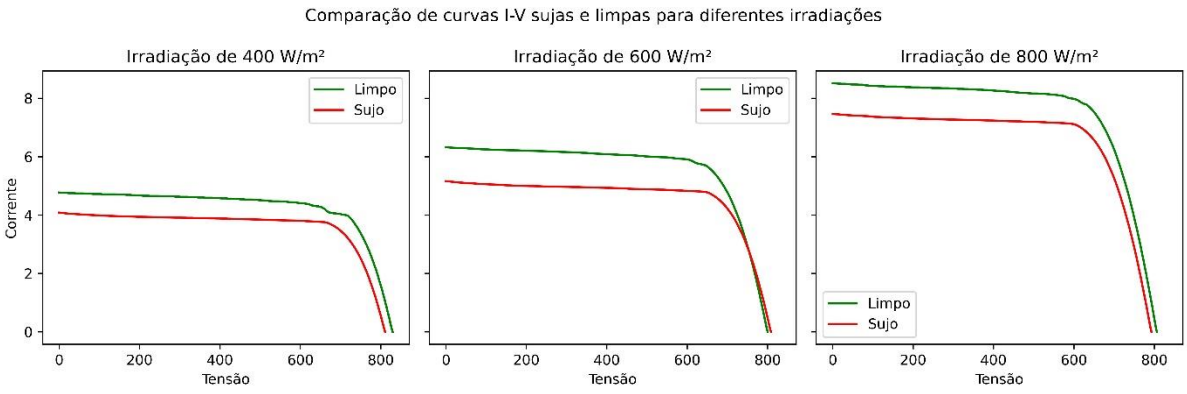
4.1 DETECÇÃO DE SUJIDADE

4.1.1 Influência da Sujidade e da Irradiação nas Curvas I-V

Através da análise exploratória de dados, é possível reconhecer alguns padrões de comportamento acerca dos módulos sujos. Com os recursos de bibliotecas de visualização do Python, é possível verificar as relações existentes entre as curvas sujas e limpas.

Na análise das influências da irradiação nas medições de curva I-V, como comentado na seção 2.1.2, é possível verificar um acréscimo na corrente de curto-circuito de acordo com a quantidade de irradiação incidida sobre o módulo. Isto se evidencia através da Figura 19 em que são mostradas as curvas I-V com diferentes perfis de irradiação. O efeito esperado de aumento da corrente de curto-circuito acontece, bem como existe uma distinção entre as curvas de cada classe predita.

Figura 19 - Relações entre a corrente, tensão e temperatura na análise de sujidade.



Fonte: Autor.

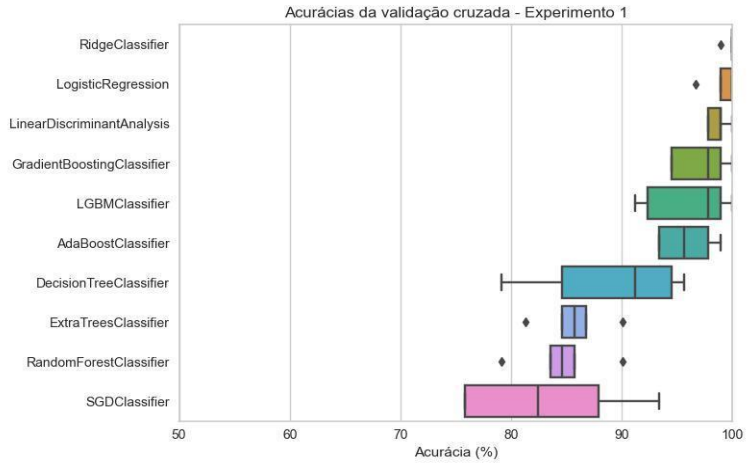
4.1.2 Experimento 1

O primeiro experimento realizado propõe o treinamento do modelo contendo 998 pontos de medição entre corrente e tensão (449 pontos de cada). Nota-se que o tamanho de variáveis preditoras (1000) é significativamente alto, podendo algumas não serem relevantes para o modelo. Além disso, muitas destas variáveis possuem correlação alta entre si, podendo ocasionar em problemas de multicolinearidade.

A multicolinearidade acontece quando uma variável preditora em um modelo de regressão pode ser linearmente predita por outras com alto grau de acurácia. Por natureza, os modelos baseados em árvores de decisão são imunes a este efeito por dividirem seus nós de acordo com apenas uma das variáveis. Porém, outros algoritmos são sensíveis a este problema e tendem a ter melhores resultados após a remoção de variáveis linearmente dependentes.

Em diversos treinamentos, o LGBM se destacou como um dos modelos com melhor capacidade de generalização. Como característica, esse modelo apresenta treinamentos rápidos e resulta em acurácias maiores do que a árvore de decisão, por exemplo, em função da construção de árvores mais complexas. Além disso, utilizando o LGBM é possível, de forma fácil, é possível analisar as variáveis preditoras que possuem maior peso para as classificações do problema. Na Figura 20 é possível comparar a acurácia de treinamento para diferentes modelos de aprendizagem de máquina aplicado ao mesmo conjunto de dados.

Figura 20 - Acurácias da validação cruzada do experimento 1.



Fonte: Autor.

No treinamento, como acurácia média da validação cruzada, o algoritmo LGBM alcançou 97,58%. Para avaliar se o modelo possui boa generalização é utilizado o conjunto de validação. Aqui, alcançou-se 98,25% de acurácia. Outras métricas de avaliação como revocação e F1-score (98,3%) indicam que não ocorreu sobre ajuste aos dados de treinamento.

4.1.3 Experimento 2

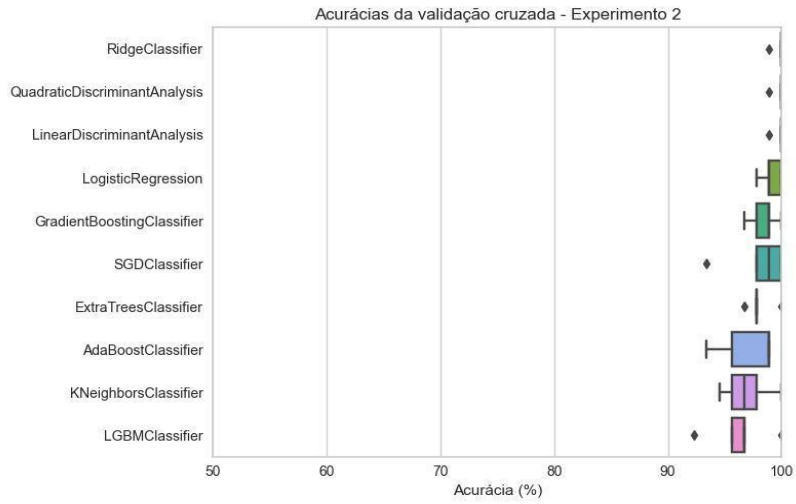
Como forma de reduzir a multicolinearidade, algumas variáveis preditoras foram filtradas de forma a reduzir a correlação entre as variáveis. Verificando as variáveis que mais

se destacam no impacto do treinamento do modelo, pode-se inferir que algumas delas podem ser filtradas de forma a não comprometerem os resultados finais. Com análises de importância das variáveis preditoras, as responsáveis por impactarem mais nas predições do modelo são: temperatura, irradiação, “corrente_1” (referente a I_{SC}) e “tensão_499” (V_{OC}), que serão utilizadas no experimento 2.

A etapa seguinte consiste em treinar os modelos empregando o novo conjunto de dados com apenas quatro variáveis. Como resultado, evidencia-se o ganho na acurácia deste experimento, podendo ser visto através de treinamentos que resultam em menores desvios padrão, isto é, indicando que as métricas de cada dobra resultaram em valores próximos a média.

De acordo com o experimento 2, obteve-se uma média de acurácia na validação cruzada de 96,0% e no conjunto de validação de 96,5%. As Figuras 20 e 21 comparam as acurácias de 10 diferentes modelos para os experimentos 1 e 2.

Figura 21 - Acurácias da validação cruzada do experimento 2.



Fonte: Autor.

Como síntese, a Tabela 2 busca reunir as acurácias obtidas na validação cruzada e no conjunto de validação do LGBM e do melhor modelo encontrado para a problemática. A demonstração de resultados já indica uma performance satisfatória na classificação da sujidade.

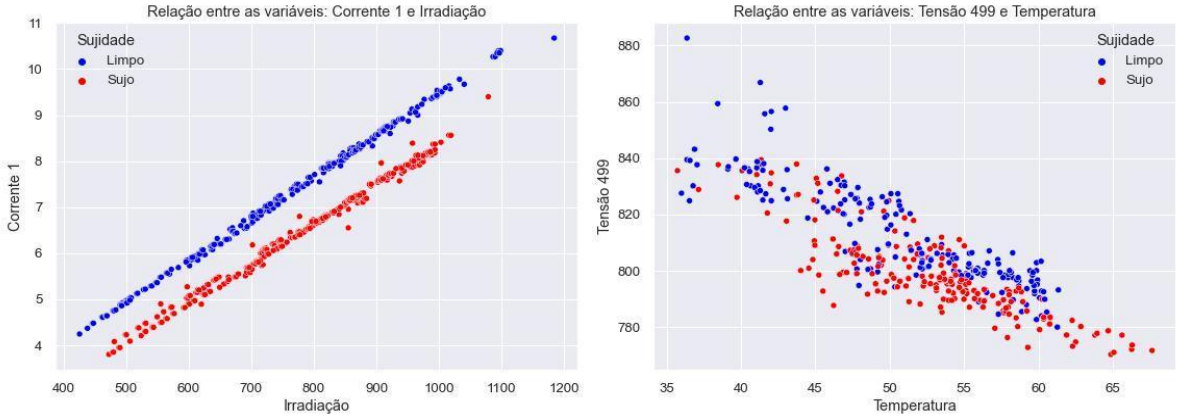
Tabela 2 - Resultados dos modelos RidgeClassifier e LGBM de ambos experimentos.

Experimento	**Modelo**	**Acurácia (validação cruzada)**	**Acurácia (conjunto de validação)**
1	LGBM	97,58%	98,25%
1	RidgeClassifier	99,78%	100%
2	LGBM	97,36%	96,49%
2	RidgeClassifier	99,78%	100%

Fonte: Autor.

Ao verificar as relações entre as variáveis filtradas, encontram-se algumas relações que podem ajudar o modelo a ter resultados satisfatórios. De acordo com a Figura 22, na relação entre a I_{SC} e a irradiação, percebe-se uma região de separação evidente entre os módulos sujos e limpos. Esta é a relação mais importante na predição da sujidade. Porém, ainda é possível identificar uma região de separação quando analisada a relação entre o último ponto de tensão e a temperatura, por mais que essa relação seja menos impactante para o modelo. Em cada gráfico da figura está representada a relação entre cada par de variável preditora das 569 amostras.

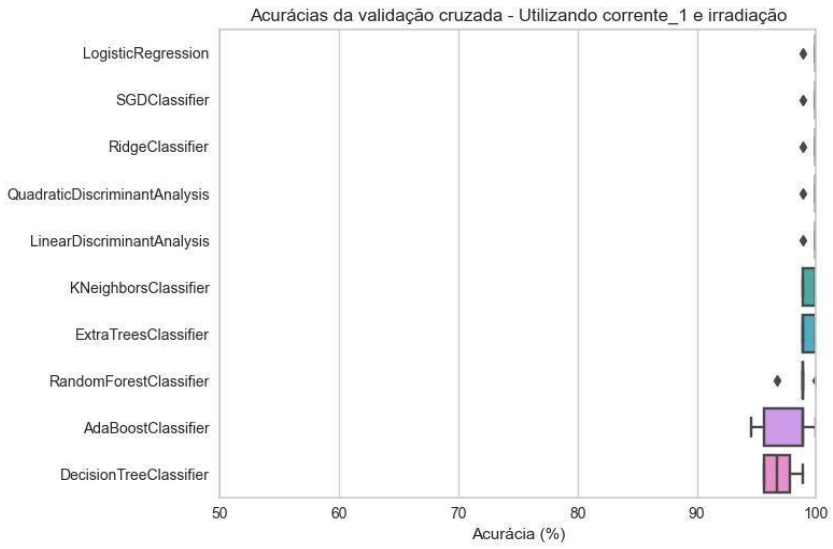
Figura 22 - Relações entre variáveis de acordo com a sujidade.



Fonte: Autor.

Filtrando apenas as variáveis de irradiação e do primeiro ponto de medição da corrente, resulta-se nos treinamentos conforme a Figura 23, que indica que somente estas variáveis já são suficientes para predizer com alta acurácia a sujidade do módulo FV.

Figura 23 - Performance dos modelos de sujidade utilizando apenas corrente e irradiação.



Fonte: Autor.

Na medida que o problema se torna mais simples com a redução de variáveis, a acurácia do LGBM é incrementada, porém outros métodos de aprendizagem de máquina se tornam mais viáveis. Fica mais evidente isso se a acurácia do LGBM for comparada a de outros algoritmos que não fazem uso de árvores de decisão e que possuem menor complexidade. O incremento das acurácias pode ser observado nas Tabelas 3 e 4, em que a primeira tabela estão dispostas as acurácias do experimento 1, enquanto na segunda tabela estão dispostas as acurácias utilizando apenas irradiação e “corrente_1” como variáveis.

Tabela 3 - Comparações de acurácias do experimento 1.

Algoritmo	Acurácia (treinamento)	Acurácia (validação)
LGBM	96,04%	96,49%
SVM	93,85%	96,49%
KNN	66,59%	67,54%
Regressão logística	99,12%	100%

Fonte: autor.

Tabela 4 - Comparações de acurácias utilizando irradiação e “corrente_1”.

Algoritmo	Acurácia (treinamento)	Acurácia (validação)
LGBM	96,48%	99,12%
SVM	99,56%	100%
KNN	99,56%	100%
Regressão logística	99,78%	100%

Fonte: autor.

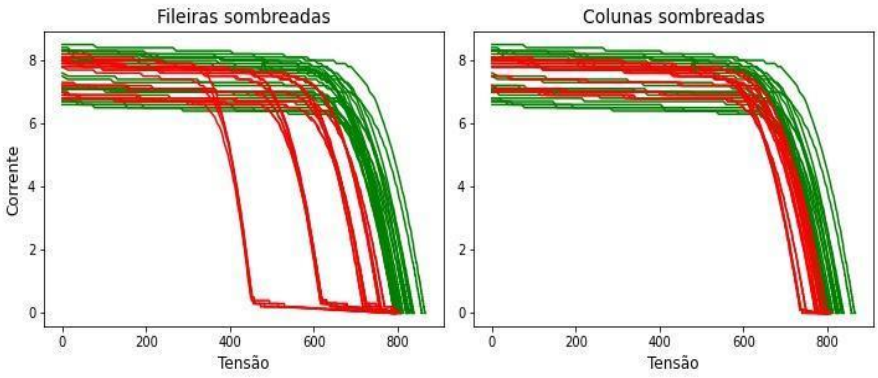
4.2 DETECÇÃO DE SOMBREAMENTO

4.2.1 Influência do Sombreamento nas Curvas I-V

A análise que se apresenta nas Figuras 24 25 26 tem como premissa o uso de amostras que possuem medições de irradiação entre 800 e 900 W/m². Aqui busca-se entender as influências do sombreamento nas curvas I-V que reduzirá a tensão de circuito aberto do módulo. Outro ponto importante nesta análise é que se deseja comparar a classe não sombreada com relação às demais. O impacto é maior quando uma classe não sombreada é predita erroneamente, pois evita-se que operadores de manutenção sejam mobilizados para corrigir uma falha que não existe de fato.

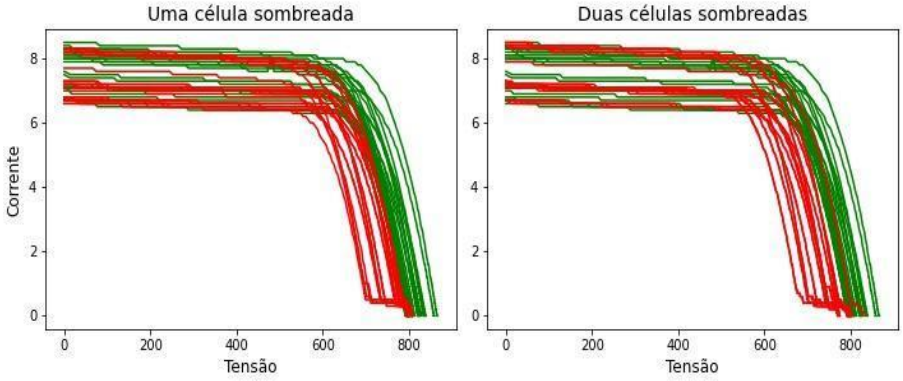
Nas figuras em questão, as curvas em verde representam as amostras que não possuem nenhum tipo de sombreamento, enquanto as curvas vermelhas estão relacionadas aos sombreamentos dos ensaios. Nas comparações das figuras, dois pontos chamam a atenção: a redução de P_{max} e de V_{OC}. A Figura 24 busca comparar as curvas I-V de módulos não sombreados com fileiras e colunas sombreadas. Na Figura 25, que busca comparar as curvas com células sombreadas, pode-se notar visualmente uma menor redução de V_{OC} e P_{max} quando comparadas às curvas com fileiras e colunas sombreadas. O sombreamento parcial, visualizado na Figura 26, será a maior dificuldade de classificação para os algoritmos devido à similaridade com os módulos não sombreados. Portanto, espera-se que a maior quantidade de classificações erradas seja nessa classe.

Figura 24 - Comparação de módulos não sombreados com fileiras e colunas sombreadas.



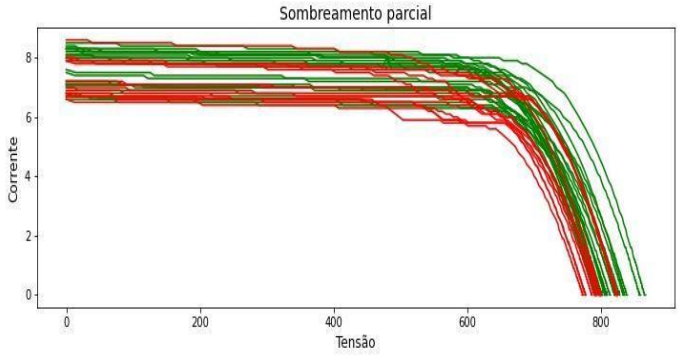
Fonte: Autor.

Figura 25 - Comparação de módulos não sombreados com uma ou duas células sombreadas.



Fonte: Autor.

Figura 26 - Comparação de módulos não sombreados com sombreamento parcial.



Fonte: Autor.

Verifica-se que já existe uma certa sobreposição das curvas quando comparadas as amostras que possuem fileiras ou colunas sombreadas com as curvas com apenas células sombreadas. A similaridade das curvas indica que a maior dificuldade do modelo será classificar o sombreamento parcial com o não sombreamento.

4.2.2 Composição do Banco de Dados

Um ponto a se atentar é que o banco de dados utilizado para a detecção de sombreamento possui um desbalanceamento acerca das classes conforme a Tabela 5. Mesmo que seja pouco desbalanceamento, pode existir interferência no resultado final, principalmente nas predições dos ensaios 3 e 5.

Tabela 5 - Contagem de amostras por ensaio.

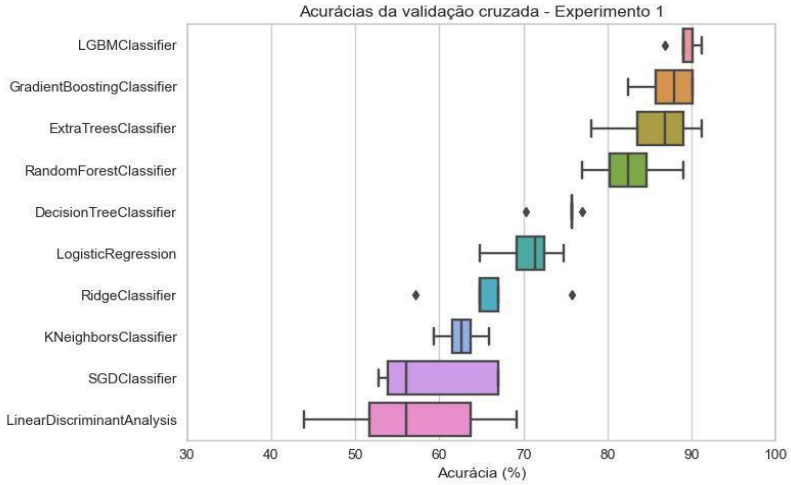
Ensaio	Número de amostras
Ensaio 1 – Sem sombreamento	107
Ensaio 2 – Sombreamento em uma fileira	100
Ensaio 3 – Sombreamento em uma coluna	88
Ensaio 4 – Sombreamento em uma célula	95
Ensaio 5 – Sombreamento em duas células	81
Ensaio 6 – Sombreamento parcial de uma célula	98

Fonte: Autor.

4.2.3 Experimento 1

O primeiro experimento da detecção do tipo de sombreamento indicou que o LGBM é o modelo que melhor consegue generalizar o problema com uma acurácia na validação cruzada de 89,89%. Outros modelos obtiveram resultados semelhantes ao LGBM, porém a maioria deles não conseguiram obter acurácias relevantes. Na Figura 27 é possível observar que alguns dos modelos resultaram em treinamentos fracos nas iterações da validação cruzada.

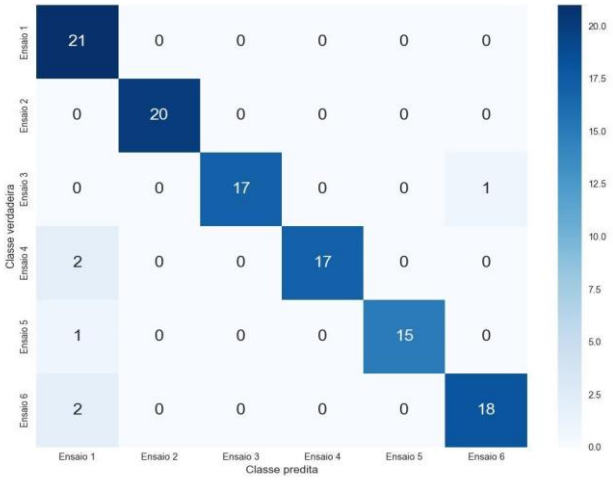
Figura 27 - Performance da validação cruzada de modelos para a predição de sombreamento.



Fonte: Autor.

Como próximo passo, o conjunto de validação foi utilizado como forma de analisar a generalização do problema para novos dados. O resultado pode ser visto através da matriz de confusão da Figura 28. Nesta validação, obteve-se acurácia de 89,47%.

Figura 28 - Matriz de confusão do LGBM utilizando 6 classes.



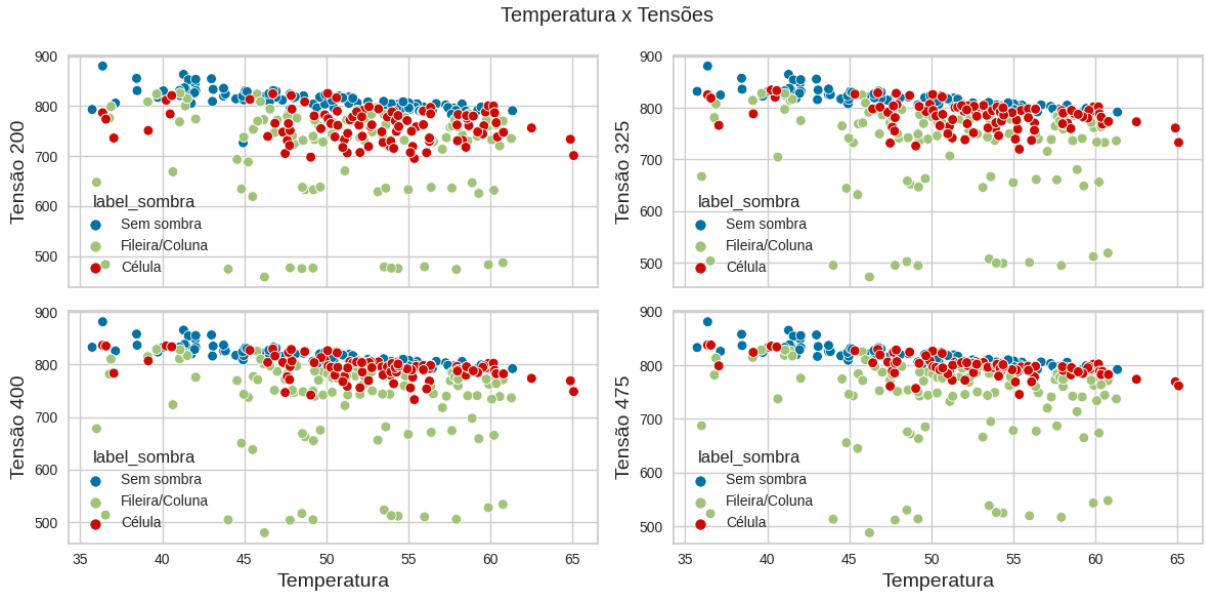
Fonte: Autor.

4.2.4 Experimento 2

Com a finalidade de facilitar a interpretação do treinamento do modelo LGBM, decidiu-se agrupar em menor quantidade as classes preditas originais. Dessa forma, foram agrupadas as classes de acordo com: as classes “Fileiras” e “Colunas” em “Fileiras/Colunas”; as classes “1 Célula” e “2 Células” em “Células”; as classes “Parcial” e “Sem sombra” permaneceram iguais. Portanto, o novo número de classes será quatro. O sombreamento foi agrupado em graus similares de forma a passar uma noção de gravidade do problema.

Uma análise que mostra a similaridade do comportamento de algumas variáveis pode ser realizada, em que a Figura 29 demonstra a relação da temperatura com as variáveis “tensao_200”, “tensao_325”, “tensao_400” e “tensao_475. As variáveis 200, 325, 400 e 475 representam a medição de tensão no respectivo ponto de tempo. Dessa forma, a variável “tensao_200” se encontra antes da metade da curva, enquanto a “tensao_475” se apresenta como um dos últimos pontos de medição (próximo do V_{OC}).

Figura 29 - Gráfico de dispersão da relação entre a temperatura e quatro tensões.



Fonte: Autor.

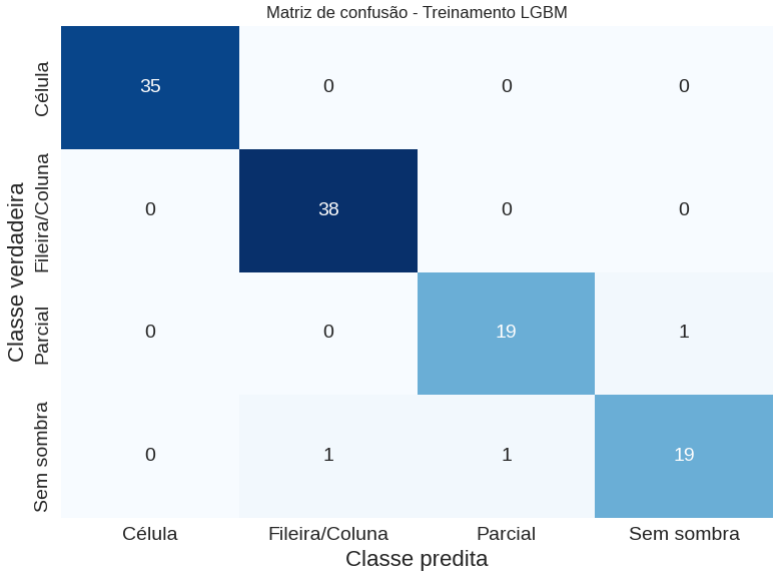
O número grande de variáveis dificulta a análise e a visualização dos dados. Para lidar com esse problema, optou-se por reestruturar o conjunto de dados, selecionando um subconjunto de variáveis por meio de *slicing*.

O *slicing* é uma técnica para selecionar uma parte de uma sequência, por exemplo como uma lista na linguagem *python*. A sintaxe para o slicing é "*start*:*stop*:*step*", onde *start* é o índice de início, *stop* é o índice de parada e *step* é o tamanho do passo entre os elementos.

No estudo, optou-se por selecionar um subconjunto de variáveis a cada 10 colunas, utilizando um *slicing* com intervalo de 10. Como resultado, reduziu-se significativamente o número de variáveis preditoras para 102 (considerando a irradiação e a temperatura). Ao realizar a reamostragem, observou-se que variáveis com comportamentos mais diversos surgiram ao analisar o impacto delas no modelo.

No novo treinamento do LGBM, obteve-se uma acurácia de 94,73% na validação cruzada e de 96,49% nos dados de validação. A matriz de confusão das predições está disposta na Figura 30.

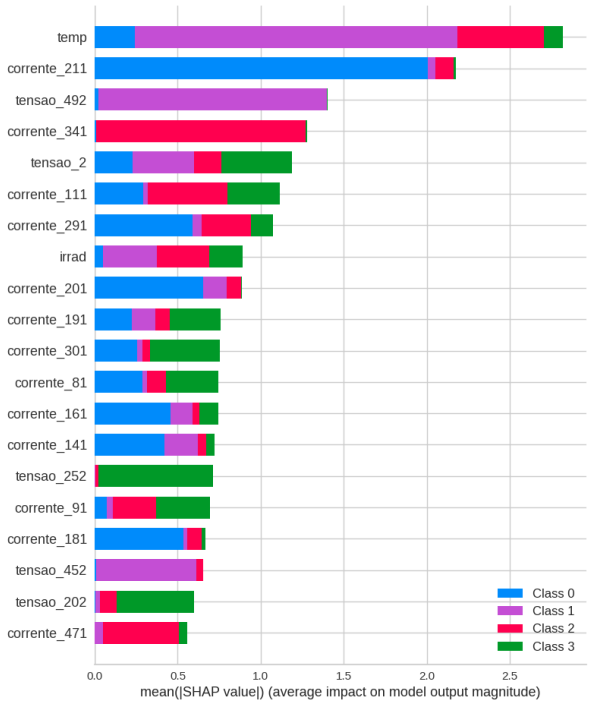
Figura 30 - Matriz de confusão do LGBM utilizando 4 classes.



Fonte: Autor.

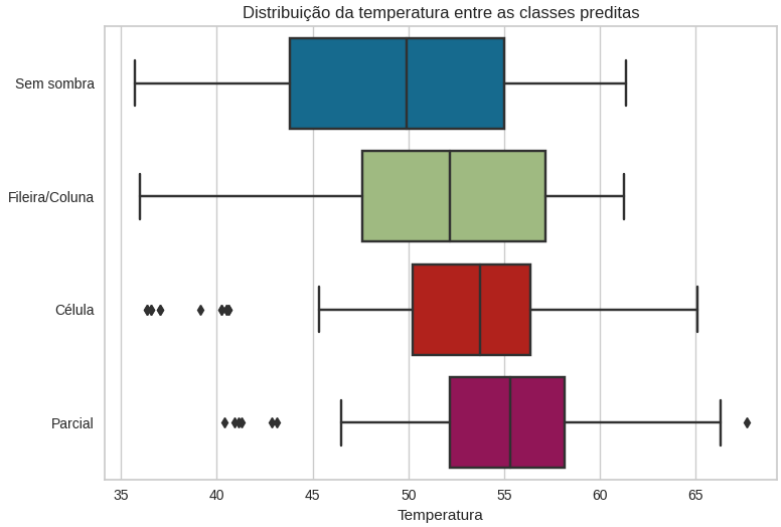
Verificou-se que a variável temperatura vinha sendo considerada a mais importante para o LGBM, conforme a Figura 31. Além disso, a distribuição da temperatura de cada classe, que pode ser visualizada na Figura 32, possui divergências. As amostras sem sombreamento, por exemplo, apresentam medições em níveis menores de temperatura.

Figura 31 - Impacto das variáveis no modelo.



Fonte: Autor.

Figura 32 - Distribuição da temperatura entre as classes preditas.

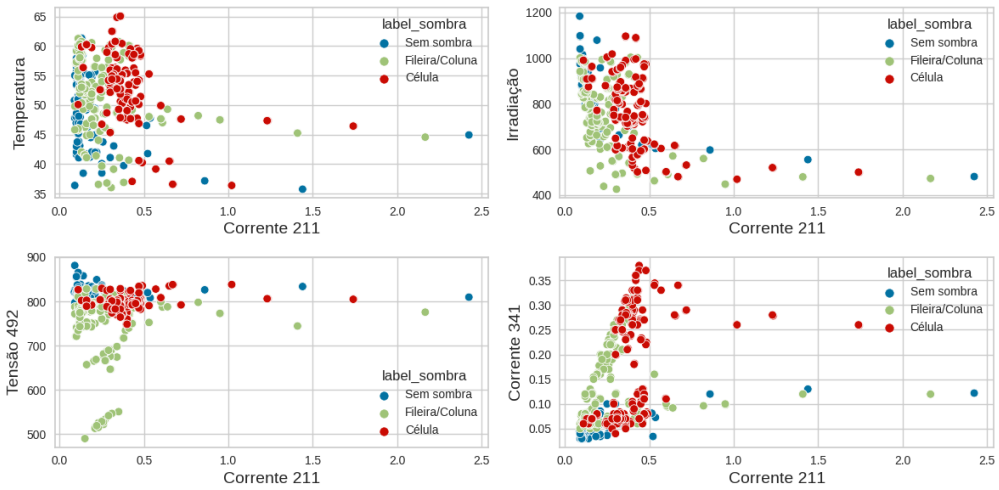


Fonte: Autor.

A complementação com novas amostras de curvas I-V com maior diversidade na temperatura auxiliará o modelo a estabelecer melhores regiões de separação.

Particularmente a variável “corrente_211” possui uma grande importância para a “Class 1”. Devido às limitações da biblioteca *pycaret*, não se tem informações de qual classe é referida. Porém de acordo com a Figura 33, ao mostrar a relação da corrente_211 com algumas outras, destaca-se sempre a distinção da classe de células sombreadas.

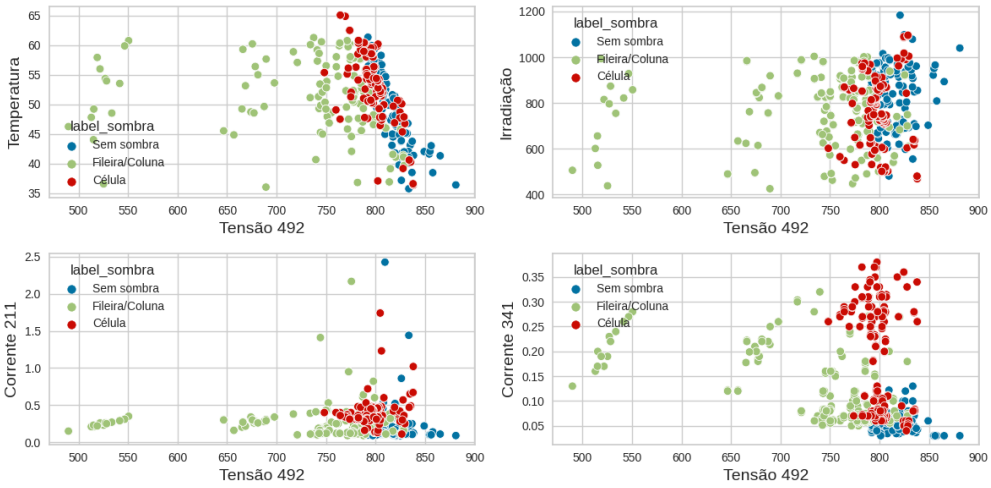
Figura 33 - Relações da "corrente_211" com outras variáveis.



Fonte: Autor.

De forma análoga, a variável “tensao_492” consegue obter uma região de separação clara da classe com fileiras sombreadas conforme a Figura 34.

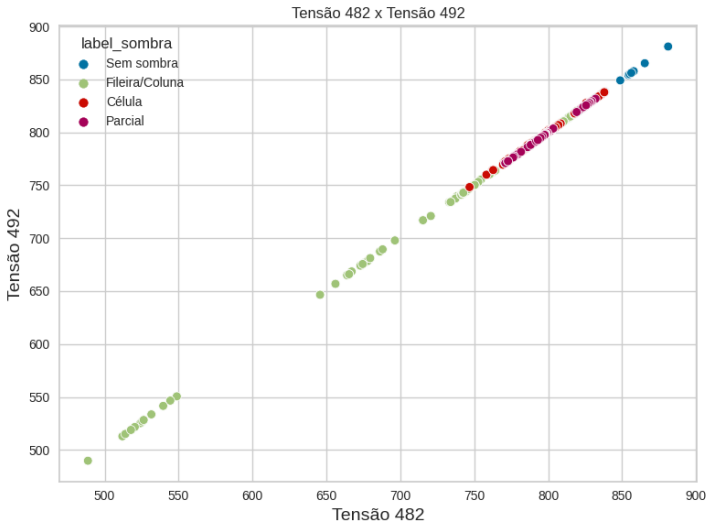
Figura 34 - Relações da "tensao_492" com outras variáveis.



Fonte: Autor.

Em razão do efeito do sombreamento na redução da V, outra relação que ajuda o modelo a distinguir as classes é entre o último valor de tensão (“tensao_492”) e do penúltimo valor (“tensao_482”). Como esperado, existirão valores menores de tensão conforme o grau de severidade do sombreamento. Esta relação pode ser vista na Figura 35, em que as classes não sombreada e de fileiras sombreadas possuem uma clara distinção. Novamente é possível analisar o efeito do sombreamento em $V_{OC,}$ em que são obtidos menores valores de tensão em maiores graus de sombreamento.

Figura 35 - Gráfico de dispersão das variáveis "tensao_482" e "tensao_492".



Fonte: Autor.

Ao utilizar a ténica de *slicing* para reduzir a dimensionalidade do problema, outros algoritmos foram treinados para serem comparados com os resultados do LGBM. O objetivo é determinar se o LGBM se manterá como o modelo com melhor capacidade de generalização neste problema. A simplificação proporcionada pelo *slicing* pode permitir que modelos não baseados em árvores de decisão obtenham resultados superiores, devido à exclusão de variáveis com alta correlação. Assim, as Tabelas 6 e 7 demonstram as acurácias das comparações.

Tabela 6 - Comparações de acurácias do experimento 1.

Algoritmo	Acurácia (validação cruzada)	Acurácia (conjunto de validação)
LGBM	93,63%	96,49%
SVM	93,85%	97,37%
KNN	77,80%	85,09%
Regressão logística	86,15%	89,47%

Fonte: autor.

Tabela 7 - Comparações de acurácias do experimento 2.

Algoritmo	Acurácia (validação cruzada)	Acurácia (conjunto de validação)
LGBM	94,73%	96,49%
SVM	94,29%	98,25%
KNN	79,56%	86,84%
Regressão logística	82,64%	85,96%

Fonte: autor.

Verifica-se que existe uma melhora marginal de acurácia com a redução de variáveis. O mais notável é que a acurácia consegue se manter, ao mesmo tempo que se consegue reduzir o custo computacional para ser treinado em máquinas mais simples. Para a detecção do tipo de sombreamento, o LGBM se manteve como o modelo com maior acurácia na validação cruzada, porém obtendo acurácia menor que o SVM no conjunto de validação.

Com o objetivo de melhorar a detecção do tipo de sombreamento, através da análise gráfica das dispersões entre variáveis é possível estimar que modelos de aprendizagem profunda possam ser utilizados. Isso se deve ao fato de que são padrões complexos que os algoritmos baseados em redes neurais possuem mais facilidade de aprender.

Uma técnica que pode alcançar acurácias superiores ao LGBM é da classificação utilizando Redes Neurais Convolucionais (CNN) juntamente com uma transformação *Gramian Angular Field* (GAF). A maior vantagem de aplicar o GAF seria em transformar as séries temporais de corrente e tensão em representações bidimensional que preservam a temporalidade

da medição. Outra possibilidade de ser aplicada devido a estrutura do conjunto de dados é a classificação utilizando o algoritmo *Long Short Term Memory* (LSTM), que, em razão das suas conexões de *feedback*, também consegue capturar dependências temporais. Um exemplo disso é que esse algoritmo está sendo muito utilizado na literatura em aplicações de classificação de doenças com sinais de eletroencefalogramas. Os algoritmos de *deep learning* possuem uma capacidade melhor de lidar com dados sequenciais, como é o exemplo dos dados utilizados para treinamento no trabalho.

A detecção de sombreamento pode desempenhar um papel fundamental na otimização dos custos de O&M de sistemas FV. Ao realizar a detecção, é possível identificar áreas especificas do sistema que estão sendo afetadas, permitindo uma localização mais ágil dos painéis afetados. Essa abordagem evita a propagação da perda de energia para outras partes do sistema.

Uma das maiores vantagens de distinguir o tipo de sombreamento reside na capacidade de identificar o grau de severidade do sombreamento incidido sobre o módulo. Caso exista uma recorrência na detecção do sombreamento, o planejamento de medidas preventivas, como a poda de árvores, por exemplo, pode ser realizado. Essas ações proativas contribuem para minimizar a ocorrência de sombreamento futuro e, consequentemente, melhorar o desempenho e a eficiência do sistema.

Uma estratégia que pode ser eficaz para minimizar o sombreamento em painéis FV é com o monitoramento periódico das curvas I-V da planta. Essa abordagem envolve a realização de medições regulares que, juntamente com a detecção de sombreamento e de sujidade, podem ser utilizados para analisar as variáveis das curvas ao longo do tempo, permitindo a identificação de padrões recorrentes na produção de energia.

Com base nas séries históricas das medições, é possível estabelecer um limite de perda aceitável de produção para determinar o momento adequado de agir. A informação da perda pode ser facilmente definida com a diferença entre P_{MAX} – que pode ser uma medida empírica dos valores máximos das séries – e a potência medida. Diferentes ações podem ser tomadas como a poda de uma árvore ou a realocação do painel.

5 CONCLUSÃO

O trabalho propôs utilizar técnicas de análise e processamento de dados para resolver dois problemas de classificação por meio do algoritmo LGBM. O primeiro problema consiste na classificação do tipo de sombreamento incidente sobre os painéis solares. O segundo problema se caracteriza pela classificação binária da sujidade dos módulos, buscando predizer se há ou não sujeira. Para ambas as tarefas, o LGBM foi implementado e comparado com outros algoritmos. A escolha do LGBM se deu a partir da sua habilidade em aprender padrões complexos nos dados e pela disponibilidade de analisar a importância das variáveis de treinamento.

O banco de dados inicial utilizado possui medições de curvas I-V, temperatura e irradiação. Uma estrutura do conjunto de dados foi criada de modo a satisfazer os requisitos dos treinamentos comumente aceitados em bibliotecas *python*. O conjunto de dados final resultou em 998 pontos de corrente e tensão – sendo 499 cada – e dos valores de irradiação e temperatura, sendo o total de 569 amostras com indicações de sujidade e do tipo de sombreamento incidente sobre o painel.

Durante o primeiro experimento da detecção de sujidade, o LGBM alcançou acurácia de 98,25%. Pôde-se notar que o modelo treinado infere grandes pesos para as variáveis irradiação e $I_{SC.}$ O experimento 2 foi proposto para facilitar a análise da relação entre as variáveis preditoras do LGBM. Como resultado, o LGBM alcançou 96,49% de acurácia, sendo inferior ao primeiro experimento. De forma geral, outros modelos melhoraram sua performance, com alguns deles ultrapassando o LGBM. Durante a análise gráfica da distribuição das classes, é possível notar uma região evidente de separação entre as variáveis I_{SC} e irradiação. De acordo com as predições deste conjunto de dados, existem modelos treinados com apenas essas duas variáveis que foram suficientes para alcançar 100% de acurácia nos dados de validação.

A classificação do tipo de sombreamento possui uma dificuldade maior por se tratar de uma classificação com múltiplas classes que são semelhantes entre si. Inicialmente, foi proposto no primeiro experimento a classificação de 6 tipos de sombreamento, resultando em 89,47% de acurácia com o LGBM. Embora o modelo tenha apresentado boa capacidade de predição, foi observado um alto custo computacional que resultou em treinamentos demorados. Para simplificar, no experimento 2, optou-se por agrupar as classes em apenas 4 devido à similaridade do grau de sombreamento entre elas. Uma reamostragem também foi utilizada para reduzir a dimensão do problema. Como resultado, o LGBM alcançou 97,37% de acurácia

(contra 96,49% do experimento 1 considerando 4 classes). Devido a simplificação, foi possível estimar algumas variáveis que possuem bom poder preditivo, de forma a ilustrar a complexidade das regiões de separação.

Diante disso, para projetos futuros, a aplicação de algoritmos baseados em aprendizagem profunda pode melhorar a detecção, juntamente com uma amostragem mais diversa de temperatura para diferentes graus de sombreamento. Também é necessário validar os resultados em diferentes arranjos FV, considerando que diferentes associações podem afetar o desempenho preditivo do modelo treinado. Por fim, é possível propor uma análise de tempo ótimo de manutenção devido a algum obstáculo natural. Dessa forma, o objetivo é otimizar operações de manutenção estabelecendo um limite máximo aceitável de perda de geração.

REFERÊNCIAS

ALESSANDRINI, Stefano; MONACHE, Luca Delle; SPERATI, Simone; CERVONE, Guido. **An analog ensemble for short-term probabilistic solar power forecast**. [S. l]: Applied Energy, v. 157, p. 95-110, nov. 2015.

ALMEIDA, Eliane et al. **Energia solar fotovoltaica: revisão bibliográfica**. [S. l.]: E. Fumec, v. 1, 2016. Disponível em: <http://revista.fumec.br/index.php/eol/article/view/3574>.

BARBOSA, Elismar Ramos; FARIA, Merlim dos Santos Ferreira; Gontijo, Fabio de Brito. **Influência da sujeira na geração fotovoltaica**. Gramado, RS, VII Congresso Brasileiro de Energia Solar, abr. 2018.

BATRA, Mridula; AGRAWAL, Rashmi. **Comparative analysis of decision tree algorithms**. Springer Singapore, Singapore, p. 31-36, 2018.

BENTÉJAC, Candice; CSÖRGO, Anna; MARTÍNEZ-MUÑOZ, Gonzalo. **A comparative analysis of xgboost**. Univeristy of Bordeaux, France, 2019.

CLEMENTE-HARDING, Laura; CERVONE, Guido; ALESSANDRINI, Stefano; MONACHE, Luca Delle. **Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble**. [S. l.]: Renewable Energy, v. 108, p. 274-286, 2017.

COUTINHO, Carlos Roberto; NARDOTO, Adriano Fazolo; PROVETI, José Rafael Cápua; COURA, Daniel José Custódio. **Efeito do sombreamento em módulos fotovoltaicos**. Belo Horizonte, MG, VI Congresso Brasileiro de Energia Solar, abr. 2016.

EMPRESA DE PESQUISA ENERGÉTICA. **Balanço energético nacional**. Rio de Janeiro, 2021. Disponível em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-675/topico-638/BEN2022.pdf>.

EMPRESA DE PESQUISA ENERGÉTICA. **Projeção da demanda de energia elétrica para os próximos 10 anos**. Rio de Janeiro, 2017.

EXAME. **Fonte solar avança, mas ainda há potencial para ser explorado**. 24 jun. 2022. Disponível em: <https://exame.com/bussola/fonte-solar-fotovoltaica-avanca-mas-ainda-ha-potencial-a-ser-explorado/>.

HÜBNER, Guilherme Ricardo. **Diagnóstico de desequilíbrio de massa para rotores de aerogeradores utilizando máquina de vetores de suporte**. Universidade Federal de Santa Maria, Santa Maria, RS, 2021.

KE, Guolin; MENG, Qi; FINLEY, Thomas; WANG, Taifeng; CHEN, Wei; MA, Weigong; YE, Qiwei; LIU, Tie-Yan. **Lightgbm: a highly efficient gradient boosting decision tree**. [S. l.]: Curran Associates, Inc., v. 30, 2017.

LI, Baojie; DELPHA, Claude; MIGAN-DUBOIS, Anne; DIALLO, Demba. **Fault diagnosis of photovoltaic panels using full i-v characteristics and machine learning techniques**. [S. l.]: Energy Conversion and Management, 2021.

MING, Du; WANG, Shu-mei; GONG, Gu. **Research on decistion tree algorithm based on information entropy**. China: Advanced Materials Research, v. 267, p. 732-737, jun. 2011.

NREL. **Renewable energy: an Overview**. 2001. Disponível em: <https://www.nrel.gov/docs/fy01osti/27955.pdf>.

ONS. **Avaliação das condições de atendimento eletroenergético do sistema interligado nacional**. Rio de Janeiro, Brasil, 2021. Disponível em: <https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/NT-ONS%20DGL%200093-2021%20-%20Estudo%20Prospectivo%20Agosto-Novembro_VF%20(1).pdf>.

PAPAGEORGAS, Panagiotis et al. **A low-cost and fast pv i-v curve tracer based on an open source platform with m2m communication capabilities for preventive monitoring**. [S. l.]: Energy Procedia, v. 74, p. 423-438, 2015.

PINHO, João Tavares; GALDINO, Marco Antônio. **Manual de engenharia para sistemas fotovoltaicos**. Rio de Janeiro: Cepel, Cresesb, mar. 2014.

QU, Zongxi; ZHANG, Kequan. **Research and application of ensemble forecasting based on a novel multi-objective optimization algorithm for wind-speed forecasting**. [S. l.]: Energy Conversion and Management, v. 154, p. 440-454, 2017.

REN21. **Renewables 2021: global status report. 2021**. Disponível em: <https://www.ren21.net/wp-content/uploads/2019/05/GSR2021_Full_Report.pdf>.

ROKACH, Lior; MALMON, Oded. **Decision trees**. The Data Mining and Knowledge Discovery Handbook, v. 6, p. 165-192, jan. 2005.

RÜTHER, Ricardo. **Edifícios solares fotovoltaicos**. Universidade Federal de Santa Catarina, Florianópolis, 2004.

SHAH, Rahul. **Tune hyperparameters with gridsearchcv**. [S. l.], jun. 2021. Disponível em: <https://www.analyticsvidhya.com/blog/2021/06/tune-hyperparameters-with-gridsearchcv/>.

TORTELLI, Carla. **O efeito do sombreamento na eficiência do sistema fotovoltaico do escritório verde da UTFPR**. Curitiba, PR: Universidade Tecnológica Federal do Paraná, 2016.
TRETER, Marcos. Monitoramento e avaliação de desempenho integrado a usinas fotovoltaicas. Joinville, 2022.

WANG, Zhanshan; LU, Huchuan; TANG, Huajin. **Advances in Neural Networks**. Moscou, Russia: 16th International Symposium on Neural Networks, jul. 2019.

WILLOUGHBY, Alexander; OSINOWO, Muritala. **Development of an electronic load i-v curve tracer to investigate the impact of harmattan aerossol loading on pv module pern2tkformance in southwest Nigeria**. Nigeria: Solar Energy, v. 166, 2018.

ZHANG, Wenyu; WANG, Jujie; WANG Jianzhou; ZHAO, Zengbao; TIAN, Meng. **Short-term wind-speed forecasting based on a hybrid model**. [S. l]: Applied Soft Computing, v. 13, n. 7, p. 3225-3233, 2013.

ANEXO A – CÓDIGO DE PROCESSAMENTO DAS CURVAS I-V

```
from pathlib import Path

import numpy as np
import pandas as pd
from tqdm import tqdm


def get_single_value_infos(dframe: pd.DataFrame) -> tuple:

    temp = dframe.at[0, 'tempModulo']
    irrad = dframe.at[0, 'irrad']

    return temp, irrad

def get_csvs_list(path: Path, child_folder: str) -> list:

    return list((path / child_folder).glob('**/*.csv'))

def extract_timeseries(csv_dir: str, label_sombra: int) -> list:

    dframe = pd.read_csv(csv_dir)
    temp, irrad = get_single_value_infos(dframe)

    label_sujeira = csv_dir.__str__().split('\\')[-3]
    if 'Série FV limpa' in label_sujeira:
        label_sujidade = 'Limpo'
    else:
        label_sujidade = 'Sujo'

    dframe = dframe.drop(0)

    tensao_values = dframe.tensao.values.tolist()
```

```
    corrente_values = dframe.corrente.values.tolist()

    output_dict = {}

    row_values = zip(tensao_values, corrente_values)

    for i, values in enumerate(row_values):
        output_dict[f'tensao_{i}'] = values[0]
        output_dict[f'corrente_{i}'] = values[1]

    output_dict['temp'] = temp
    output_dict['irrad'] = irrad
    output_dict['label_sombra'] = label_sombra
    output_dict['label_sujidade'] = label_sujidade

    return output_dict

root_data_dir = Path('./data/ensaios/')

ensaios_paths = {
    'ensaio_1': get_csvs_list(root_data_dir, child_folder='Ensaio 1 - Sem sombreamento'),
    'ensaio_2': get_csvs_list(root_data_dir, child_folder='Ensaio 2 - Fileiras de células sombreadas'),
    'ensaio_3': get_csvs_list(root_data_dir, child_folder='Ensaio 3 - Colunas de células sombreadas'),
    'ensaio_4': get_csvs_list(root_data_dir, child_folder='Ensaio 4 - Uma célula sombreada por módulo'),
    'ensaio_5': get_csvs_list(root_data_dir, child_folder='Ensaio 5 - Duas células sombreadas por módulo'),
    'ensaio_6': get_csvs_list(root_data_dir, child_folder='Ensaio 6 - Sombreamento parcial de duas células por módulo'),
}

processed_data = []
```

```
for ensaio, csvs_dirs in tqdm(ensaios_paths.items()):
    for csv_dir in csvs_dirs:
        processed_data.append(extract_timeseries(csv_dir, label_sombra=ensaio))

df = pd.DataFrame(processed_data)
```

ANEXO B – CÓDIGO DE TREINAMENTO DOS MODELOS

```
from google.colab import drive

import pandas as pd
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split
import lightgbm as lgbm
from pycaret.classification import *

drive.mount('/content/gdrive/', force_remount=True)

df = pd.read_parquet('gdrive/MyDrive/TCC/dataset_final_sem_reamostragem.parquet')

mapping = {
    'Ensaio 1': 'Sem sombra',
    'Ensaio 2': 'Fileira/Coluna',
    'Ensaio 3': 'Fileira/Coluna',
    'Ensaio 4': 'Célula',
    'Ensaio 5': 'Célula',
    'Ensaio 6': 'Parcial',
}

df['label_sombra'] = df['label_sombra'].replace(mapping)

# Configuração inicial do experimento
exp = setup(
    data=df,
    target='label_sombra',
    ignore_features=['label_sujeira'],
    train_size=0.8,
    fold=5,
    data_split_shuffle=True,
    data_split_stratify=True,
```

```
    remove_perfect_collinearity=False,
    remove_outliers=True,
    normalize=True,
    session_id=42,
    use_gpu=True
)

# Comparação de modelos
from sklearn.svm import SVC

models = ['lightgbm', SVC(C=100), 'knn', 'lr']

models_results = []

for model in tqdm(models):
    clf = create_model(model)
    res_train = pull()
    res_train_acc = res_train.loc['Mean', 'Accuracy']

    preds = predict_model(clf)
    res_test = pull()
    res_test_acc = res_test.loc[0, 'Accuracy']

    models_results.append(
        {'model': model.__str__(), 'res_train_acc': res_train_acc, 'res_test_acc': res_test_acc}
    )

df_results = pd.DataFrame(models_results)



# Treinamento do LGBM padrão
lgbm = create_model('lightgbm')
```

```
res = predict_model(lgbm)

plot_model(lgbm, 'confusion_matrix')
plot_model(lgbm, 'feature')
interpret_model(lgbm, 'summary')
```