



UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

REGINA MARIA DE LIMA NETA

**MULTIPREVISÃO DE GERAÇÃO FOTOVOLTAICA BASEADA EM
CLUSTERIZAÇÃO HIERÁRQUICA E TEMPORAL FUSION TRANSFORMER**

Recife

2025

REGINA MARIA DE LIMA NETA

**MULTIPREVISÃO DE GERAÇÃO FOTOVOLTAICA BASEADA EM
CLUSTERIZAÇÃO HIERÁRQUICA E TEMPORAL FUSION TRANSFORMER**

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do grau de Doutor em Engenharia Elétrica.

Área de concentração: Processamento de Energia.

Orientador: Prof. Dr. Ronaldo Ribeiro Barbosa de Aquino.

Coorientadora: Prof.^a Dra. Milde Maria da Silva Lira.

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Lima Neta, Regina Maria de.

Multiprevisão de geração fotovoltaica baseada em clusterização hierárquica e temporal fusion transformer / Regina Maria de Lima Neta. - Recife, 2025.

102f.: il.

Tese (Doutorado) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Programa de Pós-Graduação em Engenharia Elétrica, 2025.

Orientação: Ronaldo Ribeiro Barbosa de Aquino.

Coorientação: Milde Maria da Silva Lira.

Inclui Referências.

1. Multiprevisão; 2. Clusterização Hierárquica; 3. Temporal Fusion Transformer; 4. Geração Fotovoltaica; 5. Validação com Janela Deslizante. I. Aquino, Ronaldo Ribeiro Barbosa de. II. Lira, Milde Maria da Silva. III. Título.

UFPE-Biblioteca Central

REGINA MARIA DE LIMA NETA

**MULTIPREVISÃO DE GERAÇÃO FOTOVOLTAICA BASEADA EM
CLUSTERIZAÇÃO HIERÁRQUICA E TEMPORAL FUSION TRANSFORMER**

Tese apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Engenharia Elétrica, na área de concentração em Processamento de Energia.

Aprovada em: 28/08/2025.

BANCA EXAMINADORA

Prof. Dr. Pedro André Carvalho Rosas
(Examinador Interno)
Universidade Federal de Pernambuco

Profª. Dra. Milde Maria da Silva Lira
(Coorientadora e Examinadora Externa)
Universidade Federal de Pernambuco

Prof. Dr. Manoel Afonso de Carvalho Júnior
(Examinador Externo)
Universidade Federal de Pernambuco

Profª. Dra. Aida Araújo Ferreira
(Examinadora Externa)
Instituto Federal de Educação, Ciência e Tecnologia de Pernambuco

Prof. Dr. Jeydson Lopes da Silva
(Examinador Externo)
Universidade Federal de Pernambuco

AGRADECIMENTOS

A Deus toda honra e toda glória e toda a minha gratidão. Obrigado Senhor, pelo dom precioso da vida, pelo seu amor infinito e por realizar todos os meus sonhos.

À minha família, principalmente minha mãe, Madalena Lima, cujas orações e apoio incondicional tornaram possível a realização desse sonho. Mãe, te amo para sempre, sei que ninguém torce tanto por mim. Às minhas irmãs, Ruth e Rebeca, que sempre me encorajaram a seguir em frente. À minha querida sobrinha Ayla Clarice que chegou para encantar nossa vida e tornou tudo mais feliz.

A todos os meus amigos e familiares que me encorajaram com palavras de conforto e otimismo e intercederam a Deus por mim.

Ao meu orientador, Prof. Ronaldo Ribeiro, pela confiança, apoio, conselhos e, principalmente, por me conceder a oportunidade de realizar este trabalho.

À minha coorientadora, Profa. Milde Lira, pelo apoio, amizade, conversas, compreensão e conselhos acadêmicos.

Ao meu companheiro de Doutorado, Murilo Araújo, pelo apoio, incentivo e ajuda com o trabalho.

Aos meus colegas de trabalho no IFPE Campus Pesqueira que me motivaram e incentivaram sempre com palavras acolhedoras. Eu amo fazer parte da família IFPE Campus Pesqueira.

Aos meus colegas do LDSP, principalmente André Vitor, pela convivência diária, amizade e trocas de experiências que tornaram esse fardo mais fácil de carregar.

Aos docentes e servidores do programa de Pós-Graduação PPGEE, eu agradeço a atenção e o profissionalismo.

Aos professores da banca examinadora que muito me ensinaram com suas valiosas experiências e contribuições.

A todos que compartilharam dessa experiência enriquecedora e de plena superação que é o processo para obtenção do título de Doutorado, que muitas vezes pareceu ser uma tarefa interminável e enigmática e que somente se tornou realidade graças às muitas pessoas que participaram, direta ou indiretamente. A estes quero prestar o meu sincero agradecimento. Obrigada!

RESUMO

A alta variabilidade da geração solar fotovoltaica (FV) dificulta a previsão de curto prazo, levando o Operador Nacional do Sistema Elétrico (ONS) a manter reservas operativas elevadas e, consequentemente, custos sistêmicos maiores. Esta tese propõe uma abordagem de previsão multissérie que integra Clusterização Hierárquica com distância DTW e a arquitetura Temporal Fusion Transformer (TFT), utilizando exclusivamente dados públicos de geração (ONS) e meteorologia (NASA POWER). A metodologia inclui: imputação de dados faltantes guiada pela curva de radiação solar, clusterização de dez usinas FV para identificar grupos com padrões de geração similares e desenvolvimento de quatorze modelos TFT, contemplando configurações univariadas e multipreditivas. A avaliação, baseada em validação por janela deslizante, dez inicializações independentes e métricas MAE, RMSE e nMAE na escala original (MW), mostrou que os melhores modelos univariados atingem nMAE em torno de 10%, combinando variáveis meteorológicas, atributos derivados da série de geração e codificações sazonais. Nos modelos multipreditivos, a previsão simultânea de duas ou três usinas manteve nMAEs muito próximos aos univariados, indicando que a acurácia relativa por usina praticamente não se deteriora ao adotar uma modelagem conjunta, ao mesmo tempo em que se reduz o número de modelos a serem treinados e operados. Comparativos com redes MLP, LSTM e com o método de Persistência evidenciaram a superioridade consistente do TFT. Os resultados estabelecem uma base reproduzível e escalável para previsão FV em múltiplas usinas, com potencial aplicação na melhoria da eficiência e da gestão do sistema elétrico brasileiro.

Palavras-chave: Multiprevisão. Clusterização Hierárquica. Temporal Fusion Transformer. Geração Fotovoltaica. Validação com Janela Deslizante.

ABSTRACT

The high variability of solar photovoltaic (PV) generation poses a significant challenge for short-term forecasting, compelling the National Electric System Operator (ONS) to maintain high operating reserves and, consequently, incurring higher systemic costs. This thesis proposes a multi-series forecasting approach that integrates Hierarchical Clustering with Dynamic Time Warping (DTW) distance and the Temporal Fusion Transformer (TFT) architecture, relying exclusively on public generation (ONS) and meteorological (NASA POWER) data. The methodology includes missing data imputation guided by the solar radiation curve; clustering of ten PV plants to identify groups with similar generation patterns; and the development of fourteen TFT models, encompassing both univariate and multi-target configurations. The assessment, based on walk-forward validation, ten independent initializations, and MAE, RMSE, and nMAE metrics calculated on the original scale (MW), showed that the best univariate models achieve an nMAE of approximately 10% by combining meteorological variables, generation-derived features, and seasonal encodings. In multi-target models, the simultaneous forecasting of two or three plants maintained nMAE levels very close to those of univariate models, indicating that the relative accuracy per plant suffers virtually no degradation under joint modeling, while reducing the number of models to be trained and operated. Comparisons with MLP and LSTM networks, as well as the Persistence method, highlighted the consistent superiority of the TFT. The results establish a reproducible and scalable framework for multi-plant PV forecasting, with potential applications in enhancing the efficiency and management of the Brazilian power system.

Keywords: Multiseries Forecasting. Hierarchical Clustering. Temporal Fusion Transformer. PV Power Generation. Sliding Window Validation.

LISTA DE FIGURAS

FIGURA 1 – EXPANSÃO DA OFERTA DE GERAÇÃO DE ENERGIA ELÉTRICA EM 2024	16
FIGURA 2 - DESAFIOS ASSOCIADOS À PREVISÃO DE GERAÇÃO FV	18
FIGURA 3 – HIERARQUIA CONCEITUAL ENTRE IA, ML E DL	29
FIGURA 4 - ARQUITETURA DE UMA REDE MLP	30
FIGURA 5 – FLUXOGRAMA DO TREINAMENTO DA REDE MLP	31
FIGURA 6 - ESQUEMA DE FUNCIONAMENTO DE UMA RNR	33
FIGURA 7 - DIAGRAMA ESQUEMÁTICO DA REDE LSTM	36
FIGURA 8 - DIAGRAMA ESQUEMÁTICO DA ARQUITETURA <i>TRANSFORMER</i>	37
FIGURA 9 - ARQUITETURA DA REDE TFT	41
FIGURA 10 – ESTRUTURA DA REDE GRN	45
FIGURA 11 - ALGORITMO DE CLUSTERIZAÇÃO <i>K-MEANS</i>	48
FIGURA 12 - DIAGRAMA ESQUEMÁTICO DA CLUSTERIZAÇÃO	49
FIGURA 13 - SÉRIES TEMPORAIS COMPARADAS	50
FIGURA 14 - DIFERENÇAS COM DISTÂNCIA EUCLIDIANA	51
FIGURA 15 - DIFERENÇAS COM DISTÂNCIAS DTW	51
FIGURA 16 – EXEMPLO DE DENDROGRAMA NA CLUSTERIZAÇÃO HIERÁRQUICA	53
FIGURA 17 - DISTRIBUIÇÃO DE DADOS EM QUARTIS	55
FIGURA 18 - JANELA DESLIZANTE	56
FIGURA 19 - METODOLOGIA DA PESQUISA	58
FIGURA 20 – LOCALIZAÇÃO DAS USINAS FV SELECIONADAS NO MAPA DO BRASIL	59
FIGURA 21 - FLUXOGRAMA DO MÉTODO PROPOSTO PARA IMPUTAÇÃO DE DADOS	62
FIGURA 22 - ANÁLISE DE CORRELAÇÃO DAS VARIÁVEIS COM A GERAÇÃO FV	64
FIGURA 23 - COMPARATIVO DA TRANSFORMAÇÃO LOG NA GERAÇÃO FV	66
FIGURA 24 - DIAGRAMA ESQUEMÁTICO DA MODELAGEM TFT	69
FIGURA 25 - ESTRUTURA DE DEPENDÊNCIA TEMPORAL DA GERAÇÃO FV	73
FIGURA 26 - VALIDAÇÃO DA IMPUTAÇÃO DE DADOS NA GERAÇÃO FV	77
FIGURA 27 - DENDROGRAMA RESULTANTE DA CLUSTERIZAÇÃO HIERÁRQUICA	78
FIGURA 28 - INTERPRETABILIDADE DAS VARIÁVEIS DO MODELO M7	83
FIGURA 29 - COMPARATIVO REAL X PREVISTO PARA O MODELO M8	84
FIGURA 30 - INTERPRETABILIDADE DAS VARIÁVEIS DO MODELO M12	89
FIGURA 31 – COMPARATIVO 1 REAL X PREVISTO PARA O MODELO M12	90
FIGURA 32 - COMPARATIVO 2 REAL X PREVISTO PARA O MODELO M12	91

LISTA DE TABELAS

TABELA 1 - ÍNDICES DE CORRELAÇÃO ENTRE METEOROLOGIA E GERAÇÃO FV	17
TABELA 2 - PROBLEMAS IDENTIFICADOS NA LITERATURA E SOLUÇÕES PROPOSTAS	19
TABELA 3 - ESTUDOS CORRELATOS SOBRE TFT NA PREVISÃO FV	27
TABELA 4 – INFORMAÇÕES TÉCNICAS DAS USINAS FV	60
TABELA 5 - INTERAÇÃO DE ATRIBUTOS	64
TABELA 6 - ATRIBUTOS SELECIONADOS PARA A MODELAGEM TFT	65
TABELA 7 - ATRIBUTOS SELECIONADOS PARA O MODELO TFT	68
TABELA 8 - CONFIGURAÇÃO DE ENTRADA DOS MODELOS M1 A M8	69
TABELA 9 - CONFIGURAÇÃO DE ENTRADA DOS MODELOS M9 A M14	71
TABELA 10 - HIPERPARÂMETROS DOS MODELOS TFT.....	73
TABELA 11 - DADOS FALTANTES EM CADA USINA FV	76
TABELA 12 - MATRIZ DE CORRELAÇÃO DE PEARSON DAS USINAS FV	79
TABELA 13 – COMPARATIVO DE DESEMPENHO - MODELOS M1 A M8.....	81
TABELA 14 - RESULTADOS DE MAE POR <i>RUN</i> PARA OS MODELOS M1 A M8.....	85
TABELA 15 – TESTE DE SIGNIFICÂNCIA ESTATÍSTICA PAREADA DOS MODELOS M1 A M8	86
TABELA 16 - DESCRIÇÃO DAS USINAS DOS MODELOS M9 A M14	87
TABELA 17 - COMPARATIVO DE DESEMPENHO - MODELOS M9 A M14	87
TABELA 18 - RESULTADOS DE MAE POR <i>RUN</i> PARA OS MODELOS M9 A M14.....	92
TABELA 19 – TESTE DE SIGNIFICÂNCIA ESTATÍSTICA PAREADA DOS MODELOS M9 A M14	92
TABELA 20 - HIPERPARÂMETROS DOS MODELOS MLP E LSTM	93
TABELA 21 – COMPARATIVO DO DESEMPENHO DAS TOPOLOGIAS TFT, MLP E LSTM.....	94

LISTA DE ABREVIATURAS E SIGLAS

ACF	<i>Autocorrelation Function</i> – Função de Autocorrelação
Add & Norm	Camada de Normalização
ANN	<i>Artificial Neural Network</i> – Rede Neural Artificial
ARIMA	<i>AutoRegressive Integrated Moving Average</i> – Modelo Autorregressivo Integrado de Médias Móveis
ARIMAX	<i>AutoRegressive Integrated Moving Average with Exogenous Variables</i> – Modelo ARIMA com Variáveis Exógenas
ARMA	<i>AutoRegressive Moving Average</i> – Modelo Autorregressivo de Médias Móveis
CNN	<i>Convolutional Neural Network</i> – Rede Neural Convolucional
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i> – Agrupamento Espacial Baseado em Densidade
DHI	<i>Diffuse Horizontal Irradiance</i> – Irradiância Horizontal Difusa
DL	<i>Deep Learning</i> – Aprendizado Profundo
DNI	<i>Direct Normal Irradiance</i> – Irradiância Direta Normal
dp-MAE	Desvio-padrão do erro médio absoluto
dp-RMSE	Desvio-padrão da raiz do erro médio quadrático
DTW	<i>Dynamic Time Warping</i> – Deformação Dinâmica no Tempo
FV	Fotovoltaico (a)
GLU	<i>Gated Linear Unit</i> - Unidade Linear <i>Gated</i>
GRN	<i>Gated Residual Network</i> - Rede Residual <i>Gated</i>
GRU	<i>Gated Residual Unit</i> - Unidade Recorrente <i>Gated</i>
GW	Gigawatt
INMET	Instituto Nacional de Meteorologia
IQR	Intervalo Interquartil
KNN	<i>k-Nearest Neighbors</i> – k-Vizinhos Mais Próximos
LANN	<i>Local Average of Nearest Neighbors</i> - Média Local dos Vizinhos Mais Próximos
LayerNorm	Camada de normalização <i>Standard</i>
LMD	Dados Medidos Localmente
LSTM	<i>Long Short-Term Memory</i> – Memória de Longo e Curto Prazo
MAE	<i>Mean Absolute Error</i> – Erro Absoluto Médio
MAPE	<i>Mean Absolute Percentage Error</i> – Erro Percentual Absoluto Médio
ML	<i>Machine Learning</i> – Aprendizagem de Máquina
MLP	<i>Multilayer Perceptron</i> – Rede Multicamadas
MCI	Multi-cabeça Interpretável
MW	Megawatt

MWMED	Média Horária em Megawatt
MMGD	Micro e Minigeração Distribuída
NASA	<i>National Aeronautics and Space Administration</i>
NASA POWER	<i>Prediction of Worldwide Energy Resources</i> – Banco de Dados Meteorológicos da NASA
nMAE	Erro médio absoluto normalizado
nRMSE	Raiz do erro médio quadrático
NWP	<i>Numerical Weather Prediction</i> – Previsão Numérica do Tempo
PGMG	Programação Genética Multi-gene
ONS	Operador Nacional do Sistema Elétrico
PV	<i>Photovoltaic</i> – Fotovoltaico
R ²	Coeficiente de Determinação
RELU	Unidade Linear Retificada
RMSE	<i>Root Mean Squared Error</i> – Raiz do Erro Quadrático Médio
RNR	Rede Neural Recorrente
SEP	Sistema Elétrico de Potência
SIN	Sistema Interligado Nacional
SVR	<i>Support Vector Regression</i>
tanh	Tangente Hiperbólico
TFT	<i>Temporal Fusion Transformer</i>

LISTA DE SÍMBOLOS

X	Vetor de entrada da rede neural
Y	Valor real da saída esperada
Y'	Valor previsto pela rede neural
W_k	Pesos da rede neural
b	Viés (<i>bias</i>) da rede neural
Y_k	Saídas dos neurônios
N	Tamanho da amostra
t	Tempo
x_t	Entrada no instante t
y_t	Saída no instante t
M	Dimensão do vetor de entrada
$h_{(t)}$	Estado oculto no instante t
$h_{(t-1)}$	Estado oculto no instante anterior
a	Função de ativação
W	Matriz de Peso
U	Matriz de Peso
$c_{(t)}$	Estado da célula na LSTM no instante t
σ	Função Sigmoid
$f_{(t)}$	Vetor de ativação do <i>forget gate</i> na LSTM
$i_{(t)}$	Vetor de ativação do <i>input gate</i> na LSTM
$o_{(t)}$	Vetor de ativação do <i>output gate</i> na LSTM
$g_{(t)}$	Célula candidata na LSTM
Q	Vetor de consultas (<i>queries</i>) no mecanismo de atenção
K	Vetor de chaves (<i>keys</i>) no mecanismo de atenção
V	Vetor de valores (<i>values</i>) no mecanismo de atenção
d_k	Dimensão das chaves no mecanismo de atenção
$Atenção(Q, K, V)$	Peso de atenção atribuído a cada elemento
$softmax()$	Função <i>softmax</i>
$tanh()$	Função tangente hiperbólica
$RELU()$	Função de ativação Unidade Linear Retificada
$ELU()$	Função de ativação Linear Exponencial
z_i	Vetor
i	Elemento de um vetor
j	Número de classes
τ_{max}	Horizonte de previsão
d_m	Dimensão da Matriz M
p	Número de elementos da série temporal M
q	Número de elementos da série temporal N
$\xi_t^{(j)}$	Vetor de características processadas para a variável j
$v_t^{(j)}$	Peso de seleção de variáveis
$\tilde{\xi}_t$	Vetor final de características processadas
m	Número total de variáveis do modelo

a_p	Entrada primária
c_x	Vetor de Contexto
ω	Índice de compartilhamento de pesos
η_1	Camada intermediária da GRN
η_2	Camada intermediária da GRN
$y_i(q, t, \tau)$	Q-ésimo percentil da variável-alvo
F_q	Função de previsão
$y_i(t - k: t)$	Histórico temporal da variável-alvo
$Z_i(t - k: t)$	Covariáveis passadas
$X_i(t: t + \tau)$	Covariáveis conhecidas no futuro
S_i	Covariáveis estáticas
\tilde{H}	Média das saídas de atenção
h'	Índice de cada cabeça
m_H	Número total de cabeças
f	Função objetivo
k	Número de clusters
c_i	Centroide de cada cluster
$C(i, j)$	Custo acumulado DTW entre pontos i e j
v_{ij}	Distância local entre os pontos i e j
x_{norm}	Variável normalizada
x_{min}	Valor mínimo da variável x
$x_{máx}$	Valor máximo da variável x
μ	Média da variável x
$\bar{\sigma}$	Desvio padrão da variável x
med	Mediana da variável x
IQR	Índice interquartil
$Q1$	Primeiro quartil
$Q3$	Terceiro quartil
$^{\circ}\text{C}$	Grau Celsius
W/m^2	Watt por metro quadrado
h	Hora
$h - 1$	Defasagem de 1 hora
$h - 2$	Defasagem de 2 horas
$h - 3$	Defasagem de 3 horas
$h - 11$	Defasagem de 11 horas
$G(h)$	Geração FV na hora h
$G(h - 1)$	Geração FV na hora anterior
G_{med_h}	Média da Geração na última hora
$G_{máx_h}$	Valor Máximo da Geração na última hora
G_{min_h}	Valor Mínimo da Geração na última hora
$hora_sen$	Hora representada na função seno
$hora_cos$	Hora representada na função cosseno
mes_sen	Mês representado na função seno
mes_cos	Mês representado na função cosseno
Mes	Mês do ano
$Estacao$	Estação do ano
N	Tamanho do conjunto de teste

<i>CAP</i>	Capacidade instalada da usina
<i>hora_sen</i> \times <i>G(h)</i>	Interação de G (h) com hora_sen
<i>hora_cos</i> \times <i>G(h)</i>	Interação de G (h) com hora_cos
<i>hora_sen</i> \times <i>G(h - 1)</i>	Interação de G (h - 1) com hora_sen
<i>hora_cos</i> \times <i>G(h - 1)</i>	Interação de G (h - 1) com hora_cos
<i>hora_sen</i> \times <i>G(h - 2)</i>	Interação de G (h - 2) com hora_sen
<i>hora_cos</i> \times <i>G(h - 2)</i>	Interação de G (h - 2) com hora_cos
<i>hora_sen</i> \times <i>G(h - 3)</i>	Interação de G (h - 3) com hora_sen
<i>hora_cos</i> \times <i>G(h - 3)</i>	Interação de G (h - 3) com hora_cos
<i>hora_sen</i> \times <i>G(h - 11)</i>	Interação de G (h - 11) com hora_sen
<i>hora_cos</i> \times <i>G(h - 11)</i>	Interação de G (h - 11) com hora_cos
<i>Rad</i>	Radiação solar
<i>Temp</i>	Temperatura ambiente
<i>hora_sen</i> \times <i>Rad</i>	Interação da radiação solar com hora_sen
<i>hora_cos</i> \times <i>Rad</i>	Interação de radiação solar com hora_cos
<i>hora_sen</i> \times <i>Temp</i>	Interação da temperatura ambiente com hora_sen
<i>hora_cos</i> \times <i>Temp</i>	Interação de temperatura ambiente com hora_cos
<i>Rad</i> \times <i>Temp</i>	Interação da radiação solar com a temperatura ambiente
<i>ID</i>	Covariável categórica estática de identificação de uma usina FV
<i>Usina</i>	Covariável categórica estática que distingue cada uma das usinas dentro do mesmo conjunto de entrada
<i>time_idx</i>	Índice temporal do TFT
<i>relative_time_idx</i>	Índice temporal relativo do TFT

SUMÁRIO

1	INTRODUÇÃO	16
1.1	Contextualização e Descrição do Problema	18
1.2	Objetivo-Geral	20
1.3	Organização do Trabalho	21
2	ESTADO DA ARTE	22
3	FUNDAMENTAÇÃO TEÓRICA	29
3.1	Inteligência Artificial	29
3.1.1	Redes Neurais MLP	30
3.1.2	Redes LSTM	32
3.2	Arquitetura <i>Transformer</i>	36
3.3	<i>Temporal Fusion Transformer</i>	39
3.3.1	Arquitetura TFT	40
3.3.2	Rede de Seleção de Variáveis	43
3.3.3	Redes GRNs	44
3.3.4	Função Preditiva	46
3.3.5	Interpretabilidade do modelo TFT	47
3.4	Clusterização	48
3.4.1	Clusterização de Séries Temporais	49
3.5	Conceitos Adicionais	53
3.5.1	Normalização dos Dados	53
3.5.2	Validação com Janela Deslizante	56
4	METODOLOGIA	58
4.1	Base de Dados	59
4.1.1	Dados de Geração FV	59
4.1.2	Dados Meteorológicos	60
4.1.3	Dados Temporais e Sazonalidade	61
4.2	Pré-processamento dos Dados	61
4.2.1	Eliminação de Dados Noturnos	61
4.2.2	Imputação de Dados Faltantes	62

4.2.3	Seleção de Atributos	63
4.2.4	Transformação Log.....	65
4.2.5	Normalização dos Dados	66
4.2.6	Clusterização Hierárquica	67
4.3	Algoritmo TFT	68
4.4	Treinamento e Definição de Hiperparâmetros	72
4.5	Avaliação	74
5	RESULTADOS E DISCUSSÃO.....	76
5.1	Imputação de Dados Faltantes.....	76
5.2	Clusterização Hierárquica da Base de Dados de Geração FV	78
5.3	Análise de Desempenho dos Modelos M1 a M8.....	80
5.3.1	Desempenho médio e variabilidade	80
5.3.2	Testes de significância estatística	84
5.4	Análise de Desempenho dos Modelos Multipreditivos M9 a M14	86
5.4.1	Desempenho médio e variabilidade	86
5.4.2	Testes de significância estatística	91
5.5	Comparativo com outras Topologias	93
6	CONCLUSÃO.....	96
6.1	Sugestão para trabalhos futuros.....	98
	REFERÊNCIAS	99

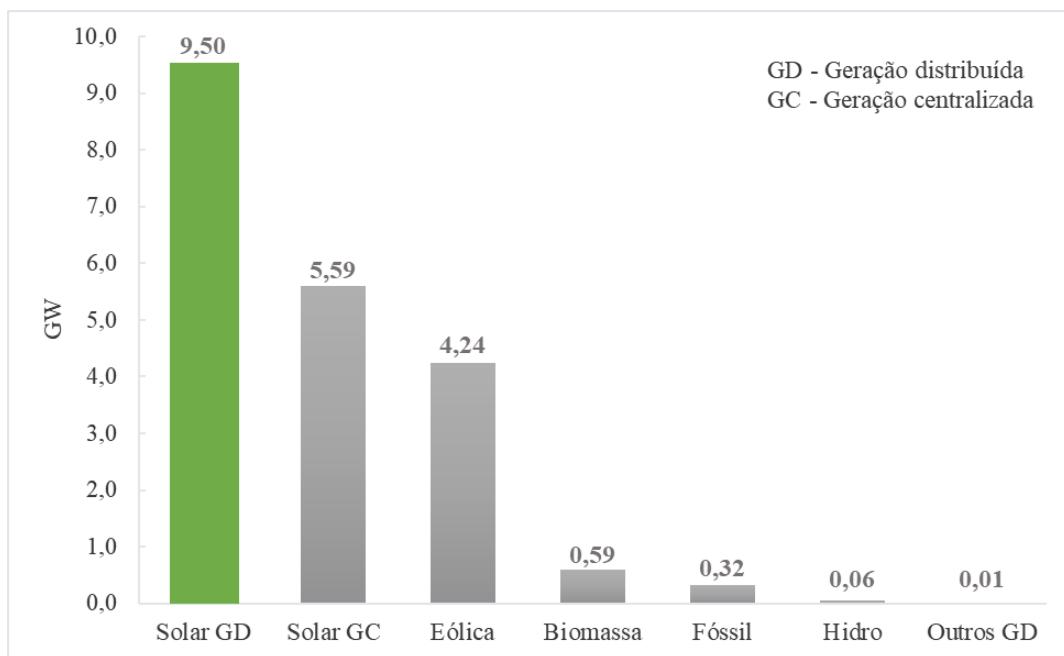
1 INTRODUÇÃO

No Brasil, a energia solar fotovoltaica (FV) tornou-se a fonte energética renovável com maior crescimento anual, pois sua capacidade instalada gerada aumentou de 49 MW em 2015 para 53.113 MW no ano de 2024 (*International Renewable Energy Agency*, 2025). Esse fato se explica pelo decréscimo dos preços dos insumos verificados nos últimos anos devido à alta oferta no mercado, à inovação tecnológica tendo em vista que os fabricantes garantem muitos anos de funcionamento em seus projetos, ao vasto potencial técnico existente e a minimização da emissão de gases de efeito estufa (Empresa de Pesquisa Energética, 2020).

Devido à sua localização geográfica, o Brasil recebe elevados índices de radiação solar, que são relativamente uniformes em todo o território nacional, o que permite o desenvolvimento de projetos solares viáveis em diferentes regiões. Assim, dada a redução de custos considerada, a fonte FV se apresenta como alternativa competitiva no fornecimento de energia. Além disso, a versatilidade e modularidade da tecnologia FV permitem o desenvolvimento de projetos em diferentes escalas, tanto centralizados quanto distribuídos.

A Empresa de Pesquisa Energética (EPE, 2025) aponta que a Micro e Minigeração Distribuída (MMGD) é a principal responsável pela expansão da capacidade instalada de energia elétrica no Brasil. Pelo quarto ano consecutivo, em 2024, a fonte solar distribuída superou as demais fontes em termos de expansão da oferta de geração, como pode ser verificado na Figura 1.

Figura 1 – Expansão da Oferta de Geração de Energia Elétrica em 2024



Fonte: Adaptado de EPE (2025).

Considerando o cenário energético atual e a crescente participação da geração FV na matriz elétrica brasileira, projeta-se a continuidade do aumento dos investimentos nessa tecnologia. Esse avanço tem sido impulsionado pela queda dos custos dos módulos no mercado internacional e pela maior conscientização sobre os benefícios ambientais e econômicos das fontes renováveis. Nesse contexto, emerge o desafio de desenvolver um modelo preditivo robusto e preciso para a geração FV. Tal modelo é crucial para reduzir a necessidade de reserva operativa do sistema elétrico, acomodar as especificidades da geração distribuída e otimizar a operação do setor. Além disso, apoia decisões de despacho e gestão de risco, sobretudo frente às flutuações de preços no mercado de energia (Lee; Kim, 2019).

O equilíbrio entre a produção e demanda de energia deve existir para que o Sistema Elétrico de Potência (SEP) opere com estabilidade, contudo, ao adicionar as fontes de energia alternativas, cuja geração varia ao longo das horas e de acordo com as condições meteorológicas, é vital que formas eficientes de previsão da produtividade energética, inclusive a FV, sejam concebidas, principalmente devido à característica estocástica desse formato de geração e para otimizar a operação do sistema elétrico. Para isso, modelos confiáveis de previsão, baseados em dados de medição das usinas geradoras e bases meteorológicas são necessários (Nespoli *et al.*, 2019).

A geração de energia FV está estreitamente relacionada à radiação solar e temperatura ambiente, conforme destacado por Cui *et al.* (2019). A Tabela 1 mostra os índices de correlação de variáveis meteorológicas, demonstrando a grande compatibilidade da radiação com a geração (0,82), seguida pela temperatura ambiente (0,60). Por outro lado, a umidade apresenta um comportamento inverso (-0,49) e, por fim, a velocidade do vento apresenta pouca representatividade (0,10). Portanto, é essencial utilizar séries históricas das variáveis meteorológicas para prever com maior precisão.

Tabela 1 - Índices de correlação entre Meteorologia e Geração FV

	Geração FV	Radiação solar	Temperatura ambiente	Umidade	Velocidade do vento
Geração FV	1,00	0,82	0,60	-0,49	0,10
Radiação solar	0,82	1,00	0,64	-0,40	0,08
Temperatura ambiente	0,60	0,64	1,00	-0,30	-0,30
Umidade	-0,49	-0,40	-0,30	1,00	-0,02
Velocidade do vento	0,10	0,08	-0,30	-0,02	1,00

Fonte: Adaptado de Francisco *et al.* (2019).

1.1 Contextualização e Descrição do Problema

A crescente inserção da energia FV na matriz elétrica brasileira, especialmente nas regiões Nordeste, Sul e Sudeste, aumenta a necessidade de modelos preditivos capazes de lidar com a intermitência e variabilidade dessa fonte. Essa variabilidade impõe desafios à operação do Sistema Interligado Nacional (SIN), exigindo que o Operador Nacional do Sistema Elétrico (ONS) mantenha reserva operativa elevada e recorra frequentemente ao despacho térmico. Essas ações elevam os custos operacionais, que tendem a ser repassados ao consumidor final. A Figura 2 sintetiza esse conjunto de efeitos.

Figura 2 - Desafios associados à previsão de Geração FV



Fonte: A Autora (2025).

O aprimoramento da acurácia das previsões de geração FV permitem também planejar a expansão da matriz elétrica com maior confiança, aproveitando plenamente o potencial da energia solar. Além disso, um modelo capaz de prever simultaneamente a geração de múltiplas usinas FV, desde que agrupadas por similaridade de perfil, pode trazer ganhos operacionais diretos ao SIN, harmonizando o despacho térmico e reduzindo o esforço computacional necessário para execução de múltiplos processos de previsão isolados.

Outro desafio é a escassez de bases públicas completas e integradas, tanto de geração quanto de variáveis meteorológicas, o que prejudica a reproduzibilidade dos resultados. Nesta pesquisa, foram utilizados dados públicos de geração horária média (MWmed) fornecidos pelo ONS. Como alternativa às medições meteorológicas locais, geralmente indisponíveis, recorreu-se à base de dados NASA POWER, que disponibiliza dados meteorológicos em escala horária ou diária de acordo com a coordenada geográfica. Essa base, derivada de sensoriamento remoto e modelos numéricos globais, supre a ausência de medições locais com ampla cobertura espacial e temporal.

As séries de geração FV, entretanto, frequentemente apresentam lacunas resultantes de falhas operacionais, limitações de sensoriamento ou interrupções na coleta de dados. Técnicas tradicionais de imputação, como médias móveis, interpolação linear ou k-vizinhos, nem sempre preservam a sazonalidade e a alta variabilidade da geração FV, podendo introduzir distorções. Neste trabalho, propõe-se uma heurística fundamentada na correlação entre perfis de radiação solar e geração, utilizando dias com padrões semelhantes de radiação para preencher lacunas de forma mais realista.

Do ponto de vista metodológico, a presente pesquisa busca desenvolver modelos mais precisos, robustos e generalizáveis por meio da integração entre Clusterização Hierárquica e *Temporal Fusion Transformer* (TFT). A primeira agrupa usinas com padrões de geração similares, permitindo a construção de modelos especializados por grupo e reduzindo a variância entre amostras. O TFT, por sua vez, incorpora mecanismos de atenção, codificação temporal e inferência probabilística, sendo especialmente adequado para lidar com séries temporais multivariadas e heterogêneas.

Assim, o problema central deste trabalho consiste em desenvolver e avaliar modelos preditivos multissérie, baseados em dados públicos. Os modelos devem ser capazes de aprender e prever de forma eficiente a geração FV em múltiplas usinas, mesmo diante de dados incompletos ou ruidosos. Nesse contexto, a arquitetura TFT, que é relativamente nova (Lim *et al.*, 2021) e vem sendo adotada de forma crescente em previsão de séries temporais, com aplicações pontuais em geração FV. As configurações multissérie ainda são pouco exploradas (López Santos *et al.*, 2022; Tao *et al.*, 2024) e, até o presente momento, não foram encontrados trabalhos que combinam explicitamente o TFT com a Clusterização Hierárquica para orientar o aprendizado entre usinas FV. Essa lacuna fundamenta a contribuição original deste trabalho.

Com base em estudos recentes sobre previsão de geração FV, a Tabela 2 resume de forma objetiva os principais problemas identificados na literatura e as soluções propostas nesta pesquisa, evidenciando as contribuições originais do estudo.

Tabela 2 - Problemas identificados na literatura e soluções propostas

Problema identificado na literatura	Solução proposta nesta pesquisa
Uso predominante de modelos de previsão de séries temporais individuais (Wu & Phan, 2022; Islam <i>et al.</i> , 2023).	Desenvolvimento de modelo multissérie capaz de prever simultaneamente múltiplas usinas, aproveitando padrões comuns entre elas.
Baixa disponibilidade de bases públicas completas e integradas de geração FV e meteorologia, dificultando a reproduzibilidade (López Santos <i>et al.</i> , 2022).	Uso exclusivo de dados públicos tanto para geração FV (ONS) quanto para variáveis meteorológicas (NASA POWER), garantindo transparência e reproduzibilidade.

Ausência de abordagem sistemática para lidar com lacunas nas séries de geração FV (Zainuddin <i>et al.</i> , 2022; Park <i>et al.</i> , 2023).	Heurística baseada na similaridade entre perfis de radiação solar e geração FV para imputação mais realista.
Escassez de estudos para previsão de geração FV com uso combinado de TFT e Clusterização Hierárquica em cenários multissérie (López Santos <i>et al.</i> , 2022; Islam <i>et al.</i> , 2023; Tao <i>et al.</i> , 2024).	Proposta original de combinar Clusterização Hierárquica e TFT para previsão FV multissérie, gerando modelos especializados por grupo.
Validação restrita a partições fixas (treino/teste), com partição aleatória do conjunto de dados (López Santos <i>et al.</i> , 2022).	Validação com janela deslizante fixa para robustez estatística da avaliação dos modelos.
Poucos estudos exploram o enriquecimento estruturado de covariáveis para capturar sazonalidade e interações não lineares (López Santos <i>et al.</i> , 2022; Islam <i>et al.</i> , 2023).	Geração de atributos derivados e interações para enriquecer as entradas do TFT.

Fonte: A Autora (2025).

1.2 Objetivo-Geral

Desenvolver e avaliar modelos de previsão de geração fotovoltaica baseados na arquitetura *Temporal Fusion Transformer*, integrados a técnicas de Clusterização Hierárquica para a otimização do desempenho preditivo em múltiplas usinas.

1.2.1 Objetivos Específicos

- Selecionar séries históricas de geração horária de dez usinas fotovoltaicas e as variáveis meteorológicas correlatas, provenientes de bases de dados públicas.
- Implementar método próprio de imputação baseado na correlação entre radiação solar e geração FV para o tratamento de lacunas nas séries de geração.
- Agrupar as usinas por meio da Clusterização Hierárquica com *Dynamic Time Warping* (DTW), identificando padrões de geração similares para a modelagem conjunta.
- Implementar o enriquecimento de dados mediante a criação de atributos derivados e interações, visando capturar sazonalidades e relações não lineares.
- Desenvolver e validar os modelos de previsão baseados em TFT, para uma horizonte de 11 horas, realizando a análise estatística do desempenho e a comparação com topologias preditivas de referência.

1.3 Organização do Trabalho

Esta tese de doutorado está organizada em seis capítulos. Este primeiro constitui-se da introdução que contextualiza a geração solar FV no cenário mundial e brasileiro, além de apresentar a motivação para a pesquisa, a descrição do problema e os objetivos do trabalho.

O Capítulo 2 apresenta o Estado da Arte em previsão de geração fotovoltaica, contextualizando o tema com as principais pesquisas e as suas contribuições mais recentes na literatura.

O Capítulo 3 apresenta a fundamentação teórica necessária para a compreensão do estudo. Dividido em cinco subseções, o capítulo enfatiza conceitos que são essenciais para a compreensão do trabalho.

No Capítulo 4, é apresentada a metodologia para desenvolvimento do modelo proposto nessa pesquisa. Nesse capítulo são detalhados os métodos e técnicas utilizadas.

Os resultados e as discussões da pesquisa são apresentados no Capítulo 5, que é dividido em cinco subseções, sendo cada uma delas referente às contribuições anteriormente propostas.

O Capítulo 6 refere-se à conclusão do trabalho, destacando os principais resultados alcançados e suas implicações no desenvolvimento de modelo para previsão de geração FV, contribuições do trabalho e sugestões para trabalhos futuros.

2 ESTADO DA ARTE

Na literatura existem diversos trabalhos que abordam de diferentes formas o problema da previsão da geração FV. A seleção adequada de técnicas de pré-processamento tem um grande impacto na precisão, a eliminação das horas noturnas e valores faltantes do conjunto de dados é muito necessário antes do aprendizado da rede. A escolha das variáveis de entrada é totalmente dependente do usuário, assim como os fatores climatológicos e parâmetros geográficos. O uso de dados de treinamento para um modelo é, no entanto, um processo experimental, uma base de dados de dois a três anos é considerada adequada para treinamento de uma modelagem (Singla; Duhan; Saroha, 2021).

A ausência de dados em séries temporais impacta severamente a categorização e a previsão, pois métodos tradicionais de processamento podem introduzir vieses e suposições imprecisas. Para abordar essa questão, a pesquisa de Zainuddin *et al.* (2022) visa categorizar os tipos de dados e mecanismos de dados ausentes em séries temporais, além de revisar diversas abordagens para preencher essas lacunas. O estudo foca nas técnicas atuais de tratamento de valores ausentes no pré-processamento de dados para séries univariadas e multivariadas, bem como nos métodos para avaliar a performance da imputação.

Zainuddin *et al.* (2022) descrevem os principais métodos para preencher valores ausentes em séries temporais, incluindo: a) *k-Nearest Neighbors* (kNN), que estima valores pela média ponderada de vizinhos próximos (Ivo Goltz *et al.*, 2023); b) *Local Average of Nearest Neighbors* (LANN), que preenche lacunas com a média dos vizinhos mais próximos (Flores; Tito; Silva, 2019); o Modelo de Índice Único, que simplifica a estimativa para uma única dimensão via abordagem de *kernel* (Ling *et al.*, 2022); Cadeias de Markov de Ordem Superior, para grandes lacunas univariadas baseadas em padrões de transição (Velasco-Gallego; Lazakis, 2022); e Imputação de Dados Sazonais, que utiliza redes neurais como MLP para lacunas longas em dados não lineares (Park *et al.*, 2023).

Considerando as particularidades de cada método de imputação de dados faltantes, percebeu-se que nenhum era ideal para a série temporal de geração FV estudada. Isso porque os dados ausentes ocorriam, na maioria das vezes, em horas consecutivas, o que inviabilizava técnicas baseadas em vizinhos próximos. Além disso, não foram identificadas lacunas longas o suficiente para justificar o uso de Cadeias de Markov ou Imputação de Dados Sazonais.

Assim, a abordagem heurística adotada neste trabalho consistiu em equiparar a série temporal de geração FV com a de radiação solar global da mesma localidade. O objetivo foi

utilizar o padrão da curva diária de radiação global para inferir os valores de geração nas horas ausentes, aproveitando a forte correlação entre essas duas variáveis.

Sobre a previsão da geração FV, na literatura são descritos diferentes métodos os quais podem ser divididos: físicos, estatísticos, por aprendizagem de máquina (Van Der Meer *et al.*, 2018) e modelos híbridos (Wang *et al.*, 2019). De acordo com o processo de previsão, este pode ser segmentado por análise pontual ou regional, além disso, sendo necessário definir a escala de tempo de averiguação (Wang; Qi; Liu, 2019).

Os modelos físicos simulam a dinâmica atmosférica com base em princípios físicos e condições de fronteira, utilizando dados meteorológicos e geográficos (Zhao *et al.*, 2018; Wang *et al.*, 2019), e são adequados para previsões de longo prazo, mas exigem alto esforço computacional (Hao; Tian, 2019). Os métodos estatísticos empregam relações matemáticas derivadas de dados históricos ambientais e de geração. Técnicas como ARMA, ARIMA, SMA e persistência são comuns, mas a persistência não é ideal para FV devido à não-linearidade da irradiação solar. De modo geral, esses modelos lineares têm limitações para previsões de longo prazo (Abdel-Nasser; Mahmoud, 2019; Wang *et al.*, 2019; Cunha, 2021). Por fim, os modelos de previsão por aprendizagem de máquina são amplamente utilizados, oferecendo melhores resultados devido à sua capacidade de tratar dados complexos. Abordagens como redes neurais artificiais, redes *fuzzy* adaptativas e DL são eficazes para lidar com relações não lineares entre entrada e saída, minimizando erros (Wang; Qi; Liu, 2019).

Shi *et al.* (2012) propuseram um modelo para prever a potência FV com 15 minutos de antecedência no dia seguinte. Essa abordagem se baseia na classificação meteorológica (céu nublado, claro, chuvoso e com neblina) e utiliza *Support Vector Regression* (SVR), um algoritmo de aprendizado de máquina supervisionado. O modelo utilizou como entradas o histórico de potência FV de 15 minutos do dia mais próximo com temperaturas máximas, mínimas e médias semelhantes às do dia seguinte, obtidas de um boletim meteorológico. Como esperado, os melhores resultados de previsão foram alcançados para dias ensolarados e nublados.

Zhou *et al.* (2019) desenvolveram um método híbrido para previsão de energia FV utilizando uma rede LSTM combinada com um mecanismo de atenção (ALSTM). Esse mecanismo, uma rede neural que calcula uma soma ponderada das entradas, seleciona de forma adaptativa as informações mais relevantes, aprimorando a precisão. O estudo empregou duas redes LSTM, uma para prever a potência e outra para a temperatura dos módulos, usando dados de uma usina FV de 20 kWp, registrados a cada 7,5 minutos. A análise comparativa,

realizada nas quatro estações do ano, mostrou que o ALSTM superou o Modelo Persistente, ARIMAX, MLP e LSTM tradicional em todos os horizontes de previsão, de 7,5 a 60 minutos à frente, ao focar nos recursos de entrada mais significativos.

Abdel-Nasser e Mahmoud (2019) propuseram e testaram cinco arquiteturas de rede LSTM para prever a potência de sistemas fotovoltaicos com uma hora de antecedência, utilizando apenas dados de produtividade. As variações incluíram: a) LSTM tradicional (modelo 1), que previa a geração da hora seguinte a partir da hora atual, treinada com diferentes épocas; b) LSTM com técnica da janela (modelo 2), que usava múltiplos instantes de tempo como entrada do modelo; c) LSTM com *time step* (modelo 3), que concentrava etapas de tempo anteriores como uma única entrada para prever a potência subsequente, considerada um enquadramento mais preciso do problema; d) LSTM com memória entre lote (modelo 4), que permitia manter o estado interno da rede através dos lotes de treinamento; e e) LSTM empilhada com memória entre lotes (modelo 5), que agrupava camadas LSTM para extrair mais conceitos e potencialmente melhorar as previsões. A análise de desempenho, usando a métrica RMSE, revelou que o modelo 3 (LSTM com *time step*) obteve os melhores resultados (Abdel-Nasser; Mahmoud, 2019).

Paiva (2021) explorou a previsão intradiária da irradiância solar utilizando dois algoritmos de aprendizado de máquina: Programação Genética Multi-gene (PGMG), que deriva funções de regressão analítica, e redes neurais artificiais MLP. Ambos foram aplicados a dados de seis locais em três países, comparando previsões com horizontes de 15 a 120 minutos. Os resultados das simulações indicaram que a precisão do modelo dominante é influenciada pela localização, horizonte de previsão e erro de avaliação. PGMG e MLP demonstraram desempenho semelhante e consistente. Além disso, as previsões melhoraram significativamente com a inclusão de variáveis climáticas exógenas (como irradiância e temperatura ambiente) nas entradas dos modelos, reiterando a influência da localidade, horizonte de previsão e métrica de erro na acurácia.

Phan, Wu e Phan (2022) propuseram um promissor modelo para prever a geração de energia FV com uma hora de antecedência, utilizando a arquitetura de rede *Transformer*. A seleção das variáveis de entrada foi realizada com o coeficiente de correlação de Pearson, que identificou nove dados relevantes, incluindo informações de Previsão Numérica do Tempo (NWP) e medições diretas de um parque solar em Taiwan. Conforme esperado, observou-se uma alta correlação (cerca de 0,99) entre a irradiância solar e a potência gerada.

As simulações indicaram que o modelo *Transformer* superou outras arquiteturas como ANN, LSTM e GRU em métricas de desempenho (NRMSE e NMAPE). Além disso, sua

capacidade de processamento paralelo e a consequente redução do tempo de computação são vantagens significativas, especialmente em cenários onde a rapidez na previsão é crucial (Phan; Wu; Phan, 2022). Os resultados gerais demonstram que o mecanismo de atenção do *Transformer* pode aprender padrões complexos em séries temporais. Isso sugere que a exploração de modelos híbridos para previsão de geração de energia FV, combinando diferentes arquiteturas de redes neurais ou integrando informações de diversas fontes, pode levar a modelos mais robustos e precisos, dada a complexidade e variabilidade dos dados de energia FV.

Phan, Wu e Phan (2023) propuseram um modelo para prever a geração FV com um dia de antecedência, combinando pré-processamento, pós-processamento e a arquitetura *Transformer*. No pré-processamento, dados ausentes foram imputados utilizando a rede XGBoost. A estrutura *Transformer* foi aprimorada com camadas adicionais de normalização e *Dropout*, além de atenção multi-cabeça. Para aumentar a precisão da previsão, uma técnica de pós-processamento foi aplicada, combinando curvas de ajuste numérico diário e uma Unidade Recorrente *Gated* (GRU). Simulados com dados históricos de dez fazendas solares em Taiwan e dados NWP, o modelo superou ANN, LSTM, XGBoost e GRU, demonstrando a importância do pré e pós-processamento para um bom desempenho preditivo.

Al-Ali *et al.* (2023) propõem um modelo híbrido para previsão de geração FV que combina CNN, LSTM e *Transformer* em uma arquitetura codificador-decodificador. Na fase de pré-processamento, utilizaram mapas auto-organizáveis para agrupar e selecionar características relevantes oriundas de dados históricos de geração e variáveis meteorológicas, organizadas em quatro categorias. A CNN extrai padrões espaciais, a LSTM captura dependências temporais e o *Transformer* integra as informações para realizar a previsão. Os resultados demonstraram desempenho superior em relação a modelos como ARIMA, GRU-CNN e *Prophet*, com menor complexidade computacional, evidenciando a eficácia da combinação entre pré-processamento por agrupamento e modelagem híbrida.

Quanto à aplicação do modelo TFT para previsão de geração FV, as principais pesquisas são resumidas nos parágrafos seguintes.

O estudo de López Santos *et al.* (2022) utilizou o TFT para prever a geração horária de energia FV com 24 horas de antecedência, utilizando dados históricos de produção, variáveis meteorológicas e informações de calendário de seis instalações na Alemanha e Austrália. A metodologia envolveu um pré-processamento de dados rigoroso, incluindo a detecção de *outliers* (DBSCAN), a seleção de atributos por *Backward Elimination* guiada pelo TFT (destacando irradiação solar horizontal e ângulos solares como cruciais) e padronização das

entradas contínuas com aplicação de logaritmo natural e escalonamento para a produção FV. Comparado a modelos como ARIMA, MLP, LSTM e XGBoost, o TFT demonstrou superioridade consistente, apresentando os menores erros e reduzindo-os em até 54% em relação ao LSTM. Sua interpretabilidade confirmou a irradiação solar horizontal e o ângulo de zênite como as variáveis mais influentes, com padrões de atenção cílicos. Apesar da sua complexidade, o desempenho e a interpretabilidade do TFT o tornam uma ferramenta valiosa para previsões precisas de energia FV.

Além disso, no trabalho de López Santos *et al.* (2022), foram empregadas covariáveis meteorológicas de uma estação próxima (irradiância solar, ângulo zenital, ângulo azimutal, temperatura e umidade relativa), transformação Log na variável-alvo e codificação da sazonalidade via seno e cosseno do mês, sem detalhar a distância entre usinas e estação. Ademais, a partição dos dados de treinamento, validação e teste foi realizada por amostragem aleatória.

O artigo de Islam *et al.* (2023) propõe um modelo de previsão de energia FV utilizando o TFT, para prever a geração horária de energia. A metodologia empregada utilizou um conjunto de dados real de uma usina solar na Índia, com granularidade de 15 minutos ao longo de 34 dias, e que compreende 23 inversores. O estudo incorpora dados geográficos, como irradiação solar e temperatura dos módulos, além de características horárias e ID dos inversores como entrada do modelo TFT. Um ponto crucial destacado pelos autores é a capacidade do TFT de interpretar a importância de diferentes variáveis no processo de previsão. Os resultados da comparação com modelos como ARIMA, LSTM e *Prophet* demonstraram que o TFT superou consistentemente todos os outros modelos em todas as métricas de desempenho. Além da precisão, a interpretabilidade do TFT revelou que a irradiação é a característica mais importante no codificador e o ID do inversor a variável estática mais significativa, fornecendo recursos valiosos sobre os fatores que influenciam a previsão.

O estudo de Tao *et al.* (2024) desenvolveu o PTFNet (Rede Paralela de Extração de Informação Temporal) para aprimorar a previsão de energia FV. A metodologia inovou ao apresentar um método de aumento de dados que incorpora informações de modelagem física específicas do sistema solar, como ângulos solares e irradâncias (DHI, DNI, GTI), enriquecendo os Dados de Medição Local (LMD) e Previsão Numérica do Tempo (NWP). O PTFNet utiliza uma arquitetura baseada em *Transformer* com uma estrutura paralela para extrair explicitamente dependências temporais e entre características. A técnica emprega segmentação de dados para otimizar a extração de informações e a associação entre dados

passados e previsões futuras. As previsões são geradas para horizontes de 24, 36 e 40 horas, com resolução de 15 minutos. Os resultados demonstram que o aumento de dados foi eficaz em 78,57% dos casos, reduzindo o RMSE e o MAE. O PTFNet superou ou foi altamente competitivo em comparação com outros modelos populares de previsão de séries temporais longas, apresentando os valores ótimos de MAE e RMSE na maioria dos cenários e prevendo picos diurnos com maior precisão. A segmentação de dados também se mostrou eficaz na melhoria do desempenho. A pesquisa também ressaltou a importância da qualidade dos dados NWP para a precisão das previsões.

A abordagem heurística desta pesquisa, quanto à seleção de atributos, combina a autocorrelação para uma pré-seleção inicial de variáveis relevantes com uma otimização iterativa do TFT baseada em métricas de perda. Essa abordagem difere do *Backward Elimination* guiado pelo TFT adotado por López Santos *et al.* (2022) e da Rede Paralela proposta por Tao *et al.* (2024) por incluir uma triagem temporal prévia, que direciona a seleção para atributos intrinsecamente relacionados à natureza de séries temporais. Essa metodologia híbrida, que une análise estatística clássica e refinamento por TFT, mostra-se robusta, prática, com menor esforço computacional e maximiza a precisão preditiva do TFT.

Nas pesquisas recentes, o TFT tem apresentado resultados competitivos (e por vezes superiores) frente a modelos clássicos, nos conjuntos e configurações avaliados. A Tabela 3 sintetiza os três trabalhos representativos acima e suas principais escolhas metodológicas.

Tabela 3 - Estudos Correlatos sobre TFT na Previsão FV

Fonte	Modelo (Horizonte Previsão)	Escopo	Entradas (covariáveis)	Principais Métodos	Modelos utilizados	Valida- ção	Resultados e observações
López Santos <i>et al.</i> (2022)	TFT (24 h)	Multissérie 06 Usinas (Alemanha Austrália)	Irradiação Solar / Ângulo Solar Zenital / Temperatura / Ângulo Solar Azimute / Mes_seno / Mes_cosseno / Umidade Relativa	DBSCAN (outliers); <i>Backward Elimination</i> guiado pelo TFT; Transform. <i>log</i>	TFT, ARIMA, MLP, LSTM, XGBoost	Partição aleatória (treino/ val/teste)	TFT com menores erros; irradiância e zênite como variáveis mais influentes; risco de <i>leakage</i> pelo <i>split</i> aleatório
Islam <i>et al.</i> (2023)	TFT (Hora-à- frente)	Unissérie; 1 usina (Índia); 23 inversores	Irradiação Solar / Temperatura do Módulo / Hora / Mês / Energia Inversor	Temperatura dos módulos; engenharia de atributos simples.	TFT, ARIMA, LSTM, Prophet	Temporal	TFT supera modelos clássicos; irradiância e ID do inversor (estática); duração do conjunto de dados (34 dias).

Tao <i>et al.</i> (2024)	PTFNet (24 h, 36 h, 40h)	Multissérie 02 Usinas (China)	Previsão Numérica do Tempo / Modelos Climáticos Globais / Irradiação Solar / Ângulo Solar / Temperatura	Modelagem física específicas do sistema solar; segmentação de dados	PTFNet (Transformer paralelo), 10 modelos tradicionais	Temporal	Aumento de dados eficaz em 78,6%; predição de picos melhor; sensível à qualidade do NWP.
--------------------------	--------------------------	-------------------------------	---	---	--	----------	--

Fonte: A Autora (2025).

• Conclusão do Capítulo

Em conclusão, pode-se afirmar que o estudo da previsão da geração de energia FV é um campo em constante evolução, impulsionado pelo avanço das tecnologias de inteligência artificial e pela crescente disponibilidade de dados. Ao longo deste capítulo, foi explorada uma variedade de abordagens e modelos utilizados na literatura para prever a produção de energia solar.

Foi observado que as redes neurais têm se destacado como uma ferramenta poderosa para lidar com a complexidade dos dados de séries temporais e capturar padrões não lineares na geração de energia FV. Arquiteturas como LSTM, CNN, MLP, *Transformer* e TFT têm sido amplamente empregados com sucesso em diferentes contextos e cenários de previsão. Além disso, é válido destacar a importância do pré-processamento adequado dos dados, incluindo técnicas para imputação de dados faltantes, seleção de variáveis relevantes e análise de correlação. Essa etapa é essencial para garantir a qualidade dos dados de entrada e melhorar a precisão dos modelos de previsão. Por fim, o desenvolvimento de modelos de previsão mais precisos e eficazes é crucial para promover e ampliar a integração da energia solar na matriz energética global.

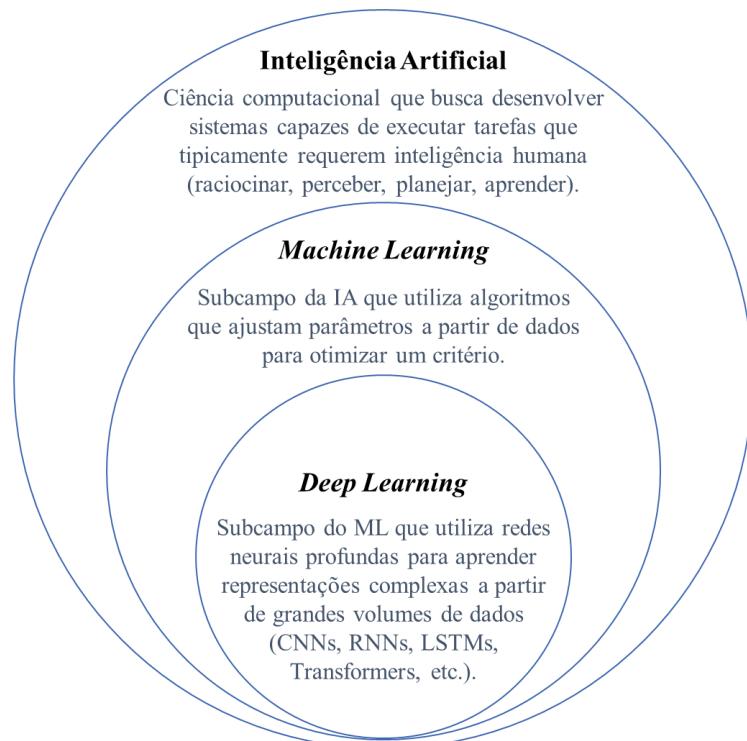
3 FUNDAMENTAÇÃO TEÓRICA

3.1 Inteligência Artificial

A expressão Inteligência Artificial (IA) foi definida em 1956 por J. McCarthy, um dos pioneiros da Ciência da Computação, que a definiu como a ciência e engenharia voltadas ao desenvolvimento de “máquinas inteligentes” (Gurkaynak; Yilmaz; Haksever, 2016). Apesar de sua origem na computação, a IA representa um campo de estudos multidisciplinar, que inclui engenharia, matemática, cibernetica, psicologia, entre outros (Valdati, 2020). Sua definição, contudo, permanece desafiadora e multifacetada. Segundo Haugeland (1985), a IA representa um esforço interessante em fazer os computadores ou máquinas pensarem como mentes humanas no sentido total e literal.

Com o avanço das pesquisas, a IA passou a abranger diferentes subáreas, destacando-se entre elas o *Machine Learning* (ML), que consiste em algoritmos capazes de aprender padrões a partir de dados para melhorar o desempenho em tarefas específicas e o *Deep Learning* (DL), um ramo do ML baseado em redes neurais profundas, representadas por várias camadas que conseguem aprender representações complexas a partir de grandes volumes de dados. A Figura 3 ilustra, de forma esquemática, a hierarquia conceitual entre IA, ML e DL.

Figura 3 – Hierarquia Conceitual entre IA, ML e DL



Fonte: Adaptado de Chollet (2021).

O subcampo ML está relacionado com a elaboração de algoritmos que aprendem e se aperfeiçoam automaticamente através da experiência, com o intuito de encontrar padrões em um conjunto de dados (Mitchell, 1997). Tais lógicas de programação envolvem a autoaprendizagem para, por exemplo, realizar previsões, substituindo atividades manuais por modelos construídos para analisar grande volume de elementos em uma base. Dessa forma, é possível afirmar que o ML se caracteriza pela utilização de códigos para extrair informações de dados brutos e representá-los através de algum modelo matemático.

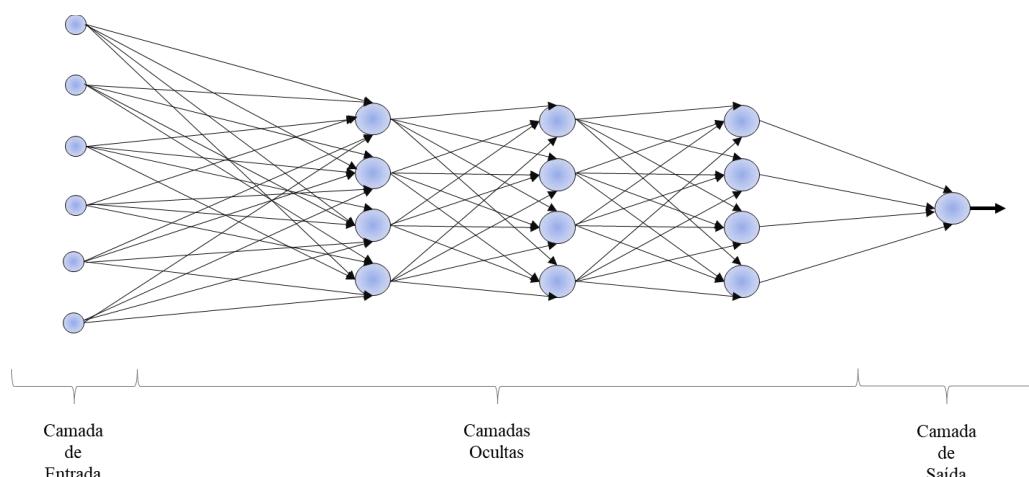
Por sua vez, o subconjunto DL representa uma forma de aprendizagem de representações hierárquicas ao longo de múltiplas camadas (Chollet, 2021). Cada uma dessas camadas processa as informações de entrada, criando uma interpretação distinta da base de dados (Alzubaidi *et al.*, 2021).

No campo da previsão de séries temporais, quanto ao avanço histórico, pode-se destacar: as redes *Multilayer Perceptron* (MLP), as redes *Long Short-Term Memory* (LSTM), as redes *Transformer* e, por fim, as redes *Temporal Fusion Transformer* (TFT).

3.1.1 Redes Neurais MLP

As redes MLP, propostas por Rumelhart, Hinton e Williams em 1986, são redes neurais clássicas, formadas pela camada de entrada, uma ou mais camadas ocultas, assim chamadas por não serem diretamente observáveis, e, uma camada de saída. Nessas redes, cada neurônio em qualquer camada (exceto a de entrada) está conectado a todos os nós ou neurônios da camada anterior. A Figura 4 mostra um exemplo de rede MLP com três camadas ocultas, onde o fluxo de sinal sempre progride no sentido da camada de entrada em direção à camada neural de saída.

Figura 4 - Arquitetura de uma Rede MLP



Fonte: Adaptado de Haykin (2001).

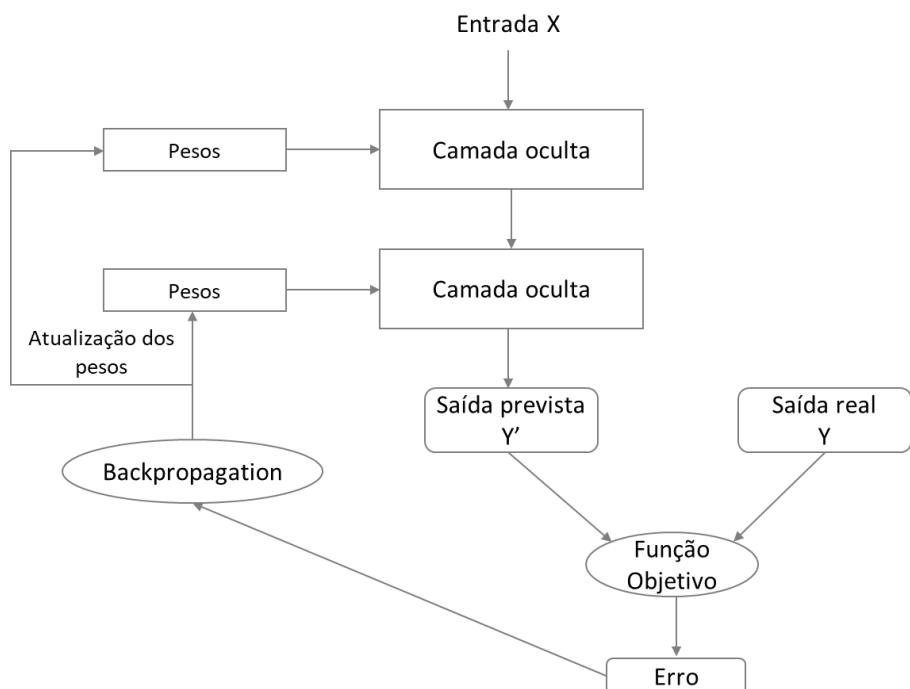
As redes MLP podem ter seu funcionamento explicado da seguinte forma: os dados de entrada (X) são processados nas camadas ocultas, nas quais há neurônios que aplicam uma mesma função de ativação, parametrizada por pesos e viés (bias). Cada neurônio calcula uma combinação linear de suas entradas e, em seguida, aplica uma função de ativação não linear, o que permite à rede aproximar relações complexas entre as variáveis de entrada e saída. Entre as funções de ativação mais utilizadas nas camadas ocultas destacam-se a ReLU (Unidade Linear Retificada), a tangente hiperbólica (\tanh) e a função sigmoide, cuja escolha depende do tipo de problema e das características dos dados.

A extração de características relevantes dos dados de entrada é feita por meio de um algoritmo que detecta automaticamente recursos significativos da base fornecida, para fins de treinamento, aprendizado e generalização.

Para uma dada saída prevista (Y'), a função objetivo tem a missão de comparar as previsões da rede com o valor real esperado para aquela saída (Y) e calcular a distância entre ambas (erro). Essa informação é utilizada para, se necessário, realizar o ajuste dos valores dos pesos, de forma a minimizar o erro, por meio do algoritmo de retropropagação do erro (*Backpropagation*). Inicialmente, os valores dos pesos são aleatórios, resultando em um erro (ou perda) elevado, pois a saída prevista é, provavelmente, distinta do valor real (Goodfellow; Bengio; Courville, 2016).

A Figura 5 mostra o fluxograma de treinamento da rede MLP.

Figura 5 – Fluxograma do treinamento da rede MLP



Fonte: Adaptado de Chollet (2021).

Apesar do avanço de arquiteturas sofisticadas, estudos recentes (Chen *et al.*, 2023) sugerem que modelos lineares e baseados em MLP apresentam desempenho competitivo e, em alguns casos, superior. O estudo de Chen *et al.* (2023) demonstrou empiricamente que modelos lineares podem superar arquiteturas sequenciais mais complexas em tarefas de previsão de séries temporais, especialmente no caso univariado (com uma única variável). Isso ocorre porque modelos lineares possuem capacidade de representação adequada para capturar dependências temporais quando há suavidade ou periodicidade nos dados.

A rede MLP apresenta formulação simples e eficiente. A saída do único neurônio da camada de saída, que representa uma função linear, é definida conforme Eq. (1):

$$y' = \sum_{k=1}^N W_k y_k + b \quad (1)$$

em que W_k representa os pesos que conectam as saídas dos neurônios da camada oculta ao neurônio de saída, y_k são as saídas dos neurônios, b representa o viés (bias) e y' é a saída da rede.

Considerando a finalidade de previsão, sabe-se que as redes MLP não apresentam dependência temporal direta, assim, essa dependência temporal precisa ser simulada ao apresentar um conjunto de valores passados como entrada para o modelo, de modo que a informação temporal esteja implícita, e não incorporada diretamente na estrutura da rede (Lazcano; Jaramillo-Morán; Sandubete, 2024).

Dessa forma, ao se utilizar uma MLP para previsão de séries temporais, o vetor de entrada é formado por um conjunto de observações passadas da série, que são processadas para gerar a previsão. Em termos formais, a rede recebe, no instante t , um vetor de entrada $x_t = [x_{t-1}, x_{t-2}, \dots, x_{t-M}]$, o qual é processado para fornecer uma saída y'_t . Para viabilizar esse processo, a série temporal deve ser reorganizada em uma estrutura matricial, em que cada linha corresponde a um vetor de entrada de dimensão M , e essas linhas são apresentadas sequencialmente à rede para a geração das previsões correspondentes.

3.1.2 Redes LSTM

As Redes Neurais Recorrentes (RNR), são redes neurais artificiais que funcionam por meio de sequências ou *loops*, nas quais as informações úteis são armazenadas em memórias, persistem durante o treinamento e influenciam diretamente na saída resultante. Esse tipo de arquitetura foi elaborado para reconhecer padrões em uma série de dados, considerando tempo

e ordem, o aprendizado é feito a cada passo de treinamento, de modo que a saída no instante ‘ $t - 1$ ’ é alimentada de volta à rede o que interfere no resultado do instante ‘ t ’ e assim sucessivamente.

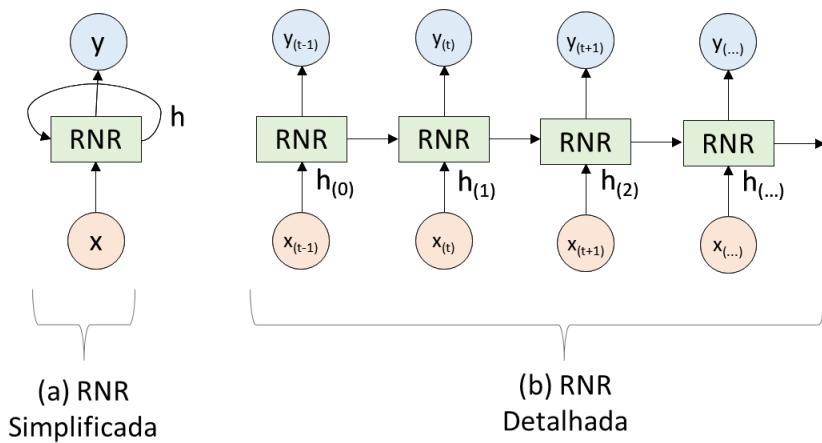
Uma RNR utiliza a ideia de que o estado oculto (ou memória) é atualizado a cada instante de tempo, de forma que o valor atual dependa do evento anterior, uma vez que tanto a entrada quanto a memória impactam na decisão (Bianchi *et al.*, 2018). Considerando h_t como o estado oculto da RNR, a equação que define essa dinâmica pode ser expressa pela Eq. (2), conforme Goodfellow, Bengio e Courville (2016):

$$h_{(t)} = a(\mathbf{W}x_{(t)} + \mathbf{U}h_{(t-1)} + b) \quad (2)$$

em que $a(\cdot)$ representa a função de ativação não linear, $h_{(t-1)}$ representa o estado oculto no instante anterior $t - 1$, $x_{(t)}$ representa o vetor de entrada no instante t , b representa o viés (bias), e, \mathbf{W} e \mathbf{U} representam as matrizes de peso de entrada–estado oculto e de estado oculto–estado oculto.

Sendo assim, pode-se afirmar que a RNR processa a informação da entrada no instante t $x_{(t)}$, mas incorpora o estado oculto anterior $h_{(t-1)}$ para definir o sinal de saída. Considerando que X e Y são os vetores de entrada e saída da rede, a Figura 6 ilustra esse processo. Na parte (a), apresenta-se a representação genérica da RNR; na parte (b), evidencia-se que a saída em cada instante t é determinada tanto pela entrada nesse instante quanto pela memória do instante anterior.

Figura 6 - Esquema de Funcionamento de uma RNR



Fonte: Adaptado de Goodfellow, Bengio e Courville (2016).

Apesar de seu ótimo desempenho, a RNR apresenta limitações quando a sequência de entrada é muito longa, pois as primeiras informações da cadeia são perdidas com o passar do tempo durante o processo de *Backpropagation*, que se constitui de uma função de otimização definida por derivadas parciais de um conjunto de parâmetros de entrada, com a finalidade de atualizar os valores dos pesos da rede neural. Dessa forma, se o valor do gradiente for muito pequeno significa que haverá pouca contribuição daquele dado para o aprendizado da rede (Lazzeri, 2021).

Nesse contexto, o modelo de rede LSTM foi proposto por Hochreiter e Schmidhuber (1997) para mitigar o problema da dependência de longo prazo em RNRs convencionais, introduzindo portões (*gates*) de controle e um estado de célula (*cell state*) na arquitetura recorrente. Essa modificação torna a LSTM especialmente adequada para classificar, processar e prever séries temporais com intervalos de duração desconhecida, evitando a perda de informações úteis ao longo do tempo.

A arquitetura da rede LSTM possui três camadas principais: entrada, oculta e saída. Entretanto, a camada oculta contém blocos de memória, chamados células, organizados em uma estrutura em cadeia que engloba quatro portões internos (Hochreiter; Schmidhuber, 1997). A informação é retida nas células, enquanto as manipulações de memória são realizadas pelos *gates*.

Uma rede LSTM utiliza três portões principais para controlar o fluxo de informação, mantendo o que é relevante e descartando o que é irrelevante. Em termos gerais, valores próximos de 0 indicam que determinada informação deve ser bloqueada, enquanto valores próximos de 1 indicam que ela deve ser preservada para o próximo passo temporal (Torres *et al.*, 2020). Esses portões são conhecidos como *Forget Gate*, *Input Gate* e *Output Gate*. Em seu estudo, Wang *et al.* (2018) descrevem de forma simplificada o papel de cada um desses componentes.

a) *Forget Gate*: nesse portão, as informações que não são mais úteis para o estado da célula são atenuadas. O estado oculto do instante anterior $h_{(t-1)}$ e a entrada atual $x_{(t)}$ são concatenados, multiplicados pela matriz de peso \mathbf{W}_f e adicionados ao *bias* b_f . Em seguida, o resultado passa por uma função de ativação sigmoide σ , que produz valores na faixa (0,1), aplicados elemento a elemento ao estado da célula anterior $c_{(t-1)}$. Valores próximos de 1 indicam que a informação deve ser preservada; valores próximos de 0 indicam que deve ser esquecida. A Eq. (3) representa a expressão do *Forget Gate* ($f_{(t)}$):

$$f_{(t)} = \sigma(\mathbf{W}_f[h_{(t-1)}, x_{(t)}] + b_f) \quad (3)$$

b) *Input Gate*: a adição de novas informações relevantes ao estado da célula é feita por esse portão. Primeiro, calcula-se um vetor de ativação sigmoide $i_{(t)}$, que controla quais componentes da nova informação serão incorporados, e, em paralelo, um vetor candidato $g_{(t)}$, obtido pela aplicação da função tanh, cujos valores pertencem ao intervalo $[-1,1]$. As Eq. (4) e (5) descrevem as expressões de $i_{(t)}$ e $g_{(t)}$, respectivamente:

$$i_{(t)} = \sigma(\mathbf{W}_i[h_{(t-1)}, x_{(t)}] + b_i) \quad (4)$$

$$g_{(t)} = \tanh(\mathbf{W}_g[h_{(t-1)}, x_{(t)}] + b_g) \quad (5)$$

em que \mathbf{W}_i e \mathbf{W}_g representam as matrizes de peso associados a cada etapa e b_i e b_g indicam o vetor *bias* de cada passo.

Ao final das duas operações, o estado da célula é atualizado para o instante atual, combinando as contribuições de $f_{(t)}$, $i_{(t)}$ e $g_{(t)}$, conforme Eq. (6):

$$c_{(t)} = f_{(t)} \odot c_{(t-1)} + i_{(t)} \odot g_{(t)} \quad (6)$$

em que $c_{(t-1)}$ representa o estado da célula no instante anterior e \odot denota o produto elemento a elemento.

c) *Output Gate*: a extração de informações relevantes do estado da célula atual é feita nessa etapa. Primeiramente, o estado oculto do instante anterior $h_{(t-1)}$ e a entrada atual $x_{(t)}$ são concatenados, multiplicados pela matriz de peso \mathbf{W}_o , somados ao viés b_o e, em seguida, passados por uma função de ativação sigmoide, gerando o vetor $o_{(t)}$, que controla a saída da rede, conforme a Eq. (7):

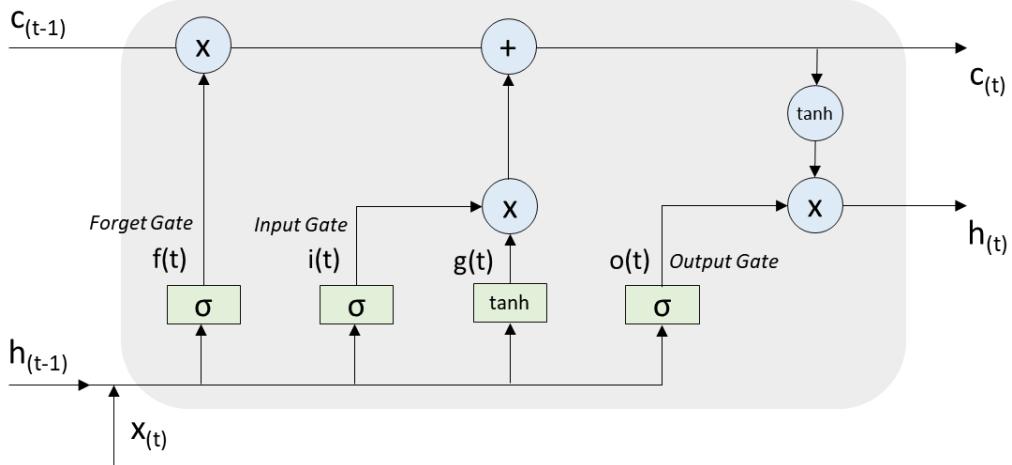
$$o_{(t)} = \sigma(\mathbf{W}_o[h_{(t-1)}, x_{(t)}] + b_o) \quad (7)$$

em que \mathbf{W}_o e b_o representam, respectivamente, a matriz de peso e o *bias* dessa fase. Por fim, aplcia-se a função tanh a $c_{(t)}$ e a saída (estado oculto atual) é determinada conforme a Eq. (8):

$$y_{(t)} = h_{(t)} = o_{(t)} \odot \tanh(c_{(t)}) \quad (8)$$

De forma resumida, a rede LSTM recebe duas entradas: o estado oculto do instante anterior $h_{(t-1)}$ e o vetor de entrada atual $x_{(t)}$. A memória de longo prazo é simbolizada pelo estado da célula $c_{(t)}$, que é modificado pelo *Forget Gate* e pelo *Input Gate*. O par $(h_{(t)}, c_{(t)})$ compõe o estado interno da rede no instante atual. Assim, a Figura 7 retrata a arquitetura da rede LSTM.

Figura 7 - Diagrama Esquemático da rede LSTM



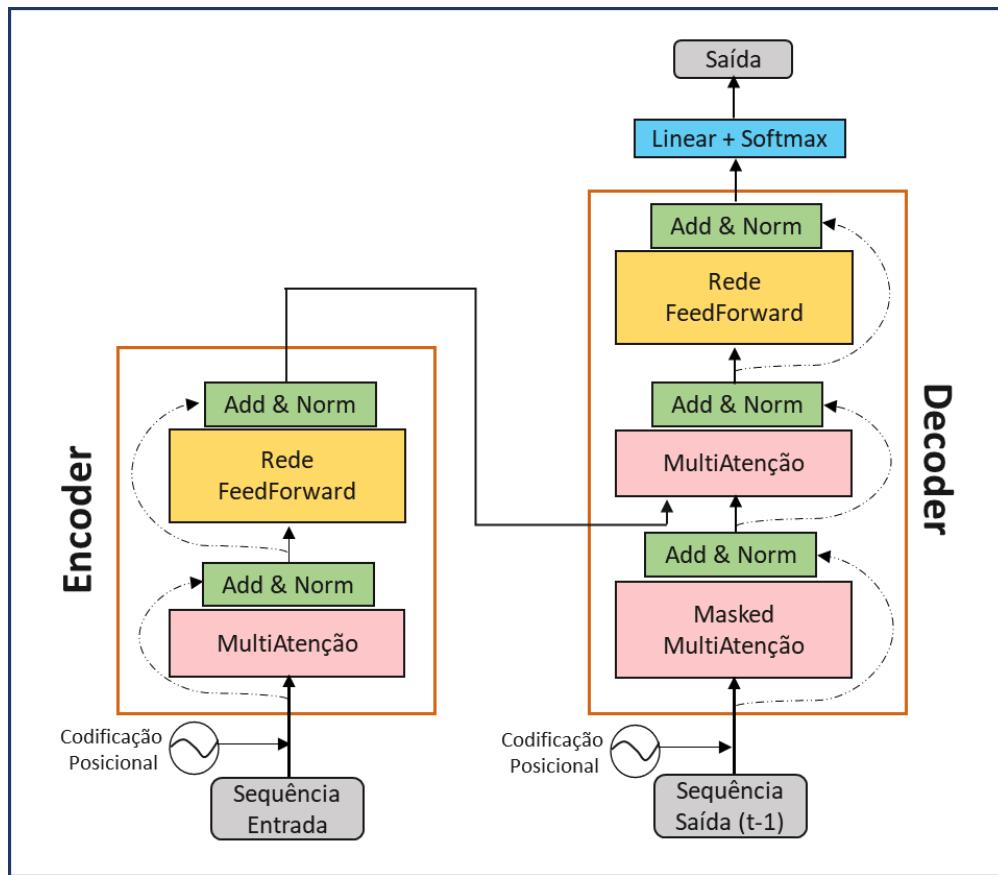
Fonte: A Autora (2025).

3.2 Arquitetura *Transformer*

Até o ano de 2017, arquiteturas como MLPs, RNRs e, em especial, suas variantes LSTM figuravam entre as principais abordagens para o processamento de sequências temporais, ainda que continuem amplamente utilizadas em diversas aplicações. No entanto, esses modelos apresentavam problemas como: dependências sequenciais que tornavam o treinamento e a inferência lentos; dificuldade em capturar dependências longas de forma eficiente e limitação de paralelismo que dificultava a análise simultânea de elementos nas séries. Para superar essas barreiras, Vaswani *et al.* (2017) propuseram o modelo *Transformer*, que revolucionou o campo do processamento de linguagem natural e aprendizado de máquina e que depende inteiramente de mecanismos de atenção e elimina o uso de recorrência.

O *Transformer* é um modelo DL cuja arquitetura é composta pela configuração de codificadores (Encoder) e decodificadores (Decoder), empilhados sequencialmente onde todas as camadas são idênticas. Cada um desses componentes desempenha um papel vital no processamento da sequência de entrada e na geração da sequência de saída, respectivamente. A Figura 8 mostra o diagrama esquemático do modelo desenvolvido por Vaswani *et al.* (2017).

Figura 8 - Diagrama Esquemático da Arquitetura Transformer

Fonte: Adaptado de Vaswani *et al.* (2017).

Analizando o Encoder, tem-se:

- Sequência de Entrada: representa a transformação dos dados de entrada em representações vetoriais de dimensão fixa e contínua, adequadas para serem processadas por modelos de aprendizado profundo. Essa técnica permite que variáveis categóricas ou temporais sejam representadas de forma densa e semântica no espaço vetorial.
- Codificação posicional: consiste no posicionamento dos elementos de entrada em uma sequência, para preservar a noção de temporalidade.
- Mecanismo de MultiAtenção: constitui-se de múltiplas “cabeças” de atenção que operam em paralelo para capturar diferentes padrões de dependência na sequência.
- Bloco “Add & Norm”: representa a conexão residual, que soma a entrada original da subcamada à sua saída. Em seguida, essa soma é processada por uma operação de normalização por camadas, que estabiliza o aprendizado ao normalizar as ativações dentro de cada amostra. Isso é crucial para mitigar problemas de instabilidade em redes profundas.
- Rede Feedforward: é uma subcamada aplicada a cada posição da sequência de entrada, composta por duas camadas densas totalmente conectadas com uma função de ativação

ReLU, que produz 0 como saída quando $x < 0$, e uma saída linear com inclinação unitária quando $x > 0$. Sua função é refinar e transformar as representações produzidas pela subcamada de atenção, permitindo ao modelo aprender transformações não lineares sobre as sequências de entrada.

Da mesma forma, analisando o Decoder, tem-se:

- a) Sequência de Saída ($t - 1$): o modelo recebe os valores de saída anteriores como parte do seu contexto para prever o próximo valor, esse deslocamento garante que, no treinamento, o modelo aprenda a prever o valor no tempo t com base apenas nas informações disponíveis até $t - 1$, evitando assim vazamento de informação.
- b) Codificação posicional: consiste no posicionamento dos elementos de saída em uma sequência, para preservar a noção de temporalidade.
- c) *Masked MultiAtenção*: é uma variação do mecanismo de MultiAtenção, o termo *Masked* refere-se à aplicação de uma máscara que impede o modelo de acessar posições futuras na sequência durante o treinamento, garantindo que a previsão de um determinado passo temporal utilize apenas informações passadas, o que é essencial em tarefas de previsão de séries temporais.
- d) Bloco “Add & Norm”, Mecanismo de MultiAtenção e Rede Feedforward: tem as mesmas funções do Encoder, utilizando a Sequência de Saída ($t-1$).
- e) Linear + Softmax: após passar as etapas anteriores, a saída final do Decoder é uma sequência de vetores contínuos. Para transformar esses vetores em saídas interpretáveis probabilísticas de valores discretos ou categóricos aplica-se a função Linear que projeta cada vetor para uma dimensão menor e a função Softmax, expressa matematicamente conforme Eq. (9), que transforma os vetores de valores reais em vetores de probabilidades (0 a 1), onde as probabilidades de cada valor são proporcionais à escala relativa de cada valor no vetor de origem (Ren; Wang, 2023).

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \quad (9)$$

em que z_i é o elemento i no vetor z e j é o número total de classes. O denominador assegura que a soma das probabilidades é unitária.

Por fim, o principal objetivo da arquitetura *Transformer* é assegurar que o modelo possua a habilidade de focar nos elementos mais relevantes durante o processamento de cada elemento da sequência de entrada (Vaswani *et al.*, 2017). O foco seletivo é realizado pelo

mecanismo de atenção, que permite ao modelo ponderar dinamicamente diferentes partes da entrada.

Mais especificamente, o cálculo da matriz de saída de atenção envolve três componentes fundamentais: Consulta (Q), Chave (K) e Valor (V). A matriz de atenção é determinada ao aplicar a função Softmax sobre o produto escalar entre Q e K, escalado pela raiz quadrada da dimensionalidade da chave (d_k), que permite a ponderação dos valores, conforme expresso na Eq. (10):

$$\text{Atenção}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (10)$$

em que QK^T representa o produto entre a matriz Q e a transposta da matriz K, possibilitando a compatibilidade dimensional para o cálculo.

Para permitir o aprendizado paralelo em diferentes representações contextuais, a arquitetura *Transformer* utiliza o mecanismo MultiAtenção. Nesse mecanismo, o modelo projeta as matrizes *Atenção* (Q, K, V) em diferentes subespaços lineares, nos quais cada “cabeça” de Atenção atua de forma autônoma, capturando padrões específicos de dependência entre os elementos da sequência. Ao final, todas as saídas são concatenadas e processadas por uma camada linear, resultando em uma representação final robusta da entrada.

Apesar da inegável contribuição dos *Transformers* para tarefas de processamento de linguagem natural e visão computacional, sua aplicação para previsão de séries temporais não apresentou resultados tão satisfatórios quanto o esperado. Estudos recentes indicam que, para diversas tarefas de previsão temporal, modelos mais simples, como redes recorrentes ou mesmo modelos lineares, podem superar os *Transformers* em termos de desempenho e estabilidade (Zeng *et al.*, 2023).

Isso se deve, em parte, ao fato de que a arquitetura original não foi projetada para lidar com características específicas das séries temporais, como variáveis exógenas, sazonalidade explícita e dependências multivariadas de longo prazo. Diante dessas limitações, em 2021 foi proposta a arquitetura TFT, projetada especificamente para tarefas de previsão temporal, combinando mecanismos de atenção com codificações temporais explícitas, modelagem de incertezas e suporte a variáveis estáticas e dinâmicas.

3.3 Temporal Fusion Transformer

A estrutura TFT foi proposta inicialmente por Lim *et al.* (2021) como uma solução específica para os desafios da previsão de séries temporais em múltiplos horizontes, focando na sua interpretabilidade e flexibilidade em contextos complexos e multivariados. A ideia desse novo modelo foi integrar os pontos fortes das duas modelagens anteriores: as redes neurais recorrentes, que possuem excelente capacidade de capturar dependências de longo prazo, e os mecanismos de atenção dos *Transformers*, que permitem a seleção dinâmica das variáveis mais relevantes a cada momento. Assim, a TFT oferece uma abordagem unificada e robusta para lidar com séries temporais compostas por variáveis estáticas, variáveis observadas no passado e variáveis conhecidas no futuro.

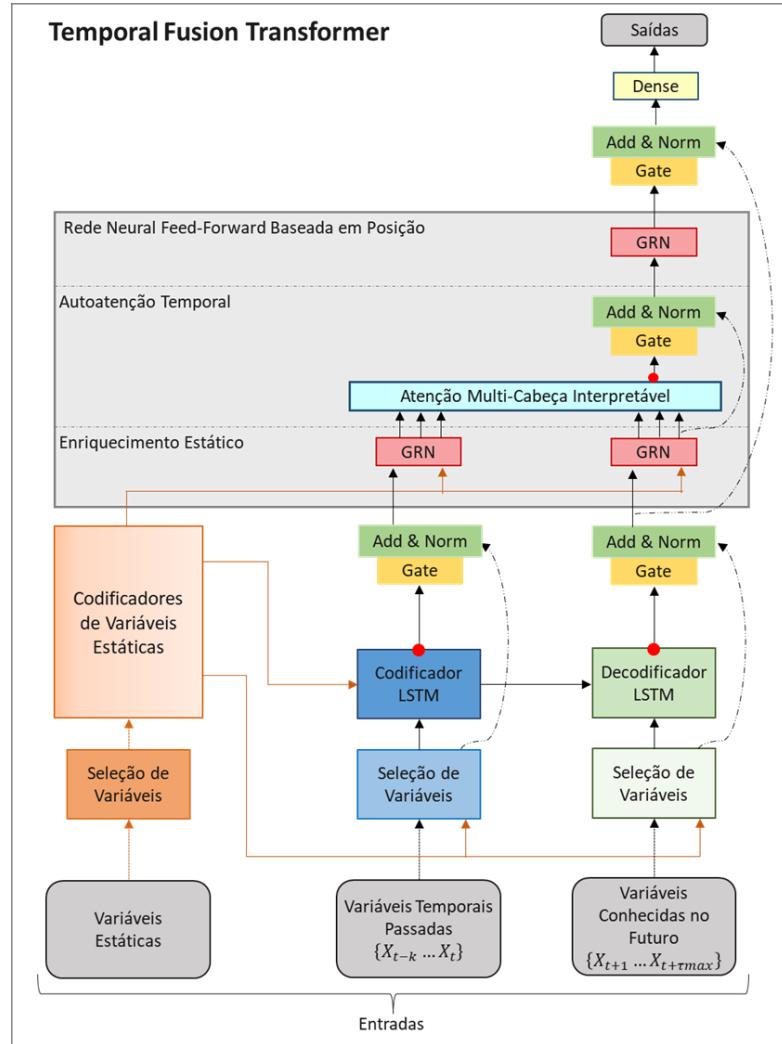
O modelo TFT foi projetado para lidar com múltiplos tipos de entradas, incluindo entradas passadas, entradas futuras e covariáveis estáticas. Ele emprega um mecanismo de controle de acesso para selecionar variáveis de entrada relevantes em cada passo de tempo, garantindo que apenas as informações mais pertinentes sejam usadas para previsões. Tal recurso ajuda a suprimir componentes desnecessários, melhorando seu desempenho.

As entradas do modelo TFT são divididas em três categorias principais: variáveis estáticas, variáveis temporais passadas e variáveis temporais conhecidas no futuro. As variáveis estáticas representam informações invariantes ao longo do tempo, como a sazonalidade ou localização geográfica. Essas variáveis são processadas por codificadores específicos e utilizadas para gerar vetores de contexto que enriquecem as representações temporais ao longo de toda a sequência, fornecendo uma base personalizada para cada série. As variáveis temporais passadas incluem aquelas disponíveis apenas até o tempo presente $\{x_{t-k} \dots x_t\}$ como dados históricos de geração FV, enquanto as variáveis conhecidas no futuro $\{x_{t+1} \dots x_{t+\tau_{max}}\}$ englobam informações previsíveis ou agendadas, como hora do dia, dia da semana, feriados e eventos externos previamente definidos. Ao estruturar e processar essas diferentes categorias de entrada de forma independente e especializada, o TFT consegue capturar relações complexas e aprimorar sua capacidade de generalização em contextos reais e multivariados.

3.3.1 Arquitetura TFT

A arquitetura TFT é composta por diferentes módulos interligados, cada um responsável por processar um tipo específico de entrada ou realizar uma operação fundamental para a previsão de séries temporais multivariadas e multi-horizonte. A Figura 9 ilustra a estrutura TFT e seus principais componentes.

Figura 9 - Arquitetura da Rede TFT



Fonte: Adaptado de Lim *et al.* (2021).

A seguir, são descritos os principais componentes do modelo TFT:

a) Entradas: Variáveis Estáticas

As variáveis estáticas são processadas inicialmente por codificadores específicos, que geram representações vetoriais densas utilizadas ao longo de toda a arquitetura, tanto no processamento das variáveis temporais passadas quanto das conhecidas no futuro. Esses vetores de contexto são incorporados em diversas camadas do modelo, atuando como informações condicionantes globais. Essa estrutura permite que o TFT adapte seu comportamento preditivo às características específicas de cada série temporal, promovendo maior robustez e interpretabilidade.

b) Entradas: Variáveis Temporais Passadas

As variáveis observadas no passado são processadas por codificadores baseados em redes LSTM, responsáveis por capturar dependências temporais de longo prazo. O codificador

concentra-se exclusivamente nos padrões históricos da série. Após essa etapa, a camada “Add & Norm” é aplicada para estabilizar o fluxo de gradiente e acelerar o treinamento, ao normalizar as ativações e mitigar o surgimento de valores extremos, contribuindo para maior estabilidade numérica.

c) Entradas: Variáveis Temporais Conhecidas no Futuro

As variáveis conhecidas no futuro são tratadas separadamente e processadas por decodificadores LSTM, cuja função é gerar a sequência de saída. Essas entradas são fundamentais para capturar padrões sazonais e estruturais com antecedência. Assim como no caso anterior, a camada de normalização “Add & Norm” é utilizada para estabilizar as ativações. Além disso, conexões de salto (representadas na Figura 9 como setas tracejadas) preservam informações originais relevantes, reforçando a influência direta dos atributos de entrada e aumentando a precisão das previsões.

d) Enriquecimento Estático

O bloco de enriquecimento estático tem como função propagar o vetor de contexto gerado a partir das variáveis estáticas para os módulos de variáveis temporais. Esse mecanismo garante que o contexto fixo de cada instância influencie o processamento sequencial da série temporal. A interação entre variáveis estáticas e temporais é mediada por Redes Residuais *Gated* (GRNs), que permitem a seleção dinâmica dos atributos mais relevantes em cada instante de tempo. As GRNs são componentes centrais do TFT, responsáveis por transformar seletivamente os dados, filtrando ruídos e reforçando padrões significativos com alta flexibilidade e estabilidade durante o treinamento.

e) Autoatenção Temporal

A camada de autoatenção interpretável multi-cabeça permite que o modelo concentre sua atenção em diferentes momentos do passado, capturando dependências de longo alcance. Cada “cabeça” da atenção aprende padrões distintos de relevância temporal, tornando possível interpretar quais momentos exerceram maior influência sobre a previsão final. Esse componente é essencial para a transparência do modelo, contribuindo para a compreensão dos fatores que impactaram na análise preditiva.

f) Rede Neural *Feed-Forward* Baseada em Posição

Esta camada atua como um dos estágios finais do TFT e é responsável por refinar as representações enriquecidas ao longo do tempo. Diferente das redes recorrentes, a rede *Feed-Forward* baseada em posição aplica a mesma transformação não linear de forma paralela a cada instante de tempo. O termo “baseada em posição” indica que as informações temporais são previamente preservadas por vetorizações e mecanismos de atenção. Essa arquitetura

permite o paralelismo computacional e reduz o custo de treinamento. Além disso, utiliza camadas GRN adicionais para garantir estabilidade e assegurar que os padrões aprendidos sejam adequadamente transformados em previsões robustas.

g) Camada *Dense*

Após a integração de todas as informações temporais e contextuais, os vetores resultantes são encaminhados para uma camada *Dense* (totalmente conectada), que tem como função gerar as previsões para cada horizonte temporal futuro. Essa camada pode ser adaptada para produzir estimativas pontuais ou distribuições probabilísticas, dependendo da abordagem adotada para modelagem da incerteza.

h) Saídas

A saída final do modelo corresponde aos valores previstos da variável-alvo ao longo do horizonte futuro definido. A modularidade da arquitetura TFT garante que as previsões sejam não apenas precisas, mas também interpretáveis, pois é possível rastrear a contribuição de cada variável e momento para o resultado.

3.3.2 Rede de Seleção de Variáveis

Esse mecanismo tem a função de identificar dinamicamente, em cada passo temporal, quais variáveis são mais informativas para o modelo, sendo aplicado tanto às covariáveis estáticas quanto às covariáveis dependentes do tempo (passadas e futuras). Além de oferecer interpretação sobre a importância de cada entrada, esse mecanismo tem um papel crucial em eliminar entradas irrelevantes ou ruidosas que poderiam comprometer o desempenho do modelo. Em cenários reais, conjuntos de dados frequentemente incluem atributos com baixo poder preditivo, e a capacidade do TFT de direcionar seletivamente a atenção às variáveis mais relevantes permite melhor utilização da capacidade de aprendizado e melhoria na generalização (Lim *et al.*, 2021).

Para representar as variáveis categóricas, o TFT utiliza processos codificação vetorial, que transformam cada categoria em um vetor contínuo em um espaço de dimensão fixa. No caso das variáveis dependentes do tempo são aplicadas transformações lineares. Assim, cada entrada é convertida em um vetor de dimensão d_m , correspondente ao tamanho das representações internas do modelo. As entradas são organizadas em três grupos distintos (estáticas, temporais passadas e futuras), cada um utilizando sua própria rede de seleção de variáveis (Ferreira; Leite, 2025).

A cada passo de tempo t , as entradas transformadas são combinadas linearmente para formar um vetor de entradas achatado. Esse vetor é então submetido a uma GRN e a uma

camada da função Softmax, juntamente com o vetor de contexto obtido por meio da codificação das covariáveis estáticas. No caso de variáveis estáticas, o vetor de contexto é omitido, uma vez que essas já contêm a informação contextual necessária.

Cada variável possui sua própria GRN, com pesos compartilhados entre todos os instantes de tempo. Assim, cada vetor transformado é processado individualmente por sua própria GRN, resultando em $\tilde{\xi}_t^{(j)}$, o vetor de características processadas para a variável j . Finalmente, todas os vetores processados são ponderados pelos respectivos pesos de seleção de variáveis $v_t^{(j)}$ e combinados, formando o vetor final $\tilde{\xi}_t$, conforme Eq. (11):

$$\tilde{\xi}_t = \sum_{j=1}^m v_t^{(j)} \tilde{\xi}_t^{(j)} \quad (11)$$

em que m representa o número total de variáveis do modelo.

Esse processo permite que, a cada instante temporal, o TFT selecione de forma adaptável os atributos mais relevantes, atribuindo um grau de influência individual a cada entrada no momento da previsão. Tal capacidade não só melhora o desempenho preditivo como também confere transparência e interpretabilidade, facilitando a análise dos fatores mais impactantes no processo decisório do modelo e possibilitando ajustes estratégicos no uso de dados de entrada.

3.3.3 Redes GRNs

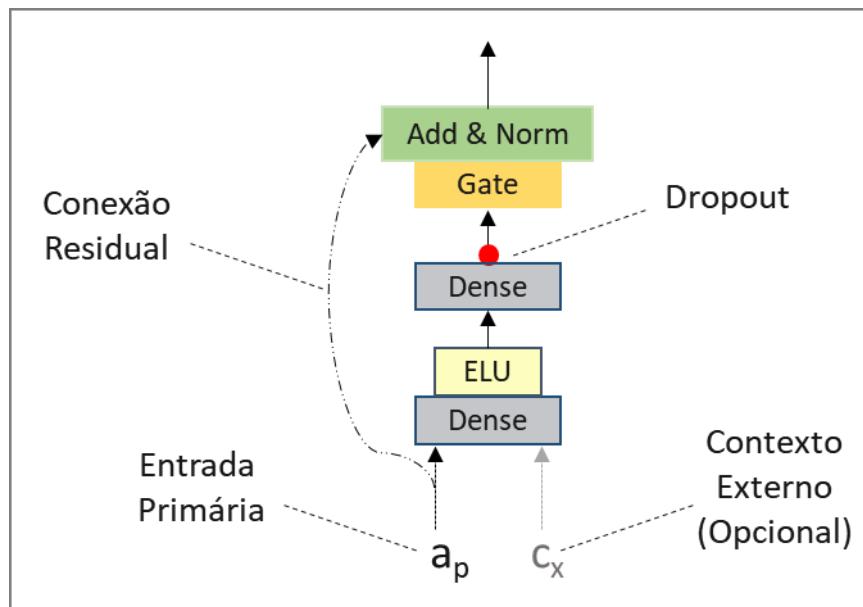
A GRN constitui uma sub-rede *FeedForward* composta por camadas totalmente conectadas, ativações não lineares e mecanismos de *gating* (portas). Estes mecanismos permitem que o modelo controle dinamicamente a quantidade de informação a ser propagada adiante. O componente de *gating* determina quanto da saída transformada e quanto da entrada original serão preservados, o que estabiliza o treinamento e possibilita o aprendizado de relações complexas sem a perda de informações úteis. Adicionalmente, a GRN implementa uma conexão residual ponderada, capacitando o modelo a aprender a combinar efetivamente a entrada com a saída da transformação, ou, se necessário, a ignorar a transformação quando esta não for útil.

É comum que a relação exata entre variáveis de entrada (exógenas) e as variáveis-alvo seja desconhecida, dificultando a identificação das entradas mais relevantes. Além disso, nem sempre é claro o quão complexo deve ser o processamento não linear de um modelo, e em certos casos (como com poucos dados ou dados ruidosos), modelos mais simples podem ser

mais eficazes. Para dar ao modelo a flexibilidade de aplicar processamento não linear apenas quando necessário, os autores propuseram a GRN como um bloco de construção fundamental da topologia TFT (Lim *et al.*, 2021).

As GRNs permitem que o TFT aplique processamento não linear de forma seletiva, garantindo uma adaptação ideal a diversos conjuntos de dados e cenários. Essas camadas utilizam o bloqueio de componentes para eliminar blocos desnecessários na arquitetura, ajustando-se à complexidade da aplicação. A Figura 10 ilustra a estrutura de uma GRN. Ela recebe duas entradas: uma entrada primária (**a**) e um vetor de contexto opcional (**c**).

Figura 10 – Estrutura da Rede GRN



Fonte: Adaptado de Lim *et al.* (2021).

O processo da GRN ocorre conforme descrito nas Eq. (12) – (14):

$$GRN_{\omega} (\mathbf{a}_p, \mathbf{c}_x) = LayerNorm \left(\mathbf{a}_p + GLU_{\omega}(\eta_1) \right) \quad (12)$$

$$\eta_1 = W_{1,\omega} \eta_2 + b_{1,\omega} \quad (13)$$

$$\eta_2 = ELU(W_{2,\omega} \mathbf{a}_p + W_{3,\omega} \mathbf{c}_x + b_{2,\omega}) \quad (14)$$

em que \mathbf{a}_p é o vetor da entrada primária, \mathbf{c}_x é o vetor de contexto opcional, ω é um índice que denota o compartilhamento de peso, LayerNorm é uma camada de normalização *Standard*, a Unidade Linear *Gated* (GLU) é uma camada que controla o fluxo de transformações não lineares com capacidade para suprimir qualquer parte da arquitetura que não é requerida para um dado conjunto de dados, η_1 e η_2 representam camadas intermediárias, $W_{(.)}$ são os pesos,

ELU é uma função de ativação que age como função identidade quando $W_{2,\omega} \mathbf{a}_p + W_{3,\omega} \mathbf{c}_x + b_{2,\omega} \gg 0$ e como uma constante quando ocorre o contrário e $b_{(\cdot)}$ representa o bias.

A conexão residual, conforme expresso na Figura 10, funciona como um atalho, preservando informações e facilitando o treinamento das redes profundas ao mitigar o problema do gradiente. O *Dropout* atua como regularização, desativando neurônios aleatoriamente e aprimorando a generalização do modelo, antes da etapa final da GRN. O mecanismo de *Gate*, representado por uma porta sigmoide, controla o fluxo da informação transformada, ponderando de forma adaptativa o quanto da saída não linear (que vem de camadas densas com ativação ELU) deve ser retido.

Por fim, a GRN também permite a injeção de contexto externo opcional (\mathbf{c}_x) para condicionar a previsão. Essa entrada auxiliar pode representar, por exemplo, variáveis estáticas categóricas, sendo incorporada à transformação por meio de uma segunda camada densa, o que favorece a condicionalidade da previsão ao contexto específico da série temporal. Essa combinação de transformações não lineares controladas, preservação de sinal e condicionamento contextual torna as GRNs essenciais para a expressividade e interpretabilidade do TFT.

3.3.4 Função Preditiva

A função preditiva do TFT, conforme proposta por Lim *et al.* (2021), que integra diversos tipos de entrada (temporal e estática) para a estimativa probabilística multi-horizonte, é expressa conforme a Eq. (15):

$$y_i(q, t, \tau) = F_q\{\tau, y_i(t - k:t), Z_i(t - k:t), X_i(t:t + \tau), S_i\} \quad (15)$$

Horizonte de Previsão
 ↑
 $y_i(q, t, \tau)$
 ↓ Previsões
 Função TFT
 Histórico Temporal Variável Alvo
 Covariáveis Passadas
 ↓
 Covariáveis Conhecidas no Futuro
 Covariáveis Estáticas

em que $y_i(q, t, \tau)$ representa o q -ésimo percentil da variável-alvo para a série i , previsto para o instante $t + \tau$, e F_q é a função de previsão implementada pelo modelo. Essa função leva em consideração:

- O histórico temporal da variável-alvo $y_i(t - k: t)$;
- As covariáveis passadas $Z_i(t - k: t)$, observadas até o instante t ;
- As covariáveis conhecidas no futuro $X_i(t: t + \tau)$, disponíveis para o período de previsão;
- E as covariáveis estáticas S_i que são invariantes no tempo.

3.3.5 Interpretabilidade do modelo TFT

A atenção multi-cabeça, conforme originalmente proposta no modelo *Transformer* (ver Eq. (10)), utiliza múltiplos mecanismos de atenção para processar um mesmo conjunto de dados de entrada. Essa abordagem permite analisar as informações em diferentes subespaços e posições, facilitando a identificação de padrões relevantes e de longo prazo. No entanto, na modelagem TFT, esse processo é modificado para torná-lo interpretável.

Em vez de concatenar as saídas das diferentes cabeças de atenção, os resultados são calculados por média e, em seguida, submetidos a uma transformação linear. Além disso, os pesos de Valor W_V são compartilhados entre todas as cabeças, o que aumenta a interpretabilidade e reduz a complexidade computacional.

Matematicamente, a função de atenção multi-cabeça interpretável (MCI) é definida pela Eq. (16):

$$MCI(Q, K, V) = \tilde{H}W_H \quad (16)$$

em que W_H representa o mapeamento linear final, e \tilde{H} é a média das saídas de atenção de cada cabeça, calculada de acordo com a Eq. (17):

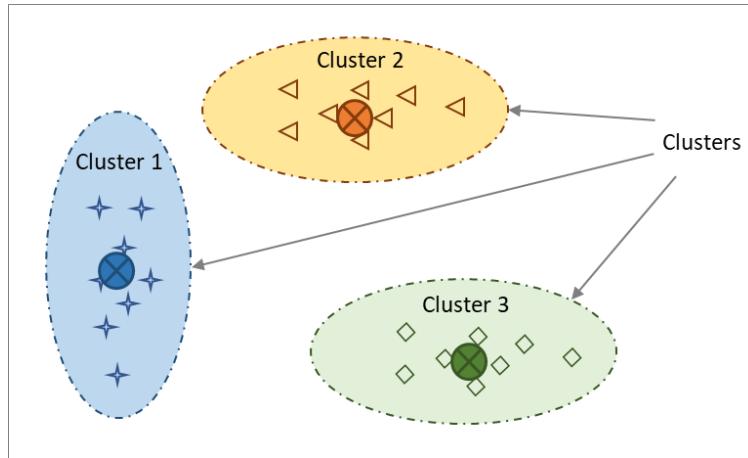
$$\tilde{H} = \frac{1}{m_H} \sum_{h=1}^H A(QW_Q^{(h')}, KW_K^{(h')})VW_V \quad (17)$$

em que A representa a função de atenção escalar, e Q , K e V representam Consulta, Chave e Valor, conforme explicado anteriormente. O peso W_V é compartilhado em todas as cabeças, enquanto $W_Q^{(h')}$ e $W_K^{(h')}$, são os pesos específicos de Consulta e Chave, respectivamente, para cada cabeça h' . Por fim, m_H e h' representam o número total de cabeças e o índice de cada cabeça, respectivamente.

3.4 Clusterização

A clusterização consiste em uma técnica de aprendizado não supervisionado que agrupa dados semelhantes em conjuntos (clusters). Em algoritmos baseados em centróides, como o *k-means*, cada cluster é representado por um centróide, normalmente localizado próximo ao “centro” do conjunto de pontos, e cada elemento da base de dados é associado ao cluster cujo centróide é o mais próximo, com base em uma medida de similaridade (por exemplo, distância). A Figura 11 demonstra, de forma simplificada, esse processo, enfatizando três clusters, seus centróides e a proximidade dos dados em cada grupo.

Figura 11 - Algoritmo de Clusterização *k-means*



Fonte: A Autora (2025).

Uma função objetivo pode ser utilizada para expressar formalmente o problema de clusterização. Considere um conjunto de dados $X = \{x_1, x_2, \dots, x_n\} \subset \mathbb{R}^d$, que deve ser particionado em k clusters. No Algoritmo *k-means* tradicional, cada cluster é representado por um centróide c_i , dado pela média dos pontos atribuídos a esse grupo. A função objetivo f minimiza a soma do quadrado da distância Euclidiana de cada ponto ao centróide do cluster ao qual pertence, conforme Eq. (18) (Tan; Steinbach; Karpatne; Kumar, 2019):

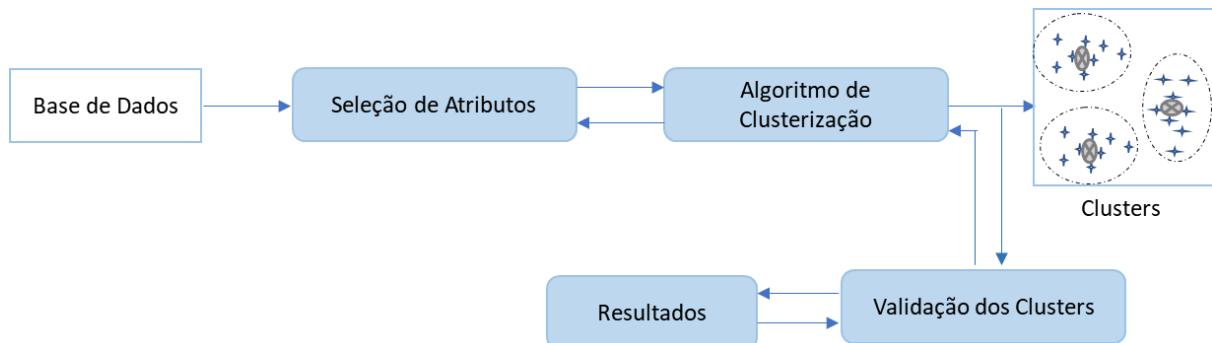
$$f = \sum_{i=1}^k \sum_{x \in C_i} \text{dist}(x, c_i)^2 \quad (18)$$

em que $\text{dist}(x, c_i)$ representa a distância Euclidiana entre o ponto x e o seu centroide c_i , k é o número total de clusters, C_i representa o i -ésimo cluster e c_i é o centroide associado a esse cluster.

Essa técnica é amplamente utilizada em uma variedade de áreas, em casos de regressão, classificação, análise de imagens, entre outras, devido principalmente à sua simplicidade e eficácia. No entanto, tem desvantagens como a sensibilidade à inicialização dos centroides, o que pode resultar na convergência para um mínimo local, para minimizar isso, é usual executar o algoritmo várias vezes com diferentes inicializações e escolher o melhor resultado. O método não apresenta bons resultados quando os clusters não são globulares e possui alta sensibilidade a *outliers* ou valores discrepantes (Wu, 2012).

A implementação computacional da clusterização basicamente ocorre em quatro etapas: 1) seleção de atributos que consiste na definição dos atributos da base de dados que serão agrupados; 2) implementação do algoritmo e realização de diversas iterações para garantia do ponto mínimo das distâncias; 3) avaliação do desempenho do agrupamento e 4) interpretação dos resultados (Xu; Wunsch, 2008). A Figura 12 sintetiza o fluxograma dessa técnica.

Figura 12 - Diagrama Esquemático da Clusterização



Fonte: Adaptado de Xu; Wunsch (2008).

3.4.1 Clusterização de Séries Temporais

A clusterização de séries temporais é realizada avaliando a similaridade entre elas, o que envolve medir a distância entre as séries para identificar padrões ou agrupamentos baseados em suas características.

O procedimento refere-se ao processo de agrupar séries temporais que compartilham padrões ou características semelhantes ao longo do tempo. É uma técnica amplamente usada em diversas áreas, como previsão, análise financeira, saúde, entre outras. Dada sua característica, em vez de se concentrar em variáveis isoladas, a clusterização de séries temporais considera as sequências de dados em seu conjunto, levando em conta a dependência temporal e a correlação entre os pontos da série. A clusterização *k-means*, não é particularmente eficaz para séries temporais devido à sua dependência da distância Euclidiana, que não se aplica a esse caso.

O objetivo principal da clusterização de séries temporais consiste em agrupar séries que evoluem de forma semelhante ao longo do tempo, independentemente das diferenças nos valores absolutos. Assim, características como médias ou variâncias são substituídas por suas dinâmicas temporais.

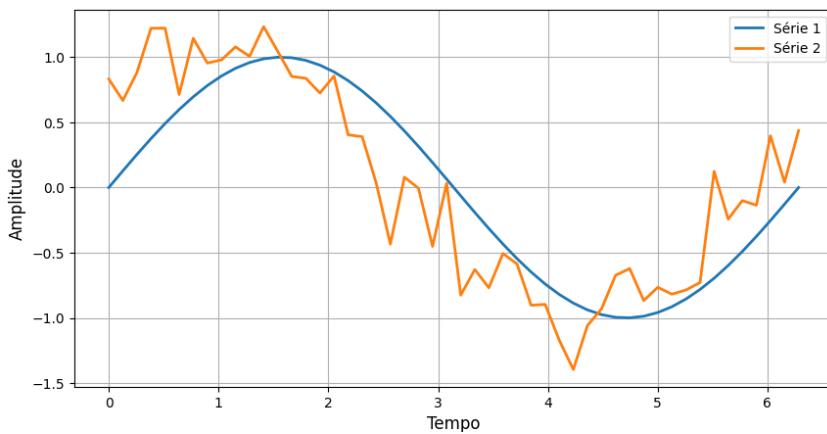
Os desafios incluem: a variabilidade de tempo, pois as séries temporais podem ter comprimentos diferentes, apresentando desafios em termos de comparação direta. O deslocamento temporal, considerando que os padrões temporais podem estar desfasados, ou seja, os máximos e mínimos podem ocorrer em momentos diferentes nas séries. O ruído ou variação aleatória que ocorre muitas vezes e que pode dificultar a identificação de padrões reais. E, por fim, a não linearidade, em que as séries temporais podem apresentar relações complexas e não lineares.

Uma parte crucial da clusterização de séries temporais é definir uma métrica de similaridade ou distância entre as séries temporais. Embora a Distância Euclidiana seja simples e fácil de calcular, não é adequada para séries com diferentes escalas de tempo ou variações de fase. Assim, surge o conceito da Derivação Temporal Dinâmica (DTW), que é uma técnica utilizada para comparar sequências de pontos de dados que podem variar em comprimento ou velocidade. É particularmente útil na análise de dados de séries temporais, onde duas sequências podem ter padrões semelhantes, mas diferem em termos de temporização ou velocidade (Sakoe; Chiba, 1978).

A DTW encontra um alinhamento ótimo entre duas sequências, distorcendo o eixo do tempo para minimizar a diferença entre elas. Isso permite a comparação de sequências que podem ter comprimentos diferentes ou podem estar ligeiramente fora de sincronia.

Para ilustrar essa diferença, a Figura 13 retrata duas séries temporais aleatórias (Série 1 e Série 2), as quais possuem formatos, picos, vales e frequências distintos entre si.

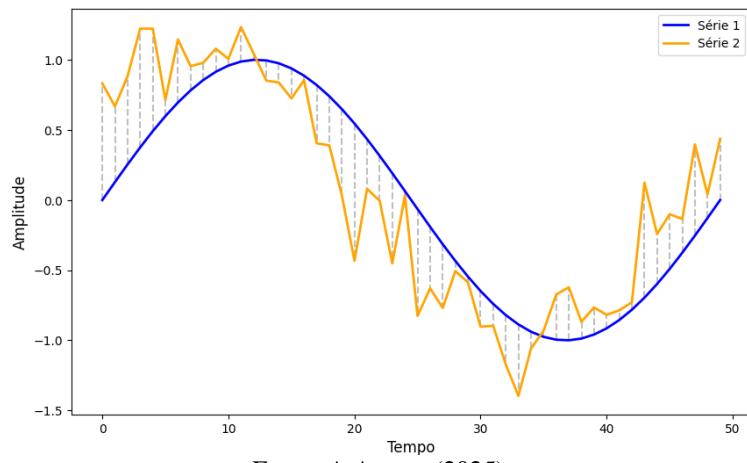
Figura 13 - Séries Temporais Comparadas



Fonte: A Autora (2025).

A Distância Euclidiana é representada pela linha tracejada na Figura 14 e a Distância DTW pela linha contínua na Figura 15. É possível perceber que a primeira calcula a distância entre as séries ponto a ponto, enquanto a segunda realiza um alinhamento entre as séries de forma não linear para encontrar a correspondência ótima entre seus pontos. Em vez de comparar apenas os pontos correspondentes no tempo (como ocorre no caso da Distância Euclidiana), a DTW ajusta o alinhamento das séries para minimizar a diferença acumulada entre elas.

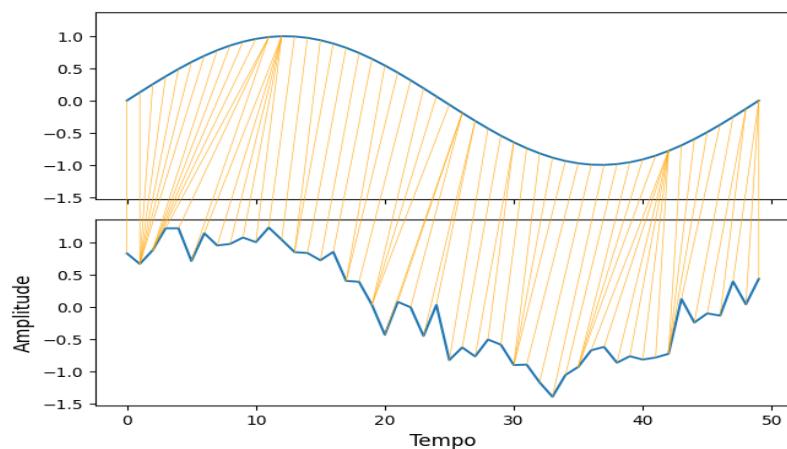
Figura 14 - Diferenças com Distância Euclidiana



Fonte: A Autora (2025).

A DTW ainda permite que um ponto de uma série seja comparado com múltiplos pontos da outra série, conforme pode ser observado na Figura 15. Isso a torna robusta a mudanças na escala temporal e a ruídos, que é o caso da Série 2. Assim, é possível afirmar que a DTW tem maior flexibilidade para lidar com deformações temporais sendo aplicável para análises de séries temporais em que o alinhamento temporal contém variação.

Figura 15 - Diferenças com Distâncias DTW



Fonte: A Autora (2025).

Para a expressão de cálculo da DTW, considerando duas séries temporais $M = [m_1, m_2, \dots, m_p]$ com p elementos e $N = [n_1, n_2, \dots, n_q]$ com q elementos, o alinhamento ótimo entre os elementos de M e N minimiza a distância acumulada entre elas. O objetivo da métrica consiste no cálculo da menor "distância de deformação" entre as duas séries, considerando que elementos de uma série podem ser comparados com múltiplos elementos da outra (Müller, 2007).

Primeiramente é construída uma matriz $p \times q$ onde cada elemento (v_{ij}) representa o custo local, ou diferença absoluta quadrática, do ponto m_i da série M com o ponto n_j da série N , conforme Eq. (19):

$$v_{ij} = |m_i - n_j|^2 \quad (19)$$

em que m_i representa um ponto da série M , com i variando de 1 a p e n_j representa um ponto da série N com j variando de 1 a q .

Em seguida, calcula-se a matriz DTW que calcula o custo acumulado ótimo para alinhar as duas séries até cada ponto (i, j) , utilizando uma programação dinâmica calculada recursivamente para minimizar o custo total de alinhamento, conforme Eq. (20).

$$C(i, j) = v(i, j) + \min\{C_{i-1, j-1}, C_{i-1, j}, C_{i, j-1}\} \quad (20)$$

em que $C(i, j)$ representa o elemento da matriz de custo acumulado ou de menor distância entre o alinhamento das duas séries M e N , considerando o custo local $v(i, j)$.

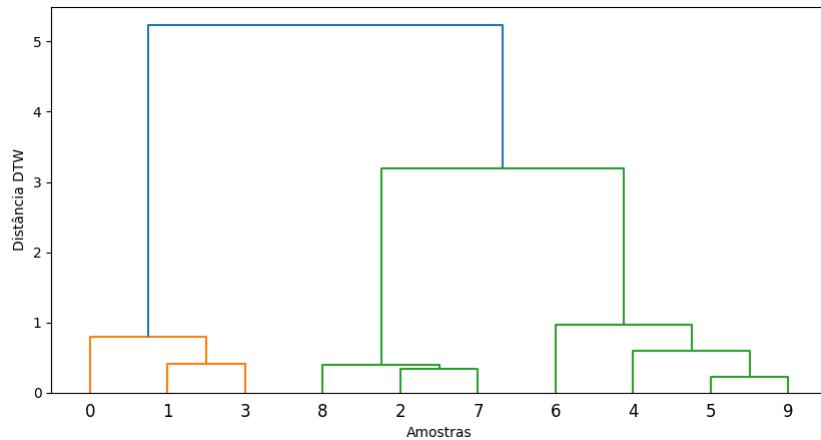
Após a determinação da matriz C , o menor caminho entre as duas séries é uma sequência de índices que define como as duas séries estão alinhadas, começando do último elemento das séries $C(p, q)$ e terminando em $C(1, 1)$, minimizando o custo total. Assim, a distância DTW final é o custo acumulado até o último ponto (Müller, 2007).

Quanto ao algoritmo, a clusterização hierárquica quando adaptada com a métrica específica DTW, torna-se uma ferramenta eficaz para identificar padrões em séries temporais, especialmente em contextos em que se deseja explorar relações estruturais e padrões de similaridade entre dados que variam ao longo do tempo, pois, dada sua característica, uma estrutura interpretável que permite identificar agrupamentos em diferentes níveis de hierarquias (Liao, 2005). A clusterização hierárquica constrói uma árvore de clusters e permite que diferentes níveis de agrupamento sejam analisados.

A aplicação da clusterização hierárquica requer etapas específicas, a primeira fase consiste no Pré-Processamento de Dados que se caracteriza na normalização das séries temporais, sendo essencial para evitar que diferenças de escala influenciem na análise dos resultados, e na suavização ou remoção de tendências para melhorar a qualidade do agrupamento. Em seguida a métrica DTW é então aplicada para o cálculo da distância entre as séries temporais. O agrupamento iterativo identifica os dois clusters mais próximos com base na DTW e combina esses clusters em um único cluster; o processo é repetido recursivamente até que todos os clusters estejam agrupados.

Por fim, o Dendrograma representa graficamente o processo de agrupamento, mostrando como os clusters são combinados em diferentes níveis de distância. A Figura 16 exemplifica um dendrograma para dez amostras agrupadas em níveis hierárquicos com base na distância DTW. Observa-se, por exemplo, que os pares (1, 3), (2, 7) e (5, 9) são unidos nos níveis de distância mais baixos, indicando maior similaridade entre esses elementos dentro de seus respectivos clusters.

Figura 16 – Exemplo de Dendrograma na Clusterização Hierárquica



Fonte: A Autora (2025).

3.5 Conceitos Adicionais

3.5.1 Normalização dos Dados

A presença de picos e componentes não estacionários nos dados de entrada (como informações meteorológicas e histórico de geração) prejudica o treinamento de modelos de previsão, elevando os erros. Isso ocorre porque essas variáveis são imprevisíveis e flutuantes devido às condições climáticas. Assim, o pré-processamento desses dados é crucial para reduzir problemas de treinamento, otimizar o custo computacional e melhorar significativamente a precisão da modelagem (Das *et al.*, 2018).

Dados de energia FV costumam ter variáveis em escalas muito diferentes, o que reflete sua grande variabilidade. Essa diferença de escalas pode levar o modelo a dar prioridade indevida a certas informações, prejudicando o desempenho do algoritmo durante o treinamento. Por isso, a normalização dos dados é essencial para equalizar todos os valores e garantir um treinamento mais eficaz. Entre as principais técnicas de normalizar pode-se citar o Escalonamento MinMax, Escalonamento *Z-Score* ou *Standard* e Escalonamento Robusto.

a) Escalonamento MinMax

Este método, formalmente expresso na Eq. (21), aplica uma transformação linear que reescalas os dados para o intervalo [0,1], preservando a ordem relativa das observações, mas alterando sua escala e amplitude. Como o MinMax depende diretamente dos valores mínimo e máximo observados no conjunto de treinamento, ele é mais indicado quando as variáveis possuem limites bem definidos e baixa incidência de valores discrepantes (*outliers*); caso contrário, esses extremos podem comprimir a maior parte dos dados e distorcer o reescalonamento.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (21)$$

em que x_{min} e x_{max} representam, respectivamente, os valores máximos e mínimos da variável x , e x_{norm} a variável normalizada.

b) Escalonamento *Standard*

Este método de normalização, expresso pela Eq. (22), envolve a transformação das variáveis pela subtração de suas respectivas médias, seguida da divisão pelo desvio padrão. Ao contrário do escalonamento MinMax, este método não impõe um limite definido aos valores resultantes. Sua aplicação é particularmente vantajosa em cenários nos quais as variáveis exibem uma distribuição aproximadamente normal. Contudo, é importante notar que o Escalonamento *Standard* é sensível à presença de outliers, que podem distorcer os valores da média e do desvio padrão, afetando a transformação.

$$x_{norm} = \frac{x - \mu}{\bar{\sigma}} \quad (22)$$

em que μ e $\bar{\sigma}$ representam, respectivamente, a média e o desvio padrão da variável x , e x_{norm} a variável normalizada.

c) Escalonamento Robusto

Este método de normalização, expresso pela Eq. (23), distingue-se dos anteriores por sua menor sensibilidade à influência de outliers. Esta técnica utiliza a mediana e o intervalo

interquartil (IQR), o qual se constitui na diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), para realizar a transformação. Essa característica confere ao Escalonamento Robusto uma vantagem significativa, tornando-o uma escolha ideal em cenários onde os dados contêm uma quantidade considerável de valores discrepantes.

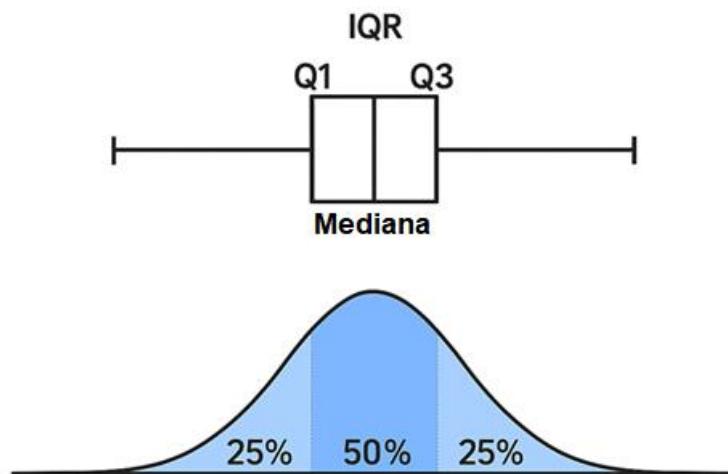
$$x_{norm} = \frac{x - med}{IQR} \quad (23)$$

em que med e IQR representam, respectivamente, a mediana e a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) da distribuição da variável x , e x_{norm} a variável normalizada.

A Figura 17 sintetiza, em duas representações complementares, a dispersão central dos dados. Considerando que o quartil consiste na divisão dos dados em quatro partes, no *boxplot* (parte superior), o primeiro quartil (Q1) e o terceiro quartil (Q3) delimitam o intervalo que contém 50% dos valores intermediários. A amplitude desse intervalo é dada pelo Intervalo Interquartil (IQR = Q3 – Q1), uma métrica robusta à presença de outliers. A mediana, por sua vez, situa-se no centro desse bloco, dividindo a amostra em duas metades.

Complementarmente, na curva inferior, a mesma ideia é expressa em termos de área sob uma distribuição aproximadamente normal: 25% dos dados estão abaixo de Q1, 50% concentram-se entre Q1 e Q3 e 25% acima de Q3. Isso evidencia que o IQR corresponde à região de maior densidade observacional e, portanto, constitui um indicador de variabilidade mais resiliente do que medidas baseadas na média e no desvio padrão.

Figura 17 - Distribuição de dados em quartis



Fonte: A Autora (2025).

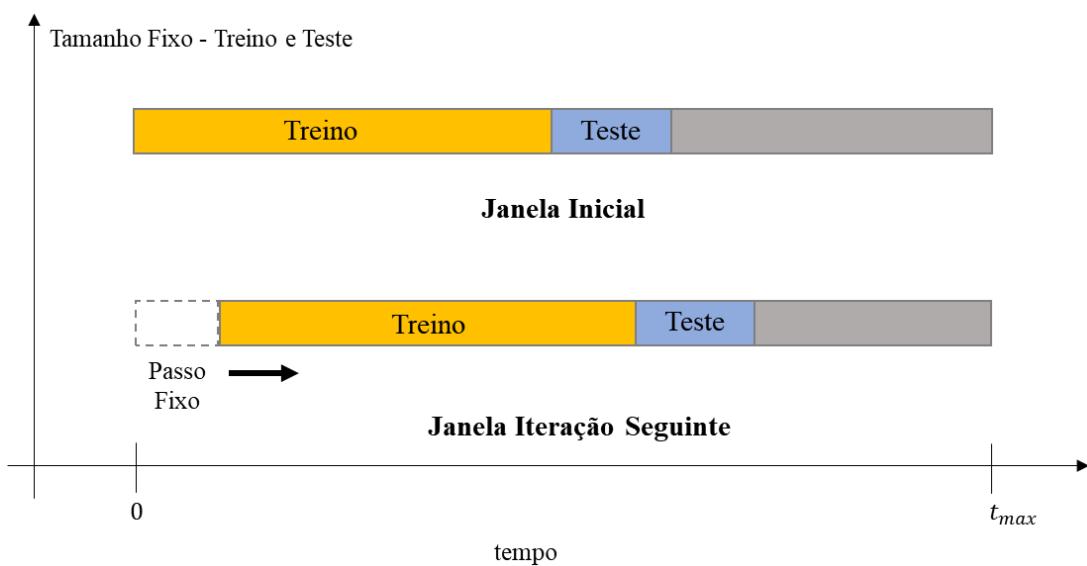
3.5.2 Validação com Janela Deslizante

A validação com janela deslizante é uma técnica robusta e amplamente utilizada na literatura para avaliar o desempenho de modelos de séries temporais. Em vez de uma única divisão estática entre dados de treino e teste, como na validação cruzada tradicional, a janela deslizante simula um cenário mais realista de previsão contínua, onde o modelo é avaliado sequencialmente no tempo em toda a amostra de dados.

Esse é um procedimento de estimação de desempenho para séries temporais em que o conjunto de treino “desliza” ao longo do tempo, preservando a ordem temporal e evitando vazamento de dados. O processo começa com uma janela inicial de treinamento $[t_1, \dots, t_k]$. O modelo é ajustado com esses dados e, em seguida, testado no bloco de tempo imediatamente seguinte $[t_{k+1}, \dots, t_{k+h}]$. Em seguida, a janela é deslocada (por expansão ou tamanho fixo), o modelo é reajustado e um novo erro de previsão é calculado. Ao final, os erros de todas as janelas são combinados (por meio de média, mediana etc.), resultando em uma avaliação de desempenho mais estável e realista para dados fora da amostra. Essa estratégia é recomendada para problemas de previsão porque respeita a dependência temporal, diferindo da validação cruzada aleatória tradicional, inadequada para dados autocorrelacionados já que misturaria informações futuras com passadas, comprometendo a validade da avaliação. (Hota; Handa; Shrivastava, 2017).

A Figura 18 detalha o processo de validação com janela deslizante em um processo de análise de desempenho de série temporal.

Figura 18 - Janela deslizante



Fonte: A Autora (2025).

- **Conclusão do Capítulo**

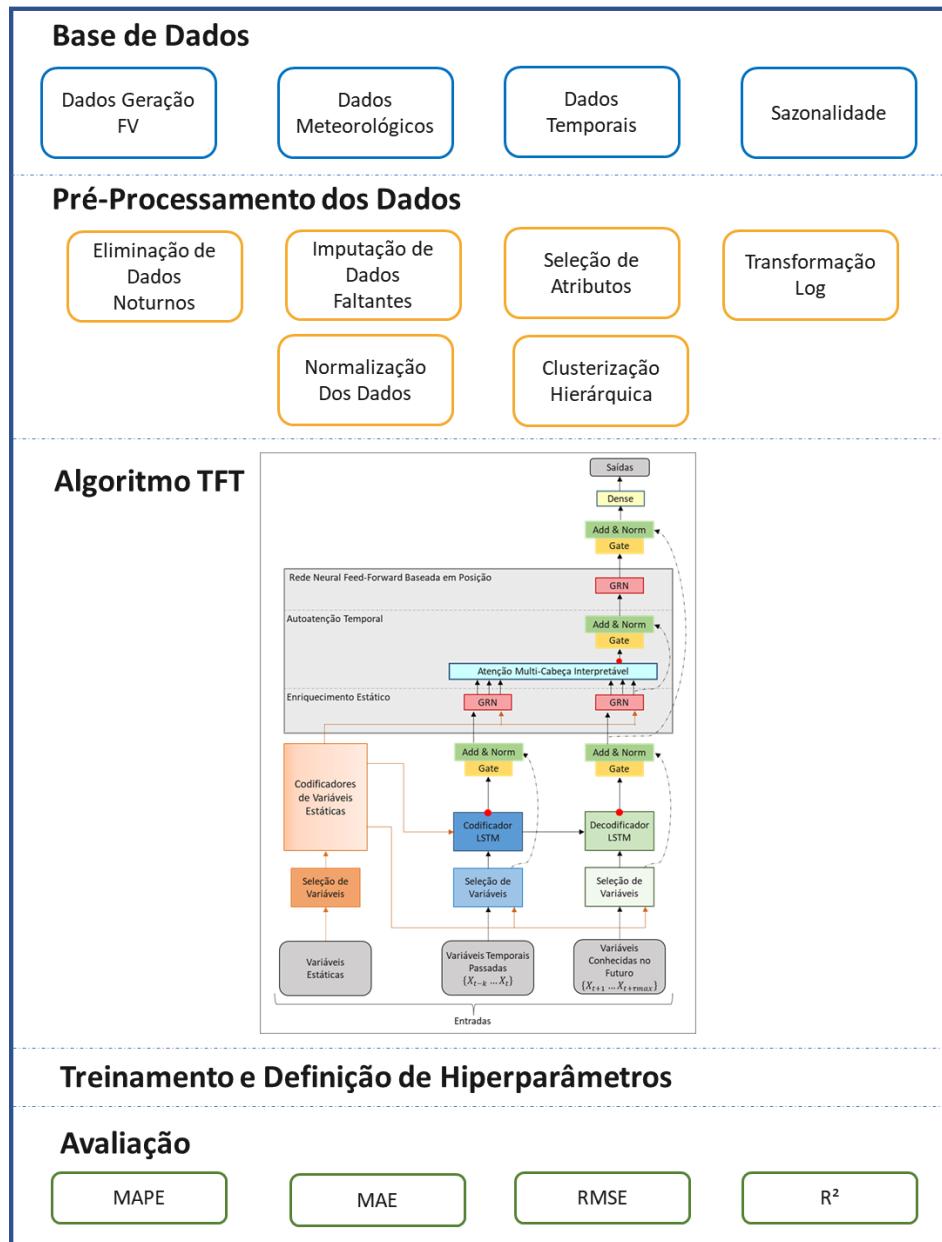
Este capítulo apresentou os fundamentos teóricos que embasam a proposta metodológica desta tese. Foram abordados inicialmente os conceitos de Inteligência Artificial e redes neurais, com ênfase nas arquiteturas MLP e LSTM, amplamente utilizadas em tarefas de previsão. Em seguida, foi discutida a arquitetura *Transformer*, que serve de base para o modelo central deste trabalho, o TFT. Detalhou-se sua estrutura interna, incluindo os mecanismos de seleção de variáveis, as redes GRNs, a função preditiva e os recursos que conferem interpretabilidade ao modelo. Complementarmente, foram abordados métodos de clusterização hierárquica e conceitos auxiliares essenciais à modelagem de séries temporais, como a normalização dos dados e a validação com janela deslizante. A fundamentação apresentada neste capítulo oferece o embasamento teórico necessário para o delineamento da metodologia e a realização dos experimentos descritos nos capítulos seguintes.

4 METODOLOGIA

Neste capítulo, serão abordados com maior detalhamento os seguintes aspectos: a estrutura da base de dados, as técnicas utilizadas na etapa de pré-processamento dos dados, a topologia de rede e a forma escolhida para avaliação do desempenho dos modelos desenvolvidos.

A Figura 19 apresenta a estrutura proposta para o estudo, resumindo as etapas a serem consideradas. Cada item dessa estrutura será abordado com maior detalhamento, individualmente, nas próximas seções deste capítulo.

Figura 19 - Metodologia da Pesquisa



Fonte: A Autora (2025).

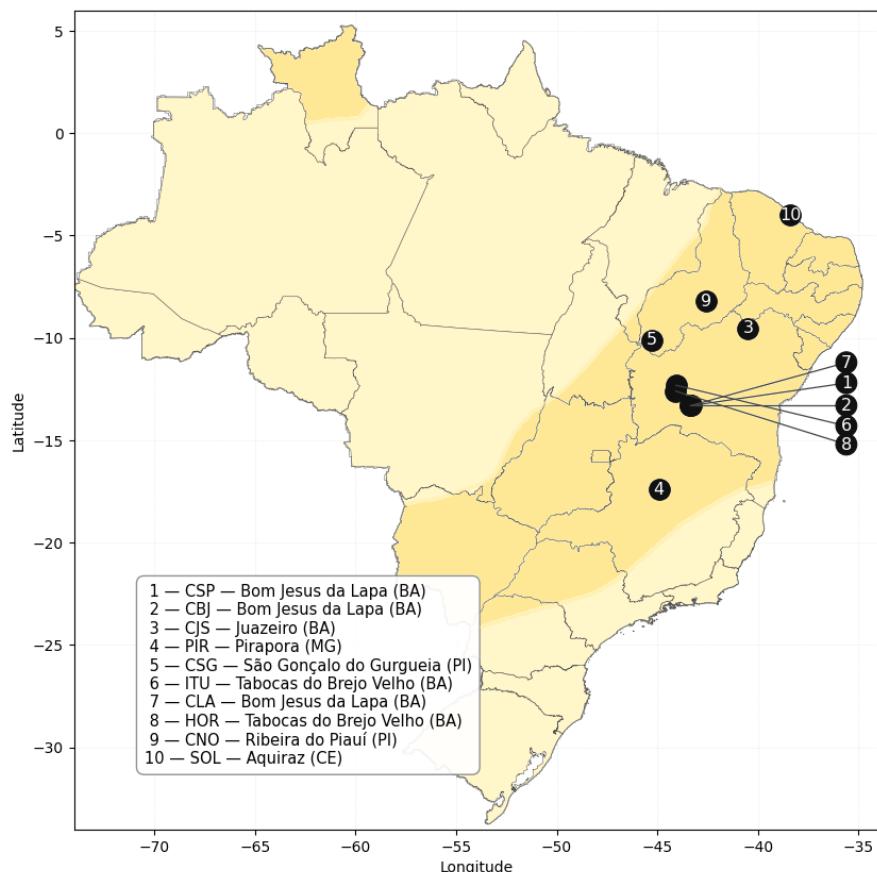
Todos os códigos foram desenvolvidos no Google Colab®, utilizando a linguagem de programação Python. O ambiente de execução foi um *desktop* com sistema operacional Windows 11, processador Intel Core i7 de 8^a geração, 16 GB de memória RAM e uma placa de vídeo Nvidia GeForce com 2 GB de memória dedicada e GPU T4. Os modelos foram criados principalmente utilizando as bibliotecas *Pandas*, *Numpy*, *Scikit-learn* e *PyTorch Forecasting*.

4.1 Base de Dados

4.1.1 Dados de Geração FV

As usinas de geração FV utilizadas nesta pesquisa foram coletadas no domínio público da base de Históricos de Operação do ONS. Foram selecionadas dez usinas FV estrategicamente distribuídas em regiões de elevado potencial solar, conhecidas como ‘cinturão solar brasileiro’ (Pereira *et al.*, 2017). Esse cinturão estende-se principalmente pelo semiárido nordestino e por trechos das regiões Centro-Oeste e Sudeste, apresentando uma irradiação solar média diária superior à das demais regiões brasileiras. A Figura 20 destaca a posição geográfica das dez usinas FV no território brasileiro e suas posições no cinturão solar brasileiro.

Figura 20 – Localização das Usinas FV selecionadas no Mapa do Brasil



Fonte: Adaptado de Pereira *et al.* (2017).

O critério central de seleção dessas dez usinas FV foi justamente o fato de estarem localizadas no cinturão solar brasileiro, o que implica níveis de radiação solar próximos entre si e, portanto, maior comparabilidade nos experimentos. Adicionalmente, considerou-se a disponibilidade e qualidade dos dados operacionais necessários à pesquisa.

A Tabela 4 detalha as principais informações técnicas de cada conjunto FV. Esta inclui a sigla de identificação da usina, sua respectiva localidade, coordenadas geográficas (latitude e longitude), e a capacidade instalada, expressa em megawatts (MW). Entre os empreendimentos analisados, destacam-se usinas de grande porte, com capacidade instalada igual ou superior a 60 MW, bem como usinas de médio porte, que contribuem significativamente para o Sistema Interligado Nacional (SIN).

Tabela 4 – Informações Técnicas das Usinas FV

Usina	Sigla Usina	Localidade	Latitude (°)	Longitude (°)	Capacidade Instalada (MW)
1	CSP	Bom Jesus da Lapa (BA)	-13,31	-43,35	67
2	CBJ	Bom Jesus da Lapa (BA)	-13,30	-43,33	80
3	CJS	Juazeiro (BA)	-9,53	-40,49	120
4	PIR	Pirapora (MG)	-17,40	-44,92	321
5	CSG	São Gonçalo do Gurgueia (PI)	-10,10	-45,27	864
6	ITU	Tabocas do Brejo Velho (BA)	-12,31	-44,07	254
7	CLA	Bom Jesus da Lapa (BA)	-13,31	-43,35	60
8	HOR	Tabocas do Brejo Velho (BA)	-12,60	-44,08	103
9	CNO	Ribeira do Piauí (PI)	-8,20	-42,55	292
10	SOL	Aquiraz (CE)	-3,98	-38,39	81

Fonte: A Autora (2025).

4.1.2 Dados Meteorológicos

As variáveis meteorológicas consideradas neste estudo consistem em radiação solar global (W/m²) e temperatura ambiente (°C).

Considerando que todas as usinas analisadas foram obtidas em domínio público e que as estações meteorológicas do INMET localizadas nas proximidades dos empreendimentos apresentavam lacunas significativas nas variáveis de interesse, em alguns casos, com anos inteiros sem registros de radiação e/ou temperatura, optou-se por utilizar os dados meteorológicos da base internacional NASA POWER.

Essa base fornece séries temporais derivadas de observações por satélite e de esquemas de assimilação de dados em modelos atmosféricos, com cobertura espacial global e parametrização a partir das coordenadas geográficas (latitude e longitude) informadas pelo usuário. Neste trabalho, os dados foram extraídos para as coordenadas de cada usina fotovoltaica, em resolução horária no padrão UTC, sendo posteriormente convertidos para o horário local.

A decisão pelo uso da NASA POWER é corroborada pelo estudo de Faccin *et al.* (2024), que analisou 476 locais no Brasil e demonstrou a boa concordância entre os dados de radiação solar global e temperatura ambiente do NASA POWER e aqueles medidos pelo INMET.

4.1.3 Dados Temporais e Sazonalidade

A análise abrangente realizada neste estudo baseou-se em dados de geração FV e em dados meteorológicos, ambos com registros em intervalos horários. O período de coleta e análise estendeu-se de 1º de janeiro de 2020 a 31 de dezembro de 2023, totalizando quatro anos de informações. Essa granularidade horária é crucial para a compreensão detalhada das variações diárias e sazonais na produção de energia e nas condições climáticas. A consistência temporal e a alta resolução dos dados são fundamentais para a acurácia do modelo.

Esse intervalo (2020 – 2023) foi selecionado de forma a reduzir a influência de episódios recentes de *curtailment* no sistema elétrico brasileiro, buscando contemplar um horizonte temporal em que a geração FV estivesse menos sujeita a restrições operativas que pudessem distorcer o comportamento natural das séries de geração.

Quanto à sazonalidade, o ano foi segmentado em suas quatro estações climáticas tradicionais: verão, primavera, outono e inverno. Essa divisão considerou tanto as variações climáticas predominantes em cada período quanto os meses específicos que as caracterizam.

4.2 Pré-processamento dos Dados

4.2.1 Eliminação de Dados Noturnos

Os dados de geração FV coletados incluem valores noturnos, caracterizados por registros nulos ou muito próximo de zero no período das 00:00h às 06:00h e das 18:00h às 23:00h. Essa esparsidão de dados representa um desafio, pois a grande quantidade de informações zeradas pode resultar em um modelo inadequadamente treinado, comprometendo seu desempenho global.

Para mitigar esse problema, os valores noturnos foram removidos de toda a base de dados, retendo-se apenas as informações correspondentes ao intervalo das 07:00h às 17:00h. Essa filtragem resultou em 11 horas diárias de dados. Consequentemente, o conjunto de dados para cada usina FV, considerando o período de quatro anos, totalizou 16.071 observações.

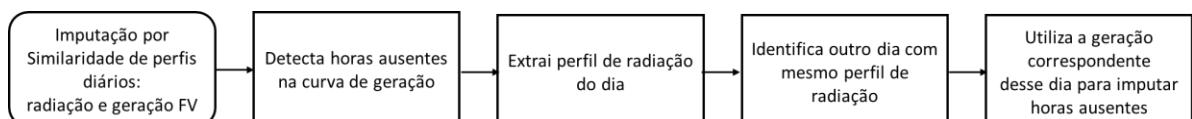
4.2.2 Imputação de Dados Faltantes

Os dados ausentes frequentemente são resultantes de falhas no processo de coleta ou de sensores defeituosos e representam um desafio significativo na análise de séries temporais. Em muitos cenários práticos, a integridade das séries pode ser comprometida por interrupções no funcionamento dos equipamentos de medição ou por erros de registro.

Para o tratamento dos dados ausentes nas séries de geração FV, a proposta metodológica deste trabalho consistiu em uma imputação baseada na curva de radiação solar global, considerando a alta correlação entre as duas variáveis. O processo adotado identifica, para cada horário ausente, um outro dia com perfil similar de radiação (que não apresenta lacunas) e utiliza a geração correspondente desse dia como valor imputado.

Para verificar o perfil de similaridade da radiação solar, foram analisados os cinco dias imediatamente anteriores ao dia com dado ausente. Para cada um desses dias, calculou-se o erro médio absoluto entre o perfil horário de radiação e a curva correspondente ao dia com falha de registro, selecionando-se como dia mais semelhante, aquele que apresentou o menor valor desse erro. O fluxograma da Figura 21 mostra o processo de imputação proposto neste trabalho.

Figura 21 - Fluxograma do método proposto para imputação de dados



Fonte: A Autora (2025).

Para avaliar a qualidade do método de imputação proposto, também foram testadas abordagens clássicas, a saber: interpolação linear; imputação pela média temporal, na qual os dados horários ausentes em um determinado dia são substituídos por um valor único igual à média dos dados não ausentes daquele dia; e o método dos *k*-vizinhos mais próximos (KNN), considerando-se a média dos cinco vizinhos mais próximos.

4.2.3 Seleção de Atributos

A seleção de atributos é uma etapa crítica em modelos de previsão de séries temporais, pois reduz a complexidade, melhora a interpretabilidade e evita sobreajuste. No contexto desta pesquisa, foram selecionados quatro grupos iniciais de variáveis: (i) estáticas, como sazonalidade (estações e meses do ano); (ii) meteorológicas, como radiação solar global e temperatura ambiente; (iii) histórico de geração, representado por defasagens de geração ($h - 1, h - 2, \dots, h - 11$) e estatísticas diárias; e (iv) horários cíclicos, todas passíveis de serem conhecidas ao longo do horizonte de previsão.

As defasagens de geração podem ser definidas como o histórico da geração em horários específicos, nesta pesquisa foram considerados as defasagens horárias de geração $h - 1, h - 2, h - 3$ e $h - 11$. As estatísticas horárias de geração (valor médio, máximo e mínimo), foram definidas, conforme pode ser visto nas Eq. (24) - (26):

$$G_{med_h} = \frac{G(h) + G(h - 1)}{2} \quad (24)$$

$$G_{máx_h} = \max\{G(h), G(h - 1)\} \quad (25)$$

$$G_{mín_h} = \min\{G(h), G(h - 1)\} \quad (26)$$

em que h representa a hora e $G(h)$ a geração FV na hora h .

Os horários cíclicos foram transformados em funções seno e cosseno para que o modelo capte naturalmente a periodicidade diária e mensal, conforme Eq. (27) e (28):

$$\left\{ hora_sen = \sin\left(\frac{2\pi \times h}{24}\right) \mid hora_cos = \cos\left(\frac{2\pi \times h}{24}\right) \right\} \quad (27)$$

$$\left\{ mes_sen = \sin\left(\frac{2\pi \times Mes}{12}\right) \mid mes_cos = \cos\left(\frac{2\pi \times Mes}{12}\right) \right\} \quad (28)$$

em que Mes representa o mês do ano.

Para enriquecer o conjunto de dados, além dos atributos já definidos, foram geradas novas variáveis por meio da interação (produto) entre pares de atributos. A inclusão dessas interações é validada por Zheng e Casari (2018), que indicam um potencial ganho de precisão no modelo em relação ao uso de características isoladas. Essa abordagem permite que o modelo compreenda e utilize relações complexas que emergem da combinação de condições,

superando a análise de variáveis separadas. As interações específicas que foram consideradas podem ser consultadas na Tabela 5.

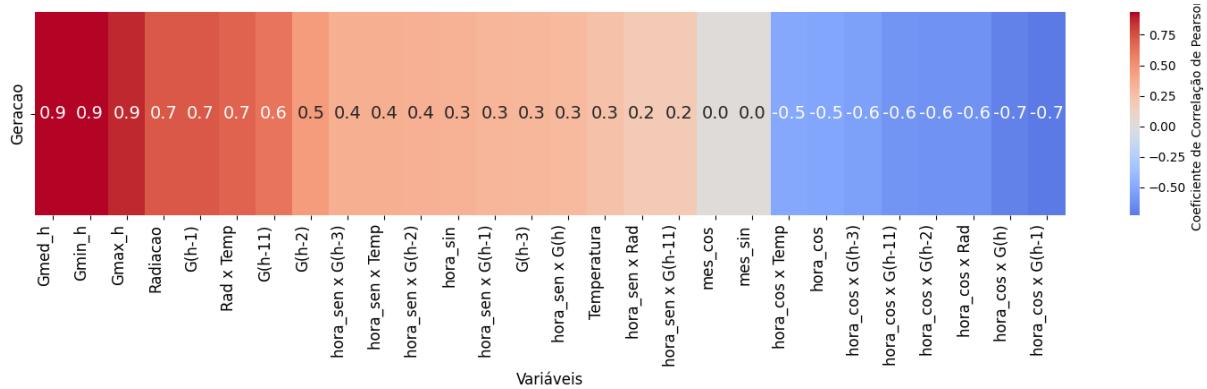
Tabela 5 - Interação de Atributos

Atributos		
$hora_sen \times G(h)$	$hora_sen \times Rad$	$hora_sen \times Temp$
$hora_cos \times G(h)$	$hora_cos \times Rad$	$hora_cos \times Temp$
$hora_sen \times G(h-1)$	$hora_cos \times G(h-1)$	$Rad \times Temp$
$hora_sen \times G(h-2)$	$hora_cos \times G(h-2)$	$hora_sen \times G(h-3)$
$hora_cos \times G(h-3)$	$hora_sen \times G(h-11)$	$hora_cos \times G(h-11)$

Fonte: A Autora (2025).

Na primeira etapa da seleção de atributos, foi realizada uma análise de correlação de Pearson entre a geração FV (variável-alvo) e as demais 28 variáveis. Atributos como G_{med_h} , G_{min_h} , $G_{máx_h}$, Rad e variáveis derivadas como $G(h - 1)$ apresentaram correlações positivas fortes ($r > 0,7$), indicando alta relevância para o modelo. Interações específicas como $Rad \times Temp$, $G(h - 11)$ e $G(h - 2)$ também se destacaram, com coeficientes de correlação acima de 0,5. Por outro lado, correlações negativas expressivas ($r < -0,5$) ocorreram como foi o caso das interações $hora_cos \times G(h)$, $hora_cos \times G(h - 1)$, $hora_cos \times Rad$, $hora_cos \times G(h - 2)$, $hora_cos \times G(h - 11)$, $hora_cos \times Temp$, $hora_cos$ e $hora_cos \times G(h - 3)$. O resultado completo dessa análise pode ser visto na Figura 22.

Figura 22 - Análise de Correlação das Variáveis com a Geração FV



Fonte: A Autora (2025).

Essa análise inicial permitiu priorizar os atributos mais relevantes e que foram o ponto de partida para a próxima etapa consistiu em treinar o modelo de base TFT de forma iterativa, analisando a importância das variáveis tanto no encoder quanto no decoder. Para cada treinamento, foram extraídas as métricas internas de importância de variáveis fornecidas pelo TFT. Essa análise permitiu identificar quais atributos exerciam maior influência na etapa de

codificação das entradas passadas (encoder), responsável por capturar dependências históricas, e na etapa de decodificação das entradas futuras (decoder), que incorpora variáveis conhecidas ao longo do horizonte de previsão.

Com base nesses resultados, procedeu-se à remoção gradual das variáveis de menor relevância, guiada pelas informações de importância do próprio TFT. Esse processo buscou reduzir a dimensionalidade do conjunto de entrada, eliminando atributos com contribuição mínima para o desempenho do modelo, preservando aqueles capazes de capturar padrões relevantes para a previsão da geração FV. A eliminação sistemática de variáveis redundantes ou pouco informativas contribuiu para diminuir o risco de sobreajuste, reduzir o custo computacional e aumentar a interpretabilidade do modelo final. A relação final dos 11 atributos considerados na pesquisa pode ser vista na Tabela 6.

Tabela 6 - Atributos Selecionados para a Modelagem TFT

Componentes do Tempo	Covariáveis Meteorológicas	Estatísticas de Geração FV
<i>hora_sen</i>	<i>hora_sen</i> \times <i>Rad</i>	G_{med_h}
<i>hora_cos</i>	<i>hora_cos</i> \times <i>Rad</i>	G_{max_h}
	<i>hora_sen</i> \times <i>Temp</i>	G_{min_h}
	<i>hora_cos</i> \times <i>Temp</i>	<i>hora_sen</i> \times $G(h - 1)$
		<i>hora_cos</i> \times $G(h - 1)$

Fonte: A Autora (2025).

4.2.4 Transformação Log

A alta variabilidade e as falhas nas séries temporais de geração diária, observadas nesta pesquisa, levaram à aplicação da transformação logarítmica. O objetivo foi minimizar ruídos e otimizar o desempenho do modelo preditivo.

Essa abordagem é justificada por Zheng e Casari (2018), que explicam que a transformação logarítmica comprime a cauda longa na extremidade superior da distribuição, aproximando valores mais altos, enquanto expande a parte inferior, dispersando valores mais baixos. Esse processo resulta em uma distribuição mais simétrica e menos concentrada, normalizando distribuições assimétricas. Ao lidar melhor com dados de cauda longa e ruídos, a transformação torna os dados mais manejáveis, revela padrões ocultos na escala original e permite que os modelos capturem relações subjacentes com maior precisão.

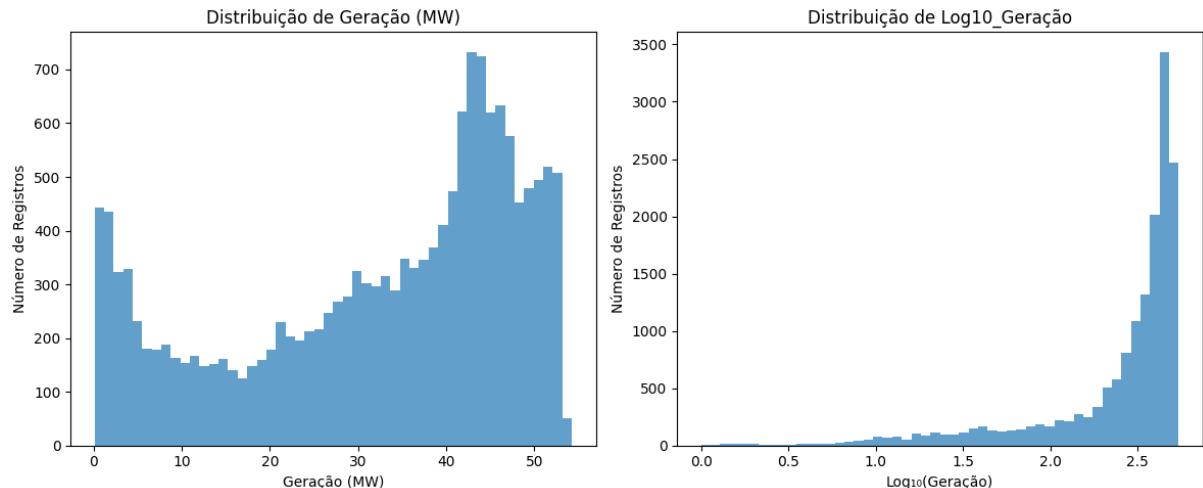
Para estabilizar a variância da base de dados de geração FV e evitar a ocorrência de valores negativos após a transformação, aplicou-se a escala logarítmica definida pela Eq. (29):

$$G_{log10} = \log_{10}(G(h) + 1) \quad (29)$$

em que $G(h)$ representa a geração FV em MW e o termo “+1” desloca toda a distribuição para valores não-negativos, garantindo $G_{log10} \geq 0$ mesmo quando $\log_{10}(G(h) + 1)$ for negativo.

A Figura 23 mostra histogramas que comparam a distribuição da geração FV em toda a base de dados de uma usina FV antes e após a transformação logarítmica na base 10. No gráfico à esquerda, observa-se uma forte assimetria, com concentração em valores elevados e uma cauda longa de geração. Após aplicar o logaritmo, a distribuição torna-se muito mais simétrica e próxima de uma curva normal, o que reduz a influência de ruídos e facilita o ajuste e a interpretação dos modelos. Esse 'achatamento' da cauda direita é fundamental para melhorar a estabilidade numérica e a capacidade preditiva de algoritmos sensíveis a desvios extremos.

Figura 23 - Comparativo da Transformação Log na Geração FV



Fonte: A Autora (2025).

É importante destacar que a Transformação Log foi aplicada exclusivamente para o treinamento dos modelos TFT. Posteriormente, durante a etapa de avaliação de desempenho dos modelos, as métricas foram calculadas com os dados revertidos à sua escala original em MW.

4.2.5 Normalização dos Dados

Para assegurar a comparabilidade entre as diversas entradas e promover estabilidade numérica durante o treinamento do TFT, empregou-se uma estratégia de normalização adaptada à distribuição empírica de cada variável. As covariáveis meteorológicas ($hora_sen \times Rad$, $hora_cos \times Rad$, $hora_sen \times Temp$, $hora_cos \times Temp$) e

componentes do tempo (*hora_sen*, *hora_cos*) apresentaram distribuições aproximadamente normal após tratamento inicial, por isso, foram submetidas à padronização *Standard*, que centra as observações em média zero e desvio-padrão unitário, facilitando a extração de padrões lineares e não lineares sem distorções significativas causadas por ruídos moderados.

Por outro lado, a variável-alvo ($G(h)$) e suas estatísticas, que exibiram caudas longas e picos acentuados, foram escalonados por Escalonamento Robusto. Essa técnica baseada em mediana e IQR, é especialmente apropriada por ser insensível a valores extremos e por preservar a escala central dos dados.

As variáveis categóricas sazonais e dependentes do tempo (*Estacao* e *Mes*) não foram escalonadas, sendo a variável *Estacao* representada no intervalo [1, 2, 3, 4], correspondente às quatro estações do ano, e a variável *Mes* na faixa [1, 2, ..., 11, 12] que representa os doze meses do ano.

4.2.6 Clusterização Hierárquica

A clusterização hierárquica foi empregada com o objetivo de identificar grupos de usinas FV com padrões de geração similares, de forma a explorar relações de dependência entre séries e subsidiar a etapa de modelagem multissérie apresentada no capítulo seguinte. A análise foi conduzida a partir das séries de geração FV das dez usinas selecionadas, considerando os dados previamente tratados e imputados conforme descrito nas seções anteriores.

Para definir a medida de similaridade entre as usinas, foram construídas séries de médias diárias de geração referentes a 50 dias do ano de 2020 para cada usina. Esse conjunto de dias, pertencente ao primeiro ano do período de análise, foi utilizado como amostra representativa do comportamento de geração de cada empreendimento, de modo a viabilizar a comparação entre usinas em uma base temporal comum. Assim, para cada usina obteve-se uma sequência temporal de 50 pontos, cada um correspondente à média diária da geração registrada naquele dia.

A partir da matriz de distâncias DTW entre todas as usinas, aplicou-se o algoritmo de clusterização hierárquica, no qual as usinas mais similares são agrupadas iterativamente em níveis sucessivos, formando uma estrutura em árvore.

O resultado desse processo é representado por um dendrograma, que ilustra graficamente as fusões entre usinas (ou grupos de usinas) em diferentes níveis de distância. Esse dendrograma permite visualizar de maneira intuitiva quais usinas apresentam comportamento mais próximo entre si e em que ponto novos agrupamentos passam a incorporar usinas com

padrões mais distintos. A classificação final dos agrupamentos, bem como o dendrograma resultante da aplicação da distância DTW para cada usina, será apresentada e discutida no Capítulo 5.

4.3 Algoritmo TFT

Considerando a expressão da modelagem TFT vista na Eq. (15), para sua modelagem foram utilizadas as seguintes entradas por categorias: variável-alvo ($G(h)$), correspondente ao histórico de geração FV; covariáveis estáticas, que neste caso, variam com o tempo; covariáveis passadas e covariáveis conhecidas no futuro. A Tabela 7 sintetiza os atributos utilizados em cada uma dessas categorias.

Tabela 7 - Atributos Selecionados para o Modelo TFT

Variável-alvo	Covariáveis Estáticas	Covariáveis Passadas	Covariáveis Conhecidas no Futuro
$G(h)$	<i>Estacao</i>	G_{med_h}	<i>hora_sen</i>
	<i>Mes</i>	G_{max_h}	<i>hora_cos</i>
		G_{min_h}	<i>hora_sen</i> \times <i>Rad</i>
		<i>hora_sen</i> \times $G(h-1)$	<i>hora_cos</i> \times <i>Rad</i>
		<i>hora_cos</i> \times $G(h-1)$	<i>hora_sen</i> \times <i>Temp</i>
			<i>hora_cos</i> \times <i>Temp</i>

Fonte: A Autora (2025).

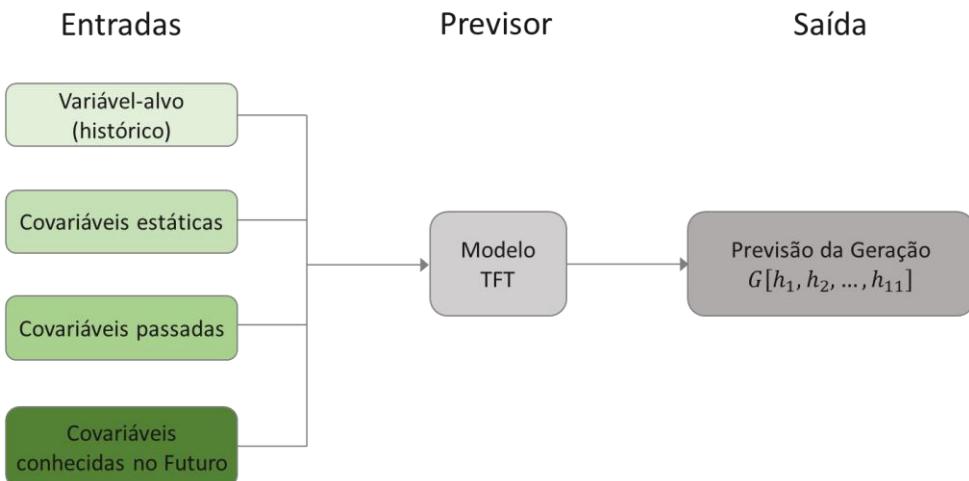
Para uma avaliação abrangente do desempenho preditivo do TFT em distintas condições operacionais e ambientais, 14 modelos distintos foram treinados. Essa abordagem multifacetada permitiu investigar a robustez e a adaptabilidade do modelo a diferentes cenários de dados, otimizando sua capacidade de generalização.

Além dos atributos listados na Tabela 7, foi incluída uma covariável categórica estática responsável por identificar a usina em cada série temporal. Nos modelos de previsão de uma única usina, essa identificação é feita pela variável *ID*, que corresponde ao nome da usina. Já nos modelos multipreditivos, em que são previstas simultaneamente duas ou três usinas, utiliza-se a variável *Usina*, que distingue cada uma das usinas dentro do mesmo conjunto de entrada.

Foram adicionados índices temporais absolutos e relativos (*time_idx* e *relative_time_idx*) para orientar o TFT sobre a cronologia das séries temporais. Tais variáveis são cruciais para a estrutura do modelo, pois permitem que o modelo reconheça a posição absoluta de cada ponto na linha do tempo.

A Figura 24 ilustra, de forma condensada, o fluxo de informação do sistema de previsão proposto. No núcleo do modelo TFT, camadas de atenção multi-cabeça e mecanismos de fusão temporal identificam padrões de curto e longo prazo, ao mesmo tempo em que quantificam a contribuição de cada variável ao longo do horizonte de previsão. Por fim, a saída da modelagem TFT fornece, para cada usina, um vetor de estimativas de geração FV para as 11 horas futuras consideradas, $G[h_1, h_2, \dots, h_{11}]$, permitindo quantificar não apenas os valores previstos, mas também a relevância relativa das diferentes fontes de informação empregadas.

Figura 24 - Diagrama Esquemático da Modelagem TFT



Fonte: A Autora (2025).

Os modelos M1 a M8 foram configurados para avaliar o desempenho preditivo do TFT em uma única usina FV, com foco no impacto da transformação logarítmica da variável-alvo. Nos casos M1, M3, M5 e M7, o modelo foi treinado com geração transformada (G_{log10}), combinando diferentes configurações de variáveis. Nos casos M2, M4, M6 e M8, utilizou-se a variável-alvo em sua escala original (MW). A Tabela 8 apresenta em detalhes as configurações de cada um desses oito modelos.

Tabela 8 - Configuração de entrada dos Modelos M1 a M8

Modelo	Variável-alvo	Covariáveis estáticas	Covariáveis passadas	Covariáveis conhecidas no Futuro
M1	Log10	<i>Estacao</i> <i>Mes</i>	-	<i>hora_sen</i> <i>hora_cos</i>
M2	MW	<i>Estacao</i> <i>Mes</i>	-	<i>hora_sen</i> <i>hora_cos</i>
M3	Log10	<i>Estacao</i> <i>Mes</i>	-	<i>hora_sen</i> <i>hora_cos</i> <i>Rad</i> <i>Temp</i>
M4	MW	<i>Estacao</i>	-	<i>hora_sen</i>

			<i>Mes</i>	<i>hora_cos</i> <i>Rad</i> <i>Temp</i>
M5	Log10	<i>Estacao</i> <i>Mes</i>	G_{med_h} G_{max_h} G_{min_h} $hora_sen \times G(h-1)$ $hora_sen \times G(h-1)$	<i>hora_sen</i> <i>hora_cos</i>
M6	MW	<i>Estacao</i> <i>Mes</i>	G_{med_h} G_{max_h} G_{min_h} $hora_sen \times G(h-1)$ $hora_sen \times G(h-1)$	<i>hora_sen</i> <i>hora_cos</i>
M7	Log10	<i>Estacao</i> <i>Mes</i>	G_{med_h} G_{max_h} G_{min_h} $hora_sen \times G(h-1)$ $hora_sen \times G(h-1)$	<i>hora_sen</i> <i>hora_cos</i> $hora_sen \times Rad$ $hora_cos \times Rad$ $hora_sen \times Temp$ $hora_cos \times Temp$
M8	MW	<i>Estacao</i> <i>Mes</i>	G_{med_h} G_{max_h} G_{min_h} $hora_sen \times G(h-1)$ $hora_sen \times G(h-1)$	<i>hora_sen</i> <i>hora_cos</i> $hora_sen \times Rad$ $hora_cos \times Rad$ $hora_sen \times Temp$ $hora_cos \times Temp$

Fonte: A Autora (2025).

Nota 1: “-” indica variável não utilizada no modelo correspondente.

Nota 2: Em todas as configurações, é utilizada uma covariável categórica estática de identificação da usina (*ID*) para os modelos de usina única (M1–M8).

É importante destacar que a transformação logarítmica foi aplicada apenas no treinamento do TFT; contudo, durante a avaliação, as métricas foram calculadas com os valores reconvertidos à escala original em MW. Adicionalmente, todos os oito modelos foram treinados com os dados de geração da Usina 1 (CSP).

Os modelos M9 a M14 foram configurados para avaliar o desempenho multipreditivo do TFT na previsão simultânea da geração de duas ou três usinas. Para isso, foram formadas combinações de pares ou trios de usinas com base nos resultados da clusterização hierárquica. Em todos esses modelos, as variáveis-alvo foram transformadas para a escala logarítmica (G_{log10}) durante o processo de treinamento.

Especificamente, os modelos M9, M10, M11 e M12 consideraram usinas FV que compartilham o mesmo conjunto de variáveis meteorológicas, dada a localização geográfica em comum entre elas. Em contraste, os modelos M13 e M14 empregaram dois e três conjuntos distintos de bases de dados meteorológicos, respectivamente. A Tabela 9 detalha as configurações de entrada adotadas em cada um desses seis modelos.

Tabela 9 - Configuração de entrada dos Modelos M9 a M14

Modelos	Variáveis-alvo	Covariáveis estáticas	Covariáveis passadas	Covariáveis conhecidas no Futuro
M9, M10, M11	Log10 ^(I)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	<i>hora_sen</i> <i>hora_cos</i>
	Log10 ^(II)		$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen \times Rad</i> <i>hora_cos \times Rad</i> <i>hora_sen \times Temp</i> <i>hora_cos \times Temp</i>
	Log10 ^(III)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	
	Log10 ^(I)		$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen</i> <i>hora_cos</i>
	Log10 ^(II)		$G_{med_h}^{(III)}$ $G_{max_h}^{(III)}$ $G_{min_h}^{(III)}$ $hora_sen \times G(h-1)^{(III)}$ $hora_sen \times G(h-1)^{(III)}$	<i>hora_sen \times Rad</i> <i>hora_cos \times Rad</i> <i>hora_sen \times Temp</i> <i>hora_cos \times Temp</i>
	Log10 ^(III)		$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	
M12	Log10 ^(I)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen</i> <i>hora_cos</i>
	Log10 ^(II)		$G_{med_h}^{(III)}$ $G_{max_h}^{(III)}$ $G_{min_h}^{(III)}$ $hora_sen \times G(h-1)^{(III)}$ $hora_sen \times G(h-1)^{(III)}$	<i>hora_sen \times Rad</i> <i>hora_cos \times Rad</i> <i>hora_sen \times Temp</i> <i>hora_cos \times Temp</i>
	Log10 ^(III)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	
	Log10 ^(I)		$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen \times Rad</i> <i>hora_cos \times Rad</i> <i>hora_sen \times Temp</i> <i>hora_cos \times Temp</i>
M13	Log10 ^(I)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	<i>hora_sen</i> <i>hora_cos</i>
	Log10 ^(II)		$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen \times Rad</i> ^(I) <i>hora_cos \times Rad</i> ^(I) <i>hora_sen \times Temp</i> ^(I) <i>hora_cos \times Temp</i> ^(I)
	Log10 ^(III)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(III)}$ $G_{max_h}^{(III)}$ $G_{min_h}^{(III)}$ $hora_sen \times G(h-1)^{(III)}$ $hora_sen \times G(h-1)^{(III)}$	<i>hora_sen \times Rad</i> ^(II) <i>hora_cos \times Rad</i> ^(II) <i>hora_sen \times Temp</i> ^(II) <i>hora_cos \times Temp</i> ^(II)
	Log10 ^(I)		$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	
	Log10 ^(II)		$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen \times Rad</i> ^(I) <i>hora_cos \times Rad</i> ^(I) <i>hora_sen \times Temp</i> ^(I) <i>hora_cos \times Temp</i> ^(I)
	Log10 ^(III)		$G_{med_h}^{(III)}$ $G_{max_h}^{(III)}$ $G_{min_h}^{(III)}$ $hora_sen \times G(h-1)^{(III)}$ $hora_sen \times G(h-1)^{(III)}$	<i>hora_sen \times Rad</i> ^(II) <i>hora_cos \times Rad</i> ^(II) <i>hora_sen \times Temp</i> ^(II) <i>hora_cos \times Temp</i> ^(II)
M14	Log10 ^(I)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(I)}$ $G_{max_h}^{(I)}$ $G_{min_h}^{(I)}$ $hora_sen \times G(h-1)^{(I)}$ $hora_sen \times G(h-1)^{(I)}$	<i>hora_sen</i> <i>hora_cos</i>
	Log10 ^(II)		$G_{med_h}^{(II)}$ $G_{max_h}^{(II)}$ $G_{min_h}^{(II)}$ $hora_sen \times G(h-1)^{(II)}$ $hora_sen \times G(h-1)^{(II)}$	<i>hora_sen \times Rad</i> ^(I) <i>hora_cos \times Rad</i> ^(I) <i>hora_sen \times Temp</i> ^(I) <i>hora_cos \times Temp</i> ^(I)
	Log10 ^(III)	<i>Estacao</i> <i>Mes</i>	$G_{med_h}^{(III)}$ $G_{max_h}^{(III)}$ $G_{min_h}^{(III)}$ $hora_sen \times G(h-1)^{(III)}$ $hora_sen \times G(h-1)^{(III)}$	<i>hora_sen \times Rad</i> ^(II) <i>hora_cos \times Rad</i> ^(II) <i>hora_sen \times Temp</i> ^(II) <i>hora_cos \times Temp</i> ^(II)
	Log10 ^(I)		$G_{med_h}^{(III)}$	<i>hora_sen \times Rad</i> ^(III)

$G_{max_h}^{(III)}$	$hora_cos \times Rad^{(III)}$
$G_{min_h}^{(III)}$	$hora_sen \times Temp^{(III)}$
$hora_sen \times G(h - 1)^{(III)}$	$hora_cos \times Temp^{(III)}$
$hora_sen \times G(h - 1)^{(III)}$	

Fonte: A Autora (2025).

Nota 1: Os sobrescritos (I), (II) e (III) indicam, respectivamente, a primeira, segunda e terceira usinas contempladas em cada modelo multipreditivo (M9 a M14), não estando relacionados à numeração global das dez usinas do estudo.

Nota 2: Em todas as configurações, é utilizada uma covariável categórica estática de identificação das usinas: *Usina* para os modelos multipreditivos para previsão de duas ou três usinas (M9–M14).

4.4 Treinamento e Definição de Hiperparâmetros

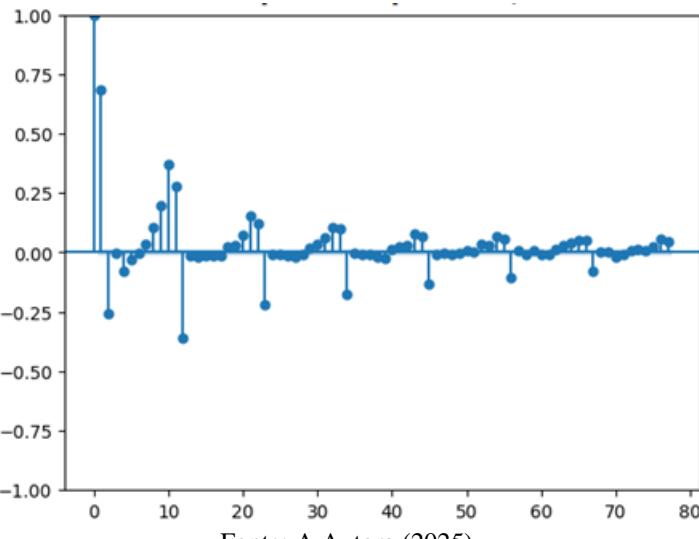
Para garantir a robustez e a capacidade de generalização dos modelos, adotou-se uma divisão temporal dos dados em três conjuntos sequenciais.

O conjunto de treinamento abrange o período de 1º de janeiro de 2020 a 31 de outubro de 2022 (aproximadamente 70,8% dos registros). Essa extensão forneceu ao TFT histórico suficiente para aprender padrões sazonais e a variabilidade intradiária. Em seguida, reservou-se o intervalo de 1º de novembro a 31 de dezembro de 2022 (cerca de 4,2% dos dados) como conjunto de validação. Este conjunto foi utilizado para o ajuste de hiperparâmetros e para o monitoramento de sobreajuste por meio de uma abordagem de validação por janela deslizante.

Por fim, o conjunto de teste corresponde ao ano completo de 2023 (os 25% restantes). Sua avaliação foi realizada por meio de janelas deslizantes, nas quais cada janela foi dividida em uma parte de codificação (as 33 horas anteriores) e uma parte de decodificação (as 11 horas previstas). Esse processo gerou 372 janelas ao longo de todos os experimentos, o que contribuiu significativamente para a robustez dos cálculos das métricas de avaliação dos modelos.

A Figura 25 apresenta a estrutura de dependência temporal de uma das séries de geração FV utilizada na pesquisa, obtida a partir da função de autocorrelação (ACF). Observa-se correlações fortemente positivas nas primeiras horas e até a defasagem de aproximadamente 33 horas, seguidas de uma queda brusca em períodos superiores. Essa janela de persistência indica que a informação passada até a 33^a hora anterior ainda carrega sinal útil para a previsão, enquanto a contribuição de instantes mais remotos se torna desprezível.

Figura 25 - Estrutura de Dependência Temporal da Geração FV



Fonte: A Autora (2025).

Com base nesse diagnóstico, o Encoder do modelo foi configurado para ingerir as 33 horas antecedentes, garantindo a captura completa dos efeitos de memória. O Decoder, por sua vez, foi configurado para prever as próximas 11 horas, o que, para esta pesquisa, equivale à previsão do próximo dia, considerando que apenas 11 horas diárias de dados foram utilizadas.

Para a definição dos hiperparâmetros dos modelos, utilizou-se a ferramenta de otimização automática Optuna. Os detalhes dos parâmetros adotados nas modelagens são descritos na Tabela 10.

Tabela 10 - Hiperparâmetros dos modelos TFT

Hiperparâmetro	Valor	Hiperparâmetro	Valor
Hidden_Size	128	Dropout	0.2
Batch_Size	128	Loss	Quantile Loss ()
Otimizador	AdamW	Attention_Head Size	4
Learning Rate	0.0001	Épocas	100

Fonte: A Autora (2025).

No presente trabalho, os modelos TFT foram treinados utilizando a função de perda *Quantile Loss*, com o objetivo de estimar diferentes quantis da distribuição condicional da geração FV. Entretanto, dado que o foco da análise comparativa reside na qualidade das previsões pontuais, o quantil mediano (0,5) foi adotado como estimativa pontual da geração. Dessa forma, todas as métricas de desempenho reportadas nos resultados são calculadas a

partir desse quantil, permitindo uma comparação direta com os demais modelos determinísticos avaliados.

O otimizador AdamW é uma variação do Adam em que a decaimento de peso é aplicado apenas após o ajuste do tamanho do passo individual de cada parâmetro. Ele foi adotado por ser apropriado para o treinamento de modelos maiores, com dados complexos e de alta dimensão, pois a redução de peso desacoplada ajuda a obter melhor generalização e convergência estável (LLUGSI *et al.*, 2021).

4.5 Avaliação

O desempenho dos métodos de previsão é analisado por meio de métricas que estabelecem a diferença entre o valor real (y) e o valor previsto (\hat{y}). Os índices utilizados nesta pesquisa foram o erro médio absoluto (MAE), a raiz do erro médio quadrático (RMSE), o erro médio absoluto normalizado (nMAE) e a raiz do erro médio quadrático normalizada (nRMSE).

O MAE e o RMSE quantificam, em unidades da variável-alvo, o erro médio e sua dispersão, respectivamente. Já o nMAE expressa o erro em termos percentuais, normalizados pela capacidade instalada de cada usina, o que facilita a comparação entre empreendimento com diferentes potências e escalas de geração. As expressões desses indicadores estão apresentadas nas Eq. (30) a (32):

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (30)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (31)$$

$$nMAE = \frac{100\%}{CAP} \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (32)$$

em que y_i representa o valor real, \hat{y}_i o valor previsto, CAP a capacidade instalada da usina e N o tamanho do conjunto de teste.

Para aumentar a confiabilidade dos resultados, cada um dos 14 modelos (M1 a M14) foi treinado e avaliado dez vezes, utilizando dez inicializações distintas (*run 0*, *run 1*, ..., *run 9*). Dessa forma, obteve-se, para cada modelo, uma distribuição de valores de MAE e RMSE ao longo das repetições, a partir da qual foram calculados o desvio-padrão do erro médio absoluto (dp-MAE) e o desvio-padrão da raiz do erro médio quadrático (dp-RMSE). Esses

indicadores descritivos permitem avaliar a estabilidade do treinamento e a variabilidade do desempenho entre diferentes inicializações.

Além da análise descritiva, foi conduzida uma análise estatística inferencial para comparar o desempenho dos modelos TFT dois a dois. Para cada par de modelos, foram consideradas as dez observações de MAE obtidas em cada *run*, formando amostras pareadas. A partir dessas amostras, foram calculados: a diferença média entre os modelos; o intervalo de confiança de 95% (IC95%) para essa diferença média; o valor do teste t pareado, para avaliar se a diferença média é estatisticamente distinta de zero sob a suposição de normalidade das diferenças; e, o teste não paramétrico de Wilcoxon para amostras pareadas, que não exige a hipótese de normalidade e serve como verificação complementar da robustez dos resultados.

A combinação das duas abordagens (teste t e Wilcoxon) permite verificar se as diferenças observadas nas métricas de erro entre dois modelos são estatisticamente significativas ou se podem ser atribuídas apenas à variabilidade decorrente das diferentes inicializações. Gráficos de dispersão (MAE de um modelo em função do MAE de outro) e boxplots da distribuição de erros também foram utilizados para auxiliar na interpretação visual dessas diferenças.

O procedimento detalhado para a reproduzibilidade desta pesquisa foi apresentado integralmente neste capítulo. No próximo capítulo, os resultados numéricos e gráficos, bem como as comparações estatísticas entre os modelos, serão discutidos em detalhe.

5 RESULTADOS E DISCUSSÃO

5.1 Imputação de Dados Faltantes

No contexto desta pesquisa, as bases de dados de geração FV apresentaram uma parcela de dados faltantes, cujos percentuais estão detalhados na Tabela 11. Em contrapartida, a base de dados meteorológicos utilizada manteve-se íntegra, sem qualquer registro de dados ausentes, o que constitui um ponto favorável para a confiabilidade das informações climáticas empregadas na modelagem preditiva. A necessidade de tratar adequadamente essas lacunas na série de geração FV é crucial para garantir a robustez e a precisão dos modelos de previsão desenvolvidos, evitando vieses e assegurando a representatividade das séries temporais analisadas.

Tabela 11 - Dados Faltantes em Cada Usina FV

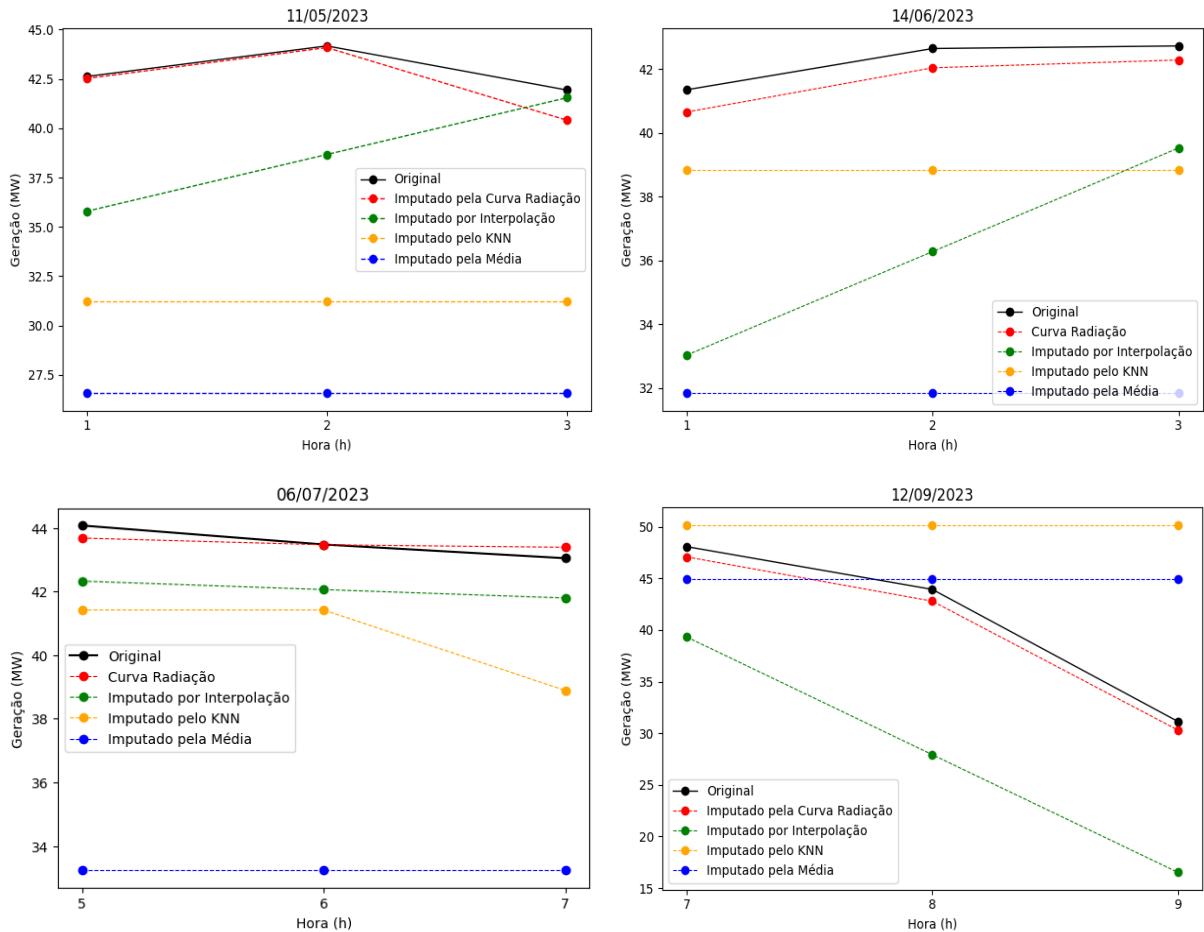
Usina	Sigla	Dados faltantes (%)	Usina	Sigla	Dados faltantes (%)
1	CSP	0,01%	6	ITU	0,07%
2	CBJ	0,03%	7	CLA	0,19%
3	CJS	0,08%	8	HOR	0,07%
4	PIR	0,01%	9	CNO	0,10%
5	CSG	0,07%	10	SOL	0,26%

Fonte: A Autora (2025).

A Figura 26 apresenta a validação da metodologia de imputação de dados ausentes na geração FV por meio de um comparativo com as três técnicas clássicas citadas na seção 4.2.2: interpolação linear, imputação pela média temporal e método dos cinco vizinhos mais próximos (KNN). Foram analisados, quatro exemplos de dias com três horas consecutivas ausentes, em diferentes datas do conjunto de dados: 06/07/2023, 14/06/2023, 12/09/2023 e 11/05/2023. Essas datas foram selecionadas aleatoriamente dentre os dias em que a série de referência (Usina 1 – CSP) apresentava as 11 horas de geração FV completas, permitindo simular artificialmente as lacunas e comparar diretamente os métodos de imputação em relação à curva original.

As cores na Figura 26 representam: curva preta – série original (dados completos, utilizada como referência); curva vermelha – imputação baseada na curva de radiação; curva verde – imputação por interpolação linear; curva laranja – imputação via KNN; e curva azul – imputação pela média temporal.

Figura 26 - Validação da Imputação de Dados na Geração FV



Fonte: A Autora (2025).

Os gráficos comparativos evidenciam que, em todas as datas analisadas, a imputação pela curva de radiação apresentou maior proximidade dos valores originais, com desvios reduzidos e preservando o comportamento real da curva de geração. Em contrapartida, os métodos tradicionais mostram limitações, especialmente em situações com variações mais abruptas na produção. Por exemplo, no dia 12/09/2023, tanto a interpolação linear quanto o KNN apresentam discrepâncias significativas em relação ao padrão observado, suavizando ou distorcendo a queda de geração, enquanto a imputação baseada na radiação mantém trajetória coerente com a série original. A imputação pela média apresentou o pior desempenho em todos os cenários analisados. A imputação pela média temporal é a que apresenta pior desempenho em todos os cenários, por ignorar completamente a dinâmica de flutuações da geração FV ao longo do dia.

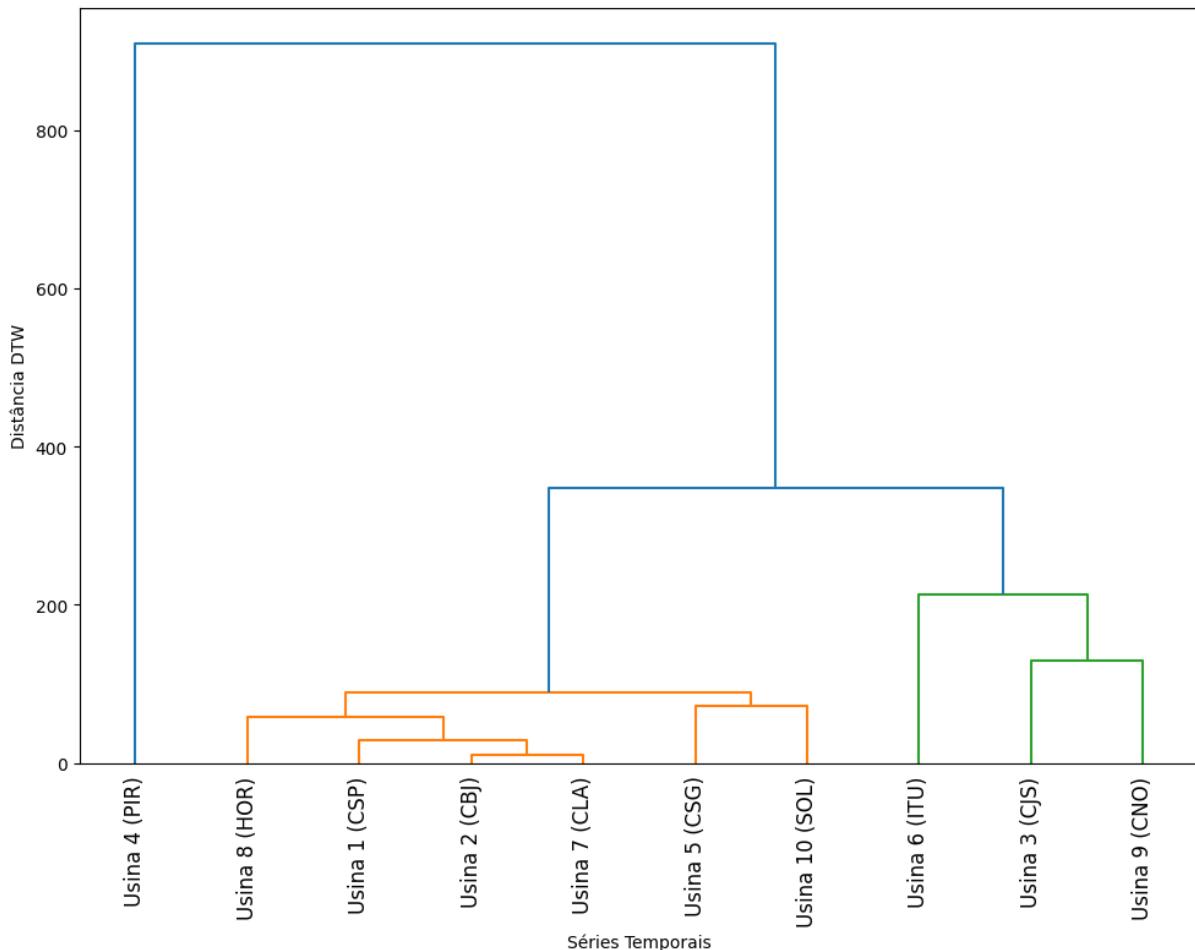
Esses resultados ilustram, de forma empírica, a eficácia da metodologia proposta, demonstrando que a incorporação explícita de informações meteorológicas auxilia na reconstrução mais fidedigna das séries de geração FV e supera técnicas que desconsideram

tais variáveis contextuais. Assim, a estratégia adotada neste trabalho se mostra mais robusta, sobretudo em trechos com múltiplos intervalos de dados ausentes e maior variabilidade da radiação solar.

5.2 Clusterização Hierárquica da Base de Dados de Geração FV

A análise da clusterização hierárquica na base de dados de geração FV, com dez usinas, foi realizada utilizando a distância DTW e tomando como referência a Usina 1 (CSP). O dendrograma, apresentado na Figura 27, sintetiza a análise de similaridade entre as séries temporais.

Figura 27 - Dendrograma resultante da Clusterização Hierárquica



Fonte: A Autora (2025).

O dendrograma evidencia que a Usina 1 (CSP) apresenta alta similaridade com a Usina 2 (CBJ), seguida pela Usina 7 (CLA). Esse agrupamento inicial ocorre em níveis reduzidos de distância DTW, indicando que essas usinas possuem padrões temporais de geração muito

próximos. Esse resultado é justificável, pois as três usinas têm localização geográfica próxima e, consequentemente, condições climáticas semelhantes.

Um segundo agrupamento é formado pela Usina 8 (HOR), que apresenta similaridade moderada com a usina de referência Usina 1 (CSP). Isso sugere que, embora compartilhem tendências gerais de variação da geração, exibem diferenças mais significativas na amplitude e/ou na dinâmica intradiária. O terceiro agrupamento compreende a Usina 5 (CSG) e a Usina 10 (SOL), que apresentam uma distância DTW considerável (~100), sendo seguidas pelas Usinas 6 (ITU), 3 (CJS) e 9 (CNO). Essas diferenças são justificadas pela maior distância em relação à usina de referência e podem ser decorrentes de diferentes configurações de arranjos FV, variação na inclinação dos módulos, regimes distintos de manutenção ou particularidades operacionais.

A Usina 4 (PIR) aparece em uma posição mais distante na árvore hierárquica e se une às demais apenas em patamares de distância muito elevados. Essa distinção acentuada pode decorrer da localização geográfica no sudeste brasileiro com regime solar e padrão de nebulosidade significativamente diferentes da usina de referência localizada no nordeste brasileiro, de particularidades no projeto da usina ou até mesmo de eventos operacionais que afetaram o desempenho no período analisado.

Em síntese, a clusterização hierárquica permite identificar um núcleo de alta similaridade composto por CSP (1), CBJ (2) e CLA (7), um grupo intermediário com usinas de comportamento moderadamente similar, como HOR (8), CSG (5) e SOL (10), e um subconjunto mais afastado, no qual PIR (4) se destaca como um caso atípico no conjunto avaliado.

A correlação de Pearson foi calculada com o objetivo de checar os resultados da clusterização hierárquica, oferecendo uma segunda opinião sobre o grau de proximidade entre as usinas. O resultado pode ser visto na Tabela 12.

Tabela 12 - Matriz de Correlação de Pearson das Usinas FV

	Usina 1	Usina 2	Usina 3	Usina 4	Usina 5	Usina 6	Usina 7	Usina 8	Usina 9	Usina 10
	CSP	CBJ	CJS	PIR	CSG	ITU	CLA	HOR	CNO	SOL
CSP	1,00	0,99	0,59	0,67	0,53	0,91	0,98	0,87	0,34	-0,15
CBJ	0,99	1,00	0,59	0,65	0,51	0,89	0,99	0,86	0,32	-0,14
CJS	0,59	0,59	1,00	0,41	0,61	0,65	0,60	0,67	0,38	0,03
PIR	0,67	0,65	0,41	1,00	0,41	0,61	0,66	0,63	0,14	-0,07
CSG	0,53	0,51	0,61	0,41	1,00	0,53	0,50	0,54	0,47	0,02
ITU	0,91	0,89	0,65	0,61	0,53	1,00	0,90	0,97	0,34	-0,19
CLA	0,98	0,99	0,60	0,66	0,50	0,90	1,00	0,87	0,34	-0,17
HOR	0,87	0,86	0,67	0,63	0,54	0,97	0,87	1,00	0,29	-0,18

CNO	0,34	0,32	0,38	0,14	0,47	0,34	0,34	0,29	1,00	-0,17
SOL	-0,15	-0,14	0,03	-0,07	0,02	-0,19	-0,17	-0,18	-0,17	1,00

Fonte: A Autora (2025).

A análise da primeira coluna da matriz de correlação indica que a Usina 1 (CSP) apresenta altíssima correlação com CBJ (0,99) e CLA (0,98), confirmando que essas três usinas compõem o núcleo mais próximo em termos de variação linear da geração. Esses resultados são coerentes com o dendrograma, no qual CSP se agrupa primeiramente com CBJ e CLA em níveis reduzidos de distância DTW.

A Usina HOR apresenta correlação elevada com a CSP (0,87) e aparece próxima no dendrograma, embora um pouco mais distante que CBJ e CLA. Isso corrobora a consistência entre as duas análises. Porém, foram observadas algumas inconsistências. As usinas PIR (0,67) e CSG (0,53) têm correlação moderada com a CSP. No dendrograma, a PIR aparece mais afastada, enquanto a CSG se mostra mais próxima, o que contradiz a ordem de correlação. No caso da SOL, a correlação com a CSP é negativa (-0,15), indicando um comportamento oposto de geração, mas ela se agrupa na mesma hierarquia que a CSG no dendrograma.

Portanto, enquanto as análises de correlação de Pearson (que descrevem a relação linear entre as séries) e Clusterização Hierárquica (que considera diferenças de forma e deslocamentos temporais) se mostraram consistentes em alguns casos, elas revelaram divergências em outros, destacando as limitações de se usar apenas uma métrica de similaridade. Por essa razão, os resultados da Clusterização Hierárquica foram considerados como a análise principal, pois fornecem uma visão mais completa da similaridade entre as usinas.

5.3 Análise de Desempenho dos Modelos M1 a M8

5.3.1 Desempenho médio e variabilidade

A Tabela 13 compara o desempenho dos modelos M1 a M8, cujas configurações de entrada foram descritas na seção 4.3. Em todos os casos, as métricas de avaliação (MAE, dp-MAE, RMSE, dp-RMSE e nMAE) foram calculadas na escala original de geração (MW), ainda que alguns modelos tenham sido treinados com a variável-alvo transformada em logaritmo decimal ($G_{\log 10}$). Todas as oito configurações foram aplicadas aos dados da Usina 1 (CSP) e, para o cálculo de nMAE, considerou-se a capacidade instalada de 67 MW. Os

valores apresentados correspondem à média das dez inicializações (*runs*) realizadas para cada modelo, enquanto dp-MAE e dp-RMSE quantificam a variabilidade dos erros entre essas repetições.

Tabela 13 – Comparativo de Desempenho - Modelos M1 a M8

Modelo	MAE (MW)	dp-MAE (MW)	RMSE (MW)	dp-RMSE (MW)	nMAE (%)
M1	9,24	0,26	12,29	0,39	14%
M2	8,99	0,17	12,27	0,33	13%
M3	7,24	0,20	10,80	0,29	11%
M4	7,11	0,36	10,71	0,40	11%
M5	9,29	0,29	12,43	0,44	14%
M6	9,27	0,53	12,76	0,70	14%
M7	7,13	0,22	10,64	0,31	10%
M8	7,03	0,20	10,64	0,40	10%

Fonte: A Autora (2025).

Os modelos M1 e M2 podem ser vistos como configurações de base (*baseline*). Eles utilizam apenas informações sazonais e de hora do dia como covariáveis adicionais, além do histórico de geração incorporado pela própria variável-alvo (este último presente em todas as configurações avaliadas). Nessa configuração, os erros médios demonstram-se relativamente elevados (MAE de 9,24 MW e nMAE de 14% para M1, e 8,99 MW e 13% para M2). Os desvios-padrão associados (dp-MAE de 0,26 e 0,17; dp-RMSE de 0,39 e 0,33, respectivamente) indicam uma variabilidade moderada entre as dez inicializações. A diferença entre M1 (log10) e M2 (MW) é pequena, o que sugere que a transformação logarítmica, isoladamente, não produz ganhos expressivos quando o conjunto de covariáveis adicionais é limitado.

A inclusão de variáveis meteorológicas (radiação solar global e temperatura ambiente) nos modelos M3 e M4 está associada a uma redução consistente dos erros médios em relação aos modelos de referência. O MAE diminui para 7,24 MW (M3) e 7,11 MW (M4), com nMAE em torno de 11% e RMSE próximos de 10,8 MW. Os valores de dp-MAE (0,20 e 0,36) e dp-RMSE (0,29 e 0,40) permanecem abaixo de 0,5 MW, indicando desempenho relativamente estável entre as execuções. Esses resultados apontam que as covariáveis meteorológicas contribuem de forma relevante para explicar a variabilidade da geração FV.

Nos modelos M5 e M6, foram incorporados apenas atributos derivados da própria série de geração (média, máximo, mínimo diários e termos de defasagem), sem o uso das variáveis meteorológicas. Nessa configuração, os erros médios (MAE em torno de 9,3 MW e nMAE de

14%) são muito próximos aos dos modelos *baseline* M1 e M2, não se observando melhora expressiva de precisão. Além disso, verifica-se aumento da variabilidade, especialmente em M6, que apresenta os maiores valores de dp-MAE (0,53) e dp-RMSE (0,70) dentre os oito modelos, sugerindo maior sensibilidade às diferentes inicializações.

A combinação simultânea de variáveis meteorológicas e atributos de geração histórica, implementada nos modelos M7 e M8, está associada aos menores valores médios de erro entre as configurações avaliadas. O modelo M7 (treinado na escala logarítmica) apresenta MAE de 7,13 MW, nMAE de 10% e RMSE de 10,64 MW, com dp-MAE de 0,22 e dp-RMSE de 0,31. O modelo M8, análogo em escala MW, obtém valores médios muito próximos (MAE de 7,03 MW, com dp-MAE de 0,20, mesmo RMSE de 10,64 MW e nMAE de 10%), embora com leve aumento na dispersão (dp-RMSE de 0,40). Esses resultados indicam, em termos médios, uma tendência de melhor desempenho das configurações mais completas, nas quais se combinam meteorologia, atributos históricos e transformação logarítmica.

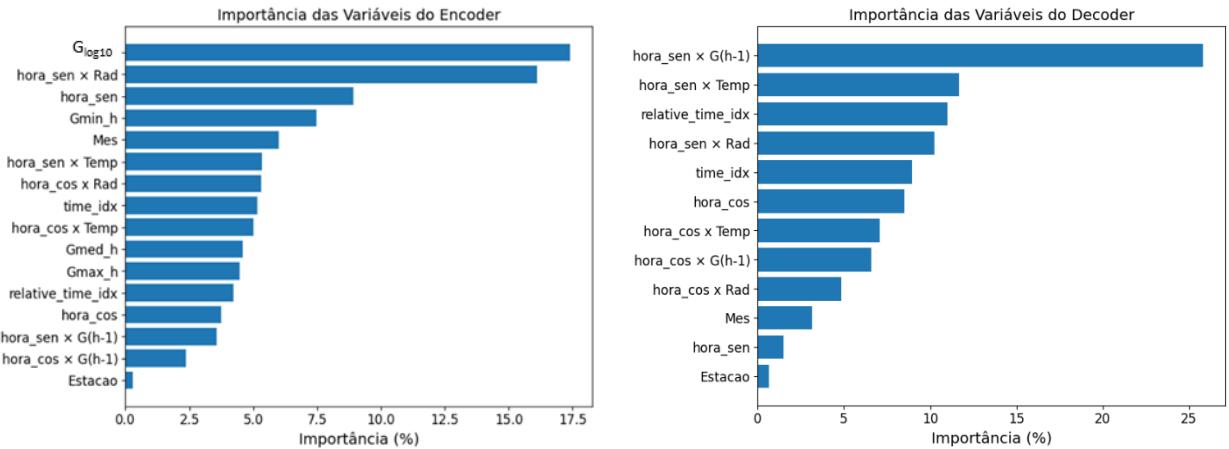
Por fim, destaca-se que as diferenças observadas na Tabela 13 refletem tendências médias de desempenho e não implicam, por si só, em superioridade estatisticamente comprovada de um modelo sobre outro. A verificação formal dessa significância é apresentada na subseção seguinte 5.3.2, por meio de testes estatísticos dois a dois (teste *t* pareado e teste de Wilcoxon), aplicados às distribuições de erro obtidas nas dez inicializações de cada modelo.

Além da comparação das métricas de erro, é importante analisar como o TFT utiliza as diferentes entradas ao construir suas previsões, de modo a verificar se o modelo está de fato explorando informações coerentes do ponto de vista físico e operacional. Nesse sentido, a interpretabilidade é investigada por meio da importância relativa das variáveis no encoder e no decoder do modelo. Entre as configurações avaliadas (M1-M8), o modelo M7 foi selecionado como estudo de caso para a análise de interpretabilidade por representar uma das configurações mais completas de entradas, combinando variáveis meteorológicas, atributos derivados da própria série de geração e transformação logarítmica da variável-alvo. Ressalta-se que os demais modelos exibem padrões de importância qualitativamente semelhantes.

A interpretabilidade do modelo M7 pode ser analisada pela importância das variáveis no encoder e no decoder, apresentada na Figura 28. No encoder, observa-se que a variável mais relevante é a própria série de geração em escala logarítmica (G_{log10}), seguida pelas interações *hora_sen* \times *Rad* e *hora_sen*, que capturam o padrão diário modulando a geração pelo ciclo diurno da radiação solar. A variável agregada da própria série, G_{min_h} , e a variável *Mes* também apresentam contribuição expressiva, indicando que o modelo utiliza simultaneamente

informação histórica de geração e sazonalidade anual para construir o contexto de longo prazo.

Figura 28 - Interpretabilidade das Variáveis do Modelo M7



Fonte: A Autora (2025).

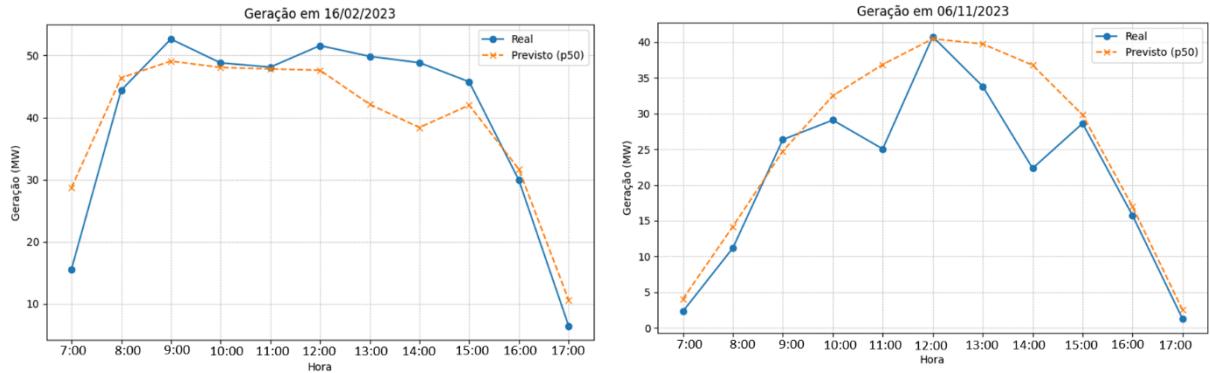
No decoder, a prioridade se desloca para as variáveis diretamente associadas ao passo de previsão corrente. A mais importante passa a ser a interação $hora_sen \times G(h - 1)$, que combina a defasagem imediata da geração com a posição intradiária, evidenciando forte dependência de curto prazo entre instantes consecutivos. Em seguida destacam-se a interação com temperatura ($hora_sen \times Temp$) e radiação solar ($hora_sen \times Rad$). Já variáveis mais agregadas ou sazonais, têm importância relativamente baixa no decoder, sugerindo que a sazonalidade é principalmente capturada na etapa de codificação, enquanto a etapa de decodificação refina a previsão com base no estado imediatamente anterior e nas condições meteorológicas locais.

A Figura 29 ilustra a comparação entre a geração real e a previsão pontual (quantil 0,5) do modelo M8 em dois dias distintos do conjunto de teste (16/02/2023 e 06/11/2023), considerando as 11 horas diárias utilizadas neste estudo. Esses dias foram selecionados como exemplos em que o perfil diário de geração FV não apresenta um comportamento “ideal”, observando-se discrepâncias pontuais possivelmente associadas à variabilidade intradiária do recurso solar.

Além disso, destacam-se pontos de decaimento que podem ser indicativos de eventos operativos não puramente relacionados à radiação solar, como episódios de *curtailment*. Este fenômeno consiste no corte ou limitação da geração por parte do ONS e tem se tornado um desafio crescente na operação do SIN. Motivados por restrições de escoamento na rede ou excesso de oferta, esses cortes forçados impõem uma redução artificial na produção que não

depende da disponibilidade solar, representando um desafio adicional para a precisão de modelos preditivos como o TFT.

Figura 29 - Comparativo Real x Previsto para o Modelo M8



Fonte: A Autora (2025).

Pode-se observar o comportamento da geração FV em 16/02/2023, em que a curva azul representa os valores medidos e a linha tracejada laranja indica a previsão pontual do modelo M8. A curva prevista do modelo segue de forma consistente o formato geral da curva diária, reproduzindo adequadamente a rápida subida matinal, o platô próximo ao pico de geração e o declínio ao final do dia. Pequenas discrepâncias ocorrem em horários específicos, com leve subestimação no pico e superestimação em parte do período vespertino, mas, no conjunto, a trajetória prevista permanece próxima da série observada.

Referente ao dia 06/11/2023, nota-se novamente que a curva prevista segue o padrão da curva real, capturando a tendência de elevação e redução da geração. Entretanto, esse exemplo evidencia flutuações mais acentuadas na curva real, especialmente em torno do horário de pico, que possivelmente pode ser devido ao *curtailment* e que não são totalmente captadas pelo modelo, resultando em uma curva prevista mais suavizada. Esse comportamento é condizente com a natureza do modelo TFT, que tende a suavizar variações muito abruptas para manter a aderência global ao perfil diário.

5.3.2 Testes de significância estatística

Com o objetivo de complementar a análise descritiva apresentada na Seção 5.3.1, esta subseção investiga se as diferenças de desempenho entre os modelos M1 a M8 são estatisticamente relevantes ou se podem ser atribuídas apenas à variabilidade decorrente das diferentes inicializações. Para isso, foram consideradas, para cada modelo, as dez observações de MAE obtidas nas *runs* independentes, de forma a compor amostras pareadas de erro para cada par de modelos comparado.

A Tabela 14 apresenta os valores de MAE de cada um dos oito modelos em cada uma das dez inicializações. Essa matriz de resultados permite visualizar, de maneira mais detalhada, o comportamento de cada configuração ao longo das diferentes execuções: observa-se que alguns modelos apresentam valores de MAE mais concentrados em torno de uma faixa estreita, indicando maior estabilidade entre *runs*, enquanto outros exibem maior espalhamento, coerente com os maiores valores de dp-MAE discutidos anteriormente. Em particular, nota-se que os modelos com melhores métricas médias (como M3, M4, M7 e M8) tendem a concentrar seus MAEs em níveis mais baixos, ao passo que modelos como M1, M2, M5 e M6 apresentam, em geral, valores mais elevados e maior variabilidade entre inicializações, o que reforça as tendências já apontadas na análise descritiva.

Tabela 14 - Resultados de MAE por *run* para os modelos M1 a M8

<i>run</i>	MAE (MW)							
	M1	M2	M3	M4	M5	M6	M7	M8
0	9,10	8,84	7,37	7,90	9,19	9,63	7,14	7,33
1	8,83	9,02	7,51	6,96	9,07	9,71	7,20	7,26
2	9,30	8,83	7,16	7,11	9,09	9,31	7,54	7,19
3	9,62	9,38	6,96	6,97	8,96	9,15	7,42	6,99
4	9,25	9,02	7,09	6,65	9,25	8,88	7,17	6,80
5	9,18	8,91	7,20	7,23	9,08	9,18	6,88	7,09
6	9,27	8,94	7,10	6,83	8,64	8,73	7,09	6,90
7	9,13	8,80	7,47	7,50	9,49	8,86	6,96	6,87
8	9,69	9,09	7,06	7,03	9,18	8,79	7,03	7,12
9	8,99	9,09	7,51	6,97	8,77	10,41	6,89	6,77
Média	9,24	8,99	7,24	7,11	9,07	9,27	7,13	7,03
dp	0,26	0,17	0,20	0,36	0,24	0,53	0,22	0,20

Fonte: A Autora (2025).

A partir dos valores de MAE por inicialização, foram construídas as comparações dois a dois entre os modelos de interesse, a saber: M1 vs. M2, M3 vs. M4, M5 vs. M6, M7 vs. M8 e, por fim, M4 vs. M8, sendo estes os dois modelos com menores MAE médios. Para cada par, calculou-se: (i) a diferença média de MAE entre os modelos; (ii) o intervalo de confiança (IC95%) dessa diferença; e (iii) os testes t pareado e de Wilcoxon. A Tabela 15 apresenta os resultados do teste de significância dois a dois dos modelos M1 a M8.

Tabela 15 – Teste de significância estatística pareada dos modelos M1 a M8

Par	Diferença média	IC 95%	p Teste t	p Teste Wilcoxon
M1 vs. M2	0,25	[0,08 – 0,42]	0,0096	0,0098
M3 vs. M4	0,13	[-0,11 – 0,36]	0,2517	0,2656
M5 vs. M6	-0,19	[-0,65 – 0,27]	0,3666	0,4316
M7 vs. M8	-0,10	[-0,07 – 0,27]	0,2104	0,2871
M4 vs. M8	0,08	[-0,14 – 0,30]	0,4138	0,7695

Fonte: A Autora (2025).

Ao analisar os resultados, verifica-se que a única diferença com relevância estatística confirmada ($p < 0,05$) ocorre na configuração mais simples (*baseline*). Conclui-se que, para modelos que utilizam apenas informações sazonais e histórico, o treinamento na escala original (MW) é superior ao uso da escala logarítmica.

Para os demais pares que envolvem variáveis meteorológicas ou atributos derivados (M3 a M8), a conclusão é de que não há evidência estatística de superioridade entre as diferentes escolhas de escala ou incrementos de atributos. Embora existam variações numéricas nas médias (como o M8 apresentando o menor MAE absoluto de 7,03 MW), os p -valores superiores a 0,05 indicam que essas diferenças podem ser fruto da variabilidade das inicializações e não de uma vantagem real do modelo.

Em suma, a análise estatística revela que, embora existam tendências numéricas favoráveis às configurações mais completas, apenas a transição da escala logarítmica para a original no modelo *baseline* (M2 vs. M1) resultou em ganho de desempenho robusto. Para os demais modelos, as diferenças observadas não possuem significância estatística, sugerindo que a capacidade preditiva do modelo TFT proposto atinge um patamar de estabilidade a partir da inclusão das covariáveis meteorológicas.

5.4 Análise de Desempenho dos Modelos Multipreditivos M9 a M14

5.4.1 Desempenho médio e variabilidade

Para analisar a aprendizagem multipreditiva do TFT, foram construídos modelos com duas usinas (M9, M10, M11 e M13) e com três usinas (M12 e M14). A formação desses pares e trios levou em conta os resultados da clusterização hierárquica apresentada na Seção 5.2, que permitiu agrupar usinas com comportamento semelhante. Os agrupamentos considerados em cada modelo, bem como a capacidade instalada de cada usina, estão descritos na Tabela 16.

Tabela 16 - Descrição das usinas dos modelos M9 a M14

Modelo	Conjunto de usinas (CAP instalada em MW)
M9	1 – CSP (67 MW); 2 – CBJ (80 MW)
M10	1 – CSP (67 MW); 7 – CLA (60 MW)
M11	2 – CBJ (80 MW); 7 – CLA (60 MW)
M12	1 – CSP (67 MW); 2 – CBJ (80 MW); 7 – CLA (60 MW)
M13	1 – CSP (67 MW); 8 – HOR (103 MW)
M14	1 – CSP (67 MW); 5 – CSG (864 MW); 8 – HOR (103 MW)

Fonte: A Autora (2025).

A Tabela 17 sintetiza o desempenho dos modelos multipreditivos M9 a M14, apresentando a média e a variabilidade de dez inicializações para cada configuração. Todas as métricas foram calculadas na escala original (MW), embora os modelos tenham sido treinados com a variável-alvo em logaritmo decimal (G_{log10}).

Tabela 17 - Comparativo de Desempenho - Modelos M9 a M14

Modelo	Usinas	MAE (MW)	dp-MAE (MW)	RMSE (MW)	dp-RMSE (MW)	nMAE (%)
M9	1 (CSP)	6,86	0,16	10,40	0,27	10%
	2 (CBJ)	7,86	0,21	11,68	0,40	10%
M10	1 (CSP)	7,04	0,33	10,60	0,33	11%
	7 (CLA)	7,30	0,40	10,93	0,31	12%
M11	2 (CBJ)	7,70	0,28	11,72	0,42	10%
	7 (CLA)	7,03	0,21	11,02	0,32	12%
M12	1 (CSP)	6,85	0,11	10,41	0,30	10%
	2 (CBJ)	7,53	0,24	11,43	0,44	9%
	7 (CLA)	6,95	0,07	10,80	0,23	12%
M13	1 (CSP)	6,84	0,09	10,56	0,19	10%
	8 (HOR)	11,06	0,30	15,56	0,30	11%
M14	1 (CSP)	6,90	0,08	10,42	0,21	10%
	5 (CSG)	52,91	0,75	77,76	1,35	6%
	8 (HOR)	10,76	0,25	15,37	0,34	10%

Fonte: A Autora (2025).

Nos modelos compostos por duas usinas que compartilham o mesmo conjunto de variáveis meteorológicas (M9, M10 e M11), observa-se que os valores médios de MAE se situam, em geral, na faixa de 6,8 MW a 7,9 MW, com nMAE variando entre 10% e 12%. Os desvios-padrão (dp-MAE e dp-RMSE) são relativamente baixos (tipicamente inferiores a 0,5 MW), o que sugere uma variabilidade moderada dos erros entre as dez inicializações. Esses resultados evidenciam que a arquitetura TFT é capaz de aprender, de forma conjunta, o

comportamento de usinas no mesmo campo meteorológico sem perda de precisão, mantendo níveis de erro semelhantes aos obtidos na modelagem individual (como visto na Tabela 13).

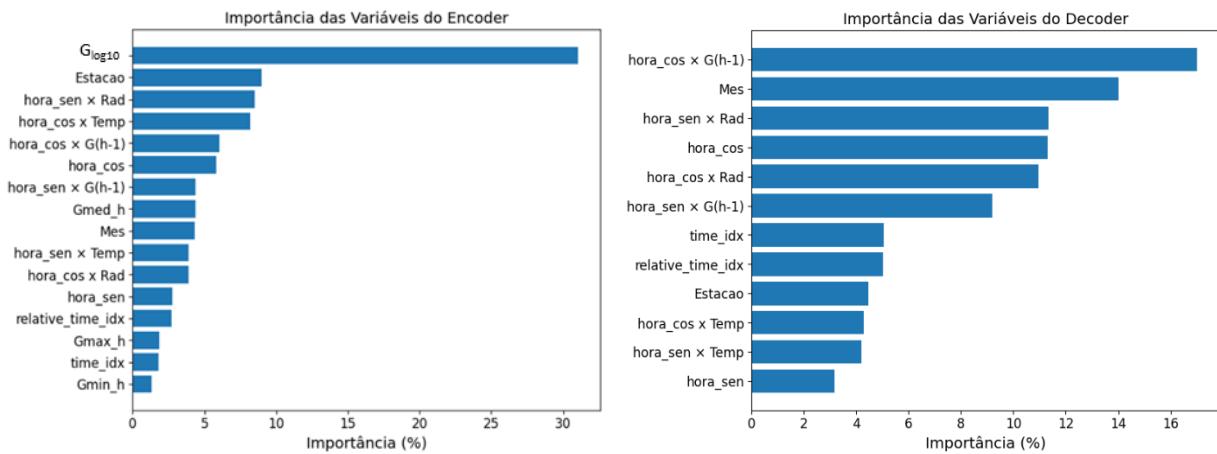
No caso do modelo M13, que passa a considerar duas usinas associadas a conjuntos distintos de dados meteorológicos, observa-se um comportamento ligeiramente mais heterogêneo entre as séries: enquanto uma das usinas mantém MAE em torno de 6,84 MW e nMAE de 10%, a outra apresenta MAE mais elevado (11,06 MW) e RMSE de 15,56 MW, embora com nMAE ainda em patamar próximo (11%). Essa diferença sugere que, quando o modelo passa a lidar simultaneamente com usinas submetidas a condições climáticas e bases meteorológicas distintas, a tarefa de previsão torna-se mais desafiadora, ainda que o desempenho relativo (em termos percentuais) permaneça adequado.

Nos modelos com três usinas, a configuração M12 representa o cenário em que todas as usinas compartilham o mesmo conjunto de variáveis meteorológicas. Nessa configuração, os valores de MAE variam aproximadamente entre 6,85 e 7,53 MW, com nMAE entre 9% e 12%. Os desvios-padrão de MAE e RMSE são particularmente baixos (dp-MAE entre 0,07 e 0,24 MW), o que indica um comportamento bastante estável entre as inicializações. Esses resultados mostram que o TFT é capaz de acomodar a previsão conjunta de três usinas sob o mesmo regime meteorológico sem aumento expressivo dos erros médios, mantendo desempenho semelhante ao observado nos modelos com duas usinas.

Por outro lado, o modelo M14 agrupa três usinas associadas a bases meteorológicas distintas, incluindo uma usina de grande porte. Nessa configuração, observa-se maior heterogeneidade nos erros absolutos: enquanto duas usinas mantêm MAE em faixas próximas às anteriores (cerca de 6,90 e 10,76 MW, com nMAE de 10%), a usina de maior capacidade apresenta MAE e RMSE bastante elevados em termos absolutos (52,91 MW e 77,76 MW, respectivamente), embora com nMAE relativamente baixo (6%). Esse comportamento indica que, para usinas de grande porte, a escala em MW amplifica naturalmente o erro absoluto, de modo que a métrica normalizada (nMAE) torna-se mais adequada para comparação entre empreendimentos de diferentes tamanhos. Ainda assim, os desvios-padrão permanecem controlados, sugerindo que o treinamento multipreditivo é estável mesmo nesse cenário mais complexo.

Quanto à interpretabilidade dos modelos multipreditivos, selecionou-se o modelo M12 como estudo de caso por representar uma das configurações mais completas de entradas; nota-se que os demais apresentam comportamento semelhante. Assim, a Figura 30 ilustra a importância das variáveis para a previsão realizada pelo modelo M12.

Figura 30 - Interpretabilidade das Variáveis do Modelo M12



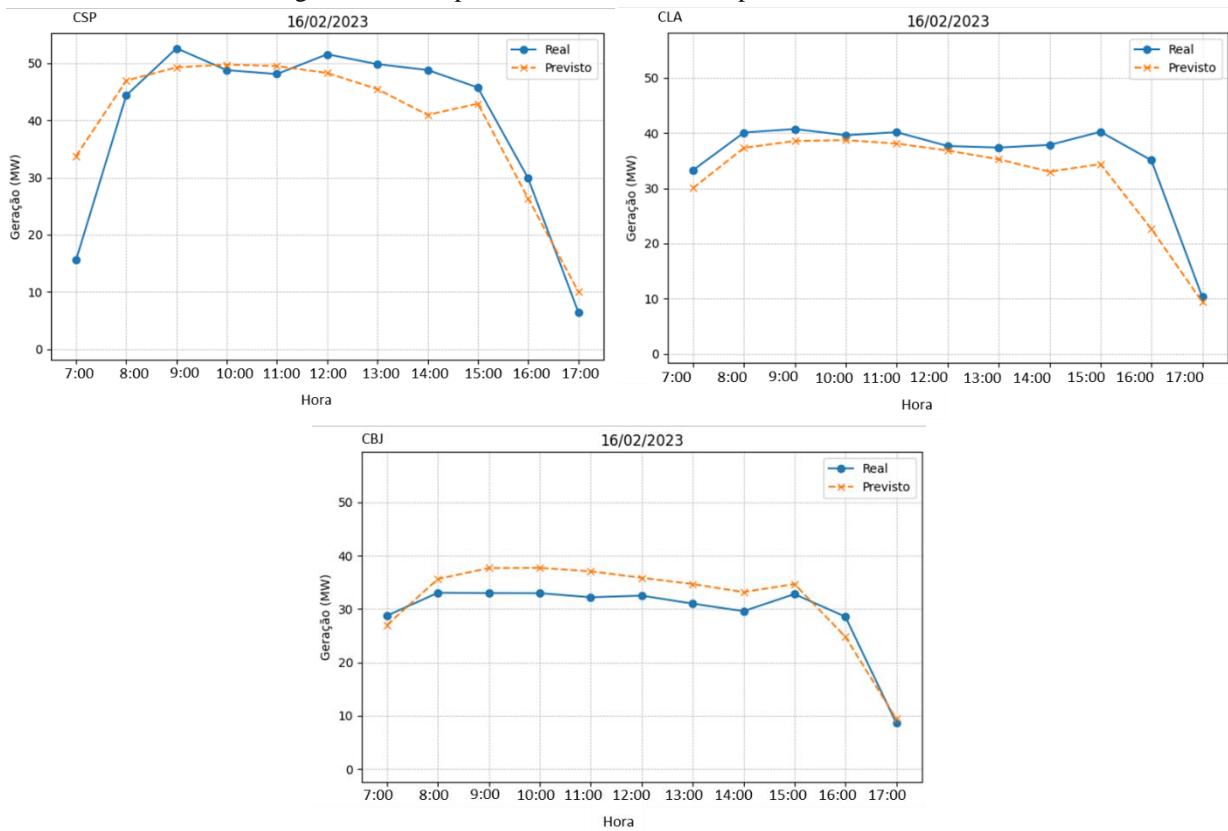
Fonte: A Autora (2025).

Quanto ao Encoder do modelo M12, observa-se que quase um terço da sua atenção é conferida à variável-alvo transformada de geração (G_{log10}), com cerca de 30%, o que confirma a primazia do histórico imediato da série no estabelecimento do estado anterior. Em seguida, a covariável *Estacao* assume papel relevante (~ 13%), pois a atenção visualiza padrões repetidos nas três usinas 1(CSP), 2 (CBJ) e 3 (CLA), tornando fatores de calendário mais relevantes para distinguir variações que não são capturadas por defasagens de tempo. A interação (*hora_sen* \times *Rad*) aparece logo após (~ 11%), indicando a importância do vínculo entre o ciclo diário e o recurso solar.

No Decoder, a maior parte da atenção (~ 17%) é dedicada à interação entre a hora e o valor de geração do passo imediatamente anterior (*hora_cos* \times $G(h - 1)$), sinalizando que este é o principal guia para cada previsão. Em seguida, a variável *Mes* (~ 14%) e a interação (*hora_sen* \times *Rad*) (~ 12%) se destacam. Esse rearranjo demonstra como o TFT combina informações de curto prazo com marcadores de calendário para refinar cada previsão horária, adaptando seu foco da construção de contexto (Encoder) para a geração de previsão (Decoder).

A Figura 31 ilustra a comparação entre a geração real e a previsão pontual (quantil 0,5) do modelo M12 para o dia 16/02/2023, considerando as 11 horas diárias utilizadas neste estudo, para as três usinas consideradas (CSP, CLA e CBJ). Esse dia foi selecionado como exemplo em que o perfil diário de geração FV não apresenta um comportamento “ideal”, observando-se discrepâncias pontuais possivelmente associadas à variabilidade intradiária do recurso solar.

Figura 31 – Comparativo 1 Real x Previsto para o Modelo M12



Fonte: A Autora (2025).

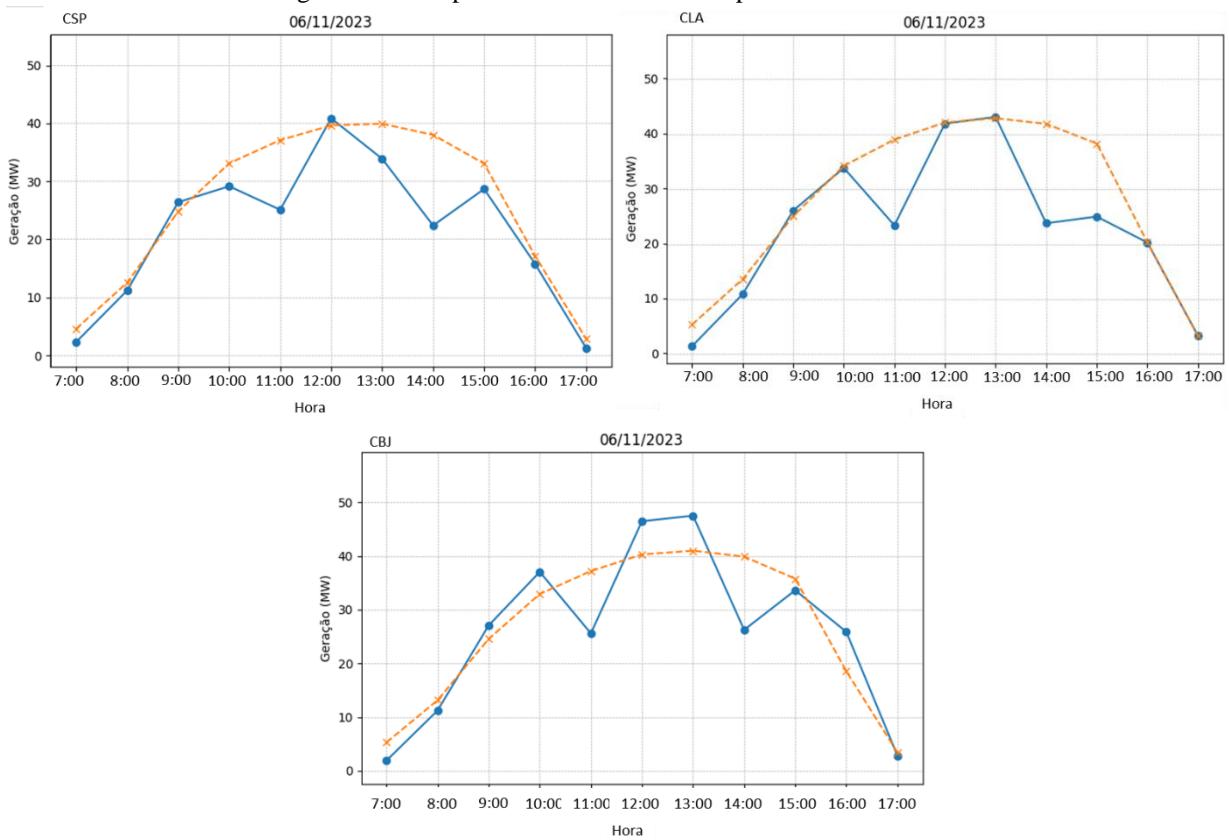
Em CSP, o modelo tende a suavizar o pico de geração e subestimar parte do período da tarde, embora permaneça próximo dos valores observados. Em CLA, nota-se um ligeiro deslocamento para baixo da curva prevista em relação à real, principalmente a partir do meio do dia, indicando subestimação moderada da produção. Já em CBJ, o comportamento é oposto pois o modelo superestima a geração ao longo do dia, mas acompanha bem o início e o final do período diurno. Essas discrepâncias pontuais ilustram a variabilidade intradiária do recurso solar e mostram que, embora o perfil diário não seja perfeitamente “ideal”, o modelo M12 consegue manter boa coerência global entre as previsões e as séries observadas nas três usinas.

A Figura 32, por sua vez, ilustra a comparação entre a geração real e a previsão pontual (quantil 0,5) do modelo M12 para o dia 06/11/2023. Observa-se que o modelo M12 continua reproduzindo o formato geral da curva diária de geração das três usinas (CSP, CLA e CBJ).

Nesse dia, em específico, verificam-se quedas abruptas de geração em horários centrais do dia, seguidas de recuperações rápidas, que não são acompanhadas pelas três curvas previstas, a qual mantém um perfil mais estável. Essas discrepâncias locais, especialmente em horários nos quais seria esperado um patamar quase constante de geração, podem estar associadas a

eventos operativos não estritamente relacionados à radiação solar, como episódios de *curtailment* ou restrições impostas pelo sistema elétrico. Nesses casos, o modelo tende a seguir o padrão médio condicionado às variáveis meteorológicas, o que explica o descolamento pontual entre a previsão e os valores medidos, mesmo mantendo boa aderência ao formato global da curva diária.

Figura 32 - Comparativo 2 Real x Previsto para o Modelo M12



Fonte: A Autora (2025).

5.4.2 Testes de significância estatística

Com o objetivo de complementar a análise descritiva, esta subseção investiga se as diferenças de desempenho entre os modelos M9 a M14 são estatisticamente relevantes ou se podem ser atribuídas apenas à variabilidade decorrente das diferentes inicializações. Para o estudo estatístico, foram considerados os modelos M9, M10, M12, M13 e M14, que incluem a Usina 1 (CSP) em sua configuração. Em cada um desses modelos, utilizaram-se as dez observações de MAE obtidas nas *runs* independentes, formando amostras pareadas de erro para cada par de modelos comparado.

A Tabela 18 apresenta os valores de MAE de cada um dos cinco modelos em cada uma das dez inicializações. Essa matriz de resultados permite visualizar, de maneira mais detalhada, o comportamento de cada configuração ao longo das diferentes execuções.

Tabela 18 - Resultados de MAE por *run* para os modelos M9 a M14

<i>run</i>	MAE (MW)				
	M9	M10	M12	M13	M14
0	6,87	6,89	6,77	6,86	6,99
1	7,08	7,05	6,69	6,92	6,93
2	6,98	7,78	7,07	7,00	6,85
3	6,98	6,92	6,93	6,91	6,87
4	6,8	6,69	6,91	6,72	6,92
5	7,1	7,43	6,86	6,79	6,76
6	6,67	7,03	6,9	6,81	6,89
7	6,7	6,75	6,91	6,8	6,95
8	6,71	7,00	6,67	6,73	7,01
9	6,74	6,91	6,87	6,86	6,82
Média	6,86	7,05	6,86	6,84	6,90
dp	0,17	0,34	0,12	0,09	0,08

Fonte: A Autora (2025).

A partir dos valores de MAE por inicialização, foram construídas as comparações dois a dois entre os modelos de interesse, a saber: M9 vs. M10, M9 vs. M13, M10 vs. M13, M12 vs. M14 e, por fim, M12 vs. M13, sendo estes os dois modelos com menores MAE médios e dp-MAE. Para cada par, calculou-se: (i) a diferença média de MAE entre os modelos; (ii) o intervalo de confiança (IC95%) dessa diferença; e (iii) os testes t pareado e de Wilcoxon. A Tabela 19 apresenta os resultados do teste de significância dois a dois dos modelos M9 a M14.

Tabela 19 – Teste de significância estatística pareada dos modelos M9 a M14

Par	Diferença média	IC 95%	<i>p</i> Teste t	<i>p</i> Teste Wilcoxon
M9 vs. M10	-0,18	[-0,38 – 0,02]	0,0657	0,1055
M9 vs. M13	0,04	[-0,06 – 0,14]	0,4082	0,4922
M10 vs. M13	0,22	[-0,01 – 0,45]	0,0599	0,0371
M12 vs. M13	0,02	[-0,07 – 0,10]	0,6445	0,4453
M12 vs. M14	-0,04	[-0,17 – 0,08]	0,4741	0,7891

Fonte: A Autora (2025).

De modo geral, observa-se que as diferenças médias de erro entre os pares avaliados são pequenas em magnitude (entre -0,18 e 0,22 MW) e que, na maior parte dos casos, os intervalos de confiança de 95% incluem o zero, com *p*-valores acima de 0,05 tanto para o

teste t pareado quanto para o teste de Wilcoxon. Isso indica que, para os pares M9 vs. M10, M9 vs. M13, M12 vs. M13 e M12 vs. M14, não há evidência estatística de que um modelo supere consistentemente o outro em termos de MAE, sendo as diferenças observadas compatíveis com a variabilidade associada às diferentes inicializações.

O único caso em que se observa algum indício de diferença é o par M10 vs. M13, cuja diferença média de 0,22 MW (M10 – M13) sugere um MAE ligeiramente maior para M10, com o teste de Wilcoxon indicando significância ao nível de 5% ($p = 0,0371$), enquanto o teste t apresenta um p-valor ligeiramente acima desse limiar ($p = 0,0599$). Esse resultado aponta para uma possível vantagem do modelo M13 em relação a M10, porém com evidência estatística mais fraca e dependente do teste considerado, o que recomenda uma interpretação cautelosa.

Em síntese, a análise estatística dos modelos multipreditivos revela que, na maior parte das comparações, o TFT apresenta desempenho semelhante entre as diferentes configurações (pares e trios de usinas, com bases meteorológicas compartilhadas ou distintas). As diferenças numéricas em MAE observadas na Tabela 19 não se traduzem, em geral, em superioridade estatisticamente robusta de um modelo sobre outro, sugerindo que o esquema multipreditivo proposto é capaz de manter um patamar de desempenho relativamente estável frente às variações na composição dos conjuntos de usinas e de covariáveis meteorológicas.

5.5 Comparativo com outras Topologias

Para avaliar o desempenho da arquitetura TFT proposta, realizou-se uma análise comparativa com as topologias Persistência, MLP e LSTM. A Tabela 20 descreve os hiperparâmetros que foram considerados nos treinamentos das duas redes MLP e LSTM, os quais foram definidos por meio da técnica de busca em grade (*grid search*). Já a configuração da Persistência adota como valor previsto a observação imediatamente anterior à janela de previsão.

Tabela 20 - Hiperparâmetros dos modelos MLP e LSTM

Hiperparâmetro	Valor	Hiperparâmetro	Valor
MLP			
Activation	$SiLU \left(SiLU = x \left(\frac{1}{1+e^{-x}} \right) \right)$	Dropout	0,0
Batch_Size	128	Patience	15
Otimizador	AdamW	Min_delta	1e-4

Learning Rate	5e-4	Épocas	200
LSTM			
Layers	1	Dropout	0,1
Batch_Size	128	Patience	15
Otimizador	AdamW	Min_delta	1e-4
Learning Rate	5e-4	Épocas	200
WeightDecay	0,0001		

Fonte: A Autora (2025).

Foram treinados os modelos M7, M9 e M12 utilizando as topologias Persistência, MLP e LSTM e os resultados médios das métricas de desempenho obtidos são apresentados na Tabela 21.

Tabela 21 – Comparativo do Desempenho das Topologias TFT, MLP e LSTM

Modelo	Usinas	MLP					LSTM				
		MAE (MW)	dp- MAE (MW)	RMSE (MW)	dp- RMSE (MW)	nMAE (%)	MAE (MW)	dp- MAE (MW)	RMSE (MW)	dp- RMSE (MW)	nMAE (%)
M7	CSP	9,29	0,11	11,99	0,08	14%	10,08	0,16	12,82	0,13	15%
M9	CSP/CBJ	10,15	0,13	12,95	0,09	15%	10,17	0,16	13,04	0,12	15%
M12	CSP/CBJ/CLA	10,20	0,10	13,00	0,10	15%	10,39	0,19	13,07	0,10	15%
Modelo	Usinas	TFT					Persistência				
		MAE (MW)	dp- MAE (MW)	RMSE (MW)	dp- RMSE (MW)	nMAE (%)	MAE (MW)	RMSE (MW)	nMAE (%)		
M7	CSP	7,13	0,22	10,64	0,31	10%					
M9	CSP/CBJ	6,86	0,16	10,40	0,27	10%	26,77	30,58	40%		
M12	CSP/CBJ/CLA	6,85	0,11	10,41	0,30	10%					

Fonte: A Autora (2025).

A Tabela 21 compara o desempenho das três topologias de redes neurais (MLP, LSTM e TFT) e do método de Persistência, para três configurações representativas, modelo univariado com uma usina (M7 – CSP), modelo multipreditivo com duas usinas (M9 – CSP/CBJ) e modelo multipreditivo com três usinas (M12 – CSP/CBJ/CLA). Em todas essas configurações, o TFT apresenta os menores valores médios de MAE e RMSE, além dos menores nMAE ($\approx 10\%$), indicando clara vantagem em termos de precisão em relação às demais redes e ao modelo de Persistência.

No modelo M7 (CSP), por exemplo, o TFT atinge MAE de 7,13 MW e RMSE de 10,64 MW (nMAE = 10%), enquanto o MLP e o LSTM apresentam erros mais elevados, com MAE de 9,29 MW e 10,08 MW, e nMAE de 14% e 15%, respectivamente. Tendência semelhante é observada nos modelos multipreditivos, em M9 (CSP/CBJ) e M12 (CSP/CBJ/CLA), o TFT mantém MAE em torno de 6,85–6,86 MW e RMSE próximos de 10,4 MW, com nMAE de 10%, ao passo que MLP e LSTM apresentam MAE em torno de 10 MW e nMAE em torno de 15%. Os desvios-padrão (dp-MAE e dp-RMSE) são baixos em todas as topologias, sugerindo que as diferenças observadas refletem, de fato, maior capacidade preditiva do TFT e não apenas flutuações entre inicializações.

Ao comparar com o método de Persistência, a superioridade do TFT torna-se ainda mais evidente, pois, enquanto o TFT apresenta MAE de 6,86 MW e nMAE de 10%, a Persistência atinge MAE de 26,77 MW, RMSE de 30,58 MW e nMAE de 40%, demonstrando que, para horizontes de 11 horas à frente, a utilização de modelos de aprendizado profundo é essencial para reduzir os erros de previsão a patamares aceitáveis.

Entre MLP e LSTM, os resultados indicam desempenhos muito próximos, com leve vantagem numérica para a MLP na maioria das configurações (MAE e RMSE ligeiramente menores). Isso sugere que, no contexto específico desta base de dados e horizonte de previsão, a complexidade adicional da LSTM não se traduz em ganhos substanciais sobre a MLP, ao passo que o TFT, ao explorar mecanismos de atenção e o uso estruturado de covariáveis temporais e estáticas, obtém melhorias consistentes em todas as configurações analisadas.

6 CONCLUSÃO

Esta tese investigou a previsão horária da geração FV em múltiplas usinas brasileiras a partir da combinação entre clusterização hierárquica (DTW) e a arquitetura TFT. De forma geral, os resultados mostraram que é possível construir um sistema de previsão multipreditivo com bom desempenho, interpretável e estatisticamente robusto, mesmo em um contexto de alta variabilidade solar e heterogeneidade entre usinas.

Do ponto de vista do pré-processamento de dados, um primeiro achado relevante foi a eficácia do método de imputação, nas séries de geração, guiado pela curva de radiação solar global. Ao incorporar diretamente a informação meteorológica na reconstrução de lacunas, essa abordagem superou, de forma consistente, técnicas clássicas como interpolação linear, média temporal e KNN.

A aplicação da clusterização hierárquica com a métrica DTW às séries de geração permitiu identificar grupos de usinas com padrões de produção semelhantes, e esses agrupamentos foram explorados na construção dos modelos multipreditivos. Os resultados indicaram que, quando usinas de um mesmo cluster e sujeitas ao mesmo conjunto de variáveis meteorológicas são previstas em conjunto, o TFT mantém erros absolutos e normalizados semelhantes aos de modelos univariados, com baixa variabilidade entre inicializações. Em contrapartida, agrupamentos que combinam usinas de clusters distintos e bases meteorológicas diferentes tendem a produzir maior heterogeneidade de desempenho entre as usinas, ainda que os erros normalizados permaneçam em níveis adequados. Esses achados sugerem que a similaridade estrutural entre as séries e a coerência das covariáveis são fatores importantes para o desenho de modelos multipreditivos mais equilibrados.

No que se refere à modelagem com TFT, a análise sistemática das 14 configurações avaliadas mostrou que a inclusão de variáveis meteorológicas e atributos derivados da própria série de geração é decisiva para reduzir os erros de previsão, enquanto o efeito isolado da escolha de escala (MW ou log10) tende a perder importância à medida que o conjunto de covariáveis se torna mais rico. A análise estatística pareada (testes t e Wilcoxon) confirmou que apenas em configurações mais simples houve diferença significativa entre escalas, ao passo que, nos modelos mais completos, as diferenças numéricas entre variantes em MW e log10 não se traduziram em superioridade estatisticamente robusta. Isso indica que a qualidade e a diversidade das covariáveis têm impacto mais consistente sobre o desempenho do que transformações pontuais da variável-alvo.

Adicionalmente, um resultado particularmente relevante diz respeito à comparação entre os modelos univariados e multipreditivos, visto que os erros das configurações multipreditivas mantiveram-se em patamares muito próximos aos das abordagens univariadas, mesmo com a previsão simultânea de duas ou três usinas. Isso indica que o desempenho preditivo individual não sofre decréscimo significativo ao adotar-se a modelagem conjunta. Na prática, essa característica permite que uma única arquitetura TFT gerencie múltiplas usinas, o que reduz o número de modelos a serem treinados e mantidos, simplificando a implantação em centros de controle e favorecendo uma visão integrada da geração FV no sistema elétrico brasileiro.

A comparação entre TFT, MLP, LSTM e Persistência constitui outra contribuição importante. Em todas as configurações representativas analisadas, o TFT apresentou, de forma consistente, os menores valores de MAE, RMSE e nMAE, superando as redes MLP e LSTM e distanciando-se fortemente do modelo de Persistência. Além disso, MLP e LSTM mostraram desempenhos muito próximos entre si, o que sugere que, neste contexto específico, a simples adoção de uma arquitetura recorrente não garante ganhos significativos frente a redes densas, enquanto o TFT, ao combinar atenção multi-cabeça, tratamento estruturado de covariáveis e mecanismo encoder-decoder temporal, oferece ganhos tangíveis em precisão. Assim, o trabalho reforça o TFT como uma alternativa para previsão FV, com desempenho superior a arquiteturas amplamente utilizadas na literatura.

Por fim, a análise de interpretabilidade do TFT mostrou que o modelo faz uso coerente das informações disponíveis: no encoder, a importância recai majoritariamente sobre o histórico de geração, sobre a interação com a radiação solar, sobre variáveis sazonais e de padrão diário; no decoder, ganham destaque defasagens imediatas e interações com radiação solar e temperatura ambiente. Esse comportamento é compatível com a natureza do problema físico, no qual tanto a memória recente quanto a sazonalidade e as condições meteorológicas locais desempenham papel central na determinação da geração FV. Dessa forma, além de preciso, o modelo proposto é interpretável em termos energéticos, aspecto fundamental para sua adoção em ambientes operacionais do setor elétrico.

Em síntese, as principais contribuições deste trabalho podem ser assim destacadas: (i) a proposição de um método de imputação de dados faltantes baseado em radiação, adequado a séries de geração FV; (ii) a utilização da clusterização hierárquica DTW como ferramenta para estruturar modelos multipreditivos em grupos de usinas; (iii) a construção e avaliação sistemática de um conjunto de modelos TFT univariados e multipreditivos, abrangendo análises de desempenho médio, variabilidade e significância estatística das diferenças observadas; (iv) a demonstração quantitativa da superioridade do TFT em relação a MLP,

LSTM e Persistência em cenários realistas de previsão; e (v) a exploração da interpretabilidade do TFT, evidenciando como o modelo combina informação histórica, sazonal e meteorológica para produzir previsões consistentes. Juntos, esses resultados reforçam o potencial da abordagem proposta para apoiar o planejamento e a operação de sistemas elétricos com elevada penetração de geração FV, ao mesmo tempo em que apontam caminhos promissores para o aperfeiçoamento de modelos preditivos baseados em aprendizado profundo.

Apesar dos resultados promissores, este estudo apresenta limitações que delimitam seu escopo e apontam caminhos para pesquisas futuras. Primeiramente, as variáveis meteorológicas foram obtidas exclusivamente da base NASA POWER; embora global e acessível, trata-se de uma estimativa via satélite que pode não capturar fenômenos de microclima com a mesma precisão de uma estação meteorológica no local. Além disso, a modelagem concentrou-se em um horizonte fixo de 11 horas. Extensões naturais deste trabalho incluem a avaliação de horizontes múltiplos, que são críticas para a estabilidade da rede em curtíssimo prazo. Por fim, o modelo não incorporou variáveis operativas, como estados de manutenção ou restrições de despacho (*curtailment*), que poderiam explicar discrepâncias residuais entre a geração real e a prevista. A integração dessas dimensões constitui uma linha promissora para o aperfeiçoamento da previsão de geração FV e para a sua aplicação direta em decisões operacionais no sistema elétrico brasileiro.

6.1 Sugestões para trabalhos futuros

Com relação à continuidade da pesquisa, sugere-se, para pesquisas futuras:

- a) Integrar dados meteorológicos de sensoriamento local para aumentar o desempenho da modelagem TFT;
- b) Avaliar os modelos em horizontes de previsão múltiplos, mais curtos e mais longos que o horizonte de 11 horas adotado neste trabalho, a fim de investigar a precisão da arquitetura TFT em diferentes escalas temporais;
- c) Incorporar variáveis operativas do setor elétrico (por exemplo, estados de manutenção e *curtailment*) nos modelos preditivos, de forma a capturar discrepâncias nas séries de geração FV;
- d) Explorar arquiteturas híbridas que combinem o TFT com outras topologias ou modelos físicos, avaliando possíveis ganhos adicionais de desempenho.

REFERÊNCIAS

- ABDEL-NASSER, M.; MAHMOUD, K. Accurate photovoltaic power forecasting models using deep LSTM-RNN. **Neural Computing and Applications**, v. 31, n. 7, p. 2727-2740, 2019.
- AHMED, R. *et al.* A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization. **Renewable and Sustainable Energy Reviews**, v. 124, p. 109792, 2020.
- AL-ALI, E. M. *et al.* Solar energy production forecasting based on a hybrid CNN-LSTM-transformer model. **Mathematics**, v. 11, n. 3, p. 676, 2023.
- ALZUBAIDI, L. *et al.* Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. **Journal of Big Data**, v. 8, n. 1, p. 1-74, 2021.
- BHOLOWALIA, P.; KUMAR, A. EBK-means: A clustering technique based on elbow method and k-means in WSN. **International Journal of Computer Applications**, v. 105, n. 9, 2014.
- BIANCHI, F. M. *et al.* An overview and comparative analysis of recurrent neural networks for short term load forecasting. **arXiv preprint arXiv:1705.04378**, 2017.
- CHEN, Si-An *et al.* Tsmixer: An all-mlp architecture for time series forecasting. **arXiv preprint arXiv:2303.06053**, 2023.
- CHOLLET, F. **Deep learning with Python**. 2. ed. New York: Manning Publications, 2021.
- CUI, S. *et al.* Residential virtual power plant with photovoltaic output forecasting and demand response. **Asian Journal of Control**, v. 21, n. 4, p. 1906-1917, 2019.
- CUNHA, B. A. **Previsão intra-diária de geração fotovoltaica usando redes neurais recorrentes do tipo LSTM e dados históricos de energia**. 2021. Tese (Doutorado em Engenharia Elétrica) – Universidade Estadual Paulista, Bauru, 2021.
- DAS, U. K. *et al.* Forecasting of photovoltaic power generation and model optimization: A review. **Renewable and Sustainable Energy Reviews**, v. 81, p. 912-928, 2018.
- DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, n. 2, p. 224-227, 1979.
- EMPRESA DE PESQUISA ENERGÉTICA (Brasil). **Plano Nacional de Energia 2050**. Brasília, DF: MME, 2020.
- EMPRESA DE PESQUISA ENERGÉTICA (Brasil). **Plano Decenal de Expansão de Energia 2035**. Rio de Janeiro: EPE, 2025. Disponível em: [\[https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/plano-decenal-de-expansao-de-energia-2035\]](https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/plano-decenal-de-expansao-de-energia-2035). Acesso em: 27 jul. 2025.
- FACCIN, C. *et al.* Comparison of Climate Data from NASA POWER Reanalysis and Data Measured at Surface Weather Stations for Application in Brazilian Paving Projects. **Anuário do Instituto de Geociências**, v. 47, 2024.

FERREIRA, A. BA; LEITE, J. B.; SALVADEO, D. HP. Power substation load forecasting using interpretable transformer-based temporal fusion neural networks. **Electric Power Systems Research**, v. 238, p. 111169, 2025.

FLORES, A.; TITO, H.; SILVA, C. Local average of nearest neighbors: Univariate time series imputation. **International Journal of Advanced Computer Science and Applications**, v. 10, n. 8, 2019.

FRANCISCO, A. *et al.* Influência de parâmetros meteorológicos na geração de energia em painéis fotovoltaicos: um caso de estudo do Smart Campus Facens, SP, Brasil. **urbe. Revista Brasileira de Gestão Urbana**, v. 11, 2019.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. 2. ed. Cambridge: MIT press, 2016.

GURKAYNAK, G.; YILMAZ, I.; HAKSEVER, G. Stifling artificial intelligence: Human perils. **Computer Law & Security Review**, v. 32, n. 5, p. 749-758, 2016.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

HAO, Y.; TIAN, C. A novel two-stage forecasting model based on error factor and ensemble method for multi-step wind power forecasting. **Applied Energy**, v. 238, p. 368-383, 2019.

HAUGELAND, J. **Artificial Intelligence: The Very Idea**. 1. ed. Massachusetts: MIT Press, 1989.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735-1780, 1997.

HOTA, H. S.; HANNA, R.; SHRIVAS, A. K. Time series data prediction using sliding window-based RBF neural network. **International Journal of Computational Intelligence Research**, v. 13, n. 5, p. 1145-1156, 2017.

IKOTUN, A. *et al.* K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. **Information Sciences**, v. 622, p. 178-210, 2023.

INTERNATIONAL RENEWABLE ENERGY AGENCY (IRENA). **Renewable Capacity Statistics 2025**. Abu Dhabi, 2025.

ISLAM, M. *et al.* Forecasting of PV plant output using interpretable temporal fusion transformer model. In: 2023 **North American Power Symposium (NAPS)**. IEEE, 2023. p. 01-06.

IVO GOLTZ, B. *et al.* Comparação de métodos de preenchimento de dados de fluxo de CO₂. **Revista Ciência e Natura**, v. 45, 2023.

LAZCANO, A.; JARAMILLO-MORÁN, M. A.; SANDUBETE, J. E. Back to basics: The power of the multilayer perceptron in financial time series forecasting. **Mathematics**, v. 12, n. 12, p. 1920, 2024.

LAZZERI, F. **Machine learning for time series forecasting with Python**. 1. ed. Indianapolis: John Wiley & Sons, 2021.

- LEE, D.; KIM, K. Recurrent neural network-based hourly prediction of photovoltaic power output using meteorological information. **Energies**, v. 12, n. 2, p. 215, 2019.
- LIAO, T. W. Clustering of time series data—a survey. **Pattern recognition**, v. 38, n. 11, p. 1857-1874, 2005.
- LIM, B. *et al.* Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. **International Journal of Forecasting**, v. 37, n. 4, p. 1748-1764, 2021.
- LING, N. *et al.* Missing responses at random in functional single index model for time series data. **Statistical Papers**, p. 1-28, 2022.
- LIU, B. Y.H.; JORDAN, R. C. The long-term average performance of flat-plate solar-energy collectors: with design data for the US, its outlying possessions and Canada. **Solar Energy**, v. 7, n. 2, p. 53-74, 1963.
- LLUGSI, R. *et al.* Comparison between Adam, AdaMax and Adam W optimizers to implement a Weather Forecast based on Neural Networks for the Andean city of Quito. In: **2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM)**. IEEE, 2021. p. 1-6.
- LÓPEZ SANTOS, M. *et al.* Application of temporal fusion transformer for day-ahead PV power forecasting. **Energies**, v. 15, n. 14, p. 5232, 2022.
- MITCHELL, T. **Machine Learning**. 1. ed. Ohio: McGraw Hill, 1997.
- MÜLLER, M. Dynamic Time Warping. **Information retrieval for music and motion**, Berlin: Springer, 2007. p. 69-84.
- NESPOLI, A. *et al.* Day-ahead photovoltaic forecasting: A comparison of the most effective techniques. **Energies**, v. 12, n. 9, p. 1621, 2019.
- PAIVA, G. M. **Métodos de aprendizagem de máquinas aplicados em previsão solar intra-díaria para operação de sistemas de potência**. 2021. 105 f. Tese (Doutorado em Engenharia Elétrica e da Computação) - Universidade Federal de Goiás, Goiânia, 2021.
- PARK, J. *et al.* Long-term missing value imputation for time series data using deep neural networks. **Neural Computing and Applications**, v. 35, n. 12, p. 9071-9091, 2023.
- PEREIRA, Enio Bueno *et al.* **Atlas Brasileiro de Energia Solar**. 2. ed. São José dos Campos: INPE, 2017. 80 p. E-book. Disponível em: <http://doi.org/10.34024/978851700089>. Acesso em: 22 abr. 2025.
- PHAN, Q.; WU, Y.; PHAN, Q. An approach using transformer-based model for short-term PV generation forecasting. In: **2022 8th International Conference on Applied System Innovation (ICASI)**. IEEE, 2022. p. 17-20.
- PHAN, Q.; WU, Y.; PHAN, Q. Application of a new Transformer-based model and XGBoost to improve one-day-ahead solar power forecasts. In: **2023 IEEE/IAS 59th Industrial and Commercial Power Systems Technical Conference (I&CPS)**. IEEE, 2023. p. 1-7.
- REN, J.; WANG, H. Calculus and Optimization. In: **Mathematical Methods in Data Science**. Elsevier, 2023. p. 51-89.

- RUMELHART, D.; HINTON, G.; WILLIAMS, R. Learning representations by back-propagating errors. **Nature**, v. 323, n. 6088, p. 533-536, 1986.
- SAKOE, H.; CHIBA, S. Dynamic programming algorithm optimization for spoken word recognition. **IEEE transactions on acoustics, speech, and signal processing**, v. 26, n. 1, p. 43-49, 1978.
- SHAHAPURE, K.; NICHOLAS, C. Cluster quality analysis using silhouette score. In: **2020 IEEE 7th international conference on data science and advanced analytics (DSAA)**. IEEE, 2020. p. 747-748.
- SHI, J. *et al.* Forecasting power output of photovoltaic systems based on weather classification and support vector machines. **IEEE Transactions on Industry Applications**, v. 48, n. 3, p. 1064-1069, 2012.
- SINGLA, P.; DUHAN, M.; SAROHA, S. A comprehensive review and analysis of solar forecasting techniques. **Frontiers in Energy**, p. 1-37, 2021.
- SOARES, L. D. **Redes neurais artificiais BIGRU_CNN aplicadas à previsão de demanda de energia elétrica de curto prazo**. 2021. Dissertação (Mestrado em Engenharia Elétrica e Computação) - Universidade Estadual do Oeste do Paraná, Foz do Iguaçu, 2021.
- TAN, P.; STEINBACH, M.; KARPATNE, A.; KUMAR, V. **Introduction to Data Mining**. 2. ed. Harlow: Pearson, 2019.
- TAO, K. *et al.* Operational day-ahead photovoltaic power forecasting based on transformer variant. **Applied Energy**, v. 373, p. 123825, 2024.
- TORRES, J. F. *et al.* Deep learning for time series forecasting: a survey. **Big Data**, v. 9, n. 1, p. 3-21, 2020.
- VALDATI, A. B. **Inteligência Artificial**. 1. ed. Curitiba: Contentus, 2020.
- VAN DER MEER, D. W. *et al.* Probabilistic forecasting of electricity consumption, photovoltaic power generation and net demand of an individual building using Gaussian Processes. **Applied Energy**, v. 213, p. 195-207, 2018.
- VASWANI, A *et al.* Attention is all you need. **Advances in Neural Information Processing Systems**, 2017.
- VELASCO-GALLEGOS, C.; LAZAKIS, I. A novel framework for imputing large gaps of missing values from time series sensor data of marine machinery systems. **Ships and Offshore Structures**, v. 17, n. 8, p. 1802-1811, 2022.
- WANG, F. *et al.* Wavelet decomposition and convolutional LSTM networks based improved deep learning model for solar irradiance forecasting. **Applied Sciences**, v. 8, n. 8, p. 1286, 2018.
- WANG, H. *et al.* A review of deep learning for renewable energy forecasting. **Energy Conversion and Management**, v. 198, 2019.

- WANG, K.; QI, X.; LIU, H. A comparison of day-ahead photovoltaic power forecasting models based on deep learning neural network. **Applied Energy**, v. 251, 2019.
- WU, J. **Advances in K-Means Clustering**: a data mining thinking. 1. ed. Beijing: Springer, 2012.
- XU, R.; WUNSCH, D. **Clustering**. 1. ed. John Wiley & Sons, 2008.
- ZAINUDDIN, A. *et al.* Time Series Data and Recent Imputation Techniques for Missing Data: A Review. In: **2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)**. IEEE, 2022. p. 346-350.
- ZENG, A. *et al.* Are Transformers Effective for Time Series Forecasting? In: **Proceedings of the AAAI conference on artificial intelligence**. 2023. p. 11121-11128.
- ZHAO, X. *et al.* One-day-ahead probabilistic wind speed forecast based on optimized numerical weather prediction data. **Energy Conversion and Management**, v. 164, p. 560-569, 2018.
- ZHENG, A.; CASARI, A. **Feature Engineering for Machine Learning**. Sebastopol: O'Reilly Media, 2018.
- ZHOU, H. *et al.* Short-term photovoltaic power forecasting based on long short-term memory neural network and attention mechanism. **IEEE Access**, v. 7, p. 78063-78074, 2019.