Universidade Federal de Juiz de Fora
Instituto de Ciência Exatas
Bacharelado em Ciência da Computação

Integração de um modelo de tratamento de dados faltantes à técnica de *Deep Learning* para a estimação de cargas de radiação solar

Yuri Dias de Azevedo

JUIZ DE FORA
DEZEMBRO, 2023

Integração de um modelo de tratamento de dados faltantes à técnica de *Deep Learning* para a estimação de cargas de radiação solar

Yuri Dias de Azevedo

Universidade Federal de Juiz de Fora

Instituto de Ciência Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Luciana Conceição Dias Campos

Coorientador: Samuel da Costa Alves Basilio

JUIZ DE FORA

DEZEMBRO, 2023

INTEGRAÇÃO DE UM MODELO DE TRATAMENTO DE DADOS FALTANTES À TÉCNICA DE *DEEP LEARNING* PARA A ESTIMAÇÃO DE CARGAS DE RADIAÇÃO SOLAR

Yuri Dias de Azevedo

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIA EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Luciana Conceição Dias Campos
Doutora em Engenharia Elétrica

Samuel da Costa Alves Basilio
Mestre em Ciência da Computação

Heder Soares Bernardino
Doutor em Modelagem Computacional

Leonardo Goliatt da Fonseca
Doutor em Modelagem Computacional

JUIZ DE FORA

8 DE DEZEMBRO, 2023

Resumo

A eficiência na geração de energia solar e o planejamento energético dependem crucialmente da precisão na previsão da radiação solar. Este estudo, focado na Zona da Mata Mineira, aborda a complexidade e os desafios inerentes à previsão da radiação solar, um fenômeno marcado por sua variabilidade e influenciado por múltiplos fatores ambientais. O trabalho ressalta, em particular, a problemática dos dados faltantes e ruidosos nas séries temporais meteorológicas, que constituem um obstáculo significativo para a precisão das previsões.

Para superar esses desafios, propomos uma abordagem que integra técnicas de imputação de dados e modelos de *Deep Learning*, especificamente redes neurais do tipo *Long Short Term Memory* (LSTM). Esta metodologia não apenas aborda a questão da imputação de dados faltantes mas também refina o processo de previsão. Nossas análises revelaram que métodos como Random Forest e Dynamic Mode Decomposition (DMD) são eficazes em preservar padrões de dados, contribuindo positivamente para o treinamento dos modelos de previsões.

Os resultados obtidos demonstram que, ao tratar efetivamente os dados faltantes, os modelos baseados em LSTM proporcionam previsões mais precisas em comparação com modelos que não utilizam técnicas adequadas de imputação. Este avanço na previsão da radiação solar tem implicações diretas na promoção da energia solar como uma fonte sustentável, contribuindo para a estabilidade da rede elétrica e impulsionando a chamada 'Economia Verde'. Portanto, este estudo não apenas fornece insights para a modelagem preditiva no campo meteorológico, mas também desempenha um papel crucial no avanço da sustentabilidade energética.

Palavras-chave: Radiação Solar, Previsão, Dados Faltantes, Deep Learning, Energia Renovável.

Abstract

Efficiency in solar energy generation and energy planning crucially depends on the accuracy of solar radiation forecasting. This study, focused on the Zona da Mata Mineira region, addresses the complexity and inherent challenges in predicting solar radiation, a phenomenon marked by its variability and influenced by multiple environmental factors. The work particularly highlights the issue of missing and noisy data in meteorological time series, which pose a significant obstacle to forecast accuracy.

To overcome these challenges, we propose an approach that integrates data imputation techniques and *Deep Learning* models, specifically *Long Short Term Memory* (LSTM) neural networks. This methodology not only addresses the issue of imputing missing data but also refines the forecasting process. Our analyses revealed that methods such as *Random Forest* and *Dynamic Mode Decomposition* (DMD) are effective in preserving data patterns, contributing positively to the training of forecasting models.

The results obtained demonstrate that, by effectively treating missing data, LSTM-based models provide more accurate predictions compared to models that do not utilize appropriate imputation techniques. This advancement in solar radiation forecasting has direct implications in promoting solar energy as a sustainable source, contributing to the stability of the electrical grid, and boosting the so-called 'Green Economy'. Therefore, this study not only provides insights for predictive modeling in the meteorological field but also plays a crucial role in advancing energy sustainability.

Keywords: Solar Radiation, Forecasting, Missing Data, Deep Learning, Renewable Energy.

Agradecimentos

À minha família, o alicerce fundamental de todas as minhas realizações.

Aos meus amigos por estarem sempre ao meu lado nesta jornada.

Aos meus orientadores, cujo apoio constante e incentivo foram essenciais durante todo o desenvolvimento deste projeto.

Ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pela bolsa de estudos e auxílio financeiro, que possibilitou a dedicação integral ao desenvolvimento deste trabalho de conclusão de curso e a operacionalização do estudo.

Conteúdo

Lista de Figuras **8**

Lista de Tabelas **9**

Lista de Abreviações **10**

1 Introdução **11**
1.1 Motivação . . . 11
1.2 Objetivos . . . 13
1.3 Organização do Trabalho . . . 13

2 Fundamentação Téorica **14**
2.1 Séries Temporais . . . 14
2.2 Aprendizado de *Machine Learning* . . . 16
2.2.1 Aprendizado Supervisionado . . . 16
2.2.2 Aprendizado Não Supervisionado . . . 16
2.2.3 Aprendizado por Reforço . . . 16
2.3 *Dynamic Mode Decomposition* (DMD) . . . 17
2.3.1 O Algoritmo . . . 18
2.4 *Random Forest* . . . 19
2.5 *K-Nearest Neighbors* (KNN) . . . 20
2.6 Interpolação Linear . . . 21
2.7 Interpolação Sazonal . . . 22
2.8 Redes Neurais Artificial e *Deep Learning* . . . 23
2.8.1 Fundamentos de Redes Neurais Artificiais . . . 23
2.8.2 Funções de Ativação . . . 24
2.8.3 Aprendizado de Redes Neurais . . . 27
2.8.4 *Deep Learning* . . . 29
2.9 Métricas de Avaliação de Dados . . . 30
2.9.1 Erro Quadrático Médio (MSE) . . . 31
2.9.2 Raiz do Erro Quadrático Médio (RMSE) . . . 31
2.9.3 Erro Absoluto Médio (MAE) . . . 32
2.9.4 Erro Percentual Absoluto Médio (MAPE) . . . 32
2.9.5 Coeficiente de Determinação (R^2) . . . 33
2.10 Considerações Finais . . . 33

3 Trabalhos Relacionados **35**

4 Metodologia **38**
4.1 Coleta de dados . . . 38
4.2 Pré-processamento dos Dados . . . 42
4.2.1 Redimensionamento . . . 42
4.2.2 Tratamento de dados faltantes . . . 43
4.2.3 Conjunto de dados . . . 44
4.2.4 Normalização dos Dados de Entrada . . . 45

4.3 Janelamento 45
4.4 Construção do Modelo de *Deep Learning* 46
4.5 Avaliação dos Modelos 47
4.6 Metodologia do 1º caso de estudo 48
4.7 Metodologia do 2º caso de estudo 52
4.7.1 Otimização dos Hiperparâmetros da Rede Neural 53

5 Estudos de caso **54**
5.1 Estudo de Caso 1: Avaliação dos Modelos de Imputação de Dados Faltantes 54
5.1.1 Análise dos Modelos de Imputação para o cenário 1 54
5.1.2 Análise dos Modelos de Imputação para cenário 2 57
5.1.3 Conclusão 60
5.2 Estudo de Caso 2: Modelo LSTM para Previsão de Radiação Solar 61
5.2.1 Cenário de Barbacena - MG 61
5.2.2 Cenário de Viçosa - MG 63
5.2.3 Cenário de Muriaé - MG 65
5.2.4 Cenário de Juiz de Fora - MG 67
5.2.5 Cenário de São João del Rei - MG 69

6 Conclusão e Trabalhos Futuros **72**
6.1 Conclusão 72
6.2 Trabalhos Futuros 73

Bibliografia **75**

Lista de Figuras

2.1 Demostração de componentes de uma série temporal. (OLIVEIRA, 2019) . 15
2.2 Esquema de um neurônio artificial (CAMPOS, 2010). 24

4.1 Mapa da EMAs do estado de Minas Gerais com as cidades selecionadas em destaque. 39
4.2 Correlação das variáveis disponível com a radiação solar global. 40
4.3 Gráfico representado a distribuição de dados faltantes em uma porção da série temporal de radiação da cidade de São João Del Rei. Elaborada pelo Autor. 42
4.4 Gráfico representado a distribuição de dados faltantes em uma porção da série temporal de temperatura da cidade de Juiz de Fora. Elaborada pelo Autor. 42
4.5 Arquitetura da Rede Neural. Elaborada pelo Autor. 47
4.6 Subsérie sem dados faltantes do período de Ago/2010 à Mai/2013. 49
4.7 Subsérie com dados faltantes do período de Ago/2010 à Mai/2013. 50
4.8 Subsérie com dados faltantes do período de Ago/2010 à Mai/2013. 51

5.1 Previsão realizada como o modelo treinado com os dados completados pela técnina Random Forest. 56
5.2 Previsão realizada como o modelo treinado com os dados completados pela técnina KNN. 56
5.3 Previsão realizada como o modelo treinado com os dados completados pela técnina Random Forest. 59
5.4 Previsão realizada como o modelo treinado com os dados completados pela técnina DMD. 59
5.5 Previsão, com dados faltantes preenchidos pelo método Interpolação Linear, realizada para cidade de Barbacena no período de Janeiro de 2021 à Junho de 2021. 63
5.6 Previsão, com dados faltantes preenchidos pelo método *Random Forest*, realizada para cidade de Viçosa no período de Abril de 2021 à Agosto de 2021. 65
5.7 Previsão, com dados faltantes preenchidos pelo método DMD, realizada para cidade de Muriaé no período de Maio de 2021 à Outubro de 2021. . . 67
5.8 Previsão, com dados faltantes preenchidos pelo método *Random Forest*, realizada para cidade de Juiz de Fora no período de Junho de 2021 à Novembro de 2021. 69
5.9 Previsão, com dados faltantes preenchidos pelo método *Random Forest*, realizada para cidade de São João del Rei no período de Maio de 2021 à Outubro de 2021. 71

Lista de Tabelas

4.1 Período dos dados coletados da cidades selecionadas. 39
4.2 Variáveis disponíveis no banco de dados meteorológico do INMET 40
4.3 Percentual de dados faltantes por cidade. 41
4.4 Período de cada porção dos dados seguindo o modelo *"Houldout"*. 44
4.5 Tabela demostrando a maior subsérie de dados sem a presença de dados faltantes. 48
4.6 Período de cada porção dos dados seguindo o modelo *"Houldout"*. 51
4.7 Configuração do modelo de previsão para avalição das técnicas de imputação de dados. 52
4.8 Intervalo de hiperparâmetros usados para otimização do modelo de previsão desenvolvido. 53

5.1 Tabela apresenta os valores das métricas de erro do processo de inputação para lacunas pequenas. 54
5.2 Qualidade da previsão com os dados após o processo de inputação para lacunas pequenas. 55
5.3 Métricas de erro do processo de inputação para lacunas de dados extensas. 57
5.4 Qualidade da previsão com os dados após o processo de inputação para lacunas extensas. 58
5.5 Configuração do modelo de previsão de Barbacena - MG. 62
5.6 Resultado das previsões realizadas com os dados da cidade de Barbacena - MG. 62
5.7 Configuração do modelo de previsão de Viçosa - MG. 64
5.8 Resultado das previsões realizadas com os dados da cidade de Viçosa - MG. 64
5.9 Configuração do modelo de previsão de Muriaé - MG. 66
5.10 Resultado das previsões realizadas com os dados da cidade de Muriaé - MG. 66
5.11 Configuração do modelo de previsão de Juiz de Fora - MG. 68
5.12 Resultado das previsões realizadas com os dados da cidade de Juiz de Fora - MG. 68
5.13 Configuração do modelo de previsão de São João del Rei - MG. 70
5.14 Resultado das previsões realizadas com os dados da cidade de São João del Rei - MG. 70
5.15 Os melhores resultados das previsões realizadas para as cidades selecionadas. 71

Lista de Abreviações

DCC	Departamento de Ciência da Computução
UFJF	Universidade Federal de Juiz de Fora
RNA	Rede Neural Artificial
DMD	Dynamic Mode Decomposition
RS	Radiação Solar
EMA	Estação Meteorológica Automática
KNN	*K-Nearest Neighbors*
LSTM	*Long Short Term Memory*

1 Introdução

A crescente demanda por energia limpa, juntamente com a urgente necessidade de reduzir as emissões de gases de efeito estufa, tem catalisado a expansão e a adoção de fontes de energia renovável em âmbito global. Entre as diversas opções disponíveis, a energia solar emerge como uma das alternativas mais promissoras, graças à sua abundância e à sua capacidade de gerar energia limpa e sustentável (CUNHA et al., 2021).

No cenário brasileiro e, mais especificamente, no estado de Minas Gerais, o setor de energia solar tem testemunhado um notável crescimento nos últimos anos. Isso se deve, em grande parte, às condições geográficas favoráveis da região e às políticas governamentais que incentivam a adoção de fontes de energia limpa. Com isso, a matriz energética de Minas Gerais tem passado por um processo de diversificação progressiva, à medida que a energia solar amplia sua parcela de contribuição na produção total de energia (MENDES, 2022) (Agência Nacional de Energia Elétrica (ANEEL), 2023).

Nesse contexto de transição energética, a previsão da radiação solar assume um papel fundamental para o planejamento eficaz da geração de energia, assim como para as tomadas de decisões relacionadas à operação e manutenção de usinas solares. A capacidade de antecipar a radiação solar permite a otimização dos processos de produção, distribuição e armazenamento de energia, contribuindo para a estabilidade das redes elétricas e evitando a necessidade de recorrer a fontes de energia mais onerosas e poluentes para atender à demanda energética. Além disso, as previsões precisas de radiação solar desempenham um papel crucial na avaliação do potencial de investimentos em novos projetos solares, reduzindo, assim, os riscos e incertezas inerentes à implementação e operação de usinas solares (SILVA, 2021).

1.1 Motivação

No Brasil, existem iniciativas alinhadas aos Objetivos de Desenvolvimento Sustentável da Organização das Nações Unidas (ONU), visando combater desafios globais como a

proteção ambiental e climática, promoção de energia limpa e acessível, e o desenvolvimento de cidades e comunidades sustentáveis (Nações Unidas no Brasil, 2022). Esses objetivos, parte da Agenda 2030 da ONU, representam um plano global para alcançar um mundo melhor até 2030. O Relatório sobre Clima e Desenvolvimento para o Brasil, publicado pelo Banco Mundial, destaca a posição privilegiada do Brasil em termos de acesso a energias renováveis, incluindo a energia solar, que é uma das mais promissoras fontes renováveis, dada a abundante irradiação solar do país. Esta posição confere ao Brasil uma grande vantagem competitiva no crescente mercado global de bens e serviços mais verdes (Banco Mundial, 2023).

Porém a adoção em larga escala da energia solar ainda enfrenta barreiras, e uma delas é a habilidade de prever com precisão a radiação solar, que é dificultada por sua natureza complexa e variável. Esta é influenciada por diversos fatores, como: posição geográfica, clima, estação do ano, cobertura de nuvens e a presença de partículas atmosféricas (FAWAZ et al., 2019).

Adicionalmente, a constante ocorrência de dados faltantes nas séries temporais climáticas intensifica esse desafio. A ausência de dados pode criar lacunas significativas nas séries temporais, dificultando a análise e prejudicando a precisão das previsões (GARCÍA; LUENGO; HERRERA, 2014). Além disso, os dados ruidosos, que apresentam variações desproporcionais, podem introduzir distorções adicionais, tornando a tarefa de previsão ainda mais desafiadora. Uma estratégia eficaz para lidar com esses dados ruidosos é removê-los e tratá-los como se fossem dados faltantes. Essa abordagem simplifica o processo de tratamento, pois permite a aplicação das técnicas de imputação de dados para preencher essas lacunas.

Embora várias técnicas de *Deep Learning* tenham sido aplicadas com sucesso na previsão da radiação solar, ainda há espaço para melhorar a precisão e a eficiência dessas previsões, principalmente no contexto de tratamento dos dados. A proposta de combinar o modelo de tratamento de dados faltantes com as técnicas de *Deep Learning* visa superar as limitações dos métodos existentes na literatura, oferecendo uma abordagem mais eficiente para a previsão da radiação solar, podendo impulsionar a adoção e a expansão da energia solar, contribuindo para a sustentabilidade energética e a mitigação das mudanças

climáticas.

1.2 Objetivos

Este estudo tem como objetivo principal desenvolver um modelo para tratamento de dados ausentes e desenvolver um modelo usando técnicas de *deep learning* com a finalidade de estimar a carga de radiação solar para as cidades da Zona da Mata no estado de Minas Gerais.

Em um aspecto secundário, este trabalho almeja realizar uma comparação criteriosa dos métodos de tratamento de dados ausentes. O objetivo é identificar as técnicas mais eficazes, visando otimizar o desempenho do modelo preditivo. Adicionalmente, o estudo busca determinar a configuração mais adequada do modelo de previsão, fundamentado na arquitetura de Redes Neurais Long Short-Term Memory (LSTM), adaptada especificamente para cada cidade incluída no estudo.

1.3 Organização do Trabalho

O presente trabalho está estruturada da seguinte maneira: o Capítulo 2 dedica-se a estabelecer a fundamentação teórica, fornecendo uma base conceitual para os temas explorados no trabalho. Seguindo, o Capítulo 3 realiza uma revisão crítica da literatura, examinando estudos relevantes e relacionados ao assunto central desta pesquisa. No Capítulo 4, é detalhada a metodologia empregada na componente prática do estudo, delineando as técnicas e os procedimentos adotados. O Capítulo 5 apresenta dois estudos de caso: o primeiro investiga métodos de imputação de dados faltantes e o segundo se concentra na previsão da radiação solar em cidades específicas da Zona da Mata Mineira, englobando cenários de teste e a análise dos resultados alcançados. Finalmente, o Capítulo 6 conclui a monografia, sintetizando os achados principais e propondo direções para futuras pesquisas na área.

2 Fundamentação Téorica

Este capítulo apresenta uma visão geral dos principais conceitos e técnicas relacionadas ao estudo e análise de séries temporais, bem como as abordagens de inteligência artificial que podem ser aplicadas para extrair informações úteis desses dados. A análise de séries temporais é uma área de pesquisa crucial, pois muitos problemas do mundo real envolvem a compreensão e previsão de dados sequenciais.

2.1 Séries Temporais

Segundo Wei (2013), uma série temporal é a realização de um processo estocástico, que é uma família de variáveis aleatórias indexadas no tempo. Em outras palavras, séries temporais são conjuntos de observações realizadas sequencialmente no decorrer do tempo, podendo ser contínuas ou discretas (ADHIKARI; AGRAWAL, 2013). Uma série temporal contínua representa as observações que ocorrem de maneira ininterrupta, como exemplo, a medição de radiação solar ao longo do tempo. Enquanto uma série discreta são as observações que ocorrem em intervalos de tempo específicos, como exemplo, a população de uma cidade medida em um período específico. As séries temporais podem ser univariadas, quando apenas uma variável é medida, ou multivariadas, quando mais de uma variável é observada ao longo do tempo.

Ao analisar séries temporais, é importante considerar os seguintes componentes principais (ADHIKARI; AGRAWAL, 2013), que estão ilustrados na Figura 2.1:

- **Tendência**: representa a direção geral de crescimento ou decrescimento da série ao longo do tempo. A tendência pode ser ascendente, descendente ou estacionária, dependendo do comportamento da variável analisada;
- **Sazonalidade**: refere-se a padrões de comportamento que se repetem periodicamente em intervalos regulares de tempo, geralmente em um período menor que um ano. Podemos identificar sazonalidade em séries temporais relacionadas ao turismo,

onde a demanda por hospedagem e atividades turísticas aumenta durante períodos de férias e feriados prolongados, retornando aos níveis normais após esses períodos;

- **Ciclos**: são variações regulares que ocorrem em períodos, normalmente maiores que um ano, e incluem tendências e variações sazonais. É possível identificar ciclos em séries econômicas, exemplificados por fases de prosperidade, declínio, recessão e retomada do crescimento.

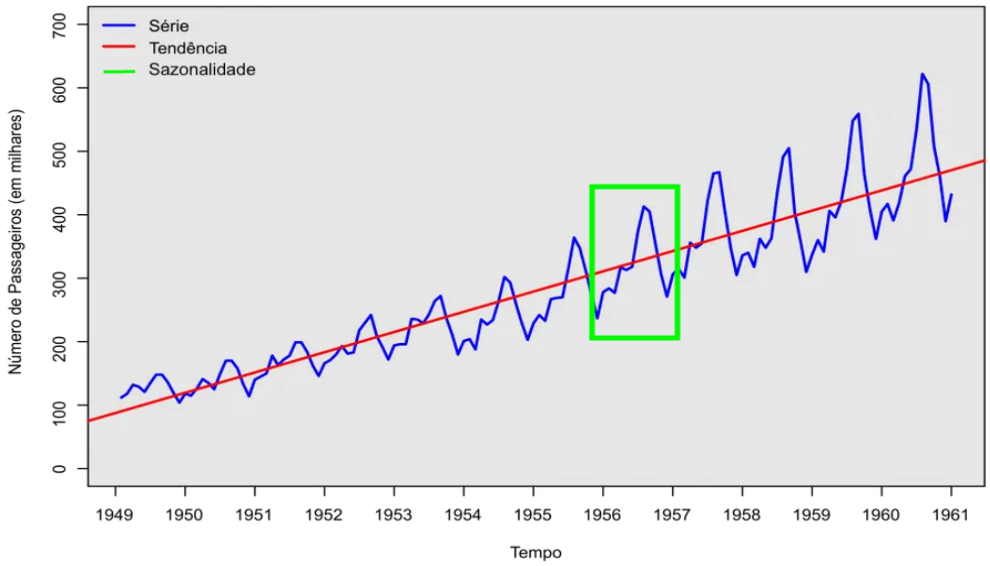

Figura 2.1: Demostração de componentes de uma série temporal. (OLIVEIRA, 2019)

Além desses componentes, é importante levar em consideração que séries temporais podem ser compostas por elementos determinísticos e não determinísticos. Os componentes não determinísticos são aqueles que não podem ser previstos com certeza, enquanto os determinísticos seguem um padrão previsível (FILDES; MAKRIDAKIS, 1995).

A estacionariedade é outra característica relevante no estudo de séries temporais. Uma série temporal é considerada estacionária quando suas propriedades estatísticas, como média e variância, permanecem constantes ao longo do tempo.

Portanto, ao estudar séries temporais, é fundamental analisar a tendência, sazonalidade, ciclos e estacionariedade, bem como considerar os componentes determinísticos e não determinísticos presentes na série. Essa análise possibilita o desenvolvimento de modelos de previsões mais acurados (FILDES; MAKRIDAKIS, 1995).

2.2 Aprendizado de *Machine Learning*

A aprendizagem de máquina constitui um pilar essencial no desenvolvimento de sistemas computacionais capazes de aprimorar seu desempenho de forma autônoma, baseando-se na experiência acumulada. Este domínio da inteligência artificial é categorizado, primordialmente, em três modalidades distintas de aprendizado: supervisionado, não supervisionado e por reforço. Cada uma dessas categorias possui características e aplicações específicas, moldando a maneira como os algoritmos aprendem e evoluem a partir dos dados disponíveis goodfellow2016deep.

2.2.1 Aprendizado Supervisionado

No aprendizado supervisionado, a modelo é treinada com base em um conjunto de dados rotulados, onde cada exemplo de treinamento é associado a uma resposta ou rótulo (HAYKIN, 2009). O objetivo é aprender uma função que mapeie as entradas para as saídas corretas, de modo que o modelo possa realizar previsões precisas em dados não vistos bishop2006pattern.

2.2.2 Aprendizado Não Supervisionado

No aprendizado não supervisionado, o modelo é treinado com base em um conjunto de dados não rotulados, onde o objetivo é aprender a estrutura subjacente dos dados sem utilizar informações de saída específicas (GOODFELLOW; BENGIO; COURVILLE, 2016). O aprendizado não supervisionado é útil para problemas como agrupamento, redução de dimensionalidade e detecção de anomalias, onde os rótulos de saída podem não estar disponíveis ou são difíceis de obter (HINTON; SALAKHUTDINOV, 2006).

2.2.3 Aprendizado por Reforço

No aprendizado por reforço, o modelo é treinado para tomar decisões em um ambiente dinâmico, onde o objetivo é aprender uma política que maximize a recompensa acumulada ao longo do tempo (SUTTON; BARTO, 2018). O aprendizado por reforço difere do aprendizado supervisionado e não supervisionado, pois o modelo recebe apenas feedback

esporádico e atrasado na forma de recompensas, em vez de rótulos ou reconstruções.

2.3 *Dynamic Mode Decomposition* (DMD)

DMD, que em português é decomposição de modos dinâmicos, é uma metodologia de decomposição e redução de dimensionalidade aplicada a conjuntos de dados sequenciais (Schmid., 2022). Em suma, ela processa medições de alta dimensão, identifica padrões e isola dinâmicas comportamentais. A DMD foi concebida através da análise de **Koopman** e, desde então, tem sido empregada em análises de dados temporais, desde sistemas de fluidos básicos até complexos, e também impactou áreas além da dinâmica dos fluidos.

A análise de previsão do comportamento de sistemas dinâmicos são tarefas fundamentais em diversas áreas da ciência. Os métodos comumente usados, como a análise de *Fourier* e a decomposição em valores singulares (SVD) (TU et al., 2014), têm sido aplicados para entender a dinâmica de tais sistemas. No entanto, esses métodos ao serem aplicados em análise de sistemas não-lineares e não-estacionários demostram suas limitações. Neste contexto, a DMD surge como uma ferramenta para analisar e modelar sistemas não-lineares, sendo motivada pela necessidade de superar as limitações desses métodos anteriormente citados (SCHMID, 2010).

Embora a DMD tenha sido desenvolvida inicialmente para a análise de comportamento de fluidos, logo se mostrou útil em outras áreas de aplicação, como aeroespacial, biológica, financeira e de controle (TAIRA et al., 2017). A motivação para a criação da DMD está intimamente ligada à teoria do operador de *Koopman*, que oferece uma abordagem linear para a análise de sistemas não-lineares. A teoria de *Koopman* foi proposta por Bernard Koopman em 1931 e reformulada por Igor Mezić em 2005, que mostrou que a análise do operador de *Koopman* é capaz de fornecer informações valiosas sobre a dinâmica de sistemas não-lineares (MEZIĆ, 2013).

A DMD é uma técnica baseada em dados que fornece uma aproximação linear, o qual atua sobre funções observáveis do estado do sistema (ROWLEY et al., 2009). Essa aproximação é obtida a partir da decomposição das matrizes de dados obtidos (TU et al., 2014). Ao aplicar a DMD, é possível extrair modos espaciais e temporais que descrevem a dinâmica do sistema em estudo, fornecendo assim uma representação compacta e coerente

da evolução do sistema ao longo do tempo. (SCHMID, 2010).

O procedimento básico da DMD consiste em decompor os dados em uma base de modos dinâmicos, onde cada modo possui uma estrutura espacial e uma frequência temporal associada (TU et al., 2014). Esses modos podem ser interpretados como blocos de construção que, quando combinados, reconstroem a dinâmica do sistema. (TAIRA et al., 2017).

2.3.1 O Algoritmo

As entradas do DMD consistem em um conjunto de pares de instantâneos nos quais cada par de instantâneos é separado por um intervalo de tempo fixo. Geralmente, esses dados provêm de uma série temporal (TAIRA et al., 2017).

As saídas incluem autovalores e modos DMD. Os modos representam estruturas espaciais que oscilam e/ou aumentam/diminuem em taxas determinadas pelos autovalores associados. Esses valores resultam da autodecomposição de um operador linear de ajuste ideal que aproxima a dinâmica observada nos dados (TAIRA et al., 2017).

Diante disso, temos o input X_1 e X_2 apresentado na equação 2.1:

$$X_1 = \begin{bmatrix} x_1, & x_2, & x_3, & \dots, & x_{n-1} \end{bmatrix} \quad e \quad X_2 = \begin{bmatrix} x_2, & x_3, & x_4, & \dots, & x_n \end{bmatrix} \tag{2.1}$$

Na DMD, a relação entre a matriz X_1 e a matriz X_2 pode ser definida de maneira linear, como pode ser observado na equação 2.2:

$$X_1 = AX_2 \tag{2.2}$$

A matriz A pode ser definida por $A = X_2X_1^{\dagger}$, onde $X_1^{\dagger}$ é a pseudo-inversa de X_1. Os autovalores e modos DMD são então definidos como os autovalores e autovetores de A. Para calcular a matriz de *Koopman*, precisamos definir a decomposição de valores singulares da matrix X_1, conforme a equação 2.3.

$$X_1 = U\Sigma V^T \tag{2.3}$$

onde U consiste em vetores singulares a esquerda, V consiste em vetores singulares a direita e Σ consiste em valores singulares em sua diagonal. Implementa-se um valor de decomposição singular truncado com um valor de *rank* r predefinido, para calcular a matriz *Koopman*, apresentada na equação 2.4:

$$\tilde{A} = U_r^T A U_r = U_r^T X_2 V_r \Sigma_r^{-1} \in \mathbb{R}^{rXr} \tag{2.4}$$

Após, encontra-se os autovalores μ_j e autovetores $\tilde{v}_j$ de $\tilde{A}$, onde $\tilde{A}\tilde{v}_j = \mu_j \tilde{v}_j$. Todo μ_i diferente de zero é um autovalor da DMD, com o modo DMD dado pela equação 2.5:

$$v_i = \mu_i^{-1} X_2 V_r \Sigma_r^{-1} \tilde{v}_i \tag{2.5}$$

O modo DMD projetado é dado por $Pv_i = U_r \tilde{v}_i$, onde $P = U_r U_r^T$ é a projeção ortogonal no primeiro r. Note, podemos inferir as taxas de crescimento/decaimento e as frêquencias dos modos DMD a partir da análise das componentes apresentado na equação 2.6:

$$\lambda_j = \frac{1}{\Delta t} \log(\mu_j) \qquad (\Delta t : \text{variação temporal}) \tag{2.6}$$

Em resumo, a motivação para a utilização da DMD é a sua capacidade de fornecer uma representação compacta e coesa do comportamento dos sistemas complexos e não-lineares. A teoria do operador de *Koopman* desempenha um papel fundamental na base teórica da DMD, permitindo a aproximação e análise de um sistema não-linear(MEZIĆ, 2013). Há aplicação bem-sucedida da DMD em diversas áreas, como já explicitado, o que evidencia a versatilidade desta técnica para melhorar nossa compreensão e previsão desses sistemas. (TAIRA et al., 2017).

2.4 *Random Forest*

Random Forest é uma técnica de aprendizado de máquina baseada em árvores de decisão, que tem se mostrado eficaz na análise de séries temporais devido à sua capacidade de modelar complexas relações não-lineares e interações entre variáveis (BREIMAN, 2001a)(ATHEY; TIBSHIRANI; WAGER, 2019). Esta técnica envolve a construção de

múltiplas árvores de decisão durante o treinamento, com cada árvore sendo gerada a partir de uma amostra aleatória do conjunto de dados. A decisão final é feita pela média das previsões de todas as árvores, o que reduz o risco de sobreajuste e aumenta a robustez do modelo (LIAW; WIENER, 2002).

No contexto de séries temporais, *Random Forest* pode ser aplicado para previsão e classificação. A abordagem tradicional para séries temporais envolve a utilização de características temporais, como tendências e sazonalidades, como entradas para o modelo (CHATFIELD, 2000). No entanto, *Random Forest* permite a incorporação de uma variedade mais ampla de características, incluindo aquelas derivadas de métodos estatísticos e de aprendizado de máquina, para capturar padrões complexos nos dados (TYRALIS; PAPACHARALAMPOUS, 2017).

Além das aplicações mencionadas anteriormente, a *Random Forest* também se destaca na imputação de dados faltantes em séries temporais (STEKHOVEN; BüHLMANN, 2011). A sua capacidade de lidar com grandes conjuntos de dados e acomodar múltiplas variáveis torna-a uma ferramenta eficaz para preencher lacunas nos dados. Este processo envolve a utilização de padrões observados nos dados existentes para estimar valores faltantes, mantendo a integridade e a estrutura temporal dos dados (TANG; ISHWARAN, 2017).

Isso é particularmente útil em aplicações práticas onde séries temporais são caracterizadas por alta dimensionalidade e complexidade (FISCHER; KRAUSS; TREICHEL, 2018). Além disso, a natureza não-paramétrica do *Random Forest* o torna menos suscetível a suposições sobre a distribuição dos dados, o que é uma limitação comum em métodos tradicionais de séries temporais (BREIMAN, 2001b).

2.5 *K-Nearest Neighbors* (KNN)

O K-Nearest Neighbors (KNN) é um método de aprendizado de máquina, que tem sido amplamente utilizado em várias aplicações, incluindo análise de séries temporais (ALTMAN, 1992). O KNN baseia-se no princípio de que instâncias semelhantes dentro de um conjunto de dados tendem a ter saídas ou classificações semelhantes. Em séries temporais, o KNN pode ser empregado para prever valores futuros com base na similaridade dos padrões

temporais (RATANAMAHATANA; KEOGH,). A eficácia do KNN em séries temporais decorre de sua capacidade de capturar a natureza dinâmica dos dados, adaptando-se a mudanças e tendências ao longo do tempo.

Um dos principais desafios no uso do KNN para séries temporais é a definição de uma medida de distância apropriada. A distância euclidiana padrão pode não ser adequada para séries temporais devido à sua sensibilidade a deslocamentos e distorções no eixo do tempo (BERNDT; CLIFFORD, 1994).

Além de previsão e classificação, o KNN também é utilizado para a imputação de dados faltantes em séries temporais (TROYANSKAYA et al., 2001). A imputação baseada no KNN envolve a identificação dos k-vizinhos mais próximos de um ponto de dados faltante e a utilização de seus valores para estimar o valor faltante. Esta abordagem é particularmente útil em séries temporais, onde a correlação temporal e a continuidade dos dados são cruciais. O KNN pode efetivamente utilizar a informação contida nos padrões temporais próximos para preencher lacunas nos dados, preservando assim a estrutura e as tendências inerentes à série temporal (LIEW; LAW; YAN, 2010). A escolha de k (o número de vizinhos) e a medida de distância são cruciais para a eficácia da imputação, exigindo uma cuidadosa calibração com base nas características específicas da série temporal em questão.

2.6 Interpolação Linear

A interpolação linear é um técnica empregada para calcular valores desconhecidos dentro de um intervalo, utilizando-se de pontos de dados já conhecidos (BURDEN; FAIRES, 2010). Essa técnica pressupõe que a variação entre dois pontos consecutivos é linear, facilitando assim a determinação de valores intermediários. Devido à sua simplicidade e eficácia, a interpolação linear é amplamente adotada em várias áreas, como no processamento de imagens e na análise de dados.

Em séries temporais, essa técnica é particularmente útil para preencher espaços em dados incompletos ou para refinar dados com baixa frequência de amostragem (CHATFIELD, 2003). A interpolação linear conecta pontos de dados por meio de segmentos lineares, mantendo as tendências gerais e variações locais nos dados. Essa estratégia é eficaz

quando os dados mostram uma variação linear ou quase linear entre amostras. Contudo, é necessário cautela ao utilizá-la em dados com variações não lineares significativas, pois pode resultar em estimativas inexatas e na perda de detalhes cruciais dos dados.

Embora seja uma ferramenta útil, a interpolação linear tem suas desvantagens. Sua principal limitação é a inabilidade de representar adequadamente curvas complexas ou padrões não lineares nos dados (GAUTSCHI, 2011). Adicionalmente, pode gerar artefatos indesejados, como oscilações e descontinuidades, particularmente em conjuntos de dados muito variáveis. Para contornar esses desafios, métodos de interpolação mais sofisticados, como interpolação polinomial e splines cúbicas, são muitas vezes preferidos. Esses métodos oferecem uma modelagem mais precisa para dados complexos, mas requerem maior esforço computacional e são mais complexos de implementar.

2.7 Interpolação Sazonal

A interpolação sazonal é uma técnica estatística utilizada para analisar e preencher lacunas em séries temporais que exibem padrões sazonais (CLEVELAND et al., 1990). Esta abordagem é particularmente relevante em contextos onde os dados são influenciados por fatores sazonais, como economia, meteorologia e agricultura. A interpolação sazonal não se limita a simplesmente preencher valores faltantes, mas também leva em consideração as variações periódicas inerentes aos dados. Isso é feito através da decomposição da série temporal em componentes, geralmente incluindo tendência, sazonalidade e resíduos, e então aplicando técnicas de interpolação que respeitam esses padrões sazonais.

Um dos métodos mais comuns para interpolação sazonal é o uso de modelos de suavização, como o suavizador sazonal de Loess (STL) (CLEVELAND et al., 1990). O STL é uma técnica robusta que permite a decomposição de uma série temporal em componentes, ajustando-se flexivelmente a diferentes tipos de padrões sazonais e tendências. A interpolação sazonal através do STL envolve primeiro a decomposição da série temporal e, em seguida, a aplicação de técnicas de interpolação aos componentes sazonais e de tendência separadamente. Isso permite que a interpolação preserve as características únicas dos padrões sazonais dos dados, resultando em uma imputação mais precisa e significativa.

2.8 Redes Neurais Artificial e *Deep Learning*

As Redes Neurais Artificiais (RNAs) são modelos computacionais baseados na estrutura neurológica dos seres humanos. Essas estruturas foram desenvolvidas com o intuito de possibilitar com que as máquinas pudessem operar de forma semelhante a um cérebro humano e tivessem a capacidade de aprender com dados. Com o avanço das técnicas de aprendizado e o aumento da capacidade computacional, as RNAs evoluíram para arquiteturas mais complexas, conhecidas como *Deep Learning* (LECUN et al., 1998).

2.8.1 Fundamentos de Redes Neurais Artificiais

Inspiradas na composição neurológica do cérebro humano, as redes neurais são formadas por neurônios artificiais, estruturas essas que são similares em funcionamento aos neurônios biológicos. O modelo de comunicação entre os neurônios é inspirado no modelo neurológico, onde os neurônios são organizados em camadas e interconectados de forma que a saída de um neurônio serve como entrada de outros neurônios. Essas conexões entre os neurônios são chamadas de pesos sinápticos, que simulam as sinapses que ocorrem no cérebro, e servem para determinar o impacto que o valor de cada entrada do neurônio tem no valor de saída do mesmo, o que caracteriza o conhecimento aprendido pela rede neural (MCCULLOCH; PITTS, 1943).

A Figura 2.2 apresenta o modelo do neurônio artificial e sua equação matemática é aprensentada na equação 2.7. Observe que o neurônio recebe um vetor de entrada X e produz um valor de saída y. Internamente, as entradas são ponderadas pelos pesos sinápticos, representado por ω_{in} (onde o índice in representa a entrada n do neurônio i).

Posteriormente, as entradas ponderadas passam por um processo de combinações lineares, representado pelo núcleo do neurônio, simbolizado por Σ. De acordo com Haykin (1994), para modificar a influência do valor da combinação linear, utiliza-se o valor do bias, representado por θ_i. O resultado dessa combinação linear gera um valor que na figura é representado por net_i, que é introduzido em uma função de ativação.

A função de ativação, representado por φ, é uma função matemática que determina a amplitude da saída do neurônio representado por y_i. Fazendo um paralelo com o neurônio biológico, a função de ativação faz o papel de definir o limiar de excitação

minímo para a transmissão do pulso, a diferença é que no neurônio artificial esse processo serve para modular o valor de saída do neurônio. A função de ativação é fundamental para introduzir a não-linearidade ao modelo, permitindo que a rede neural aprenda e generalize padrões complexos e não-lineares presentes nos dados.

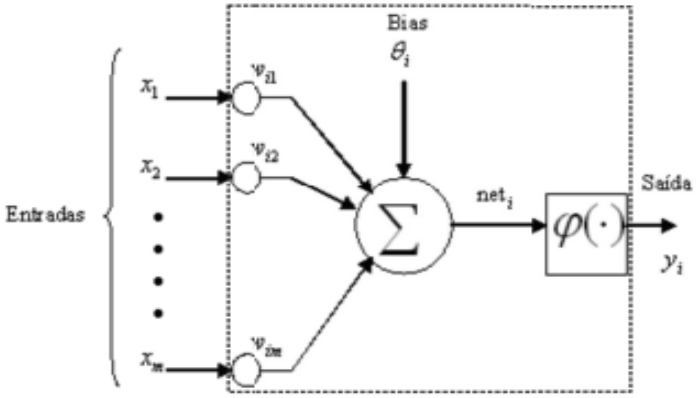

Figura 2.2: Esquema de um neurônio artificial (CAMPOS, 2010).

$$y_i = \varphi(net_i) = \varphi\left(\sum_{j=1}^{n} w_{ij}x_j + \theta_i\right), \text{ onde } i \text{ refere-se ao neurônio.} \tag{2.7}$$

Sem uma função de ativação, o neurônio seria apenas um somador linear, limitando a capacidade da rede neural de resolver problemas complexos e não-lineares (LECUN; BENGIO; HINTON, 2015).

Os pesos sinápticos são ajustado durante o processo de treinamento da rede neural, de modo a alcançar os objetivos preestabelecidos do treinamento.

2.8.2 Funções de Ativação

Conforme apresentado, as funções de ativação tem papel essencial nas RNAs porque ela é capaz de fornecer a não-linearidade ao modelo. Existem várias funções de ativação comumente usadas em redes neurais na literatura. Algumas das mais notáveis incluem: a função sigmoide, a função tangente hiperbólica (tanh), a função de ativação linear retificada (ReLU) e a função de ativação linear retificada parametrizada (PReLU).

Função Sigmoide

A função sigmoide é uma função de ativação que mapeia a entrada para um valor entre 0 e 1 (BISHOP, 2006), cuja fórmula é apresentada na equação 2.8.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{2.8}$$

A função sigmoide tem a vantagem de ser contínua e diferenciável, o que facilita o treinamento de redes neurais usando algoritmos de otimização baseados em gradientes, como a descida do gradiente estocástico (SGD) (BOTTOU, 2010).

Função Tangente Hiperbólica

A função tangente hiperbólica (tanh) é uma função de ativação que mapeia a entrada para um valor entre -1 e 1 (LECUN et al., 2012), cuja fórmula é apresentada na equação 2.9.

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{2.9}$$

A função tanh também é contínua e diferenciável, facilitando o treinamento das redes neurais com algoritmos baseados em SGD.

Função de Ativação Linear Retificada (ReLU)

A função de ativação linear retificada (ReLU) é uma função simples e não-linear que se tornou extremamente popular devido à sua eficácia no treinamento de redes neurais profundas (KRIZHEVSKY; SUTSKEVER; HINTON, 2017). A ReLU mapeia a entrada para o valor máximo entre a entrada e zero, cuja fórmula é apresentada na equação 2.10.

$$\text{ReLU}(x) = \max(0, x) \tag{2.10}$$

A ReLU possui algumas desvantagens, como a não diferenciabilidade em zero e o fato de que neurônios com saídas ReLU negativas não são ativados durante o treinamento, o que pode levar a "neurônios mortos"(HE et al., 2015).

Para mitigar essas desvantagens, a função de ativação Leaky ReLU foi proposta.

Diferente da ReLU tradicional, a Leaky ReLU permite um pequeno gradiente quando a unidade não está ativa. cuja fórmula é apresentada na equação 2.11.

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{se } x > 0 \\ \alpha x, & \text{caso contrário} \end{cases} \tag{2.11}$$

onde α é um pequeno coeficiente que proporciona a "fuga"(*leak*) para valores negativos, mantendo assim uma pequena ativação mesmo quando a unidade não está ativa. Esta característica ajuda a evitar o problema dos neurônios mortos e melhora a capacidade de generalização da rede (MAAS; HANNUN; NG, 2013).

Função de Ativação Linear Retificada Parametrizada (PReLU)

A função de ativação linear retificada parametrizada (PReLU) é uma variação da ReLU que permite uma pequena inclinação negativa para entradas negativas, o que pode ajudar a evitar o problema dos neurônios mortos (HE et al., 2015). A fórmula é apresentada na equação 2.12.

$$\text{PReLU}(x) = \max(\alpha x, x) \tag{2.12}$$

onde α é um parâmetro aprendido durante o treinamento. Inicialmente, α é atribuído um pequeno valor positivo, comumente 0.01, permitindo que a função PReLU opere de maneira semelhante à Leaky ReLU para entradas negativas (HE et al., 2015). Durante o treinamento, o ajuste de α é realizado por meio do algoritmo de backpropagation, onde o gradiente da perda em relação a α é calculado e utilizado para atualizar o valor de α utilizando métodos de otimização como o gradiente descendente (GOODFELLOW; BENGIO; COURVILLE, 2016). Tal mecanismo permite que a PReLU se ajuste dinamicamente às características específicas do conjunto de dados, potencialmente conduzindo a uma performance aprimorada do modelo em comparação com funções de ativação não adaptativas (HE et al., 2015; GOODFELLOW; BENGIO; COURVILLE, 2016).

2.8.3 Aprendizado de Redes Neurais

O aprendizado em redes neurais artificiais (RNAs) é um processo fundamental que permite que esses modelos aprendam a partir de dados e generalizem padrões complexos, possibilitando a resolução de uma ampla variedade de problemas, como a previsão de séries temporais.

O aprendizado em redes neurais envolve a atualização dos pesos sinápticos e bias dos neurônios com base nos dados de treinamento, de modo a otimizar o modelo de acordo com um objetivo pré-estabelecido (BISHOP, 2006). Os tipos de aprendizado em redes neurais seguem o padrão detalhado na seção 2.2.

Divisão de Dados por Separação *"Holdout"*

O modelo de divisão de dados *Holdout* é uma das formas mais simples de avaliar o desempenho de modelos de machine learning. Em modelos de redes neurais, este método envolve a divisão do conjunto de dados em três subconjuntos distintos: um para treinamento, um para validação e outro para teste (KOHAVI, 1995).

- Conjunto de treino: Esse conjunto de dados é utilizado para ajustar os pesos da rede neural durante a fase de treinamento. Geralmente, é uma porcentagem maior do conjunto de dados total, destinada a ensinar o modelo sobre os padrões e relações nos dados;

- Conjunto de validação: É outra parte do conjunto de dados original, separada do conjunto de treinamento. Esse conjunto é utilizado para monitorar o desempenho do modelo durante a fase de treinamento, ajudando a evitar o *overfitting* e a identificar o ponto de parada de treinamento ideal;

- Conjunto de teste: É a última parte do conjunto de dados original, separada dos conjuntos de treinamento e validação. Esse conjunto de dados não é usado durante o treinamento do modelo e foi reservado para avaliar a qualidade do modelo após o treinamento, fornecendo uma avaliação do desempenho de generalização do modelo de predição.

Overfitting em Aprendizado de Máquina

Overfitting é um fenômeno comum em aprendizado de máquina, onde um modelo treinado se ajusta excessivamente aos dados de treinamento, perdendo a capacidade de generalizar para novos dados. Este problema ocorre quando o modelo aprende padrões específicos e ruídos presentes no conjunto de treinamento, em vez de capturar tendências generalizáveis que seriam aplicáveis a dados não vistos (HAWKINS, 2004).

Um modelo que sofre overfitting tende a ter um desempenho excepcionalmente bom nos dados de treinamento, mas falha ao prever novos dados de forma precisa. Isso é particularmente problemático em aplicações práticas, onde o objetivo é desenvolver modelos que funcionem bem em situações reais, e não apenas no conjunto de dados usado para treinamento (BISHOP, 2006).

Normalização dos Dados

A preparação dos dados de entrada é uma etapa importante para realizar o treinamento dos modelos preditivos. Este processo envolve a normalização dos dados, uma técnica que busca ajustar a escala dos dados a um intervalo definido. A normalização garante que nenhuma variável de entrada domine sobre as outras, permitindo assim que o modelo de previsão aprenda de forma mais eficiente e produza resultados mais precisos (GOODFELLOW; BENGIO; COURVILLE, 2016).

Geralmente essa transformação aplica uma escala que transpõe cada recurso individualmente para o intervalo definido. A fórmula utilizada para a normalização é apresentada na equação 2.13:

$$x_{norm} = \frac{(x - x_{min})(l_{max} - l_{min})}{x_{max} - x_{min}} + l_{min} \tag{2.13}$$

Os parâmetros nesta equação são:

x_{norm} é o valor normalizado.

x é o valor original a ser normalizado.

x_{max} e x_{min} são, respectivamente, os valores máximo e mínimo da série de dados.

l_{max} e l_{min} são os limites máximo e mínimo do intervalo de normalização.

Como já foi dito, a normalização é essencial para evitar a prevalência de qualquer variável sobre as outras durante o ajuste dos pesos. Essa precaução é necessária para garantir a robustez do modelo de redes neurais e que o modelo aprenda e detecte padrões mais complexos nos dados.

2.8.4 *Deep Learning*

Deep learning é uma subárea do aprendizado de máquina que explora o uso de redes neurais artificiais profundas para aprender representações hierárquicas de dados (LECUN; BENGIO; HINTON, 2015). Essa abordagem permite que os modelos de *deep learning* capturem abstrações de alto nível a partir de dados brutos, como imagens e texto, melhorando a capacidade de generalização e desempenho em comparação com as redes neurais rasas.

Arquiteturas de *Deep Learning*

As arquiteturas de *deep learning* são caracterizadas pela presença de várias camadas ocultas, que permitem a aprendizagem de características em diferentes níveis de abstração (GOODFELLOW; BENGIO; COURVILLE, 2016). Algumas das arquiteturas populares na literatura incluem:

1. **Redes Neurais Convolucionais (CNNs)**: Essas redes são projetadas para processar dados com estrutura de grade, como imagens, utilizando convoluções em vez de conexões densas entre camadas. Isso permite a identificação de características locais e reduz o número de parâmetros a serem aprendidos, melhorando a eficiência e o desempenho (LECUN et al., 1998);

2. **Transformers**: Os Transformers, propostos por Vaswani et al. (2017), são arquiteturas baseadas em mecanismos de atenção que dispensam a recorrência e a convolução. Eles têm sido amplamente utilizados em tarefas de processamento de linguagem natural, como tradução automática e geração de texto;

3. **Redes Neurais Recorrentes (RNNs)**: Essas redes neurais são comumente usadas para o processamento de sequências de dados, sendo amplamente aplicadas em tarefas como reconhecimento de fala, tradução automática e análise de séries temporais (GOODFELLOW; BENGIO; COURVILLE, 2016; GRAVES, 2012). Diferentemente das redes neurais feedforward, as RNNs possuem conexões cíclicas entre seus neurônios, permitindo a manutenção de um estado interno ou memória que reflete a influência de entradas anteriores na sequência (ELMAN, 1990).

   A arquitetura de uma RNN é projetada para considerar não apenas a entrada atual, mas também o estado gerado no passo anterior (ELMAN, 1990) que atua como uma memória de curto prazo da rede, capturando informações relevantes de todos os passos anteriores, permitindo que a RNN capture dependências temporais nos dados.

   No entanto, as RNNs tradicionais enfrentam desafios significativos, como o *Vanish Gradient Problem*, que dificulta a aprendizagem de dependências de longo prazo (BENGIO; SIMARD; FRASCONI, 1994). Para mitigar esses problemas, foram desenvolvidas variantes como as *Long Short Term Memory* (LSTM) e as *Gated Recurrent Unit* (GRU). As LSTMs, introduzidas por Hochreiter e Schmidhuber (HOCHREITER; SCHMIDHUBER, 1997), incorporam células de memória e portões de controle (entrada, saída e esquecimento) para regular o fluxo de informações, permitindo que a rede aprenda dependências de longo prazo mais efetivamente. As GRUs, propostas por Cho et al. (2014), oferecem uma arquitetura mais simplificada que as LSTMs, mas ainda assim eficaz para capturar dependências temporais.

2.9 Métricas de Avaliação de Dados

A avaliação de modelos de previsão é uma etapa essencial em qualquer estudo de séries temporais, pois fornece uma quantificação objetiva do desempenho do modelo. Diversas métricas são utilizadas na literatura para avaliar a precisão e a eficácia dos modelos de previsão. Nesta seção, discutiremos as métricas que fora utilizadas nesse trabalho: MSE, RMSE, MAE, R^2 e MAPE.

2.9.1 Erro Quadrático Médio (MSE)

O Erro Quadrático Médio (MSE) é uma métrica frequentemente utilizada para avaliar a precisão de modelos de previsão. Ele mede a média dos quadrados das diferenças entre os valores previstos e os valores reais (HYNDMAN; KOEHLER, 2006a). A equação 2.14 apresenta o cálculo do MSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2.14}$$

onde n é o número de elementos, Y_i é o valor real e $\hat{Y}_i$ é o valor previsto.

O MSE é particularmente útil em cenários onde é importante penalizar erros maiores de forma mais severa, uma vez que o quadrado das diferenças tende a amplificar desvios maiores (HYNDMAN; KOEHLER, 2006a). Esta métrica é ideal para comparar e ajustar modelos de regressão, especialmente em conjuntos de dados onde a distribuição dos erros é simétrica e os outliers não são predominantemente problemáticos. No entanto, sua aplicabilidade pode ser limitada em situações com presença significativa de outliers, devido à sua sensibilidade a desvios extremos (H et al., 2016).

2.9.2 Raiz do Erro Quadrático Médio (RMSE)

A Raiz do Erro Quadrático Médio (RMSE) é uma métrica derivada do Erro Quadrático Médio (MSE) e é comumente utilizada para avaliar a precisão de modelos de previsão. Enquanto o MSE mede a média dos quadrados das diferenças entre os valores previstos e reais, o RMSE representa a raiz quadrada desse valor médio, conforme demonstrado na equação 2.15.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{2.15}$$

onde n é o número de elementos, Y_i é o valor real e $\hat{Y}_i$ é o valor previsto.

O RMSE oferece uma medida da dispersão dos erros, expressa na mesma unidade que a variável-alvo, facilitando a interpretação dos resultados. Assim como o MSE, o RMSE é sensível a desvios maiores, sendo útil em cenários onde se deseja penalizar erros

mais significativos de maneira proporcional. No entanto, é importante observar que o RMSE também pode ser influenciado por outliers, embora em menor grau em comparação com o MSE, proporcionando uma avaliação mais robusta da performance do modelo em diferentes cenários (HYNDMAN; KOEHLER, 2006a; H et al., 2016).

2.9.3 Erro Absoluto Médio (MAE)

O Erro Absoluto Médio (MAE) é outra métrica comumente usada para avaliar modelos de previsão. Diferentemente do MSE, o MAE mede a média das diferenças absolutas entre os valores previstos e os valores reais (HYNDMAN; KOEHLER, 2006a). Essa diferença fornecendo uma medida mais robusta em presença de desvios extremos (CHAI; DRAXLER, 2014). A equação 2.16 apresenta o cálculo do MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \tag{2.16}$$

onde n é o número de elementos, Y_i é o valor real e $\hat{Y}_i$ é o valor previsto.

O MAE é particularmente eficaz em contextos onde a resistência a outliers é crucial (WILLMOTT; MATSUURA, 2005). Portanto, o MAE é recomendado em situações onde a simplicidade, a interpretabilidade e a resistência a outliers são prioritárias na avaliação do desempenho do modelo de regressão.

2.9.4 Erro Percentual Absoluto Médio (MAPE)

O Erro Percentual Absoluto Médio (MAPE) é uma métrica que expressa a precisão como uma porcentagem. Ele mede a média das diferenças percentuais absolutas entre os valores previstos e os valores reais, oferecendo uma perspectiva percentual clara (HYNDMAN; KOEHLER, 2006a). A equação 2.17 apresenta o cálculo do MAPE.

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| \tag{2.17}$$

onde n é o número de elementos, Y_i é o valor real e $\hat{Y}_i$ é o valor previsto.

O MAPE é uma métrica valiosa na avaliação de modelos, especialmente útil em contextos onde a comparação relativa dos erros é mais informativa do que a magnitude

absoluta dos erros (MAKRIDAKIS, 1993). No entanto, existe um problema quando Y_i é um valor próximos a zero, isso faz com que o valor do MAPE estoure e ele se torna uma métrica irrelevante para comparação. Uma abordagem comum para mitigar esse problema é a adição de uma constante pequena aos valores reais, evitando assim a divisão por valores extremamente baixos (HYNDMAN; KOEHLER, 2006b).

2.9.5 Coeficiente de Determinação (R^2)

O coeficiente de determinação, também conhecido como R^2, é uma métrica que indica a proporção da variação na variável dependente que é previsível a partir das variáveis independentes (MONTGOMERY; PECK; VINING, 2015). A equação 2.18 apresenta o cálculo do R2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{2.18}$$

onde n é o número de elementos, Y_i é o valor real, $\hat{Y}_i$ é o valor previsto e $\bar{Y}$ é a média dos valores previstos.

O R^2 mede a proporção da variância na variável dependente que é previsível a partir das variáveis independentes, fornecendo uma indicação de quão bem as entradas explicam a saída (DRAPER; SMITH, 1998). Esta métrica é particularmente útil em contextos onde é importante quantificar a força da relação entre as variáveis e o modelo. Um valor de R^2 próximo de 1 indica melhores resultados. O R^2 é mais apropriado para modelos lineares, onde a interpretação de proporção da variância explicada é mais clara. No entanto, o R2 pode ser usado em modelos não lineares. Contudo, a aplicação direta do R2 em modelos não lineares pode ser limitada, uma vez que o R2 pode não capturar totalmente a complexidade de modelos não lineares mais avançados. O que pode levar a interpretações enganosas (Legates; McCabe, 1999).

2.10 Considerações Finais

Ao longo deste capítulo de fundamentação teórica, foram explorados os conceitos e técnicas relacionados à análise de séries temporais, com ênfase nas redes neurais artificiais que têm

se mostrado extremamente eficazes na previsão de séries temporais, oferecendo uma alternativa robusta às técnicas estatísticas clássicas. Além disso, foram explicados conceitos, como: DMD, KNN, *Random Forest*, Interpolação Linear, Interpolação Sazonal, que foram a base da construção de modelos de tratamento de dados faltantes, descrito no capítulo 4.

3 Trabalhos Relacionados

Nesse capítulo foi feito uma revisão da literatura de estudos que são relevantes para a previsão de radiação solar, o cerne da presente monografia. Essas referências bibliográficas ajudam a construir um sólido embasamento teórico, além de oferecer a oportunidade de se realizar uma avaliação comparativa das metodologias empregadas, proporcionando uma justificativa para as técnicas usadas e permitindo a identificação de possíveis lacunas no conhecimento que ainda precisam ser exploradas.

O trabalho de Peng et al. (2021) apresentou uma abordagem baseada em *Deep Learning*, denominada *CEN-SCA-BiLSTM*. Esta estrutura combina o *complete ensemble empirical mode decomposition with adaptive noise* (CEEMDAN), o *Sine Cosine Algorithm* (SCA), e o *Bi-directional long short-term memory* (BiLSTM) para aprimorar a previsão da radiação solar. O modelo híbrido CEN-SCA-BiLSTM demonstrou superioridade em precisão sobre modelos tradicionais, alcançando resultados significativos em várias métricas de avaliação. Uma comparação entre o presente trabalho e o modelo integrado CEN-SCA-BiLSTM revela diferenças fundamentais nas metodologias adotadas. O modelo CEN-SCA-BiLSTM não apresenta uma método para tratamento de dados faltante, se concentrando no uso de uma combinação sofisticada de técnicas de aprendizado profundo e algoritmos de otimização para melhorar a precisão das previsões. Enquanto isso, o foco deste trabalho reside na resolução de problemas relacionados a dados faltantes, uma etapa preliminar crítica para qualquer análise preditiva, que também influi na precisão das previsões realizadas pelos modelos.

O trabalho de Narvaez et al. (2021) também explora o uso de técnicas de aprendizado de máquina com a adição do uso *site-adaptation* para a mesma finalidade. A metodologia envolve a coleta de dados de várias fontes, incluindo informações de satélite e sensores terrestres, seguida do treinamento de modelos de aprendizado de máquina com uma combinação de técnicas de aprendizado supervisionado e não supervisionado. Essa referência apresenta um nova perspectiva na construção de modelos de previsão de radiação com base em diferentes fontes de dados. No entanto, não é apresentado nenhum

processo para o tratamento dos dados faltantes dessas múltiplas fontes.

O estudo de Cannizzaro et al. (2021) adotou uma abordagem diversificada. A pesquisa se concentrou na aplicação de redes convolucionais (CNN) e técnicas de aprendizado conjunto para prever a radiação solar. Foram exploradas arquiteturas híbridas que mesclam CNN, LSTM e Random Forest. A metodologia adotada combinou duas CNNs para processar sinais decompostos e modelos de regressão para processar outras medidas meteorológicas, resultando em previsões de radiação de curto e longo prazo. Além disso, ao lidar com dados ausentes, o estudo optou por uma técnica de imputação simples, preenchendo valores ausentes com interpolação linear. Porém, enquanto nossa pesquisa dá ênfase em análisar diversos métodos de imputação de dados e ao uso específico de LSTM, o estudo de Cannizzaro et al. (2021) foca em explorar o uso de redes mais complexar para realizar a previsão de maneira mais precisa.

O trabalho intitulado *"A critical overview of the (Im)practicability of solar radiation forecasting models"* de Babatunde et al. (2023) oferece um panorama amplo, destacando a relevância e os desafios dos modelos de previsão de radiação solar em aplicações práticas. Este estudo discute a diversidade dos modelos existentes, desde aqueles baseados em dados de satélite até os impulsionados por inteligência artificial, e evidencia lacunas de pesquisa, particularmente na generalização desses modelos para locais sem instrumentação adequada. Esse estudo forneceu uma base para compreender a importância e os desafios acerca do tema: "previsão de radiação solar".

No trabalho realizado por Basílio, Saporetti e Goliatt (2023), apresenta-se um modelo evolutivo de aprendizado de máquina com o objetivo de aprimorar as previsões de radiação solar. Esse modelo combina algoritmos evolutivos com técnicas de aprendizado de máquina, efetuando uma otimização integrada que inclui tanto a seleção de subconjuntos de variáveis quanto a calibração de hiperparâmetros, alcançando resultados que prometem superar os métodos tradicionais. De forma complementar, em Basílio et al. (2023), o mesmo autor introduz um modelo com foco específico no estado de Minas Gerais, constituindo uma referência comparativa valiosa. No entanto, ao comparar com o presente estudo, os trabalhos divergem em termos metodológicos; enquanto Basílio et al. (2023) se concentra na construção de um modelo evolutivo para previsão, ele não detalha

um método para o tratamento de dados ausentes.

Adicionalmente, o estudo conduzido por Parra-Plazas, Gaona-Garcia e Plazas-Nossa (2023) concentra-se no tratamento de séries temporais no contexto de dados meteorológicos. O trabalho se destaca ao propor uma estratégia numérica que visa o tratamento de valores atípicos e o tratamento de dados faltantes em séries temporais provenientes de estações meteorológicas, empregando três métodos distintos: a média, a regressão linear sem parâmetros de incerteza e a transformada discreta de Fourier (DFT). No entanto, o estudo realizado por Parra-Plazas, Gaona-Garcia e Plazas-Nossa (2023) limita-se à exploração de métodos para o tratamento de dados e não analisa o impacto desse tratamento nas previsões feitas por modelos de previsão.

4 Metodologia

Neste capítulo, é apresentada uma descrição detalhada da metodologia que foi empregada para conduzir o estudo de previsão de radiação solar, desde a coleta de dados até a implementação e a avaliação do modelo de aprendizado profundo.

4.1 Coleta de dados

A coleta de dados é um aspecto essencial de qualquer pesquisa, pois o tipo e a qualidade dos dados coletados podem influenciar significativamente os resultados. Neste estudo, utilizamos dados meteorológicos horários gerados pelas estações de medição automáticas (EMAs) presentes no banco de dados meteorológico do portal do Instituto Nacional de Meteorologia (INMET)[1]. O INMET é fonte confiável e amplamente utilizada para estudos relacionados ao clima e condições atmosféricas.

O presente estudo se limita as cidades da Zona da Mata Mineira, para tal, iremos selecionar os dados das EMAs das cidades indicadas na Figura 4.1 e na Tabela 4.1.

[1] https://bdmep.inmet.gov.br/

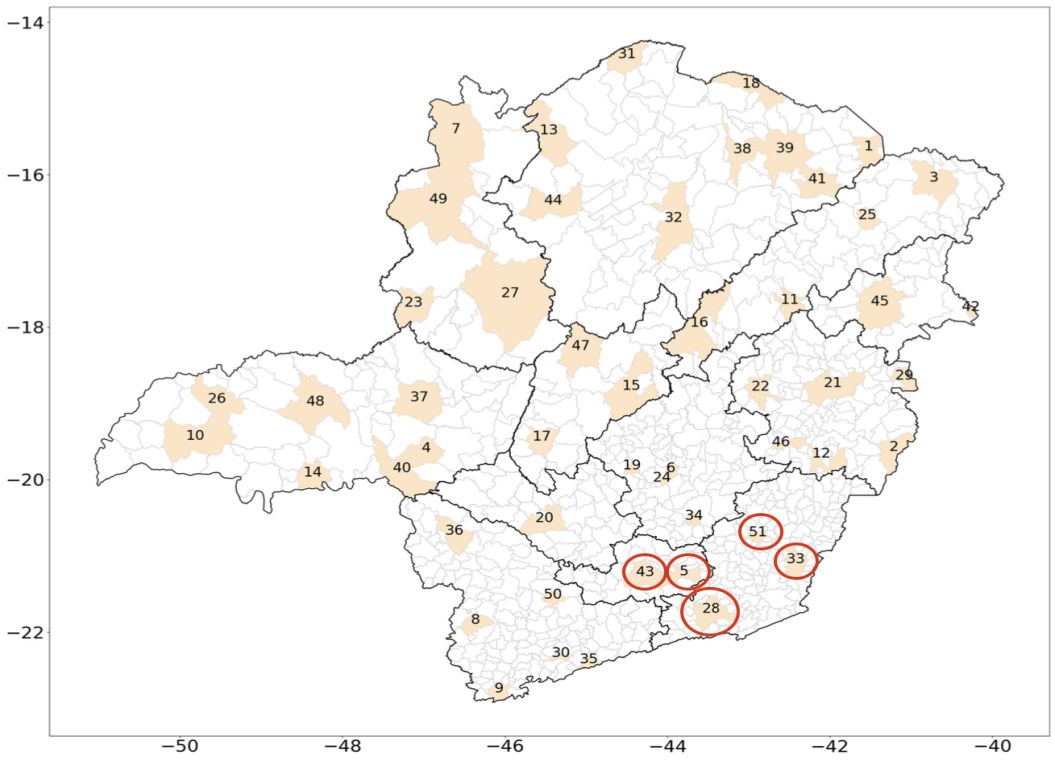

Figura 4.1: Mapa da EMAs do estado de Minas Gerais com as cidades selecionadas em destaque.

Cidade Selecionadas	**Latitude (S)**	**Longitude (W)**	**Altitude (m)**	**Período**
Barbacena	-21.228	-43.767	1,169	Janeiro/2003 à Dezembro/2022
Juiz de Fora	-21.769	-43.364	937	Maio/2007 à Dezembro/2022
Muriaé	-21.104	-42.375	283	Agosto/2006 à Dezembro/2022
São João Del Rei	-21.106	-44.250	930	Junho/2006 à Dezembro/2022
Viçosa	-20.762	-42.864	698	Setembro/2005 à Dezembro/2022

Tabela 4.1: Período dos dados coletados da cidades selecionadas.

As variáveis climáticas disponível nesse banco de dados estão indicas na Tabela 4.2.

Variáveis Disponíveis	
PRECIPITAÇÃO TOTAL (PRECIPITATION)	PRESSÃO ATMOSFÉRICA (PRESSURE_r)
PRESSÃO ATMOSFÉRICA MIN. (PRESSURE_rMin)	PRESSÃO ATMOSFÉRICA MAX. (PRESSURE_rMax)
TEMPERATURA DO AR (Temp)	TEMPERATURA DO PONTO DE ORVALHO (Temp_dew)
TEMPERATURA MÁXIMA (Temp_max)	TEMPERATURA MÍNIMA (Temp_min)
TEMPERATURA ORVALHO MAX (Temp_dew_max)	TEMPERATURA ORVALHO MIN. (Temp_dew_max)
UMIDADE RELATIVA MAX. (Humidity_hMax)	UMIDADE RELATIVA MIN. (Humidity_hMin)
UMIDADE RELATIVA DO AR (Humidity_h)	VENTO, DIREÇÃO HORÁRIA (gr) (Wind_d)
VENTO, RAJADA MÁXIMA (Wind_max)	VENTO, VELOCIDADE (Wind_s)
RADIAÇÃO SOLAR GLOBAL - GHI (Radiation)	

Tabela 4.2: Variáveis disponíveis no banco de dados meteorológico do INMET

Depois de coletar as bases de dados, analisou-se a correlação das variáveis disponível com a variável de radiação solar. Esta correlação está indicada na Figura 4.2.

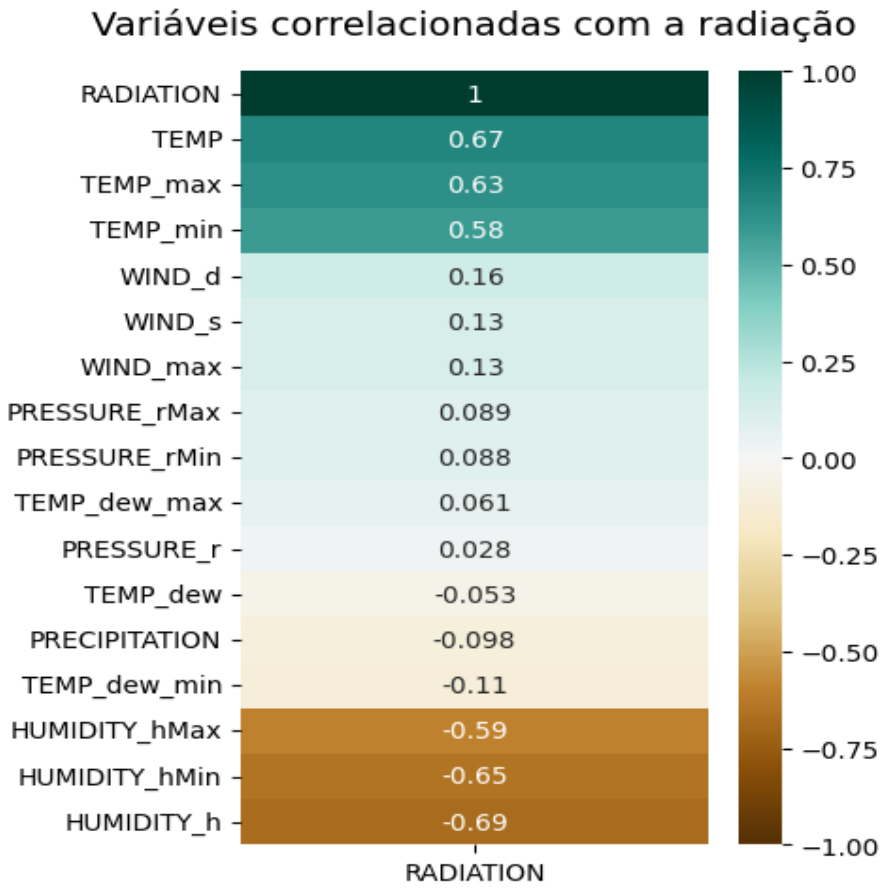

Figura 4.2: Correlação das variáveis disponível com a radiação solar global.

Ao analisar a correlação presente na Figura 4.2, foram selecionadas para compor a base de dados do trabalho as variáveis:

- Radiação Solar Global (GHI): Esta é a quantidade total de radiação solar, direta

e difusa, que incide em uma superfície horizontal. Medida em quilojoule por metro quadrado (KJ/m^2), a GHI varia de acordo com o ângulo de incidência dos raios solares e é uma variável importante para entender a quantidade de energia solar disponível em um determinado local e momento;

- Temperatura Média ($Temp$): A temperatura média fornece informações valiosas sobre as condições climáticas de um local. Medida em grau celsius C°. Ela pode influenciar vários fatores, como a eficiência dos painéis solares;

- Umidade Relativa do Ar Média ($Umid$): A umidade relativa do ar, medida em termos percentuais, pode afetar a quantidade de radiação solar que alcança a superfície terrestre, influenciando a GHI.

Durante a fase de análise deste estudo, identificamos um aspecto crítico que merece atenção especial: a ocorrência de dados faltantes nas variáveis meteorológicas escolhidas para os períodos indicados na Tabela 4.1 de cada cidade selecionada. Esta é uma preocupação comum em estudos que envolvem séries temporais.

A Tabela 4.3 apresenta um visão clara da quantidade de lacunas nos dados para cada cidade selecionada no estudo.

	Radiação	**Temperatura**	**Umidade**
Barbacena	48.79%	6.53%	12.05%
Juiz de Fora	44.54%	1.55%	1.59%
Muriaé	48.34%	3.68%	3.88%
São João Del Rei	48.33%	4.83%	4.99%
Viçosa	47.74%	2.72%	3.71%

Tabela 4.3: Percentual de dados faltantes por cidade.

Além disso, nas figuras 4.3 e 4.4 temos um exemplo representado graficamente de como essas lacunas estão distribuidas em uma porção dos dados.

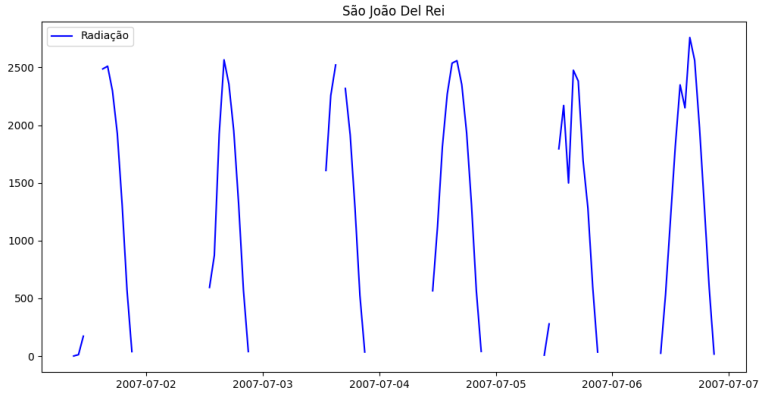

Figura 4.3: Gráfico representado a distribuição de dados faltantes em uma porção da série temporal de radiação da cidade de São João Del Rei. Elaborada pelo Autor.

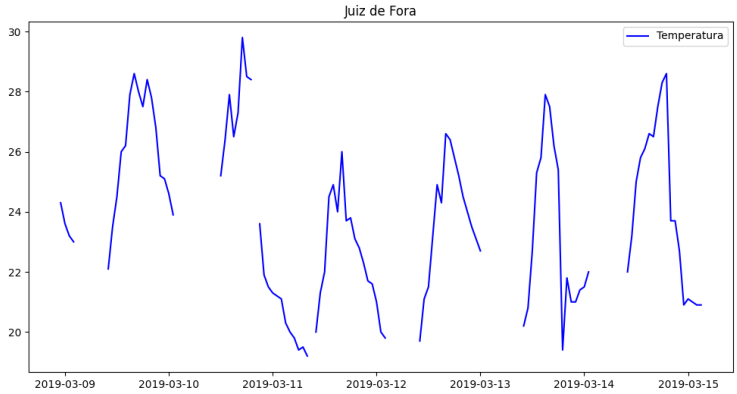

Figura 4.4: Gráfico representado a distribuição de dados faltantes em uma porção da série temporal de temperatura da cidade de Juiz de Fora. Elaborada pelo Autor.

Nas seções subsequentes, abordamos as estratégias e metodologias que implementamos para tratar esses dados faltantes.

4.2 Pré-processamento dos Dados

4.2.1 Redimensionamento

Conforme apresentado na seção 4.1, os dados coletados possuem frequência horária. Contudo, com o objetivo de estabelecer uma comparação com outros estudos da literatura, como por exemplo, o estudo conduzido por Basílio et al. (2023) que usam dados diários, decidiu-se pelo redimensionamento dos dados para uma escala diária. A metodologia de

agregação adotada segue o padrão descrito a seguir:

- Irradiância Solar Global (GHI): é a soma dos valores horários do dia;
- Temperatura Média ($Temp$): é a média dos valores horários do dia;
- Umidade Relativa do Ar ($Umid$): é a média dos valores horários do dia.

É relevante salientar que lacunas foram identificadas nas faixas horárias dos dados coletados. Assim, propomos uma análise que contemple a hipótese de tratar esses dados faltantes previamente ao redimensionamento, com o intuito de avaliar o impacto, na previsão, decorrente dessa variação no processo de pré-processamento dos dados.

4.2.2 Tratamento de dados faltantes

Como mencionado na seção 4.1, o conjuntos de dados temporais coletados da base de dados do INMET apresentou lacunas de dados faltante. Essas ausências podem surgir devido a variadas razões, sejam elas falhas de equipamento, manutenções programadas ou mesmo erros no processo de coleta. Sendo a presença dessas lacunas notória, sem um tratamento adequado, pode comprometer a precisão dos modelos preditivos elaborados a partir desses dados.

Diante do desafio imposto pelas lacunas nesse trabalho, decidimos fazer um estudo de caso aprofundado a fim de escolher as melhores técnicas de tratamento de dados faltantes e ruidosos.

A metodologia adotada para este estudo de caso inicial é descrita na seção 4.6. Em nossas investigações preliminares, focamos nas seguintes técnicas para o tratamento de dados faltantes:

- **DMD (Dynamic Mode Decomposition)**: Esta técnica, que constitui a proposta inicial deste estudo, é parte de um projeto apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).
- **Random Forest para Imputação de Dados**: Reconhecida na literatura por sua eficácia na imputação de dados faltantes (STEKHOVEN; BüHLMANN, 2011).

- **Interpolação Linear**: Amplamente empregada no tratamento de dados faltantes por sua simplicidade (CHATFIELD, 2003).
- **Interpolação Sazonal**: Esta técnica tem ganhado destaque pela sua habilidade em completar dados sazonais através da decomposição sazonal.
- **KNN (K-Nearest Neighbors)**: Recomendada por várias bibliotecas de aprendizado de máquina.

O primeiro estudo de caso deste trabalho consiste em analisar essas técnicas de imputação de dados afim de selecionar os modelos mais eficientes para o tratamento de dados faltantes nas séries temporais utilizadas neste trabalho.

4.2.3 Conjunto de dados

A série temporal foi dividida, para ser usada no treinamento dos modelos de predição, seguindo o modelo "*holdout*", mantendo o aspecto temporal da série, conforme detalhado na seção 2.8. As proporções adotadas para esta divisão foram escolhidas baseada em experimentos preliminares, com intuito de otimizar o desempenho do modelo de predição:

- Conjunto de treino: Este conjunto é composto por 70% dos dados da série temporal;
- Conjunto de validação: Contém 10% da série temporal;
- Conjunto de teste: Contém 20% do total de dados da série temporal.

A Tabela 4.4 detalha o período dessa divisão para cada cidade da Zona da Mata Mineira.

	Treino	**Validação**	**Teste**
Barbacena	Jan/2003 à Dez/2016	Jan/2017 à Dez/2018	Jan/2019 à Dez/2022
Juiz de Fora	Mai/2007 à Abr/2018	Mai/2018 à Nov/2019	Dez/2019 à Dez/2022
Muriaé	Ago/2006 à Jan/2018	Fev/2018 à Out/2019	Nov/2019 à Dez/2022
São João Del Rei	Jun/2006 à Jan/2018	Fev/2018 à Out/2019	Nov/2019 à Dez/2022
Viçosa	Set/2005 à Out/2017	Nov/2017 à Jul/2019	Ago/2019 à Dez/2022

Tabela 4.4: Período de cada porção dos dados seguindo o modelo *"Houldout"*.

4.2.4 Normalização dos Dados de Entrada

Neste estudo, os dados foram normalizados para um intervalo de $[0, 1]$. Esta transformação foi realizada utilizando a função *MinMaxScaler* do pacote de pré-processamento da biblioteca *Scikit-learn* [2]. A transformação *MinMaxScaler* implementa uma normalização que ajusta cada característica de forma independente para um intervalo especificado como descrito na seção 2.8. Vale ressaltar que o treinamento do normalizador leva em consideração somente os dados de treinamento, no entando, o normalização é aplicada na série completa.

4.3 Janelamento

Para treinar a rede neural para prever uma série temporal é imprescindível preparar os dados de acordo com o janelamento. Esse processo requer a organização dos dados em uma estrutura chamada janela deslizante, que configura a série temporal em matrizes de entrada X e saída Y apropriadas para o modelo (HAYKIN, 2009).

A janela deslizante é essencialmente uma técnica que permite a utilização de uma sequência fixa de observações anteriores para prever a variável de interesse em um período subsequente. Neste estudo, o objetivo é construir um modelo que possa prever a carga de radiação solar ghi_n (onde n representa o dia em questão), com base nos dados coletados durante os 5 dias anteriores. Essa configuração do janelamento foi baseada em estudos preliminares. Em outras palavras, os dados no intervalo de $[ghi_{n-5}, ..., ghi_{n-1}]$ são usados para prever ghi_n.

Esse trabalho foi realizado com série temporal multivariada conforme apresentado na seção 4.1, e a representação matricial da janela deslizante é expressa pelas Equações 4.1 e 4.2. A matriz X consiste em um conjunto de vetores, cada um contendo as observações dos 5 dias anteriores para as variáveis *ghi* (radiação), *temp* (temperatura) e *umid* (umidade relativa do ar). Assim:

[2] https://scikit-learn.org

$$X = \begin{bmatrix} ghi_{n-1} & temp_{n-1} & umid_{n-1} \\ ghi_{n-2} & temp_{n-2} & umid_{n-2} \\ & ... & \\ ghi_{n-5} & temp_{n-5} & umid_{n-5} \end{bmatrix} \tag{4.1}$$

A matriz Y é um vetor que contém o valor da carga de radiação ghi_n do dia que estamos tentando prever:

$$Y = \begin{bmatrix} ghi_n \end{bmatrix} \tag{4.2}$$

Em resumo, a matriz de entrada X representa os dados que alimentarão os modelos de previsão e a matriz Y representa o valor esperado de saída, que o modelo irá prever.

4.4 Construção do Modelo de *Deep Learning*

A arquitetura de *Deep Learning* proposta para esse estudo é baseada no uso de camadas LSTM que, como explicado na seção 2.8, têm a capacidade de aprender e reter as dependências de longo prazo nos dados. Esta escolha é justificada pela vasta aplicabilidade e eficácia das redes LSTM na previsão de séries temporais, conforme amplamente documentado na literatura científica. A habilidade das LSTM de capturar sequências temporais complexas e suas interdependências as torna particularmente adequadas para o contexto deste trabalho, onde as nuances e padrões intrínsecos nos dados são cruciais para previsões precisas. Além disso, o foco principal deste estudo é investigar a integração de modelos avançados de tratamento de dados com a arquitetura LSTM. Juntamente com a camada LSTM é introduzida uma camada de *dropout* para ajudar a prevenir o *overfitting*. Essa combinação de camada LSTM com camada de *dropout* formam a camada oculta do modelo proposto neste estudo, essa camada oculta pode aparecer N vezes a depender da configuração escolhida do modelo. Por fim, uma camada densa é usada para

a geração do *output* final do modelo. Essa configuração pode ser visualizada na Figura 4.5.

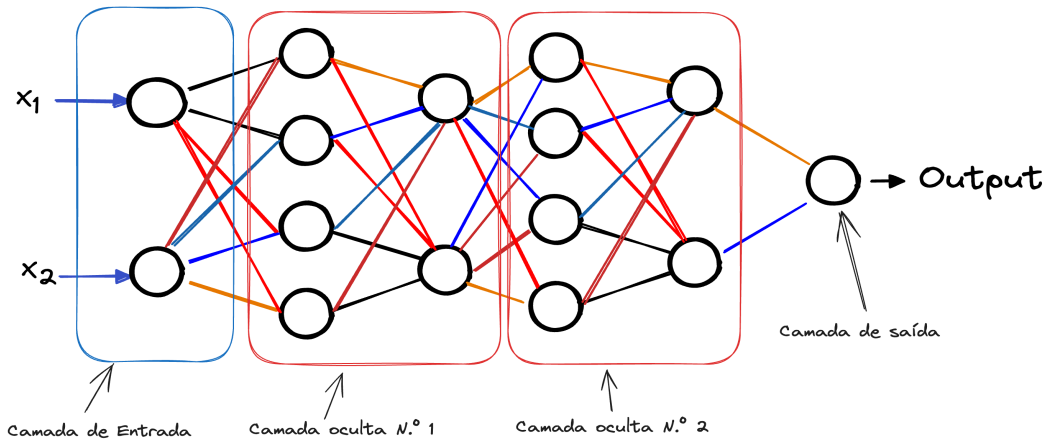

Figura 4.5: Arquitetura da Rede Neural. Elaborada pelo Autor.

4.5 Avaliação dos Modelos

A avaliação dos modelos de previsão de séries temporais do presente estudo visa quantificar o desempenho de cada modelo e compará-lo com outros modelos. Nesta seção, detalhamos o processo de avaliação utilizado, destacando as métricas de desempenho adotadas.

Foram empregados quatros métricas de desempenho, que fora explicitadas na seção 2.9: o Erro Quadrático Médio (MSE), o Raiz do Erro Quadrático Médio (RMSE), o Erro Absoluto Médio (MAE), O Erro Percentual Absoluto Médio (MAPE) e o coeficiente de determinação (R^2).

As métricas selecionadas desempenham uma função crucial na avaliação e comparação das previsões, desempenhando um papel significativo na compreensão da eficácia de cada modelo de previsão em diferentes cenários, particularmente em contextos meteorológicos. A escolha criteriosa dessas métricas permite uma análise aprofundada da precisão dos modelos, proporcionando insights valiosos sobre o desempenho relativo de cada abordagem em termos de acurácia e confiabilidade. Essa avaliação comparativa é essencial para embasar conclusões e informar decisões relacionadas à escolha do modelo mais adequado para aplicações práticas.

4.6 Metodologia do 1º caso de estudo

Como destacado, na seção 4.2.2, nosso estudo de caso vai se concentrar em analisar diferentes técnicas de imputação de dados faltantes: DMD, *Random Forest*, a Interpolação Linear, a Interpolação Sazonal e a técnica baseada em KNN.

Para cada método de imputação, analisamos os cenários de teste conforme os seguintes critérios:

1. **Qualidade da Imputação:** Como o método preenche os dados faltantes e sua capacidade de manter a estrutura original da série temporal;

2. **Qualidade da Previsão:** Uma análise comparativa entre os modelos de previsão treinados com os dados tratados pelos diferentes métodos de imputação e o modelo treinado com os dados completos.

Nossa metodologia para avaliar os modelos de imputação segue um procedimento estruturado. Inicialmente, identificamos em nossa base de dados a mais extensa sequência de dados sem lacunas para ser usada para comparar as técnicas de imputação de dados. Para essa finalidade, optamos pela série temporal da cidade de Viçosa, visto que esta dispõe da maior subsérie de dados completos, conforme ilustrado na Tabela 4.5.

Cidades	**Maior subsérie de dados (em dias)**
Viçosa	**3839**
Juiz de fora	3221
Barbacena	3126
Muriaé	1841
São João Del Rei	1509

Tabela 4.5: Tabela demostrando a maior subsérie de dados sem a presença de dados faltantes.

A Figura 4.6 representa uma porção da subsérie selecionada.

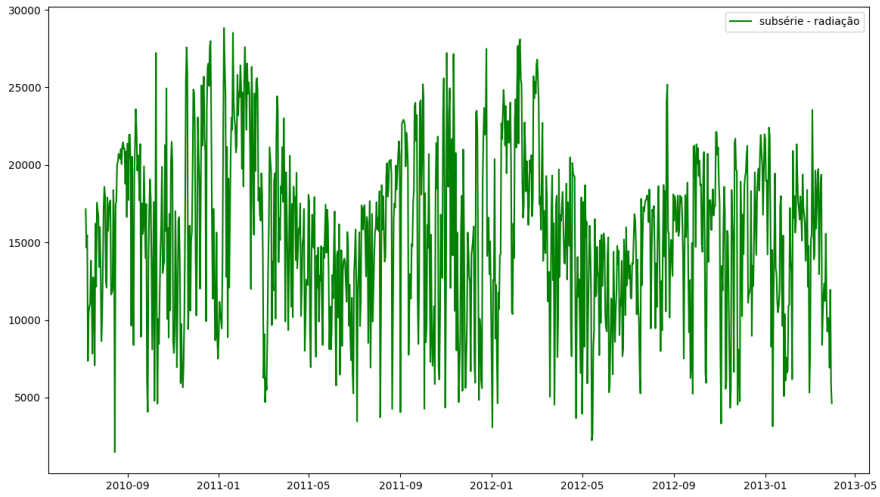

Figura 4.6: Subsérie sem dados faltantes do período de Ago/2010 à Mai/2013.

Com a maior subsérie de dados completos selecionada, inserimos deliberadamente lacunas em posições aleatórias da série temporal, reproduzindo os cenários de dados faltantes que ocorrem na prática. A simulação de lacunas de dados faltantes foi realizada de maneira aleatória e distribuída ao longo da subsérie selecionada para evitar qualquer viés na localização temporal das lacunas. Dessa forma, pudemos criar um cenário realista e desafiador para testar a capacidade dos modelos de imputação em preencher dados ausentes mantendo a integridade e continuidade da série temporal.

No primeiro cenário de teste, procedemos à simulação de dados faltantes na subsérie da cidade de Viçosa correspondendo a 20% do conjunto de dados total, com as lacunas variando aleatoriamente entre 1 e 4 dias. Esta escolha é representativa de situações reais em que pode ocorrer falhas de curta duração podem ocorrer devido a problemas técnicos ou manutenções periódicas que ocorrem nas EMAs.

A Figura 4.7 representa uma porção da subsérie selecionada, após o processo de inserção de dados faltantes do primeiro cenário.

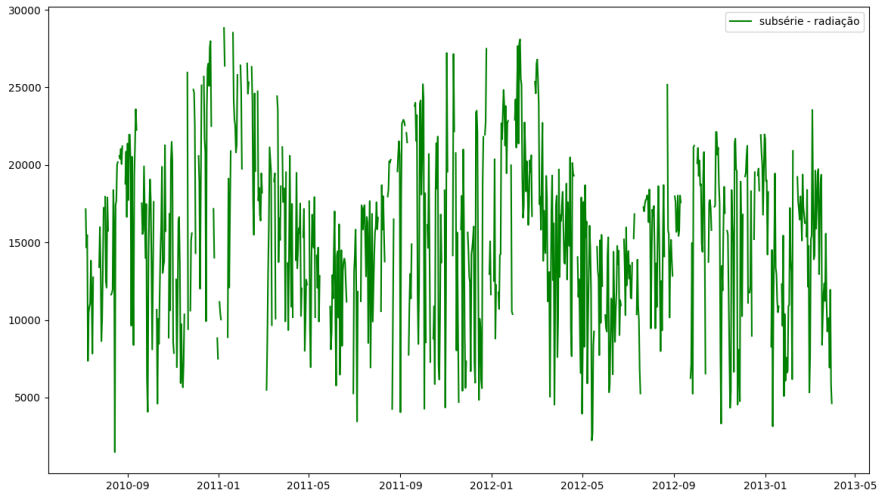

Figura 4.7: Subsérie com dados faltantes do período de Ago/2010 à Mai/2013.

No segundo cenário de teste avançamos nosso estudo de caso para testar a eficácia dos modelos de imputação diante de lacunas mais extensas na série temporal. Para esta análise, simulamos a ocorrência de dados faltantes que correspondem a 20% do volume total, com lacunas que variam de 5 a 40 dias. Esta simulação pode ser a representativa de interrupções prolongadas que podem ocorrer devido a eventos não previstos, como falhas extensas de equipamentos ou condições climáticas adversas que impeçam a coleta de dados, que ocorrem nas EMAs.

A Figura 4.8 representa uma porção da subsérie selecionada, após o processo de inserção de dados faltantes do primeiro cenário.

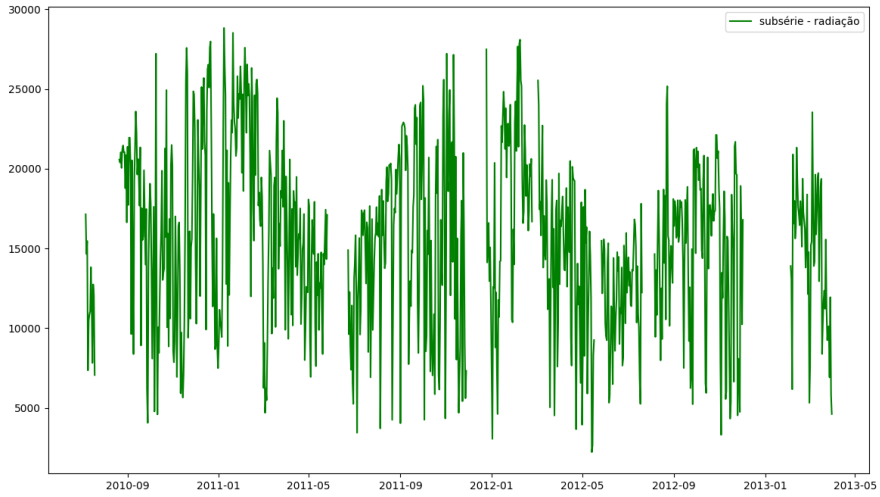

Figura 4.8: Subsérie com dados faltantes do período de Ago/2010 à Mai/2013.

A partir da elaboração do escopo do estudo de caso, submetemos as amostras dos cenários aos diferentes modelos de imputação propostos, visando restaurar a sequência de dados completos.

Ao término desta etapa, comparamos a amostra da sequência original intacta, sem a presença de dados faltantes, com a amostra de dados faltantes intencionalmente inseridos e posteriormente completos para cada modelo de imputação aplicado. Essa comparação nos dará a qualidade da imputação dos métodos de imputação de dados analisados.

O estágio subsequente envolve realizar previsões com base nessas séries completadas. E, então, aplicar critérios específicos de avaliação para comparar o desempenho dos modelos de prévisão treinados com a série completada por cada técnica de imputação de dados.

A Tabela 4.6 apresenta detalhadamente o período correspondente à divisão da subsérie selecionada, seguindo o padrão *"Holdout"*. Esta divisão é essencial para o treinamento do modelo de previsão.

	Treino	**Validação**	**Teste**
Subsérie	Jul/2010 à Nov/2017	Dez/2017 à Nov/2018	Dez/2018 à Jan/2021

Tabela 4.6: Período de cada porção dos dados seguindo o modelo *"Houldout"*.

Esse modelo de previsão segue o padrão explicitado na seção 4.4 e sua configuração está descrita na Tabela 4.7. Para efeito comparativo, vamos usar a mesma configuração do modelo de previsão para todos os cenários analisados. Os parâmetros dessa configuração foram selecionados através de um estudo preliminar com a série temporal completa.

Configuração do modelo de previsão	**Valores**
Número de neurônios LSTM (1° camada oculta)	200
Taxa Dropout (1° camada oculta)	0.1
Número de neurônios na camada de saída	1
Função de ativação da camada de saída	Sigmoid
Otimizador	Adam

Tabela 4.7: Configuração do modelo de previsão para avalição das técnicas de imputação de dados.

Ao analisar quais foram os melhores modelos de imputação de dados, com base nesse resultado de qualidade de previsão, obtemos a informação do impacto de cada técnica de imputação de dados no treinamento dos modelos de *Deep Learning*.

4.7 Metodologia do 2º caso de estudo

O segundo estudo de caso deste trabalho é dedicado à elaboração de modelos de previsão para as cidades selecionadas da Zona da Mata Mineira. O propósito central é desenvolver modelos eficientes para a previsão da radiação solar em cada uma dessas cidades.

Este experimento adota a metodologia detalhada ao longo do Capítulo 4. A seleção das técnicas de imputação de dados para este estudo foi influenciada pelas descobertas obtidas no primeiro estudo de caso, baseando-se na análise aprofundada dos resultados, conforme exposto na seção 5.1. Vale ressaltar que o tratamento de dados faltantes será aplicado, também, ao conjunto de teste. Isso se deve pelo percentual baixo de dados faltantes nessa porção dos dados. Então o tratamento dessa faixa de dados provavelmente não impactará significativamente os resultados.

A construção do modelo de previsão segue o protocolo estabelecido na seção 4.4. Contudo, visando aprimorar a eficácia do modelo e determinar os conjuntos de hiperparâmetros mais apropriados para cada série temporal, foi conduzido um processo de otimização. Este processo de otimização é descrito em detalhes na seção 4.7.1, enfatizando a busca por uma configuração que visa a precisão e eficiência nas previsões.

4.7.1 Otimização dos Hiperparâmetros da Rede Neural

A configuração dos hiperparâmetros da rede neural do modelo de previsão de cada cidade selecionada foi realizada por meio da otimização utilizando o pacote *Optuna*[2]. Esta ferramenta oferece recursos eficazes para ajustar os hiperparâmetros dos modelos de previsão de maneira eficiente. A estratégia de otimização empregada baseou-se no algoritmo *Tree-structured Parzen Estimator* (TPESampler) (WATANABE, 2023), e a métrica utilizada para a comparação entre os modelos durante o processo de otimização foi o Erro Quadrático Médio (MSE). Esse procedimento sistemático permitiu uma busca eficaz no espaço de hiperparâmetros, que estão listados na Tabela 4.8, resultando em configurações otimizadas para cada cenário do estudo, contribuindo assim para a robustez e desempenho aprimorado na tarefa de previsão de radiação solar.

A Tabela 4.8 apresenta os intervalos de valores escolhidos para a otimização dos hiperparâmetros no modelo de previsão empregado. Esses intervalos foram escolhidos com base em estudos preliminares.

Hiperparâmetros	**Intervalo de valores**
Número de camadas ocultas	1, 2, 3, 5
Unidades LSTM (por camada)	10, 32, 64, 128, 200, 300
Taxa de Dropout (por camada)	0, 0.1, 0.25, 0.35
Ativador da camada de saída	relu, softmax, sigmoid
Batch size	10, 16, 32, 64

Tabela 4.8: Intervalo de hiperparâmetros usados para otimização do modelo de previsão desenvolvido.

[2]https://optuna.readthedocs.io

5 Estudos de caso

5.1 Estudo de Caso 1: Avaliação dos Modelos de Imputação de Dados Faltantes

5.1.1 Análise dos Modelos de Imputação para o cenário 1

Avaliação da Qualidade da Imputação para o cenário 1

A série de lacunas pequenas foi tratada pelas técnicas de imputação de dados para completar essas lacunas de dados faltantes. Após, pegou-se a série completada de cada método de imputação de dados e comparou com a série original sem dados faltantes, utilizando as métricas de erro decritas na seção 2.9.

Métodos de Imputação	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
DMD	0.212	5206.44	1186.25	9.7%
Interpolação Linear	**0.878**	**2049.01**	**664.31**	**5.8%**
Interpolação Sazonal	0.430	4429.10	1573.08	10.2%
KNN	0.818	2503.45	855.19	7.2%
Random Forest	0.807	2576.20	895.76	7.8%

Tabela 5.1: Tabela apresenta os valores das métricas de erro do processo de inputação para lacunas pequenas.

A Tabela 5.1 apresenta os resultados que refletem a eficácia de cada método de imputação no tratamento de dados faltantes em lacunas pequenas. Uma inspeção cuidadosa desses resultados revela que os métodos de Interpolação Linear, *Random Forest* e KNN apresentaram um bom desempenho, destacando a Interpolação Linear como o melhor método para esse cenário de teste.

Por outro lado, os métodos DMD e Interpolação Sazonal não alcançaram o mesmo nível de desempenho nesse cenário específico. As limitações destes métodos tornaram-se evidentes, pois mostraram-se menos eficientes em lidar com lacunas pequenas. Essa

constatação é crucial, pois indica que, embora possam ser adequados em outros contextos, para lacunas pequenas eles podem não ser a escolha mais confiável.

Avaliação da Qualidade da Previsão com Dados Imputados para o cenário 1

Após o tratamento dos dados faltantes da subsérie selecionda, os dados foram usados para treinar o modelo de previsão.

A Tabela 5.2 fornece uma perspectiva crucial sobre a qualidade das previsões realizadas após o processo de imputação para lacunas pequenas. Nesta análise, vamos comparar a precisão dos modelos treinados sob três diferentes condições: primeiro, com a série original que não apresenta dados faltantes; segundo, utilizando séries temporais com lacunas de dados faltantes; e, finalmente, com séries temporais que foram submetidas às técnicas de imputação de dados.

	R2	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Dados originais	0.360	4624.33	3600.96	0.323
Dados com lacunas	0.316	4653.13	3620.55	0.350
DMD	0.326	4650.64	3619.79	**0.348**
Interpolação Linear	0.328	4649.08	3634.68	0.349
Interpolação Sazonal	0.331	4635.76	3625.15	**0.348**
KNN	0.322	4667.50	**3616.80**	**0.348**
Random Forest	**0.333**	**4630.85**	3659.95	0.349

Tabela 5.2: Qualidade da previsão com os dados após o processo de inputação para lacunas pequenas.

As Figuras 5.1 e 5.2 representam a previsão realizada pelos modelos de previsão treinados com os dados tratados pelos 2 melhores modelos de imputação de dados, no período de Jan/2020 à Jun/2020. É importante destacar que, embora visualmente as previsões ilustradas pareçam bastante similares, as métricas de erro utilizadas para avaliação quantitativa revelam diferenças sutis entre elas. Esta observação sublinha a importância de uma avaliação quantificação de erro para uma interpretação precisa do desempenho dos modelos de previsão.

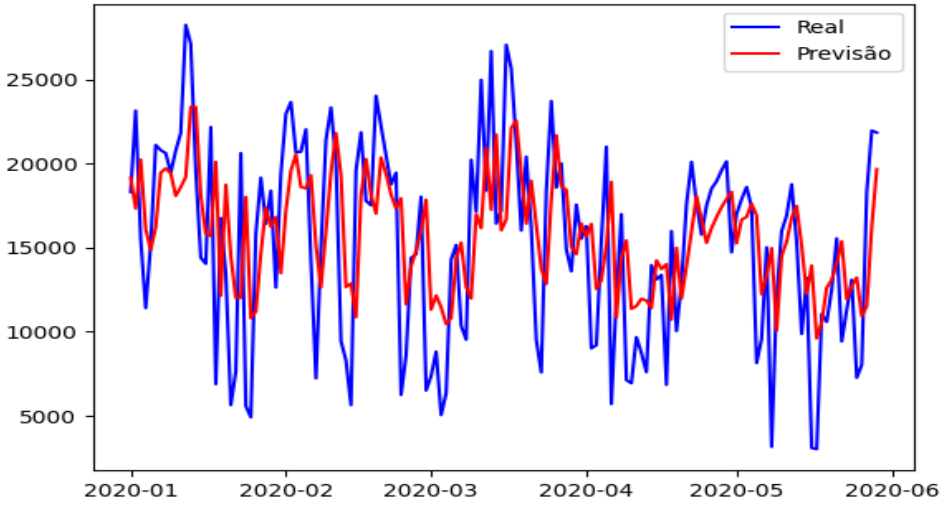

Figura 5.1: Previsão realizada como o modelo treinado com os dados completados pela técnina Random Forest.

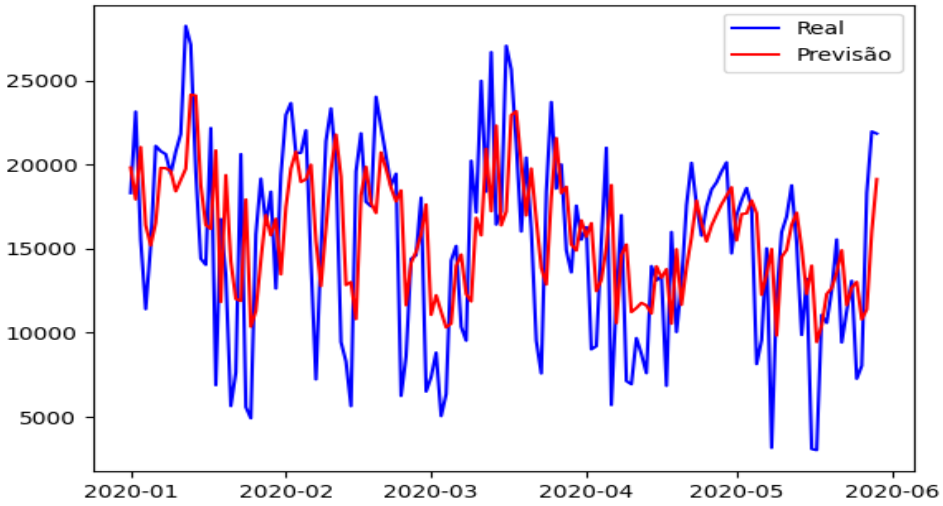

Figura 5.2: Previsão realizada como o modelo treinado com os dados completados pela técnina KNN.

Ao analisar os resultados de previsão obtidos após a imputação, observamos uma mudança significativa na desempenho das técnicas de imputação. No contexto deste cenário específico, não se observou uma técnica de imputação de dados faltantes que se sobressaísse em todas as métricas avaliadas. Contudo, a técnica de *Random Forest* demonstrou superioridade nas métricas de R2 e RMSE, enquanto a abordagem baseada

em KNN destacou-se nas métricas de MAE e MAPE. Dessa forma, ambas as técnicas foram identificadas como as mais eficazes neste cenário. Esta constatação é fundamental, pois revela que a análise puramente baseada na qualidade da imputação não é suficiente para determinar o melhor modelo de tratamento para lacunas menores. A habilidade dos modelos de imputação de preservar a integridade dos padrões dos dados para previsões precisas é um fator igualmente importante.

Ademais, é fundamental destacar um resultado específico que amplia a compreensão desse resultado. A interpolação linear, embora tenha demonstrado a melhor qualidade de imputação ao ser comparada com outras técnicas, enfrentou dificuldades quando submetida aos testes de previsão. Este aspecto exemplifica a complexidade na escolha do método de imputação, uma vez que o desempenho em cenários de previsão pode divergir consideravelmente da qualidade de imputação observada.

5.1.2 Análise dos Modelos de Imputação para cenário 2

Avaliação da Qualidade da Imputação para o cenário 2

Esta simulação desafia os modelos a lidar com cenários de perda de dados que podem afetar significativamente a tendência e sazonalidade da série. A aleatoriedade na distribuição e tamanho das lacunas garante que a simulação reflita uma ampla gama de possíveis interrupções reais, fornecendo um teste para as técnicas de imputação de dados.

A Tabela 5.3 detalha as métricas de erro para cada método de imputação aplicado a lacunas extensas.

Métodos de Imputação	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
DMD	0.723	3084.79	1090.25	8.6%
Interpolação Linear	0.706	3176.94	1156.73	8.8%
Interpolação Sazonal	0.208	5217.38	1917.61	12.0%
KNN	0.689	3269.26	1192.51	8.8%
Random Forest	**0.771**	**2784.64**	**1023.35**	**8.1%**

Tabela 5.3: Métricas de erro do processo de inputação para lacunas de dados extensas.

Os resultados indicam que os modelos *Random Forest*, DMD e Interpolação Linear

apresentam melhor desempenho na imputação de lacunas extensas. O método de KNN ficou um pouco abaixo, mas ainda apresentou resultados relevantes. Por outro lado, a Interpolação Sazonal revela limitações significativas neste cenário, sugerindo que pode não ser a opção mais eficaz para imputação.

Avaliação da Qualidade da Previsão com Dados Imputados para o cenário 2

A tabela 5.4 fornece uma avaliação dos modelos de *Deep Learning* treinados com os dados tratados por cada método de imputação. Esta análise visa compreender a eficácia dos dados imputados quando utilizados em modelos preditivos.

	R2	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Dados originais	0.360	4624.33	3600.96	0.323
Dados com lacunas	0.305	4693.13	3694.63	0.368
DMD	0.330	4659.56	3629.79	**0.348**
Interpolação Linear	0.330	4678.87	3636.87	0.349
Interpolação Sazonal	0.303	4697.76	3698.05	0.368
KNN	0.322	4667.50	3616.80	**0.348**
Random Forest	**0.335**	**4652.06**	**3614.90**	**0.348**

Tabela 5.4: Qualidade da previsão com os dados após o processo de inputação para lacunas extensas.

As Figuras 5.3 e 5.4 representa a previsão realizada pelos modelos de previsão treinados com os dados tratados pelos 2 melhores modelos de imputação de dados, no período de Jan/2020 à Jun/2020. É essencial ressaltar que, apesar das previsões representadas nas ilustrações aparentarem ser visualmente semelhantes, as métricas de erro aplicadas na avaliação quantitativa indicam diferenças entre elas. Tal constatação enfatiza a necessidade de uma análise quantitativa do erro para uma interpretação acurada do desempenho dos modelos preditivos.

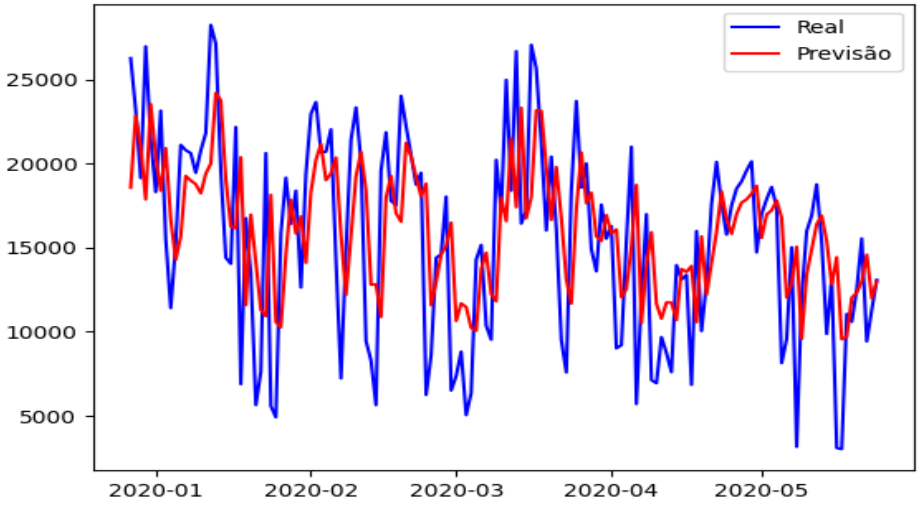

Figura 5.3: Previsão realizada como o modelo treinado com os dados completados pela técnina Random Forest.

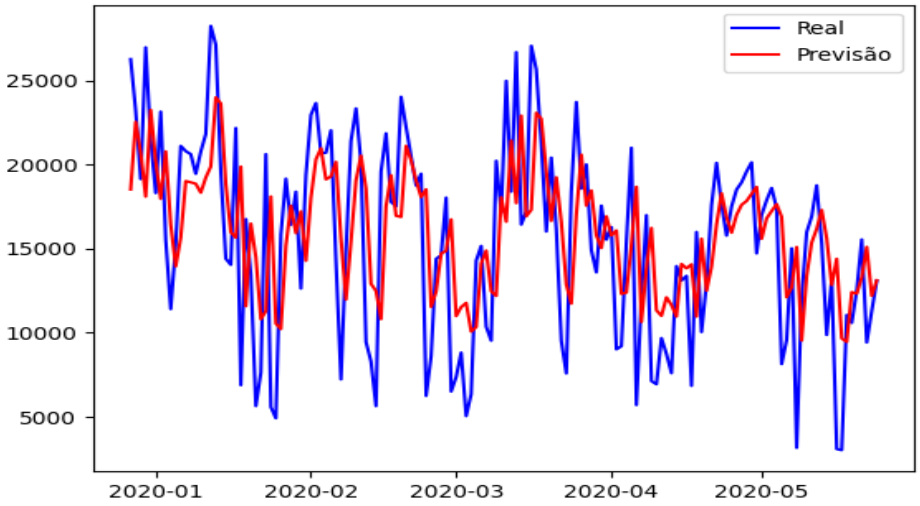

Figura 5.4: Previsão realizada como o modelo treinado com os dados completados pela técnina DMD.

Interessantemente, ao aplicar os modelos de previsão treinados com dados imputados, observamos que o Random Fores alcança os melhores resultados assim como aconteceu no cenário 1. Observe que o DMD também teve um resultado promissor nesse cenário. Estes pontos sugere que a análise da eficiência do processo de imputação, usando as métricas de erro, pode não ser o suficiente para entender todos os aspectos da im-

putação. Esse fenômeno ressalta a importância de avaliar os dados imputados não apenas pelo seu erro de imputação, mas também pelo seu impacto nas previsões finais, especialmente em cenários de lacunas extensas onde a integridade dos dados é crítica.

5.1.3 Conclusão

Este estudo proporcionou um entendimento valiosos sobre a imputação de dados em séries temporais, especialmente no contexto de previsão de radiação solar. Embora, na prática, não seja possível avaliar o erro de imputação diretamente em comparação com os dados reais, nossa análise preliminar revelou aspectos críticos a serem considerados ao selecionar métodos de imputação de dados para modelos de *Deep Learning* para previsão de séries temporais.

Um dos principais aprendizados é que a avaliação da qualidade da imputação baseada apenas em métricas de erro pode não ser suficiente para assegurar a eficácia dos dados para o treinamento de modelos de *Deep Learning*. Isso se deve, em parte, à possibilidade de que os padrões gerados por certos modelos de imputação de dados possam afetar negativamente o desempenho dos modelos de predição. No entanto, vale ressaltar que é necessário conduzir um estudo com maior quantidade de amostras com o objetivo de entender se esse comportamento se mantém em todos os cenários de forma consistente.

O modelo baseado em *Random Forest* mostrou-se particularmente eficaz, equilibrando a captura de padrões tanto de curto quanto de longo prazo, pois destacou-se por sua habilidade de manter a consistência dos dados tanto em lacunas pequenas quanto extensas de dados ausentes. Portanto, proporcionando um tratamento adequado dos dados faltantes.

O modelo DMD teve um desempenho notável em lacunas de dados mais extensas, mas revelou-se menos eficiente para lacunas menores. Isso sugere que o DMD pode ser mais apropriado para situações onde as lacunas de dados são substanciais.

Os modelos de KNN e Interpolação Linear, apesar de apresentarem bons resultados nas métricas de erro de imputação, não garantiram as melhores previsões. Uma possível explicação para isso é que esses métodos, sendo mais simples, tendem a criar um padrão de imputação que se alinha à média projetada dos dados, o que pode não ser

ideal para o treinamento das redes neurais. Contudo, esses métodos foram eficientes em cenários de lacunas pequenas, demonstrando serem suficientes para cobrir essas lacunas sem prejudicar o treinamento dos modelos preditivos.

Em resumo, a escolha do método de imputação deve ser feita com cautela, considerando não apenas a precisão da imputação, mas também o impacto subsequente na modelagem preditiva.

5.2 Estudo de Caso 2: Modelo LSTM para Previsão de Radiação Solar

No segundo estudo de caso, a metodologia adotada é detalhadamente descrita na seção 4.7. Conforme mencionado anteriormente, a escolha das técnicas de imputação de dados foi fundamentada nas constatações provenientes do primeiro estudo de caso. Neste contexto, a técnica *Random Forest* foi selecionada devido à sua destacada eficácia na previsão, evidenciada em ambos os cenários avaliados. Adicionalmente, o método DMD foi incorporado em virtude de seu desempenho promissor e relevância demonstrados nas previsões de lacunas extensas, o que acontece muito nos dados das séries reais analisadas. Complementarmente, a Interpolação Linear foi escolhida, tendo em vista sua eficiência na imputação de dados em lacunas de dados pequenas, que pode ocorrer nos dados reais.

5.2.1 Cenário de Barbacena - MG

No contexto do estudo realizado na cidade de Barbacena - MG, identificamos o melhor conjunto de hiperparâmetros, cujos detalhes estão delineados na Tabela 5.5.

Configuração do modelo de previsão de Barbacena - MG	
Parâmetros	**Valores**
Número de camadas ocultas	2
Número de neurônio LSTM (1° camada oculta)	350
Taxa Dropout (1° camada oculta)	0.35
Número de neurônio LSTM (2° camada oculta)	320
Taxa Dropout (2° camada oculta)	0
Função de ativação das camadas oculta	Tanh
Função de ativação da camada de saída (Dense)	Sigmoid
Otimizador	Adam

Tabela 5.5: Configuração do modelo de previsão de Barbacena - MG.

Na Tabela 5.6, são apresentados os resultados consolidados das métricas de erro: R2, RMSE, MAE e MAPE, obtidos a partir da média de 10 execuções independentes. Estes valores, que são exclusivamente do conjunto de teste, oferecem uma avaliação minuciosa do desempenho dos modelos de previsão. Eles são comparados em dois contextos distintos: primeiro, quando treinados com séries temporais que foram submetidas a diferentes técnicas de imputação de dados e, segundo, quando treinados com séries temporais que permaneceram sem qualquer tratamento de dados faltantes

Resultado das previsões realizadas com os dados da cidade de Barbacena - MG				
Técnica de tratamento de dados faltantes	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Sem tratamento de dados faltante	0.447	4728.48	3692.53	0.379
DMD	0.487	4470.10	3592.12	0.371
Interpolação Linear	**0.554**	**4180.98**	**3179.73**	**0.301**
Random Forest	0.507	4536.65	3483.30	0.311

Tabela 5.6: Resultado das previsões realizadas com os dados da cidade de Barbacena - MG.

Uma análise criteriosa dos resultados obtidos revela que, no cenário específico de Barbacena - MG, o modelo que demonstrou desempenho superior foi aquele treinado com a série temporal cujas lacunas foram preenchidas através da técnica de Interpolação Linear. O Random Forest emergiu como a segunda opção de melhor modelo, convergindo com as descobertas do estudo de caso 1.

A Figura 5.5 detalha a previsão com o modelo treinado com os dados completados pela técnica de imputação Interpolação Linear.

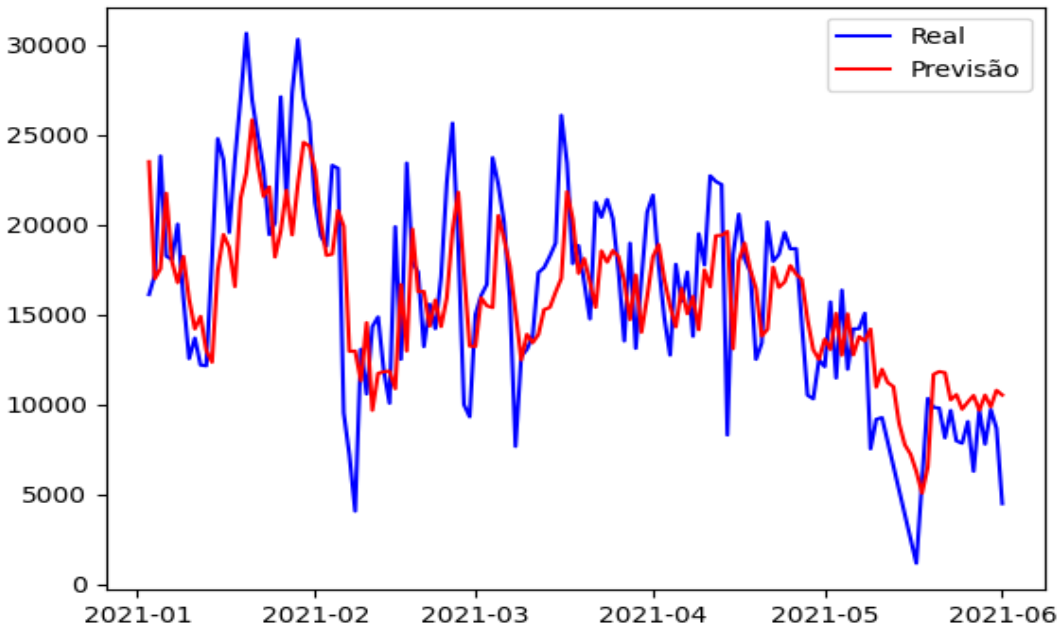

Figura 5.5: Previsão, com dados faltantes preenchidos pelo método Interpolação Linear, realizada para cidade de Barbacena no período de Janeiro de 2021 à Junho de 2021.

5.2.2 Cenário de Viçosa - MG

No contexto do estudo realizado na cidade de Viçosa - MG, identificamos o melhor conjunto de hiperparâmetros, cujos detalhes estão delineados na Tabela 5.7.

Configuração do modelo de previsão de Viçosa - MG	
Parâmetros	**Valores**
Número de camadas ocultas	1
Número de neurônio LSTM (1° camada oculta)	200
Taxa Dropout (1° camada oculta)	0.1
Função de ativação das camadas oculta	Tanh
Função de ativação da camada de saída (Dense)	Sigmoid
Otimizador	Adam

Tabela 5.7: Configuração do modelo de previsão de Viçosa - MG.

Na Tabela 5.8, são apresentados os resultados consolidados das métricas de erro: R2, RMSE, MAE e MAPE, obtidos a partir da média de 10 execuções independentes. Estes valores, que são exclusivamente do conjunto de teste, oferecem uma avaliação minuciosa do desempenho dos modelos de previsão. Eles são comparados em dois contextos distintos: primeiro, quando treinados com séries temporais que foram submetidas a diferentes técnicas de imputação de dados, e segundo, quando treinados com séries temporais que permaneceram sem qualquer tratamento de dados faltantes

Resultado das previsões realizadas com os dados da cidade de Viçosa - MG				
Técnica de tratamento de dados faltantes	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Sem tratamento de dados faltante	0.316	4463.85	3632.61	0.298
DMD	0.431	4557.86	3485.39	0.262
Interpolação Linear	0.429	4701.65	3579.73	0.293
Random Forest	**0.531**	**4416.05**	**3422.12**	**0.259**

Tabela 5.8: Resultado das previsões realizadas com os dados da cidade de Viçosa - MG.

Após uma avaliação dos resultados alcançados, observa-se que, no contexto particular de Viçosa - MG, o modelo que se destacou em termos de desempenho foi o que utilizou a série temporal com as lacunas preenchidas pela técnica de imputação de dados baseada em *Random Forest*. Notavelmente, o método de *Dynamic Mode Decomposition* (DMD) ficou em segundo lugar, também apresentando boas métricas de erro, reforçando sua eficácia como uma técnica de imputação viável e confiável neste contexto específico.

A Figura 5.6 detalha a previsão com o modelo treinado com os dados completados pela técnica de imputação *Random Forest*.

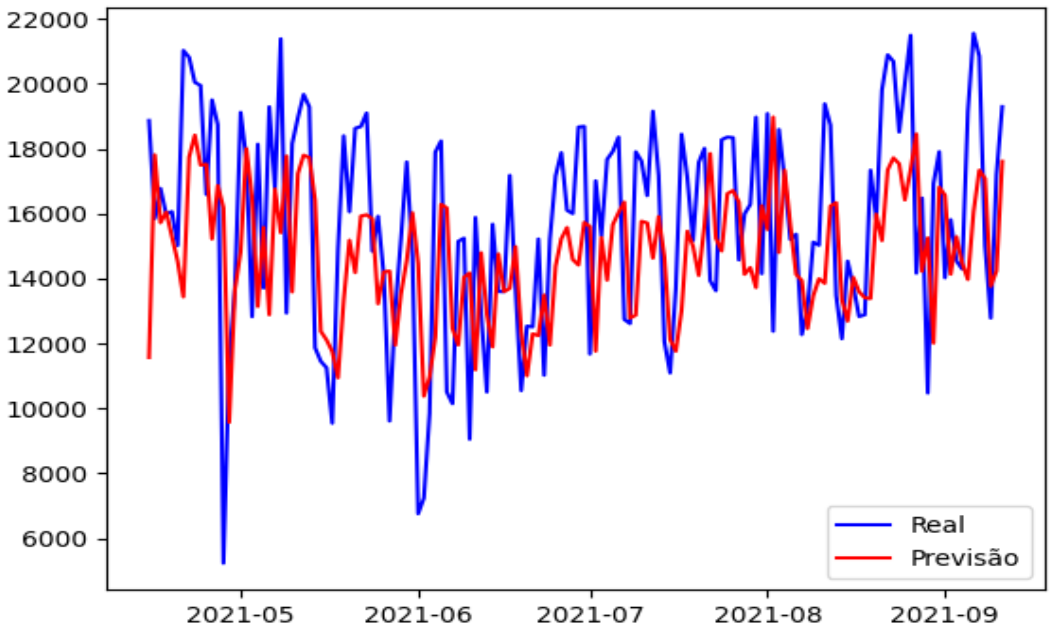

Figura 5.6: Previsão, com dados faltantes preenchidos pelo método *Random Forest*, realizada para cidade de Viçosa no período de Abril de 2021 à Agosto de 2021.

5.2.3 Cenário de Muriaé - MG

No contexto do estudo realizado na cidade de Muriaé - MG, identificamos o melhor conjunto de hiperparâmetros, cujos detalhes estão delineados na Tabela 5.9.

Configuração do modelo de previsão de Muriaé - MG	
Parâmetros	**Valores**
Número de camadas ocultas	3
Número de neurônio LSTM (1° camada oculta)	128
Taxa Dropout (1° camada oculta)	0
Número de neurônio LSTM (2° camada oculta)	64
Taxa Dropout (2° camada oculta)	0.25
Número de neurônio LSTM (3° camada oculta)	128
Taxa Dropout (3° camada oculta)	0.1
Função de ativação das camadas oculta	Tanh
Função de ativação da camada de saída (Dense)	Sigmoid
Otimizador	Adam

Tabela 5.9: Configuração do modelo de previsão de Muriaé - MG.

Na Tabela 5.10, são apresentados os resultados consolidados das métricas de erro: R2, RMSE, MAE e MAPE, obtidos a partir da média de 10 execuções independentes. Estes valores, que são exclusivamente do conjunto de teste, oferecem uma avaliação minuciosa do desempenho dos modelos de previsão. Eles são comparados em dois contextos distintos: primeiro, quando treinados com séries temporais que foram submetidas a diferentes técnicas de imputação de dados, e segundo, quando treinados com séries temporais que permaneceram sem qualquer tratamento de dados faltantes

Resultado das previsões realizadas com os dados da cidade de Muriaé - MG				
Técnica de tratamento de dados faltantes	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Sem tratamento de dados faltante	0.429	4572.89	3446.36	0.341
DMD	**0.483**	**4091.90**	**3089.97**	**0.295**
Interpolação Linear	0.467	4164.23	3183.18	0.323
Random Forest	0.473	4123.16	3140.77	0.310

Tabela 5.10: Resultado das previsões realizadas com os dados da cidade de Muriaé - MG.

A análise detalhada dos dados coletados indica que, no caso específico de Muriaé - MG, o modelo que apresentou um desempenho mais eficaz foi o treinado com a série temporal, onde as lacunas foram adequadamente preenchidas utilizando a técnica de DMD. Interessantemente, o *Random Forest* ficou em segundo lugar, demonstrando também um

resultado bastante positivo. Nota-se, contudo, que o teste realizado sem qualquer tratamento de dados faltantes obteve o pior desempenho, destacando a importância crucial das técnicas de imputação para a precisão e eficácia dos modelos de previsão em séries temporais.

A Figura 5.7 detalha a previsão com o modelo treinado com os dados completados pela técnica de imputação DMD.

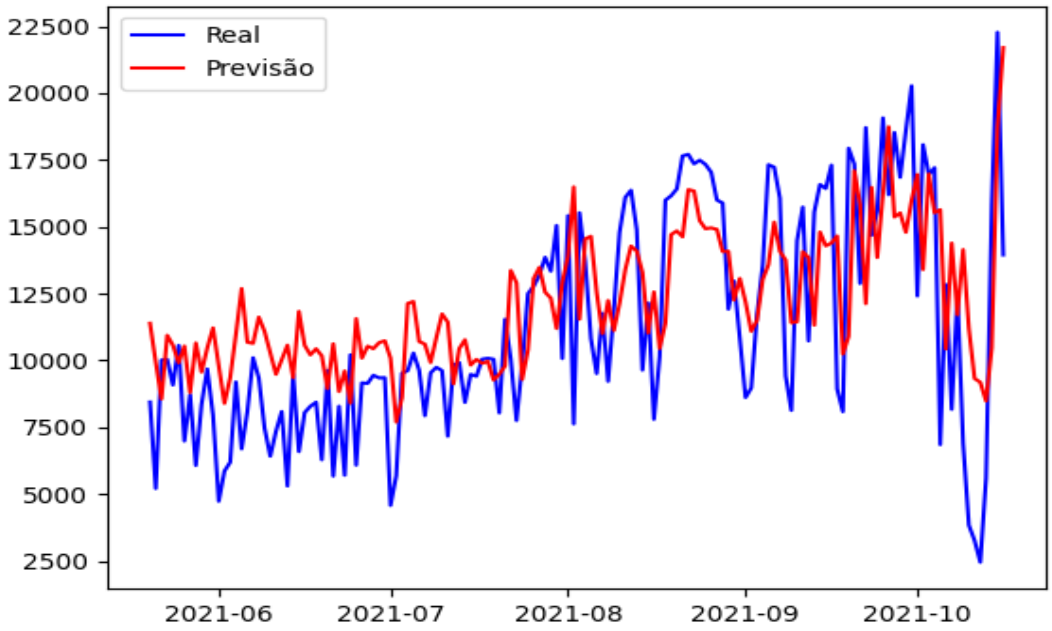

Figura 5.7: Previsão, com dados faltantes preenchidos pelo método DMD, realizada para cidade de Muriaé no período de Maio de 2021 à Outubro de 2021.

5.2.4 Cenário de Juiz de Fora - MG

No contexto do estudo realizado na cidade de Juiz de Fora - MG, identificamos o melhor conjunto de hiperparâmetros, cujos detalhes estão delineados na Tabela 5.11.

Configuração do modelo de previsão de Juiz de Fora - MG	
Parâmetros	**Valores**
Número de camadas ocultas	2
Número de neurônio LSTM (1° camada oculta)	100
Taxa Dropout (1° camada oculta)	0.25
Número de neurônio LSTM (2° camada oculta)	200
Taxa Dropout (2° camada oculta)	0
Função de ativação das camadas oculta	Tanh
Função de ativação da camada de saída (Dense)	Sigmoid
Otimizador	Adam

Tabela 5.11: Configuração do modelo de previsão de Juiz de Fora - MG.

Na Tabela 5.12, são apresentados os resultados consolidados das métricas de erro: R2, RMSE, MAE e MAPE, obtidos a partir da média de 10 execuções independentes. Estes valores, que são exclusivamente do conjunto de teste, oferecem uma avaliação minuciosa do desempenho dos modelos de previsão. Eles são comparados em dois contextos distintos: primeiro, quando treinados com séries temporais que foram submetidas a diferentes técnicas de imputação de dados, e segundo, quando treinados com séries temporais que permaneceram sem qualquer tratamento de dados faltantes

Resultado das previsões realizadas com os dados da cidade de Juiz de Fora - MG				
Técnica de tratamento de dados faltantes	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Sem tratamento de dados faltante	0.336	4542.55	3618.90	0.319
DMD	0.357	4364.34	3482.27	0.312
Interpolação Linear	0.379	4314.75	3409.28	0.310
Random Forest	**0.390**	**4293.79**	**3401.26**	**0.310**

Tabela 5.12: Resultado das previsões realizadas com os dados da cidade de Juiz de Fora - MG.

Após uma avaliação minuciosa dos resultados alcançados, constatou-se que, no contexto particular de Juiz de Fora - MG, o modelo que exibiu um desempenho notavelmente superior foi o que utilizou a série temporal complementada pela técnica de imputação de dados faltantes baseada em *Random Forest*. Interessante observar que a técnica de Interpolação Linear se posicionou em segundo lugar, demonstrando também re-

sultados satisfatórios. Contudo, é importante destacar que o teste realizado sem qualquer tratamento dos dados faltantes resultou no pior desempenho dentre todos.

A Figura 5.8 detalha a previsão com o modelo treinado com os dados completados pela técnica de imputação *Random Forest*.

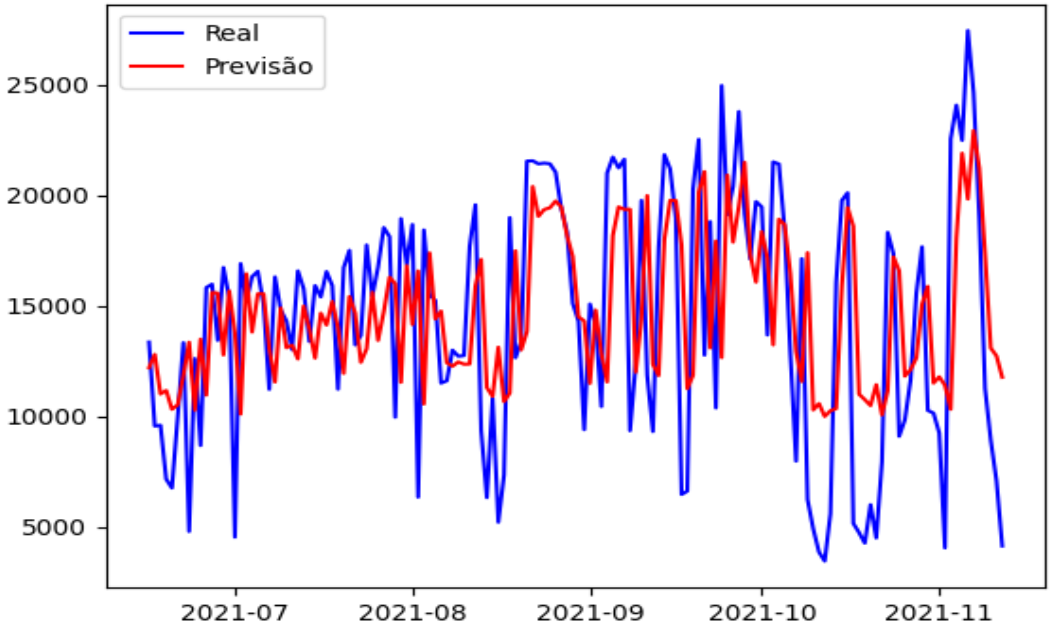

Figura 5.8: Previsão, com dados faltantes preenchidos pelo método *Random Forest*, realizada para cidade de Juiz de Fora no período de Junho de 2021 à Novembro de 2021.

5.2.5 Cenário de São João del Rei - MG

No contexto do estudo realizado na cidade de São João del Rei - MG, identificamos o melhor conjunto de hiperparâmetros, cujos detalhes estão delineados na Tabela 5.13.

Configuração do modelo de previsão de São João del Rei - MG	
Parâmetros	**Valores**
Número de camadas ocultas	1
Número de neurônio LSTM (1° camada oculta)	300
Taxa Dropout (1° camada oculta)	0.25
Função de ativação das camadas oculta	Tanh
Função de ativação da camada de saída (Dense)	Sigmoid
Otimizador	Adam

Tabela 5.13: Configuração do modelo de previsão de São João del Rei - MG.

Na Tabela 5.14, são apresentados os resultados consolidados das métricas de erro: R2, RMSE, MAE e MAPE, obtidos a partir da média de 10 execuções independentes. Estes valores, que são exclusivamente do conjunto de teste, oferecem uma avaliação minuciosa do desempenho dos modelos de previsão. Eles são comparados em dois contextos distintos: primeiro, quando treinados com séries temporais que foram submetidas a diferentes técnicas de imputação de dados, e segundo, quando treinados com séries temporais que permaneceram sem qualquer tratamento de dados faltantes

Resultado das previsões realizadas com os dados da cidade de São João del Rei - MG				
Técnica de tratamento de dados faltantes	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Sem tratamento de dados faltante	0.389	4453.55	3409.38	0.265
DMD	0.396	4393.25	3363.63	0.260
Interpolação Linear	0.396	4393.84	3311.65	0.262
Random Forest	**0.406**	**4370.71**	**3270.42**	**0.260**

Tabela 5.14: Resultado das previsões realizadas com os dados da cidade de São João del Rei - MG.

A análise detalhada dos resultados indica que, no caso específico de São João del Rei - MG, o modelo com melhor performance foi aquele treinado utilizando a série temporal tratada pela técnica de *Random Forest*. Notavelmente, o *Dynamic Mode Decomposition* (DMD) ocupou o segundo lugar, exibindo também um bom desempenho, mas com uma diferença marginal em relação ao terceiro colocado, que foi a Interpolação Linear. Entretanto, é crucial destacar que o modelo testado sem qualquer tratamento de dados faltantes apresentou o pior resultado de todos.

A Figura 5.9 detalha a previsão com o modelo treinado com os dados completados pela técnica de imputação *Random Forest*.

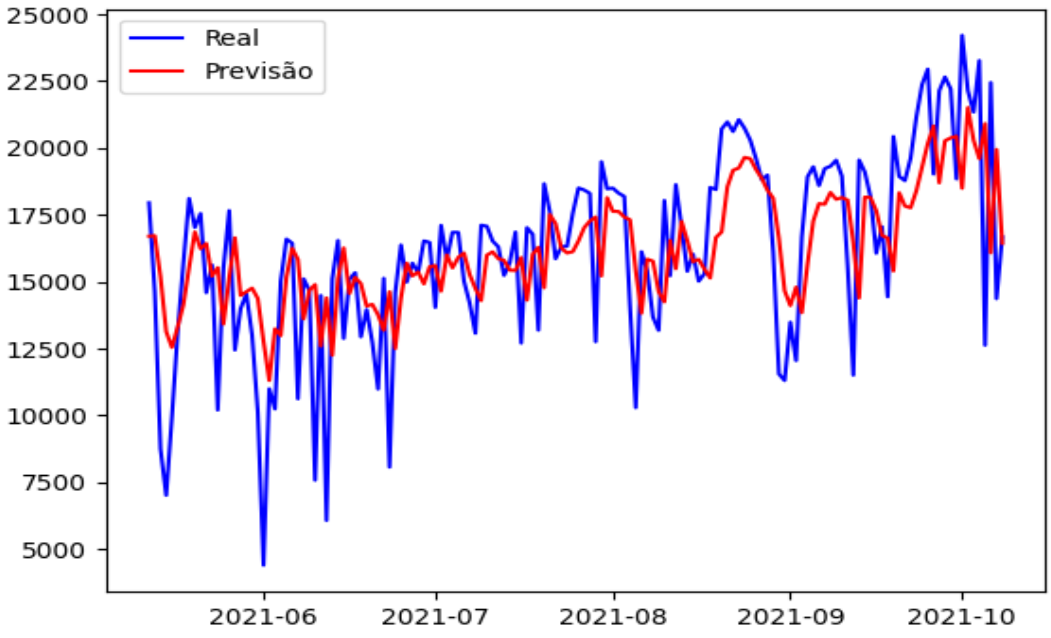

Figura 5.9: Previsão, com dados faltantes preenchidos pelo método *Random Forest*, realizada para cidade de São João del Rei no período de Maio de 2021 à Outubro de 2021.

A Tabela 5.15 apresenta os melhores de previsão para cada cidade selecionada.

Cidade	**R2**	**RMSE (KJ/m²)**	**MAE (KJ/m²)**	**MAPE (%)**
Barbacena	0.554	4180.98	3179.73	0.301
Viçosa	0.531	4416.05	3422.12	0.259
Muriaé	0.483	4091.90	3089.97	0.295
Juiz de Fora	0.390	4293.79	3401.26	0.310
São João del Rei	0.406	4370.71	3270.42	0.260

Tabela 5.15: Os melhores resultados das previsões realizadas para as cidades selecionadas.

6 Conclusão e Trabalhos Futuros

6.1 Conclusão

Este trabalho teve como objetivo principal desenvolver modelos para a previsão da Radiação Solar na Zona da Mata Mineira. Ao longo deste estudo destacamos aproblemática dos dados faltantes nas séries temporais meteorológicas selecionadas. Esta questão emergiu como um elemento central, conduzindo o foco do estudo para o tratamento eficaz destes dados.

A imputação de dados faltantes revelou-se um tópico de estudo independente devido à sua relevância no treinamento efetivo do modelo preditivo. A literatura existente ainda não estabelece um consenso sobre a técnica mais apropriada para esta imputação, refletindo a complexidade do problema, que abrange tanto a natureza intrínseca dos dados quanto os objetivos específicos da imputação.

Nossas análises indicaram que não há uma única abordagem universalmente superior para todas as situações. A eficiência de um método de imputação está intrinsicamente ligada ao contexto específico e às características dos dados faltantes. Identificou-se, por exemplo, que a Interpolação Linear é particularmente eficiente para lacunas menores. No entanto, sua eficácia na imputação não se traduz necessariamente em previsões mais precisas, uma teoria para isso é que o processo de imputação gera ruído que impactam negativamente no treinamento do modelo de previsão. Isso ilustra a complexidade envolvida na escolha do método de imputação mais adequado.

Foi observado que a escolha do método de imputação tem impacto na precisão dos modelos preditivos. Técnicas como Random Forest e DMD, apesar de não apresentarem as melhores métricas de erro de imputação, forneceram dados que contribuíram para previsões mais acuradas. Isso sugere que a capacidade de um método de preservar a integridade dos padrões dos dados é tão vital quanto a precisão da imputação em si.

O foco principal do estudo foi o desenvolvimento de um modelo de *Deep Learning* para a previsão de séries temporais de radiação solar para cada cidade analisada, utilizando

a arquitetura de Redes Neurais Recorrentes, especificamente o modelo *Long Short Term Memory* (LSTM).

Os resultados obtidos demonstraram melhorias nas previsões quando os dados de treinamento eram submetidos ao processo de tratamento de dados faltantes. Isso reforça a ideia de que o tratamento de dados é uma etapa essencial na construção de modelos preditivos, especialmente no contexto da previsão de radiação solar.

Concluímos, portanto, que modelos de previsão baseados em redes LSTM, quando combinados com técnicas adequadas de imputação de dados faltantes, mostram uma melhoria substancial em comparação aos modelos treinados com séries temporais que apresentam lacunas de dados.

6.2 Trabalhos Futuros

Com base nos resultados e conclusões obtidos neste estudo, diversas direções promissoras emergem para pesquisas futuras.

Para ampliar a confiança dos resultados obtidos, sugere-se a realização de simulações adicionais do primeiro estudo de caso, empregando uma amostra maior de séries temporais em cenários similares. Essa expansão do estudo ajudaria a verificar a consistência dos resultados encontrados neste trabalho e garantiria que o comportamento observado se mantém em um espectro mais amplo de dados. Essa abordagem asseguraria a generalização dos achados e a validação da eficácia dos modelos e técnicas empregados.

Para fortalecer a pesquisa, podemos amplicar nosso referencial teórico sobre o tratamento de dados faltantes e comparar os resultados obtidos neste estudo com os de literaturas recentes. Isso não só validaria as descobertas atuais em um contexto mais amplo, mas também identificaria oportunidades de melhoria e inovação, contribuindo para o avanço no campo da imputação de dados em séries temporais.

Além disso, é recomendo avaliar a possibilidade de tratar os dados antes de redimensioná-los de horários para diários. Essa consideração é crucial, pois pode haver ocorrências de dados faltantes em intervalos horários ao longo de um dia inteiro, o que poderia afetar significativamente a precisão dos dados quando consolidados em uma escala diária. Ao identificar e tratar essas lacunas nos dados horários antecipadamente, é

possível melhorar a integridade e a confiabilidade da série temporal.

Outra recomendação valiosa para este estudo é a inclusão do cálculo do desvio médio nos resultados dos erros médios das previsões. Essa métrica proporcionaria uma compreensão mais aprofundada da variabilidade dos erros de previsão, complementando as métricas de erro médio já utilizadas.

Ademais, uma área promissora é a aplicação de Redes Geradoras Adversárias (GANs) para a imputação em séries temporais. A utilização de GANs pode oferecer uma abordagem inovadora para o preenchimento de dados faltantes em séries temporais, superando os métodos tradicionais em termos de precisão e eficácia. Esta técnica, ao simular a distribuição estatística dos dados existentes, pode gerar imputações que preservam as características intrínsecas das séries temporais, tais como tendências e padrões sazonais (GOODFELLOW et al., 2014).

Outro avanço promissor é a exploração de redes do tipo Transformer para a criação de modelos de previsão. Em particular, a técnica *Time Fusion Transformer* apresenta-se como um candidato promissor. Esta abordagem combina a eficiência dos Transformers com um mecanismo de fusão temporal, o que potencialmente pode melhorar a precisão na previsão de séries temporais (LIM et al., 2021). A capacidade dos Transformers de processar sequências de dados de maneira eficiente e sua habilidade de capturar dependências de longo alcance tornam-nos adequados para a análise complexa de séries temporais, como as envolvidas na previsão de radiação solar.

Bibliografia

ADHIKARI, R.; AGRAWAL, R. K. *An Introductory Study on Time Series Modeling and Forecasting*. 2013.

Agência Nacional de Energia Elétrica (ANEEL). *ANEEL sinaliza novo recorde para expansão da geração em 2023*. 2023. ⟨https://www.gov.br/aneel/pt-br/assuntos/noticias/2023/aneel-sinaliza-novo-recorde-para-expansao-da-geracao-em-2023⟩. Expansão, se alcançada, será a maior já verificada no Brasil desde a fundação da ANEEL em 1997.

ALTMAN, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, [American Statistical Association, Taylor Francis, Ltd.], v. 46, n. 3, p. 175–185, 1992. ISSN 00031305. Disponível em: ⟨http://www.jstor.org/stable/2685209⟩.

ATHEY, S.; TIBSHIRANI, J.; WAGER, S. Generalized random forests. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 47, n. 2, p. 1148 – 1178, 2019. Disponível em: ⟨https://doi.org/10.1214/18-AOS1709⟩.

BABATUNDE, O.; MUNDA, J.; HAMAM, Y.; MONYEI, C. A critical overview of the (im)practicability of solar radiation forecasting models. *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, v. 5, p. 100213, 2023. ISSN 2772-6711. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S2772671123001080⟩.

Banco Mundial. *Brasil: Desafios e Oportunidades para o Desenvolvimento Climático*. 2023. ⟨https://www.worldbank.org/pt/country/brazil/brief/brasil-ccdr⟩. Acesso em [data de acesso].

BASÍLIO, S. d. C. A.; SAPORETTI, C. M.; GOLIATT, L. An interdependent evolutionary machine learning model applied to global horizontal irradiance modeling. *Neural Computing and Applications*, Springer, v. 35, n. 16, p. 12099–12120, 06 2023. ISSN 1433-3058. Disponível em: ⟨https://doi.org/10.1007/s00521-023-08342-1⟩.

BASíLIO, S. da C. A.; PUTTI, F. F.; CUNHA, A. C.; GOLIATT, L. An evolutionary-assisted machine learning model for global solar radiation prediction in minas gerais region, southeastern brazil. *Earth Science Informatics*, v. 16, n. 3, p. 2049–2067, 9 2023. ISSN 1865-0481. Disponível em: ⟨https://doi.org/10.1007/s12145-023-00990-0⟩.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, v. 5, n. 2, p. 157–166, 1994.

BERNDT, D. J.; CLIFFORD, J. Using dynamic time warping to find patterns in time series. In: *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*. [S.l.]: AAAI Press, 1994. (AAAIWS'94), p. 359–370.

BISHOP, C. M. [S.l.]: Springer, 2006. ISBN 978-0-387-31073-2.

BOTTOU, L. Large-scale machine learning with stochastic gradient descent. In: LECHEVALLIER; YVES; SAPORTA, G. (Ed.). *Proceedings of COMPSTAT'2010*. [S.l.]: Physica-Verlag HD, 2010. p. 177–186. ISBN 978-3-7908-2604-3.

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001. ISSN 1573-0565. Disponível em: ⟨https://doi.org/10.1023/A:1010933404324⟩.

BREIMAN, L. Statistical Modeling: The Two Cultures. *Statistical Science*, Institute of Mathematical Statistics, v. 16, n. 3, p. 199 – 231, 2001. Disponível em: ⟨https://doi.org/10.1214/ss/1009213726⟩.

BURDEN, R. L.; FAIRES, J. D. *Numerical Analysis*. 9. ed. Boston: Brooks/Cole, Cengage Learning, 2010.

CAMPOS, L. *"Modelo Estocástico Periódico baseado em Redes Neurais", Tese (Doutorado em Engenharia Elétrica) - PUC-Rio, Departamento de Engenharia Elétrica, Rio de Janeiro, 2010.* 2010.

CANNIZZARO, D.; ALIBERTI, A.; BOTTACCIOLI, L.; MACII, E.; ACQUAVIVA, A.; PATTI, E. Solar radiation forecasting based on convolutional neural network and ensemble learning. *Expert Systems with Applications*, v. 181, p. 115167, 2021. ISSN 0957-4174. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0957417421006060⟩.

CHAI, T.; DRAXLER, R. R. Root mean square error (rmse) or mean absolute error (mae)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, Copernicus GmbH, v. 7, n. 3, p. 1247–1250, 2014.

CHATFIELD, C. *Time-series forecasting*. [S.l.]: CRC press, 2000. ISBN 1584880635.

CHATFIELD, C. *The Analysis of Time Series: An Introduction, Sixth Edition*. 6. ed. [S.l.]: Chapman and Hall/CRC, 2003.

CHO, K.; MERRIENBOER, B. van; GULCEHRE, C.; BAHDANAU, D.; BOUGARES, F.; SCHWENK, H.; BENGIO, Y. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014.

CLEVELAND, R. B.; CLEVELAND, W. S.; MCRAE, J. E.; TERPENNING, I. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, v. 6, n. 1, p. 3–73, 1990.

CUNHA, A. C.; FILHO, L. R. A. G.; TANAKA, A. A.; PUTTI, F. F. Performance and estimation of solar radiation models in state of minas gerais, brazil. *Modeling Earth Systems and Environment*, v. 7, n. 1, p. 603–622, 2021. ISSN 2363-6211.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. [S.l.]: John Wiley & Sons, 1998. v. 326.

ELMAN, J. L. Finding structure in time. *Cognitive Science*, v. 14, n. 2, p. 179–211, 1990. ISSN 0364-0213. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/036402139090002E⟩.

FAWAZ, H. I.; FORESTIER, G.; WEBER, J.; IDOUMGHAR, L.; MULLER, P.-A. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery*, v. 33, n. 4, p. 917–963, 2019. ISSN 1573-756X.

FILDES, R.; MAKRIDAKIS, S. The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review / Revue Internationale de Statistique*, [Wiley, International Statistical Institute (ISI)], v. 63, n. 3, p. 289–308, 1995. ISSN 03067734, 17515823. Disponível em: ⟨http://www.jstor.org/stable/1403481⟩.

FISCHER, T.; KRAUSS, C.; TREICHEL, A. *Machine learning for time series forecasting - a simulation study*. [S.l.], 2018.

GARCÍA, S.; LUENGO, J.; HERRERA, F. *Data Preprocessing in Data Mining*. [S.l.]: Springer International Publishing, 2014. (Intelligent Systems Reference Library). ISBN 9783319102474.

GAUTSCHI, W. *Numerical Analysis*. Birkhäuser Boston, 2011. (SpringerLink : Bücher). ISBN 9780817682590. Disponível em: ⟨https://books.google.com.br/books?id=-fgjJF9yAIwC⟩.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

GOODFELLOW, I. J.; POUGET-ABADIE, J.; MIRZA, M.; XU, B.; WARDE-FARLEY, D.; OZAIR, S.; COURVILLE, A.; BENGIO, Y. *Generative Adversarial Networks*. 2014.

GRAVES, A. *Supervised Sequence Labelling with Recurrent Neural Networks*. [S.l.: s.n.], 2012. v. 385. ISBN 978-3-642-24796-5.

H, I.; FRANK, E.; HALL, M. A.; PAL, C. J. *Data Mining: Practical machine learning tools and techniques*. [S.l.]: Morgan Kaufmann, 2016.

HAWKINS, D. M. The problem of overfitting. *Journal of chemical information and computer sciences*, ACS Publications, v. 44, n. 1, p. 1–12, 2004.

HAYKIN, S. *Neural Networks: A Comprehensive Foundation*. [S.l.]: Macmillan, 1994. ISBN 9780132265560.

HAYKIN, S. *Neural Networks and Learning Machines*. [S.l.]: Prentice Hall, 2009. (Neural networks and learning machines, v. 10). ISBN 9780131471399.

HE, K.; ZHANG, X.; REN, S.; SUN, J. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015.

HINTON, G. E.; SALAKHUTDINOV, R. R. Reducing the dimensionality of data with neural networks. *Science*, v. 313, p. 504–507, 2006.

HOCHREITER, S.; SCHMIDHUBER, J. Long Short-Term Memory. *Neural Computation*, v. 9, n. 8, p. 1735–1780, 11 1997. ISSN 0899-7667. Disponível em: ⟨https://doi.org/10.1162/neco.1997.9.8.1735⟩.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, v. 22, n. 4, p. 679–688, 2006. ISSN 0169-2070. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0169207006000239⟩.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, v. 22, n. 4, p. 679–688, 2006. ISSN 0169-2070. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0169207006000239⟩.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: . San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995. (IJCAI'95), p. 1137–1143. ISBN 1558603638.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 60, n. 6, p. 84–90, may 2017. ISSN 0001-0782.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 2015.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, IEEE, v. 86, n. 11, p. 2278–2324, 1998.

LECUN, Y. A.; BOTTOU, L.; ORR, G. B.; MÜLLER KLAUS-ROBERT", e. G.; ORR, G. B.; MÜLLER, K.-R. Efficient backprop. In: _____. *Neural Networks: Tricks of the Trade: Second Edition*. [S.l.]: Springer, 2012. p. 9–48. ISBN 978-3-642-35289-8.

Legates, D. R.; McCabe, G. J. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water Resources Research*, v. 35, n. 1, p. 233–241, jan. 1999.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. *R News*, v. 2, n. 3, p. 18–22, 2002. Disponível em: ⟨http://CRAN.R-project.org/doc/Rnews/⟩.

LIEW, A. W.-C.; LAW, N.-F.; YAN, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings in Bioinformatics*, v. 12, n. 5, p. 498–513, 12 2010. ISSN 1467-5463. Disponível em: ⟨https://doi.org/10.1093/bib/bbq080⟩.

LIM, B.; ARıK, S. ; LOEFF, N.; PFISTER, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, v. 37, n. 4, p. 1748–1764, 2021. ISSN 0169-2070. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0169207021000637⟩.

MAAS, A. L.; HANNUN, A. Y.; NG, A. Y. Rectifier nonlinearities improve neural network acoustic models. In: *Proceedings of the 30th International Conference on Machine Learning*. [s.n.], 2013. v. 28, n. 3. Disponível em: ⟨https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf⟩.

MAKRIDAKIS, S. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, v. 9, n. 4, p. 527–529, 1993. ISSN 0169-2070. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/0169207093900793⟩.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. springer, v. 5, p. 115–133, 1943.

MENDES, B. Energia solar no brasil: conheça os estados que produzem mais. *Solar Volt*, 2022. Disponível em: ⟨https://www.solarvoltenergia.com.br/blog/energia-solar-no-brasil-maiores-produtores⟩.

MEZIĆ, I. Analysis of fluid flows via spectral properties of the koopman operator. v. 45, p. 357–378, 2013.

MONTGOMERY, D.; PECK, E.; VINING, G. *Introduction to Linear Regression Analysis*. Wiley, 2015. (Wiley Series in Probability and Statistics). ISBN 9781119180173. Disponível em: ⟨https://books.google.com.br/books?id=27kOCgAAQBAJ⟩.

NARVAEZ, G.; GIRALDO, L. F.; BRESSAN, M.; PANTOJA, A. Machine learning for site-adaptation and solar radiation forecasting. *Renewable Energy*, v. 167, p. 333–342, 2021. ISSN 0960-1481. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0960148120318395⟩.

Nações Unidas no Brasil. *Objetivos de Desenvolvimento Sustentável*. 2022. ⟨https://brasil.un.org/pt-br/sdgs⟩. Último acesso em 08 de Novembro de 2022.

OLIVEIRA, B. Características das séries temporais. 2019. Disponível em: ⟨https://statplace.com.br/blog/caracteristicas-das-series-temporais⟩.

PARRA-PLAZAS, J.; GAONA-GARCIA, P.; PLAZAS-NOSSA, L. Time series outlier removal and imputing methods based on colombian weather stations data. *Environmental Science and Pollution Research*, v. 30, n. 28, p. 72319–72335, 6 2023. ISSN 1614-7499. Disponível em: ⟨https://doi.org/10.1007/s11356-023-27176-x⟩.

PENG, T.; ZHANG, C.; ZHOU, J.; NAZIR, M. S. An integrated framework of bi-directional long-short term memory (bilstm) based on sine cosine algorithm for hourly solar radiation forecasting. *Energy*, v. 221, p. 119887, 2021. ISSN 0360-5442. Disponível em: ⟨https://www.sciencedirect.com/science/article/pii/S0360544221001365⟩.

RATANAMAHATANA, C. A.; KEOGH, E. Making time-series classification more accurate using learned constraints. In: _____. *Proceedings of the 2004 SIAM International Conference on Data Mining (SDM)*. [s.n.]. p. 11–22. Disponível em: ⟨https://epubs.siam.org/doi/abs/10.1137/1.9781611972740.2⟩.

ROWLEY, C. W.; MEZIć, I.; BAGHERI, S.; SCHLATTER, P.; HENNINGSON, D. S. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, Cambridge University Press, v. 641, p. 115–127, 2009.

SCHMID, P. J. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, Cambridge University Press, v. 656, p. 5–28, 2010.

SILVA, L. P. V. da. Desenvolvimento de um modelo para a estimação da carga de radiação solar com base em variáveis climáticas. *Programa de pós-graduação em modelagem computacional, Universidade Federal de Juiz de Fora*, 2021.

STEKHOVEN, D. J.; BüHLMANN, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, v. 28, n. 1, p. 112–118, 10 2011. ISSN 1367-4803. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/btr597⟩.

SUTTON, R.; BARTO, A. *Reinforcement Learning, second edition: An Introduction*. [S.l.]: MIT Press, 2018. (Adaptive Computation and Machine Learning series). ISBN 9780262039246.

TAIRA, K.; BRUNTON, S. L.; DAWSON, S. T.; ROWLEY, C. W.; COLONIUS, T.; MCKEON, B. J.; SCHMIDT, O. T.; GORDEYEV, S. K.; THEOFILIS, V.; UKEILEY, L. S. Modal analysis of fluid flows: An overview. American Institute of Aeronautics and Astronautics, v. 55(12), p. 4013–4041, 2017.

TANG, F.; ISHWARAN, H. Random forest missing data algorithms. *Statistical Analysis and Data Mining*, v. 10, n. 6, p. 363–377, 2017.

TROYANSKAYA, O.; CANTOR, M.; SHERLOCK, G.; BROWN, P.; HASTIE, T.; TIBSHIRANI, R.; BOTSTEIN, D.; ALTMAN, R. B. Missing value estimation methods for DNA microarrays . *Bioinformatics*, v. 17, n. 6, p. 520–525, 06 2001. ISSN 1367-4803. Disponível em: ⟨https://doi.org/10.1093/bioinformatics/17.6.520⟩.

TU, J. H.; ROWLEY, C. W.; LUCHTENBURG, S. L. B. D. M.; ; KUTZ, J. N. On dynamic mode decomposition: Theory and applications. American Institute of Mathematical Sciences, v. 1(2), p. 391–421, 2014.

TYRALIS, H.; PAPACHARALAMPOUS, G. Variable selection in time series forecasting using random forests. *Algorithms*, v. 10, n. 4, 2017. ISSN 1999-4893. Disponível em: ⟨https://www.mdpi.com/1999-4893/10/4/114⟩.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, L.; POLOSUKHIN, I. *Attention Is All You Need*. 2017.

WATANABE, S. *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance*. 2023.

WEI, W. W. 458Time Series Analysis. In: *The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2: Statistical Analysis*. [S.l.]: Oxford University Press, 2013. ISBN 9780199934898.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, Inter-Research Science Center, v. 30, n. 1, p. 79–82, 2005. ISSN 0936577X, 16161572. Disponível em: ⟨http://www.jstor.org/stable/24869236⟩.