

**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**

**LUCAS ROSSO TOKARSKI  
SOFIA MASSANEIRO**

**ANÁLISE DE SÉRIES TEMPORAIS POR MACHINE LEARNING PARA ESTUDO E  
COMPARAÇÃO DA GERAÇÃO DE ENERGIA POR DIFERENTES TECNOLOGIAS  
DE MÓDULOS FOTOVOLTAICOS DO LABENS DO CAMPUS UTFPR-CT.**

**CURITIBA  
2025**

**LUCAS ROSSO TOKARSKI  
SOFIA MASSANEIRO**

**ANÁLISE DE SÉRIES TEMPORAIS POR MACHINE LEARNING PARA ESTUDO E  
COMPARAÇÃO DA GERAÇÃO DE ENERGIA POR DIFERENTES TECNOLOGIAS  
DE MÓDULOS FOTOVOLTAICOS DO LABENS DO CAMPUS UTFPR-CT.**

**Analysis of Time Series with Machine Learning for the Study and Comparison  
of Energy Generation by Different Photovoltaic Module Technologies at  
LABENS of the UTFPR-CT Campus.**

Trabalho de Conclusão de Curso de Graduação,  
apresentado como requisito para obtenção do título  
de Bacharel(a) em Engenharia Elétrica do curso de  
Engenharia Elétrica da Universidade Tecnológica  
Federal do Paraná (UTFPR).

Orientadora: Profa. Dra. Angélica Maria Tortola  
Ribeiro

Coorientador: Prof. Dr. Gerson Máximo Tiepolo

**CURITIBA  
2025**



[4.0 Internacional](https://creativecommons.org/licenses/by/4.0/)

Esta licença permite compartilhamento, remixe, adaptação e criação a partir do trabalho, mesmo para fins comerciais, desde que sejam atribuídos créditos ao(s) autor(es). Conteúdos elaborados por terceiros, citados e referenciados nesta obra não são cobertos pela licença.

**LUCAS ROSSO TOKARSKI**  
**SOFIA MASSANEIRO**

**ANÁLISE DE SÉRIES TEMPORAIS POR MACHINE LEARNING PARA ESTUDO E  
COMPARAÇÃO DA GERAÇÃO DE ENERGIA POR DIFERENTES TECNOLOGIAS  
DE MÓDULOS FOTOVOLTAICOS DO LABENS DO CAMPUS UTFPR-CT.**

Trabalho de Conclusão de Curso de Graduação,  
apresentado como requisito para obtenção do título  
de Bacharel(a) em Engenharia Elétrica do curso de  
Engenharia Elétrica da Universidade Tecnológica  
Federal do Paraná (UTFPR).

Data de aprovação: 17/novembro/2025

---

Angélica Maria Tortola Ribeiro  
Doutorado  
Universidade Tecnológica Federal do Paraná

---

Gerson Máximo Tiepolo  
Doutorado  
Universidade Tecnológica Federal do Paraná

---

Jair Urbanetz Junior  
Doutorado  
Universidade Tecnológica Federal do Paraná

---

José Carlos Pereira Coninck  
Doutorado  
Universidade Tecnológica Federal do Paraná

**CURITIBA**

**2025**

## **AGRADECIMENTOS**

Agradecemos a nossa orientadora, Profa. Dra. Angélica Maria Tortola Ribeiro, pela sabedoria com que nos guiou nesta trajetória. Seu conhecimento e sua didática são diferenciais, e auxiliaram não somente com o desenvolvimento deste trabalho, mas também com nosso desenvolvimento pessoal.

Agradecemos também aos professores Gerson Maximo Tiepolo e Jair Urbanetz Junior, que orientaram, corrigiram e apoiaram nosso trabalho, especialmente nos detalhes que deixamos passar e/ou em equívocos que foram despercebidos por nós.

Agradecemos imensamente aos nossos colegas e amigos, sejam estes do curso, dos estágios ou externos, que torceram por nosso sucesso e auxiliaram em nossos momentos de alívio de tensão.

Eu, Lucas, também agradeço grandemente a minha namorada, Sophia, que acompanhou cada etapa deste trabalho, me apoiou em todos os momentos difíceis, ouviu todas as minhas reclamações, comentários, desabafos, e acima de tudo torceu pelo êxito deste trabalho e pela conclusão do curso superior.

Gostaríamos de deixar registrado também, os nossos reconhecimentos às nossas famílias, pois acreditamos que, sem o apoio deles, seria muito difícil vencer esse desafio.

## RESUMO

Em virtude da representatividade da geração solar fotovoltaica no Brasil, com grande potência instalada e geração, é importante para o Operador Nacional do Sistema (ONS) ter visibilidade da estimativa de geração de energia por essa fonte, para conseguir planejar os processos de expansão, manutenção, controle e substituição de usinas. Nesse sentido, o presente trabalho estima a geração de energia entre os meses de janeiro e setembro de 2022, da planta instalada pelo LABENS (Laboratório de Energia Solar), no campus Neoville, da UTPFR-CT. Para isso, foram avaliados modelos de *Machine Learning (ML)*, *Random Forest (RF)* e KNN (*K-Nearest Neighbors*), para prever a potência gerada pelos sistemas mono e policristalinos. O desenvolvimento das análises de ML contemplou o estudo da importância de variáveis, para o algoritmo RF. Pelos resultados foi possível notar que a irradiância foi a covariável de maior relevância para o modelo, enquanto o tipo de tecnologia foi menos relevante. Apesar disso, o sistema monocristalino apresentou resultados maiores para a estimativa de energia. O ajuste para ambos os modelos foi alto, sendo que o KNN superou o Random Forest, com menor raiz do erro quadrático médio e maior coeficiente de correlação, métricas utilizadas para averiguar o desempenho dos ajustes.

Palavras-chave: Previsão; Energia; Fotovoltaica; Aprendizado de máquina.

## ABSTRACT

Given the significance of solar photovoltaic generation in Brazil, with its large installed capacity and generation, it is important for the National System Operator (ONS) to have visibility into the estimated energy generation from this source, in order to plan the processes of expansion, maintenance, control, and replacement of power plants. In this context, the present study estimates the energy generation between the months of January and September 2022 from the plant installed by LABENS (Laboratório de Energia Solar), on the Neoville campus of UTPFR-CT. For this, Machine Learning (ML) models, Random Forest (RF) and KNN (*K-Nearest Neighbors*), were evaluated to predict the power generated by monocrystalline and polycrystalline systems. The development of the ML analyses included the study of variable importance for the RF algorithm. From the results, it was possible to note that irradiance was the covariate of greatest relevance for the model, while the technology type was less relevant. Despite this, the monocrystalline system showed higher results for the energy estimation. The fit for both models was high, with KNN outperforming Random Forest, showing a lower root mean square error and a higher correlation coefficient, metrics used to verify the performance of the fits.

Keywords: Forecasting; Power; Photovoltaic; Machine learning.

## LISTA DE ILUSTRAÇÕES

Figura 1 - Evolução da capacidade total instalada de potência fotovoltaica .....	17
Figura 2 - Evolução da produção de módulos fotovoltaicos de diferentes tecnologias .....	18
Figura 3 - Processo de formação do lingote de silício .....	18
Figura 4 - Células fotovoltaicas das tecnologias: a) silício monocristalino; b) silício policristalino .....	19
Figura 5 - Esquemático de um sistema fotovoltaico conectado à rede (SFVCR) .....	20
Figura 6 - Espalhamento da irradiância em suas componentes .....	21
Figura 7 - Exemplo de um gráfico <i>BoxPlot</i> .....	24
Figura 8 - (a) Dispersão de $S_{xy}$ positiva; (b) Dispersão de $S_{xy}$ negativa .....	26
Figura 9 - Exemplo de um diagrama de ordem de decisão .....	27
Figura 10 - Diagrama de funcionamento do método <i>Random Forest</i> .....	29
Figura 11 - Diagrama de busca do método KNN, para $k = 3$ .....	30
Figura 12 - Histograma da variável pac-mon1CT .....	43
Figura 13 - Histograma da variável pac-pol2CT .....	43
Figura 14 - Comparação entre os <i>boxplots</i> das duas tecnologias analisadas .....	45
Figura 15 – Diagrama de dispersão entre pac-mon1CT vs tilt_avg-so01CT .....	46
Figura 16 – Diagrama de dispersão entre pac-pol2CT vs tilt_avg-so01CT .....	46
Figura 17 - Diagrama de dispersão entre mono-tempCT vs pac-mon1CT .....	46
Figura 18 - Diagrama de dispersão entre poli-tempCT vs pac-pol2CT .....	46
Figura 19 - Resultados da métrica IncNodePurity .....	47
Figura 20 - Resultado da predição pelo algoritmo Random Forest .....	48
Figura 21 - Dispersão da potência prevista para as duas tecnologias, utilizando o método <i>Random Forest</i> .....	50
Figura 22 - Gráfico de comparação da estimativa de energia média com a energia média real, para a tecnologia monocristalina pelo modelo Random Forest .....	51
Figura 23 – Gráfico de comparação da estimativa de energia média com a energia média real, para a tecnologia policristalina pelo modelo Random Forest .....	51
Figura 24 - Desvio percentual por mês de análise para o sistema monocristalino .....	52
Figura 25 - Desvio percentual por mês de análise para o sistema policristalino .....	53
Figura 26 - Resultado da predição pelo algoritmo KNN .....	54
Figura 27 - Dispersão da potência prevista para as duas tecnologias, utilizando o método KNN .....	54

<b>Figura 28 - Gráfico de comparação da estimativa de energia média com a energia real, para a tecnologia monocristalina, pelo modelo KNN .....</b>	<b>55</b>
<b>Figura 29 - Gráfico de comparação da estimativa de energia média com a energia real, para a tecnologia policristalina, pelo modelo KNN .....</b>	<b>56</b>
<b>Figura 30 - Desvio percentual por mês de análise para o sistema monocristalino .....</b>	<b>57</b>
<b>Figura 31 - Desvio percentual por mês de análise para o sistema policristalino .....</b>	<b>57</b>



## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>9</b>
<b>1.1</b>	<b>Tema</b>	<b>9</b>
1.1.1	Delimitação do tema	11
<b>1.2</b>	<b>Problemas e premissas</b>	<b>11</b>
<b>1.3</b>	<b>Objetivos</b>	<b>12</b>
1.3.1	Objetivo geral	12
1.3.2	Objetivos específicos	13
<b>1.4</b>	<b>Justificativa</b>	<b>13</b>
<b>1.5</b>	<b>Abordagem metodológica e metodologia</b>	<b>14</b>
<b>1.6</b>	<b>Estrutura do trabalho</b>	<b>14</b>
<b>2</b>	<b>REVISÃO DE LITERATURA</b>	<b>16</b>
<b>2.1</b>	<b>Análise fotovoltaica</b>	<b>16</b>
<b>2.2</b>	<b>Estatística descritiva</b>	<b>22</b>
<b>2.3</b>	<b>Análise de correlação</b>	<b>25</b>
<b>2.4</b>	<b><i>Machine Learning</i></b>	<b>26</b>
<b>3</b>	<b>METODOLOGIA</b>	<b>31</b>
<b>3.1</b>	<b>Análise descritiva dos dados</b>	<b>31</b>
3.1.1	Análise descritiva do sistema físico instalado	31
3.1.2	Análise descritiva do banco de dados	34
<b>3.2</b>	<b>Análise estatística dos dados</b>	<b>38</b>
<b>3.3</b>	<b>Análise por <i>Machine Learning</i></b>	<b>39</b>
<b>4</b>	<b>RESULTADOS E DISCUSSÕES</b>	<b>42</b>
<b>4.1</b>	<b>Análises estatísticas básicas</b>	<b>42</b>
<b>4.2</b>	<b>Análise via <i>Machine Learning</i></b>	<b>47</b>
<b>5</b>	<b>CONCLUSÃO</b>	<b>59</b>
	<b>REFERÊNCIAS</b>	<b>62</b>

## 1 INTRODUÇÃO

Neste capítulo são apresentados os conceitos iniciais sobre o tema, “Análise de séries temporais e Machine Learning para estudo e comparação da geração de energia por diferentes tecnologias de módulos fotovoltaicos do LABENS (Laboratório de Energia Solar) do campus UTFPR-CT”.

Apresentam-se também os objetivos, justificativa e estrutura do projeto, a fim de proporcionar uma visão inicial sobre o tema e a sua importância.

### 1.1 Tema

A temática ambiental e energética sempre foi de grande interesse social e político, visto a importância que refletem no cotidiano das pessoas. Nesse contexto, a busca por fontes de energia renováveis tem recebido destaque, principalmente pelo seu rápido crescimento e potencial produtivo.

Dessa forma, as matrizes energéticas de vários países têm sido alteradas, buscando maior eficiência energética (REAd, 2017). No caso do Brasil, de acordo com dados do Balanço Energético Nacional (BEN, 2025), a maior parte de toda a geração de energia do país ainda é hídrica, representando cerca de 55,3% do total - valor que tem decrescido a cada ano, com a expansão de outras fontes renováveis, como eólica, biomassa e solar. Conforme os dados, em 2024, a fonte solar fotovoltaica atingiu 41 TWh e 35,9 TW de geração e potência instalada, respectivamente, em termos de micro e minigeração distribuída, enquanto em 2023 (BEN, 2024), foram gerados 29,8 TWh e instalados 26,4 MW - representando um crescimento de 37,6 % na geração de energia e de 36,4 % na capacidade instalada, de um ano para o outro.

A partir dos dados apresentados anteriormente, nota-se uma maior representatividade da energia solar fotovoltaica, muito relacionado ao potencial energético solar do país, considerando sua localização privilegiada no planeta.

No entanto, existem outros tipos de tecnologias que podem ser aplicadas para aproveitamento de energia solar. Dentre elas, podem-se citar os coletores solares térmicos, responsáveis pelo aquecimento de fluidos, a partir dos raios infravermelhos e a geração indireta de eletricidade. A geração indireta trata-se de mecanismos responsáveis por concentrar o calor que transforma a água em vapor, para que seja transformado em eletricidade por meio da ação de turbinas e de um gerador. Por outro

lado, a geração de eletricidade direta consiste em utilizar o sistema de módulos fotovoltaicos a fim de transformar a faixa de luz visível em energia útil para consumo.

Essa última forma de geração de energia segue o princípio do efeito fotovoltaico, o qual acontece quando um material semicondutor (silício, por exemplo) é incidido pela luz solar, fazendo com que os elétrons sejam excitados (Chapin; Fuller; Pearson, 1954). Como consequência dessa agitação, gera-se corrente elétrica de natureza contínua, a qual é então convertida, por um inversor, para uma corrente elétrica alternada, com tensão e frequência compatíveis com a rede à qual o sistema estará conectado.

Dependendo do nível de radiação solar incidente na superfície dos módulos, a produtividade energética varia, principalmente durante as estações do ano. Em meses com menor tempo diário de sol, além de menores valores de irradiação, como no inverno, assim como em períodos de maior nebulosidade, a produtividade é menor, enquanto nos meses de verão, a produção de energia é maior. Fatores como a irradiação, a temperatura e inclinação do módulo, por exemplo, interferem na produtividade do sistema instalado (Sameera; Tariq; Rihan, 2024).

Além disso, a escolha da tecnologia dos módulos, sendo de silício monocristalino ou policristalino, também é uma variável importante nas considerações de geração de energia (CBENS, 2020). Somando todas as variáveis externas é possível entender o comportamento energético de um sistema fotovoltaico (ZAMPIVA, 2024). Nesse sentido, são válidas as análises estatísticas que trabalham combinando covariáveis, a fim de identificar aquelas de maior interferência em um conjunto de dados, que podem ser utilizadas para analisar o comportamento de algum fenômeno, considerando a passagem de tempo.

Para isso, destacam-se as análises de séries temporais. Uma série temporal é um conjunto de dados que estão dispostos e organizados no tempo, e sua análise procura identificar as repetições de ciclos, bem como a tendência de ocorrência de algum tipo de evento (Montgomery; Jennings; Kulahci, 2007). Com isso, são dispostos modelos estatísticos que explicam com maior confiabilidade uma determinada ocorrência. Dessa forma, é possível embasar as melhores decisões e prever comportamentos a partir de dados anteriores.

Dessa forma, serão relacionados os conceitos e metodologias estatísticas, de análise de séries temporais com o *machine learning*, para o estudo aprofundado de produtividade de sistemas fotovoltaicos. Foi escolhida a usina de módulos

fotovoltaicos instalada pelo LABENS, na sede Neville, do campus UTFPR-CT. Foram selecionados os painéis compostos por células de silício monocristalino e policristalino, a fim de inferir seu comportamento em um espaço de tempo e comparar suas produtividades.

#### 1.1.1 Delimitação do tema

A pesquisa estudará, por meio de *machine learning*, a produtividade das plantas fotovoltaicas de silício monocristalino e policristalino, instaladas pelo LABENS, na sede Neville, do campus UTFPR-CT.

*Machine Learning* é uma das classes da Inteligência Artificial, e visa ensinar aos computadores a modelar um conjunto de dados, por meio de regressão e redes neurais, por exemplo, e gerar decisões a partir da repetição de cenários (Neha; Yadav; Goyal; 2024).

A proposta de pesquisa definirá, a partir de parâmetros estatísticos, a melhor tecnologia para instalação, bem como a tendência de produtividade com o passar do tempo.

Dessa forma, pretende-se estimar valores futuros de geração de energia para cada tipo de tecnologia construtiva, que sirvam como parâmetros decisores para a escolha de tecnologia de módulo mais adequada para a instalação de um sistema fotovoltaico, assim como para auxílio de órgãos reguladores do sistema elétrico quanto ao controle de geração e despacho de carga.

### 1.2 Problemas e premissas

Durante a elaboração de um projeto de um sistema fotovoltaico (micro ou minigeração, por exemplo), é importante que haja uma estimativa de geração de energia que será alcançada após sua instalação, de modo a prever o retorno sobre o investimento (*payback*), assim como para análise da necessidade de expansão ou substituição de painéis fotovoltaicos, seja por maiores capacidades de potência de saída ou para tecnologias diferentes, que podem ser mais rentáveis. No entanto, estas estimativas são aproximações, usualmente realizadas considerando médias diárias/mensais de geração de energia obtidas previamente, e a falta da consideração de alguns parâmetros, como a tecnologia empregada na construção dos módulos, acaba tornando estas estimativas imprecisas, o que não é desejável para um estudo

ou análise mais detalhada de um sistema fotovoltaico, visto que podem ocorrer graves erros de orçamentos e especificações construtivas (elétricas e/ou mecânicas), os quais podem resultar em prejuízos tanto econômicos quanto materiais. Além da micro e minigeração, as aplicações de maior escala também requerem um controle preciso da geração de energia futura, de modo que a operação do sistema elétrico seja realizada de forma mais confiável e com menor impacto à rede.

Tendo isto em vista, é possível realizar uma análise estatística de variáveis e covariáveis (como temperatura, radiação e umidade, por exemplo) de dois sistemas fotovoltaicos que utilizam as duas diferentes tecnologias construtivas de silício cristalino em seus painéis, de modo a obter métodos precisos para estimar a geração de energia de cada sistema e realizar comparações e análises que possam fornecer maior grau de precisão e confiabilidade.

As amostras são relativas a fenômenos físicos e variam conforme as horas e dias, tem-se uma base de dados ordenada de forma cronológica, que dá origem a uma série temporal. Uma série temporal é o conjunto de dados que contém informações obtidas de forma temporal, ou seja, dentro de um espaço de tempo amostras de determinadas covariáveis são coletadas de forma periódica (Montgomery; Jennings; Kulahci, 2007). De forma geral, essas amostras são dependentes e refletem em algum fenômeno prático, logo, entender essa(s) interação(ões) é o objeto de estudo da análise de séries temporais (Box; Jenkins; Ljung; Reinsel, 2016).

Os estudos de *Machine Learning* são aplicados a fim de identificar padrões em um conjunto de dados e trabalhar com séries temporais. Os métodos que analisam uma grande base de dados podem ser chamados de *data mining*, no qual o conjunto é processado por meio de algoritmos e aprendizado da máquina, para gerar um modelo simples de descrição dos dados, de alta precisão (Alpaydin, 2004).

## 1.3 Objetivos

### 1.3.1 Objetivo geral

O objetivo geral deste trabalho é comparar, através de métodos de análise estatística, duas tecnologias construtivas de módulos fotovoltaicos - silício

monocristalino e silício policristalino – de uma usina fotovoltaica da UTFPR-CT, e analisar as diferentes variáveis que afetam a geração de energia destas tecnologias.

### 1.3.2 Objetivos específicos

Considerando o que foi analisado no objetivo geral, considera-se as seguintes etapas com a finalidade de atingir este objetivo:

- Realizar diferentes análises de importância das variáveis (irradiância média, temperatura na superfície dos módulos, tecnologia construtiva), com relação à potência, e analisar a correlação entre as mesmas utilizando ferramentas computacionais e métodos estatísticos;
- Aplicar diferentes modelos de *Machine Learning*, de modo a gerar previsões de potência média, e comparar a eficiência de cada modelo (*Random Forest* e KNN) em relação aos valores reais medidos, para cada as tecnologias dos sistemas fotovoltaicos;
- Obter resultados de previsão de energia média, comparar estes com os valores reais e discorrer sobre a importância da obtenção desses dados para o planejamento de geração de energia e despacho de cargas.

## 1.4 Justificativa

No Brasil, a busca por fontes de geração de energia limpa tem crescido consideravelmente, destacando-se a geração de energia solar fotovoltaica. Como citado previamente, de acordo com os dados do Balanço Energético Nacional (BEN, 2025), tratando-se de micro e mini geração distribuída de energia elétrica, dos 42.268 GWh gerados e 36.168 MW instalados, 97% da geração e 99,2% da carga instalada são oriundas da fonte solar fotovoltaica.

Os Sistemas Elétricos Fotovoltaicos têm se popularizado na sociedade, o que levou os brasileiros a buscar instalar tais sistemas em suas residências, comércios e empresas. No entanto, mais que reduzir o impacto ambiental, o consumidor busca entender quais os impactos na conta de energia a partir da tecnologia que será instalada.

Dessa forma, o presente estudo busca comparar e justificar ao consumidor final, a sociedade, o uso das tecnologias de módulos fotovoltaicos de silício policristalino e silício monocristalino e sua produtividade energética, a partir de dados selecionados e sistematizados em ordem cronológica (Análise de Séries Temporais por *Machine Learning*).

Como consequência, a análise estatística escolhida poderá ser utilizada para mensurar o comportamento energético futuro das tecnologias analisadas, servindo como base para futuros investimentos, orientando a tomada de decisões.

Além dos benefícios apresentados à sociedade, o tema também é relevante para a comunidade acadêmica da UTFPR e de outras instituições, já que avaliará as tecnologias dos módulos fotovoltaicos do LABENS do campus UTFPR-CT, sede Neoville. A partir dos resultados apresentados, as organizações científicas poderão analisar outros aspectos da produção de energia fotovoltaica, como as possibilidades de refino das tecnologias estudadas ou o desenvolvimento de novas.

## **1.5 Abordagem metodológica e metodologia**

Os procedimentos de pesquisa empregados neste trabalho são de cunho bibliográfico, com consultas em bases de pesquisa acadêmicas virtuais, fornecendo acesso às fontes que serão estudadas para a elaboração do trabalho, como artigos, bases de dados, teses, livros e demais trabalhos e pesquisas relacionados ao tema, e através de consultas com um profissional da área, que forneceu as bases de dados necessárias para a análise e obtenção dos resultados desejados, assim como o procedimento de estudo de caso da situação avaliada. Além disso, serão aplicados os conceitos de *Machine Learning* através de ferramentas computacionais, como a linguagem de programação *Python*.

## **1.6 Estrutura do trabalho**

O trabalho apresentado é dividido a partir da seguinte estrutura:

Capítulo 1: Introdução. É a seção que apresenta o tema e sua delimitação de estudo, o problema e as premissas, os objetivos geral e específicos da análise. Além disso, expõe a justificativa para o conteúdo estudado, a partir de uma visão social e acadêmica. Por fim, o primeiro capítulo exhibe a abordagem metodológica de análise dos dados utilizados e o detalhamento da estrutura de trabalho.

Capítulo 2: Revisão de Literatura. Apresenta os fundamentos teóricos e a revisão bibliográfica de tecnologias fotovoltaicas e geração de energia fotovoltaica, além de definições para o estudo de Análise de Séries Temporais, a fim de embasar todo o procedimento matemático que envolve o estudo estatístico apresentado no capítulo seguinte. Além disso, também será apresentada a mesma análise, mas considerando os aspectos de *Machine Learning*.

Capítulo 3: Metodologia. Apresenta de forma sequencial o escopo matemático utilizado, a partir da linguagem de programação *Python*, a fim de analisar a base de dados disponibilizada pelo LABENS, delimitada entre 01/01/2022 e 31/12/2024.

Capítulo 4: Resultados. Apresenta os resultados matemáticos e estatísticos obtidos a partir das manipulações computacionais dos dados, a fim de interpretá-las e descrever, numericamente, as diferenças de produtividade energética das duas tecnologias comparadas.

Capítulo 5: Conclusões. Apresenta as conclusões e discussões, a partir de um sistema causal, os padrões de ações temporais e a evolução da produtividade energética dos sistemas fotovoltaicos de silício monocristalino e policristalino. A seção também tem por objetivo avaliar a melhor das duas tecnologias estudadas, além de retornar uma previsão de energia média a ser gerada pelas plantas analisadas.



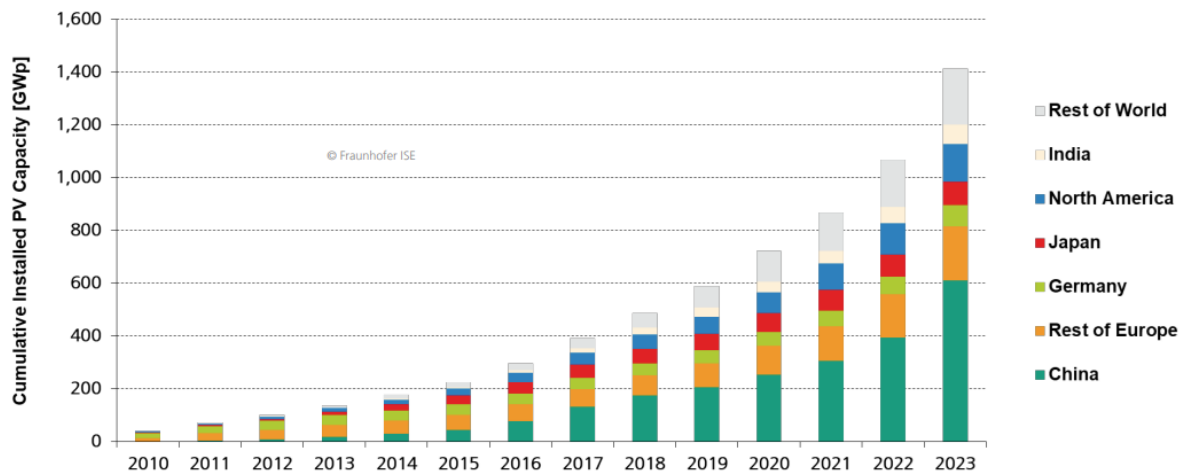
## 2 REVISÃO DE LITERATURA

De modo a desenvolver o trabalho, foi necessário um desenvolvimento teórico através de revisão de literaturas, contemplando livros, artigos, pesquisas e demais trabalhos que contemplam os conteúdos teóricos a serem analisados e discutidos no decorrer do trabalho.

### 2.1 Análise fotovoltaica

A energia solar fotovoltaica é definida como aquela gerada através da conversão direta da radiação solar em eletricidade, através de células fotovoltaicas, as quais atuam se baseando no princípio fotovoltaico (Becquerel, 1839). Este fenômeno funciona de modo que, quando é absorvida a radiação solar, é ocasionada uma diferença de potencial na estrutura do material semicondutor (Severino, Oliveira, 2010). Quando não há a presença do sol, ou quando esta é reduzida, os valores de geração de energia reduzem drasticamente ou se anulam.

O uso da tecnologia de geração de energia fotovoltaica teve um crescimento exponencial nos últimos 20 anos - entre 2000 e 2022, foi observado um aumento de 37%, em média, por ano, na produção mundial de energia solar, com essa porcentagem duplicando a cada 2,2 anos (*Ember Climate*, 2023). Assim, houve uma busca por avanços nas tecnologias empregadas nesta modalidade de geração de energia, utilizando diferentes tipos de materiais, diminuindo as perdas de potência e melhorando a eficiência, ao mesmo tempo que reduzindo custos. A figura 1 apresenta uma evolução da capacidade instalada de potência fotovoltaica, em GWp, em um período de 13 anos, comparando países que possuem as maiores capacidades instaladas, culminando em aproximadamente 1.412 GWp instalados mundialmente em 2023.

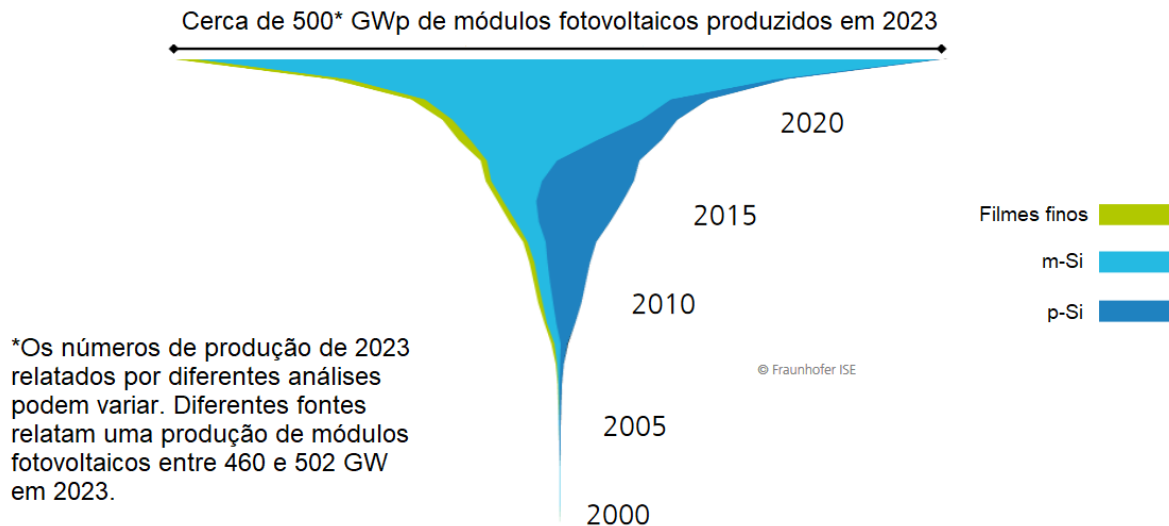
**Figura 1 - Evolução da capacidade total instalada de potência fotovoltaica**

Fonte: Fraunhofer ISE (2024)

De acordo com o relatório global REN21, de 2025, a capacidade instalada de sistemas solares fotovoltaicos ao redor do mundo alcançou 2,25 TW, aproximadamente o dobro da capacidade instalada em 2022, com previsão de chegar a marca de 6 TW em 2030 (REN21, 2025).

O material mais utilizado na fabricação das células fotovoltaicas é o silício (Si), o qual é empregado em suas formas monocristalino (m-Si), policristalino (p-Si) e amorfo (a-Si), sendo este último considerado como filme fino (Luque; Hegedus, 2011). De acordo com o estudo “Photovoltaics Report”, de Fraunhofer ISE, atualizado em julho de 2024, os módulos de silício monocristalino e policristalino compõem cerca de 97% das tecnologias presentes no mercado, devido às suas características como eficiência e confiabilidade. Na figura 2, pode ser observada a evolução da produção de módulos fotovoltaicos de diferentes tecnologias, evidenciando a predominância da tecnologia de silício monocristalino (m-Si).

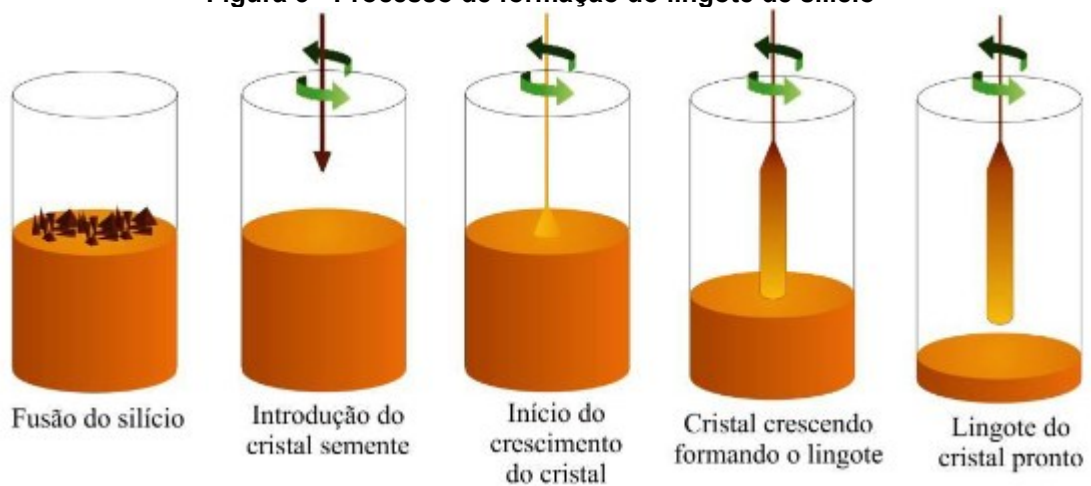
**Figura 2 - Evolução da produção de módulos fotovoltaicos de diferentes tecnologias**



**Fonte: Adaptado de Fraunhofer ISE (2024)**

As células compostas de silício monocristalino são feitas a partir um procedimento chamado de processo de Czochralski, no qual um único cristal de silício é mergulhado em silício fundido e, durante seu crescimento, é dopado com boro de modo a se tornar um semiconductor de tipo “p” (Mori; Santos; Sobral, 2007). Na figura 3, é observada uma representação da etapa de formação do lingote de silício, o qual é cortado em fatias de pequena espessura e dopado novamente, formando a junção “p-n”, tornando-as utilizáveis para a construção de células fotovoltaicas.

**Figura 3 - Processo de formação do lingote de silício**

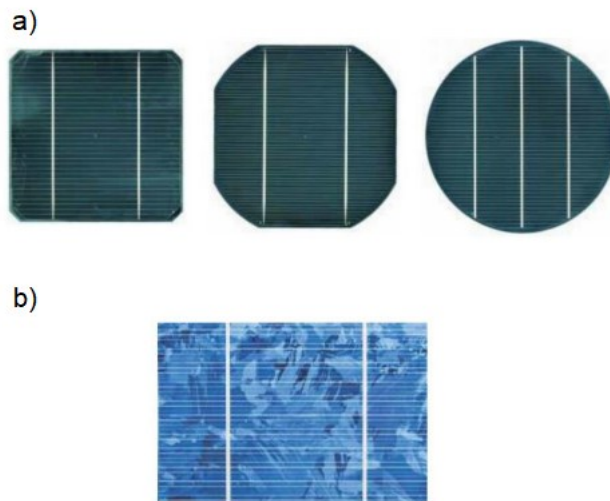


**Fonte: Adaptado de MIT (2013)**

Já as células de silício policristalino não passam pelo processo de Czochralski, sendo que o silício líquido de alta pureza do tipo “p” é resfriado, formando cristais que são diretamente dopados com fósforo, de modo a formar a junção “p-n”.

Estas células possuem eficiência entre 12% e 20%, enquanto as células de silício monocristalino possuem um rendimento na faixa de 15% a 23%, sendo que valores de aproximadamente 27% já foram alcançados experimentalmente (LONGi, 2024). A figura 4 apresenta imagens de células fotovoltaicas das tecnologias silício monocristalino e silício policristalino, respectivamente.

**Figura 4 - Células fotovoltaicas das tecnologias: a) silício monocristalino; b) silício policristalino**



**Fontes: DGS (2008); EPIA (2011)**

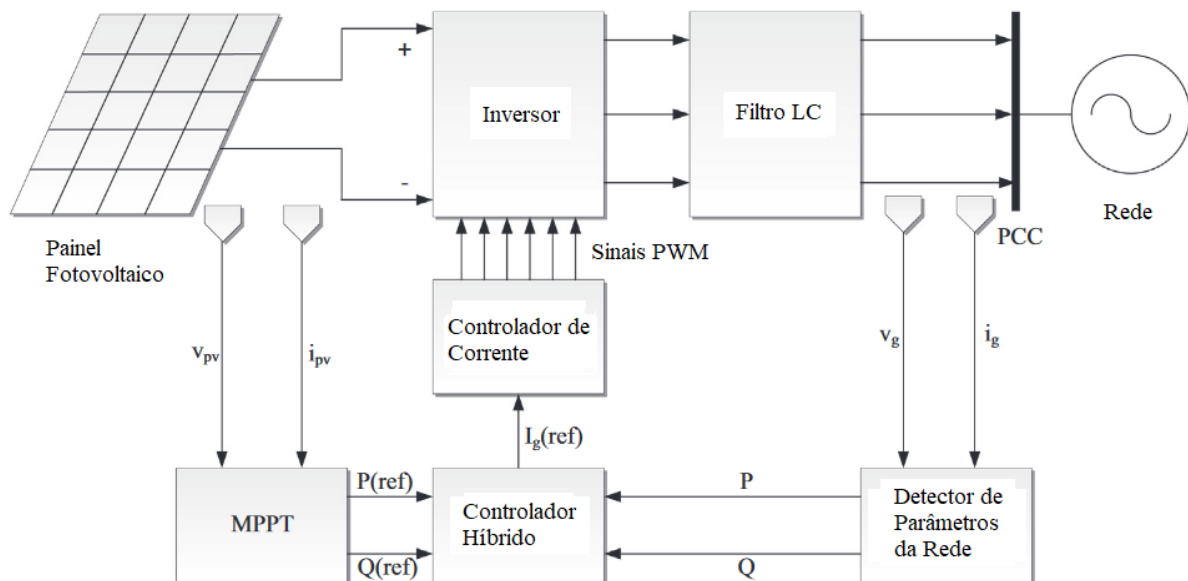
Quando expostos à luz, os elétrons da junção “p-n” das células se tornam mais energéticos, devido à absorção de fótons, gerando uma corrente elétrica através da junção, cujo fluxo dá origem à diferença de potencial entre as duas faces da junção “p-n” - fenômeno conhecido como efeito fotovoltaico (Guimarães *et al*, 1999).

Uma vez que a luz solar é capaz de gerar eletricidade, através de corrente e tensão, torna-se possível a associação entre as células fotovoltaicas, de modo a atingir valores desejados dos parâmetros elétricos, para que o resultado final seja um módulo com capacidade de captar a irradiação solar e transformá-la em energia elétrica, a qual poderá ser convertida e injetada na rede com o auxílio de um inversor de frequência (Pereira; Oliveira, 2011). Usualmente, os sistemas são projetados de modo que a energia gerada possa ser utilizada para consumo próprio, e também injetada na rede em casos de não utilização, ou seja, sem nenhum equipamento que armazena a energia gerada em excesso.

Os sistemas fotovoltaicos conectados à rede (SFVCR) são elaborados de forma que toda a energia gerada através dos módulos é injetada na rede elétrica, garantindo um elevado rendimento e tornando desnecessária a utilização de baterias

e outros equipamentos de armazenamento de energia (Pinho; Galdino, 2014). Esse tipo de sistema é usualmente composto pelos painéis fotovoltaicos, que irão converter a energia solar em energia elétrica em corrente contínua (CC), através do efeito fotovoltaico, associados a um ou mais inversores de frequência, responsáveis por converter essa energia elétrica para corrente alternada (CA), além de corrigir a amplitude e a frequência de modo a atender os requisitos da rede à qual o sistema estará conectado (110/220 V, 60 Hz) (Pinho; Galdino, 2014). Destaca-se que a melhor produtividade de um módulo fotovoltaico é atingida com a maior conversão e, conseqüentemente, absorção da energia solar. A figura 5 apresenta um esquemático de um sistema fotovoltaico conectado à rede.

**Figura 5 - Esquemático de um sistema fotovoltaico conectado à rede (SFVCR)**

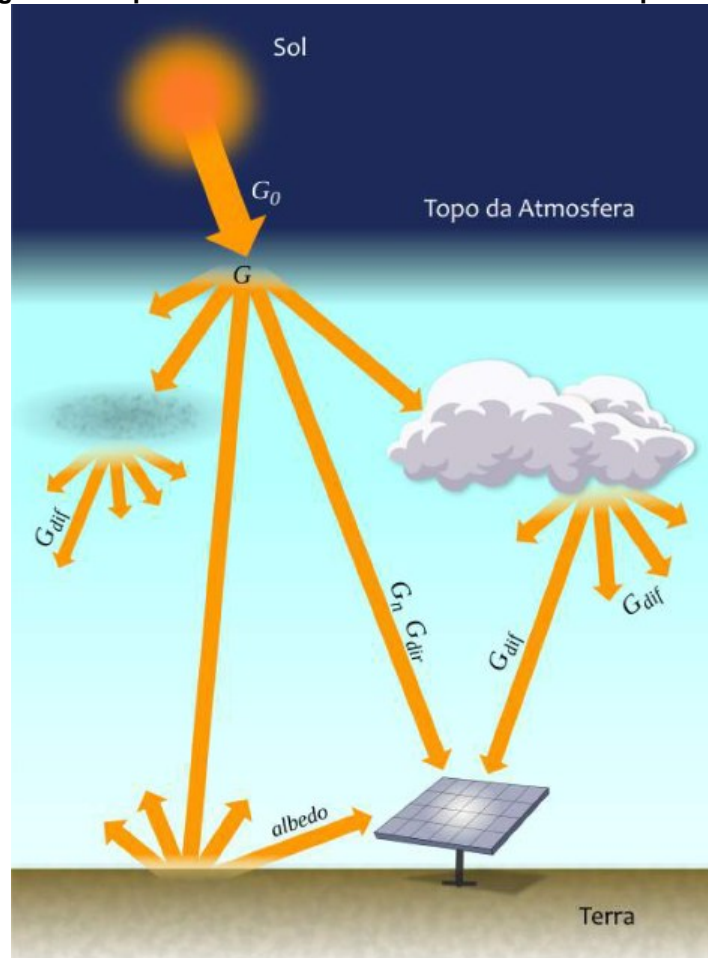


**Fonte: Adaptado de Sujatha et. al (2016)**

Desta forma, percebe-se a importância da melhor e mais eficiente captação da irradiação solar, de modo a obter melhores resultados na geração de energia de sistemas fotovoltaicos. A irradiação é a energia radiante incidente acumulada em um intervalo de tempo, e irradiância é a taxa de energia incidente por unidade de área, e possui componentes diretas e difusas, nas direções normal, horizontal e no plano inclinado (Pereira *et al.*, 2017), sendo que a condição ideal para um painel fotovoltaico captar essa energia é inclinado na latitude do local que for instalado, e voltado para a linha do Equador. Na cidade de Curitiba, no estado do Paraná, no Brasil, por exemplo, a condição ideal para instalação de um painel fotovoltaico é com este inclinado em

aproximadamente 25° e voltado para o Norte geográfico (Urbanetz Jr *et al*, 2016). A figura 6 ilustra a irradiância assim que atinge a atmosfera, e seu espalhamento.

**Figura 6 - Espalhamento da irradiância em suas componentes**



Fonte: Atlas Brasileiro de Energia Solar, Tiepolo *et. al* (2017)

Devido à dependência de fatores climáticos e outras variáveis, o uso de energia solar fotovoltaica possui um considerável grau de incerteza e complexidade técnica (Duman Altan, Diken e Kayışoğlu, 2021). Assim como outras fontes de energia renováveis, parâmetros ambientais e geográficos são impactantes na geração de energia solar. A estimativa de potência gerada por um SFVCR é de suma importância para o sistema elétrico, uma vez que este deve estar preparado para as variações de entrega de potência de diversos sistemas, incluindo grandes usinas, que trabalham com base nas características climáticas instantâneas do local onde estão instalados. Os pontos que mais requerem atenção dentro do sistema elétrico, em relação à variabilidade da geração de potência dos sistemas fotovoltaicos, são a confiabilidade, a gestão de reservas, o planejamento de produção e a integração do sistema como um todo (Raza *et al*, 2016).

Ter conhecimento das condições climáticas e ambientais de uma localização é essencial para a análise de implantação de um sistema solar fotovoltaico. Características como irradiação solar e clima na região especificada, assim como a posição de instalação dos módulos, umidade e nebulosidade, são as mais relevantes não só para o estudo de implementação de um sistema, mas também para o comportamento futuro de geração de energia.

Deste modo, é possível interpretar que existem relações diretas entre a potência gerada por um painel fotovoltaico, a qual pode ser interpretada como energia, quando avaliada no decorrer de um período de tempo, e variáveis como irradiância, irradiação, temperatura, umidade, entre outras (Onaifo *et al*, 2020). Além disso, dados intrínsecos de um módulo fotovoltaico, como tensão de circuito aberto e corrente de curto-circuito, também são importantes para a avaliação e o estudo da potência que pode ser entregue por um sistema, composto de vários módulos (Lacerda, 2014).

Tendo conhecimento dos principais fatores que influenciam na geração de energia, é possível pensar em abordagens, que podem ser de natureza física ou estatística, para que sejam elaboradas relações entre as variáveis da geração de energia solar fotovoltaica, as quais são importantes para a redução ou eliminação do risco baseado na incerteza da quantidade de energia a ser produzida (Duman Altan, Diken e Kayışoğlu, 2021).

Nas seções seguintes, serão apresentados conceitos básicos das análises a serem empregadas no trabalho.

## **2.2 Estatística descritiva**

A estatística descritiva é uma área da estatística que tem por objetivo analisar o comportamento dos dados e definir métricas para caracterizá-los. Esses estudos podem ser aplicados para uma amostra (um recorte da população) ou para uma população (conjunto de todos os pontos estudados). Quando uma amostra é responsável por descrever uma população, isso é chamado de inferência estatística, utilizando de um raciocínio indutivo (Devore; Cordeiro, 2006).

Para estudar uma base de dados algumas métricas podem ser utilizadas para caracterizar as amostras. Média, mediana e moda são valores interessantes para descrever a localização de uma base. Além delas, outros recursos gráficos como

dispersão e *boxplot* também podem ser aplicados para avaliar o comportamento dos dados e embasar decisões estratégicas para a engenharia (Devore, 2006).

Uma das formas de descrever a amostra é estudando a tendência central dos dados, por meio da média aritmética simples, resultante do cálculo da divisão entre a soma de todas as amostras pela frequência de observações. A unidade de medida da média é a mesma dos dados analisados (Devore, 2006).

A mediana é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente (Bussab, 2017). Ela também é considerada uma medida mais robusta, por não ser muito afetada por valores discrepantes (*outliers*). No entanto, definir a mediana pode ser difícil quando o conjunto de dados é expressivo, já que ela exige organização dos valores, em ordem crescente. Além dessa métrica, outra que pode indicar o comportamento de uma base de dados é a moda, que busca encontrar os valores que se repetem com maior frequência dentro do conjunto (Magalhães; Lima, 2011).

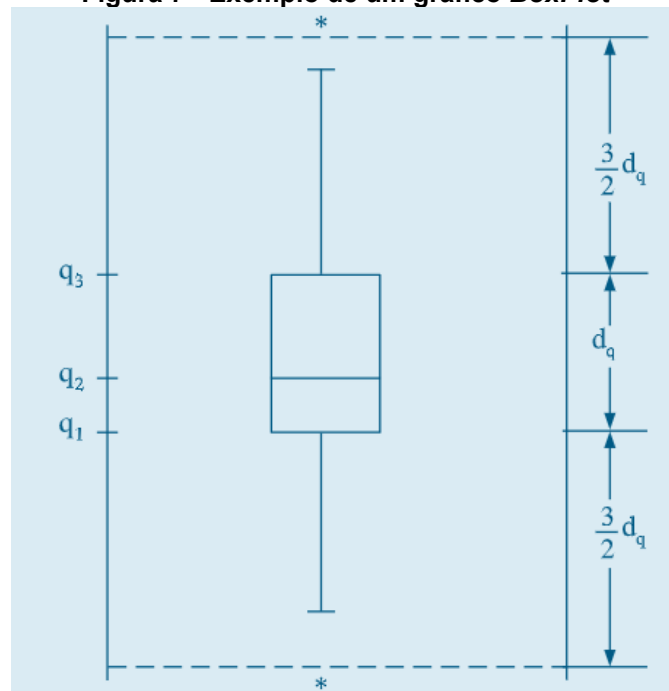
Ao considerar as medidas de média e mediana, nota-se que ambas possuem desvantagens, uma vez que a média é muito afetada pelos dados discrepantes, enquanto a mediana não é. (Devore; Cordeiro, 2006).

Outra maneira de visualizar a disposição dos dados é a partir do quartil. O primeiro critério de divisão é o segundo quartil (Q2), ou seja, a mediana, que divide o conjunto ao meio, a metade superior e inferior. Já o primeiro quartil (Q1) delimita o valor de corte para os 25% do conjunto inferior a ele, enquanto o terceiro quartil (Q3) realiza a mesma segmentação, pensando nos 75% dos dados abaixo do corte Q3. Dessa forma, o conjunto é dividido em quatro partes iguais. Realizando a diferença entre Q1 e Q3 é calculado o intervalo interquartílico (IQR), que avalia a dispersão de 50% dos dados (Devore, 2006).

Um dos artifícios gráficos que podem descrever a variabilidade dos dados são os gráficos *boxplot*. Esse recurso descreve a distribuição dos dados a partir da mediana, o segundo quartil (Q2) - semirreta no meio da caixa - e o intervalo interquartílico (IQR - entre o primeiro e o terceiro quartil), nas extremidades do retângulo. As semirretas externas à caixa se estendem até o menor e o maior valor da base analisada, dentro do intervalo de  $1,5 \cdot \text{IQR}$ . Um aspecto importante do *boxplot* é a visibilidade de assimetria e valores extremos, por conta de sua distribuição gráfica (Devore, 2006). A figura 7 apresenta um exemplo de um gráfico *BoxPlot*.



**Figura 7 - Exemplo de um gráfico *BoxPlot***



Fonte: Morettin; Bussab (2010)

Além das métricas de localização uma amostra também pode ser interpretada a partir do seu grau de dispersão. A variabilidade de um conjunto de dados pode ser caracterizada pela amplitude, que é o intervalo entre os dois valores extremos do conjunto o que também apresenta imprecisão, já que é uma medida que caracteriza um conjunto a partir de dois valores. No entanto, a dispersão também é descrita a partir do desvio das amostras em relação à média. A soma dos quadrados dos desvios, dividida pelo número de amostras menos um, é chamada de variância amostral (Devore, 2006).

A vantagem de utilizar a variância está na normalização dos dados, pela elevação ao quadrado da diferença entre a amostra e a média, que evita o estudo de valores negativos, no entanto ela falha no quesito de unidades, já que eleva ao quadrado a unidade de medida. Essa métrica pode ser definida para uma amostra ou população e quanto maior o seu valor, mais dispersa é a amostra. Para resolver essa questão, o desvio padrão pode ser aplicado, tirando a raiz quadrada da variância. Assim como a variância, o desvio padrão também é maior quanto maior é a dispersão dos dados, além de ser afetado pelos *outliers* (Devore, 2006).

### 2.3 Análise de correlação

Quando é desejável conhecer o nível de interferência entre duas ou mais variáveis são aplicados estudos de correlação. De acordo com Montgomery, Runger (2021), correlação significa “relação mútua”. Dessa forma, o foco do estudo de correlação não é prever um comportamento futuro, na verdade ele busca compreender o comportamento de uma covariável, por meio da variação de outra, para então refletir sobre o impacto dessa alteração.

Um recurso muito utilizado nas análises de correlação são os diagramas de dispersão. Esse tipo de gráfico tem como objetivo mostrar a distribuição das amostras a partir de pares de coordenadas definidas pelas covariáveis X e Y. A partir desse gráfico, é possível inferir se a tendência do relacionamento entre as duas variáveis é linear ou não, além do comportamento de pontos discrepantes, os *outliers* (Morettin; Bussab, 2010).

Nesse sentido, a partir das diferenças em relação à média, é possível verificar a distribuição dos dados além da tendência de comportamento entre elas. Os desvios  $S_x$  e  $S_y$  são dados pelas equações 1 e 2, respectivamente:

$$S_x = d_{xi} \quad (1)$$

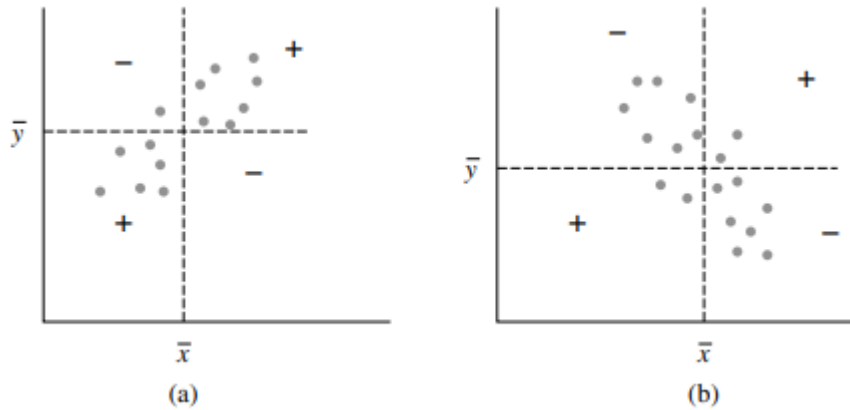
$$S_y = d_{yi} \quad (2)$$

Já o comportamento entre as covariáveis é dado pela equação 3:

$$S_{xy} = d_{xi}d_{yi} \quad (3)$$

A partir dessa métrica é possível inferir se a relação entre as covariáveis será positiva ou negativa. Se as duas amostras estiverem acima da média os desvios serão positivos, logo a relação será fortemente positiva. O mesmo é válido caso os pontos escolhidos estiverem abaixo da média. Analogamente, o produto e a relação serão negativos caso um dos desvios seja negativo. Criando um gráfico dessa relação  $S_{xy}$  é possível verificar a tendência desses pontos. Caso os pontos se distribuam entre os quatro quadrantes, pode-se afirmar que não há associação linear entre as amostras, logo  $S_{xy}$  tende a zero (Devore, 2006). A figura 8 apresenta as diferentes possibilidades de dispersão, positiva e negativa, respectivamente.

Figura 8 - (a) Dispersão de  $S_{xy}$  positiva; (b) Dispersão de  $S_{xy}$  negativa



Fonte: Devore (2006)

A soma dos produtos dos desvios, dividida pela quantidade de amostras ( $n$ ) é chamada de covariância. Essa métrica estatística descreve a dinâmica de variação conjunta de duas variáveis. No entanto, ela sofre com a interferência das unidades de medida e da variabilidade desigual das covariáveis, o que pode interferir nas análises. Para mitigar isso é aplicado o coeficiente de correlação de Pearson, que representa o grau de relação entre as amostras de covariáveis selecionadas, sem dependência das unidades de medida e da dimensão dos valores (Morettin; Bussab, 2010). As equações 4 e 5 apresentam o cálculo da covariância e do grau de correlação, respectivamente.

$$cov(X, Y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n} \quad (4)$$

$$corr(X, Y) = \frac{cov(X, Y)}{S_x * S_y} \quad (5)$$

Sendo  $S_x$  e  $S_y$  o desvio padrão de  $X$  e  $Y$ , respectivamente. Os desvios são obtidos a partir da raiz quadrada da variância. Justamente por estar metrificado sem o viés de unidades, o coeficiente de correlação é um valor entre -1 e 1, sendo que quanto mais próximo de zero, menor a correlação entre os dados, ou seja,  $X$  interfere em  $Y$  com menos intensidade (Devore, 2006).

## 2.4 Machine Learning

Um dos métodos de análise estatística de dados é o *machine learning*, no qual um sistema pode aprender e tomar decisões com base nos dados recebidos,

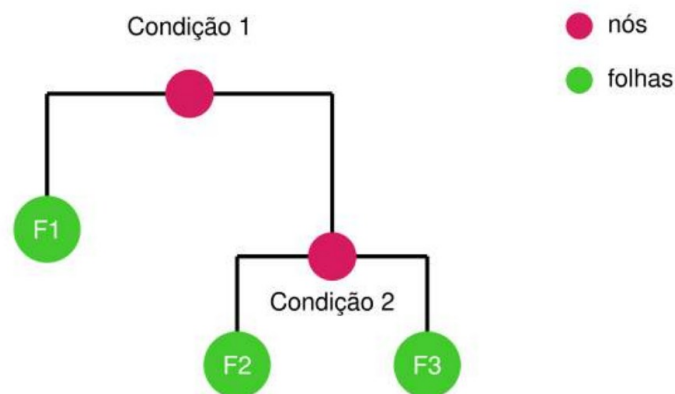
utilizando algoritmos para identificar padrões e com capacidade de fazer previsões ou classificações. Burkov (2019, p. 3) também define *machine learning* como “o processo de resolver um problema prático através de 1) obter um grupo de dados, e 2) algoritmicamente construir um modelo estatístico com base neste grupo de dados, o qual seria capaz de resolver o problema apresentado”.

É possível associar a análise através de *machine learning* com os modelos de séries temporais utilizando de ferramentas computacionais associadas a uma programação, como a linguagem *Python*, por exemplo. Para isso, podem ser utilizadas algumas bibliotecas específicas, como *pandas* - para a manipulação de *dataframes* - e *matplotlib* - para plotar gráficos e realizar operações matemáticas.

Um dos algoritmos utilizados em *Machine Learning* é o *Random Forest*. Ele é um dos métodos *ensemble*, cujo objetivo é ajustar um padrão descritivo de um fenômeno a partir do arranjo de diferentes modelos. Um desses ajustes também pode depender de um modelo anterior e assim sucessivamente. O princípio de execução desses algoritmos é a árvore de decisão, no qual uma decisão tomada leva aos mais diferentes resultados (Izbicki, 2020).

A estrutura de uma árvore de decisão é composta por nós e folhas. Os nós representam as condições a serem atendidas, que ramificam o fluxo de decisão. Para cada condição o caminho de execução segue para um sentido, chegando em uma folha, que representa um resultado, ou em outra condição e assim consecutivamente, de forma binária. Dessa forma, as covariáveis vão sendo divididas em diferentes regiões, a partir do cálculo da média das respostas observadas naquele subconjunto (Izbicki, 2020). A figura 9 apresenta um exemplo de diagrama de ordem de decisão.

**Figura 9 - Exemplo de um diagrama de ordem de decisão**



Fonte: Izbick (2020)

Essas divisões sucessivas procuram o critério do menor erro quadrático, a partir de um critério de corte ( $t$ ), até que existam menos observações por folha. A primeira etapa constrói uma estimativa para o conjunto de teste, mas pode superajustar para o conjunto de execução de amostras futuras. Para resolver isso é necessário “podar” a árvore, a fim de diminuir a variância do modelo. O mecanismo de recorte da árvore é realizado a partir da eliminação gradual de nós e posterior cálculo do erro médio, assim (RSS - *residual sum of squares*), é possível entender quais condições são mais importantes para o estimador, retirando as outras (Izbicki, 2020).

Além do RSS, outra métrica equivalente que pode definir a proximidade da predição com o fenômeno real, seu desempenho, é o RMSE - *Root Mean Squared Error*, um derivado do erro quadrático médio e é calculado a partir da raiz quadrada da média do RSS, como mostra a equação 6 (Hastie; Tibshirani; Friedman, 2009).

$$RMSE = \sqrt{\frac{RSS}{n}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \quad (6)$$

A partir da criação de diferentes árvores ( $B$  árvores, criadas pela função de *bagging* - *Bootstrap Aggregating*) é possível combiná-las para ter um estimador mais preciso de um fenômeno. A partir de uma amostra aleatória, de igual tamanho da original, que é repostada, chamada de *bootstrap*, gera-se uma função de predição resultante do *bagging* que pode ser representada pela soma de todas as  $g_b(x)$  funções a partir das  $B$  árvores, de acordo com a equação 7 (Izbicki, 2020).

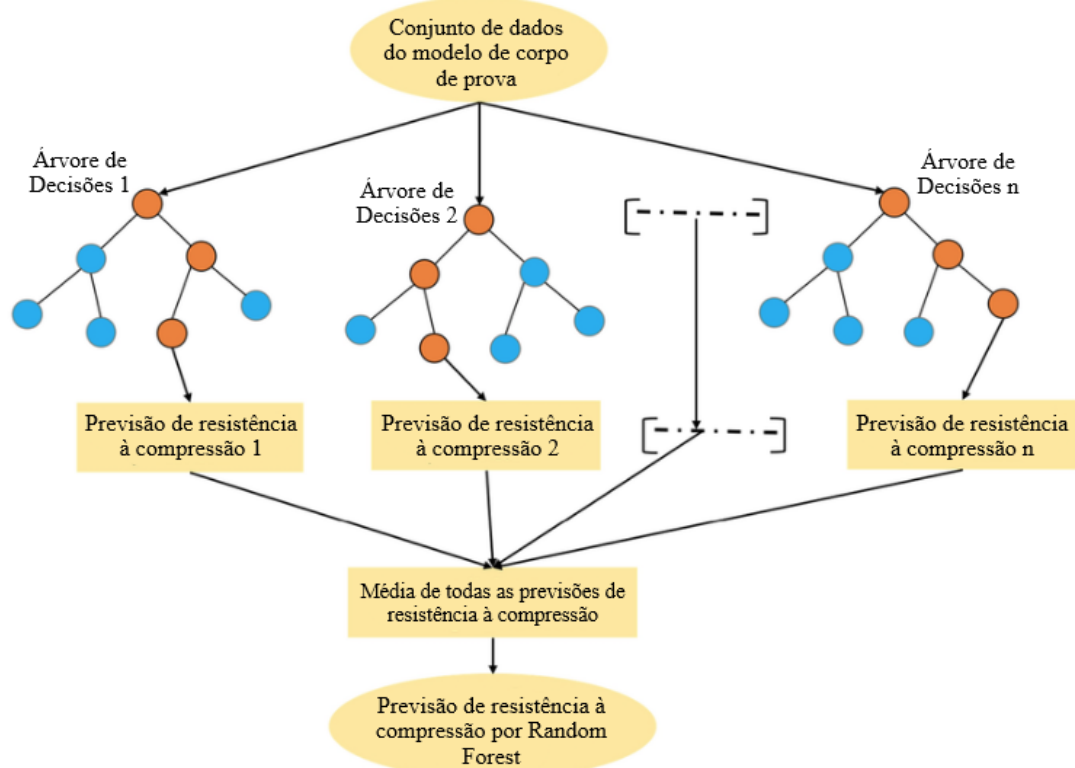
$$g(x) = \frac{1}{B} \sum_{b=1}^B g_b(x) \quad (7)$$

A função de *bagging* é, então, a média de todos os resultados previstos pelas funções que são originadas por amostras *bootstrap*. Caso  $B$  seja igual a dois e as funções  $g_1(x)$  e  $g_2(x)$  sejam não correlacionadas, não viesadas e com a mesma variância, o estimador  $g_b$  caracteriza melhor o conjunto de dados do que as estimativas  $g_1(x)$  e  $g_2(x)$  separadas (Izbicki, 2020).

No entanto, apesar de diferentes árvores representarem uma mesma série, elas podem ser muito parecidas, ou seja, serem altamente correlacionadas. O método *Random Forest* utiliza dois pressupostos: 1) os estimadores são não viesados e 2) os estimadores não são correlacionados. Esse modelo trabalha como uma expansão do

*bagging*, apresentado anteriormente. Para isso, a metodologia divide o grupo de  $d$  covariáveis em um subgrupo com  $m$  covariáveis e seleciona uma delas para cada nó de execução. Nesse caso, ao analisar o risco em função da quantidade de árvores  $B$ , nota-se que a curva resultante é estável (Izbicki, 2020). A figura 10 apresenta o diagrama de funcionamento do método *Random Forest*.

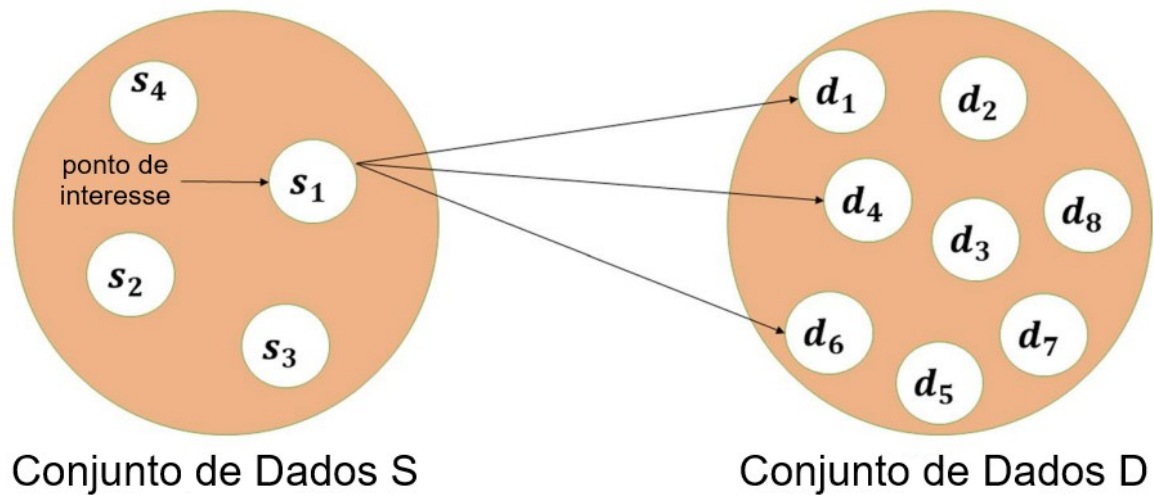
**Figura 10 - Diagrama de funcionamento do método *Random Forest***



**Fonte: Adaptado de Gogineni; Kumar; Panday; Paswan (2023)**

Além do modelo *Random Forest* a previsão por *machine learning* também pode utilizar outros algoritmos, como o KNN (*K-Nearest Neighbors*). Esse algoritmo utiliza vetores para prever a proximidade de um dado em relação ao outro e posteriormente, classificá-lo por sua similaridade com outras amostras, pertencentes a diferentes classes. Ao receber um novo dado, o princípio de funcionamento desse método é calcular distâncias entre a nova amostra e as já existentes, por meio de métodos como a distância Euclidiana e ponderada, por exemplo, a partir do parâmetro  $k$ , a quantidade de distâncias que serão calculadas (Halder *et al.*, 2024). Na figura 11, pode ser observado um exemplo de busca utilizando o método KNN, com  $k = 3$ , onde são encontrados, no conjunto de dados  $D$ , os pontos mais próximos do ponto de interesse no conjunto de dados  $S$ .

Figura 11 - Diagrama de busca do método KNN, para  $k = 3$



Fonte: Adaptado de Halder et al. (2024)

Em seguida, o algoritmo verifica as  $k$  distâncias e realiza uma média para prever o resultado da próxima amostra. Um cuidado ao utilizar esse método é usar um  $k$  adequado, caso ele seja pequeno, pode causar um *overfitting*, por conta da sensibilidade entre as amostras mais próximas. Similarmente, um  $k$  muito grande pode causar o *underfitting*, apesar de diminuir os ruídos (Halder et al., 2024).

Por se tratar de uma abordagem de distância vetorial, o algoritmo KNN pode ter seus dados normalizados ou não. No entanto, no caso de variáveis com escalas muito discrepantes essa operação se torna relevante para melhor adequação do modelo (Bhardwaj; Mishra; Desikan, 2018).

### 3 METODOLOGIA

#### 3.1 Análise descritiva dos dados

Com o objetivo de proceder com as análises iniciais do estudo, foram coletados dados referentes ao sistema físico instalado na Estação de Pesquisa em Energia Solar EPESOL do campus Neville da UTFPR, em Curitiba, assim como amostras de dados coletados pelos inversores, estação solarimétrica e sensores de temperatura, durante um intervalo de três anos, armazenados em bancos de dados do Laboratório de Energia Solar (LABENS).

##### 3.1.1 Análise descritiva do sistema físico instalado

Nos SFVCR analisados, existem 14 módulos de cada tecnologia estudada neste trabalho (m-Si e p-Si), cujas características elétricas e especificações técnicas podem ser observadas no Quadro 1.

**Quadro 1 - Especificações elétricas dos módulos fotovoltaicos em condições STC**

	<b>m-Si</b>	<b>p-Si</b>
<b>Fabricante</b>	QCells	QCells
<b>Modelo</b>	Q.Peak L-G5.0.G	Q.Power L-G5
<b>Potência nominal (Wp)</b>	365	335
<b>Tensão de operação (V)</b>	39,38	37,80
<b>Tensão de circuito aberto (V)</b>	48,16	46,30
<b>Corrente de operação (A)</b>	9,27	8,87
<b>Corrente de curto circuito (A)</b>	9,75	9,40
<b>Eficiência (%)</b>	≥18,80	≥17,20
<b>Temperatura de operação nominal na célula fotovoltaica (°C)</b>	45±3	45±3
<b>Coeficiente de temperatura em relação à potência nominal (%/°C)</b>	-0,39	-0,40



Fonte: Adaptado de Silveira (2020)

Estes módulos estão arrançados de modo que cada conjunto de 14 módulos, de cada tecnologia avaliada, está associado a dois inversores de 3 kW, da marca NHS Solar (modelo 3K-GSM1), sendo que cada inversor aloca uma *string* de 7 módulos conectados em série (Silveira, 2020). As características elétricas e especificações técnicas do modelo de inversor utilizado estão apresentadas no Quadro 2.

Quadro 2 - Especificações elétricas do inversor	
	<b>Inversor 3K - GSM1</b>
<b>Potência máxima CC (VA)</b>	3900
<b>Tensão Máxima CC (V)</b>	500
<b>Faixa de tensão do MPPT (V)</b>	80~450
<b>Tensão de partida (V)</b>	80
<b>Corrente máxima CC (A)</b>	18
<b>Proteção sobrecorrente CC (A)</b>	22,5
<b>Potência nominal CA (VA)</b>	3000
<b>Máxima potência CA (VA)</b>	3000
<b>Corrente máxima CA (A)</b>	13,5
<b>Proteção sobrecorrente CA (A)</b>	20
<b>Parâmetros nominais da rede</b>	50/60 Hz; 220/230 VAC

Fonte: Adaptado de Silveira (2020)

Além destes módulos e inversores descritos, existem também painéis das tecnologias CIGS e telureto de cádmio (CdTe), associados a inversores da mesma

marca, porém diferente modelo (menor capacidade de potência). Estes equipamentos não serão abordados neste trabalho.

Os inversores são responsáveis pela coleta e armazenamento dos dados de energia gerada dos módulos fotovoltaicos, a cada minuto. Nota-se que este é um dado cumulativo, ou seja, o valor armazenado a cada minuto é somado ao valor registrado no minuto anterior. Além disso, os inversores também coletam dados de potência instantânea, (*Watts*).

Também instalada na estação de pesquisa, há uma estação do tipo SONDA (Sistema de Organização Nacional de Dados Ambientais), responsável por adquirir dados climatológicos, como temperatura ambiente. Para o trabalho apresentado, os dados de irradiância (global horizontal e no plano inclinado), além da temperatura ambiente e na superfície dos módulos são os mais relevantes.

A coleta dos dados de irradiância é feita por dois piranômetros, instalados com diferentes inclinações, sendo que os valores medidos são diferentes para cada ângulo de inclinação. O piranômetro instalado em 0° é responsável por coletar os dados de irradiância global horizontal, enquanto o outro equipamento foi instalado na mesma inclinação dos módulos (25°), considerada a inclinação ideal para obter melhores valores de irradiação solar, e coleta os dados de irradiância no plano inclinado.

Já a coleta dos dados de temperatura na superfície dos módulos é feita através de sensores de temperatura, modelo CS240 PT-1000, presos com uma fita na parte traseira de um dos módulos para cada tecnologia, totalizando quatro sensores por estação de pesquisa. A medição destes valores ocorre a cada segundo, e a média de cada minuto é enviada para o mesmo armazenamento dos dados de irradiância. Os dados de temperatura ambiente são coletados em intervalos de dez minutos, e são enviados para o armazenamento. A fotografia 1 apresenta um exemplo de fixação de sensor de temperatura na base do módulo fotovoltaico.

**Fotografia 1 - Sensor de temperatura preso na parte traseira de um módulo fotovoltaico**



**Fonte: LABENS (2010); Silveira (2020)**

### 3.1.2 Análise descritiva do banco de dados

De modo a realizar as análises desejadas, foram fornecidas para os autores planilhas mensais de dados, coletados a cada minuto, referentes à Estação de Pesquisa em Energia Solar EPESOL, do Laboratório de Energia Solar (LABENS) da UTFPR, localizada no campus Neoville, em Curitiba. As variáveis foram selecionadas tanto por sugestão do professor responsável pela base de dados, quanto por escolha dos autores, visando escolher aquelas mais relevantes para o estudo a ser realizado. Os quadros a seguir relacionam as variáveis do banco de dados selecionadas para o estudo com suas respectivas grandezas, unidades e o código como estão apresentadas nas planilhas de dados.

**Quadro 3 - Primeiro grupo de variáveis e suas grandezas, unidades e códigos**

Código	Variável	Grandeza	Unidade	Código	Variável	Grandeza	Unidade
ene-mon1CT	Energia acumulada no inversor referente ao primeiro conjunto de módulos monocristalinos	Energia	kWh	idc-mon1CT	Corrente DC referente ao primeiro conjunto de módulos monocristalinos	Corrente	A
ene-mon2CT	Energia acumulada no inversor referente ao segundo conjunto de módulos monocristalinos	Energia	kWh	idc-mon2CT	Corrente DC referente ao segundo conjunto de módulos monocristalinos	Corrente	A
ene-pol1CT	Energia acumulada no inversor referente ao primeiro conjunto de módulos policristalinos	Energia	kWh	idc-pol1CT	Corrente DC referente ao primeiro conjunto de módulos policristalinos	Corrente	A
ene-pol2CT	Energia acumulada no inversor referente ao segundo conjunto de módulos policristalinos	Energia	kWh	idc-pol2CT	Corrente DC referente ao segundo conjunto de módulos policristalinos	Corrente	A
glo_avg-so01CT	Irradiância global média	Irradiância	W/m <sup>2</sup>	mono-tempCT	Temperatura na superfície dos módulos monocristalinos	Temperatura	°C
glo_std-so01CT	Desvio padrão da medição de irradiância global	Irradiância	W/m <sup>2</sup>				

Fonte: Adaptado de banco de dados do LABENS (2025)

**Quadro 4 - Segundo grupo de variáveis e suas grandezas, unidades e códigos**

Código	Variável	Grandeza	Unidade	Código	Variável	Grandeza	Unidade
tilt_avg-so01CT	Irradiância no plano inclinado média	Irradiância	W/m <sup>2</sup>	pac-mon1CT	Potência ativa instantânea gerada pelo primeiro conjunto de módulos monocristalinos	Potência	W
tilt_std-so01CT	Desvio padrão da medição de irradiância no plano inclinado	Irradiância	W/m <sup>2</sup>	pac-mon2CT	Potência ativa instantânea gerada pelo segundo conjunto de módulos monocristalinos	Potência	W
vdc-mon1CT	Tensão DC referente ao primeiro conjunto de módulos monocristalinos	Tensão	V	pac-pol1CT	Potência ativa instantânea gerada pelo primeiro conjunto de módulos policristalinos	Potência	W
vdc-mon2CT	Tensão DC referente ao segundo conjunto de módulos monocristalinos	Tensão	V	pac-pol2CT	Potência ativa instantânea gerada pelo segundo conjunto de módulos policristalinos	Potência	W
vdc-pol1CT	Tensão DC referente ao primeiro conjunto de módulos policristalinos	Tensão	V	poli-tempCT	Temperatura na superfície dos módulos policristalinos	Temperatura	°C
vdc-pol2CT	Tensão DC referente ao segundo conjunto de módulos policristalinos	Tensão	V	tp_sfc-so01CT	Temperatura ambiente	Temperatura	°C

**Fonte: Adaptado de banco de dados do LABENS (2025)**

Foram fornecidas planilhas com dados mensais referentes aos 12 meses dos anos 2022, 2023 e 2024, totalizando cerca de 1080 dias no total, para análise. Considerando que os dados são coletados minuto a minuto, é possível aproximar um total de 1,5 milhões de amostras de todas as variáveis propostas acima (exceto temperatura ambiente, a qual é coletada a cada 10 minutos).

É importante destacar que, como todo grande banco de dados, existem “*gaps*” de informações em algumas situações. Estes intervalos nos quais não existem dados coletados podem ter diferentes explicações, como falhas de comunicação entre servidores (dados são coletados na sede Neoville, e enviados para o servidor localizado na sede Centro), quedas de energia em alguma das sedes, defeito em algum dos equipamentos de medição, sendo que dados como energia armazenada, potência instantânea, tensões e correntes, são coletadas pelos inversores de cada sistema, enquanto dados climatológicos são adquiridos pela estação SONDA.

Outra inconsistência observada foi em relação à energia medida e armazenada pelo inversor referente ao primeiro conjunto de módulos policristalinos (*ene-pol1CT*), a qual, a partir da data de 12/06/2023, parou de ser medida, devido à possíveis problemas de comunicação entre o inversor e os servidores, ou até um problema intrínseco ao equipamento, fazendo com que os dados armazenados de energia fossem mantidos em um valor constante, quando deveria ser crescente. Desta forma, optou-se por desconsiderar a análise deste sistema quanto à energia armazenada e potência instantânea, uma vez que a falta de dados afetaria negativamente as conclusões obtidas pela análise estatística.

Para a avaliação estatística dos dados, de modo a não prejudicar os resultados obtidos, foi definido um recorte de período de tempo, definindo o “dia” com início às 06h e fim às 18h, uma vez que o intervalo de tempo restante é referente à “noite”, quando não há geração de energia solar fotovoltaica (ou os valores ainda são mínimos, podendo ser desconsiderados), e, consequentemente, não sendo relevante para o estudo realizado neste trabalho. Ademais, foi também definido um recorte de tempo referente aos meses de análise, de modo a evitar os “*gaps*” de informações, além de tornar o estudo mais confiável. Esta segmentação dos dados será abordada com mais detalhes na seção 3.2.

### 3.2 Análise estatística dos dados

Dentro da amostra recolhida, será realizada uma segmentação dos dados, para uma análise de melhor qualidade. O período analisado será de 01 de janeiro a 11 de setembro de 2022, totalizando 254 amostras. A escolha por esses períodos se deu pela maior consistência de dados na base, ou seja, foram momentos de menos falhas. Além disso, analisar o intervalo selecionado é interessante pois representa dois períodos opostos, verão (janeiro - maior irradiância e temperatura) e inverno (junho - menor irradiância e temperatura).

Dentro dos períodos selecionados o recorte do dia a ser utilizado será das 06h até 18h de todos os dias. Essa escolha foi baseada na melhor performance dos dados no âmbito da estatística e energia solar fotovoltaica, já que o período de produção de energia ocorre durante o dia. Apesar dos períodos de maior luminosidade no verão serem maiores que do inverno, optou-se por padronizar o recorte diário, para reduzir a variabilidade do conjunto de dados.

Como o objetivo é comparar diferentes tecnologias de módulos fotovoltaicos, serão separados a partir de dois sistemas da usina, monocristalino 1 e policristalino 2. As variáveis de análise serão: *pac-mon1CT*, *pac-pol2CT*, *tilt\_avg-so01CT*, *mono-tempCT* e *poli-tempCT*. Esses dados serão agrupados de forma a apresentarem uma média diária, por mês, de cada uma das covariáveis apresentadas. Para melhor performance da análise serão utilizados recursos estatísticos como tabelas, gráficos *BoxPlot* e histogramas.

As bases serão modeladas utilizando a biblioteca *Pandas*, em *Python*. Essa biblioteca de ciências de dados trabalha com estruturas de *DataFrames* e *Series*, por meio de métricas descritivas, além de facilitar o trabalho com um volume grande de dados, em detrimento de técnicas manuais utilizadas pelo *Excel*. Para visualizar os dados será utilizada a biblioteca *Matplotlib*, a qual possibilita a plotagem de gráficos 2D e 3D, que explicam o comportamento dos dados de forma visual. A *Matplotlib* é uma extensão do *NumPy*, outra biblioteca matemática do *Python*.

Inicialmente será realizada a análise descritiva, a partir do cálculo da média de irradiância e temperatura durante os meses e sistemas analisados, além da distribuição por quartis, que possibilitará reconhecer como as amostras estão concentradas. Nesse sentido, a construção dos gráficos *boxplot* possibilitará o entendimento gráfico da dinâmica das amostras, ou seja, como elas estão distribuídas

em torno da mediana. Também será calculado o desvio padrão das variáveis, por mês, o que explicará a dispersão dos dados em torno de sua média.

Além disso, ao analisar as medidas de tendência central, média e mediana, pretende-se entender se o conjunto analisado é simétrico ou não e a existência ou ausência de amostras discrepantes, os *outliers*, a partir do que é esperado para um mês típico de verão e inverno.

A fim de estudar quais os fatores interferem com maior intensidade na produção de energia, também serão analisadas as correlações entre potência e irradiância entre a potência e temperatura, tanto em dias mais ensolarados, como dias mais nublados, da mesma forma com dias quentes e frios para os meses de janeiro e junho do período delimitado (2022, 2023 e 2024). Isso possibilita estudar qual o tipo de estrutura mais se afeta com uma variação brusca das variáveis físicas e o quanto dessa alteração cada módulo converte em potência.

Qualitativamente, os aspectos da correlação serão observados a partir da plotagem dos diagramas de dispersão da temperatura, irradiância e potência e pelos coeficientes de correlação entre temperatura e potência e irradiância e potência, além do cálculo de coeficientes de correlação. Pensando na base diária de cada um dos meses, será selecionado o período entre janeiro e setembro de 2022 dos sistemas monocristalino 1 e policristalino 2, pois estes não apresentaram perdas significativas na base de dados, portanto, geraram análises mais consistentes no sentido de estudar o comportamento da potência. Esse recorte também permitirá acompanhar uma parte do verão e da primavera, assim como o outono e o inverno de forma integral.

### 3.3 Análise por *Machine Learning*

A análise por *Machine Learning* entrará como forma de estruturar um modelo de predição da potência produzida pelo sistema fotovoltaico. As variáveis *mono-tempCT* e *tilt\_avg-so01CT* serão preditoras para *pac-mon1CT*, enquanto *poli-tempCT* e *tilt\_avg-so01CT* serão as covariáveis preditoras para *pac-pol2CT*. Nesse contexto, será assumido que a potência gerada poderá ser explicada pela irradiância e temperatura dos módulos. O trabalho utiliza uma distribuição conjunta de irradiância e temperatura, onde elas modelam a variável potência.

O objetivo será estudar a capacidade da máquina em modelar dados a partir de um banco de testes e comparar com a potência realmente gerada, para validar um



possível modelo descritivo para a produtividade do SFVCR mono e policristalino. Para tanto, serão aplicados dois algoritmos de *machine learning*, o *random forest* e o KNN (*k-nearest neighbors*). A escolha desses modelos foi motivada pela sua capacidade de processar dados não lineares, além de não exigir a normalização dos dados, como o SVR (*Support Vector Regression*), ou seja, é possível trabalhar com eles sem uma manipulação prévia dentro de uma escala, o que o torna um algoritmo mais simples e representa de forma justa um fenômeno físico.

No código escrito em *Python*, será utilizada a biblioteca *sklearn.ensemble*, que contém o modelo de regressão *RandomForestRegressor(...)* e a biblioteca *sklearn.neighbors* para o algoritmo KNN com o modelo *KNeighborsRegressor(...)*. Elas serão aplicadas pela sua facilidade de comandos voltados para o segmento de *machine learning*, além de serem extremamente completas em recursos de regressão e predição.

Esses algoritmos receberão 254 amostras de cada uma das duas tecnologias, que serão testadas via *Random Forest* e KNN, por meio de uma predição *out-of-fold* (OOF). Essa técnica de validação cruzada pressupõe que o conjunto de dados será previsto a partir de 5 divisões ou “dobras” na base, que farão com que cada subconjunto seja treinado a partir de dados diferentes daquele que se pretende prever. A escolha de 5 dobras se deu pela dimensão do conjunto de dados, que contém 254 amostras por tecnologia. Ao utilizar esse valor, a base é dividida em aproximadamente 6 partes, o que cobre a sazonalidade do período, sem que o conjunto de treino seja muito pequeno. Nesse sentido, o modelo não terá visto os dados que ele está tentando prever, anteriormente. Um detalhe é que a primeira dobra é utilizada como aquecimento do modelo (*warm-up*) e não é testada via OOF. Nesse sentido, as amostras previstas se iniciaram no dia 14 de fevereiro de 2022 e terminarão junto com o limite da base segmentada, no dia 11 de setembro de 2022.

Essa base não será embaralhada, ou seja, manterá a sequência temporal das amostras, para que a análise não seja prejudicada, para isso será utilizada a técnica de reamostragem *TimeSeriesSplit*. Ou seja, amostras anteriores serão responsáveis por prever os resultados futuros, seguindo a ordem cronológica.

Para o *Random Forest*, serão utilizadas 500 árvores de decisão por dobra, para que o sistema seja mais robusto e de maior confiabilidade sendo que o total de dobras é 5. Esse valor representa o equilíbrio entre tempo de processamento e adequação do modelo. Valores menores podem não ser o suficiente para generalizar

toda a cadeia de dados, enquanto uma quantidade maior de árvores aumentará o tempo de execução do código, o que não necessariamente garante uma assertividade integral.

Diferentemente de uma regressão linear ou polinomial, por exemplo, o algoritmo *random forest* não resultará em uma equação final que explique todo o processo. Na verdade, ele é um método que, a partir das árvores de decisão, fará uma previsão de valores futuros, considerando parâmetros passados.

No caso da análise pelo algoritmo KNN será utilizada a biblioteca *sklearn.neighbors*, que contém o modelo de regressão *KNeighborsRegressor(...)*, sem a normalização dos dados. Serão utilizadas as 5 dobras citadas anteriormente, para o OFF, e 5 vizinhos ( $k = 5$ ) para prever as amostras das variáveis de potência (pac-mon1CT e pac-pol2CT), por meio do cálculo da distância euclidiana. A escolha pela distância euclidiana se deu pelo seu desempenho para o conjunto de dados, sendo ele relativamente pequeno. Nesse sentido, pontos mais próximos são favorecidos, em detrimento de desvios maiores. O algoritmo também considerará que os vizinhos mais próximos do ponto de previsão terão meio peso (*weights="distance"*) na previsão, isso pretende melhorar o ajuste do modelo.

Para mensurar o desempenho dos algoritmos utilizados será calculado o RMSE (*root mean-square error*), ou a raiz do erro quadrático médio, responsável por avaliar o desvio da previsão ao dado real registrado. Além disso, o recurso gráfico também será aplicado para evidenciar o trabalho do algoritmo, frente a série existente, a partir de gráficos de linha, que indicarão a trajetória do estudo.

Por fim, será estimada a energia média gerada por mês, para cada um dos sistemas. Esse cálculo consiste em multiplicar os resultados de potência por 12, correspondendo ao intervalo de tempo considerado como doze horas, e, em seguida, somar os valores obtidos dentro de cada um dos meses. Esses resultados serão plotados no formato de gráficos de barras, que comparam as estimativas de energia média real e energia média gerada por cada uma das tecnologias estudadas.

Adicionalmente, para enxergar o percentual de diferença entre o que foi previsto e o efetivamente gerado será gerado um gráfico de linhas, cujos pontos representam o percentual da diferença entre energia prevista e gerada, por mês.

## 4 RESULTADOS E DISCUSSÕES

### 4.1 Análises estatísticas básicas

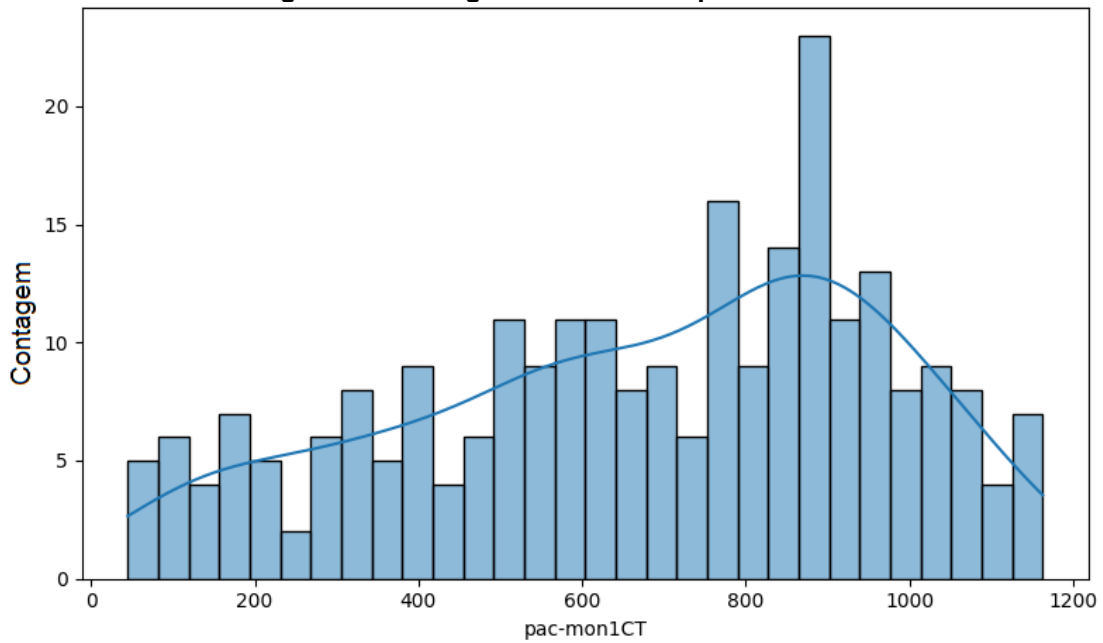
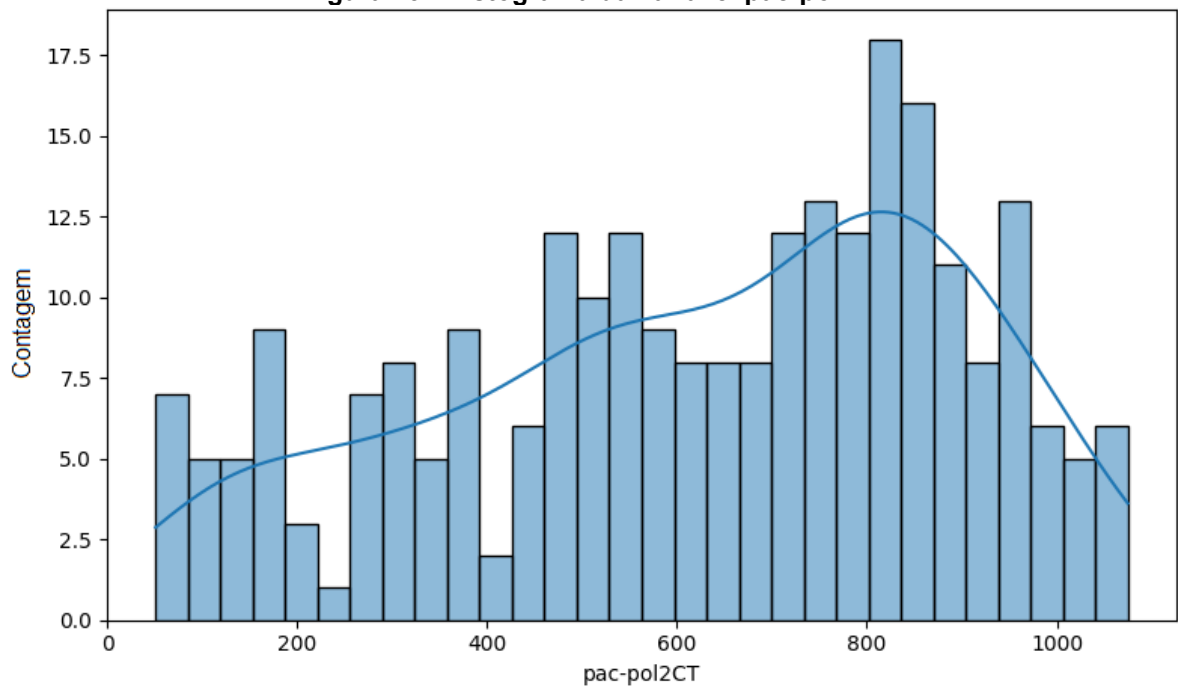
Uma vez que foram realizadas as manipulações estatísticas nos dados fornecidos pelo LABENS, com resultados de média, mediana, desvio padrão, mínimos e máximos, foram realizados agrupamentos mensais e anual (dentro do intervalo avaliado) dos dados obtidos. Os resultados encontrados para o intervalo definido (entre janeiro e setembro de 2022) podem ser visualizados na tabela 1.

**Tabela 1 – Resultados totais da análise manipulativa das planilhas de dados fornecidas aos autores.**

	<b>Média</b>	<b>Mediana</b>	<b>Desvio Padrão</b>	<b>Mínimo</b>	<b>Máximo</b>
<b>pac-mon1CT (W)</b>	672,9638473	722,3526836	290,8536471	45,02564103	1162,707692
<b>pac-pol2CT (W)</b>	624,2579163	673,1326515	270,2550907	51,08870056	1074,788462
<b>tilt_avg-so01CT (W/m<sup>2</sup>)</b>	352,0276161	379,5509637	155,4628897	29,05107349	628,5470828
<b>tp_sfc-so10CT (°C)</b>	19,20528656	18,98739654	4,530264809	7,952463615	28,24474692
<b>mono-tempCT (°C)</b>	27,99977295	28,10995513	7,768365852	9,71945641	45,89684615
<b>poli-tempCT (°C)</b>	27,40305499	27,42411603	7,468528333	9,851026923	44,79852564

**Fonte: Autores (2025)**

Os histogramas dos dados de potência referentes aos sistemas monocristalino e policristalino podem ser observados nas figuras 12 e 13, respectivamente.

**Figura 12 - Histograma da variável pac-mon1CT****Fonte: Autores (2025)****Figura 13 - Histograma da variável pac-pol2CT****Fonte: Autores (2025)**

Avaliando os histogramas, observa-se que ambos apresentam um comportamento semelhante, com uma curva resultante praticamente idêntica, que possui assimetria à esquerda, ou seja, para o período avaliado, houve maior frequência de dias com geração elevada de potência. O sistema monocristalino realizou mais medidas em valores maiores de potência, ultrapassando 20 contagens em aproximadamente 900 W, e pelo menos 5 contagens acima de 1100 W, além de

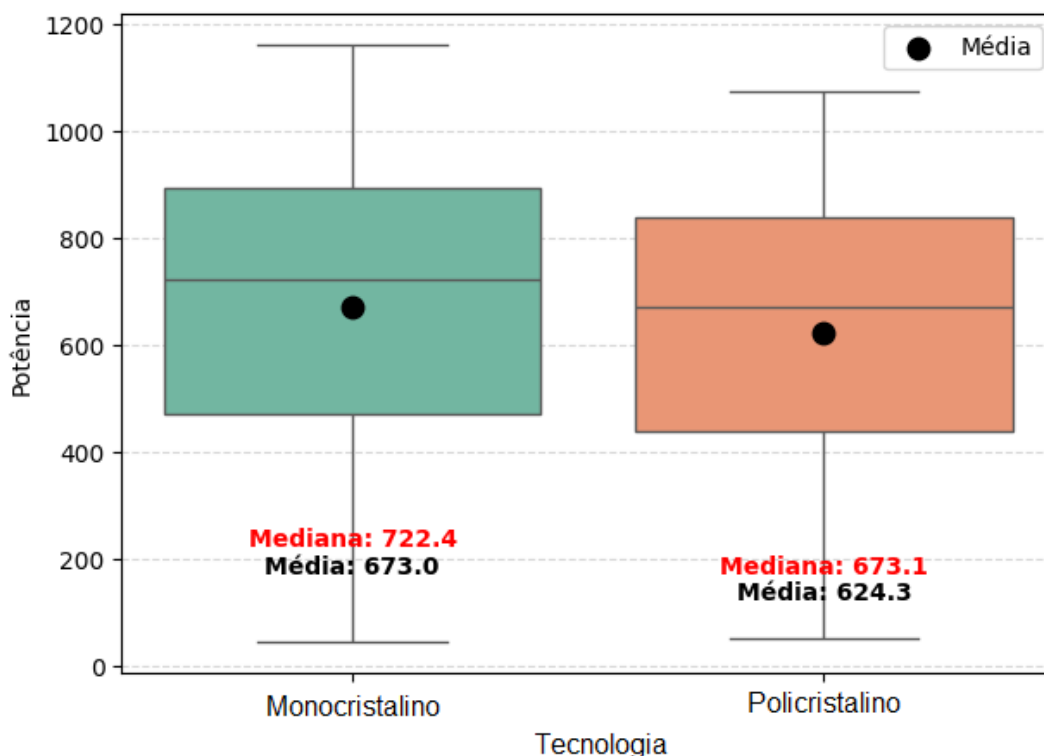
possuir valores médios maiores e uma maior variabilidade. Já o sistema policristalino teve seu maior número de medidas, cerca de 18, em aproximadamente 840 W, assim como possui uma concentração em valores médios menores. Isso indica que, no mesmo período avaliado, o sistema monocristalino gerou, em maior frequência (frequência de observações dos dados estatísticos), mais energia que o policristalino. A diferença entre a produção para as tecnologias pode se dar devido a fatores construtivos, assim como características intrínsecas aos módulos, como as perdas por alta temperatura.

A partir da Tabela 1 tem-se que as medianas são 28,11°C para o sistema monocristalino e 27,424°C para o policristalino, o que representa uma diferença de quase 10°C em comparação com a mediana da temperatura ambiente. Isso reflete o aquecimento causado pela irradiância incidente nos módulos e justifica a perda de eficiência dos sistemas em condições de altas temperaturas. Pela Tabela 1, a variação anual das médias de temperatura para os painéis monocristalino estão dentro da faixa de 9,719 °C e 45,897 °C e são um pouco mais dispersas que a variação apresentada no policristalino, que teve como mínima 9,851 °C e máxima de 44,789 °C.

A variação anual das médias diárias de irradiância (tilt\_avg-so01CT) ficou entre 29,051 W/m<sup>2</sup> e 628,547 W/m<sup>2</sup>. E a mediana foi de 379,55 W/m<sup>2</sup>. Para os painéis monocristalinos a média da potência mínima gerada foi de 45,056 W e a máxima de 1162,708 W. Já a mediana foi de 722,353 W, e a média 672,964 W (dados retirados da Tabela 1). Já no caso do sistema policristalino, a média da potência mínima gerada foi de 51,089 W e a máxima de 1.074,788 W. A mediana é 673,133 W e a média 624,258 W.

Foram comparadas as duas tecnologias apresentadas, via *boxplot*, visando elucidar a distribuição da potência. Os resultados estão descritos na figura 14.

**Figura 14 - Comparação entre os *boxplots* das duas tecnologias analisadas**  
**Distribuição da Potência por Tecnologia**



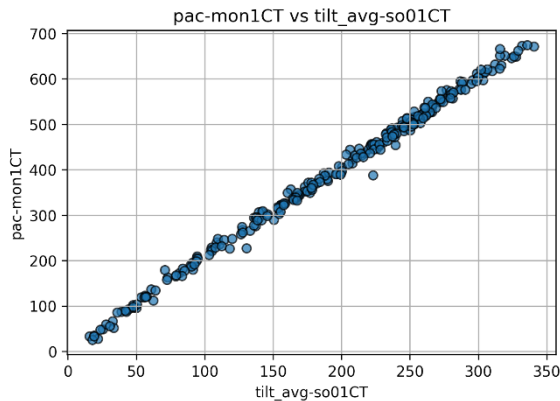
Fonte: Autores (2025)

Analisando os gráficos, é possível notar a diferença de performance entre monocristalino e policristalino, que é refletida em uma maior média, mediana e na própria distribuição dos dados, apesar do intervalo interquartil dos dois sistemas serem muito parecidos.

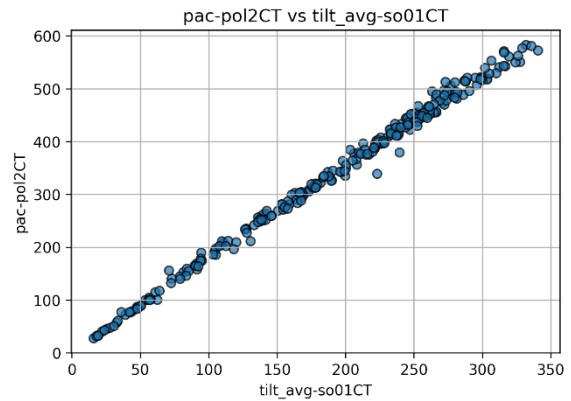
O teste de Wilcoxon para grupos independentes foi aplicado para comparar as potências entre os dois tipos de tecnologias. Com um p-valor de 0,0308 ( $p < 5\%$ ), rejeitou-se a hipótese nula de igualdade entre as amostras, incluindo diferença significativa na distribuição das potências entre os sistemas. Esse resultado confirma a distinção observada graficamente, em que as medianas e a dispersão dos dados nos *boxplots* apresentam níveis distintos.

Outra forma de analisar a relação entre variáveis é estudando os diagramas de dispersão, a partir do espalhamento dos pontos. Os gráficos gerados, apresentados nas figuras 15 e 16, relacionam as variáveis: *pac-mon1CT* e *pac-pol2CT*, *pac-pol2CT* e *tilt\_avg-so01CT*, *pac-mon1CT* e *tilt\_avg-so01CT*.

**Figura 15 – Diagrama de dispersão entre pac-mon1CT vs tilt\_avg-so01CT**



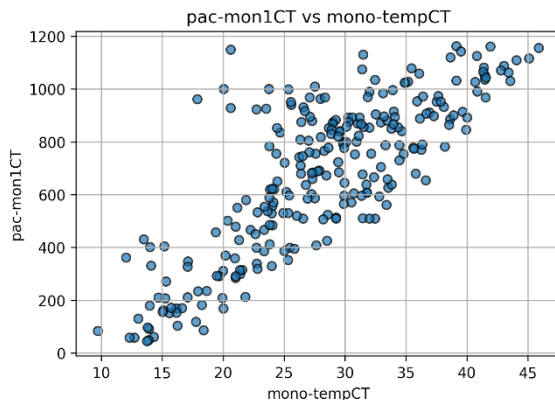
**Figura 16 – Diagrama de dispersão entre pac-pol2CT vs tilt\_avg-so01CT**



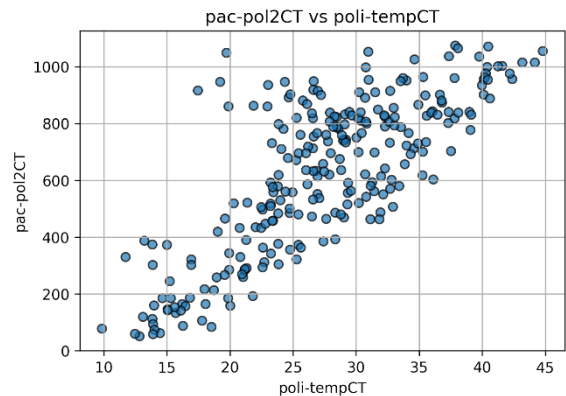
**Fonte: Autores (2025)**

Ao analisar a dispersão entre a potência dos módulos monocristalinos e policristalinos com a irradiância média (figuras 15 e 16, respectivamente), nota-se que as variáveis têm correlação positiva, ou seja, conforme a irradiância média aumenta (variável independente), a potência (variável dependente) também cresce. As figuras 17 e 18 apresentam a relação entre a potência gerada e a temperatura dos módulos monocristalino e policristalino, respectivamente, e podem ser visualizadas a seguir.

**Figura 17 - Diagrama de dispersão entre mono-tempCT vs pac-mon1CT**



**Figura 18 - Diagrama de dispersão entre poli-tempCT vs pac-pol2CT**



**Fonte: Autores (2025)**

Em termos gerais, ambas tecnologias apresentam comportamentos semelhantes, por conta do mesmo condicionante, a irradiância, sendo que o aumento da temperatura implica em maiores potências, principalmente ao analisar temperaturas entre 20°C e 35°C. Essa faixa de temperatura dos módulos é maior que a média de temperatura ambiente, conforme apresentado na Tabela 1.

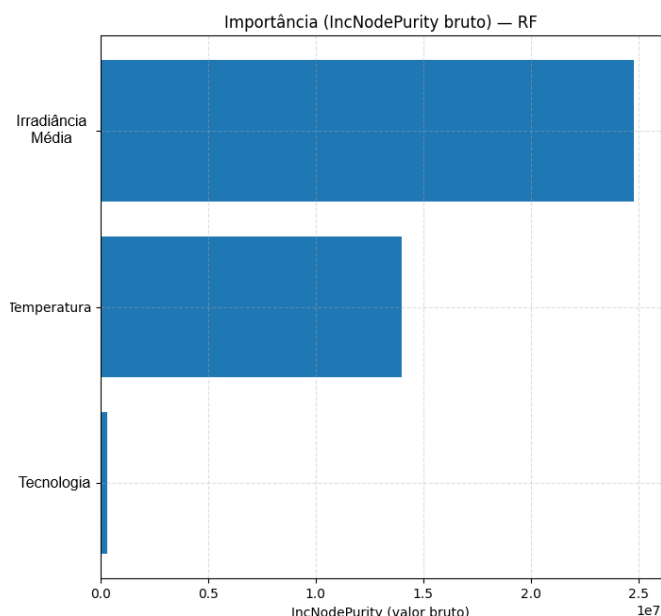
Para verificar a relação entre as variáveis de temperatura e irradiância média com as diferentes tecnologias avaliadas, monocristalina e policristalina, foram realizados testes de correlação de Spearman. Os valores do coeficiente de correlação de Spearman obtidos, entre a temperatura e a potência foram de 0,76 e 0,73, ambas com p-valor menor que 1% ( $p\text{-valor} < 0,01$ ), para as tecnologias monocristalino e policristalino, respectivamente. Com relação à correlação de Spearman entre a irradiância média e a potência, foram obtidos os valores 0,997 e 0,996, também com p-valor muito inferior a 1% ( $p\text{-valor} < 0,01$ ), para as tecnologias monocristalino e policristalino, respectivamente. Com base nos resultados, observa-se uma correlação significativa entre as variáveis resposta potência média no inversor, com a temperatura e a irradiância média.

## 4.2 Análise via *Machine Learning*

A partir da análise por *machine learning* foram estudadas as formas de predição e importância das covariáveis preditoras para os sistemas fotovoltaicos apresentados. Foram utilizados dois algoritmos: *Random Forest* e KNN.

Nesse sentido, para entender quais variáveis são mais importantes para prever a potência gerada, a partir da variação da pureza dos nós, foi realizada a análise *IncNodePurity* (*Increase in Node Purity*), para o método *Random Forest*. Os resultados são apresentados na figura 19.

**Figura 19 - Resultados da métrica IncNodePurity**



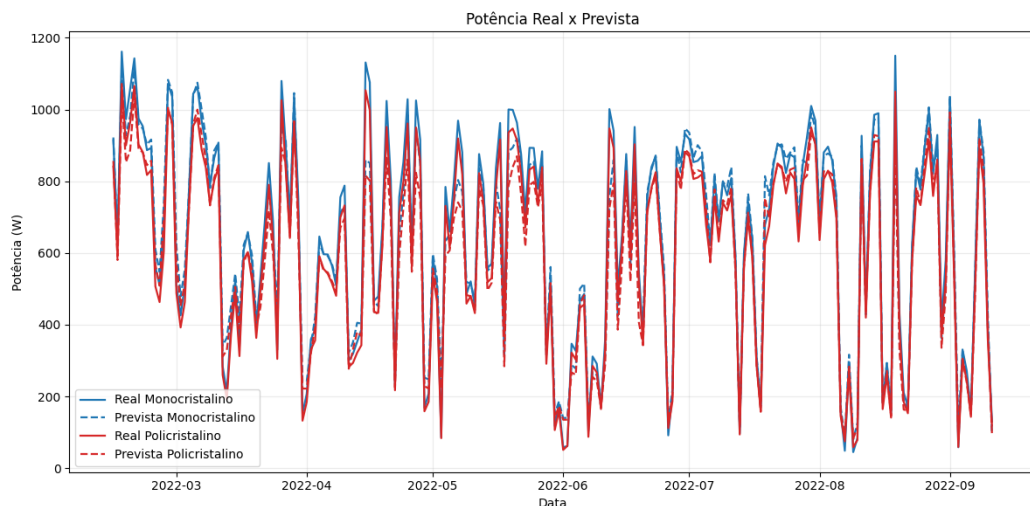
Fonte: Autores (2025)



A partir da interpretação da métrica apresentada na figura 19, nota-se que a variável que mais reduz a impureza de um nó da árvore de decisão é a irradiância média ( $2.482112e+07$ ), ou seja, ela é a covariável de maior importância, reduzindo a variância da previsão do modelo, o que também é avaliado nos gráficos das figuras 17 e 18. Em seguida, a covariável mais importante para o modelo analisado é a temperatura ( $1.401298e+07$ ), o que destaca a influência, geralmente refletida em energia dispersiva, dessa variável para a eficiência do sistema. Por fim, a tecnologia ( $3.195231e+05$ ) também possui influência nas árvores de decisão, ainda que bem menor que os outros parâmetros analisados, essa característica viabiliza o estudo comparativo das diferentes tecnologias construtivas de painéis fotovoltaicos.

A aplicação do estudo por meio do algoritmo *random forest* comparou as duas tecnologias, trazendo resultados similares a potência real, no quesito de previsão para as duas tecnologias, dentro do período analisado. Essa proximidade significa que o algoritmo utilizado conseguiu interpretar com qualidade as relações físicas entre irradiância média, temperatura e tecnologia. Os resultados da potência real versus a prevista, para os dois tipos construtivos de tecnologias, estão presentes nos gráficos da figura 20.

**Figura 20 - Resultado da previsão pelo algoritmo Random Forest**



**Fonte: Autores (2025)**

A partir da análise visual da figura 20, nota-se que o modelo não acumulou tantos erros, ou seja, não há aumento na defasagem entre previsão e dados reais, o que demonstra a capacidade do algoritmo *random forest* em registrar padrões estáveis com o passar do tempo. Além disso, o melhor desempenho do modelo ocorre próximo aos valores médios e a maior diferença entre a potência média prevista e a

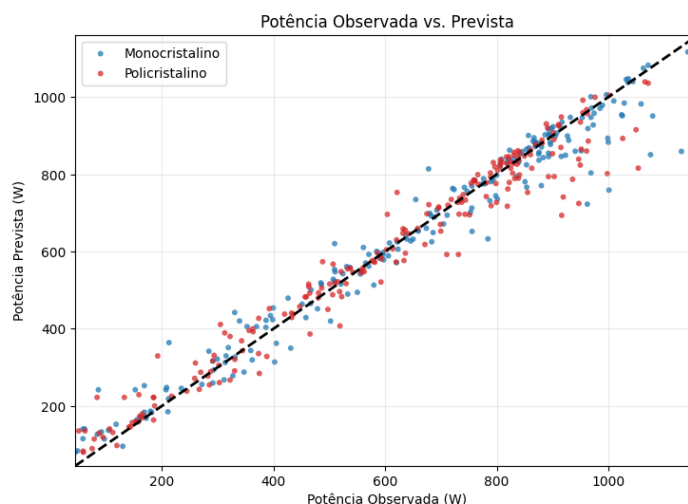
média real ocorre em situações de picos, sendo eles altos ou baixos. Essa é uma característica do modelo treinado, já que ele procura evitar o *overfitting*, buscando suavizar os valores extremos. Por esse motivo, diferenças mais significativas entre potências foram previstas com menor precisão.

Outro elemento que evidencia a boa adequação do modelo são os valores de RMSE e  $R^2$ . Para o modelo estudado, o RMSE foi igual a 58,409 W, enquanto o coeficiente de determinação  $R^2$  foi igual a 0,958. O valor elevado se dá pela relação física forte entre as covariáveis estudadas e a potência. O valor da raiz do erro quadrático médio demonstra que, em média, o algoritmo treinado contém um erro de 58,409 W, ou seja, a taxa de erro da previsão é de aproximadamente 58 unidades de potência. Ao analisar o coeficiente de determinação ( $R^2$ ), tem-se que 95,8% da variável resposta (potência) é explicada pelas covariáveis preditoras, irradiância média e temperatura. Isso demonstra que o modelo escolhido descreve de forma eficiente a variabilidade dos dados da variável de resposta.

Além de estudar a adequação do modelo via  $R^2$  e RMSE, pode-se analisar a adequação do modelo em aspectos gráficos, separando por tecnologia. Ao fazer isso nota-se que quase todos os pontos estão próximos da linha diagonal, que representa um cenário ideal, no qual a previsão é perfeita e estima de forma correta todas as amostras da base. Ambas as tecnologias foram muito bem previstas pelo modelo *random forest*, dado que existem poucos pontos superestimados ou subestimados, demonstrado pela proximidade dos pontos da linha diagonal.

Como as amostras não se afastam tanto da linha diagonal com o aumento da potência, pode-se afirmar que não há heterocedasticidade na base. Os poucos pontos que têm esse comportamento podem ser resultado de falhas ou ruídos nos sensores, pouca precisão nas medições e/ou um sombreamento momentâneo nos sistemas fotovoltaicos. O gráfico referente a essa discussão está apresentado na figura 21.

**Figura 21 - Dispersão da potência prevista para as duas tecnologias, utilizando o método *Random Forest***



**Fonte: Autores (2025)**

A partir da previsão de potência média gerada pelo algoritmo *random forest* foi possível estimar um valor médio mensal de energia para os meses entre janeiro e setembro de 2022, sendo que o mês de setembro tinha 11 dias registrados na base. Os resultados de energia em kWh são apresentados na tabela 2.

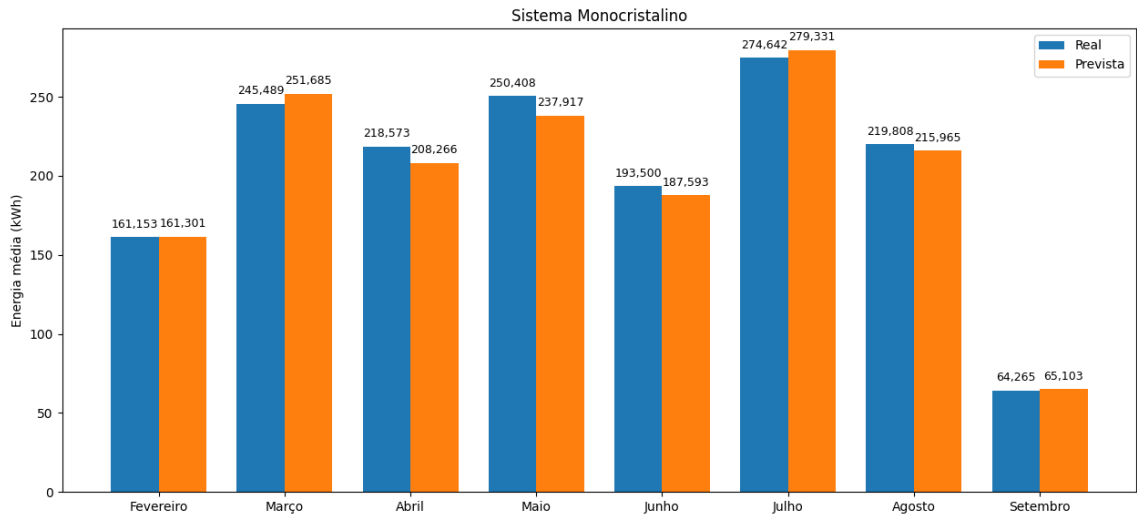
**Tabela 2 – Estatísticas de energia produzida pelos dois sistemas, em kWh, pelo método *Random Forest*.**

<b>Energia Média (kWh)</b>		
<b>Mês</b>	<b>Monocristalino</b>	<b>Policristalino</b>
<b>Fevereiro</b>	161,301	149,504
<b>Março</b>	251,685	233,629
<b>Abril</b>	208,266	192,199
<b>Maio</b>	237,917	218,731
<b>Junho</b>	187,592	173,815
<b>Julho</b>	279,330	251,134
<b>Agosto</b>	215,964	198,385
<b>Setembro</b>	65,102	60,626

**Fonte: Autores (2025)**

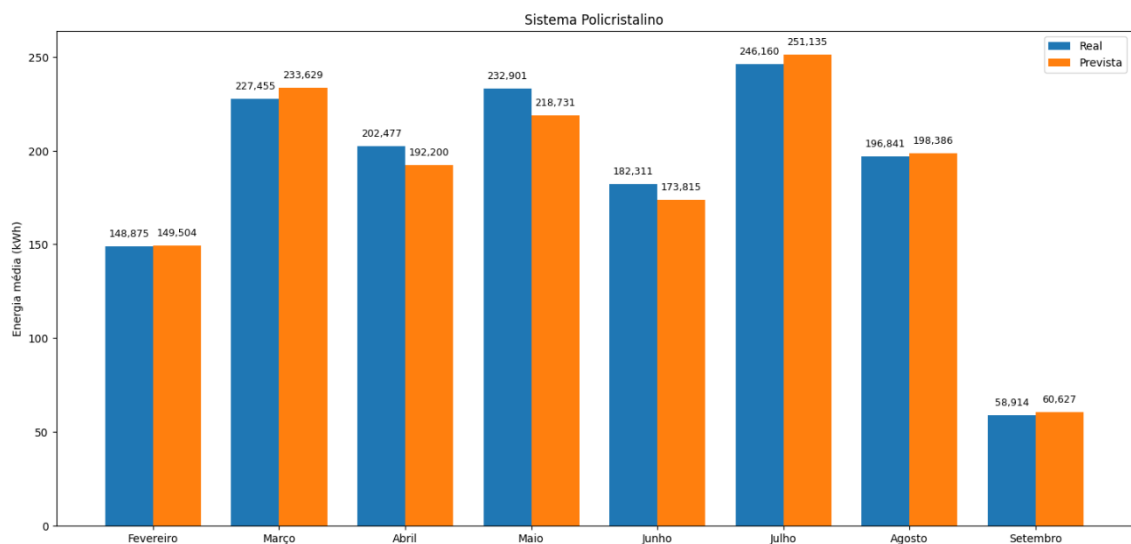
Ao comparar a energia média estimada com a energia média real, nota-se que os dois valores são muito próximos, reafirmando a qualidade e aderência do modelo de predição da potência. Os gráficos comparativos para as duas tecnologias analisadas são apresentados nas figuras 22 e 23.

**Figura 22 - Gráfico de comparação da estimativa de energia média com a energia média real, para a tecnologia monocristalina pelo modelo Random Forest**



Fonte: Autores (2025)

**Figura 23 – Gráfico de comparação da estimativa de energia média com a energia média real, para a tecnologia policristalina pelo modelo Random Forest**



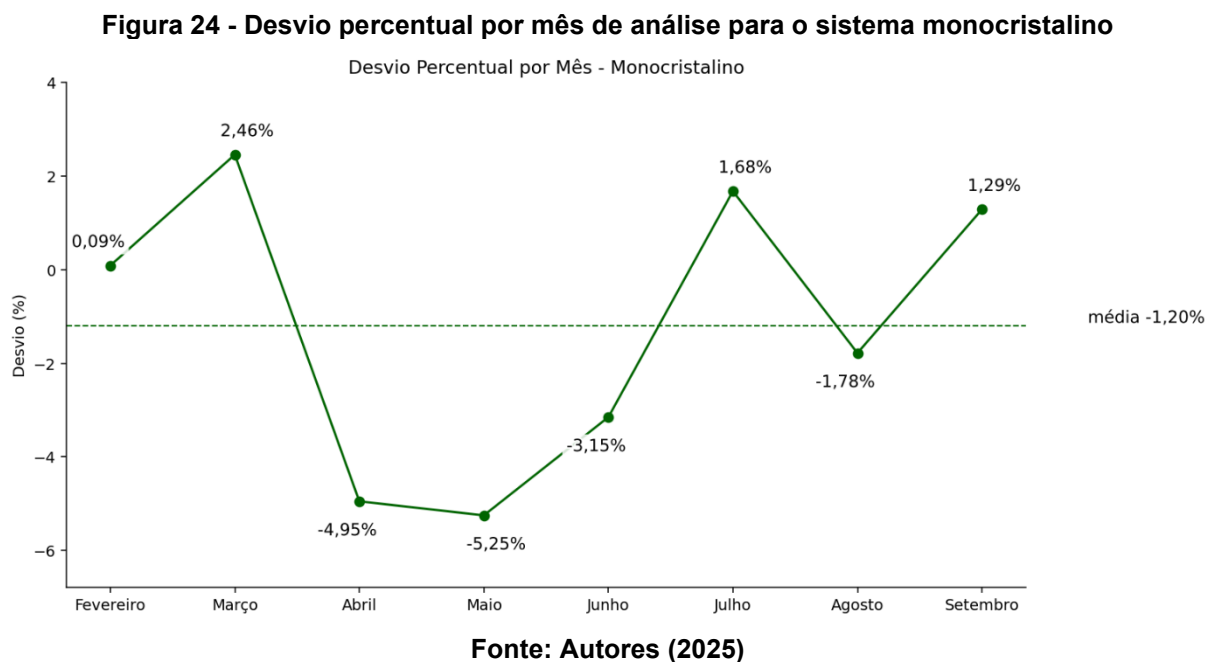
Fonte: Autores (2025)

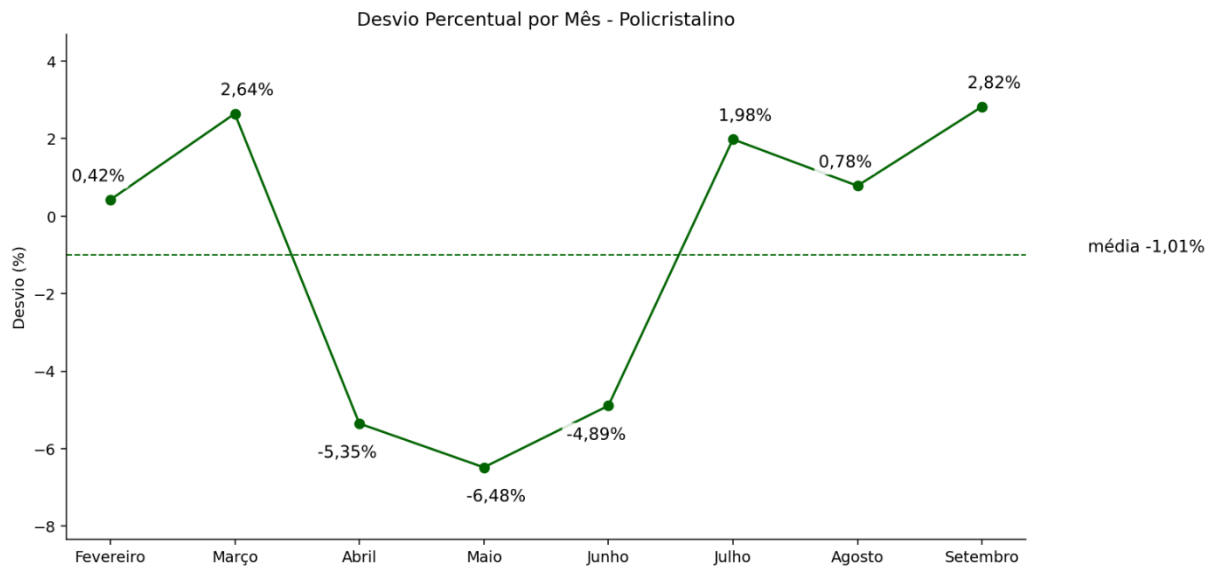
A partir do estudo dos gráficos das figuras 22 e 23, bem como da tabela 2, percebe-se que as estimativas resultantes de energia média prevista ficaram muito próximas à energia média gerada para os dois sistemas. Como as potências dos dois sistemas analisados tiveram resultados próximos, o resultado da estimativa de energia também respeitou esse critério.

Outro fator que contribui para esse resultado foi a escolha do intervalo de tempo diário estudado, das 06h até às 18h, já que existem dias em que os níveis de irradiância podem variar significativamente, com tardes ensolaradas e manhãs e

entardeceres nublados. Nesse sentido, a estimativa de energia média pode ser deslocada quando influenciada por esses eventos extremos durante o período selecionado, que apresentam relações não lineares entre irradiância e temperatura e a tecnologia.

No entanto, apesar de muito próximos, em nenhum dos meses analisados a estimativa de previsão se igualou a estimativa de energia gerada, sendo que alguns meses tiveram a estimativa de energia foi superestimada ou subestimada. Porém, o comportamento dessas diferenças foi razoavelmente semelhante para as duas tecnologias. Os gráficos que ilustram essa situação para os sistemas monocristalino e policristalino estão nas figuras 24 e 25, respectivamente.

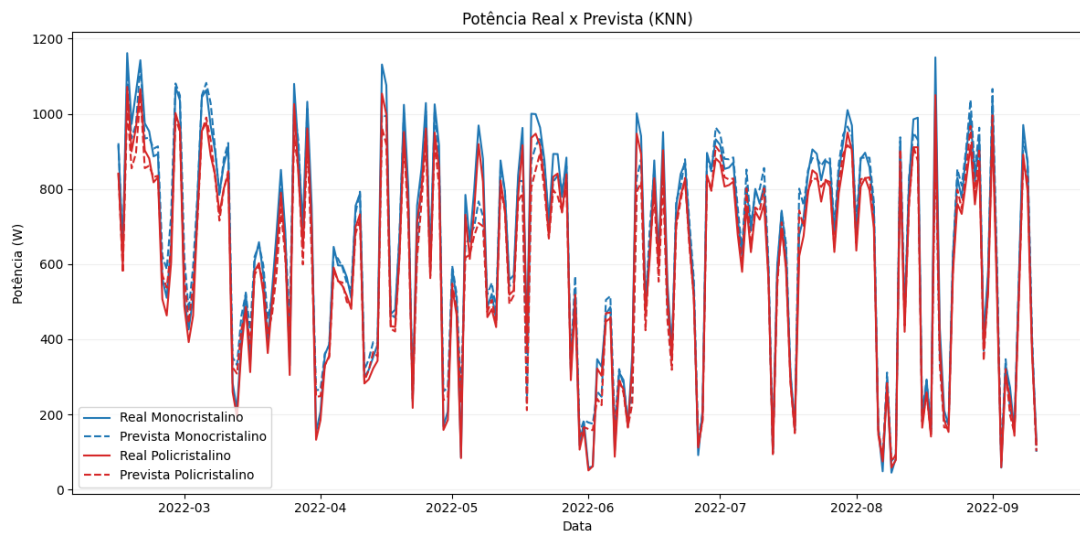


**Figura 25 - Desvio percentual por mês de análise para o sistema policristalino**

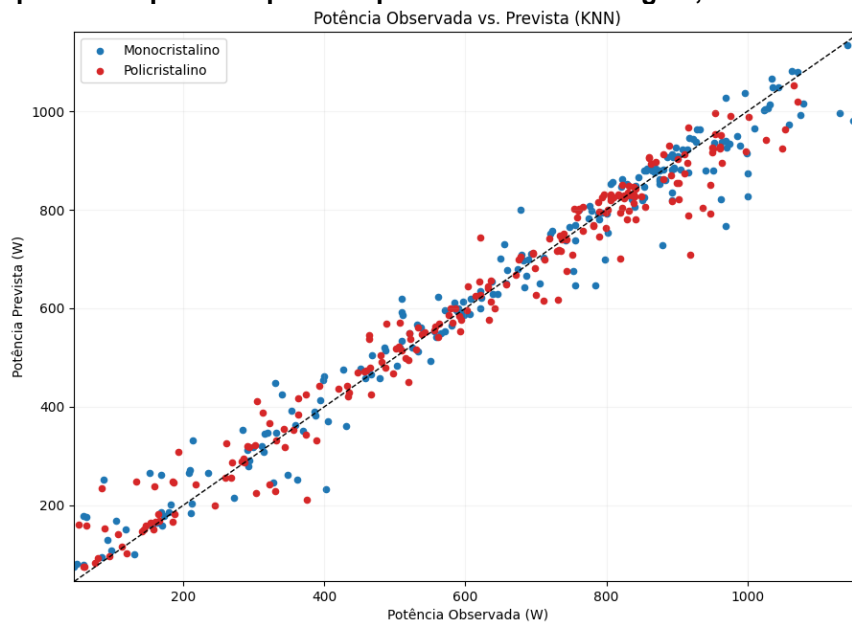
**Fonte: Autores (2025)**

A partir do estudo dos gráficos das figuras 24 e 25 é possível notar que a média do desvio de energia foi de -1,20% para o sistema monocristalino e -1,01% para policristalino. Isso significa que para ambas as tecnologias o valor real de energia média gerada foi menor que a energia média prevista para os meses, demonstrando que as estimativas superestimaram os dados reais em 0,20 e 0,01% para as tecnologias monocristalino e policristalino, respectivamente.

Para o modelo KNN (*K-Nearest Neighbors*), tem-se que o  $R^2$  foi de 0,967 e o RMSE foi igual a 51,99W. Para esse modelo, a raiz do erro quadrático médio desmontou que a predição pode divergir até aproximadamente 52 W ( $52/1200 = 4\%$ ) do valor real, o que ainda assim é um valor baixo, comparado a média da potência real dos sistemas, apresentada na tabela 1. Além disso, pela observação do  $R^2$ , foi demonstrada uma adequação melhor do algoritmo, em comparação ao estudo anterior, resultando em uma série mais ajustada aos valores reais. Isso pode ser observado no gráfico da figura 26.

**Figura 26 - Resultado da predição pelo algoritmo KNN****Fonte: Autores (2025)**

O  $R^2$  muito próximo de 1 também pode ser estudado a partir da dispersão do gráfico de dispersão das potências médias previstas em função das potências médias geradas pelos sistemas. O resultado está expresso no gráfico da figura 27.

**Figura 27 - Dispersão da potência prevista para as duas tecnologias, utilizando o método KNN****Fonte: Autores (2025)**

O gráfico da figura 27 evidencia que o ajuste do modelo, via algoritmo KNN foi quase perfeito, principalmente para as potências médias, entre 600 e 800 W. A partir dessas informações também foi estimada a energia média que cada um dos sistemas poderia gerar. Seus resultados foram concatenados dentro dos meses e estão apresentados na tabela 3.

Tabela 3 – Estatísticas de energia produzida pelos dois sistemas, em kWh, pelo método KNN.

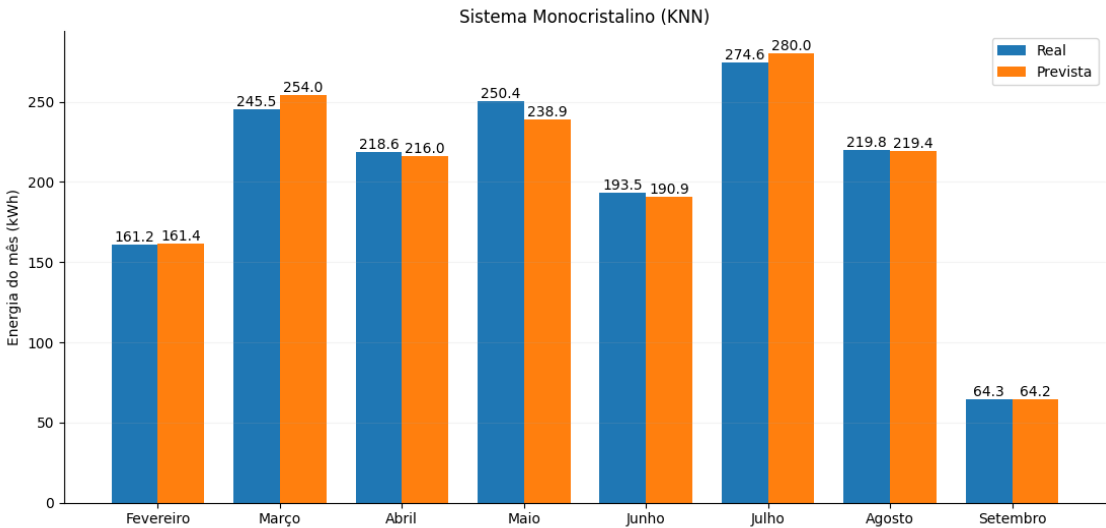
Energia Média (kWh)

Mês	Monocristalino	Policristalino
Fevereiro	161,368	147,804
Março	254,011	232,788
Abril	215,973	199,973
Maiο	238,89	223,455
Junho	190,915	176,807
Julho	279,952	262,581
Agosto	219,355	205,6
Setembro	64,195	60

Fonte: Autores (2025)

O comportamento da previsão de energia média para o algoritmo KNN mostrou valores próximos entre as duas tecnologias, o que também foi observado nas análises anteriores, relacionadas à potência média. No entanto, pequenas diferenças foram observadas, sendo que a tecnologia policristalina gerou menos energia nos meses analisados. Quanto ao comportamento das diferenças entre a energia média gerada e energia média prevista, os resultados foram plotados no gráfico da figura 28 e 29, para os sistemas monocristalino e policristalino, respectivamente.

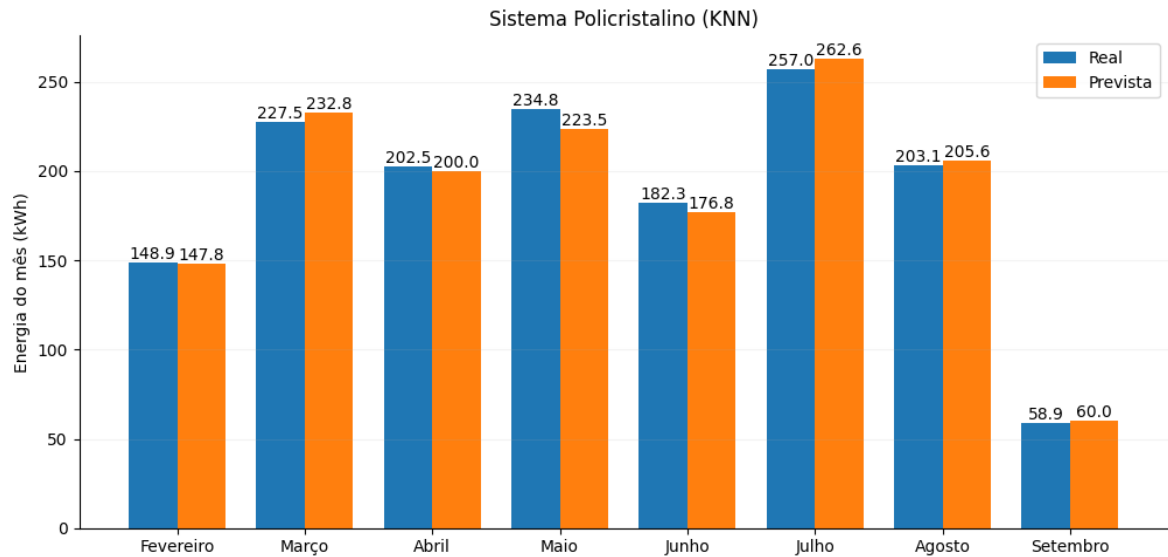
Figura 28 - Gráfico de comparação da estimativa de energia média com a energia real, para a tecnologia monocristalina, pelo modelo KNN



Fonte: Autores (2025)

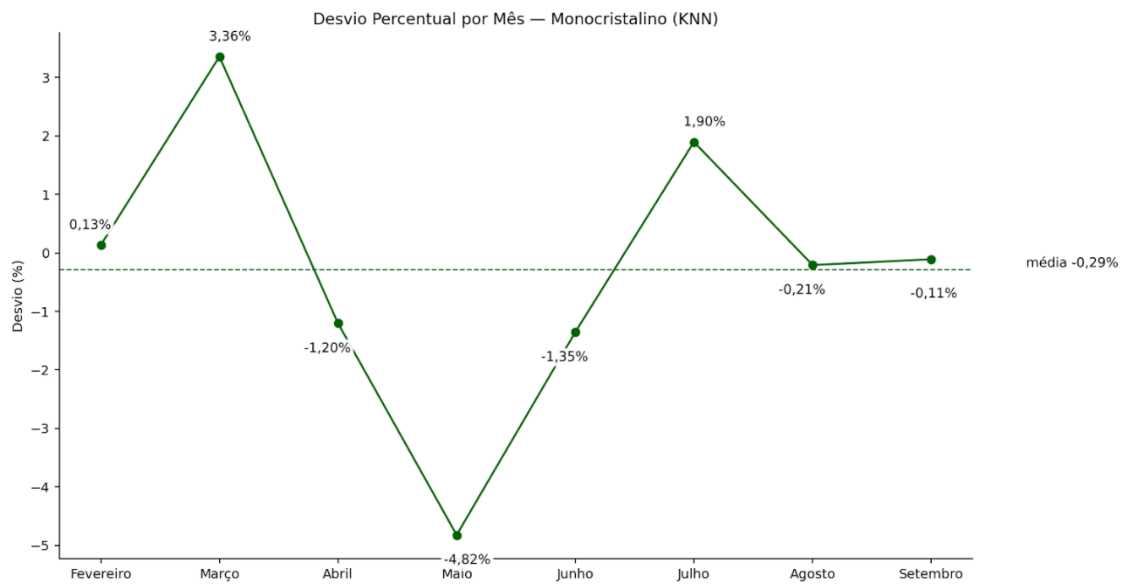
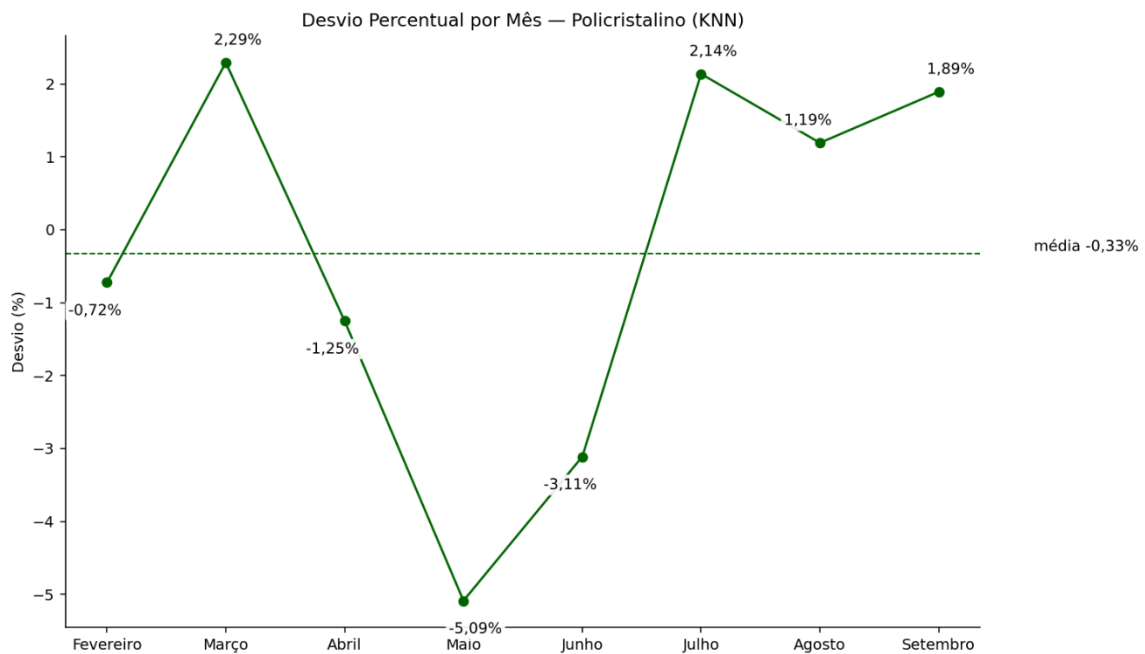


**Figura 29 - Gráfico de comparação da estimativa de energia média com a energia real, para a tecnologia policristalina, pelo modelo KNN**



**Fonte: Autores (2025)**

Ambos os gráficos demonstram as diferenças de produtividade energética entre os meses, além de trazerem a perspectiva de comparação entre o observado e o previsto. No entanto, algo em comum para as duas tecnologias foi o fato de que março e julho foram os meses de maior geração de energia, dentro do intervalo analisado. Analisando os dois algoritmos utilizados para predição, nota-se que o KNN obteve um desempenho melhor ao descrever as diferenças entre os dados previstos e observados. Esses resultados podem ser observados nos gráficos das figuras 30 e 31, que descrevem o desvio percentual de energia média, para as tecnologias monocristalino e policristalino, respectivamente.

**Figura 30 - Desvio percentual por mês de análise para o sistema monocristalino****Fonte: Autores (2025)****Figura 31 - Desvio percentual por mês de análise para o sistema policristalino****Fonte: Autores (2025)**

A partir da observação dos resultados foi possível inferir que o comportamento dos desvios seguiu uma curva semelhante para os dois sistemas analisados. Além disso, para os meses de abril, maio e junho os resultados foram negativos, o que significa que nesses meses a energia real foi menor que a prevista, destaca-se o mês

de maio, onde a energia média gerada foi 5,09% menor do que aquela estimada a partir da predição da potência.

Outro indicativo de que o algoritmo KNN teve um melhor desempenho comparado com o *Random Forest* foi a média do desvio percentual da energia média, que ficou menor que a calculada pelo algoritmo anterior. Apesar de ter apresentado resultados melhores, o modelo KNN também superestimou a energia média, assim como o *Random Forest*, já que a média percentual foi negativa. Além disso, a média de desvio foi menor do que a apresentada pelo algoritmo anterior, sendo -0,29% para o sistema monocristalino e -0,33% para o sistema policristalino.

## 5 CONCLUSÃO

A análise comparativa da geração de energia de dois sistemas fotovoltaicos do LABENS, instalados na sede Neville do campus UTFPR-CT, com tecnologias de silício monocristalino e policristalino, realizada através de abordagens estatísticas gerais, assim como de *machine learning*, visou estimar e comparar a energia gerada por ambas as tecnologias no período estudado e identificar o desempenho dos modelos preditivos testados.

Ainda que a base de dados fornecida possuísse lacunas de dados, foi possível realizar as análises estatísticas gerais buscando comparar as diferentes tecnologias construtivas e relacioná-las com as variáveis avaliadas, resultando em coeficientes de correlação na faixa de 0,75 para temperatura e 0,99 para irradiância média. Associados aos histogramas, diagramas de dispersão e *boxplots*, é possível inferir que as tecnologias possuem pequenas diferenças entre si, porém estas não são suficientes para definir qual o sistema com melhor desempenho em geração de energia elétrica.

Os resultados de modelagem apresentaram boa relação entre previsões e medições reais, sendo que o modelo de previsão KNN ( $R^2 = 0,967$  e RMSE = 51,99 W) obteve melhor desempenho que o modelo *random forest* ( $R^2 = 0,958$  e RMSE = 58,41 W), ao custo de possuir um maior tempo de processamento para obtenção de seus resultados.

Os dois modelos analisados apresentam  $R^2$  alto, significando que as variáveis selecionadas explicam de forma coerente a variabilidade da potência gerada pelo sistema. Já o RMSE é relativamente baixo, ao comparar com a média de potência no inversor dos sistemas. Apesar de assumirem um coeficiente de determinação alto, o modelo KNN alcançou a raiz dos erros quadráticos médios menor do que aqueles apresentados pelo *random forest*. Muito dessa dinâmica pode ser explicada pelo funcionamento dos algoritmos utilizados. O valor mais baixo de RMSE do KNN demonstra que esse algoritmo é mais eficiente ao avaliar não linearidades e as relações entre as covariáveis estudadas, o que é um aspecto positivo, já que fenômenos físicos e naturais nem sempre são lineares, já que depende da integração de diversos fatores.

O comportamento de potência média produzida, verificado a partir das análises por *machine learning*, mostram-se coerentes com o efetivamente gerado,

considerando que a irradiância pode variar aproximadamente 15%, para mais ou para menos, característica que impacta na potência e, conseqüentemente, na energia gerada. Além disso, para ambos os algoritmos selecionados notou-se que os resultados apresentados convergem em valores próximos ao que realmente foi observado nas plantas. Isso significa que ambos os modelos foram capazes de observar padrões dentro das séries temporais e refleti-los no treinamento dos dados. O que demonstra esse fato são os resultados dos gráficos das séries temporais resultantes, bem como o gráfico de dispersão, cujos pontos estão próximos da diagonal ideal, onde  $R^2$  é próximo de 1.

Uma vez que as previsões de potência média foram de alta precisão, a obtenção dos valores de energia prevista diariamente também resultou em um alto grau de precisão, com relação aos valores medidos, com médias de desvio de energia de -0,29% para o modelo KNN e -1% para o modelo *random forest*. Esta precisão em um modelo preditivo de energia é demasiadamente importante para controle de geração e planejamento de despacho de energia, realizado no Brasil pelo Operador Nacional do Sistema (ONS). É essencial que o órgão tenha conhecimento da quantidade de energia que será gerada em um dado intervalo de tempo, de modo a realizar planejamentos de operação de usinas geradoras de energia elétrica, além de manejar adequadamente a relação entre demanda e produção de energia.

Comparando as duas tecnologias estudadas, silício monocristalino e policristalino, notou-se que existe uma diferença em potência produzida e, por conseguinte, na energia gerada. No entanto, essa disparidade não foi tão significativa para avaliação dos resultados, quanto a irradiância, que foi um fator decisivo, seguida pela temperatura dos módulos.

De modo a expandir a abrangência da proposta que foi apresentada neste trabalho, é possível realizar análises por aprendizado de máquina utilizando outros métodos, buscando aquele que melhor se adequa à situação de estudo, assim como considerar um maior número de variáveis, como umidade e velocidade do vento, por exemplo. Além disso, a análise em um intervalo de tempo mais reduzido (5 a 15 minutos) pode ser uma alternativa mais aplicável na geração fotovoltaica, já que o fator climático pode causar alterações mais drásticas quando a avaliação é realizada em intervalos maiores de tempo. Ademais, a automatização do processo de coleta de informações dos bancos de dados, seu tratamento e a realização dos cálculos e

execução de algoritmos tornaria a utilização deste procedimento mais fácil e intuitiva, podendo ser utilizada por operadores e tomadores de decisão.

## REFERÊNCIAS

ALPAYDIN, E. ***Introduction to Machine Learning***, p. 1-14, 2004. ISBN 0-262-01211-1.

BECQUEREL, A.-E. ***Mémoire sur les effets électriques produits sous l'influence des rayons solaires***. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, v. 9, p. 561–567, 1839.

BHARDWAJ, C. A.; MISHRA, M.; DESIKAN, K. ***Dynamic feature scaling for k-nearest neighbor algorithm***. arXiv preprint arXiv:1811.05062, 2018.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. ***Time Series Analysis Forecasting and Control***, p. 2-15, 2016. ISBN 978-1-118-67502-1.

BURKOV, A.. ***The hundred-page machine learning book***. 2019.

CEMIG - COMPANHIA ENERGÉTICA DE MINAS GERAIS. ***Alternativas Energéticas: uma visão Cemig***. Belo Horizonte: CEMIG, 2012.

CEPEL – CENTRO DE PESQUISAS DE ENERGIA ELÉTRICA. ***As energias solar e eólica no Brasil***. 2013. Disponível em:

<http://cresesb.cepel.br/download/casasolar/casasolar2013.pdf>. Acesso em: janeiro de 2025.

CHAPIN, D. M.; FULLER, C. S.; PEARSON, G. L. ***A New Silicon p-n Junction Photocell for Converting Solar Radiation into Electrical Power***. *J. Appl. Phys.*, p. 676-677, 1954. Acesso em: janeiro de 2025.

CRYER, J. D.; CHAN, K. S. ***Time series analysis with applications in R***. 2. ed. New York: Springer, 2008.

DA ROCHA, A. L. S. **IMPLEMENTAÇÃO E ANÁLISE ESPAÇO-TEMPORAL DO VALOR PRESENTE LÍQUIDO (DINÂMICO) EM RESERVATÓRIOS DE PETRÓLEO**. PPGCEP - Doutorado em Ciência e Engenharia do Petróleo. Natal, Brasil: Universidade Federal do Rio Grande do Norte, 2024.

DE GIORGI, MARIA GRAZIA; CONGEDO, PAOLO MARIA; MALVONI, MARIA. **Photovoltaic power forecasting using statistical methods: impact of weather data**. IET Science, Measurement & Technology, v. 8, n. 3, p. 90-97, 2014.

DEVORE, J. L.; CORDEIRO. **Probabilidade e estatística: para engenharia e ciências**. Cengage Learning Edições Ltda., 2006.

DGS. **Planning and installing photovoltaic systems: a guide for installers, architects, and engineers**. 2 ed. London: Earthscan, 2008.

EHLERS, R. S. **ANÁLISE DE SÉRIES TEMPORAIS**. [s.l: s.n.].

Ember Climate. "Yearly electricity data". 2023. Disponível em: <https://ember-energy.org/data/yearly-electricity-data/>. Acesso em: março de 2025.

EPIA; GREENPEACE. **Solar Generation 6: solar photovoltaic electricity empowering the world**. 2011. Disponível em: [http://pvtrin.eu/assets/media/PDF/Publications/Other%20Publications/36.SolarGeneration6\\_\\_2011.pdf](http://pvtrin.eu/assets/media/PDF/Publications/Other%20Publications/36.SolarGeneration6__2011.pdf). Acesso em: maio de 2025.

FRAUNHOFER ISE. **Photovoltaics report**. 2024. Disponível em: <https://www.ise.fraunhofer.de/en/publications/studies/photovoltaics-report.html>. Acesso em: maio de 2025.

GOGINENI, Abhilash et al. **Predictive modelling of concrete compressive strength incorporating GGBS and alkali using a machine-learning approach**. *Asian Journal of Civil Engineering*, v. 25, n. 1, p. 699-709, 2024.

GRUPO DE TRABALHO DE ENERGIA SOLAR. **Manual de engenharia para sistemas fotovoltaicos**. Edição especial, CRESESB. Rio de Janeiro, 1999.

HALDER, R.K., UDDIN, M.N., UDDIN, M.A. et al. **Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications**. *J Big Data* 11, 113 (2024). <https://doi.org/10.1186/s40537-024-00973-y>. Acesso em: junho de 2025.

HASTIE, Trevor. **The elements of statistical learning: data mining, inference, and prediction**. 2009.



IMHOFF, J. **Desenvolvimento de Conversores Estáticos para Sistemas Fotovoltaicos Autônomos**. Dissertação de Mestrado apresentada à Escola de Engenharia Elétrica da Universidade Federal de Santa Maria, Santa Maria. 2007. 146 f.

*International Energy Agency. Photovoltaic Power Systems Programme.*  
Disponível em: <https://iea-pvps.org/>. Acesso em: fevereiro de 2025.

IZBICKI, R.; DOS SANTOS, T. M. **Aprendizado de Máquina - Uma Abordagem Estatística**. São Carlos. 2020. ISBN 978-65-00-02410-4

LANA, L. T. C.; ALMEIDA, E.; DIAS, F. C. L. S.; ROSA, A. C.; DO ESPÍRITO SANTO, O. C.; SACRAMENTO, T. C. B.; BAZ, K. T. M. **Energia solar fotovoltaica: revisão bibliográfica**. *Engenharias On-line*, 2015, 21-33.

LUQUE, A.; HEGEDUS, S.. **Handbook of Photovoltaic Science and Engineering**. 2. ed. Chichester: *John Wiley & Sons*, 2011. ISBN 9780470721698.

MACHADO, C. T.; MIRANDA, F. S. **Energia Solar Fotovoltaica: Uma Breve Revisão**. *Rev. Virtual Quim.*, 2015, 126-143. Data de publicação na Web: 14 de outubro de 2014.

MACHADO MARQUEZ ZAMPIVA, M. **Análise multivariada do impacto das variáveis meteorológicas locais no desempenho da geração de energia elétrica de um sistema fotovoltaico conectado à rede**. 100 f. Tese (Programa de Pós-Graduação em Engenharia de Energia na Agricultura) - Universidade Estadual do Oeste do Paraná, Cascavel, PR, 2024. Disponível online em: <https://tede.unioeste.br/handle/tede/7475>. Acesso em: janeiro de 2025.

MAGALHÃES, M. N.; LIMA, A. C. P. **Noções de Probabilidade e Estatística**. 7 ed. São Paulo: EDUSP, 2011.

MENEZES NUNES, H.; PHELIPE DA SILVA CORREA, R.; MALAGOLI BUIATTI, G.; LATUF ANDRADE, R.; KOS BOTELHO, F.; VIEIRA NETO, J.; DE CARVALHO VENANCIO, V.; SILVA CARVALHO, R. **ESTUDO COMPARATIVO TÉCNICO-FINANCEIRO DE USINAS FOTOVOLTAICAS UTILIZANDO MÓDULOS DE SILÍCIO MONOCRISTALINOS E POLICRISTALINOS**. *Anais Congresso Brasileiro*

de Energia Solar - CBENS, [S. l.], 2020. Disponível online em: <https://anaiscbens.emnuvens.com.br/cbens/article/view/855>. Acesso em: janeiro de 2025.

MIT Open Course Ware. **Fundamentals of Photovoltaics**. Disponível em: <http://ocw.mit.edu/courses/mechanical-engineering/2-626-fundamentals-of-photovoltaics-fall-2008/lecture-notes/lecture9.pdf>. Acesso em: fevereiro de 2025.

MME, 2023. Ministério de Minas e Energia. **Balanço Energético Nacional 2023: Ano base 2022**. Disponível online em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-748/topico-687/BEN2023.pdf>. Acesso em: janeiro de 2025.

MME, 2024. Ministério de Minas e Energia. **Balanço Energético Nacional 2024: Ano base 2023**. Disponível online em: <https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-819/topico-723/BEN2024.pdf>. Acesso em: janeiro de 2025.

MME, 2025. Ministério de Minas e Energia. **Balanço Energético Nacional 2025: Ano base 2024**. Disponível online em: [https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-885/topico-771/Relatório%20Final\\_BEN%202025.pdf](https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-885/topico-771/Relatório%20Final_BEN%202025.pdf). Acesso em: setembro de 2025.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. 2010.

MORETTIN, P. A.; TOLOI, C.M.C. **Análise de Séries Temporais**. 3. ed. São Paulo: Editora Blucher, 2018.

MORI, V.; SANTOS, R. L. C.; SOBRAL, L. G. S. **Metalurgia do silício: processos de obtenção e impactos ambientais**. 1a. ed., CETEM/MCT. Rio de Janeiro, 2007.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística Aplicada e Probabilidade para Engenheiros**. 7 ed. Rio de Janeiro: LTC, 2021.

MONTGOMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. **Introduction to Time Series Analysis and Forecasting**, p. 1-14, 2007. ISBN 978-0-471-65397-4.

NASCIMENTO, V. A. M. A.; TRINDADE, T.B.; CARVALHO, C.M. **Análise dos parâmetros para geração de energia solar fotovoltaica no Acre, Brasil.**

InterEspaço: Revista de Geografia e Interdisciplinaridade, v. 7, e202129, 2021.

Disponível em: <http://dx.doi.org/10.18764/2446-6549.e202129>. Acesso em: abril de 2025.

NEHA, S V S T; YADAV, Y.; GOYAL, Y.. **Introduction to Machine Learning.**

*International Journal of Advanced Research in Science, Communication and Technology* (IJARSCT), India, v. 4, n. 2, 2 mar. 2024. DOI 10.48175/IJARSCT-

15723. Disponível em: <https://ijarsct.co.in/Paper15723.pdf>. Acesso em: janeiro de 2025.

ONAIFO, F.; OKANDEJI, A.; AJETUNMOBI, O.; BALOGUN, D., 2021. **EFFECT OF TEMPERATURE, HUMIDITY AND IRRADIANCE ON SOLAR POWER**

**GENERATION.** *Journal of Engineering Studies and Research*. 26. 113-119.

10.29081/jesr.v26i4.243.

PEREIRA, E. B.; MARTINS, F. R.; GONÇALVES, A. R.; COSTA, R. S.; LIMA, F. J. L.; RÜTHER, R.; ABREU, S. L.; TIEPOLO, G. M.; PEREIRA, S. V.; SOUZA, J. G.

**Atlas brasileiro de energia solar.** 2ª edição. São José dos Campos: INPE, 2017.

88 p. ISBN 978-85-17-00090-4. IBI: 8JMKD3MGP3W34P/3PERDJE. Disponível em: <http://urlib.net/ibi/8JMKD3MGP3W34P/3PERDJE>. Acesso em: janeiro de 2025.

PEREIRA SALGADO JÚNIOR, A.; APARECIDO DOS SANTOS PIMENTEL, L;

MATTOS BORGES DE OLIVEIRA, M; CHIARETTI NOVI, J. **O IMPACTO NAS**

**VARIAÇÕES DAS MATRIZES ENERGÉTICAS E USO DA TERRA: ESTUDO**

**SOBRE A EFICIÊNCIA AMBIENTAL DO G20.** REAd (Revista Eletrônica de

Administração), Porto Alegre, Edição 86, Nº 2, p. 306 - 332, 2017. Disponível online em: <https://doi.org/10.1590/1413.2311.013.62781>. Acesso em: janeiro de 2025.

REN21. **Renewables 2025 Global Status Report: Global Overview.** REN21

Secretariat, 2025. Disponível em: <https://www.ren21.net/renewables-2025-global-status-report-global-overview/>. Acesso em: julho de 2025.

ROSA, A. R. O.; GASPARIN, F. P. **Panorama da energia solar fotovoltaica no Brasil**. Revista Brasileira de Energia Solar, São Paulo, v. 7, n. 2, p. 140–147, dez. 2016.

SAMEERA; TARIQ, M.; RIHAN, M.. ***Analysis of the impact of irradiance, temperature and tilt angle on the performance of grid-connected solar power plant***. *Measurement: Energy*, v. 2, 2024. Disponível em: <https://doi.org/10.1016/j.meaeene.2024.100007>. Acesso em: janeiro de 2025.

SAMUEL, A. L. ***Some studies in machine learning using the game of checkers***. 1959.

SEVERINO, M.& OLIVEIRA, M. **Fontes e Tecnologias de Geração Distribuída para Atendimento a Comunidades Isoladas**. Energia, Economia, Rotas Tecnológicas: textos selecionados, Palmas, ano 1, p. 265-322, 2010.

SUJATHA B.G.; ANITHA G.S.. ***Enhancement of PQ in grid connected PV system using hybrid technique***. *Ain Shams Engineering Journal*, 2016. <http://dx.doi.org/10.1016/j.asej.2016.04.007>. Acesso em: janeiro de 2025.

URBANETZ JR, J.; TIEPOLO, G. M.; CASAGRANDE JR, E. F.; TONIN F. S.; MARIANO, J. D., 2016. **Geração Distribuída Fotovoltaica: O Caso dos Sistemas Fotovoltaicos da UTFPR em Curitiba**. X CBPE, Congresso Brasileiro de Planejamento Energético, pp. 12.

VALPECOVSKI URBANETZ, I.; DE MOURA NETTO, A.; SCOLARI, B.; LEITE, V.; URBANETZ JR, J. **Panorama Atual e Cenário 2025 da Energia Solar Fotovoltaica no Brasil**. Smart Energy 2018, Paraná. Acesso em janeiro de 2025.

VIEIRA, E.; MORAIS, J.; MENDES, F.; MACÊDO, A. **Energia renovável: os fundamentos da energia solar**. Tópicos Especiais em Engenharia: inovações e avanços tecnológicos, p. 102-109, 2022. Disponível online em: [https://www.researchgate.net/publication/361786751\\_Energia\\_renovavel\\_os\\_fundamentos\\_da\\_energia\\_solar](https://www.researchgate.net/publication/361786751_Energia_renovavel_os_fundamentos_da_energia_solar). Acesso em: janeiro de 2025.