UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS / FACULDADE DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

Caian Dutra de Jesus

Criação de Base de Dados Inédita de Usinas Solares e Aplicação em Previsão de Geração de Energia com Redes Neurais Otimizadas por Aprendizado por Transferência

Juiz de Fora

2025

Caian Dutra de Jesus

Criação de Base de Dados Inédita de Usinas Solares e Aplicação em Previsão de Geração de Energia com Redes Neurais Otimizadas por Aprendizado por Transferência

Dissertação Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, na área de concentração em Modelagem Computacional, como requisito parcial à obtenção do título de Mestre em Modelagem Computacional.

Orientador: Dr. Eduardo Pestana de Aguiar

Juiz de Fora

2025

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

de Jesus, Caian Dutra.
Criação de Base de Dados Inédita de Usinas Solares e Aplicação em Previsão de Geração de Energia com Redes Neurais Otimizadas por Aprendizado por Transferência / Caian Dutra de Jesus. – 2025.
79 f. : il.

Orientador: Eduardo Pestana de Aguiar
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas / Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2025.

1. Base de Dados. 2. Usinas Solares. 3. Aprendizado de Máquina. I. de Aguiar, Eduardo Pestana, orient. II. Título.

Caian Dutra de Jesus

Criação de Base de Dados Inédita de Usinas Solares e Aplicação em Previsão de Geração de Energia com Redes Neurais Otimizadas por Aprendizado por Transferência

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional.

Aprovada em 25 de setembro de 2025.

BANCA EXAMINADORA

Prof. Dr. Eduardo Pestana de Aguiar - Orientador
Universidade Federal de Juiz de Fora

Prof. Dr. Petrônio Cândido de Lima e Silva
Instituto Federal do Norte de Minas Gerais - Campus Januária

Prof. Dr. Arthur Caio Vargas e Pinto
Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais - Campus Itabirito

Juiz de Fora, 23/09/2025.

Documento assinado eletronicamente por **Eduardo Pestana de Aguiar**, **Professor(a)**, em 25/09/2025, às 11:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Arthur Caio Vargas e Pinto**, **Usuário Externo**, em 25/09/2025, às 14:21, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Petrônio Cândido de Lima e Silva**, **Usuário Externo**, em 25/09/2025, às 14:31, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2635852** e o código CRC **BEBB51B4**.

AGRADECIMENTOS

À Universidade Federal de Juiz de Fora (UFJF) e ao Programa de Pós-Graduação em Modelagem Computacional, pela oportunidade de formação e pelo ambiente acadêmico de excelência oferecido ao longo deste percurso.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pelo apoio financeiro concedido por meio da bolsa de estudos, essencial para a realização desta pesquisa.

Ao Laboratório de Automação Industrial e Inteligência Computacional (LAIIC), que acolheu e forneceu o suporte necessário para o desenvolvimento das investigações, em especial ao meu orientador, Dr. Eduardo Pestana de Aguiar, pela orientação criteriosa, incentivo contínuo e pela dedicação à condução deste trabalho.

À empresa TECSCI, pela parceria estabelecida que resultou na base de dados utilizada nesta dissertação e que possibilitou a consolidação dos resultados aqui apresentados.

A todos que, de forma direta ou indireta, contribuíram para a concretização deste trabalho, expresso minha sincera gratidão.

RESUMO

O crescimento acelerado da geração de energia solar fotovoltaica no Brasil evidencia a necessidade de bases de dados estruturadas e de alta qualidade, capazes de sustentar o desenvolvimento de novas tecnologias e de aprimorar a confiabilidade dos estudos no setor. Entre as principais demandas, destaca-se a criação de conjuntos de dados que viabilizem a comparação e a avaliação de modelos de previsão de geração com maior rigor científico. Nesse contexto, este trabalho tem como objetivo principal a criação e disponibilização da base de dados BR-PVGen, composta por registros históricos de geração elétrica e variáveis meteorológicas de 51 usinas fotovoltaicas em operação real no território brasileiro. A base foi tratada segundo critérios de qualidade, completude e padronização, constituindo um repositório inédito voltado ao suporte de pesquisas em previsão, análise de desempenho e otimização de sistemas fotovoltaicos.

Complementarmente, o trabalho apresenta uma aplicação de modelos de aprendizado de máquina, em especial Redes Neurais Artificiais com arquiteturas otimizadas por busca de hiperparâmetros e técnicas de *Transfer Learning*, aplicadas na mesma base de dados BR-PVGen. Os experimentos realizados mostraram que o Aprendizado por Transferência entre usinas com diferentes históricos de dados reduziu o erro médio de previsão, superando modelos de referência como *Gradient Boosting*, *Bayesian Ridge* e *Kernel Ridge*, com destaque para as métricas SMAPE e NRMSE.

Os resultados obtidos evidenciam que a BR-PVGen representa um avanço significativo na consolidação de dados fotovoltaicos no Brasil, promovendo reprodutibilidade e padronização de estudos na área. Conclui-se que a combinação entre uma base de dados multivariada, padronizada e de abrangência nacional, e o uso de técnicas modernas de aprendizado de máquina, constitui um caminho promissor para o aprimoramento das previsões e para a integração inteligente da energia solar ao sistema elétrico brasileiro.

Palavras-chave: Base de Dados; Usinas Solares; Aprendizado de Máquina; Aprendizado por Transferência.

ABSTRACT

The accelerated growth of solar photovoltaic generation in Brazil highlights the need for structured and high-quality databases capable of supporting the development of new technologies and improving the reliability of studies in the sector. Among the main challenges is the creation of datasets that enable the comparison and evaluation of generation forecasting models with greater scientific rigor. In this context, the primary objective of this work is the creation and public release of the BR-PVGen database, composed of historical records of electrical generation and meteorological variables from 51 photovoltaic power plants operating in real conditions across Brazil. The database was processed according to criteria of quality, completeness, and standardization, constituting an unprecedented repository designed to support research on forecasting, performance analysis, and optimization of photovoltaic systems.

In addition, the study presents an application of machine learning models, particularly Artificial Neural Networks with architectures optimized through hyperparameter search and *Transfer Learning* techniques, applied to the same BR-PVGen database. The experiments demonstrated that Transfer Learning between plants with different data histories reduced the average forecasting error, outperforming reference models such as *Gradient Boosting*, *Bayesian Ridge*, and *Kernel Ridge*, with emphasis on the SMAPE and NRMSE metrics.

The results show that BR-PVGen represents a significant advancement in the consolidation of photovoltaic data in Brazil, promoting reproducibility and standardization of studies in the field. It is concluded that the combination of a multivariate, standardized, and nationwide database with modern machine learning techniques constitutes a promising path for improving forecasting accuracy and for the intelligent integration of solar energy into the Brazilian power system.

Keywords: Database; Solar Power Plants; Machine Learning; Transfer Learning.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
AI	*Artificial Intelligence* (Inteligência Artificial)
AI/AO	*Analog Input / Analog Output* (Entrada Analógica / Saída Analógica)
API	*Application Programming Interface*
BRT	*Brasília Time* (Horário de Brasília)
CC	Corrente Contínua
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
DHI	*Diffuse Horizontal Irradiance* (Irradiância Difusa Horizontal)
DI/DO	*Digital Input / Digital Output* (Entrada Digital / Saída Digital)
DNI	*Direct Normal Irradiance* (Irradiância Direta Normal)
DNN	*Deep Neural Network* (Rede Neural Profunda)
GB	*Gradient Boosting*
GHI	*Global Horizontal Irradiance* (Irradiância Global Horizontal)
GD	Geração Distribuída
GW	Gigawatt
HTTP/HTTPS	*Hypertext Transfer Protocol / Hypertext Transfer Protocol Secure*
IEC	*International Electrotechnical Commission*
IPOA	Irradiância no Plano dos Módulos (*In-Plane of Array Irradiance*)
IRENA	*International Renewable Energy Agency* (Agência Internacional de Energia Renovável)
kWp	*kilowatt-peak* (quilowatt de pico)
LSTM	*Long Short-Term Memory* (Memória de Curto e Longo Prazo)
MAPE	*Mean Absolute Percentage Error* (Erro Percentual Médio Absoluto)
ML	*Machine Learning* (Aprendizado de Máquina)
MLP	*Multilayer Perceptron* (Perceptron Multicamadas)
MWp	*megawatt-peak* (megawatt de pico)
nMAE	*normalized Mean Absolute Error* (Erro Médio Absoluto Normalizado)
nRMSE	*normalized Root Mean Square Error* (Raiz do Erro Quadrático Médio Normalizado)
NREL	*National Renewable Energy Laboratory*
NWP	*Numerical Weather Prediction* (Previsão Numérica do Tempo)
ONS	Operador Nacional do Sistema Elétrico
PV	*Photovoltaic* (Fotovoltaico)
PVDAQ	*Photovoltaic Data Acquisition*
RC	*Reservoir Computing* (Computação de Reservatórios)
ReLU	*Rectified Linear Unit* (Unidade Linear Retificada)
RF	*Random Forest*
RNA	Rede Neural Artificial
RMSE	*Root Mean Square Error* (Raiz do Erro Quadrático Médio)
SCADA	*Supervisory Control and Data Acquisition*

SGD	*Stochastic Gradient Descent* (Gradiente Descendente Estocástico)
SMAPE	*Symmetric Mean Absolute Percentage Error* (Erro Percentual Médio Absoluto Simétrico)
STC	*Standard Test Conditions* (Condições Padrão de Teste)
SVR	*Support Vector Regression* (Regressão por Vetores de Suporte)
TL	*Transfer Learning* (Aprendizado por Transferência)
WMA	*Weighted Moving Average* (Média Móvel Ponderada)

LISTA DE SÍMBOLOS

$\forall$	Para todo
$\in$	Pertence a
A_{mod}	Área total dos módulos fotovoltaicos (m^2)
η_{mod}	Eficiência nominal do módulo
η_{irr}	Fator de correção da eficiência em função da irradiância
$\eta_{\text{mod,T}}$	Eficiência ajustada pela temperatura
β_T	Coeficiente térmico do módulo (%/°C)
γ	Coeficiente de bifacialidade
I_{POA}	Irradiância no plano dos módulos (W/m^2)
I_{GRI}	Irradiância global refletida inclinada (W/m^2)
P_{ideal}	Potência ideal teórica (W)
P_{bif}	Potência ajustada pelo ganho bifacial (W)
P_{exp}	Potência esperada final (W)
P_{clip}	Potência perdida por *clipping* (W)
$P_{\text{inv,max}}$	Potência nominal máxima de saída do inversor (W)
f_{deg}	Fator de degradação dos módulos
T_{mod}	Temperatura do módulo (°C)
T_{STC}	Temperatura em condições padrão de teste (25°C)
ρ	Coeficiente de correlação de Pearson
ℓ	Defasagem temporal (*lag*)
n	Número total de observações
t	Índice de tempo

SUMÁRIO

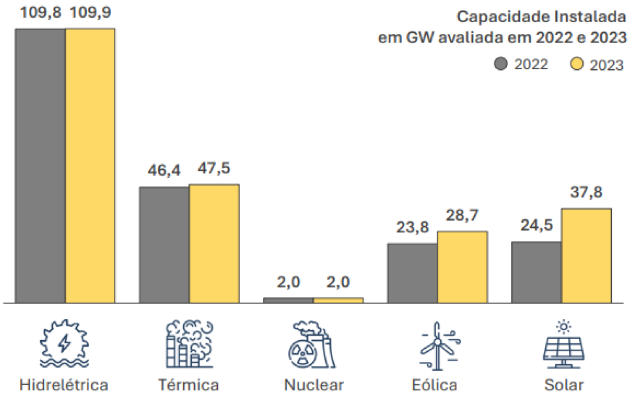
1 INTRODUÇÃO

O aumento constante da demanda mundial por energia, impulsionado pelo crescimento populacional, mudanças no estilo de vida e avanços tecnológicos, tem levado países a diversificarem suas matrizes energéticas. Essa necessidade é reforçada pela urgência em mitigar os impactos das mudanças climáticas, majoritariamente associadas à queima de combustíveis fósseis e pela escassez progressiva desses recursos, agravada por fatores geopolíticos. Nesse contexto, a energia solar se destaca como fonte limpa, renovável e abundante, com elevado potencial de aplicação em diferentes regiões do mundo.

De acordo com o relatório *Energy Transition Investment Trends 2024* (2), os investimentos globais em energia solar atingiram US$ 1,8 trilhão em 2023, representando um crescimento expressivo em relação a anos anteriores. A queda no custo dos módulos fotovoltaicos e a crescente conscientização sobre benefícios sustentáveis e econômicos impulsionam esse avanço, criando um cenário favorável ao desenvolvimento e integração de novas tecnologias para o setor.

O Brasil se destaca no uso eficiente em energias renováveis, dentro desse contexto, a energia solar vem apresentando expansão significativa. Em 2023, a capacidade instalada atingiu 37,8 GW, um aumento de 54,8% em relação ao ano anterior (3), como pode ser visto na Figura 1. Minas Gerais, em especial, ocupa posição de destaque, com mais de 5,5 GW de potência instalada, distribuídos entre Geração Centralizada (GC) e Geração Distribuída (GD).

Figura 1: Capacidade Instalada no Brasil.



Variação da capacidade
Instalada das fontes no parque gerador (MW)

Fonte	2022	2023	Δ%23/22
Hidrelétrica	109.807	109.922	0,1%
Térmica[2]	46.440	47.515	2,3%
Nuclear	1.990	1.990	0,0%
Eólica	23.761	28.682	20,7%
Solar	24.453	37.843	54,8%
Capacidade disponível	206.451	225.952	9,4%

Fonte: (3).

Apesar desse crescimento, desafios relacionados à confiabilidade e estabilidade da rede permanecem, sendo a variabilidade climática um fator determinante. A intermitência da geração fotovoltaica (FV), decorrente de variações de irradiância e condições meteorológicas, exige o uso de métodos preditivos capazes de estimar com precisão a produção

de energia. Modelos de previsão de séries temporais têm sido amplamente estudados, desde abordagens estatísticas, como o ARIMA, até técnicas avançadas de Aprendizado de Máquina e sistemas *fuzzy* (4). Dentro deste contexto, as Redes Neurais Artificiais (RNAs) se destacam pela capacidade de modelar relações não lineares e complexas; porém, a falta de dados abertos de usinas solares brasileiras e estudos consolidados ainda representam um desafio.

Este trabalho insere-se nesse contexto, divulgando dados reais de operação de usinas fotovoltaicas no Brasil, se mostrando como fator determinante para auxiliar o processo de desenvolvimento de tecnologias nacionais, incluindo modelos de Aprendizado de Máquina. Além disso, o estudo propõe o uso de RNAs, com arquiteturas otimizadas por busca de hiperparâmetros e Aprendizado por Transferência, para maximizar a acurácia e reduzir o tempo de processamento de previsão de geração de energia em usinas fotovoltaicas.

1.1 JUSTIFICATIVA

A geração de energia elétrica por meio da irradiação solar tem se tornado cada vez mais relevante como uma fonte limpa e sustentável de energia, principalmente diante da necessidade de diversificação da matriz energética e redução das emissões de gases de efeito estufa.

De acordo com dados do Relatório de Tendências Globais em Investimentos em Energia Renovável 2021, publicado pela Agência Internacional de Energia Renovável (IRENA), os investimentos globais em energia solar atingiram um recorde de US$ 160,8 bilhões em 2020 (5), representando um aumento significativo em relação aos anos anteriores. Esse crescimento exponencial reflete o reconhecimento crescente de seu potencial como uma fonte de energia confiável, acessível e sustentável, sendo um dos pilares da transição energética.

Contudo, a natureza variável e intermitente da energia solar, influenciada por fatores como condições climáticas, sazonalidade e localização geográfica, apresenta um desafio significativo para a previsão confiável (6). A falta de sincronia entre a oferta e a demanda de energia resulta em custos adicionais associados à ativação de usinas termelétricas de reserva e à compra de energia no mercado a preços mais elevados. Além disso, a operação ineficiente da rede elétrica devido à falta de previsão pode levar a problemas de qualidade de energia, interrupções no fornecimento e perdas financeiras para as concessionárias.

A capacidade de antecipar flutuações na produção de energia é fundamental para garantir a estabilidade e a confiabilidade do sistema elétrico. A implementação eficaz de modelos de previsão permite aos operadores de rede tomarem medidas para mitigar impactos negativos, como cortes de energia, e assegura um fornecimento contínuo e seguro de eletricidade para consumidores e empresas.

Nessa perspectiva, é fundamental aplicar a previsão de geração de energia solar, uma vez que isso implica em:

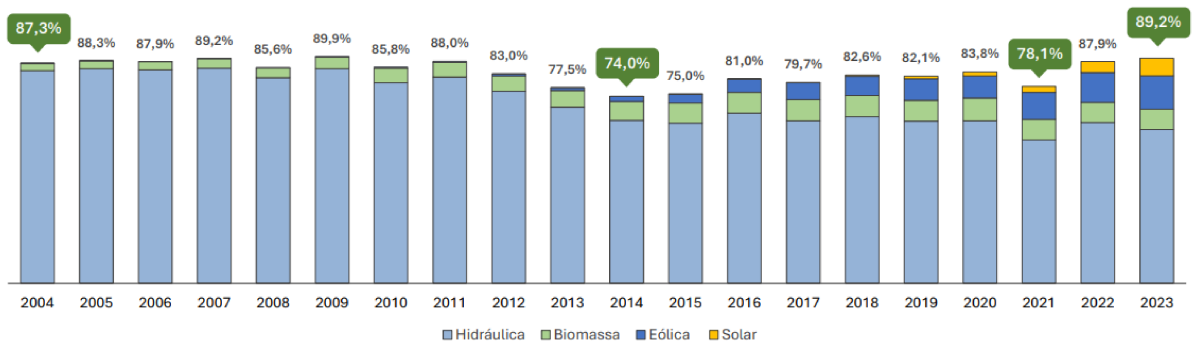
- Aumento da confiabilidade operacional;
- Otimização da alocação de recursos;
- Redução de custos operacionais e de acionamento de reservas;
- Apoio à manutenção preventiva;
- Melhoria da sustentabilidade do sistema elétrico.

1.2 FORMULAÇÃO DO PROBLEMA

O setor de geração de energia elétrica tem passado por uma transição significativa nas últimas décadas, impulsionada pela necessidade de diversificar a matriz energética, reduzir a dependência de combustíveis fósseis e mitigar as emissões de gases de efeito estufa. Nesse contexto, a energia solar fotovoltaica consolidou-se como uma das fontes renováveis de maior crescimento no mundo. De acordo com a Agência Internacional de Energia Renovável (IRENA), a capacidade instalada global de geração fotovoltaica ultrapassou 1,6 TW em 2023, representando aproximadamente 5,5% da matriz elétrica mundial e registrando um crescimento médio anual superior a 20% na última década.

O Brasil se destaca nesse cenário por conta de sua localização geográfica, extensão territorial e abundância de recursos naturais, alcançando a marca de 89,2% de representação de energias renováveis em sua matriz elétrica, um aumento de 11,1% em relação a 2021 (3).

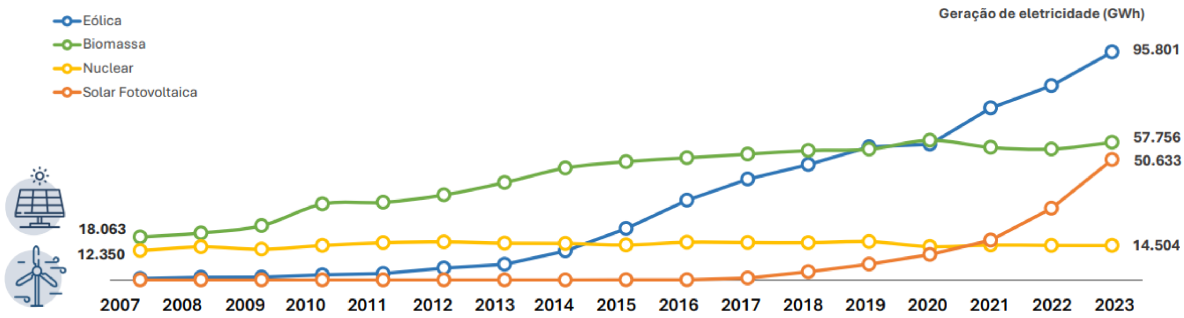
Figura 2: Evolução da Participação do Uso de Energias Renováveis na Matriz Energética Brasileira.



Fonte: (3).

Esse avanço decorre principalmente da expansão das fontes solar e eólica, conforme pode ser observado na Figura 3, que evidencia a trajetória exponencial da geração dessas fontes, que já superior à produção nuclear em GWh nos últimos anos (3).

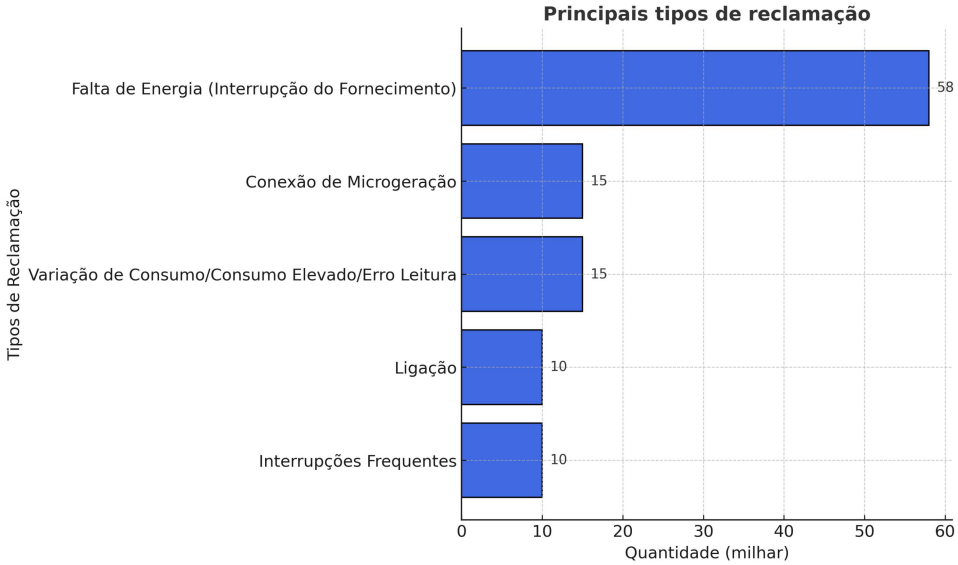
Figura 3: Geração de Eletricidade em GWh por Ano e Fonte Energética no Brasil.



Fonte: (3).

Apesar do crescimento acelerado, a rede elétrica nacional ainda apresenta limitações em absorver de forma eficiente essa nova configuração de geração. O aumento de fontes renováveis intermitentes, associado ao crescimento da demanda, tem contribuído para sobrecarga do sistema e maior número de reclamações registradas por consumidores. Apenas em 2023, foram contabilizadas 1.201.134 interações entre consumidores e a Ouvidoria da ANEEL (Agência Nacional de Energia Elétrica), um aumento de aproximadamente 38% em relação a 2022. Dentre essas interações, mais de 250 mil corresponderam a queixas sobre falta de fornecimento de energia, como descrito na Minuta do Relatório de Gestão de 2023 da ANEEL (7).

Figura 4: Reclamações nas ouvidorias da ANEEL.



Fonte: Minuta do Relatório de Gestão de 2023 da ANEEL.

Diante desse cenário, uma das principais barreiras enfrentadas no setor é a ausência de bases de dados nacionais de acesso aberto, com granularidade e detalhamento suficientes para o desenvolvimento de soluções baseadas em previsão e otimização. Embora existam informações disponibilizadas por órgãos como a ANEEL e o ONS, estas são limitadas em termos de resolução temporal e abrangência de variáveis, dificultando análises mais robustas de previsão de geração e a implementação de tecnologias de suporte à estabilidade da rede.

1.3 OBJETIVOS

Com o intuito de superar a carência de bases públicas e estruturadas de dados fotovoltaicos no contexto nacional, este trabalho, desenvolvido em parceria com a empresa TECSCI (Juiz de Fora–MG), teve como foco principal a construção, tratamento e disponibilização da base de dados **BR-PVGen**, composta por registros reais de 51 usinas solares distribuídas em diferentes regiões do Brasil. As séries temporais englobam medições de operação de inversores e variáveis meteorológicas, devidamente anonimizadas em conformidade com a Lei Geral de Proteção de Dados (LGPD), constituindo um dos primeiros esforços de abertura de dados dessa natureza no país.

As variáveis coletadas incluem medições de geração elétrica (nível de inversor), irradiância global no plano do módulo, temperatura do módulo e temperatura ambiente. Esses atributos possibilitam tanto o cálculo de indicadores de desempenho quanto a aplicação de técnicas de modelagem preditiva, ampliando o potencial de utilização do conjunto em estudos de previsão, controle e otimização da operação de sistemas fotovoltaicos.

A partir dessa base, foi conduzida uma investigação sobre métodos de previsão de geração de energia solar, definida como uma tarefa de previsão de séries temporais multivariadas, cujo objetivo é estimar a geração das usinas ($P_{t+\Delta}$) em um instante futuro $t + \Delta t$, a partir de um vetor de características $\mathbf{x}_t$ formado por observações anteriores. Foram consideradas diferentes configurações de defasagem temporal (*lags*) e variáveis correlatas de desempenho e condições ambientais.

Objetivo Geral:

Consolidar e disponibilizar uma base de dados nacional de usinas solares fotovoltaicas (**BR-PVGen**) e, a partir dela, desenvolver e avaliar modelos de previsão baseados em Redes Neurais Artificiais, com e sem Transferência de Aprendizado, aplicados à estimativa da geração de potência fotovoltaica.

Objetivos Específicos:

- Realizar a coleta, consolidação, tratamento e anonimização de dados históricos de geração e variáveis meteorológicas provenientes de 51 usinas fotovoltaicas distribuídas no Brasil;
- Estruturar e documentar a base de dados BR-PVGen, assegurando padronização, integridade e potencial de reuso científico;
- Implementar modelos de Redes Neurais Artificiais do tipo Perceptron Multicamadas (MLP), com diferentes arquiteturas e configurações, variando o número de camadas, neurônios e funções de ativação;
- Aplicar estratégias de Transfer Learning para aprimorar o desempenho preditivo em usinas com menor histórico de dados, avaliando diferentes abordagens de transferência entre plantas;
- Avaliar o desempenho dos modelos utilizando métricas quantitativas de erro, como SMAPE e NRMSE;
- Comparar o desempenho das arquiteturas propostas com métodos convencionais de previsão baseados em aprendizado supervisionado, identificando ganhos de acurácia e eficiência.

Hipóteses:

H1 A consolidação e disponibilização de uma base de dados multivariada, abrangendo variáveis de geração elétrica e meteorológicas, contribui significativamente para o avanço da pesquisa em previsão de geração fotovoltaica no contexto brasileiro, possibilitando maior reprodutibilidade e comparabilidade entre estudos.

H2 Modelos de Redes Neurais Artificiais do tipo MLP são capazes de representar relações não lineares complexas entre variáveis de geração e clima, superando o desempenho de modelos univariados e métodos tradicionais.

H3 O uso de estratégias de Transfer Learning permite transferir conhecimento entre usinas com diferentes volumes de dados, reduzindo o erro preditivo em cenários com amostras limitadas.

Além de validar essas hipóteses, este trabalho pretende fomentar a pesquisa nacional por meio da disponibilização da base BR-PVGen, fortalecendo o ecossistema de desenvolvimento de soluções baseadas em dados para o setor fotovoltaico brasileiro.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está estruturado da seguinte maneira: o **Capítulo 2** apresenta a revisão da literatura, abordando sobre trabalhos de referência na área de base de dados de usinas fotovoltaicas e de técnicas de previsão de geração de energia. O **Capítulo 3** expõe a formulação dos modelos de Redes Neurais Artificiais, que foram utilizadas como modelos de previsão de geração fotovoltaica. O **Capítulo 4** descreve a aquisição dos dados, incluindo a coleta em usinas fotovoltaicas e a composição da base proposta. Em seguida, o **Capítulo 5** apresenta as etapas de tratamento e organização dos dados. O **Capítulo 6** discute a avaliação dos modelos de aprendizado de máquina empregados. Posteriormente, o **Capítulo 7** apresenta e analisa os resultados obtidos. Por fim, o **Capítulo 8** sintetiza as conclusões do estudo e aponta perspectivas para trabalhos futuros.

1.5 PUBLICAÇÕES

Durante o desenvolvimento desta dissertação, foram produzidos e publicados (ou aceitos para publicação) os seguintes trabalhos:

1. ALVES, K. S. T. R.; **DE JESUS, C. D.**; AGUIAR, E. P. A new Takagi–Sugeno–Kang model for time series forecasting. *Engineering Applications of Artificial Intelligence*, v. 133, p. 108155, 2024. DOI: <https://doi.org/10.1016/j.engappai.2024.108155>.

2. CAMANDAROBA, C. G. M.; OLIVEIRA, C. V. M. B.; **DE JESUS, C. D.**; FERREIRA, M. A. M.; PUSSENTE, G. A. N.; NETO, G. F.; AGUIAR, E. P. Previsão de Geração de Energia em Usinas Fotovoltaicas utilizando Redes Neurais Artificiais e Aprendizado por Transferência. In: *Congresso Brasileiro de Automática – CBA 2024*, Sociedade Brasileira de Automática, 2024. Disponível em: <https://www.sba.org.br/cba2024/papers/paper_1047.pdf>.

3. CAMANDAROBA, C. G. M.; WERNECK, S. S. G.; MELLOS, J. Z. C.; **DE JESUS, C. D.**; MALTA, M.; GUIMARÃES, F. G.; AGUIAR, E. P. *A Federated Learning Approach for Distributed Solar Irradiance Forecasting.* CBIC 2025 – Congresso Brasileiro de Inteligência Computacional, 2025. (Artigo aceito para publicação nº 1175590, submetido em 28/05/2025, Área Temática: Sistemas Neurais – 6).

4. **DE JESUS, C. D.**; FERREIRA, R. N.; AGUIAR, E. P. *BR-PVGen - The Brazilian Photovoltaic Generation Dataset.* IEEE Access, 2025. (Artigo submetido em setembro de 2025, em avaliação).

2 REVISÃO DA LITERATURA

2.1 BASE DE DADOS DE USINAS FOTOVOLTAICAS

A disponibilidade de bases de dados públicas tem sido fundamental para o avanço da pesquisa em sistemas fotovoltaicos, permitindo estudos de desempenho, previsão de geração e integração à rede elétrica. Ao longo dos anos, diversas iniciativas internacionais foram estabelecidas, cada uma com suas particularidades, vantagens e limitações.

Um dos conjuntos de dados mais amplamente utilizados é o *Pecan Street Dataport* (8), que disponibiliza medições de alta frequência, com resolução de até um minuto, sobre consumo elétrico residencial e geração fotovoltaica em diferentes regiões dos Estados Unidos. Embora seja valioso para estudos em sistemas de pequena escala, carece de medições meteorológicas detalhadas e não fornece dados no nível de inversor.

O *Photovoltaic Data Acquisition* (PVDAQ) (9), mantido pelo *National Renewable Energy Laboratory* (NREL), reúne dados sincronizados de desempenho e variáveis meteorológicas de diferentes instalações fotovoltaicas nos Estados Unidos, permitindo análises comparativas e modelagem de desempenho em condições reais. Apesar de incluir irradiância e temperatura de módulo, frequentemente não contempla medições detalhadas por inversor ou metadados consistentes.

Outro exemplo de destaque é o *Desert Knowledge Australia Solar Centre* (DKASC) (10), que oferece dados de mais de 150 sistemas fotovoltaicos operando em clima desértico. O conjunto inclui medições de saída de inversores e variáveis meteorológicas, predominantemente com resolução de cinco minutos e cobertura temporal de aproximadamente nove anos. Entretanto, está restrito geograficamente a zonas áridas e a sistemas de menor porte, com capacidade média de 416 kWp e máxima de 1,8 MWp.

O *FAIR PV* (11) representa outra iniciativa relevante, contendo dois anos de dados de 316 usinas comerciais em diferentes zonas climáticas dos Estados Unidos. O diferencial está na padronização e organização dos dados segundo os princípios FAIR (*Findable, Accessible, Interoperable, and Reusable*). Apesar de incluir medições no nível de inversor, as variáveis meteorológicas são obtidas de bases externas, e não por estações locais, o que pode limitar a precisão de análises ambientais específicas.

Apesar das contribuições dessas bases, ainda há lacunas significativas, especialmente no contexto de usinas fotovoltaicas com capacidade acima de 1MW. Além disso, boa parte dos conjuntos de dados disponíveis não oferece medições simultâneas de variáveis elétricas e meteorológicas e é restrita a pequenas regiões.

Para mitigar essas limitações, este trabalho apresenta um novo conjunto de dados composto por 51 usinas fotovoltaicas de Geração Distribuída, o BR-PVGen. O banco contempla tanto sistemas com rastreadores de eixo único quanto sistemas de estrutura

fixa, com medições sincronizadas no nível de inversor e dados meteorológicos coletados por estações solarimétricas instaladas na própria usina e com dados de diferentes regiões do Brasil. As variáveis incluem potência ativa e reativa, potência em corrente contínua, temperatura interna dos inversores, irradiância no plano do arranjo, irradiância global inclinada, temperatura de módulo, temperatura ambiente, velocidade e direção do vento, precipitação acumulada, índice de albedo, coeficientes de bifacialidade e de temperatura, além de dados estruturais como tipo de suporte e área de módulos.

Os dados foram capturados de 4 em 4 minutos e são registrados no banco com resolução de 15 minutos e organizados em três arquivos principais: metadados, medições de inversores e dados solarimétricos. O processo de aquisição de dados é explicitado no Capítulo 4.

A Tabela 1 apresenta uma comparação entre o conjunto BR-PVGen e bases de referência, como DKASC, FAIR PV, PVDAQ e Pecan Street Dataport, destacando as características que o tornam especialmente relevante para pesquisas de previsão, detecção de falhas e modelagem de séries temporais em sistemas de GD de grande porte em regiões tropicais.

Tabela 1: Comparação entre os conjuntos de dados: BR-PVGen, DKASC, FAIR PV, PVDAQ e Pecan Street Dataport

Propriedade / Característica	**BR-PVGen**	**Conjunto DKASC**	**Conjunto FAIR PV**	**Conjunto PVDAQ**	**Pecan Street Dataport**
Registro temporal	✓	✓	✓	✓	✓
Potência ativa	✓	✓	✓	✓	✓
Potência reativa	✓	–	–	–	–
Potência CC	✓	–	✓	–	–
Temperatura do inversor	✓	–	–	–	–
Irradiância POA	✓	–	–	✓	–
Irradiância GRI	✓	–	–	✓	–
Irradiância GHI	✓	✓	✓	✓	✓
Temperatura do módulo	✓	✓	✓	–	–
Temperatura ambiente	✓	✓	✓	–	✓
Velocidade do vento	✓	✓	✓	–	✓
Direção do vento	✓	✓	✓	–	✓
Precipitação acumulada	✓	✓	✓	–	✓
Índice de albedo	✓	–	–	–	–
Pressão atmosférica	–	✓	✓	–	✓
Acúmulo de granizo	–	✓	–	–	–
Indicador de módulo bifacial	✓	–	–	–	–
Coeficiente de temperatura do módulo	✓	–	–	–	–
Coeficiente de bifacialidade do módulo	✓	–	–	–	–
Área do módulo	✓	–	–	✓	–
Eficiência do módulo	✓	–	–	–	–
Número de módulos	✓	–	–	–	–
Tipo de estrutura (tracker/fixa)	✓	–	–	✓	–
Dados separados por inversor	✓	–	✓	✓	✓
Cobertura temporal	1 ano	9 anos	2 anos	Não especificado	3 anos
Resoluções de amostragem	15 minutos	5 minutos	5 minutos (inversor), 15 minutos (clima)	15 minutos	15 minutos
Número de usinas fotovoltaicas	51	9	316	Não especificado	60
Capacidade média das usinas	~3 MWp	~416 kWp	Não especificado	Não especificado	Não especificado

2.2 TÉCNICAS DE PREVISÃO DE GERAÇÃO FOTOVOLTAICA

A previsão de geração de energia é componente central para a operação segura e econômica de sistemas elétricos com alta penetração de fontes renováveis. A literatura recente tem se consolidado em torno de modelos baseados em Aprendizado de Máquina (em inglês, *Machine Learning* ou ML), capazes de lidar com relações não lineares e com a variabilidade inerente às condições meteorológicas. Nesta dissertação, a revisão foi fundamentada em trabalhos de referência que reúnem e analisam estudos representativos da área, com destaque para *Machine Learning Based Solar Photovoltaic Power Forecasting: A Review and Comparison*, que apresenta um resumo das principais metodologias, tratamentos de dados, métricas de avaliação e abordagens para comparação. A Tabela 2, adaptada deste mesmo estudo, sintetiza pesquisas relevantes e práticas atuais (1).

Nos últimos anos, o estado da arte em previsão fotovoltaica tem avançado significativamente com o uso de arquiteturas neurais profundas, métodos híbridos e estratégias de aprendizado por transferência. Modelos híbridos combinando Redes Neurais Recorrentes (RNN, LSTM, GRU) com redes convolucionais (CNN) têm apresentado resultados superiores em horizontes de curto e médio prazo ao capturar simultaneamente padrões temporais e espaciais de irradiação. Além disso, técnicas de *ensemble learning*, como XGBoost, consolidaram-se como alternativas robustas em cenários com alta variabilidade e dados ruidosos. Estudos recentes também exploram a utilização de *Transfer Learning* e *Domain Adaptation* para reduzir a necessidade de grandes volumes de dados locais, facilitando a generalização entre plantas com diferentes perfis climáticos. Apesar desses avanços, ainda há desafios relevantes relacionados à padronização de bases públicas, à interpretação dos modelos de ML e à integração de previsões probabilísticas para uso operacional em sistemas elétricos.

2.2.1 Tratamento e preparação de dados

A etapa de preparação de dados é decisiva para o desempenho de modelos de previsão de potência fotovoltaica e envolve:

1. **Controle de qualidade e tratamento de faltantes**: remoção ou imputação de valores ausentes, detecção e correção de *outliers*;
2. **Normalização e escalonamento**: padronização de variáveis para estabilizar o processo de treinamento e evitar dominância de atributos com maior amplitude;
3. **Engenharia de atributos**: extração de informações temporais (mês, dia do ano, hora solar, feriados), meteorológicas (GHI, DNI, DHI, nuvens, temperatura do ar, vento, pressão) e atrasos temporais (*lags*) do próprio valor de geração;

4. **Seleção de variáveis**: redução de colinearidade e custo computacional por meio de métodos de *feature selection* (12);

5. **Divisão de dados e horizonte de previsão**: separação adequada em conjuntos de treino, validação e teste, respeitando a natureza temporal dos dados para evitar vazamento de informação, e definição do horizonte de previsão conforme a aplicação.

2.2.2 Modelos de previsão: metodologias e famílias

Modelos físicos e híbridos. Baseiam-se na modelagem termoelétrica dos módulos e inversores, associando parâmetros de campo (ângulo, azimute, perdas e temperatura) com variáveis meteorológicas previstas. Esses modelos apresentam boa interpretabilidade e desempenho consistente quando a calibração dos parâmetros é precisa, mas são suscetíveis a erros de previsão meteorológica e a desvios de medição (12, 13, 14).

Abordagens estatísticas clássicas. Métodos como *Multiple Linear Regression* (MLR), ARIMA e modelos em espaço de estados continuam amplamente empregados como referências de baixo custo computacional. São normalmente utilizados como *baselines*, mas sua capacidade de generalização é limitada em contextos com forte não linearidade e alta variabilidade climática (15, 16, 17). Trabalhos recentes exploram decomposições sazonais e estruturas hierárquicas, aprimorando a estabilidade desses modelos.

Métodos de vizinhança e *kernel*. Técnicas baseadas em similaridade, como k-NN e SVR, têm sido aplicadas para capturar não linearidades locais de forma eficiente, com boa relação entre desempenho e custo de processamento. A *Support Vector Regression* (SVR), especialmente com *kernel* radial, figura entre os melhores resultados para previsões de curto prazo quando bem ajustada (18, 19, 17). Sua limitação principal está na escalabilidade para grandes volumes de dados e na sensibilidade a hiperparâmetros.

Árvores e *ensembles*. Famílias como *Random Forest* (RF), *Gradient Boosting* (GB), *XGBoost* e *Extra Trees* apresentam excelente desempenho em contextos ruidosos e de múltiplas variáveis correlacionadas (13, 20). Os modelos baseados em *boosting* têm se destacado por combinar interpretabilidade com alta precisão e baixa necessidade de normalização. Abordagens recentes incluem *stacked ensembles* e *blending*, combinando preditores heterogêneos para melhorar a robustez das previsões.

Redes Neurais Artificiais. As RNAs do tipo Perceptron Multicamadas (MLP) continuam sendo amplamente utilizadas pela simplicidade e flexibilidade na modelagem de relações não lineares (21, 22). No entanto, as arquiteturas recorrentes, especialmente LSTM e GRU, se tornaram o padrão-ouro para previsão fotovoltaica em horizontes de 1 a 24 horas, por sua capacidade de capturar dependências temporais de longo alcance (23, 24). Além disso, estratégias de *Transfer Learning* vêm sendo aplicadas para adaptar modelos pré-treinados a usinas com poucos dados, reduzindo o tempo de treinamento e melhorando a generalização.

Tabela 2: Resumo de estudos representativos de previsão de potência fotovoltaica usando técnicas de aprendizado de máquina (1)

Trabalho	**Horizonte de Previsão**	**Resolução**	**Período de dados**	**Local**	**Pré-processamento**	**Método**
Huang e Perry (2016) (18)	1–24h	1h	01/04/2012 a 31/05/2014	NA	Filtros passa-baixa (Fourier) para tendências	*Gradient Boosting* (determinística) e k-NN (probabilística)
Li et al. (2016) (15)	24h	24h	01/01/2011 a 30/06/2012	Macau	NA	RNAs, ARIMA, ARMAX, média móvel
Nespoli et al. (2017) (25)	24h	1h	2017	Itália	Normalização aprimorada	MLP ANN *Selective Ensemble*
Mellit et al. (2018) (21)	24h	1h	01/01/2011 a 31/12/2011	Itália	NA	RNA com classificação de dias (previsão GHI)
Lin e Pai (2018) (19)	1 mês	1 mês	01/01/2010 a 30/04/2014	Taiwan	Decomposição sazonal	Previsão sazonal evolutiva LS-SVR para previsão mensal da potência FV
Ramsami e Oree (2018) (26)	24h	1h	03/02/2012 a 30/12/2013	Reino Unido	*Stepwise Regression* e remoção de registros ausentes	Rede Neural Híbrida
Gao et al. (2019) (23)	24h	1h	01/11/2016 a 28/10/2017	China	Classificação de dias por clima/estação	LSTM
Mohammed et al. (2019) (27)	1–24h	1h	01/04/2012 a 31/05/2014	NA	NA	Previsão probabilística baseada em técnicas estatísticas
Pretto et al. (2020) (13)	24h	1h	2017 a 2019	Itália	Agrupamento por previsão de irradiação	*Ensemble* probabilístico
Abdellatif et al. (2022) (20)	24h	NA	01/01/2018 a 31/12/2021	Malásia	Normalização por desvio-padrão	RF, XGBoost, AdaBoost, Extra Trees
Khan et al. (2022) (24)	1h	1h	4 anos	Holanda	Limpeza e escalonamento	ANN, LSTM, XGBoost
Wang et al. (2023) (28)	72h	1h	2018 a 2020	China	Escalonamento e extração de padrões espaciais	CNN + LSTM híbrido para previsão fotovoltaica
Alves et al. (2024) (4)	1h	1h	2016 a 2020	Brasil	Normalização e seleção de atributos	Novo modelo Takagi–Sugeno–Kang para séries temporais

3 REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais (RNAs) constituem uma classe de modelos computacionais inspirados na estrutura e funcionamento do cérebro humano. A formulação inicial foi apresentada por McCulloch e Pitts (29) em 1943, que introduziram o primeiro modelo de neurônio artificial. Esse trabalho estabeleceu as bases teóricas da computação neural.

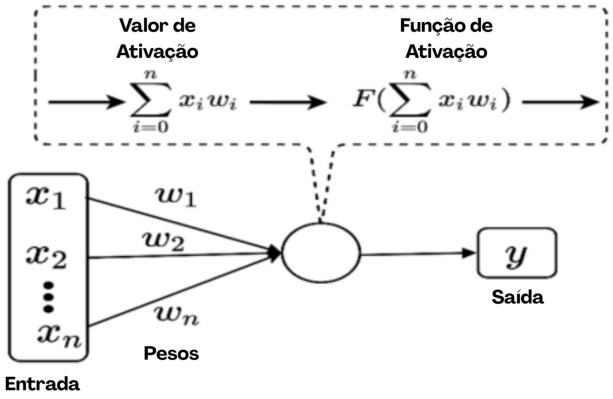


Figura 5: Representação de um neurônio artificial com pesos, soma ponderada e função de ativação.

Fonte: Adaptação de (1) .

Cada neurônio recebe um vetor de entradas $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, ponderadas por pesos $\mathbf{w} = [w_1, w_2, \ldots, w_n]$, e soma ainda um termo de polarização (*bias*) b.

A Figura 5 ilustra o funcionamento interno de um neurônio artificial, evidenciando a soma ponderada das entradas e a aplicação da função de ativação.

O valor de ativação u do neurônio é dado por:

$$u = \sum_{i=1}^{n} w_i x_i + b. \tag{3.1}$$

Esse valor é então transformado por uma função de ativação não-linear $F(\cdot)$, resultando na saída do neurônio:

$$y = F(u). \tag{3.2}$$

Posteriormente, Frank Rosenblatt (30) em 1958 propôs o perceptron, no qual os neurônios eram dispostos em uma única camada, conectados diretamente aos nós de entrada e aos de saída, sendo assim um modelo de neurônio treinável capaz de resolver problemas lineares por meio do ajuste iterativo de pesos. Este modelo representou a primeira aplicação prática do aprendizado supervisionado em redes neurais.

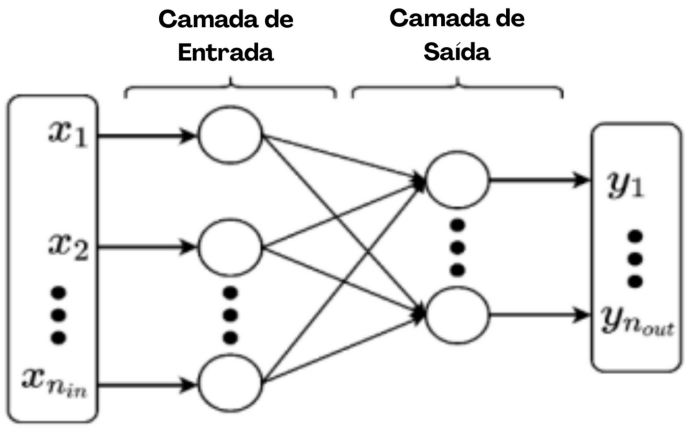


Figura 6: Estrutura de uma Rede Neural Artificial do tipo *Perceptron* com única camada).

Fonte: Adaptação de (1).

Décadas depois, David Rumelhar, Geoffrey Hinton e Ronald Williams, (31) introduziram o algoritmo de retropropagação (*backpropagation*) em 1986, que permitiu treinar redes multicamadas ao propagar o erro da saída de volta às camadas anteriores. Esse avanço viabilizou a aplicação em problemas não lineares e complexos, consolidando o modelo do perceptron multicamadas (MLP).

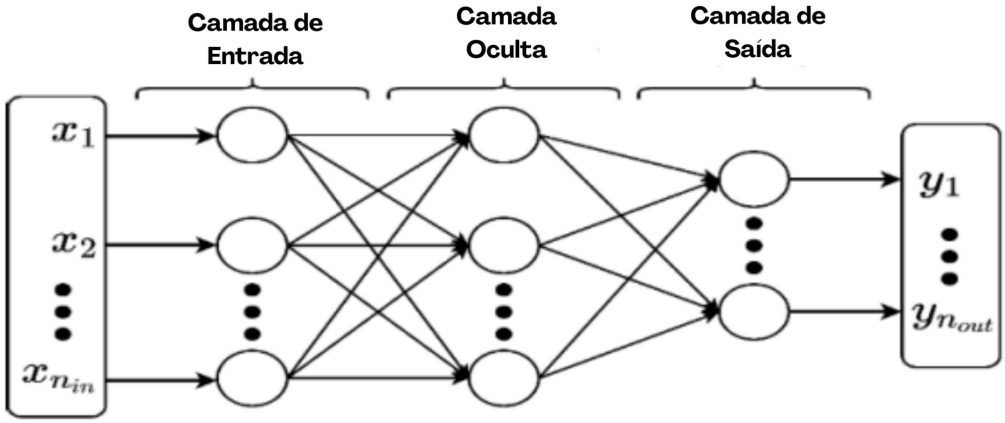


Figura 7: Estrutura de uma Rede Neural Artificial do tipo (MLP).

Fonte: Adaptação de (1).

De forma mais abrangente, Simon Haykin (32) em 1999 sistematizou o campo das RNAs em um arcabouço matemático e computacional, discutindo topologias, funções de ativação, regras de aprendizado e aplicações.

3.1 Função de Ativação

As funções de ativação têm papel central na capacidade de aprendizado da rede, pois introduzem não-linearidade ao modelo. Entre as mais utilizadas estão:

- **Sigmoide:** $F(u) = \frac{1}{1+e^{-u}}$. É útil em contextos probabilísticos, pois restringe a saída em $[0, 1]$. Contudo, apresenta saturação para valores muito positivos ou muito negativos de u, situação em que a derivada $F'(u)$ torna-se próxima de zero. Esse fenômeno é chamado de gradiente desvanecente, em que os gradientes se propagam de forma cada vez menor pelas camadas, dificultando o ajuste eficaz dos parâmetros.
- **Tangente Hiperbólica:** $F(u) = \tanh(u)$. Fornece saídas no intervalo $[-1, 1]$ e é centrada em zero. Entretanto, ainda suscetível a gradientes pequenos em valores extremos.
- **ReLU (Rectified Linear Unit):** $F(u) = \max(0, u)$. Amplamente empregada em redes profundas por sua simplicidade computacional e por reduzir parcialmente o problema do gradiente desvanecente, já que sua derivada é constante (1) para $u > 0$. No entanto, pode levar ao problema do *neurônio morto*, quando unidades permanecem com saída zero e deixam de contribuir para o aprendizado (33).

3.2 Propagação Direta

O cálculo da saída da rede é realizado por meio do processo de propagação direta (*forward propagation*), no qual as entradas percorrem sequencialmente as camadas até a obtenção da saída final.

Seja uma rede neural com L camadas, em que a entrada é representada por $\mathbf{h}^{(0)} = \mathbf{x}$. Para cada camada $l = 1, 2, \ldots, L$, a transformação é dada por:

$$\mathbf{z}^{(l)} = W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}, \tag{3.3}$$

$$\mathbf{h}^{(l)} = F^{(l)}\left(\mathbf{z}^{(l)}\right), \tag{3.4}$$

em que:

- $W^{(l)}$ é a matriz de pesos da camada l e $\mathbf{b}^{(l)}$ é o vetor de vieses;
- $\mathbf{z}^{(l)}$ representa a combinação linear dos sinais da camada anterior;

Após atravessar todas as camadas ocultas, a saída final da rede é dada por:

$$\mathbf{y} = \mathbf{h}^{(L)}. \tag{3.5}$$

3.3 Retropropagação e Treinamento

A etapa de retropropagação (*backpropagation*) é responsável por calcular os gradientes da função de custo em relação aos parâmetros da rede (pesos e vieses), permitindo a atualização iterativa desses parâmetros durante o treinamento (31).

O objetivo do treinamento é ajustar os pesos $\mathbf{w}$ e vieses $\mathbf{b}$ de modo a minimizar uma função de custo $J(\mathbf{w}, \mathbf{b})$, que quantifica o erro entre a saída prevista $\mathbf{y}$ e a saída desejada $\hat{\mathbf{y}}$.

3.3.1 Função Custo:

Em problemas de regressão, uma função amplamente utilizada é o erro quadrático médio (*Mean Squared Error*, MSE):

$$J_{\text{MSE}}(\mathbf{w}, \mathbf{b}) = \frac{1}{m} \sum_{j=1}^{m} \left\| \mathbf{y}^{(j)} - \hat{\mathbf{y}}^{(j)} \right\|^2, \tag{3.6}$$

onde m é o número de exemplos no conjunto de treinamento.

No entanto, em cenários de previsão de geração fotovoltaica, valores extremos e períodos de baixa produção (próximos de zero) podem distorcer o ajuste. Para mitigar esse efeito, adota-se a função de custo Huber, que combina as propriedades do MSE e do erro absoluto médio (MAE), sendo mais robusta a *outliers* (34). A formulação é dada por:

$$J_\delta(r) = \begin{cases} \frac{1}{2} r^2, & \text{se } |r| \leq \delta, \\ \delta \left(|r| - \frac{1}{2}\delta \right), & \text{caso contrário,} \end{cases} \tag{3.7}$$

em que $r = y - \hat{y}$ é o resíduo, e $\delta > 0$ é um hiperparâmetro que controla a transição entre o regime quadrático (MSE) e o linear (MAE).

3.3.2 Algoritmos de Otimização

O cálculo do gradiente fornece apenas a direção de ajuste; a forma como os parâmetros são atualizados depende do otimizador adotado, podendo empregar diferentes estratégias de otimização para melhorar a convergência, estabilidade e velocidade de aprendizado (33). Entre os mais utilizados destacam-se:

- **Stochastic Gradient Descent (SGD)**: Atualiza os parâmetros a partir do gradiente estimado em uma iteração da propagação direta:

$$w_i \leftarrow w_i - \eta \frac{\partial J}{\partial w_i}, \tag{3.8}$$

$$b \leftarrow b - \eta \frac{\partial J}{\partial b}, \tag{3.9}$$

em que η é a taxa de aprendizado (*learning rate*). O SGD é simples e eficiente, mas sensível à escolha de η e pode apresentar oscilações durante a convergência.

- **Adam (Adaptive Moment Estimation)**: O Adam combina os conceitos de *Momentum* e ajuste adaptativo de taxas de aprendizado. Ele mantém médias móveis do gradiente e do quadrado do gradiente para cada parâmetro:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1)\nabla J(\mathbf{w}, \mathbf{b}), \tag{3.10}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2)\nabla J(\mathbf{w}, \mathbf{b})^2, \tag{3.11}$$

onde:

  - m_t é a média móvel de primeira ordem (gradiente acumulado, similar ao *momentum*);
  - v_t é a média móvel de segunda ordem (estimativa da variância do gradiente);
  - β_1 e β_2 são hiperparâmetros de decaimento exponencial típicos ($\beta_1 = 0.9$, $\beta_2 = 0.999$).

Como essas estimativas iniciais são enviesadas em direção a zero, aplica-se uma correção de viés:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \tag{3.12}$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}. \tag{3.13}$$

A atualização final dos parâmetros é dada por:

$$\theta \leftarrow \theta - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{3.14}$$

em que:

  - η é a taxa de aprendizado;
  - ϵ é um termo pequeno (ex.: 10^{-8}) para evitar divisão por zero.

O Adam apresenta bom desempenho em uma ampla gama de aplicações, pois adapta dinamicamente a taxa de aprendizado de cada parâmetro e incorpora inércia (via β_1). Essa combinação torna-o particularmente indicado para MLPs aplicadas à regressão, pela robustez, estabilidade e baixa necessidade de ajuste fino de hiperparâmetros.

Regularização e Controle de Treinamento

O treinamento de uma rede neural ocorre de forma iterativa em épocas, nais quais podem ser definidas como quantidade de processamentos completos de todas as amostras do conjunto de treinamento.

Durante o treinamento, é recomendado utilizar também um conjunto de validação, composto por exemplos não vistos durante o ajuste dos pesos. Após cada época, a rede é avaliada no conjunto de validação, fornecendo uma estimativa da capacidade de generalização. Essa métrica é fundamental para a avaliação do treinamento.

Para reduzir o risco de sobreajuste (*overfitting*) e melhorar a capacidade de generalização da rede, foram incorporadas as seguintes estratégias:

- **Regularização L2**:

  Consiste em adicionar um termo de penalização proporcional à norma quadrática dos pesos à função de custo. A função de custo regularizada é dada por:

  $$J_{\text{reg}}(\mathbf{w}, \mathbf{b}) = J(\mathbf{w}, \mathbf{b}) + \frac{\alpha}{2m} \sum_{l=1}^{L} \|W^{(l)}\|_F^2, \tag{3.15}$$

  em que $\alpha > 0$ é o parâmetro de regularização e $\|W^{(l)}\|_F$ denota a norma de Frobenius da matriz de pesos $W^{(l)}$.

  A norma de Frobenius é definida como:

  $$\|W^{(l)}\|_F = \sqrt{\sum_{i=1}^{n_l} \sum_{j=1}^{n_{l-1}} \left(W_{ij}^{(l)}\right)^2}, \tag{3.16}$$

  Essa penalização desencoraja pesos excessivamente grandes, promovendo soluções mais suaves e generalizáveis.

- **Dropout**: Durante o treinamento, em cada iteração e para cada camada oculta, neurônios individuais são desativados com probabilidade p. Essa técnica promove diversidade e reduz a dependência da rede em neurônios específicos, aumentando sua robustez (35). Para esse estudo foi utilizado uma probabilidade de 10%.

- **Early Stopping**: Monitora a função de custo no conjunto de validação ao longo das épocas. Seja $J_{\text{val}}^{(t)}$ o valor da função de custo de validação na época t. Se não houver melhora em $J_{\text{val}}^{(t)}$ ao longo de T_{pac} épocas consecutivas (neste trabalho, $T_{\text{pac}} = 20$), o treinamento é interrompido:

  $$\text{parar se } J_{\text{val}}^{(t)} \geq \min_{\tau \in [t - T_{\text{pac}},\, t]} J_{\text{val}}^{(\tau)}. \tag{3.17}$$

  Essa estratégia evita sobreajuste e reduz o custo computacional (36).

- **ReduceLROnPlateau**: Ajusta dinamicamente a taxa de aprendizado quando a função de custo de validação entra em um *platô*. Seja η_t a taxa de aprendizado na época t. Se não houver melhora em $J_{\text{val}}^{(t)}$ por T_{pat} épocas consecutivas, a taxa de aprendizado é reduzida por um fator $\gamma \in (0, 1)$:

$$\eta_{t+1} = \begin{cases} \gamma \cdot \eta_t, & \text{se estagnado por } T_{\text{pat}} \text{ épocas,} \\ \eta_t, & \text{caso contrário.} \end{cases} \tag{3.18}$$

  Essa adaptação permite ajustes mais finos dos parâmetros sem necessidade de redefinir manualmente o *learning rate*.

3.4 Aprendizado por Transferência

O aprendizado por transferência (*Transfer Learning*, TL) é uma abordagem que visa reaproveitar o conhecimento adquirido por uma rede neural treinada em uma tarefa-fonte ($\mathcal{T}_F$) com maior volume de dados, para melhorar o desempenho em uma tarefa-alvo ($\mathcal{T}_A$), com menor volume de dados, otimizando os resultados e tempo de treinamento para este fim (37).

Seja uma rede neural treinada na tarefa-fonte com parâmetros (W_F, b_F), ajustados a partir de uma base de dados $\mathcal{D}_F$. O objetivo do TL é inicializar os parâmetros da rede na tarefa-alvo, para toda camada transferida $l \in \{1, 2, \ldots, L\}$:

$$W_A^{(l)}(0) = W_F^{(l)}, \qquad b_A^{(l)}(0) = b_F^{(l)}, \quad \forall\, l \in \{1, 2, \ldots, L\}, \tag{3.19}$$

A partir dessa inicialização, duas estratégias são possíveis:

1. **Extração de características**: as camadas transferidas da $(\mathcal{T}_F \to \mathcal{T}_A)$ são mantidas congeladas, ou seja, os pesos dessas camada não são atualizados com o treinamento e suas ativações são utilizadas como representações de entrada para novas camadas específicas da tarefa-alvo.

2. ***Fine-tuning***: os pesos transferidos servem como inicialização, mas continuam sendo atualizados durante o treinamento na tarefa-alvo. Dessa forma, as representações aprendidas são ajustadas às especificidades da nova base de dados.

O TL apoia-se no conceito de que a função de representação aprendida pela rede na tarefa-fonte aproxima-se da função ideal desejada para a tarefa-alvo. Assim, ao invés de iniciar o treinamento de forma aleatória, parte-se de um espaço de hipóteses mais próximo da solução ótima (33).

Esse paradigma é particularmente útil quando:

- há escassez de dados rotulados na tarefa-alvo;
- a tarefa-fonte e a tarefa-alvo compartilham domínios semelhantes;
- deseja-se acelerar a convergência e reduzir custos computacionais.

No contexto de previsão de geração fotovoltaica, o TL permite treinar modelos em usinas com bases de dados históricas extensas ($\mathcal{D}_F$) e transferir o conhecimento para plantas com poucos dados disponíveis ($\mathcal{D}_A$).

As camadas iniciais, que aprendem padrões gerais de sazonalidade e resposta à irradiação, podem ser mantidas fixas ou levemente ajustadas, enquanto as camadas finais são re-treinadas para refletir as condições específicas da planta-alvo. Essa estratégia resulta em modelos mais robustos em cenários de disponibilidade limitada de medições.

3.5 DEMAIS MODELOS DE REGRESSÃO UTILIZADOS

Para fins de comparação com os modelos otimizados propostos, foram selecionados quatro algoritmos clássicos de regressão amplamente utilizados em aprendizado de máquina supervisionado: *Linear Regression*, *Gradient Boosting Regressor*, *Bayesian Ridge Regression* e *Kernel Ridge Regression*. Todos os modelos utilizados estão disponíveis na biblioteca *Scikit-learn* e representam abordagens com diferentes fundamentos matemáticos e propriedades preditivas. A seguir, apresentam-se as fundamentações teóricas e principais características de cada um.

- ***Linear Regression***

  A regressão linear é um dos modelos mais antigos e fundamentais para problemas de regressão (38). Assume-se que a relação entre as variáveis independentes e a variável dependente é linear, buscando-se estimar os coeficientes que minimizam o erro quadrático médio (*Mean Squared Error*, MSE). Sua formulação básica é:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon, \tag{3.20}$$

  em que:

  - y é a variável resposta;
  - $x_1, x_2, \ldots, x_n$ são as variáveis explicativas;
  - β_0 é o intercepto;
  - $\beta_1, \beta_2, \ldots, \beta_n$ são os coeficientes;
  - ϵ é o termo de erro aleatório.

Apesar de sua simplicidade, é eficaz quando a relação entre variáveis é aproximadamente linear. Entretanto, apresenta limitações em presença de não linearidades ou multicolinearidade, que podem resultar em estimativas instáveis.

- ***Gradient Boosting Regressor***

  O *Gradient Boosting* é uma técnica de *ensemble* baseada no princípio de *boosting*, no qual modelos fracos (tipicamente árvores de decisão rasas) são treinados sequencialmente, de modo que cada modelo subsequente busca corrigir os erros residuais do anterior (39). O método constrói o modelo de forma aditiva, minimizando iterativamente uma função de perda, como MSE ou erro absoluto.

  Sua capacidade de modelar relações não lineares e lidar com alta dimensionalidade faz com que seja amplamente utilizado em problemas de previsão contínua e classificação, apresentando, em geral, alta acurácia.

- ***Bayesian Ridge Regression***

  A *Bayesian Ridge Regression* é uma versão probabilística da regressão ridge tradicional (40). Em vez de estimar coeficientes como valores fixos, assume-se que eles seguem distribuições a priori, geralmente gaussianas. O ajuste dos parâmetros é feito considerando a distribuição a posteriori, incorporando incerteza nas estimativas.

  Essa abordagem fornece intervalos de confiança para os coeficientes e tende a ser mais robusta a sobreajustes, especialmente em conjuntos de dados com multicolinearidade.

- ***Kernel Ridge Regression (KRR)***

  O *Kernel Ridge Regression* combina a regressão ridge, que aplica regularização L2 para reduzir a variância do modelo, com o *kernel trick*, permitindo projetar os dados para um espaço de características de maior dimensionalidade (41). Isso possibilita capturar relações não lineares de forma eficiente.

  A regressão ridge resolve o problema de minimização:

$$\min_{\beta} \ ||y - X\beta||^2 + \alpha||\beta||^2, \tag{3.21}$$

  em que:

  - α é o parâmetro de regularização;
  - X é a matriz de preditores;
  - β é o vetor de coeficientes.

  No *KRR*, a escolha do kernel (linear, polinomial, RBF, etc.) define como os dados são mapeados para o novo espaço, aumentando a flexibilidade do modelo em relação a padrões não lineares.

4 AQUISIÇÃO DE DADOS

Este capítulo apresenta melhor a base de dados fotovoltaica brasileira proposta BR-PVGen, descrevendo sua metodologia de aquisição, procedimentos de pré-processamento, organização estrutural e principais características estatísticas. A base foi construída a partir do registro sistemático de dados operacionais de usinas fotovoltaicas localizadas em diferentes estados do Brasil, abrangendo condições climáticas e geográficas diversas. Os dados foram disponibilizados em parceria com a empresa TECSCI (Juiz de Fora–MG), que possui um sistema supervisório web SCADA[1] com dezenas de usinas monitoradas.

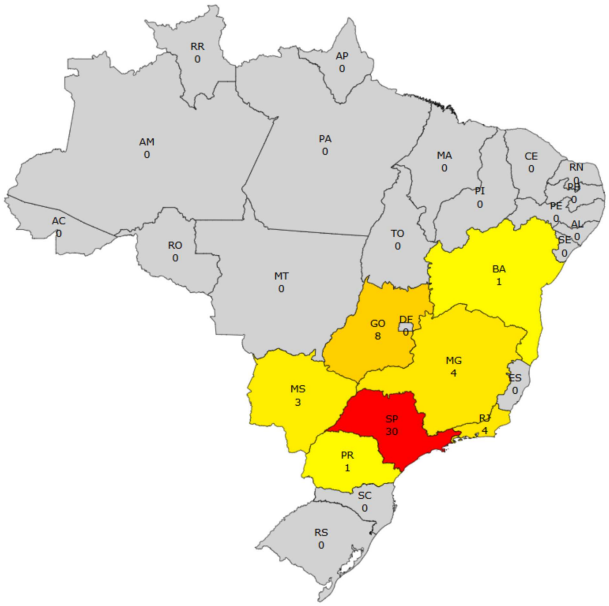


Figura 8: Distribuição geográfica das 51 usinas fotovoltaicas incluídas na base de dados proposta.

Fonte: Autoria própria (2025).

O processo de aquisição de dados é realizado através de um equipamento, chamado *datalogger*, que faz a comunicação e leitura dos dados dos demais dispositivos da planta conectados na rede ethernet da usina, a cada 4 minutos. Os dados coletados são tratados internamente no *datalogger* e depois transmitidos para um banco de dados único em um servidor em nuvem da *Amazon Web Services* (AWS).

O processo de aquisição de dados está ilustrado na Figura 9. A coleta contínua das medições dos dispositivos de campo é feita através do protocolo de comunicação *Modbus* (Seta A) da Figura 9.

[1] SCADA (Supervisory Control and Data Acquisition) é uma arquitetura de controle composta por computadores, comunicações em rede e interfaces gráficas, utilizada para monitoramento e controle de processos em alto nível.

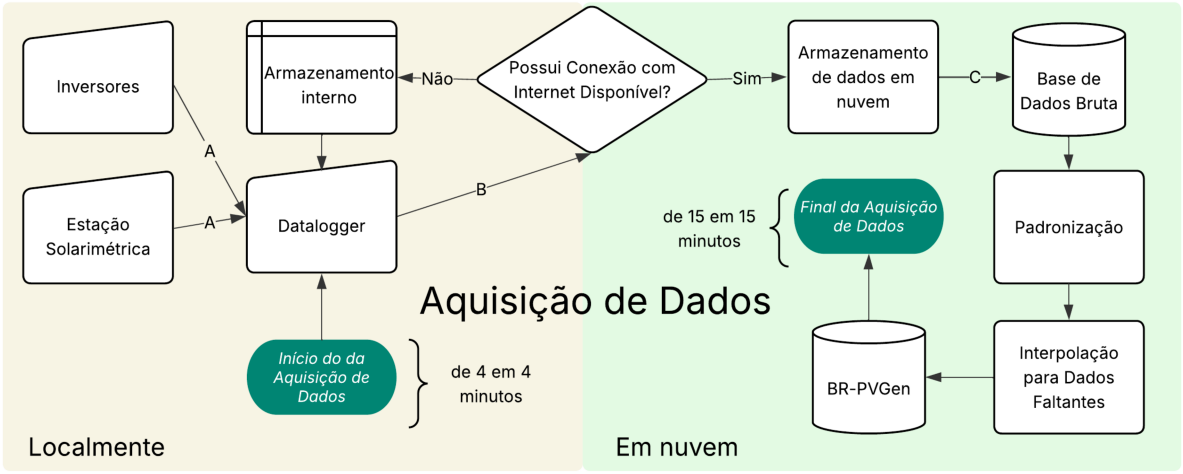


Figura 9: Diagrama da metodologia de aquisição de dados.

Fonte: Autoria própria (2025).

Após a coleta, o *datalogger* transmite os dados para uma API[2] hospedada em plataforma em nuvem (Seta B), utilizando exclusivamente o protocolo HTTPS[3].

Em situações de falha ou baixa qualidade de rede de internet, os dados são armazenados localmente e enviados automaticamente assim que a conexão é restabelecida. Os dados recebidos pela API são processados e armazenados em um banco de dados não relacional [4] na nuvem (Seta C), garantindo tolerância a falhas e confiabilidade na entrega.

O *datalogger* utilizado nesta aplicação é um Controlador Lógico Programável (CLP) da **WAGO**, chamado CC100. Este equipamento possui funcionalidades de TI e TA (Tecnologia da Informação e Automação), possibilitando integração com diversos sensores e atuadores da planta. Entre seus recursos:

- 8 Entradas Digitais (DI) e 4 Saídas Digitais (DO);
- 2 Entradas e 2 Saídas Analógicas (AI/AO);
- 2 canais PT100 e 1 porta RS485.

[2] Application Programming Interface; conjunto de rotinas e protocolos que permitem a interação entre diferentes sistemas de software.

[3] Hypertext Transfer Protocol Secure; versão segura do HTTP que garante integridade e criptografia.

[4] Banco de dados sem esquema fixo, otimizado para estruturas flexíveis como documentos e pares chave-valor.

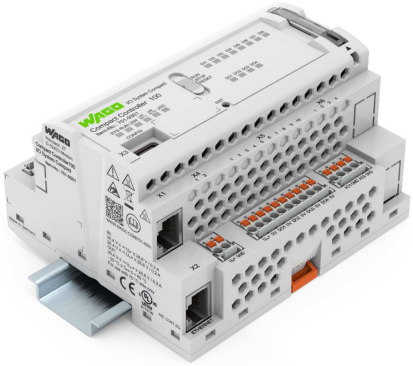


Figura 10: Controlador WAGO CC100.

Fonte: https://www.wago.com/br/novo-controlador-iot-cc100. Acessado em 10/03/2025.

4.1 PROCESSAMENTO E TRATAMENTO DE DADOS

O período coberto pela base de dados inicia-se em 26 de março de 2024 e encerra-se em 9 de junho de 2025, resultando em mais de 44 milhões de registros de inversores e mais de 5,6 milhões de registros de estações solarimétricas das 51 usinas.

Tabela 3: Resumo da base de dados bruta

Característica	**Inversor**	**Estação Solarimétrica**
Período	26/03/2024 – 09/06/2025	
Total de Registros	44.092.970	5.669.460

4.1.1 Segregação por Fonte de Dados

Os registros foram segregados por origem: dados de inversores (geração elétrica) e dados de estações solarimétricas (irradiância e variáveis meteorológicas). Essa distinção reflete as diferenças físicas e funcionais entre os sistemas de medição, permitindo pré-processamentos específicos para cada tipo de dado (42).

4.1.2 Padronização das Séries Temporais

Devido à irregularidade nos intervalos de registro, os carimbos de tempo foram padronizados para uma malha fixa de 15 minutos, abrangendo as 24 horas do dia. Essa uniformização é essencial para análises temporais, reduz ruídos e melhora a compatibilidade com algoritmos de previsão (43).

4.1.3 Agregação por Média Móvel Ponderada

Para padronizar a resolução temporal e reduzir a influência de valores atípicos, foi utilizada a média móvel ponderada (MMP), agregando dados originalmente amostrados a cada 5 minutos, aproximadamente, para intervalos fixos de 15 minutos.

A cada instante central T_k, em $\{00{:}00, 00{:}15, \ldots, 23{:}45\}$, foi considerada uma janela simétrica de 15 minutos ($h = 7,5$ minutos).

Cada valor $x(t_i)$ representa uma medição no instante t_i. Para cálculo do ponto do valor agregado $y(T_k)$ foram considerados apenas $x(t_i)$ nos quais $t_i \in [T_k - h, T_k + h]$. Cada valor $x(t_i)$ na janela recebeu peso inversamente proporcional à sua distância temporal em relação a T_k:

$$w_i = \frac{1}{1 + |t_i - T_k|}$$

O valor agregado foi calculado por:

$$y(T_k) = \frac{\sum_i w_i \cdot x_i}{\sum_i w_i}$$

Essa técnica preserva tendências locais e reduz flutuações não representativas (44).

4.1.4 Interpolação Direcionada para Valores Faltantes

Após a agregação, foi aplicada interpolação linear apenas quando havia medições válidas imediatamente antes e depois do ponto ausente, e o intervalo entre elas não excedia 15 minutos. A fórmula utilizada foi:

$$y(t) = y_1 + \left(\frac{t - t_1}{t_2 - t_1}\right) \cdot (y_2 - y_1)$$

Essa abordagem minimiza riscos de suavização excessiva e não foi aplicada a variáveis acumulativas.

4.2 PÓS-PROCESSAMENTO

A Figura 11 apresenta a evolução mensal de registros válidos, evidenciando o crescimento progressivo decorrente da inclusão de novas usinas no sistema SCADA. Observa-se uma redução em junho de 2025 em virtude do encerramento da coleta em 9 de junho.

Após todas as etapas de pré-processamento, a base final resultou em 14.400.480 registros válidos provenientes dos inversores e 1.151.232 registros oriundos das estações solarimétricas. Trata-se, portanto, de um conjunto de dados de grande escala, com consistência temporal e amplitude geográfica inédita para o contexto brasileiro, constituindo um recurso valioso tanto para pesquisas acadêmicas quanto para o desenvolvimento de novas tecnologias aplicadas à operação de usinas solares.

Além disso, foi calculado o percentual de dados ausentes em cada variável por usina, considerando apenas o período efetivo de geração (entre 06:00 e 18:00, horário de Brasília).

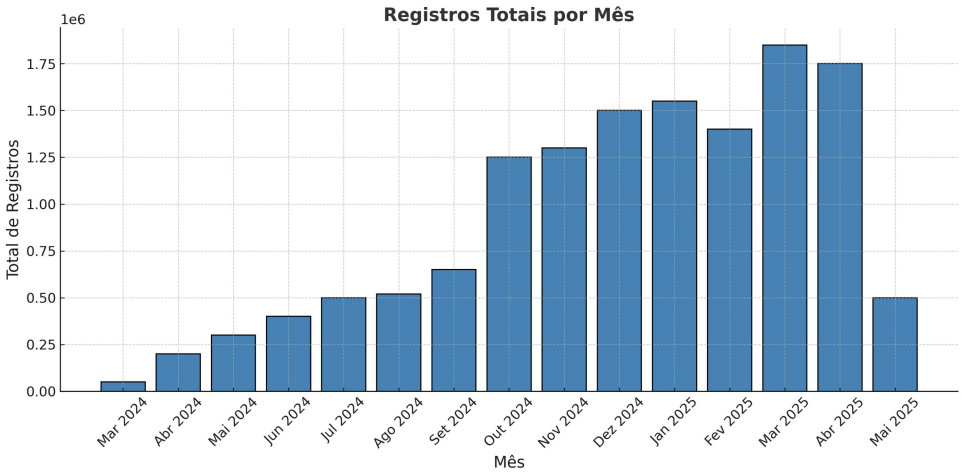


Figura 11: Registros mensais combinados de inversores e estações solarimétricas.

Fonte: Autoria própria (2025).

Essa análise, ilustrada no mapa de calor da Figura 12, permite avaliar a completude da base e orientar estratégias de imputação ou seleção de variáveis para os modelos preditivos.

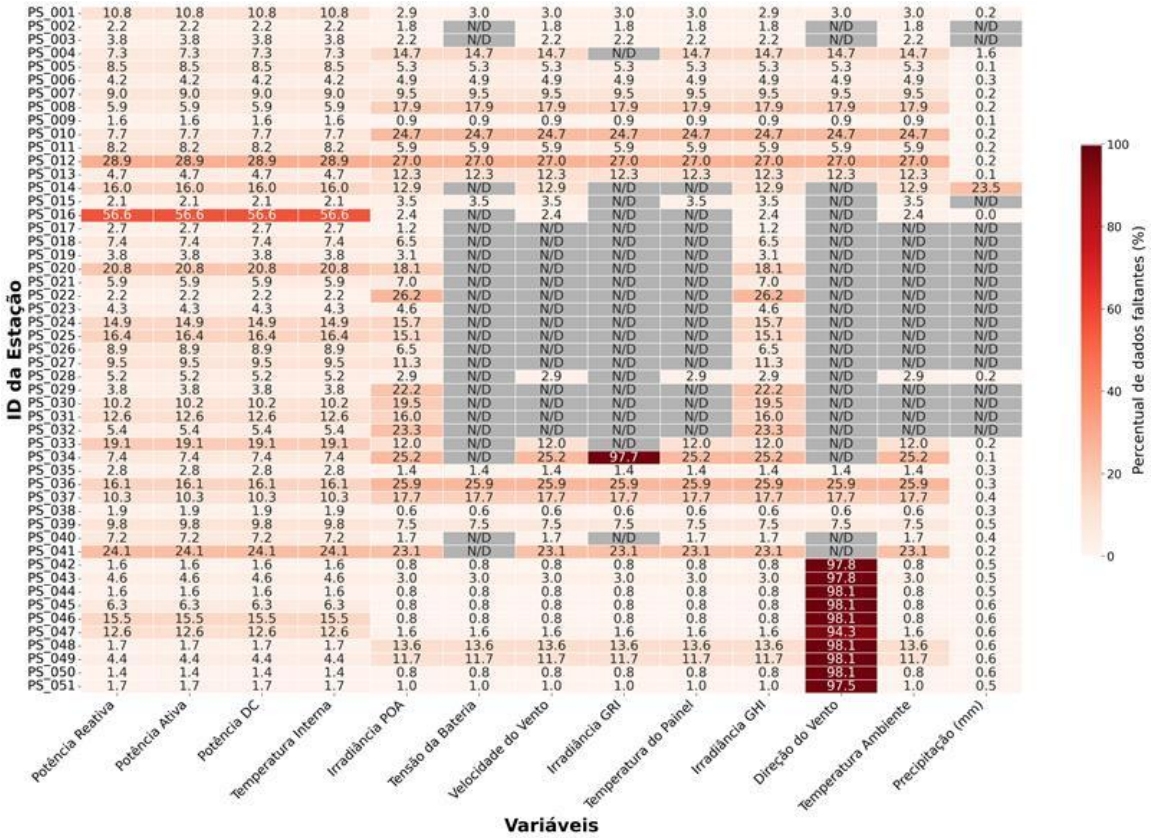


Figura 12: Percentual de dados ausentes por variável e usina durante o período de geração efetiva (06:00–18:00, BRT).

Fonte: Autoria própria (2025).

A base é organizada em três entidades principais: **Metadados**, **Inversor** e **Estação Solarimétrica**, descritas na Tabela 4.

4.3 DIFERENCIAIS E POSSÍVEIS APLICAÇÕES

Além da metodologia e estrutura previamente descritas, a base de dados proposta apresenta diferenciais relevantes em relação a outros conjuntos de dados fotovoltaicos disponíveis na literatura:

- **Dados Granulares em Nível de Inversor:** A base fornece medições elétricas detalhadas no nível individual de cada inversor, permitindo análises aprofundadas do comportamento de componentes e identificação de desvios de desempenho localizados. Embora esse recurso também esteja presente no *FAIR PV Dataset* (11), a base brasileira expande essa capacidade ao incluir variáveis adicionais, como potência reativa e temperatura interna do inversor.
- **Organização Estruturada dos Dados:** O conjunto de dados é organizado em três componentes distintos, metadados, medições de inversores e dados de estações solarimétricas. Essa estrutura modular aumenta a transparência, facilita a escalabilidade e integra-se de forma eficiente a fluxos de processamento de dados avançados.
- **Ampla Cobertura Espacial e Operacional:** Apesar de a série temporal se limitar a pouco mais de um ano, a base contempla 51 usinas fotovoltaicas distribuídas geograficamente pelo Brasil, capturando o comportamento dos sistemas em uma ampla gama de condições climáticas e operacionais da rede elétrica. A Figura 8 ilustra a distribuição geográfica das plantas.
- **Representatividade em Escala Utilitária:** A base brasileira é composta por sistemas fotovoltaicos distribuídos de grande porte, com capacidade instalada média de aproximadamente 3 MWp, variando de 1,7 MWp a 5 MWp. Esse porte contrasta com instalações menores presentes no *DKASC Dataset*.
- **Aquisição de Dados Meteorológicos Baseada em Sensores Locais:** Diferente de conjuntos de dados que dependem de APIs de terceiros para informações meteorológicas, a base brasileira utiliza estações solarimétricas instaladas em cada local, fornecendo medições ambientais diretas e em tempo real, incluindo irradiância no plano dos módulos e irradiância global inclinada.
- **Conjunto Ampliado de Variáveis:** A base inclui variáveis raramente encontradas em benchmarks públicos, como potência reativa, potência em corrente contínua, temperatura interna do inversor, índice de albedo, coeficientes bifaciais e tensão da bateria, possibilitando análises e modelagens mais abrangentes dos sistemas.

A base proposta constitui um repositório único no contexto brasileiro, integrando dados em nível de inversor com medições meteorológicas locais. Sua abrangência geográfica e temporal, aliada à diversidade de variáveis, oferece potencial para:

- Modelagem e previsão de geração fotovoltaica.
- Estudos de desempenho e perdas de sistemas.
- Análises de impacto climático e geográfico na produção de energia.
- Treinamento de modelos de aprendizado de máquina para diagnóstico e operação de plantas.

Este recurso atende a uma lacuna crítica de dados consolidados para pesquisa e operação de usinas fotovoltaicas no Brasil, alinhando-se a padrões de proteção de dados e ampliando as possibilidades de desenvolvimento científico e tecnológico no setor.

4.3.1 Disponibilidade da Base de Dados

A base proposta, denominada **BR-PVGen**, encontra-se disponível para acesso público no repositório Kaggle [5]. Os arquivos estão organizados em formato `CSV`, acompanhados de documentação descritiva (`README`). A estrutura segue a organização discutida neste capítulo.

[5] <https://www.kaggle.com/datasets/tecsci/brazilian-pv-dataset/data/data>

Tabela 4: Atributos das entidades que compõem a base de dados

Atributo	Tipo	Descrição
Entidade Metadados		
`id`	Inteiro	Identificador único da usina.
`nominal_power_mw`	Real	Potência nominal instalada em megawatts (MW).
`is_panel_bifacial`	Booleano	Indica se o painel é bifacial (`true`) ou monofacial (`false`).
`panel_temperature_coefficient`	Real	Perda percentual de potência por aumento de 1 °C.
`panel_bifaciality_coefficient`	Real	Razão de eficiência traseira/frontal de painéis bifaciais (0 a 1).
`panel_area_mm2`	Real	Área de um módulo fotovoltaico em mm^2.
`panel_efficiency_percentage`	Real	Eficiência de conversão do painel em porcentagem.
`number_of_panels`	Inteiro	Quantidade total de módulos instalados.
`brazil_federative_unit`	Texto	Unidade federativa (estado) onde a usina está localizada (ex.: "SP").
`structure_type`	Texto	Tipo de estrutura: `TRACKER` ou `FIXED`.
Entidade Inversor		
`datetime`	Texto	Data/hora no formato ISO 8601 (`YYYY-MM-DDThh:mm:ssZ`).
`total_reactive_power_var`	Real	Potência reativa total (VAR); `null` se indisponível.
`total_active_power_w`	Real	Potência ativa total (W); `null` se indisponível.
`total_dc_power_w`	Real	Potência total de entrada CC (W); `null` se indisponível.
`internal_temperature_celsius`	Real	Temperatura interna do inversor em °C.
`document_count`	Objeto<Texto, Inteiro>	Quantidade de pontos brutos usados na Média Móvel Ponderada por variável.
`interpolated_keys`	Objeto<Texto, Booleano>	Mapeia cada variável para indicar se foi interpolada (`true` se sim).
`inverter_id`	Inteiro	Identificador único do inversor na usina.
Entidade Estação Solarimétrica		
`datetime`	Texto	Data/hora no formato ISO 8601 (`YYYY-MM-DDThh:mm:ssZ`).
`poa_irradiance_wm2`	Real	Irradiância no plano dos módulos (W/m^2).
`battery_voltage`	Real	Tensão da bateria da estação (V).
`wind_speed_ms`	Real	Velocidade do vento (m/s).
`gri_irradiance_wm2`	Real	Irradiância refletida pelo solo (W/m^2).
`panel_temperature_celsius`	Real	Temperatura de superfície do módulo FV (°C).
`tracker_albedo_index`	Real	Índice de albedo do solo para sistemas com tracker.
`ghi_irradiance_wm2`	Real	Irradiância Global Horizontal (W/m^2).
`wind_direction_degrees`	Real	Direção do vento em graus.
`ambient_temperature_celsius`	Real	Temperatura ambiente (°C).
`precipitation_accumulated_mm`	Real	Precipitação acumulada (mm).
`document_count`	Objeto<Texto, Inteiro>	Quantidade de pontos brutos usados na MMP por variável.
`interpolated_keys`	Objeto<Texto, Booleano>	Mapeia cada variável para indicar se foi interpolada (`true` se sim).

5 TRATAMENTO DOS DADOS

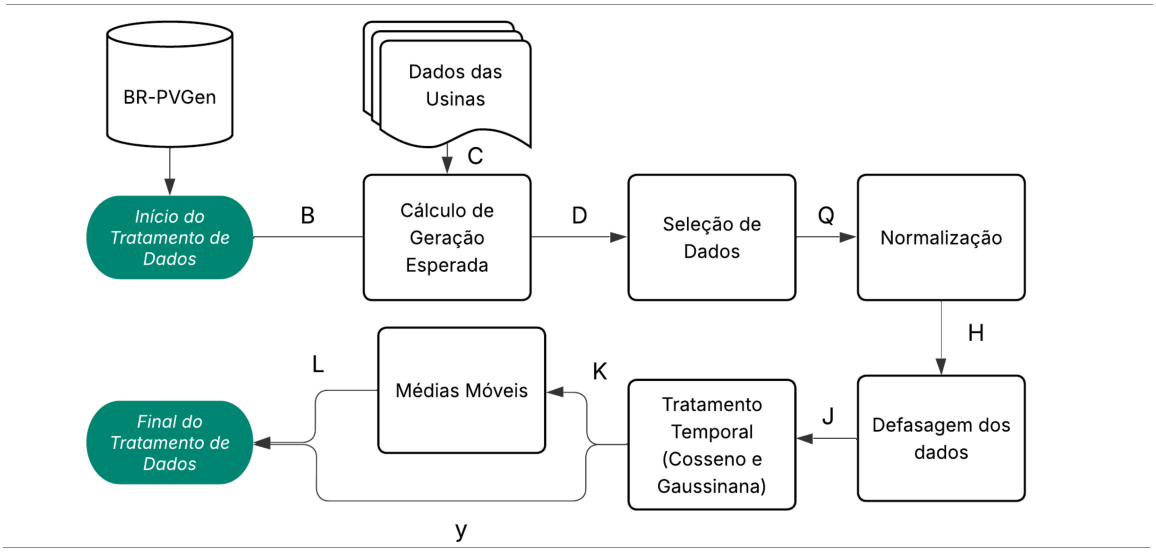


Figura 13: Fluxo geral do tratamento de dados adotado.

Fonte: Autoria própria (2025).

A estrutura geral do tratamento dos dados é apresentada na Figura 13. O processo inicia-se a partir de um conjunto de bases de dados provenientes de 51 usinas fotovoltaicas, representado pela matriz $\mathcal{B} \in \mathbb{R}^{51\times T\times V}$, em que T corresponde ao número de amostras no horizonte temporal (apresentado por usina na Tabela 5) e V ao número inicial de variáveis disponíveis, incluindo medições de inversores e estações solarimétricas.

De forma complementar, emprega-se a matriz $\mathcal{C}$, a qual reúne parâmetros técnicos das usinas, tais como capacidade de potência ativa nominal dos inversores, área total dos módulos fotovoltaicos, eficiência de referência dos módulos e demais características operacionais. A integração das informações de $\mathcal{B}$ e $\mathcal{C}$ possibilita o cálculo da geração esperada e das perdas do sistema, conforme descrito na Seção 5.2. O resultado desse processamento constitui a matriz inicial $\mathcal{D} \in \mathbb{R}^{51\times T\times 15}$, que serve como ponto de partida para as etapas subsequentes de tratamento.

A partir de $\mathcal{D}$, o fluxo de tratamento segue pelas seguintes etapas principais. Inicialmente, das 15 variáveis originais foram selecionadas apenas duas de interesse direto para a previsão: Geração Esperada e Geração Real. A matriz resultante, denotada por $\mathcal{Q} \in \mathbb{R}^{51\times T\times 2}$, preserva a modelagem nos atributos mais relevantes ao problema. No caso da previsão diária, considera-se a energia acumulada no dia; já para a previsão intradiária, utiliza-se a potência instantânea registrada a cada 15 minutos.

5.1 RESUMO ESTRUTURAL DAS MATRIZES

O processo de tratamento pode ser resumido pela seguinte sequência de transformações:

$$\mathcal{B} \in \mathbb{R}^{51\times T\times V} \xrightarrow{\text{Geração Esperada}} \mathcal{D} \in \mathbb{R}^{51\times T\times 15} \xrightarrow{\text{Seleção}} \mathcal{Q} \in \mathbb{R}^{51\times T\times 2} \xrightarrow{\text{Normalização}} \mathcal{H} \in [0,1]^{51\times T\times 2}$$

$$\mathcal{H} \xrightarrow{\text{Defasagens}} \begin{cases} \mathcal{J}_1 \in \mathbb{R}^{51\times T\times 12}, & \text{(base diária, 5 lags por variável)} \\ \mathcal{J}_2 \in \mathbb{R}^{51\times T\times 26}, & \text{(base intradiária, 12 lags por variável)} \\ y_{1,2} \in \mathbb{R}^{51\times T\times 1}, & \text{(vetor target, com } \textit{lag}\text{ 1 instante a frente)} \end{cases}$$

$$\mathcal{J}_i \xrightarrow{\text{Tratamento Temporal}} \begin{cases} \mathcal{K}_1 \in \mathbb{R}^{51\times T\times 13}, & \text{(base diária, +1 coluna de cosseno)} \\ \mathcal{K}_2 \in \mathbb{R}^{51\times T\times 28}, & \text{(base intradiária, +1 cosseno e +1 gaussiana)} \end{cases}$$

$$\mathcal{K}_i \xrightarrow{\text{Médias Móveis}} \begin{cases} \mathcal{L}_1 \in \mathbb{R}^{51\times T\times 15}, & \text{(base diária, +2 médias móveis)} \\ \mathcal{L}_2 \in \mathbb{R}^{51\times T\times 32}, & \text{(base intradiária, +4 médias móveis)} \end{cases}$$

Ao final dessas etapas, para cada usina, possui uma matriz de $\mathcal{L}$ onde cada linha corresponde a um vetor de medições no instante t, representando os vetores de entrada $x^{(v)}(t)$. A cada vetor de entrada associa-se uma variável-alvo $y(t)$, definida como a potência ou energia gerada pela usina no instante subsequente $t+1$.

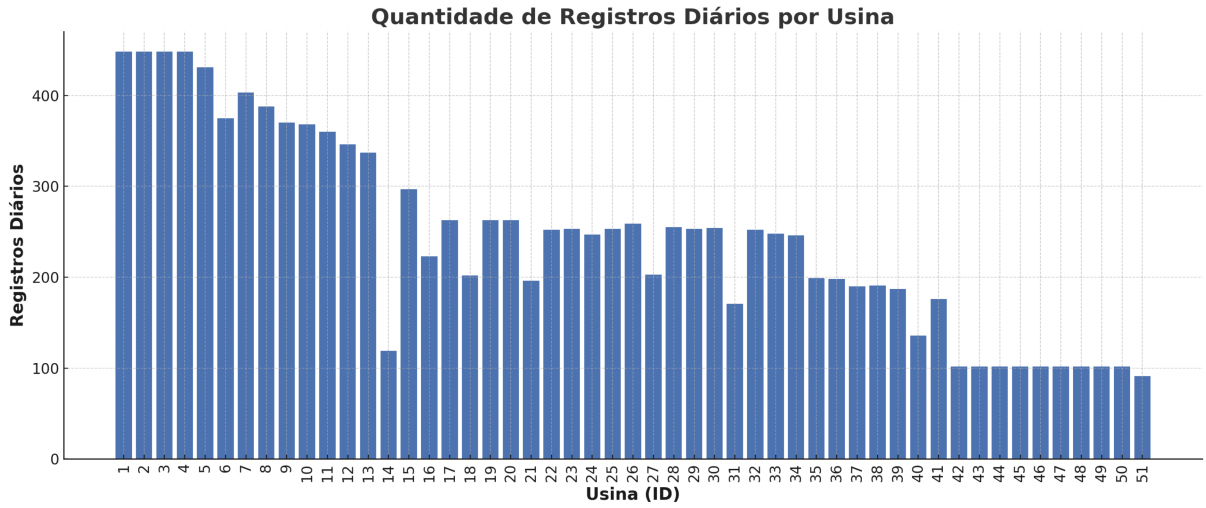


Figura 14: Quantidade de Dias na base de dados por usina

Fonte: Autoria própria (2025).

Tabela 5: Quantidade de registros por usina nas diferentes resoluções temporais.

Usina	Registros Diários	Registros Intradiários (15 min)
1, 2, 3, 4	448	21515
5	431	20730
6	375	18035
7	403	19360
8	388	18670
9	370	17790
10	368	17690
11	360	17300
12	346	16615
13	337	16220
14	119	5745
15	297	14260
16	223	10735
17	263	12645
18	202	9705
19, 20	263	12645
21	196	9410
22	252	12105
23	253	12155
24	247	11860
25	253	12155
26	259	12450
27	203	9755
28	255	12255
29	253	12155
30	254	12205
31	171	8235
32	252	12105
33	248	11910
34	246	11810
35, 36	199	9560
37,38	190	9165
39	187	9020
40	136	6570
41	176	8480
42,43, 44, 45, 46, 47, 48, 49, 50	102	4905
51	91	4415

A Tabela 5 e a Figura 14 apresentam o quantitativo de registros válidos por usina, tanto para a base de dados diária quanto para a intradiária (15 minutos).

Na sequência, são detalhados os procedimentos empregados nas diferentes etapas do tratamento de dados.

5.2 GERAÇÃO ESPERADA DE USINAS FOTOVOLTAICAS

A estimativa da geração teórica sob condições operacionais ideais é uma etapa fundamental tanto para a avaliação de desempenho de usinas fotovoltaicas quanto para o enriquecimento de bases de dados utilizadas em modelos de previsão. Esses cálculos permitem identificar desvios operacionais, quantificar perdas e criar variáveis derivadas para uso em algoritmos de inteligência artificial, especialmente em cenários com falhas de medição ou dados ausentes.

5.2.1 Potência Ideal

A potência ideal instantânea P_{ideal} foi calculada a partir da equação clássica para módulos fotovoltaicos:

$$P_{\text{ideal}} = A_{\text{mod}} \cdot I_{\text{POA}} \cdot \eta_{\text{mod}}, \tag{5.1}$$

em que:

- A_{mod} é a área total dos módulos (m^2),
- I_{POA} é a irradiância no plano dos módulos (kW/m^2),
- η_{mod} é a eficiência nominal do módulo em condições padrão de teste (STC).

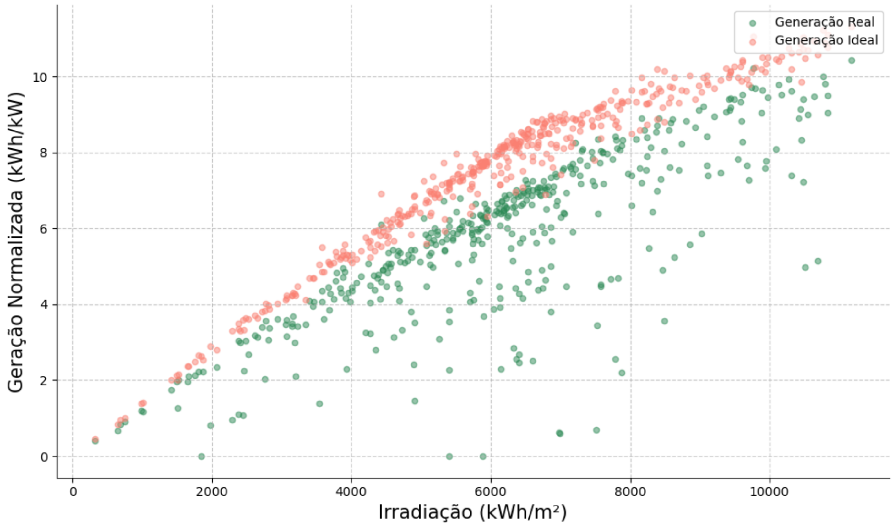


Figura 15: Comparação entre a geração medida e a geração ideal estimada.

Fonte: Autoria própria (2025).

Essa formulação assume condições ideais, incluindo módulos limpos, configuração monofacial, ausência de perdas térmicas ou elétricas e sem degradação. Assim, representa um limite superior teórico para a geração instantânea, utilizado como referência na avaliação de desempenho.

5.2.2 Componentes de Perda e de Ajuste de Performance

Para modelar a diferença entre a geração ideal e a geração efetivamente medida, foram incorporados mecanismos de perda que refletem condições reais de operação, incluindo ganho bifacial, perdas dependentes de irradiância e temperatura, degradação de módulos, perdas por *clipping* e indisponibilidade do sistema.

5.2.2.1 Ganho Bifacial

Considerando que todas as usinas do conjunto de dados utilizam módulos bifaciais, a irradiância efetiva foi ajustada para incluir a contribuição da face traseira, proveniente da irradiância refletida pelo solo I_{GRI}, ponderada pelo coeficiente de bifacialidade γ:

$$I_{\text{eff}} = I_{\text{POA}} + \gamma \cdot I_{\text{GRI}}. \tag{5.2}$$

A potência ideal ajustada para o ganho bifacial é dada por:

$$P_{\text{bif}} = A_{\text{mod}} \cdot I_{\text{eff}} \cdot \eta_{\text{mod}}. \tag{5.3}$$

5.2.2.2 Perdas Dependentes da Irradiância

A eficiência dos módulos tende a reduzir sob baixos níveis de irradiância. Para modelar este efeito, foi aplicado um fator corretivo η_{irr}. Para valores de irradiância inferiores a 500 W/m², utilizou-se um polinômio de sexto grau ajustado empiricamente (45):

$$\eta_{\text{irr}} = \begin{cases} 1, & I_{\text{POA}} \geq 500 \\ \sum_{i=0}^{6} a_i \cdot I_{\text{POA}}^i, & I_{\text{POA}} < 500 \end{cases} \tag{5.4}$$

Tabela 6: Coeficientes na Equação

a_0	a_1	a_2	a_3	a_4	a_5	a_6
0.74894	0.0031	-1.68884×10^{-5}	4.72152×10^{-8}	-6.95836×10^{-11}	5.13525×10^{-14}	-1.49454×10^{-17}

5.2.2.3 Degradação dos Módulos

A perda de eficiência ao longo do tempo foi modelada por um fator multiplicativo $f_{\text{deg}}(t)$, representando a redução acumulada em função dos anos de operação (45):

$$f_{\text{deg}}(t) = \begin{cases} 0.975, & t \leq 1 \text{ ano} \\ 0.975 - 0.00753 \cdot (t-1), & 1 < t \leq 27 \\ 0.801, & t > 27 \text{ anos} \end{cases} \tag{5.5}$$

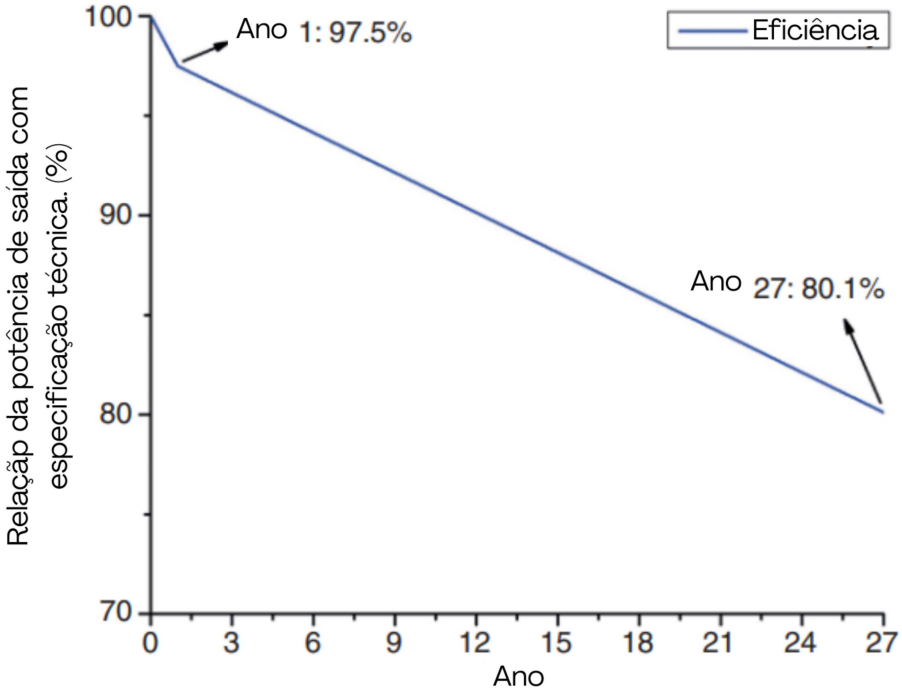


Figura 16: Eficiência Normalizada de Acordo com Nível de Irradiância.

Fonte: Adaptação de (45)

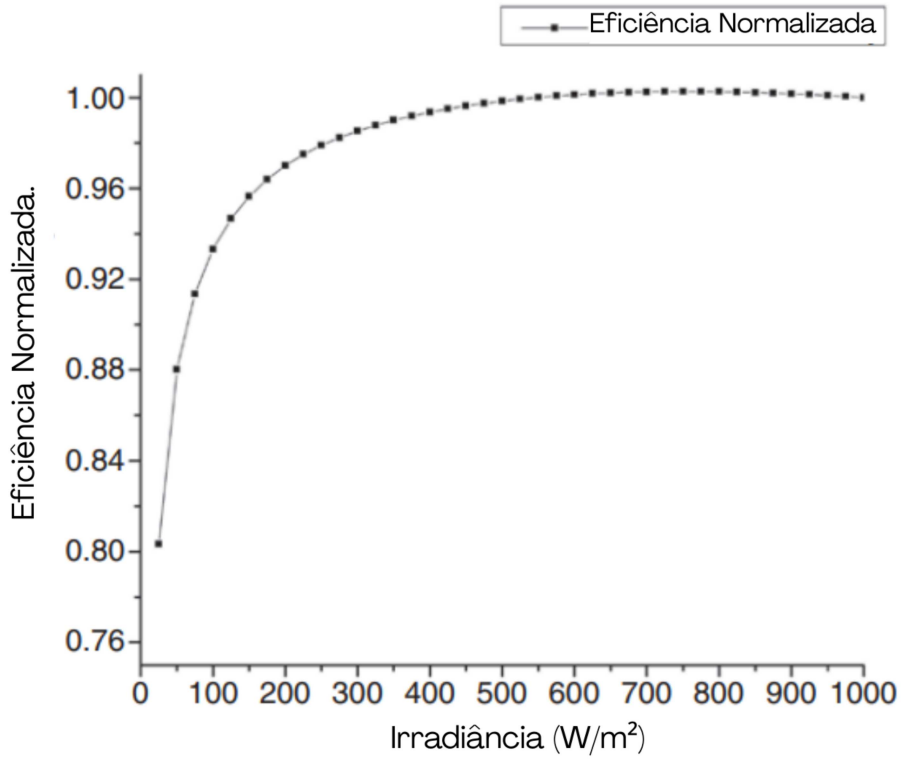


Figura 17: Degradação da Eficiência dos Módulos.

Fonte: Adaptação de (45)

5.2.2.4 Perdas Térmicas

Seguindo a norma IEC 61724-2, a eficiência foi ajustada para refletir o efeito da temperatura, resultando em:

$$\eta_{\text{mod},T} = \eta_{\text{irr}} \cdot \left[1 + \beta_T \cdot (T_{\text{mod}} - T_{\text{STC}})\right], \tag{5.6}$$

onde β_T é o coeficiente térmico do módulo (%/°C), T_{mod} é a temperatura estimada do módulo e $T_{\text{STC}} = 25°\text{C}$.

5.2.2.5 Potência Esperada Final

Combinando os fatores descritos, a potência esperada final foi obtida por:

$$P_{\text{exp}} = A_{\text{mod}} \cdot I_{\text{eff}} \cdot \eta_{\text{mod},T} \cdot f_{\text{deg}}. \tag{5.7}$$

Esse valor reflete uma estimativa mais realista do desempenho, incorporando condições meteorológicas e características do sistema.

5.2.2.6 Perdas por *Clipping*

As perdas por *clipping* ocorrem quando a potência em corrente contínua excede a capacidade nominal de saída em corrente alternada do inversor:

$$P_{\text{clip}} = \max(P_{\text{exp}} - P_{\text{inv,max}}, 0). \tag{5.8}$$

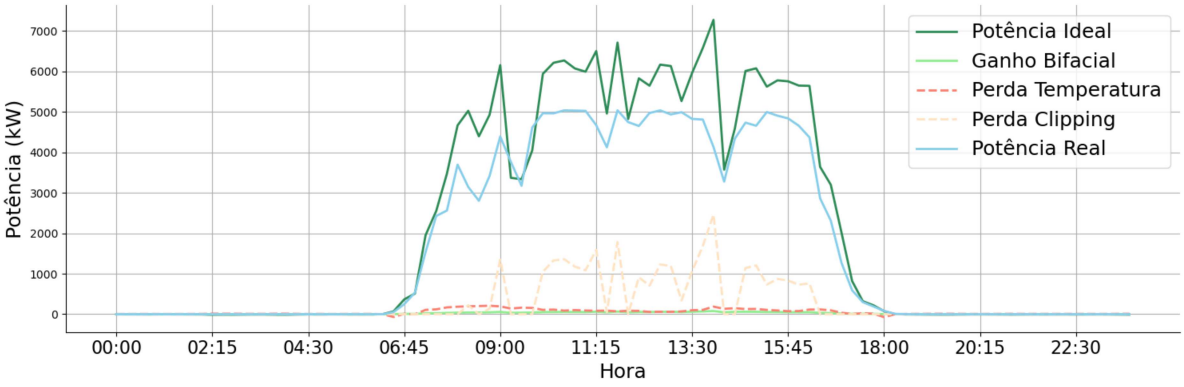


Figura 18: Geração esperada, perdas modeladas e geração medida.

Fonte: Autoria própria (2025).

5.2.2.7 Perdas por Indisponibilidade

Períodos de indisponibilidade operacional foram identificados quando $P_{\text{real}} = 0$ e $I_{\text{POA}} > I_{\min}$, sendo $I_{\min}$ um limiar mínimo de irradiância (50 W/m²). Nestes casos, P_{exp} foi utilizado para estimar a energia não gerada.

5.2.2.8 Integração na Base de Dados

Os valores de geração esperada e componentes de perda calculados foram incorporados como *features* adicionais na base de dados na matriz $\mathcal{B}$ de dados originais, resultando na matriz $\mathcal{D}$. Tendo como objetivo:

- Preenchimento de lacunas de dados com estimativas físicas consistentes;
- Referência de comparação para validação de previsões;
- Detecção de anomalias operacionais.

5.3 NORMALIZAÇÃO DOS DADOS

Para uniformizar a escala das variáveis e melhorar a convergência dos modelos, aplica-se a normalização *Min–Max* individualmente a cada coluna $x^{(v)}$ da matriz $\mathcal{D}$, considerando apenas o conjunto de treino $\mathcal{D}_{\text{treino}}$. A transformação é dada por:

$$\tilde{x}_t^{(v)} = \frac{x_t^{(v)} - \min\left(x^{(v)}\right)}{\max\left(x^{(v)}\right) - \min\left(x^{(v)}\right)}, \tag{5.9}$$

onde $\min(x^{(v)})$ e $\max(x^{(v)})$ são, respectivamente, o menor e o maior valor da variável $x^{(v)}$ no conjunto de treino. O resultado da normalização é a matriz $\mathcal{H} \in [0,1]^{T \times V}$, que preserva as tendências temporais e minimiza o impacto de diferenças de magnitude entre as variáveis.

5.4 ANÁLISE DE CORRELAÇÃO

A matriz $\mathcal{H}$ foi utilizada para avaliar a relação linear entre a variável-alvo y (geração fotovoltaica) e as variáveis explicativas $x^{(v)}$. O coeficiente de correlação de Pearson quantifica o grau de associação linear entre duas variáveis contínuas, sendo definido como:

$$\rho_{X,Y} = \frac{\sum_{t=1}^{T}\left(x_t - \bar{x}\right)\left(y_t - \bar{y}\right)}{\sqrt{\sum_{t=1}^{T}\left(x_t - \bar{x}\right)^2}\,\sqrt{\sum_{t=1}^{T}\left(y_t - \bar{y}\right)^2}}, \tag{5.10}$$

onde x_t e y_t representam as observações das variáveis no instante t, e $\bar{x}$ e $\bar{y}$ correspondem às respectivas médias amostrais. O coeficiente assume valores entre -1 e 1, indicando correlação negativa, nula ou positiva.

A matriz de correlação apresentada na Figura 19 permite observar as inter-relações entre as variáveis analisadas. Identifica-se elevada colinearidade entre os diferentes tipos de irradiância (*poa, ghi, gri*) e entre as variáveis de potência calculadas (*power_ideal, power_adjusted, power_final*). Em contrapartida, variáveis meteorológicas como direção do vento e precipitação acumulada apresentam baixos coeficientes, sugerindo influência linear reduzida sobre a geração de energia.

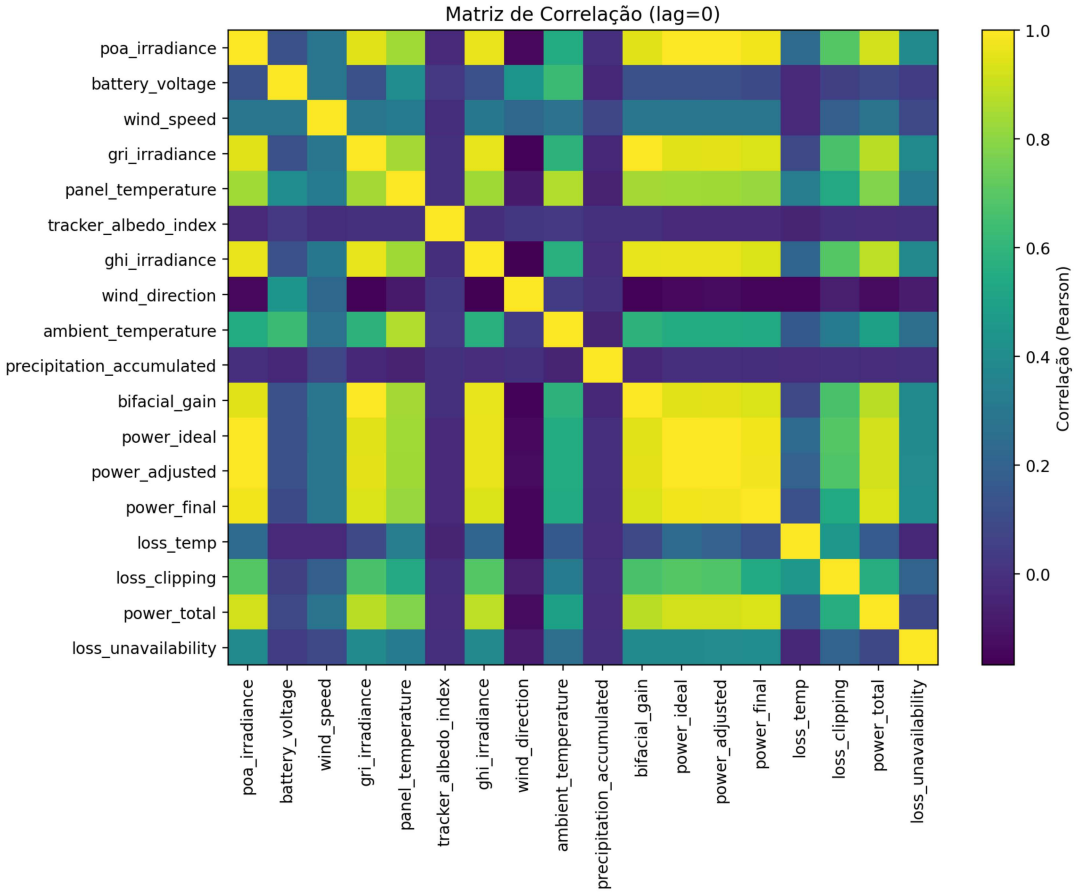


Figura 19: Matriz de correlação de Pearson ($\ell = 0$) entre as variáveis de entrada.

Fonte: Autoria própria (2025).

5.4.1 Correlação com defasagens temporais

Para investigar a dependência temporal da geração fotovoltaica, o coeficiente de Pearson foi estendido para incluir defasagens (*lags*) nas variáveis de entrada. O coeficiente para um atraso ℓ é definido como:

$$\rho_{y,x^{(v)}}(\ell) = \frac{\sum_{t=1+\ell}^{T} (y_t - \bar{y}) \left(x_{t-\ell}^{(v)} - \overline{x^{(v)}}\right)}{\sqrt{\sum_{t=1+\ell}^{T} (y_t - \bar{y})^2} \sqrt{\sum_{t=1+\ell}^{T} \left(x_{t-\ell}^{(v)} - \overline{x^{(v)}}\right)^2}}, \tag{5.11}$$

em que $\ell \in \{1, \ldots, \ell_{\max}\}$ representa o número de amostras defasadas. Foram analisadas duas escalas temporais: geração diária acumulada e geração intradiária em intervalos de 15 minutos.

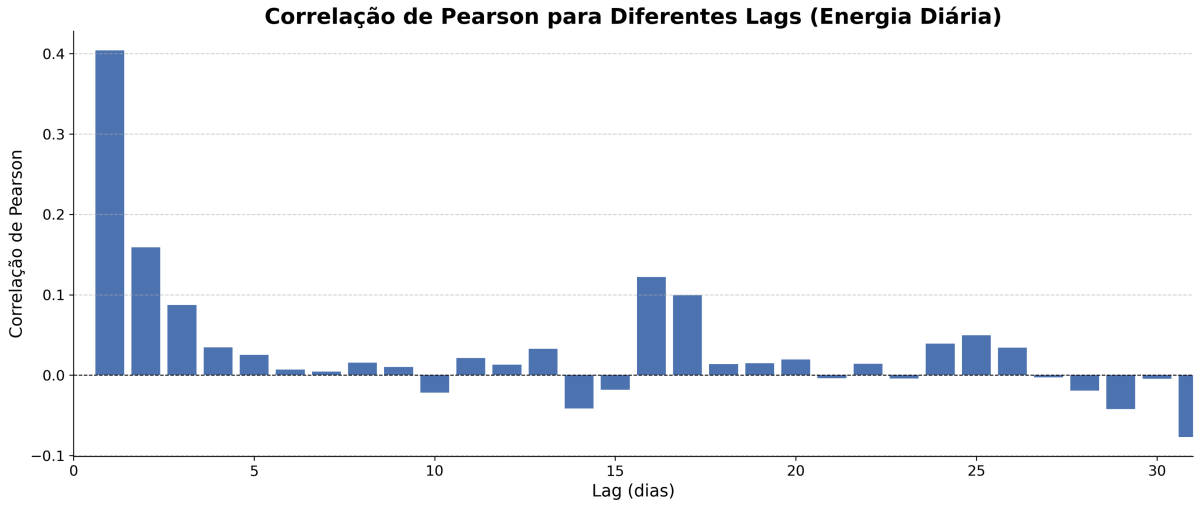


Figura 20: Análise de correlação dos lags para geração acumulada diária.

Fonte: Autoria própria (2025).

Na escala diária, observa-se correlação positiva significativa nos primeiros atrasos, com destaque para o *lag* 1 (dia anterior), o que evidencia a forte persistência temporal da geração entre dias consecutivos. Essa correlação diminui progressivamente com o aumento da defasagem, tornando-se próxima de zero após cinco dias.

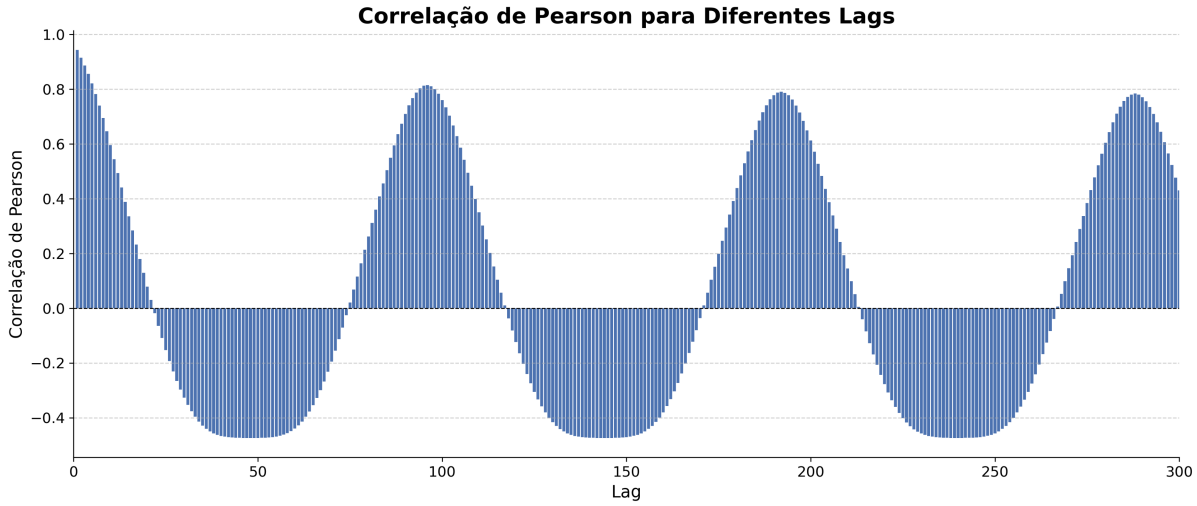


Figura 21: Análise detalhada dos lags para a geração a cada 15 minutos.

Fonte: Autoria própria (2025).

O comportamento periódico e simétrico das curvas indica forte sazonalidade diária, que pode ser representada matematicamente por funções periódicas, como o cosseno ou a gaussiana, conforme discutido na Seção 5.5.

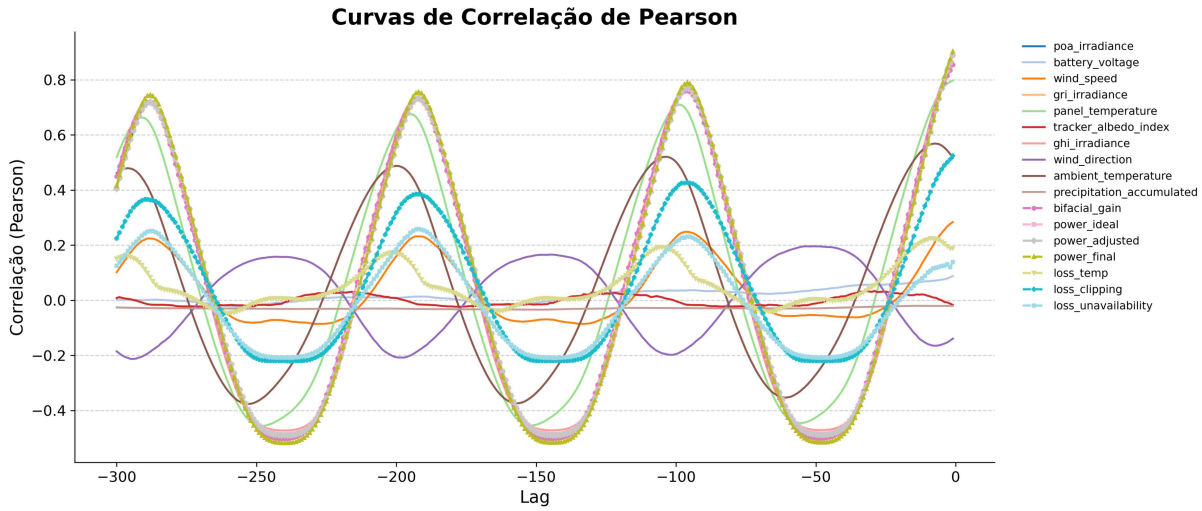


Figura 22: Curvas de correlação de Pearson considerando múltiplas variáveis e defasagens.

Fonte: Autoria própria (2025).

A Figura 22 apresenta as curvas de correlação obtidas para diferentes variáveis e defasagens na escala intradiária. Verifica-se comportamento periódico, com picos de correlação positiva em múltiplos de 96 amostras (equivalentes a 24 horas na amostragem de 15 minutos) e correlações negativas em múltiplos de 48 amostras (12 horas). Esse padrão reflete o ciclo diurno da geração solar: altos valores de potência próximos ao meio-dia estão associados a baixos valores noturnos.

5.4.2 Informação Mútua e PMI

Embora o coeficiente de Pearson quantifique dependências lineares, ele é limitado na identificação de relações não lineares entre as variáveis. Para capturar dependências mais gerais, foi empregada a Informação Mútua (*Mutual Information*, MI), que mede o grau de interdependência estatística entre duas variáveis aleatórias X e Y.

A MI é dada por:

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)}, \tag{5.12}$$

onde $p(x,y)$ representa a distribuição conjunta de X e Y, e $p(x)$ e $p(y)$ são as distribuições marginais.

Complementarmente, a Informação Mútua Pontual (PMI, *Pointwise Mutual Information*) quantifica a contribuição individual de cada par de eventos (x,y), sendo expressa como:

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)\, p(y)}. \tag{5.13}$$

Valores elevados de PMI indicam associação estatística forte, mesmo em relações não lineares. No presente estudo, as curvas de MI foram utilizadas para identificar defasagens e variáveis com maior dependência informacional em relação à geração fotovoltaica.

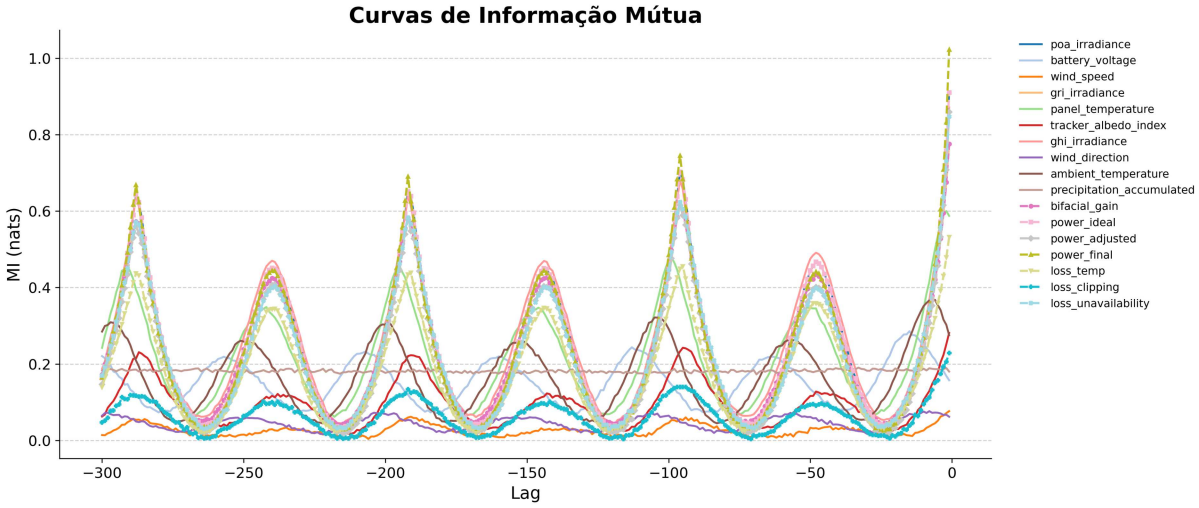


Figura 23: Curvas de Informação Mútua entre variáveis e a geração fotovoltaica.

Fonte: Autoria própria (2025).

A Figura 23 evidencia que as variáveis de irradiância e temperatura dos módulos apresentam os maiores valores de MI, reforçando sua relevância para a modelagem da geração. Assim como observado nas curvas de correlação, há periodicidade marcada, refletindo o comportamento cíclico e dependente do tempo característico da produção solar.

A combinação das análises de Pearson e MI fornece uma visão abrangente das relações entre as variáveis, permitindo identificar dependências lineares e não lineares relevantes para a previsão da geração. Com base nesses resultados, foram definidos os *lags* utilizados na expansão da matriz de entrada $\mathcal{J}$, conforme descrito na Seção subsequente.

5.4.3 Definição de variáveis defasadas

Com base na análise de correlação, foram incluídas novas colunas correspondentes às defasagens temporais. Para o caso diário, foram consideradas até $\ell_{\max} = 5$. Já para o caso intradiário, além das primeiras defasagens, foi incorporado explicitamente o *lag* 96, correspondente à geração no mesmo horário do dia anterior:

$$x_{t-\ell}^{(v)}, \quad \ell = 1, \ldots, \ell_{\max}. \tag{5.14}$$

A inclusão dessas variáveis resulta na matriz $\mathcal{J} \in \mathbb{R}^{T \times V'}$, onde V' representa o número total de variáveis após a expansão com os lags.

5.5 MODELAGENS PERIÓDICAS

5.5.1 Modelagem sazonal com função cosseno

Para representar a sazonalidade anual e a variação da irradiação ao longo do ano, de acordo com as estações do ano, é adicionada uma nova variável calculada como:

$$\text{Cosseno}_t = \cos\left(2\pi \frac{\text{Dia}_t}{\text{Ano}}\right), \tag{5.15}$$

onde Dia_t é o dia do ano no instante t e $\text{Ano} = 365$. A inserção desta coluna gera a matriz $\mathcal{K}$.

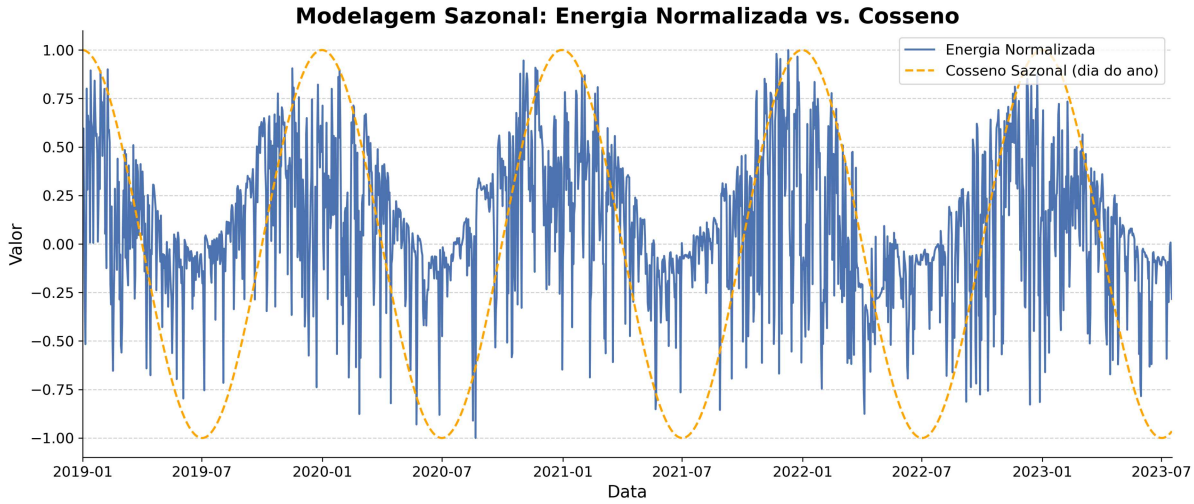


Figura 24: Análise de modelagem sozonal.

Fonte: Autoria própria (2025).

5.5.2 Modelagem intradiária com função gaussiana

A variação da potência fotovoltaica em um dia ideal, sem nuvens no céu, apresenta um comportamento típico, com máximo próximo ao meio-dia solar e valores próximos de zero no início da manhã e no final da tarde, tendo *clipping* na potência máxima dos inversores, conforme detalhado na Seção *5.2.2.6*. Para representar esse padrão, foi utilizada uma função gaussiana parametrizada, definida como:

$$G(t) = \min\left(1,\ A \cdot \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\right), \tag{5.16}$$

onde t representa a hora do dia em formato decimal, $\mu = 12$ corresponde ao horário de pico (12:00), σ foi ajustado de forma que os valores da função sejam praticamente nulos

nos extremos do intervalo de geração (entre 6:00 e 18:00), e A é um fator de amplitude calculado de modo que a função permaneça maior ou igual a 1 no intervalo de 10:00 a 13:00. O resultado passa por uma função $\min(x, y)$ que garante a modelagem do clipping dos inversores.

A coluna correspondente a esta modelagem foi adicionada à matriz de dados $\mathcal{K}$, utilizada para as previsões intradiárias com resolução de 15 minutos.

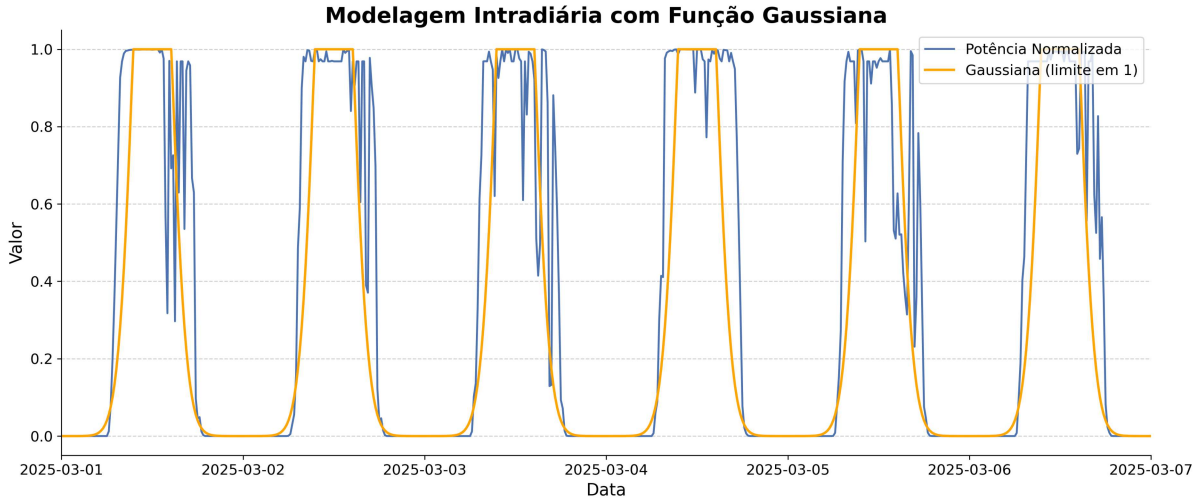


Figura 25: Função gaussiana proposta para modelagem do perfil intradiário de potência.

Fonte: Autoria própria (2025).

5.5.3 Média Móvel

Com o objetivo de ampliar o contexto histórico disponível ao modelo e suavizar flutuações pontuais, foi calculada, para cada variável $x^{(v)}$, a média móvel simples de janela j, definida como:

$$\mathrm{MA}_t^{(v)} = \frac{1}{j} \sum_{k=0}^{j-1} x_{t-k}^{(v)}. \tag{5.17}$$

No caso da previsão de geração acumulada diária, foram utilizadas janelas de 14 e 28 dias, de modo a capturar padrões de médio prazo.

Para a previsão intradiária, com registros a cada 15 minutos, foram empregadas janelas de 96 e 288 amostras, correspondentes a 1 dia e 3 dias, respectivamente, permitindo ao modelo explorar o comportamento em horizontes mais curtos.

A inclusão dessas médias móveis gera a matriz $\mathcal{L}$, que combina as variáveis originais com suas representações suavizadas, enriquecendo o conjunto de atributos e favorecendo a identificação de padrões temporais relevantes.

6 AVALIAÇÃO DE MODELOS DE APRENDIZADO DE MÁQUINA

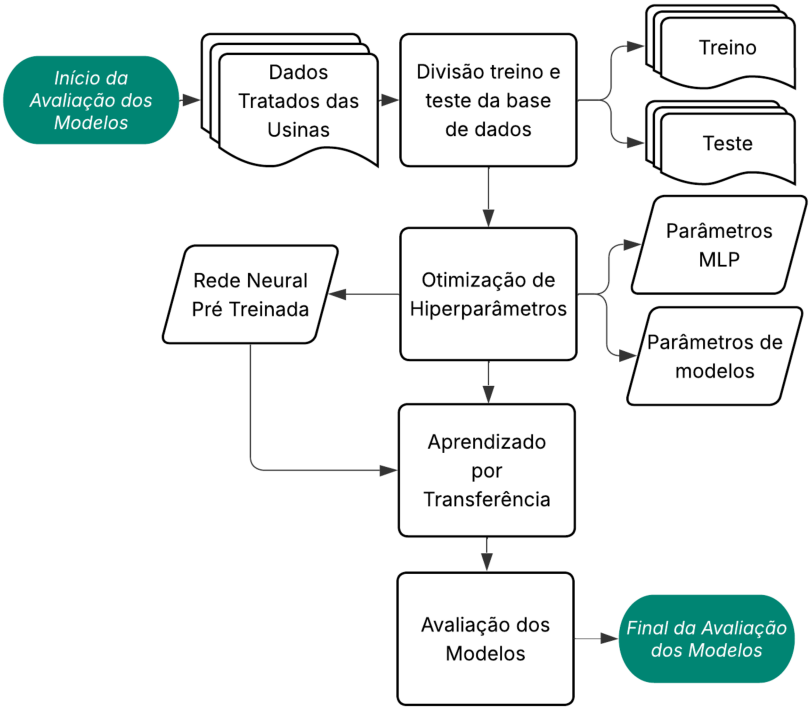


Figura 26: Diagrama da metodologia de avaliação de modelos de aprendizado de máquina.

Fonte: Autoria própria (2025).

6.1 TREINO E TESTE

A matriz final $\mathcal{L}$ foi utilizada para gerar partições de treino e teste respeitando a ordem temporal dos registros. A divisão seguiu a forma convencional empregada em séries temporais, em que a porção inicial do conjunto é destinada ao treinamento do modelo e a parte final é reservada para o teste, sem sobreposição entre as amostras. Neste trabalho foi utilizada a divisão da base de dados em treino (80%) e teste (20%) respeitando a ordem temporal.

Dessa forma, assegura-se que as previsões sejam realizadas sempre em instantes posteriores aos utilizados no ajuste do modelo, refletindo um cenário mais próximo da aplicação prática de previsão de geração em tempo real.

O conjunto de treino foi utilizado para ajuste dos parâmetros e hiperparâmetros dos modelos, enquanto o conjunto de teste permaneceu isolado, sendo empregado exclusivamente para avaliação do desempenho final. Esse procedimento reduz o risco de sobreajuste (*overfitting*) e garante maior confiabilidade na interpretação das métricas obtidas.

6.2 MÉTRICAS DE AVALIAÇÃO

Sejam $\mathbf{y} = \{y_1, y_2, \ldots, y_n\}$ os valores reais da série e $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\}$ os valores previstos correspondentes. As métricas utilizadas para avaliar o desempenho dos modelos foram:

6.2.1 Raiz do Erro Quadrático Médio (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}, \tag{6.1}$$

a qual penaliza erros de maior magnitude, sendo amplamente utilizada em problemas de previsão de séries temporais.

6.2.2 Erro Percentual Médio Absoluto (MAPE)

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|, \tag{6.2}$$

que expressa o erro médio relativo em porcentagem, facilitando a interpretação do desempenho do modelo.

6.2.3 Raiz do Erro Quadrático Médio Normalizado (NRMSE)

$$\text{NRMSE} = \frac{\text{RMSE}}{\max(\mathbf{y}) - \min(\mathbf{y})}, \tag{6.3}$$

obtido pela normalização do RMSE pela amplitude dos valores reais, permitindo comparações entre diferentes séries.

6.2.4 Erro Percentual Médio Absoluto Simétrico (SMAPE)

Para avaliar o desempenho de forma independente de escala, empregou-se o Erro Percentual Médio Absoluto Simétrico (SMAPE), definido como:

$$\text{SMAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2}. \tag{6.4}$$

O SMAPE varia de 0% (acurácia perfeita) a 200%.

6.3 BUSCA POR HIPERPARÂMETROS

A escolha adequada de hiperparâmetros é determinante para o desempenho de modelos de previsão baseados em *machine learning*.

A técnica de Busca em Grade, em inglês *Grid Search* foi utilizada para a otimização de hiperparâmetros nos modelos tradicionais de aprendizado supervisionado e nas redes neurais do tipo *Multi-Layer Perceptron* (MLP). Esse método consiste em uma busca sistemática e exaustiva sobre um conjunto previamente definido (grade) de possíveis valores para cada hiperparâmetro.

O processo de execução ocorre de forma iterativa, em que todas as combinações possíveis entre os valores dos hiperparâmetros são avaliadas. Para cada combinação, o modelo é treinado e testado utilizando o conjunto de validação, sendo o desempenho quantificado pela métrica SMAPE. Ao término da busca, seleciona-se a combinação de hiperparâmetros que apresentou o menor valor de erro, garantindo, assim, que o modelo final seja configurado para minimizar o erro percentual médio absoluto nas previsões.

Tabela 7: Hiperparâmetros utilizados no *Grid Search*

Modelo	**Hiperparâmetros**	**Faixa de Busca**
MLP	Hidden Layer Size	[(128, 64), (256, 128, 64), (128, 128, 64, 32), (8,16,32,64,128,64,32,16,8)]
	Activation	['relu', 'elu']
	Solver	['adam', 'sgd', 'nadam', 'rmsprop']
	Alpha	[1e-5, 1e-4]
	Batch Size	[32, 64]
	Max Iterations	[300, 400, 500]
RF	Max Features	['log2', 'sqrt']
	Min Samples Split	[5, 6, 7, 8, 9, 10, 11]
	N Estimators	Range (100, 350, 10)
XGB	Eta	[0.006, 0.008, 0.009, 0.01, 0.015, 0.017, 0.019, 0.02]
	Gamma	Range (150, 310, 10)
	N Estimators	Range (70, 90, 2)
	Subsample	[0.5, 0.75, 1]
GB	Learning Rate	[0.005, 0.007, 0.009, 0.01, 0.02, 0.03]
	Max Depth	[2, 4, 6]
	Min Samples Split	[5, 6, 7, 8, 9, 10, 11]
	N Estimators	Range (400, 500, 10)
ARIMA	Order (p, d, q)	[(0,0,0), (0,0,1), (0,0,2), (0,1,0), (0,1,1), (0,1,2), (1,0,0), (1,0,1), (1,0,2), (1,1,0), (1,1,1), (1,1,2), (2,0,0), (2,0,1), (2,0,2), (2,1,0), (2,1,1), (2,1,2)]
	Trend	[None, 'n', 'c']

7 RESULTADOS

7.1 CONFIGURAÇÃO COMPUTACIONAL E FERRAMENTAS

Os experimentos foram conduzidos em um computador com processador Intel Core i5-12450H, frequência de até 4,40 GHz, 16 GB de memória RAM e sistema operacional Windows 11.

O desenvolvimento foi realizado em Python 3.11, utilizando o ambiente de desenvolvimento Visual Studio Code (VS Code). Os principais pacotes empregados foram:

- `numpy` e `pandas`: manipulação e análise de dados;
- `matplotlib`: geração de gráficos e visualizações;
- `scikit-learn`: implementação de modelos de regressão, *grid search* e métricas de avaliação;
- `tensorflow` e `keras`: construção e treinamento das Redes Neurais Artificiais;

7.2 BASE DE DADOS BR-PVGEN

Um dos principais resultados deste trabalho consiste na proposição da base **BR-PVGen**, que reúne séries temporais de geração de energia elétrica de diferentes usinas fotovoltaicas em território brasileiro. Essa base foi estruturada com o objetivo de disponibilizar um repositório padronizado de dados para pesquisa em previsão de geração, avaliação de desempenho e modelagem de aprendizado de máquina aplicada a sistemas solares.

7.2.1 Descrição da Base

A BR-PVGen contempla tanto dados de geração quanto informações meteorológicas e técnicas, de forma a permitir análises completas de desempenho. Sua estrutura inclui:

- **Resolução temporal**: dados diários e intradiários (com passo de 15 minutos);
- **Variáveis disponíveis**: energia gerada (kWh), irradiância global, temperatura ambiente, temperatura do módulo, velocidade do vento e dados de operação de inversores;
- **Período de cobertura**: máximo de 26 de março de 2024 a 9 de junho de 2025, variando de acordo com a disponibilidade de cada usina;

7.2.2 Cálculo da Geração Esperada e Perdas

A geração esperada de cada usina foi estimada a partir da irradiância no plano do módulo (POA), da potência nominal dos inversores e dos parâmetros técnicos do sistema fotovoltaico. Conforme descrito na Seção **5.2.1**, a energia ideal pode ser calculada com base na irradiação do sol medida [W/m^2], total de área dos módulos [m^2] e eficiência de conversão dos módulos [%] (normalmente próximo de 21%).

Com base nesse indicador e demais perdas do sistema calculados conforme **5.2.2**, a Figura 27 apresenta o diagrama em cascata com a decomposição dessas perdas, partindo da geração ideal até a energia efetivamente entregue.

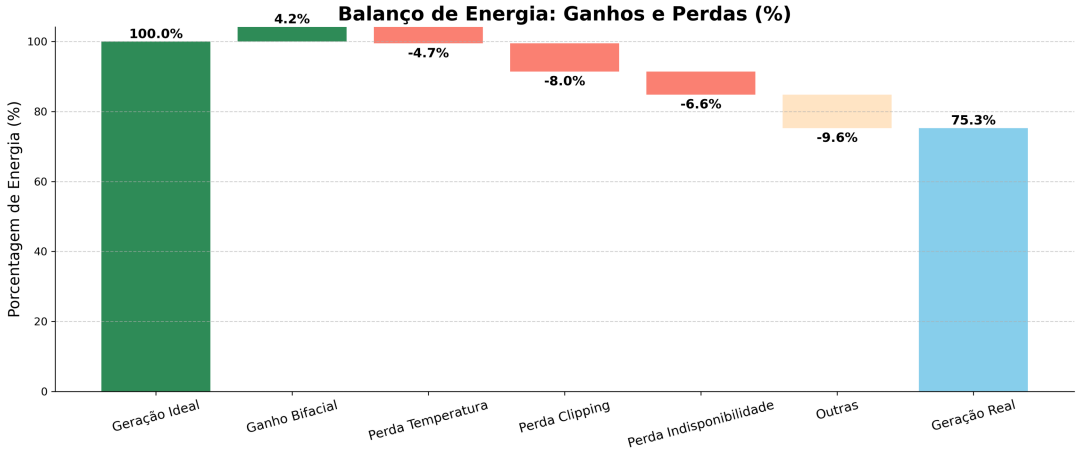


Figura 27: Diagrama em cascata das perdas: da geração ideal até a geração efetiva.

Fonte: Autoria própria (2025).

Observa-se que a energia efetiva representa em média 75,3% do montante esperado, indicador que pode ser traduzido pelo **Performance Ratio (PR)**:

$$PR = \frac{E_{\text{real}}}{E_{\text{ideal}}}. \tag{7.1}$$

As principais perdas identificadas foram:

- Temperatura: aproximadamente 4,7%;
- *Clipping* dos inversores: 8%;
- Indisponibilidade: 6,6%;
- Outros (9,6%): perdas por sujidade, sombreamentos, falhas de sensores e incertezas do cálculo de E_{ideal}.

7.3 RESULTADOS DO *GRID SEARCH*

A busca exaustiva por hiperparâmetros foi realizada por meio do *Grid Search*, contemplando 51 usinas da base BR-PVGen. A Tabela 8 apresenta os valores selecionados como ótimos.

Tabela 8: Hiperparâmetros utilizados no *Grid Search* e valores selecionados

Modelo	Hiperparâmetros	Faixa de Busca	Parâmetros Selecionados
MLP	Hidden Layer Size	[(128, 64), (256, 128, 64), (128, 128, 64, 32), (8,16,32,64,128,64,32,16,8)]	(256, 128, 64)
	Activation	['sigmoid', 'tanh', 'relu', 'elu']	Relu
	Solver	['adam', 'sgd', 'nadam', 'rmsprop']	adam
	Alpha	[1e-5, 1e-4]	1e-5
	Batch Size	[32, 64]	64
	Max Iterations	[300, 500, 800]	800
RF	Max Features	['log2', 'sqrt']	log2
	Min Samples Split	[5, 6, 7, 8, 9, 10, 11]	7
	N Estimators	Range (100, 350, 10)	100
XGB	Eta	[0.006, 0.008, 0.009, 0.01, 0.015, 0.017, 0.019, 0.02]	0.02
	Gamma	Range (150, 310, 10)	260
	N Estimators	Range (70, 90, 2)	88
	Subsample	[0.5, 0.75, 1]	0.5
GB	Learning Rate	[0.005, 0.007, 0.009, 0.01, 0.02, 0.03]	0.005
	Max Depth	[2, 4, 6]	4
	Min Samples Split	[5, 6, 7, 8, 9, 10, 11]	10
	N Estimators	Range (400, 500, 10)	410
ARIMA	Order (p, d, q)	[(0,0,0), (0,0,1), (0,0,2), (0,1,0), (0,1,1), (0,1,2), (1,0,0), (1,0,1), (1,0,2), (1,1,0), (1,1,1), (1,1,2), (2,0,0), (2,0,1), (2,0,2), (2,1,0), (2,1,1), (2,1,2)]	(0, 0, 1)
	Trend	[None, 'n', 'c']	'n'

Além dos hiperparâmetros principais, outros mecanismos foram empregados para estabilizar o treinamento:

1. **Função de perda Huber**: reduziu a sensibilidade a outliers, comuns em dados de usinas com indisponibilidade ou falhas de medição.
2. **ReduceLROnPlateau**: permitiu refinamento da taxa de aprendizado quando a validação estagnava, garantindo ajustes mais finos.
3. **Early Stopping**: interrompeu treinamentos sem ganho após 40 épocas, evitando sobreajuste e reduzindo tempo computacional.
4. **Regularização L2 adicional**: aplicada em todas as camadas da MLP, restringiu o crescimento dos pesos e melhorou generalização.

Em conjunto, essas escolhas resultaram em modelos mais robustos, com convergência estável e menor custo computacional, especialmente nas arquiteturas de MLP com *transfer learning*, que apresentaram os melhores resultados médios.

Para comparação do desempenho dos modelos com seus respectivos melhores hiperparâmetros, foi realizada a previsão da geração das usinas fotovoltaicas, forma

individual por usina e, em seguida, consolidada por meio da média, desvio padrão, valores mínimo e máximo. Os resultados estão organizados nas Tabelas 9,10, referentes à previsão de energia diária, e nas Tabelas 11 e 12 que apresentam a previsão de potência em escala intradiária.

Tabela 9: Desempenho dos modelos - SMAPE (%) - Previsão energia diária

Modelo	**Menor Erro**	**Maior Erro**	**Média ± Desvio**
ARIMA	55.47	151.45	117.90 ± 21.70
Gradient Boosting	64.77	159.01	126.16 ± 19.98
Random Forest	61.42	169.21	128.13 ± 21.76
XGBoost	67.62	168.96	127.68 ± 21.32
MLP	60.94	160.82	120.87 ± 23.08
MLP com Transfer Learning	76.08	156.50	116.26 ± 16.19

Tabela 10: Desempenho dos modelos - NRMSE (%) - Previsão energia diária

Modelo	**Menor Erro**	**Maior Erro**	**Média ± Desvio**
ARIMA	13.72	67.04	25.52 ± 9.00
Gradient Boosting	17.47	48.81	25.51 ± 5.62
Random Forest	14.56	46.12	25.38 ± 5.28
XGBoost	16.04	51.12	25.63 ± 5.89
MLP	17.04	69.66	26.08 ± 8.32
MLP com Transfer Learning	15.56	52.87	24.94 ± 5.68

Tabela 11: Desempenho dos modelos - SMAPE (%) - Previsão energia em 15 min

Modelo	**Menor Erro**	**Maior Erro**	**Média ± Desvio**
ARIMA	22.04	72.89	47.05 ± 8.66
Gradient Boosting	29.63	73.97	46.94 ± 8.40
Random Forest	33.05	76.70	47.33 ± 8.17
XGBoost	27.97	72.92	46.48 ± 8.13
MLP	26.98	77.77	46.40 ± 8.28
MLP com Transfer Learning	28.15	71.57	43.92 ± 6.29

Tabela 12: Desempenho dos modelos - NRMSE (%) - Previsão energia em 15 min

Modelo	**Menor Erro**	**Maior Erro**	**Média ± Desvio**
ARIMA	9.81	25.36	14.23 ± 2.46
Gradient Boosting	9.37	27.16	13.99 ± 2.82
Random Forest	9.34	28.77	13.82 ± 2.95
XGBoost	9.37	28.98	13.92 ± 2.89
MLP	10.36	26.76	14.49 ± 2.59
MLP com Transfer Learning	9.32	21.89	14.07 ± 2.40

7.4 TRANSFER LEARNING

Como evidenciado nas Tabelas 9,10,11 e 12, os resultados da MLP foram aprimorados com a aplicação da técnica de *Transfer Learning*. Nessa abordagem, a usina com maior histórico de registros (id=1) serviu como base para o pré-treinamento do modelo, mantendo-se congelados os pesos das duas camadas iniciais e permitindo o ajuste apenas da camada final durante o re-treinamento em cada nova usina.

A Tabela 13 resume os valores médios de SMAPE antes e depois da aplicação do *Transfer Learning*, para o caso da previsão de energia diária, considerando diferentes faixas de disponibilidade de registros. Nota-se que, para o conjunto completo de 51 usinas, houve redução média de **3,81%**, passando de 120,87 para 116.26. Os ganhos tornam-se mais expressivos à medida que diminui o volume de dados: **6,60%** de redução em usinas com menos de 200 registros e **9,03%** quando disponíveis apenas 100 registros ou menos.

Tabela 13: Comparação de resultado antes e depois do *Transfer Learning* - Energia Diária

Caso	**Usinas**	**SMAPE Médio Antes**	**SMAPE Médio Depois**	**Redução (%)**
Todas usinas	51	120.87	116.26	3.81
$<$ 400 registros	46	118.35	114.68	3.09
$<$ 300 registros	38	118.91	114.73	3.51
$<$ 200 registros	20	125.79	117.56	6.60
$\leq$ 100 registros	10	127.66	116.12	9.03

Tabela 14: Comparação de resultado antes e depois do *Transfer Learning* - Energia Diária (NRMSE)

Caso	**Usinas**	**NRMSE Médio Antes**	**NRMSE Médio Depois**	**Redução (%)**
Todas usinas	51	26.17	24.94	4.71
$<$ 400 registros	46	26.28	25.10	4.49
$<$ 300 registros	38	27.16	25.94	4.50
$<$ 200 registros	20	29.02	27.30	5.94
$\leq$ 100 registros	10	29.80	26.82	9.99

A Figura 28 ilustra essa diferença, mostrando que a maioria das usinas obteve ganhos após a aplicação da técnica.

Resultados ainda mais expressivos foram observados no cenário intradiário. A Tabela 15 mostra que, para todas as 51 usinas, o SMAPE médio caiu de 47.24 para 43.92, representando uma redução de **7,02%**. O ganho cresce em condições de menor quantidade de dados, chegando a **11,06%** para usinas com menos de 4800 registros.

A Figura 29 complementa a análise, destacando a consistência da melhoria entre diferentes usinas.

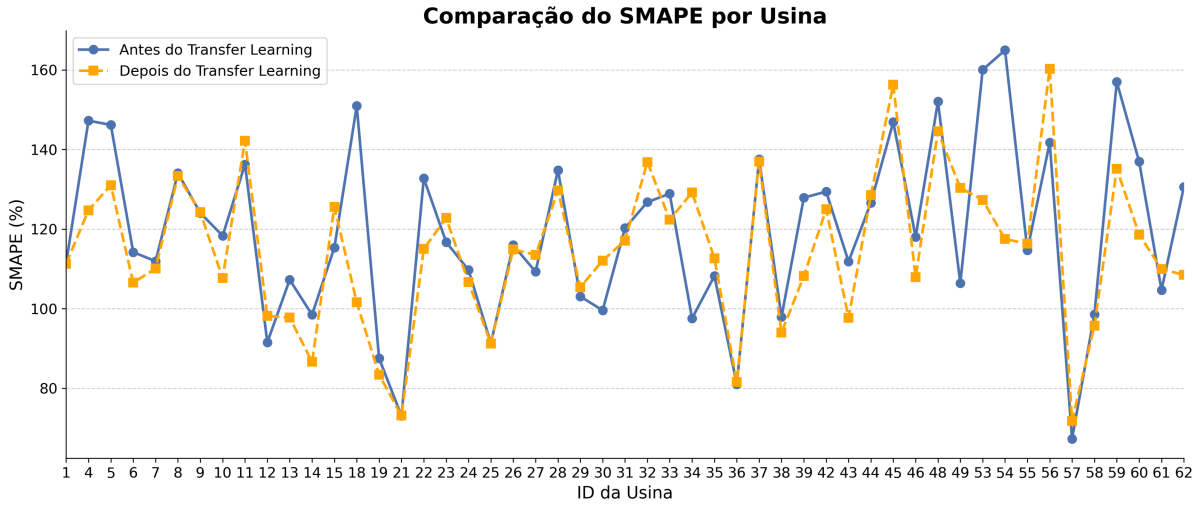


Figura 28: Comparação de resultado antes e depois do *Transfer Learning* - Energia Diária.

Fonte: Autoria própria (2025).

Tabela 15: Comparação de resultado antes e depois do *Transfer Learning* - (15 minutos)

Caso	**Usinas**	**SMAPE Médio Antes**	**SMAPE Médio Depois**	**Redução(%)**
Todas usinas	51	47.24	43.92	7.02
$<$ 19200 registros	46	46.69	42.45	9.09
$<$ 14400 registros	38	46.90	42.34	9.72
$<$ 9600 registros	20	46.07	42.54	7.68
$\leq$ 4800 registros	10	58.27	51.83	11.06

Tabela 16: Comparação de resultado antes e depois do *Transfer Learning* - (15 minutos) (NRMSE)

Caso	**Usinas**	**NRMSE Médio Antes**	**NRMSE Médio Depois**	**Redução (%)**
Todas usinas	51	14.50	14.07	2.96
$<$ 19200 registros	46	15.32	14.54	5.10
$<$ 14400 registros	38	16.07	15.21	5.34
$<$ 9600 registros	20	17.71	16.93	4.43
$\leq$ 4800 registros	10	19.66	18.53	5.76

Em ambos os horizontes de previsão, o *Transfer Learning* demonstrou ganhos consistentes, com reduções mais modestas na previsão diária e melhorias significativas na previsão intradiária.

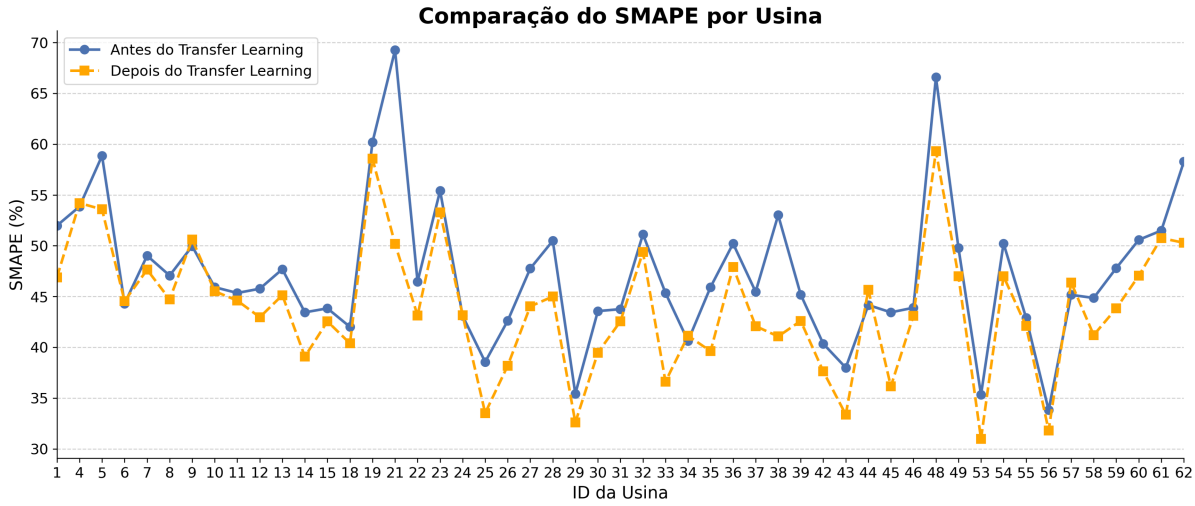


Figura 29: Comparação do SMAPE por usina antes e depois do *Transfer Learning* (15 minutos).

Fonte: Autoria própria (2025).

Além do aumento de acurácia, destaca-se a economia computacional: o tempo médio de treinamento foi aproximadamente **40% menor** em comparação ao treinamento do zero, já que o modelo inicia com pesos pré-ajustados e apenas as camadas finais passam por ajuste fino. Essa combinação de melhor desempenho e maior eficiência reforça o potencial do *Transfer Learning* para aplicações em larga escala em usinas fotovoltaicas.

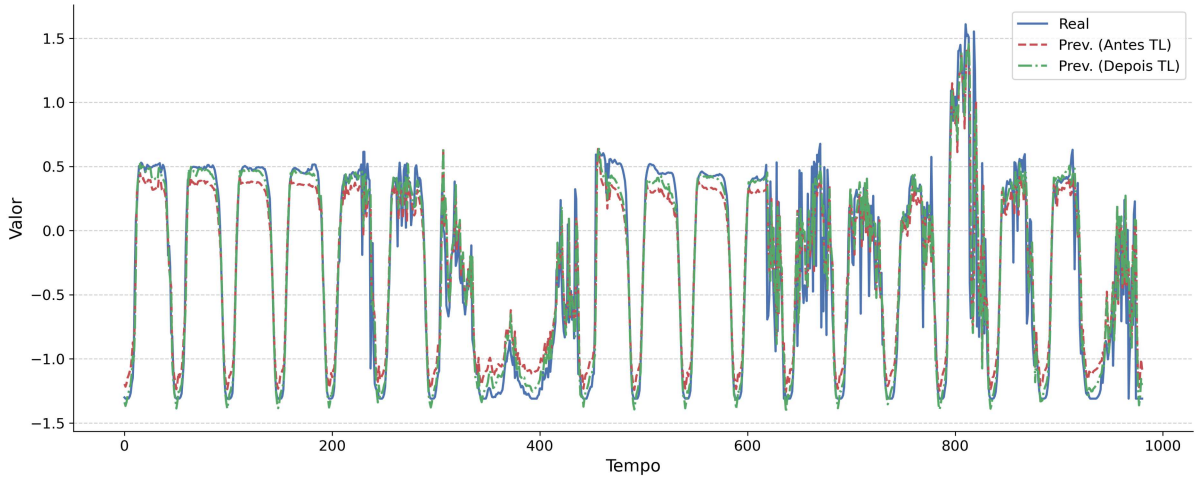


Figura 30: Exemplo de comparação do SMAPE em uma usina antes e depois do *Transfer Learning*.

Fonte: Autoria própria (2025).

7.4.1 Curvas de Aprendizado

As curvas de aprendizado constituem uma ferramenta essencial para analisar o comportamento da rede neural ao longo do processo de otimização. A Figura 31 apresenta o histórico de treinamento da usina base, enquanto a Figura 32 ilustra o comportamento do modelo durante o ajuste para a usina alvo, utilizando a estratégia de *Transfer Learning*.

No treinamento da usina base (Figura 31), observa-se uma fase inicial de maior instabilidade, com oscilações significativas na função de custo (*loss*) e na perda de validação (*val_loss*). Esse comportamento é típico do período inicial de ajuste dos pesos da rede, quando os parâmetros são atualizados de forma mais intensa em busca de uma região de convergência no espaço de otimização. A partir de aproximadamente 50 épocas, ambas as curvas tendem à estabilização, indicando que o modelo passou a generalizar de forma satisfatória.

A linha pontilhada em vermelho representa a variação da taxa de aprendizado (*learning rate*), ajustada automaticamente pelo método *ReduceLROnPlateau*. Inicialmente, são utilizados valores mais elevados de taxa de aprendizado para acelerar a descida no gradiente; à medida que a perda de validação deixa de apresentar ganhos significativos, a taxa é reduzida progressivamente, permitindo um refinamento mais preciso dos pesos e mitigando oscilações no mínimo local. O critério de parada antecipada (*Early Stopping*), indicado pela linha vertical em Época = 87, interrompeu o treinamento após a estagnação da perda em validação, prevenindo sobreajuste e garantindo o uso do melhor modelo observado.

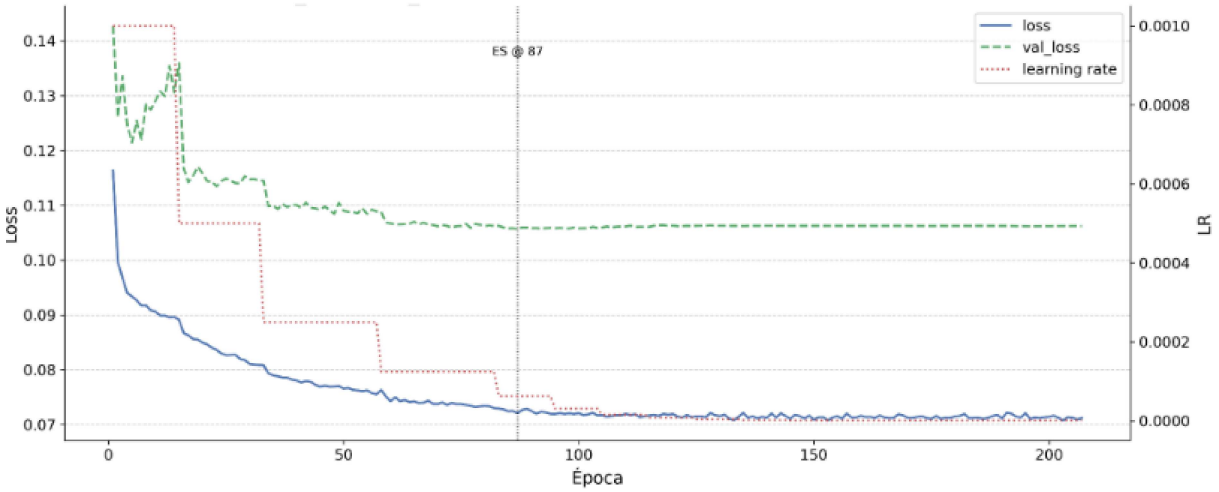


Figura 31: Curvas de treinamento e validação da rede na usina base.

Fonte: Autoria própria (2025).

Já na Figura 32, referente ao processo de adaptação do modelo pré-treinado para a usina alvo, verifica-se uma convergência consideravelmente mais rápida. Isso ocorre porque a rede parte de uma configuração inicial de pesos já ajustada com base na usina base,

reduzindo o tempo necessário para atingir uma região de ótimo local no novo domínio de dados.

A queda acentuada da função de perda nas primeiras épocas reflete o processo de ajuste fino (*fine-tuning*) das camadas finais, responsáveis por capturar as especificidades da usina alvo, como padrões locais de irradiação, temperatura e comportamento dinâmico da geração.

O *Early Stopping* foi acionado em Época = 80, momento em que a perda de validação cessou de apresentar melhorias, indicando que a rede atingiu o ponto ótimo de aprendizado para os novos dados. A taxa de aprendizado, novamente controlada por *ReduceLROnPlateau*, acompanhou essa evolução de forma adaptativa, reduzindo-se em degraus à medida que a função de custo se estabilizava.

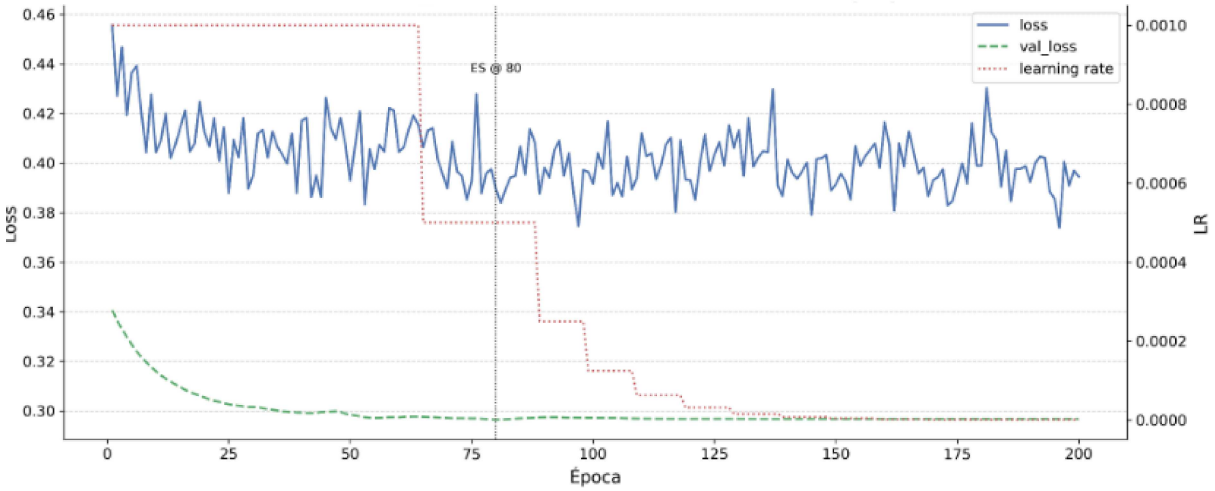


Figura 32: Curvas de treinamento e validação do modelo com *Transfer Learning* aplicado à usina alvo.

Fonte: Autoria própria (2025).

De modo geral, o comportamento observado nas curvas demonstra que o treinamento inicial na usina base foi fundamental para fornecer uma representação robusta das relações entre as variáveis meteorológicas e a geração elétrica. Esse conhecimento prévio foi reutilizado e refinado no ajuste da usina alvo, resultando em um processo de convergência mais estável e eficiente. A redução significativa da perda em menor número de épocas confirma a eficácia do *Transfer Learning* em cenários de limitação de dados, proporcionando melhor desempenho com menor custo computacional.

7.5 COMPARAÇÕES COM OUTROS MODELOS

A Tabela 17 apresenta o desempenho médio (SMAPE e NRMSE) para diferentes modelos considerando dois cenários: previsão de energia diária e previsão em intervalos de 15 minutos. Os resultados indicam uma clara evolução da MLP quando comparada aos modelos tradicionais, em especial no cenário intradiário.

Tabela 17: SMAPE e NRMSE (%) para diferentes modelos e conjuntos de dados

2*Modelo	Geração diária		Geração 15 min	
	SMAPE (%)	**NRMSE (%)**	**SMAPE (%)**	**NRMSE (%)**
MLP Transfer Learning	116.25 ± 16.19	24.94 ± 5.68	43.92 ± 6.29	14.07 ± 2.40
MLP	120.87 ± 23.08	26.17 ± 8.32	46.40 ± 8.28	14.49 ± 2.59
Linear Regression	117.90 ± 21.70	77.58 ± 24.20	49.54 ± 8.19	40.53 ± 6.61
Ridge	118.16 ± 22.06	76.39 ± 22.04	49.45 ± 8.28	40.42 ± 6.39
Lasso	174.88 ± 17.25	87.58 ± 20.38	188.04 ± 9.49	101.60 ± 10.99
ElasticNet	163.48 ± 17.78	84.92 ± 21.93	119.27 ± 8.78	75.71 ± 12.30
Decision Tree	130.62 ± 19.83	112.12 ± 25.40	57.88 ± 10.14	58.29 ± 9.19
SVR kernel	119.04 ± 24.18	80.17 ± 21.74	177.30 ± 25.03	40.96 ± 11.29
Extra Trees	130.74 ± 21.62	81.03 ± 20.22	47.11 ± 8.27	40.00 ± 8.82
AdaBoost	134.26 ± 24.80	81.85 ± 20.29	70.04 ± 15.10	49.32 ± 9.52
Bagging	127.15 ± 19.67	82.56 ± 21.07	49.77 ± 9.50	42.34 ± 8.62
KNN	130.98 ± 20.17	85.14 ± 19.56	50.11 ± 7.98	44.38 ± 9.76
Bayesian Ridge	–	74.82 ± 19.59	–	40.39 ± 6.41
Reservoir Computing	117.90 ± 21.70	–	49.54 ± 8.19	–
GB (GS)	126.16 ± 19.98	25.51 ± 5.62	46.94 ± 8.40	13.99 ± 2.82
RF (GS)	128.13 ± 21.76	25.38 ± 5.28	47.33 ± 8.17	13.82 ± 2.95
ARIMA (GS)	117.90 ± 21.70	25.52 ± 9.00	47.05 ± 8.66	14.23 ± 2.46
XGB (GS)	127.68 ± 21.32	25.63 ± 5.89	46.48 ± 8.13	13.92 ± 2.89

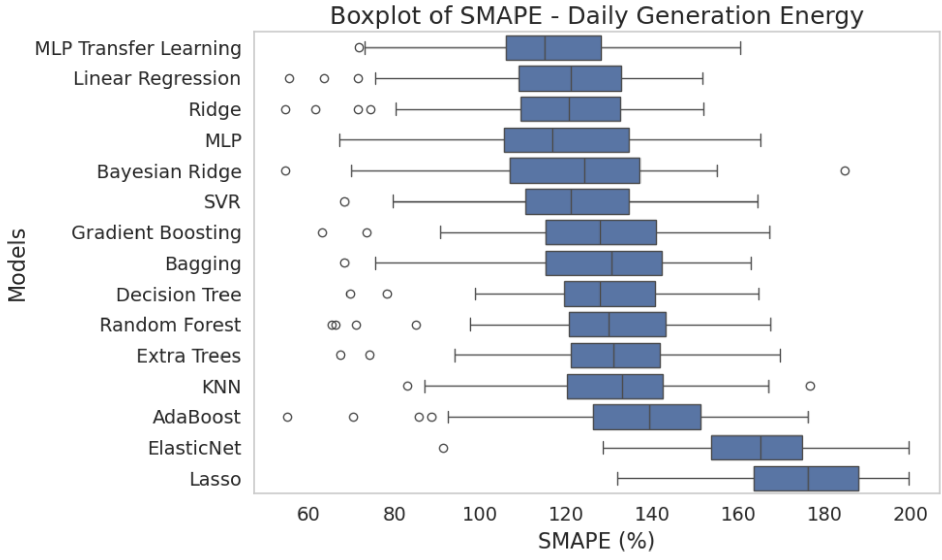


Figura 33: Distribuição do SMAPE para previsão de geração de energia diária.

Fonte: Autoria própria (2025).

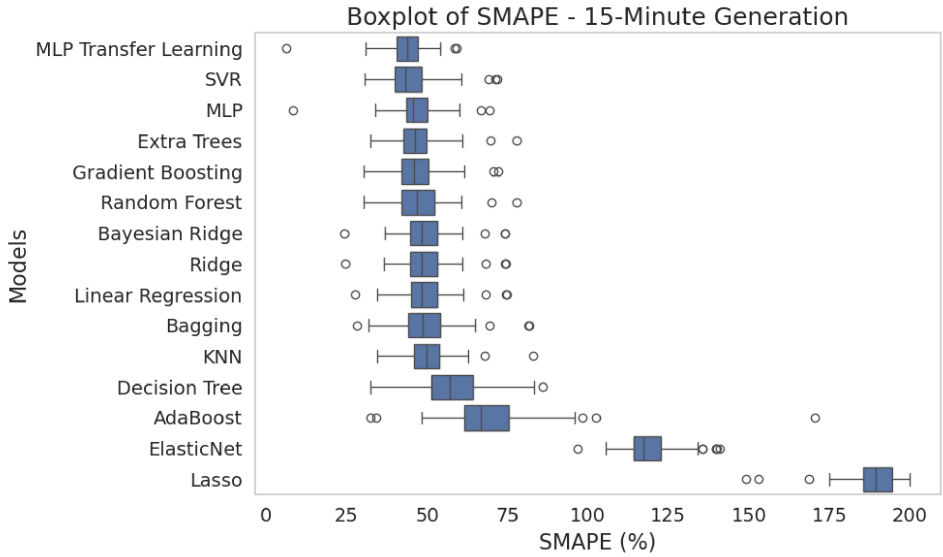


Figura 34: Distribuição do SMAPE para previsão de geração de energia em intervalos de 15 minutos.

Fonte: Autoria própria (2025).

Os boxplots das Figuras 33 e 34 confirmam que a MLP com *Transfer Learning* apresenta distribuição de erros mais concentrada e robusta frente às demais alternativas. Já os heatmaps das Figuras 35 e 36 evidenciam mostram o erro por modelo e usina.

Com o intuito de verificar a existência de diferenças estatisticamente significativas entre os modelos de previsão avaliados, foi conduzida uma análise de variância (ANOVA) de um fator, considerando as métricas de erro *SMAPE* obtidas para cada modelo ao longo das diferentes usinas.

Os resultados da ANOVA indicaram diferença altamente significativa entre os modelos, com estatística $F = 32{,}64$ e valor de $p = 4{,}10 \times 10^{-68}$. Este resultado rejeita a hipótese nula de igualdade das médias de desempenho, confirmando que pelo menos um dos modelos apresenta comportamento distinto em relação aos demais.

Para identificar quais modelos diferem entre si, foi aplicado o teste *post-hoc* de Tukey HSD, com nível de significância de 5%. Observou-se que os modelos baseados em regularização linear, como Lasso e ElasticNet, apresentaram diferenças significativas e médias de erro consideravelmente mais elevadas em relação à maioria dos demais modelos. Em contrapartida, modelos como *Random Forest*, *Ridge*, *SVR*, *Linear Regression*, *MLP* e *MLP Transfer Learning* apresentaram desempenhos estatisticamente equivalentes entre si, situando-se entre os melhores resultados obtidos.

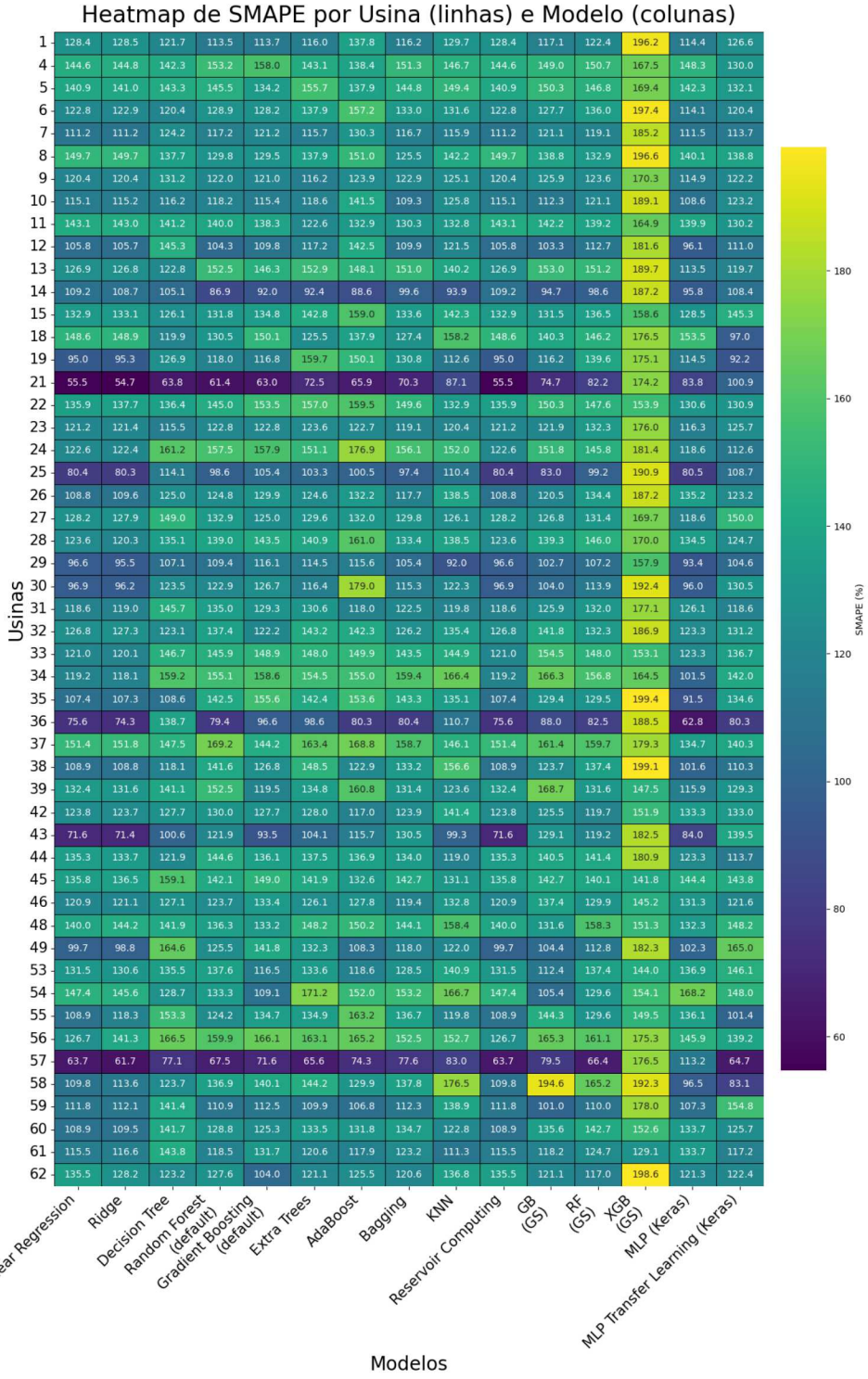


Figura 35: Heatmap de comparação por usina – Previsão diária.

Fonte: Autoria própria (2025).

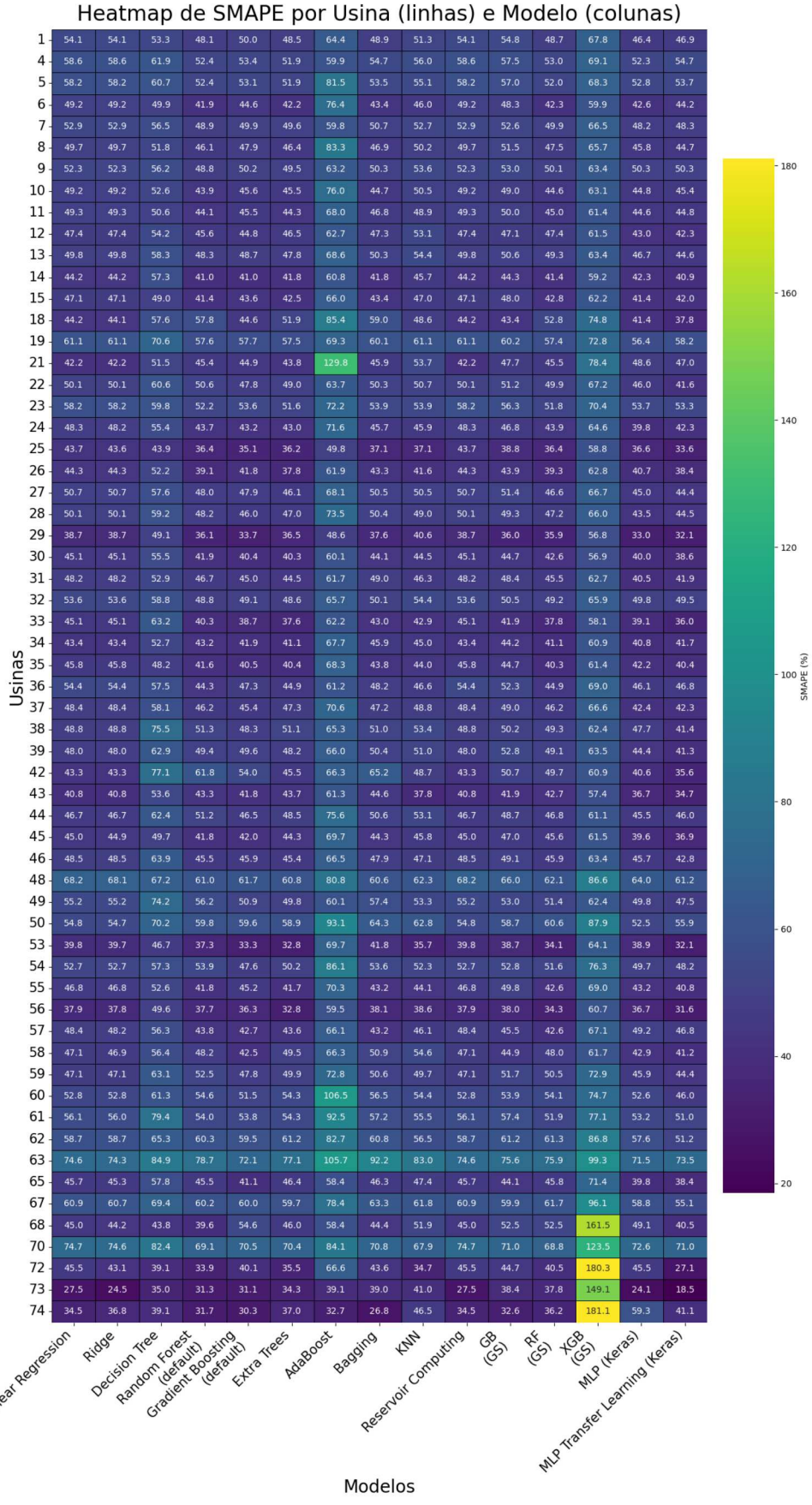


Figura 36: Heatmap de comparação por usina – Previsão intradiária (15 minutos).

Fonte: Autoria própria (2025).

A Figura 37 apresenta o *heatmap* das diferenças médias de SMAPE entre modelos, derivado das comparações múltiplas de Tukey. Cores azuladas representam diferenças negativas (modelos na linha apresentando menor erro médio que os da coluna), enquanto tons avermelhados indicam diferenças positivas (modelos com maior erro médio). A análise visual evidencia blocos de comportamento semelhante, nos quais modelos como *MLP*, *MLP Transfer Learning*, *Ridge* e *Random Forest* formam um agrupamento de alta performance, caracterizado por diferenças estatisticamente não significativas entre si. Por outro lado, observa-se que os modelos *Lasso* e *ElasticNet* se destacam por apresentarem médias de erro substancialmente superiores, indicando limitações no ajuste para este conjunto de dados.

De forma geral, os resultados confirmam a robustez dos modelos de aprendizado profundo e de regressão regularizada com penalização de segunda ordem (*Ridge*), além de reforçarem o potencial da estratégia de *transfer learning* aplicada ao modelo MLP, que se manteve consistentemente entre os melhores desempenhos sem apresentar diferenças significativas em relação às abordagens mais estáveis. Tal evidência corrobora a hipótese de que técnicas de aprendizado transferido podem contribuir para maior generalização dos modelos de previsão fotovoltaica, especialmente em cenários com variabilidade interplanta pronunciada.

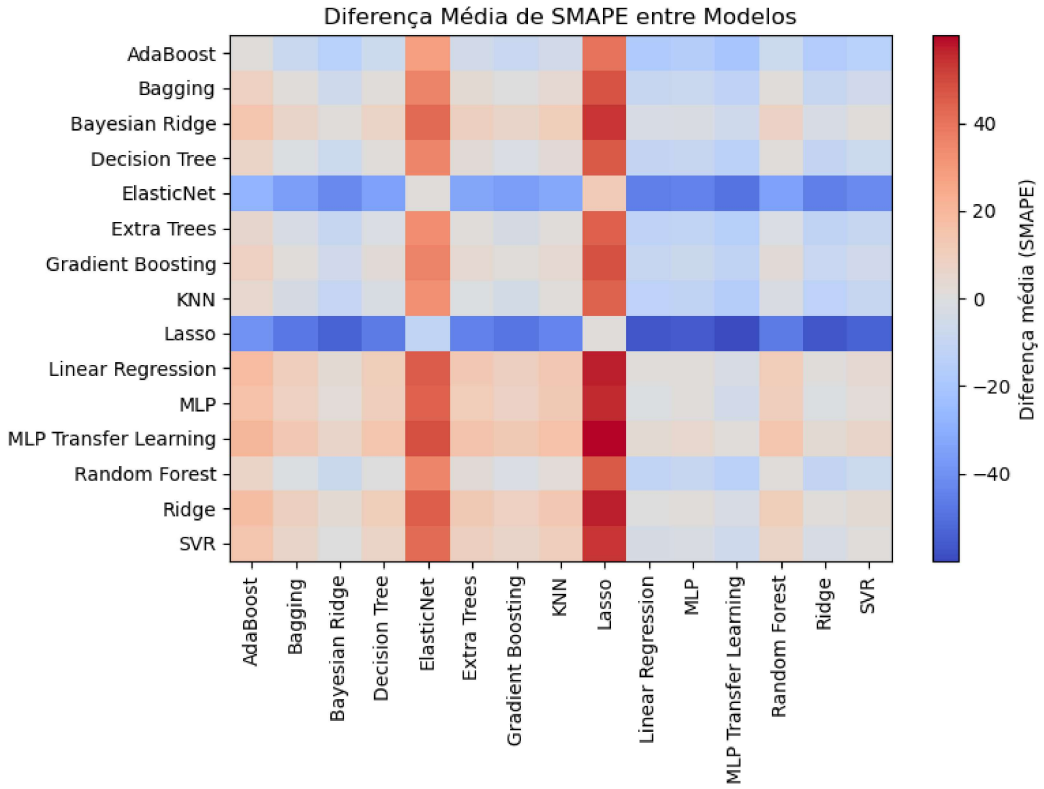


Figura 37: Diferença média de SMAPE entre modelos obtida pelo teste de Tukey HSD.

Fonte: Autoria própria (2025).

8 CONCLUSÃO

Esta dissertação apresentou o desenvolvimento e a disponibilização da base de dados **BR-PVGen**, constituindo um dos primeiros esforços nacionais de consolidação e padronização de registros de usinas fotovoltaicas conectadas à rede em operação real. O trabalho teve como objetivo principal a criação dessa base pública, estruturada e anonimizável, capaz de apoiar pesquisas em previsão, controle e otimização de sistemas solares, e, de forma complementar, a avaliação de metodologias de aprendizado de máquina aplicadas à previsão de geração de energia fotovoltaica.

A BR-PVGen reuniu séries temporais provenientes de 51 usinas fotovoltaicas distribuídas em diferentes regiões do Brasil, com medições diárias e intradiárias de potência dos inversores e variáveis meteorológicas correspondentes. O tratamento dos dados incluiu etapas de verificação de qualidade, interpolação, filtragem e normalização, assegurando consistência e reprodutibilidade. Essa base viabilizou não apenas o desenvolvimento dos modelos propostos nesta dissertação, mas também fornece um repositório de referência para futuras pesquisas na área de energia solar e aprendizado de máquina.

Com base nessas informações, foi realizada uma análise de desempenho operacional das usinas, identificando um *Performance Ratio* médio de 75,3%, com perdas associadas principalmente à temperatura (4,7%), ao *clipping* dos inversores (8,0%) e à indisponibilidade (6,6%). Esses resultados reforçam a importância de conjuntos de dados integrados para o diagnóstico e melhoria da performance de sistemas fotovoltaicos.

A segunda contribuição deste trabalho concentrou-se na aplicação de modelos de aprendizado de máquina para previsão de geração fotovoltaica, com ênfase em Redes Neurais Artificiais otimizadas por *Transfer Learning*. Os experimentos demonstraram que o uso de pré-treinamento entre usinas com históricos mais longos e aquelas com menor volume de dados resulta em ganhos consistentes de acurácia e eficiência computacional. Em horizontes diários, o SMAPE médio reduziu de 120,87% para 116,26%, enquanto em horizontes intradiários (15 minutos) houve melhora média de 8,45%, chegando a 11,06% em usinas com séries menores. Além disso, o *Transfer Learning* reduziu em aproximadamente 40% o tempo médio de treinamento, evidenciando sua aplicabilidade em contextos operacionais.

De forma geral, os resultados confirmam que a utilização de bases multivariadas e de estratégias de transferência de conhecimento permite explorar padrões de geração de forma mais generalizável e eficiente, abrindo caminho para novas abordagens no monitoramento e previsão da geração solar no contexto brasileiro.

CONTRIBUIÇÕES E TRABALHOS FUTUROS

As principais contribuições desta dissertação podem ser sintetizadas em três eixos:

- Desenvolvimento e disponibilização da base de dados **BR-PVGen**, inédita no contexto nacional, estruturada com dados meteorológicos e de geração elétrica de 51 usinas fotovoltaicas, com potencial para servir como *benchmark* em pesquisas futuras;
- Aplicação da BR-PVGen como estudo de caso para previsão de geração, demonstrando a eficiência do *Transfer Learning* em cenários de dados escassos e heterogêneos;
- Implementação e validação de um fluxo de modelagem preditiva reprodutível, que integra controle de regularização, taxa de aprendizado adaptativa e funções de custo robustas.

Como perspectivas de continuidade, destacam-se:

- Expansão da base **BR-PVGen** com novas usinas, períodos históricos e variáveis meteorológicas complementares, aprimorando a representatividade geográfica e climática;
- Investigação de arquiteturas híbridas e recorrentes, como **Long Short-Term Memory (LSTM)**, **Gated Recurrent Unit (GRU)** e **Reservoir Computing**, capazes de capturar dependências temporais de longo alcance e padrões não lineares complexos;
- Exploração de estratégias de **Aprendizado Federado (Federated Learning)**, permitindo o treinamento colaborativo entre diferentes usinas sem compartilhamento direto dos dados brutos, preservando privacidade e confidencialidade;
- Aplicação das técnicas de previsão e transferência de aprendizado em tarefas de diagnóstico de falhas e detecção de anomalias em tempo real.

Em síntese, esta dissertação contribui para o avanço da pesquisa nacional em previsão de geração fotovoltaica ao unir o desenvolvimento de uma base de dados pública, padronizada e de alta qualidade, à demonstração prática de metodologias modernas de aprendizado profundo. Espera-se que a BR-PVGen se torne um recurso de referência para a comunidade científica e para o setor energético, incentivando estudos colaborativos e o desenvolvimento de soluções inteligentes para a operação e integração da energia solar no sistema elétrico brasileiro.

REFERÊNCIAS

1 GABOITAOLELWE, J. et al. Machine learning based solar photovoltaic power forecasting: A review and comparison. *IEEE Access*, v. 11, p. 34639–34665, 2023.

2 BLOOMBERGNEF. Investimento global em energia limpa salta 172024. Acesso em: 2024-04-01. Disponível em: <https://www.bloomberg.com.br/blog/bloombergnef-investimento-global-em-energia-limpa-salta-17-e-atinge-us-18-trilhao-em-2023/>.

3 Empresa de Pesquisa Energética (EPE). *Balanço Energético Nacional 2024: Relatório Síntese - Ano Base 2023*. Rio de Janeiro: Ministério de Minas e Energia, 2023. Acesso em: 23 dez. 2024. Disponível em: <http://www.epe.gov.br>.

4 ALVES, K. S. T. R.; de Jesus, C. D.; de Aguiar, E. P. A new takagi–sugeno–kang model for time series forecasting. *Engineering Applications of Artificial Intelligence*, v. 133, p. 108155, 2024. ISSN 0952-1976. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0952197624003130>.

5 REN21. Global status report. 2021. Acesso em: 2024-04-01. Disponível em: <https://www.ren21.net/reports/global-status-report/>.

6 SPERATI, S. et al. The "weather intelligence for renewable energies" benchmarking exercise on short-term forecasting of wind and solar power generation. *Energies*, MDPI, v. 8, n. 9, p. 9594–9619, 2015. Disponível em: <https://doi.org/10.3390/en8099594>.

7 Agência Nacional de Energia Elétrica. *Minuta do Relatório de Gestão 2023: Anexo à Portaria n.º 6.891, de 2 de maio de 2024*. 2023. <https://www2.aneel.gov.br/cedoc/aprt20246891_2.pdf>. Acesso em: 28 set. 2025.

8 INC., P. S. *Pecan Street Dataport*. 2018. Base de dados com medições de consumo de energia e geração FV em nível residencial, alta resolução temporal, EUA. Disponível em: <https://www.pecanstreet.org/dataport/>.

9 (NREL), N. R. E. L. *Photovoltaic Data Acquisition (PVDAQ)*. 2024. Medições de geração, irradiância solar, temperatura de módulos e metadados de sistemas fotovoltaicos nos EUA. Disponível em: <https://www.nrel.gov/pv/data-tools>.

10 CENTRE, D. K. A. S. *DKASC Dataset*. 2024. Conjunto de dados de mais de 150 sistemas FV na Austrália, com medições de saída de inversores e variáveis meteorológicas, principalmente em intervalos de 5 minutos. Disponível em: <https://dkasolarcentre.com.au/>.

11 NIHAR, A. et al. Toward findable, accessible, interoperable and reusable (fair) photovoltaic system time series data. In: *2021 IEEE 48th Photovoltaic Specialists Conference (PVSC)*. [s.n.], 2021. p. 1701–1706. Disponível em: <https://ieeexplore.ieee.org/document/9518782>.

12 THEOCHARIDES, S. et al. Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Applied Energy*, v. 268, p. 115023, jun. 2020.

13 PRETTO, S. et al. A new probabilistic ensemble method for an enhanced day-ahead pv power forecast. *IEEE Journal of Photovoltaics*, v. 12, n. 2, p. 581–588, mar. 2022.

14 ALESSANDRINI, S. et al. An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, v. 157, p. 95–110, nov. 2015.

15 LI, Y.; SU, Y.; SHU, L. An armax model for forecasting the power output of a grid connected photovoltaic system. *Renewable Energy*, v. 66, p. 78–89, jun. 2014.

16 LONG, W. et al. Comparison of k-nearest neighbors, multiple linear regression, svm and ann for day-ahead pv power forecasting. *Energy*, v. 69, p. 435–446, 2014.

17 ZAMO, M. et al. A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. part ii: Probabilistic forecast of daily production. *Solar Energy*, v. 105, p. 804–816, 2014.

18 HUANG, J.; PERRY, M. A semi-empirical approach using gradient boosting and k-nearest neighbors regression for gefcom2014 probabilistic solar power forecasting. *International Journal of Forecasting*, v. 32, n. 3, p. 1081–1086, 2016.

19 LIN, K.-P.; PAI, P.-F. Solar power output forecasting using evolutionary seasonal decomposition least-square support vector regression. *Journal of Cleaner Production*, v. 134, p. 456–462, out. 2016.

20 ABDELLATIF, A. et al. Forecasting photovoltaic power generation with a stacking ensemble model. *Sustainability*, v. 14, n. 17, p. 11083, set. 2022.

21 MELLIT, A.; PAVAN, A. M.; LUGHI, V. Short-term forecasting of power production in a large-scale photovoltaic plant. *Solar Energy*, v. 105, p. 401–413, jul. 2014.

22 LEVA, S. et al. Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. *Mathematics and Computers in Simulation*, v. 131, p. 88–100, jan. 2017.

23 GAO, M. et al. Day-ahead power forecasting in a large-scale photovoltaic plant based on weather classification using lstm. *Energy*, v. 187, p. 115838, nov. 2019.

24 KHAN, W.; WALKER, S.; ZEILER, W. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach. *Energy*, v. 240, p. 122812, fev. 2022.

25 NESPOLI, A. et al. A selective ensemble approach for accuracy improvement and computational load reduction in ann-based pv power forecasting. *IEEE Access*, v. 10, p. 32900–32911, 2022.

26 RAMSAMI, P.; OREE, V. A hybrid method for forecasting the energy output of photovoltaic systems. *Energy Conversion and Management*, v. 95, p. 406–413, maio 2015.

27 MOHAMMED, A.; YAQUB, W.; AUNG, Z. Probabilistic forecasting of solar power: An ensemble learning approach. In: *Proceedings of the International Conference on Intelligent Decision Technologies*. [S.l.]: Springer, 2015. p. 449–458.

28 WANG, C. et al. A hybrid deep learning model for photovoltaic power forecasting based on convolutional and recurrent neural networks. *Applied Energy*, v. 344, p. 121278, 2023.

29 MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v. 5, n. 4, p. 115–133, 1943.

30 ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, n. 6, p. 386–408, 1958.

31 RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, 1986.

32 HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. 2. ed. Upper Saddle River, NJ: Prentice Hall, 1999. ISBN 978-0132733502.

33 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA: MIT Press, 2016. ISBN 978-0262035613. Disponível em: <http://www.deeplearningbook.org>.

34 HUBER, P. J. Robust estimation of a location parameter. *Annals of Mathematical Statistics*, Institute of Mathematical Statistics, v. 35, n. 1, p. 73–101, 1964. Disponível em: <https://doi.org/10.1214/aoms/1177703732>.

35 SRIVASTAVA, N. et al. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, v. 15, n. 56, p. 1929–1958, 2014. Disponível em: <http://jmlr.org/papers/v15/srivastava14a.html>.

36 PRECHELT, L. Early stopping — but when? In: ORR, G. B.; MÜLLER, K.-R. (Ed.). *Neural Networks: Tricks of the Trade*. Berlin, Heidelberg: Springer, 1998, (Lecture Notes in Computer Science, v. 1524). p. 55–69.

37 OLIVEIRA, C. G. M. C. . D. d. J. . A. M. F. . A. N. P. . F. N. . P. d. A. Camille V. M. B. de. Previsão de geração de energia em ufvs utilizando rna e aprendizado por transferência. In: SOCIEDADE BRASILEIRA DE AUTOMáTICA (SBA). *Congresso Brasileiro de Automática (CBA)*. [S.l.], 2024. Anais do CBA 2024.

38 MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2012.

39 FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, v. 29, n. 5, p. 1189–1232, 2001.

40 TIPPING, M. E. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, v. 1, n. Jun, p. 211–244, 2001.

41 SHAWE-TAYLOR, J.; CRISTIANINI, N. *Kernel methods for pattern analysis*. [S.l.]: Cambridge University Press, 2004.

42 Solargis. *Solar Data Quality: Best Practices for Quality Control in Solar Data*. 2023. Disponível em: <https://kb.solargis.com/docs/data-quality-control-analysis>.

43 SCHNIERER, B. *The pros and cons of 1-minute, 15-minute, and 60-minute solar data*. 2023. Solargis blog. Disponível em: <https://solargis.com/resources/blog/best-practices/the-pros-and-cons-of-1-minute-15-minute-and-60-minute-solar-data>.

44 HASSANI, H.; KALANTARI, M.; GHODSI, Z. Evaluating the performance of multiple imputation methods for handling missing values in time series data: A study focused on east africa, soil-carbonate-stable isotope data. *Stats*, v. 2, n. 4, p. 457–467, 2019. ISSN 2571-905X. Disponível em: <https://www.mdpi.com/2571-905X/2/4/32>.

45 GHOSH, S.; ROY, J. N.; CHAKRABORTY, C. A model to determine soiling, shading and thermal losses from pv yield data. *Clean Energy*, Oxford University Press, v. 6, n. 4, p. 372–391, 2022. Advance access publication 29 April 2022. Disponível em: <https://doi.org/10.1093/ce/zkac014>.