PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA E DE ARTES

Norton Pereira Ricardo

INOVAÇÃO EM COMPLETAÇÃO DE DADOS CLIMÁTICOS: MÉTODOS BASEADOS EM VIZINHOS, REGRESSÃO LINEAR E REDES NEURAIS DE CAMADAS DENSAS

GOIÂNIA
2024

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA E DE ARTES

Norton Pereira Ricardo

INOVAÇÃO EM COMPLETAÇÃO DE DADOS CLIMÁTICOS: MÉTODOS BASEADOS EM VIZINHOS, REGRESSÃO LINEAR E REDES NEURAIS DE CAMADAS DENSAS

Trabalho de conclusão de Curso apresentado à Escola Politécnica e de Artes, da Pontifícia Universidade Católica de Goiás.
Orientador(a): Dra. Maria José Pereira Dantas
Coorientador(a): Me. Leonardo Merelles.

GOIÂNIA
2024

Título: INOVAÇÃO EM COMPLETAÇÃO DE DADOS CLIMÁTICOS: MÉTODOS BASEADOS EM VIZINHOS, REGRESSÃO LINEAR E REDES NEURAIS DE CAMADAS DENSAS

RESUMO

Esta pesquisa teve como objetivo desenvolver e avaliar métodos de completação de dados em bases climáticas. Após uma revisão de literatura na base de dados *Web of Science* com o uso das palavras-chave "*missing data*", "*data climate*", "*climate*" e "*imputation*" e equivalentes, identificou-se uma tendência no uso de métodos de machine learning (ML) para a obtenção de resultados mais acurados. Pesquisas se seguiram para um levantamento de artigos científicos que usam ML para o entendimento dos métodos e definições das escolhas dos métodos. Esta pesquisa teve como ponto de partida uma API que coleta dados do INMET em tempo real e já disponibiliza ao usuário escolhas de estações e períodos de dados para a obtenção de gráficos de linhas e *boxplots* de variáveis climáticas da base. A referida API foi iniciada em pesquisa anterior e finalizada no início da pesquisa em questão para receber os métodos de completação. Foram avaliadas abordagens como o Método do Vizinho Mais Próximo (*Nearest Neighbor Method*), Regressão Linear e Redes Neurais de Camadas Densas. A metodologia incluiu o pré-processamento dos dados, análise estatística exploratória, tratamento de *outliers*, normalização *z-score* e separação dos dados. Os resultados obtidos foram analisados e comparados usando a métrica RMSE (*Root Mean Squared Error*). Os resultados mostraram que a técnica baseada em aprendizado de máquina apresentou maior precisão na imputação dos dados faltantes em comparação com métodos tradicionais. As redes neurais de camadas densas apresentaram um desempenho superior na captura de padrões não lineares das variáveis climáticas. A pesquisa sugere que a integração desses métodos na API existente pode melhorar significativamente a qualidade dos dados climáticos fornecidos, beneficiando uma ampla gama de aplicações.

Palavras-chave: API, imputação, *machine learning*, dados faltantes, bases climáticas.

ABSTRACT

This research aimed to develop and evaluate data completion methods for climatic databases. Following a literature review in the Web of Science database using the keywords "missing data," "data climate," "climate," and "imputation," a trend was identified in the use of machine learning (ML) methods for achieving more accurate results. Further research was conducted to survey scientific articles that use ML to understand the methods and define the research choices. The research started with an API that collects real-time data from INMET and already provides users with choices of stations and data periods to generate line graphs and box plots of climatic variables. This API was initiated in a previous study and finalized at the beginning of the current research to incorporate the completion methods. Approaches such as the Nearest Neighbor Method, Linear Regression, and Dense Layer Neural Networks were evaluated. The methodology included data preprocessing, exploratory statistical analysis, outlier treatment, normalization, and data separation. The results obtained were analyzed and compared using the RMSE (Root Mean Squared Error) metric. The findings showed that the ML-based technique achieved greater accuracy in imputing missing data compared to traditional methods. Dense layer neural networks demonstrated superior performance in capturing nonlinear patterns of climatic variables. The research suggests that integrating these methods into the existing API can significantly improve the quality of the climatic data provided, benefiting a wide range of applications.

Keywords*: API, imputation, machine learning, missing data, climate databases.*

LISTA DE FIGURAS

LISTA DE TABELAS

SUMARIO

1. INTRODUÇÃO

O Brasil apresenta uma grande diversidade de fatores climáticos, com diferentes biomas, cada um com características climáticas únicas, e é afetado pelas mudanças climáticas globais. De acordo com o relatório intergovernamental de 2018, sobre mudanças climáticas, o mundo enfrentará consequências catastróficas a menos que as emissões globais de gases de efeito estufa (GEE) sejam eliminadas em 30 anos. Diante desse cenário, o estudo climático e a capacidade de prever as condições climáticas estão se tornando cada vez mais indispensáveis para a tomada de decisão em diversas áreas do país, incluindo agricultura, produção de energia, atividades industriais e prevenção de evacuações em áreas de risco. Além disso, é fundamental para o desenvolvimento sustentável do país (IPCC, 2018).

A disponibilidade de dados climáticos precisos é fundamental para uma variedade de aplicações, incluindo previsões meteorológicas, estudos de mudanças climáticas e análises ambientais. No entanto, muitas vezes, as bases de dados climáticas contêm lacunas de dados devido a falhas de equipamentos, erros de medição ou outras razões (AVILA-DIAZ, 2020). E neste cenário, o completamento de dados faltantes é um desafio para que as bases de dados possam ser utilizadas de forma a subsidiar as decisões e a comunidade científica tem tratado o problema, e várias discussões de soluções têm sido tratadas em (AFRIFA-YAMOAH et al., 2020); (MOHAMAD et al., 2021); (TEEGAVARAPU, 2020).

O completamento de dados em variáveis climáticas é uma tarefa desafiadora devido à natureza complexa e variável dos fenômenos climáticos. Abordagens tradicionais incluem métodos de interpolação espacial e temporal, como a interpolação de *Kriging*, que utiliza a autocorrelação espacial das variáveis climáticas para preencher os valores ausentes. No entanto, esses métodos podem ser limitados em sua capacidade de capturar padrões climáticos não lineares e podem não ser adequados para todas as variáveis climáticas. (Lopez et al.); (Mohamad, 2021); (Sattari, 2017).

Um fator que vem dificultando o avanço desses estudos e desenvolvimentos tem sido a ausência de dados climáticos sólidos em diversas regiões do Brasil. Além da dificuldade em encontrar e acessar esses dados, existem problemas com os formatos de dados, variáveis, escala de coleta de dados, *outliers* e dados faltantes, especialmente, em dados mais antigos, o que torna a construção e análise dos dados um processo complexo (ESSOU, 2016).

Existem algumas instituições, nacionais ou internacionais, que fornecem dados climáticos do Brasil, através de estações meteorológicas, satélites, balões meteorológicos e boias. As principais fontes de dados climáticos são o Instituto Nacional de Meteorologia (INMET), o Centro de Previsão de Tempo e Estudos Climáticos (CPTEC) e a *National Oceanic Atmospheric Administration* (NOAA). No entanto, cada instituição trabalha com variáveis distintas, períodos de coleta de dados e unidades de medida diferentes, além de haver ausência de manual, restrições de acesso e dados ausentes ou *outliers*. Todos esses fatores dificultam a obtenção de dados com a qualidade necessária (MACHADO et al., 2019).

Nos últimos anos, o uso de técnicas de aprendizado de máquina para o completamento de dados climáticos tem se destacado como uma abordagem promissora. Algoritmos de aprendizado de máquina, como regressão, redes neurais e modelos de séries temporais, têm a capacidade de aprender relações complexas entre as variáveis climáticas e, assim, prever valores ausentes de forma mais precisa (BASAKIN, 2023). Alguns estudos têm explorado o uso de redes neurais convolucionais (CNNs) para o completamento de imagens de radar meteorológico e séries temporais climáticas. Esses modelos demonstraram resultados promissores na previsão de valores ausentes.

O objetivo desta pesquisa é desenvolver e avaliar métodos eficazes para completar dados perdidos em bases climáticas, a fim de melhorar a qualidade e a utilidade dessas informações. Com a inclusão de técnicas de aprendizado de máquina, esta pesquisa visa explorar abordagens avançadas para o completamento de dados climáticos, visando uma melhor precisão e eficácia na previsão e análise climática. Entre os métodos disponíveis, existe uma tendência para ferramentas *Machine Learning* (ML), que faz parte da inteligência artificial. A questão de pesquisa a ser respondida é: “Como desenvolver algoritmos de ML para completamento de dados perdidos que sejam adequados para diferentes tipos de variáveis climáticas (por exemplo, temperatura, umidade, precipitação, pressão atmosférica) e grau de comprometimento da base, ao mesmo tempo, como avaliar a acurácia desses métodos em termos de precisão e robustez?”

Esta questão de pesquisa aborda a necessidade de desenvolver métodos eficazes de completamento de dados em bases climáticas, levando em consideração a variedade de variáveis climáticas e suas características únicas. Além disso, a questão busca compreender como avaliar de forma abrangente a acurácia desses métodos,

garantindo que eles produzam resultados confiáveis para aplicações meteorológicas e climatológicas. Para responder esta questão de pesquisa definiu-se os seguintes objetivos:

- Evoluir a API já desenvolvida agregando o tratamento de *outliers* e análise dos dados faltantes quanto às suas características;
- Desenvolver algoritmo de ML para completamento de dados perdidos que seja adequado para diferentes tipos de variáveis climáticas e grau de comprometimento dos dados;
- Avaliar o desempenho dos métodos de completamento de dados em termos de precisão e robustez.
- Investigar o impacto do completamento de dados perdidos na precisão das previsões meteorológicas e em análises climáticas.
- Comparar diferentes abordagens de completamento de dados com a abordagem de aprendizado de máquina.

1.1 JUSTIFICATIVA

Para identificar tendências climáticas ao longo do tempo, é necessário ter séries temporais contínuas e confiáveis de dados. A falta de dados pode levar a lacunas nas análises e dificultar a detecção de mudanças climáticas significativas. O completamento de dados é fundamental para avaliar eventos climáticos extremos, como secas, inundações e ondas de calor. A falta de dados pode prejudicar a capacidade de compreender a frequência e a intensidade desses eventos.

Os modelos climáticos são usados para projetar cenários futuros de mudanças climáticas. Para validar esses modelos, é necessário compará-los com dados observados. O completamento de dados ajuda a melhorar a qualidade das comparações e aprimorar a confiabilidade das projeções climáticas.

O completamento de dados é essencial para apoiar a tomada de decisão em relação à adaptação às mudanças climáticas. Isso inclui planejamento urbano, gestão de recursos hídricos, agricultura e políticas de mitigação de riscos. A qualidade e disponibilidade de dados climáticos são cruciais para entender os padrões climáticos, prever eventos meteorológicos e analisar tendências climáticas. No entanto, a coleta de dados climáticos é suscetível a lacunas devido a falhas em instrumentos de

medição, problemas de transmissão de dados e falta de estações de monitoramento em algumas regiões.

1.2 CONTRIBUIÇÃO DA PESQUISA

Pretende-se incorporar métodos de completamento de dados como resultado da pesquisa, que já possui uma API como ponto de partida, que atualiza dados em tempo real do Instituto Nacional de Meteorologia (INMET), além de análises dos dados com gráficos e tabelas detalhadas.

O Completamento de dados faltantes pode melhorar a qualidade dos dados meteorológicos, tornando os gráficos e análises mais confiáveis. Isso é especialmente importante em aplicações meteorológicas onde a precisão dos dados é crucial. Com dados mais completos, você pode realizar análises mais abrangentes e extrair informações mais significativas sobre o comportamento das variáveis climáticas. Isso pode levar a *insights* mais profundos e úteis para os usuários da API. Dados climáticos completos e de alta qualidade pode ser valiosos para diversas aplicações, como previsão do tempo, agricultura, gerenciamento de recursos hídricos, entre outras. A imputação de dados faltantes pode tornar a API mais útil para um público mais amplo.

A implementação de métodos de completamento de dados é uma oportunidade de aprendizado e aplicação de conhecimentos em métodos de processamento de dados. Isso pode enriquecer sua experiência acadêmica e profissional. Se a API é acessada por outros pesquisadores, profissionais ou entusiastas do clima, a incorporação de métodos de completamento de dados pode ser uma contribuição valiosa para a comunidade, tornando os dados mais úteis e acessíveis. A documentação da API mostrará claramente as técnicas utilizadas, comunicando qualquer incerteza associada à imputação de dados e fornecer aos usuários informações transparentes sobre como os dados foram tratados. Isso é fundamental para garantir a confiabilidade e a transparência da API e das análises resultantes.

Este trabalho tem o potencial de contribuir significativamente para a ciência climática e o meio acadêmico, fornecendo *insights* valiosos para pesquisadores, tomadores de decisão e a sociedade como um todo. A conclusão bem-sucedida desta pesquisa não só avançará o conhecimento acadêmico, mas também enriquecerá a ferramenta desenvolvida anteriormente, que já automatiza a coleta de dados climáticos de hora em hora da plataforma do INMET. Com a integração dos métodos de imputação de dados propostos, esta ferramenta será capaz de oferecer uma série

histórica completa e tratada, pronta para ser utilizada em uma variedade de aplicações. Esta melhoria facilitará ações práticas frente aos crescentes impactos das mudanças climáticas e destacará a importância de se completar este estudo com o mais alto nível de rigor e precisão.

2. REFERENCIAL TEÓRICO

2.1 SÉRIES TEMPORAIS

A escolha das fontes de dados climáticos adequadas é fundamental. As Estações terrestres coletam dados diretos, mas podem ser limitadas em áreas remotas. Os Satélites fornecem dados abrangentes, mas podem ter limitações temporais. Compreender as características e limitações dessas fontes é crucial para coletar, preparar as séries temporais.

As séries temporais são conjuntos de dados organizados em uma sequência temporal regular, onde as observações são registradas em intervalos específicos ao longo do tempo. Uma revisão sobre séries temporais pode ser obtida no livro *Análise de Séries Temporais: Modelos Multivariados e Não Lineares* (MORETTIN; TOLOI, 2020).

No contexto da climatologia e meteorologia, as séries temporais desempenham um papel fundamental, pois essas disciplinas estão intrinsecamente ligadas ao estudo das condições atmosféricas e climáticas ao longo do tempo. As variáveis climáticas, como temperatura, umidade, precipitação, pressão atmosférica e ventos, são frequentemente medidas em séries temporais. As séries temporais climáticas apresentam várias características distintas:

• **Sazonalidade**: Muitas variáveis climáticas exibem padrões sazonais, repetindo-se ao longo das estações do ano. Por exemplo, a temperatura média tende a aumentar no verão e diminuir no inverno.

• **Tendências**: Além da sazonalidade, as séries temporais climáticas podem apresentar tendências de longo prazo, como o aumento gradual das temperaturas devido às mudanças climáticas.

• **Ruído**: As séries temporais também são afetadas por ruídos, variações aleatórias que podem dificultar a identificação de padrões significativos.

• **Autocorrelação**: Valores passados de uma variável climática geralmente estão correlacionados com valores futuros, o que torna as séries temporais auto correlacionadas.

• **Valores Ausentes**: Lacunas de dados são comuns devido a falhas na coleta ou transmissão de dados, tornando o completamento de dados essencial.

A análise de séries temporais envolve técnicas estatísticas e matemáticas para extrair informações úteis das séries. Alguns métodos comuns incluem:

• **Decomposição**: Separar uma série temporal em seus componentes de tendência, sazonalidade e ruído para melhor compreensão.

• **Autocorrelação**: Avaliar a correlação entre os valores passados e futuros para identificar padrões temporais.

• **Médias Móveis**: Calcular médias em intervalos específicos para suavizar flutuações e identificar tendências.

2.2 VARIÁVEIS CLIMÁTICAS

Cada variável climática possui características distintas que devem ser consideradas ao desenvolver métodos de completamento de dados. Algumas variáveis são descritas a seguir.

Temperatura: A temperatura é uma variável climática que exibe padrões sazonais e diurnos. Portanto, ao completar dados ausentes de temperatura, é importante considerar variações sazonais, ciclos diários e anomalias de curto prazo, como frentes frias ou quentes.

Umidade: A umidade do ar pode variar significativamente em diferentes regiões geográficas e momentos do dia. O completamento de dados de umidade deve levar em conta essas variações, bem como a influência da temperatura na capacidade do ar de conter umidade.

Precipitação: A precipitação é altamente variável no tempo e no espaço. Ao lidar com dados de precipitação, é necessário considerar a natureza intermitente e pontual das chuvas, bem como as diferenças sazonais na distribuição da precipitação.

Pressão Atmosférica: A pressão atmosférica é uma variável que varia com a altitude e pode ser influenciada por sistemas meteorológicos, como ciclones e anticiclones. O completamento de dados de pressão atmosférica deve levar em conta essas variações verticais e horizontais.

Vento: A velocidade e a direção do vento são influenciadas por fatores locais e regionais, incluindo relevo, proximidade de corpos d'água e padrões climáticos maiores. O completamento de dados de vento deve levar em conta essas variáveis para fornecer uma representação precisa das condições de vento em uma área específica.

Velocidade do Vento: A velocidade do vento é uma variável crítica para diversas aplicações, incluindo energia eólica e previsão de tempo. Deve-se considerar

as flutuações diurnas, sazonais e os padrões de circulação atmosférica ao completar dados de velocidade do vento.

Radiação Solar: A radiação solar é influenciada pela localização geográfica, a estação do ano e a hora do dia. Fatores como cobertura de nuvens e poluição também afetam os níveis de radiação solar. Ao completar dados de radiação solar, é importante considerar essas variações temporais e espaciais.

Os artigos de Machado et al. (2020) e Moura & Fortes (2020) apresentam as variáveis climáticas que serão utilizadas neste trabalho e estão associadas à base de dados do INMET.

2.3 REGRESSÃO LINEAR MÚLTIPLA

A regressão linear múltipla é uma técnica estatística amplamente utilizada para modelar a relação entre uma variável dependente e duas ou mais variáveis independentes. O modelo de regressão linear múltipla assume que existe uma relação linear entre a variável dependente Y e um conjunto de variáveis independentes $X_1, X_2, ..., X_p$. Essa relação é expressa pela equação $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...+ \beta_p Xp + \epsilon$, em que β_0 é o intercepto, β_1, β_2, ..., β_p são os coeficientes angulares associados às variáveis independentes, e ϵ é o termo de erro que captura as influências de outras variáveis não incluídas no modelo. O objetivo da regressão linear múltipla é estimar os coeficientes $\beta_0, \beta_1, ..., \beta_p$ de forma que a soma dos quadrados dos resíduos (as diferenças entre os valores observados e os valores previstos) seja minimizada (Montgomery, Peck & Vining, 2012).

O processo de ajuste do modelo de regressão linear múltipla envolve a utilização do método dos mínimos quadrados ordinários (OLS - *Ordinary Least Squares*). Este método calcula os valores dos coeficientes $\beta_0, \beta_1, ..., \beta_p$ que minimizam a soma dos quadrados dos resíduos. A eficácia do modelo é avaliada através de várias métricas, como o coeficiente de determinação ajustado (R^2 ajustado), que indica a proporção da variabilidade total da variável dependente explicada pelas variáveis independentes, ajustada pelo número de preditores no modelo. Um R^2 ajustado próximo de 1 indica que o modelo explica a maior parte da variabilidade dos dados, enquanto um valor próximo de 0 indica que o modelo tem pouca capacidade explicativa (Montgomery, Peck & Vining, 2012). A regressão linear múltipla é aplicada em uma ampla gama de situações tais como em economia, engenharia, ciências sociais e ciências naturais. Em economia, por exemplo, é utilizada para modelar a

relação entre o consumo e a renda. Na engenharia, pode ser usada para prever a resistência dos materiais com base em suas propriedades físicas. Em ciências sociais, a regressão linear é frequentemente empregada para estudar a relação entre variáveis sociais, como a educação e o rendimento. Aqui, a regressão será utilizada para permitir o completamento dos dados faltantes de uma variável do clima em funções de outras variáveis completas. A simplicidade e a interpretabilidade do modelo tornam a regressão linear uma ferramenta valiosa para analisar dados e fazer previsões (Montgomery, Peck & Vining, 2012).

Apesar de sua utilidade, a regressão linear tem algumas limitações. A principal suposição do modelo é a linearidade, ou seja, que a relação entre as variáveis é linear. No entanto, muitas relações no mundo real são não lineares. Além disso, o modelo assume que os resíduos são homocedásticos (têm variância constante) e que não há multicolinearidade (altas correlações entre variáveis independentes). Quando essas suposições são violadas, os resultados da regressão linear podem ser enganosos. Portanto, é importante diagnosticar e tratar quaisquer violações das suposições antes de confiar nas inferências do modelo (Montgomery, Peck & Vining, 2012).

2.4 INTELIGÊNCIA ARTIFICIAL

A área de inteligência artificial (IA) engloba um conjunto de técnicas e algoritmos que permitem aos sistemas computacionais realizar tarefas que normalmente requerem inteligência humana. A IA abrange desde a capacidade de processamento e interpretação de grandes volumes de dados até o aprendizado automático e a tomada de decisões inteligentes. Essa área de estudo tem sido impulsionada por avanços em áreas como aprendizado de máquina, processamento de linguagem natural, visão computacional e robótica, entre outras. Com a aplicação de técnicas de IA, têm sido alcançados resultados notáveis em diversas áreas, como saúde, finanças, transporte, manufatura e assistência pessoal, trazendo impactos significativos na sociedade (RUSSEL, S.; NORVIG, 2020).

A IA baseia-se em modelos e algoritmos que permitem que os sistemas aprendam a partir de dados, identifiquem padrões, façam previsões e tomem decisões com base em informações disponíveis. O aprendizado de máquina, em particular, tem desempenhado um papel fundamental na IA, com algoritmos capazes de aprender com exemplos e melhorar seu desempenho ao longo do tempo. Além disso, avanços recentes em redes neurais artificiais, como as redes neurais profundas, permite o

processamento de dados não estruturados e a realização de tarefas complexas, como reconhecimento de imagens e processamento de linguagem natural (LECUN, Y.; BENGIO, Y.; HINTON, 2015).

Em Kadow et al. (2020) os autores exploram como um método de IA pode se aplicar na preparação de dados robustos para alimentar métodos de previsão de clima. Os métodos desenvolvidos não exigem hipóteses sobre os dados como ocorre com os métodos estatísticos e têm apresentado resultados promissores.

2.5 MACHINE LEARNING

O aprendizado de máquina (*machine learning*) é uma subárea da inteligência artificial dedicada ao desenvolvimento de algoritmos e modelos que capacitam sistemas computacionais a aprender e melhorar a partir de dados. Esse campo abrange a criação de modelos estatísticos e computacionais que podem identificar padrões nos dados e fazer previsões ou tomar decisões com base nesses padrões. A aplicação do aprendizado de máquina é vasta, englobando áreas como reconhecimento de padrões, processamento de linguagem natural, análise de dados e decisões automatizadas, resultando em avanços significativos que afetam diversos setores (Mitchell, 1997)

Os modelos de aprendizado de máquina são treinados com base em conjuntos de dados relevantes, o que lhes permite adquirir conhecimento e ajustar-se aos padrões encontrados nos dados. As principais técnicas de aprendizado de máquina incluem aprendizado supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, os modelos são treinados com dados rotulados, nos quais os rótulos fornecem a informação correta para cada instância de treinamento. No aprendizado não supervisionado, os modelos buscam identificar estruturas ou padrões nos dados sem a orientação de rótulos prévios. O aprendizado por reforço, por sua vez, envolve a aprendizagem a partir da interação com um ambiente, utilizando tentativa e erro para maximizar uma recompensa (Goodfellow, Bengio & Courville, 2016)

O artigo de Alkabbani (2022) explora a aplicação de métodos de aprendizado de máquina para a imputação de dados. Existem muitas opções disponíveis, e uma forma de potencializar esses métodos é através da hibridização. Em um estudo realizado por Bayma e Pereira (2018), os autores investigaram diversas técnicas de aprendizado de máquina para preencher lacunas em dados climáticos. Entre os métodos analisados estavam redes neurais, árvores de regressão *bagged* e florestas

aleatórias (*random forest*). A pesquisa foi conduzida no estado de Minas Gerais, Brasil, utilizando uma base de dados de 48 estações meteorológicas automáticas, com dados fornecidos pelo Instituto Nacional de Meteorologia (INMET). O estudo focou na imputação de dados faltantes em séries temporais climáticas, como temperatura máxima e mínima e precipitação. A técnica de floresta aleatória foi identificada como a mais eficaz para a imputação de dados climáticos faltantes no estado de Minas Gerais.

2.6 REDES NEURAIS

As redes neurais artificiais representam um componente crucial da inteligência artificial, demonstrando eficácia notável em diversas aplicações, como na climatologia e meteorologia. Dentro desse contexto, as redes neurais são empregadas para modelar, prever e analisar dados climáticos e meteorológicos complexos, aprimorando a compreensão do clima e aumentando a precisão das previsões.

Inspiradas no funcionamento do cérebro humano, as redes neurais são compostas por camadas de neurônios artificiais interconectados. Cada neurônio realiza operações matemáticas básicas, processando informações recebidas de neurônios na camada anterior e transmitindo os resultados para neurônios na camada seguinte.

Uma referência é o livro de “Redes Neurais: Princípios e Prática” de Simon Haykin. 2ª. Ed. 2000.Edição Português. Paulo Martins Engel (Tradutor). Ed. Bookman.

Os principais componentes incluem:

- **Camada de Entrada**: Recebe os dados de entrada, que podem ser variáveis climáticas, como temperatura, umidade, pressão atmosférica etc.
- **Camadas Ocultas**: Neurônios interconectados que realizam transformações matemáticas nos dados de entrada.
- **Camada de Saída**: Produz a saída final, que pode ser uma previsão meteorológica ou uma análise climática.

A força das redes neurais reside na sua capacidade de aprender padrões complexos a partir dos dados, adaptando-se aos padrões presentes nas séries

temporais climáticas, por exemplo. As redes neurais têm várias aplicações no campo da climatologia e meteorologia: Elas capturam padrões complexos, como a influência de fenômenos climáticos, e melhoram a precisão das previsões.

Embora as redes neurais ofereçam muitas vantagens, elas também apresentam desafios:

- **Quantidade de Dados**: Redes neurais geralmente requerem grandes volumes de dados para treinamento eficaz. No entanto, em climatologia e meteorologia, os dados históricos podem ser limitados em algumas regiões.
- **Interpretabilidade**: As redes neurais são frequentemente consideradas como "caixas pretas" devido à complexidade de suas operações. Interpretar como uma rede neural chega a uma previsão pode ser desafiador.
- ***Overfitting***: Redes neurais podem ser suscetíveis ao *overfitting*, onde o modelo se ajusta demasiadamente aos dados de treinamento, prejudicando sua capacidade de generalização.

Em resumo, as redes neurais são ferramentas poderosas na climatologia e meteorologia, permitindo a modelagem e previsão de fenômenos climáticos complexos. No entanto, seu sucesso depende da disponibilidade de dados de qualidade, da capacidade de interpretação dos resultados e do controle adequado do overfitting. Com avanços contínuos em técnicas de aprendizado profundo, espera-se que as redes neurais continuem a desempenhar um papel crucial na compreensão e previsão do clima.

Aqui estão alguns dos hiperparâmetros mais comuns que são otimizados em redes neurais:

- **Taxa de Aprendizado (*Learning Rate*):** Determina o tamanho dos passos no processo de atualização dos pesos da rede durante o treinamento. Uma taxa de aprendizado muito alta pode fazer com que a rede salte a solução ótima, enquanto uma taxa muito baixa pode levar a um processo de treinamento muito lento ou a ficar preso em mínimos locais.
- **Número de Épocas**: Quantidade de vezes que o algoritmo de aprendizado trabalhará através de todo o conjunto de dados de treinamento. Mais épocas nem sempre significam melhores resultados, pois podem levar a *overfitting*.

- **Tamanho do *Batch* (*Batch Size*):** Número de amostras de treinamento usadas em uma iteração. Valores menores tornam o treinamento mais lento, mas podem melhorar a generalização. Valores maiores aceleram o treinamento, mas podem levar a uma convergência instável.
- **Número de Camadas e Neurônios em Cada Camada**: Redes mais profundas (com mais camadas) e redes com mais neurônios podem capturar padrões mais complexos, mas também são mais propensas a *overfitting* e são computacionalmente mais caras.
- **Funções de Ativação**: Como *ReLU, sigmoid, tanh* etc. Essas funções introduzem não-linearidades no modelo, permitindo que a rede aprenda padrões mais complexos.
- **Inicialização dos Pesos**: A escolha de como os pesos iniciais são definidos pode afetar a velocidade de convergência do treinamento e a capacidade de encontrar uma solução ótima.
- **Regularização (L1, L2, *Dropout*):** Técnicas para evitar *overfitting*, adicionando uma penalidade aos pesos ou desativando aleatoriamente neurônios durante o treinamento.
- **Otimizador**: Define o algoritmo usado para atualizar os pesos da rede (ex: SGD, Adam, RMSprop).
- **Taxa de Decaimento (*Decay Rate*):** Diminui a taxa de aprendizado ao longo do tempo, que pode ser útil para refinar a solução no final do treinamento.

A escolha e ajuste desses hiperparâmetros dependem fortemente do problema específico, da arquitetura da rede e dos dados disponíveis. Uma abordagem comum é começar com valores padrões recomendados na literatura e depois ajustar de acordo com o desempenho do modelo em dados de validação. A otimização de hiperparâmetros pode ser realizada manualmente, por *grid search*, *random search*, ou por métodos mais avançados como algoritmos genéticos ou otimização Bayesiana.

Quando se trata de imputar dados faltantes usando redes neurais, a escolha da arquitetura de rede neural pode depender de vários fatores, incluindo a natureza dos dados, o tamanho do conjunto de dados e os objetivos específicos da imputação. Não há uma única rede neural "melhor" para todos os cenários de imputação de dados

faltantes, mas existem algumas arquiteturas que são frequentemente usadas e podem ser adequadas em diferentes contextos. Aqui estão algumas delas:

Redes de Camadas Densas: As redes neurais com camadas densas, também conhecidas como redes totalmente conectadas, são uma arquitetura fundamental em aprendizado de máquina. Em uma camada densa, cada neurônio está conectado a todos os neurônios da camada anterior e da camada seguinte. Essas conexões permitem que a rede aprenda uma representação detalhada dos dados de entrada, transformando-os em saídas através de combinações lineares e não lineares. A fórmula matemática básica para uma camada densa é $y = f(Wx_{\square} + b)$, onde *W* são os pesos, *x* e a entrada, *b* é o viés, e *f* é a função de ativação.

O funcionamento das redes neurais com camadas densas envolve o treinamento da rede usando um algoritmo de otimização, como o gradiente descendente. Durante o treinamento, os pesos das conexões são ajustados para minimizar a função de perda, que mede a diferença entre as previsões da rede e os valores reais. A retropropagação é a técnica usada para calcular os gradientes dos pesos, permitindo que o algoritmo de otimização atualize os pesos de maneira eficiente. Esse processo iterativo continua até que a função de perda atinja um valor aceitável, indicando que a rede aprendeu a mapear as entradas para as saídas corretamente (Goodfellow, Bengio & Courville, 2016).

As redes neurais com camadas densas são aplicáveis em diversas áreas, incluindo reconhecimento de imagem, processamento de linguagem natural, e análise preditiva. Por exemplo, em visão computacional, essas redes são usadas para tarefas de classificação de imagens, onde cada pixel da imagem serve como uma entrada para a rede. No processamento de linguagem natural, são aplicadas em modelos de tradução automática e análise de sentimento, onde palavras ou frases são transformadas em vetores numéricos que a rede pode processar. A flexibilidade das camadas densas permite que elas sejam usadas em qualquer problema onde exista uma relação complexa entre as variáveis de entrada e saída (Goodfellow, Bengio & Courville, 2016).

Apesar de suas vantagens, as redes neurais com camadas densas apresentam desafios significativos. Elas requerem grandes quantidades de dados para evitar *overfitting*, onde a rede aprende padrões específicos do conjunto de treinamento em vez de generalizar para novos dados. Além disso, o treinamento dessas redes é

computacionalmente intensivo, muitas vezes necessitando de hardware especializado como GPUs para acelerar o processo. Técnicas como regularização, *dropout* e validação cruzada são usadas para mitigar esses problemas, melhorando a capacidade de generalização da rede e a eficiência do treinamento (Goodfellow, Bengio & Courville, 2016).

2.6 MÉTRICAS PARA AVALIAÇÃO DOS MÉTODOS

A precisão refere-se à medida de quão próximo o valor estimado está do valor verdadeiro de um dado ausente. Para avaliar a precisão dos métodos de completamento de dados climáticos, podem ser utilizadas diversas métricas, incluindo:

- **Erro Médio Quadrático (RMSE):** Essa métrica calcula a média dos quadrados dos erros entre os valores estimados e os valores verdadeiros. Quanto menor o RMSE, maior a precisão do método.
- **Coeficiente de Correlação**: Mede a relação linear entre os valores estimados e os valores verdadeiros. Um coeficiente de correlação próximo de 1 indica alta precisão.
- **Erro Médio Absoluto (MAE)**: Essa métrica calcula a média dos valores absolutos dos erros entre os valores estimados e os valores verdadeiros. O MAE fornece uma medida mais robusta à presença de valores discrepantes.
- **Erro Percentual Médio Absoluto (MAPE):** Calcula a média dos erros percentuais entre os valores estimados e os valores verdadeiros. É especialmente útil para avaliar o desempenho em escalas diferentes.

A robustez refere-se à capacidade dos métodos de completamento de dados de manter seu desempenho em diferentes cenários e condições. Ao avaliar a robustez, os seguintes aspectos podem ser considerados:

- **Sensibilidade a Diferentes Tipos de Variáveis Climáticas**: Verificar se o método mantém sua eficácia ao lidar com diversas variáveis climáticas, como temperatura, umidade, precipitação e pressão atmosférica. Métodos robustos devem ser aplicáveis a uma ampla gama de dados climáticos.
- **Variação Temporal e Espacial**: Avaliar como os métodos lidam com variações temporais (diárias, sazonais) e espaciais (locais, regionais) nas variáveis climáticas. A robustez pode ser testada por meio da aplicação dos métodos a diferentes locais e períodos.

- **Presença de Dados Ausentes em Grande Escala**: Verificar como os métodos se comportam quando há uma grande quantidade de dados ausentes em um período específico ou em uma região geográfica. Os métodos robustos devem ser capazes de lidar com cenários desafiadores.
- **Mudanças Climáticas**: Considerar como os métodos de completamento de dados respondem a mudanças climáticas e eventos extremos, que podem resultar em padrões climáticos não usuais.

A avaliação do desempenho em termos de precisão e robustez pode ser realizada por meio de testes controlados e análises estatísticas, incluindo validação cruzada, análise de sensibilidade e simulações. Além disso, é importante usar conjuntos de dados independentes para validar os métodos em condições do mundo real.

2.7 TRABALHOS CORRELATOS

As mudanças climáticas incluem o aumento da frequência e intensidade de eventos extremos, como ondas de calor, secas, inundações, tempestades, derretimento das geleiras polares, além de mudanças nos padrões de precipitação e na distribuição de espécies. O impacto dessas mudanças é global, mas apresenta variações regionais significativas, afetando de maneira desigual diversos ecossistemas e comunidades ao redor do mundo. Especificamente na costa oeste do Atlântico Sul, uma pesquisa recente de Sanches et al. (2023) evidencia um aumento notável na intensidade e frequência de extremos de temperatura do ar na superfície, demonstrando que estas alterações climáticas não são uniformes e possuem implicações distintas em diferentes latitudes. Este estudo destaca a importância de compreender as mudanças climáticas em um contexto regional para melhor apreciar seu impacto diversificado em ecossistemas e comunidades locais.

O relatório intergovernamental de 2023 sobre mudanças climáticas alerta para os riscos cada vez maiores de tais mudanças serem irreversíveis e estima que o mundo enfrentará consequências catastróficas, como o desaparecimento de ecossistemas inteiros nas regiões polares, costeiras e montanhosas e um risco muito elevado de extinção de 14% de todas as espécies terrestres (IPCC, 2023).

No contexto brasileiro, a complexidade dos impactos das mudanças climáticas é intensificada pela vasta diversidade dos biomas do país e sua posição geográfica estratégica, tornando-o um elo crucial na dinâmica climática global. O Brasil,

abrigando biomas diversos como Amazônia, Cerrado, Mata Atlântica, Pantanal, Caatinga e Pampas, enfrenta desafios específicos devido a incêndios florestais e elevadas emissões de carbono, como demonstrado pelo estudo "*Persistent fire foci in all biomes undermine the Paris Agreement in Brazil*". Este estudo ressalta a urgência de políticas públicas eficazes para prevenir incêndios e reduzir emissões em todos os biomas, enfatizando a importância de uma abordagem detalhada e regionalizada na compreensão e resposta às mudanças climáticas no Brasil.

Essa compreensão, no entanto, enfrenta desafios significativos, principalmente pela necessidade de um extenso período de dados climáticos para análises confiáveis e significativas. Como indicado por Lovejoy e Schertzer (2022), a utilização de intervalos de tempo mais longos, no mínimo de 20 anos, é crucial para identificar tendências, variabilidades e extremos climáticos, uma vez que as taxas de mudança climática são frequentemente subestimadas em períodos mais curtos, destacando a importância de um olhar de longo prazo para a modelagem precisa de cenários futuros e para a tomada de decisões informadas para mitigação e adaptação.

A construção de uma base de dados climáticos robusta e abrangente para o Brasil é um desafio complexo (MOURA e FORTES, 2021). Enquanto o Instituto Nacional de Meteorologia (INMET) oferece dados cruciais, existem lacunas notáveis, como dados faltantes, outliers, e interrupções devido à desativação temporária de equipamentos ou falta de manutenção. Este cenário é ainda mais complicado pelos problemas de dados incompletos em regiões com monitoramento recente, inferior ao período necessário para uma análise abrangente. Tal realidade, comum em estudos ambientais e climáticos no Brasil, pode comprometer significativamente a integridade, precisão e tempo de pesquisa. O artigo ressalta a necessidade de métodos avançados de controle de qualidade e preenchimento de lacunas para assegurar a confiabilidade dos dados climáticos, um aspecto crucial para a validade das análises e pesquisas em um país de vasta diversidade climática como o Brasil.

A identificação, tratamento ou completamento de dados faltantes são etapas cruciais no processo de pesquisa. Dados ausentes, resultantes de diversos fatores como falhas em equipamentos, erros na coleta de dados ou lacunas temporais e espaciais na cobertura de monitoramento, podem comprometer significativamente a integridade das análises. Como ilustrado no estudo de Sattari et al. (2017), que aborda a questão de dados de precipitação ausentes em regiões áridas, o reconhecimento e o tratamento correto de dados faltantes são essenciais para manter a qualidade e a

confiabilidade das análises. Este estudo destaca a relevância de abordagens meticulosas para lidar com a ausência de dados, enfatizando a necessidade de garantir a completude dos dados para análises precisas e confiáveis.

Métodos estatísticos envolvem técnicas como interpolação, regressão linear e análise de séries temporais. A interpolação, por exemplo, é usada para estimar valores de dados em locais onde não foram coletados, baseando-se em valores conhecidos de locais próximos. Um estudo significativo neste campo é o realizado por Sattari et al. (2017), que avaliou o uso da média aritmética, uma forma de interpolação, para preencher dados de precipitação ausentes, demonstrando a eficácia dessa abordagem em condições geográficas similares.

Outra técnica comum é a regressão linear, que permite estimar dados faltantes com base em correlações entre variáveis. No mesmo estudo, Sattari et al. (2017) aplicaram a regressão linear múltipla, destacando sua utilidade na previsão de dados de precipitação. A análise de séries temporais, por sua vez, é útil para preencher lacunas temporais. O trabalho de Sattari et al. (2017) também exemplifica isso, ao aplicar métodos avançados como algoritmos de árvore de decisão para analisar e preencher dados temporais de precipitação.

Com o avanço da computação, algoritmos de aprendizado de máquina têm sido cada vez mais aplicados para tratar dados faltantes. Técnicas como Redes Neurais Artificiais (RNA) e Máquinas de Vetores de Suporte (SVM) são empregadas para modelar e prever valores faltantes com base em padrões complexos nos dados existentes. Um exemplo notável é o trabalho de Kadow, Hall e Ulbrich (2020), que utiliza técnicas de inteligência artificial para reconstruir informações climáticas ausentes, demonstrando precisão superior em comparação com métodos estatísticos tradicionais. Neste estudo, técnicas de *'image inpainting'* foram aplicadas para preencher lacunas em dados de temperatura, mostrando altos coeficientes de correlação e baixos erros quadráticos médios em comparação com os dados originais.

Métodos híbridos, que combinam técnicas estatísticas e de aprendizado de máquina, são eficazes em tirar proveito das forças de ambos os domínios. Por exemplo, o estudo de Breve et al. (2023) emprega uma abordagem estatística para a redução da dimensionalidade através da Análise de Componentes Principais (PCA) e dos Mínimos Quadrados Parciais (PLS), seguida pela aplicação do método *Analog Ensemble (AnEn)* para a reconstrução de dados meteorológicos. Esta combinação permite uma análise inicial eficaz dos padrões nos dados e, subsequentemente, utiliza

algoritmos de aprendizado de máquina para prever valores faltantes de forma precisa. Este método híbrido, conforme apresentado no estudo, mostrou-se particularmente eficiente para variáveis meteorológicas instáveis, como a velocidade do vento, demonstrando a utilidade prática e a precisão dos métodos híbridos em situações complexas de previsão de dados.

Diante do exposto, torna-se evidente a necessidade crítica de métodos eficientes para o preenchimento de lacunas em bases de dados climáticos, especialmente em um país com a complexidade geográfica e climática do Brasil. A precisão desses dados é vital para a compreensão e modelagem das mudanças climáticas e para a implementação de políticas públicas eficazes (AVILA-DIAZ, Álvaro et al. 2020).’

3. PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa, segundo sua natureza é um resumo de assunto, buscando explicar a área do conhecimento do projeto, indicando sua evolução histórica, como resultado da investigação das informações obtidas, levando ao entendimento de suas causas e explicações (WAZLAWICK, 2014).

Segundo os objetivos é uma pesquisa exploratória e descritiva. A pesquisa descritiva busca dados mais consistentes sobre determinado assunto, porém, não ocorre a interferência do pesquisador, apenas expõe os fatos como realmente são (WAZLAWICK, 2014). As pesquisas descritivas descrevem as características de certo fenômeno ou população. Também pode ser elaborada com o intuito de identificar as relações entre as variáveis (GIL, 2017).

A pesquisa exploratória muitas vezes é considerada como a primeira parte do processo de pesquisa, porque não necessariamente o autor tem um objetivo ou hipótese definida (WAZLAWICK, 2014). Essa pesquisa tem como objetivo a maior familiaridade do autor com o problema, tornando mais explícito ou facilitar a construção de hipóteses. Geralmente é uma pesquisa flexível, porque considera os variados aspectos referentes aos fatos ou fenômeno estudado (GIL, 2017).

Quanto aos procedimentos técnicos, será uma pesquisa bibliográfica e experimental. A pesquisa bibliográfica requer o estudo de teses, artigos, entre outros. A pesquisa experimental é caracterizada por ter uma ou mais variáveis experimentais que podem ser coordenadas pelo pesquisador (WAZLAWICK, 2014).

A pesquisa bibliográfica, foi elaborada a partir de materiais já publicados, podendo incluir livros, teses, materiais disponibilizados na Internet, revistas, entre outros. A principal vantagem é permitir uma sucessão de fenômenos maior do que seria capaz de pesquisar diretamente (GIL, 2017).

A pesquisa experimental consiste que o pesquisador provoque mudanças no ambiente de pesquisa, observando se as alterações realizadas são de acordo com os resultados esperados (WAZLAWICK, 2014).

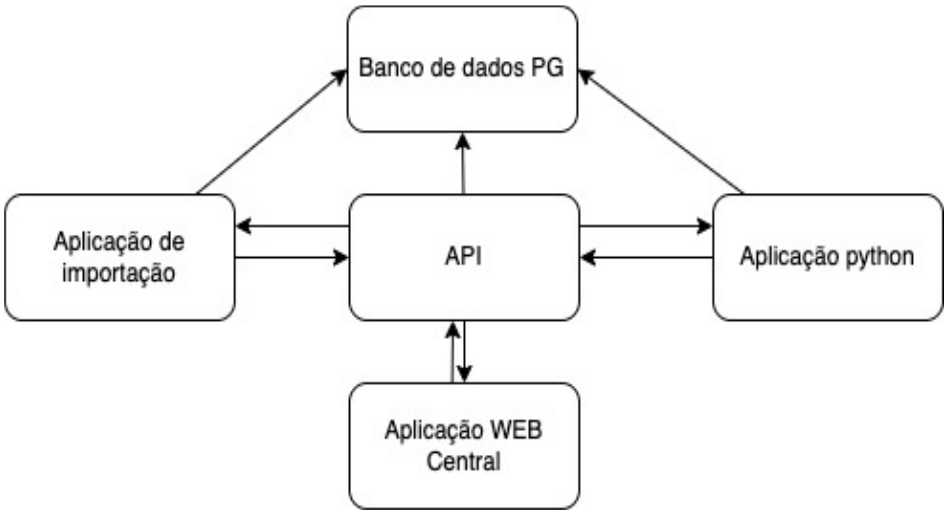
A pesquisa experimental consiste em estabelecer um objeto de estudo, escolher as variáveis que a influenciam e determinar as formas de controle e observar os efeitos que a variável gera no objeto. Realiza pelo menos um dos elementos que julga ser responsável pela circunstância que está sendo pesquisado (GIL, 2017).

No presente trabalho, a pesquisa exploratória foi conduzida utilizando a base científica *Web of Science*. Para avaliar as direções da pesquisa, foi realizada uma busca utilizando as palavras-chave "*missing data*", "*climate*" e "*imputation*".

Nesta pesquisa serão desenvolvidos o tratamento, conversão e completamento de dados automaticamente, fornecendo uma base de dados planejados, previamente tratados, padronizados, documentados e de fácil acesso para todos que necessitem desses dados. Utilizando tecnologias mais recentes para desenvolvimento web, pretende-se desenvolver algumas aplicações separadamente, mas que trabalharão em conjunto para importar, tratar, validar e entregar os dados climáticos com qualidade e rapidez.

O primeiro grupo contém os dados brutos, importados diretamente de cada fonte de dado climático (INMET, NOAA, CPTEC, entre outros), e cada dado importado terá sua respectiva característica, seja variável, unidade de medida, tempo de coleta, tipo de sensor e origem. O segundo grupo contém os dados processados, que passarão por tratamento, correção de *outliers*, completamento de dados faltantes com o uso de métodos estatísticos ou inteligência artificial (IA). Este trabalho tem como objetivo principal desenvolver métodos de completamento para integrar com um sistema de dados climáticos brasileiros já existente com uma API de fácil acesso, integrando as principais bases climáticas, conforme Figura 1.

Figura 1- API de dados do Clima



Fonte: Autoria propria

Foram realizadas a importação e visualização das informações em gráficos das diferentes variáveis climáticas. Para fazer todo o processo de comunicação e importação dos dados das fontes externas, a base do INMET foi cadastrada para

importação por meio de API, *webservice* e até mesmo por arquivo de texto, Excel. Os consumos via API ou webservice serão automáticos para que os dados sejam importados em tempo real de cada plataforma, tendo dados, atualizados em tempo real. Para a importação o INMET exigiu que o projeto fosse cadastrado para a liberação de dados por um período que deverá ser renovado.

O Instituto Nacional de Meteorologia (INMET) é um órgão do governo brasileiro responsável por fornecer previsões do tempo, informações climáticas e outros dados meteorológicos para o país. O INMET opera uma rede de estações meteorológicas em todo o Brasil, que recolhe dados sobre temperatura, precipitação, vento, humidade e outras variáveis meteorológicas.

O INMET também fornece acesso a dados históricos de tempo e clima através de seu site, incluindo dados sobre temperatura, precipitação e outras variáveis meteorológicas para vários locais em todo o Brasil. Os dados podem ser acessados em vários formatos, como CSV ou NetCDF. Os dados do INMET podem ser usados para pesquisa, planejamento e tomada de decisões em vários setores, como agricultura, energia, transporte e saúde. Ele tem uma ampla cobertura de sensores onde as variáveis medidas são:

• **Temperatura**: Isso inclui a temperatura do ar, que é frequentemente medida em graus Celsius (°C) ou Fahrenheit (°F), bem como a temperatura do solo em diferentes profundidades.

• **Umidade Relativa**: A umidade relativa do ar, geralmente expressa em porcentagem (%), que indica a quantidade de vapor de água presente no ar em relação à quantidade máxima que o ar pode conter a uma determinada temperatura.

• **Precipitação**: Informações sobre a quantidade de chuva ou outras formas de precipitação, geralmente medida em milímetros (mm) ou centímetros (cm).

• **Pressão Atmosférica**: A pressão atmosférica ao nível do mar, frequentemente medida em hectopascais (hPa) ou milibares (mb).

• **Velocidade do Vento**: Dados sobre a velocidade do vento, geralmente em metros por segundo (m/s) ou quilômetros por hora (km/h).

• **Direção do Vento**: A direção de onde o vento está soprando, geralmente expressa em graus, onde 0° representa o norte, 90° leste, 180° sul e 270° oeste.

• **Radiação Solar**: Informações sobre a quantidade de radiação solar incidente, medida em watts por metro quadrado (W/m²).

• **Radiação Ultravioleta (UV):** A intensidade da radiação ultravioleta do sol, geralmente apresentada em unidades de índice UV.

• **Neblina e Visibilidade**: Dados relacionados à visibilidade, especialmente importante em áreas de tráfego e aviação.

• **Índices Climáticos**: O INMET também pode calcular índices climáticos específicos, como índice de calor, índice de resfriamento, índice de seca, entre outros.

• **Nível de Água e Vazão de Rios**: Em algumas estações, dados sobre o nível de água em rios e lagos, bem como a vazão dos rios, podem estar disponíveis.

A disponibilidade dessas variáveis pode variar de estação para estação, dependendo da localização e da finalidade da estação meteorológica. Além disso, o INMET oferece acesso a esses dados por meio de sua plataforma online e API, tornando-os amplamente acessíveis para pesquisa e aplicação em várias áreas, incluindo agricultura, previsão do tempo, monitoramento ambiental e muito mais.

Os dados gerados ou tratados têm um sinalizador e um descritivo demonstrando o que foi feito. Para o tratamento dos dados importados, será desenvolvida uma aplicação em Python que verifica possíveis outliers, corrigindo-os e preenchendo os dados faltantes e padronizando as unidades de medidas.

Para acesso aos dados por meio externo, existe uma API dedicada para que os usuários possam acessar grandes intervalos de dados sem limitação. Para uma interface de usuário, a aplicação web agrupa todos os recursos anteriores. Nesta aplicação, o usuário poderá se cadastrar, verificar documentos de usabilidade e documentos do fluxo, quais plataformas externas fornecem os dados (inicialmente, só o INMET), como os dados são tratados e concluídos. Os usuários também poderão ter acesso diretamente aos dados climáticos, podendo fazer algumas análises prévias com gráficos já disponibilizados, aplicando filtros e fazendo o download dos dados no formato desejado. Com isso, espera-se uma aplicação estável e acessível, fornecendo dados climáticos de fontes diversas com a qualidade necessária para apoiar o desenvolvimento de pesquisas que demandem dados do clima.

Fases detalhadas:

Coleta e preparação dos dados climáticos: Reunir dados climáticos de fontes confiáveis e identificar lacunas nos conjuntos de dados. Já existe uma API disponível em http://weather.iadadev.com que coleta dados em tempo real do base de dados INMET, projeto de iniciação científica “API WEB PARA INTEGRAR DIFERENTES BASES COM DADOS DO CLIMA DO BRASIL”, 08-2022ª 06-2023.

3.1 PRÉ-PROCESSAMENTO DE DADOS

Normalizar os dados, tratar valores ausentes identificados previamente e realizar transformações necessárias, como a criação de características derivadas.

Desenvolvimento de métodos de completamento de dados usando aprendizado de máquina:

Seleção de recursos: Identificar as variáveis climáticas mais relevantes para o completamento de dados.

Modelos de regressão: Treinar modelos de regressão, como regressão linear, regressão logística ou regressão de árvore de decisão, para prever valores ausentes com base em dados disponíveis. Usar técnicas de validação cruzada para avaliar o desempenho dos modelos.

Aprendizado profundo: Explorar o uso de redes neurais, como redes neurais convolucionais (CNNs) ou redes neurais recorrentes (RNNs), para capturar relações complexas entre variáveis climáticas e prever valores ausentes.

Transferência de aprendizado: Investigar a aplicabilidade de técnicas de transferência de aprendizado, treinando modelos em dados de climas semelhantes e adaptando-os para a base de dados específica em estudo.

Avaliação do desempenho: Avaliar o desempenho dos modelos de aprendizado de máquina usando métricas adequadas, como erro médio quadrático (RMSE), R-*squared*, MAE (Erro Médio Absoluto) e validação cruzada. Comparar o desempenho dos modelos de aprendizado de máquina com métodos estatísticos.

Otimização de hiperparâmetros: Realizar ajustes nos hiperparâmetros dos modelos de aprendizado de máquina para otimizar o desempenho.

Visualização de resultados: Criar visualizações e gráficos para comunicar eficazmente os resultados aos interessados e tomar decisões informadas.

Documentação e replicabilidade: Documentar cuidadosamente todo o processo, incluindo escolhas de modelos, hiperparâmetros e resultados, para permitir a replicação da pesquisa por outros cientistas.

Descrever as descobertas, limitações, conclusões e recomendações.

Localização de dados faltantes.

A análise para localizar dados faltantes, muitas vezes referida como "análise de dados ausentes", é uma etapa importante na preparação de dados climáticos antes

de aplicar técnicas de preenchimento ou completamento de dados. Uma das abordagens eficazes é o uso de um "mapa de calor" ou matriz de correlação para visualizar padrões de dados ausentes.

3.2 PREPARAÇÃO DOS DADOS

Deve-se reunir todos os dados climáticos relevantes em um formato adequado, como um DataFrame em Python usando a biblioteca Pandas.

Fazer o uso de funções como isnull () ou isna () em seu DataFrame para identificar valores ausentes em seus dados. Isso criará uma máscara booleana que indica quais pontos nos dados estão ausentes.

Criar um mapa de calor de correlação: Após identificar os dados ausentes, um mapa de calor de correlação entre as variáveis climáticas não mostrará diretamente os dados ausentes, mas ajudará a entender como as variáveis estão relacionadas.

Para visualizar diretamente os dados ausentes, pode criar um gráfico de barras ou um gráfico de dispersão que mostre a distribuição dos valores ausentes em relação ao tempo ou à localização geográfica.

Para avaliar a quantidade de dados faltantes em seu conjunto de dados climáticos, deve-se usar uma métrica chamada "Taxa de Dados Ausentes" (*Missing Data Rate*) ou "Porcentagem de Dados Ausentes" (*Missing Data Percentage*). Essa métrica expressa a proporção de valores ausentes em relação ao número total de observações ou entradas no conjunto de dados.

Com base na taxa de dados ausentes, decidir como lidar com os dados ausentes. Isso pode incluir o preenchimento dos valores ausentes, a remoção das observações ou a aplicação de técnicas mais avançadas de imputação de dados.
Calcular a taxa de dados ausentes para cada variável individualmente. Isso ajudará a identificar quais variáveis são mais afetadas pela falta de dados. Variáveis com uma alta taxa de dados ausentes pode ser mais problemáticas.

Criar gráficos ou visualizações que mostrem a distribuição dos dados ausentes ao longo do tempo ou da localização geográfica, se aplicável. Isso pode revelar padrões temporais ou espaciais nas lacunas de dados.

Considerar a importância de cada variável climática nas análises ou modelos. Se uma variável com muitos dados ausentes for crítica para os objetivos, isso pode indicar um maior comprometimento dos dados.

Analisar a correlação entre variáveis climáticas e a presença de dados ausentes. Variáveis que têm uma alta correlação com a presença de dados ausentes podem indicar uma relação significativa.

Entender por que os dados estão faltando. Isso pode envolver verificar se há padrões sazonais nas lacunas de dados, problemas técnicos na coleta de dados ou outras razões específicas.

Decidir se é possível e apropriado preencher os dados ausentes. Dependendo da quantidade e da natureza dos dados ausentes, técnicas de imputação como média, mediana, regressão ou métodos de aprendizado de máquina podem ser aplicados.

Realizar análises de sensibilidade para entender como diferentes abordagens para lidar como os dados ausentes afetam os resultados. Comparar os resultados de análises com dados ausentes imputados e análises com dados removidos.

Fazer Validação Cruzada para avaliar o desempenho do modelo com diferentes conjuntos de treinamento e teste, levando em consideração os dados ausentes.

Comunicar claramente os resultados da análise em relação aos dados ausentes, destacando qualquer impacto potencial nos resultados e nas conclusões.

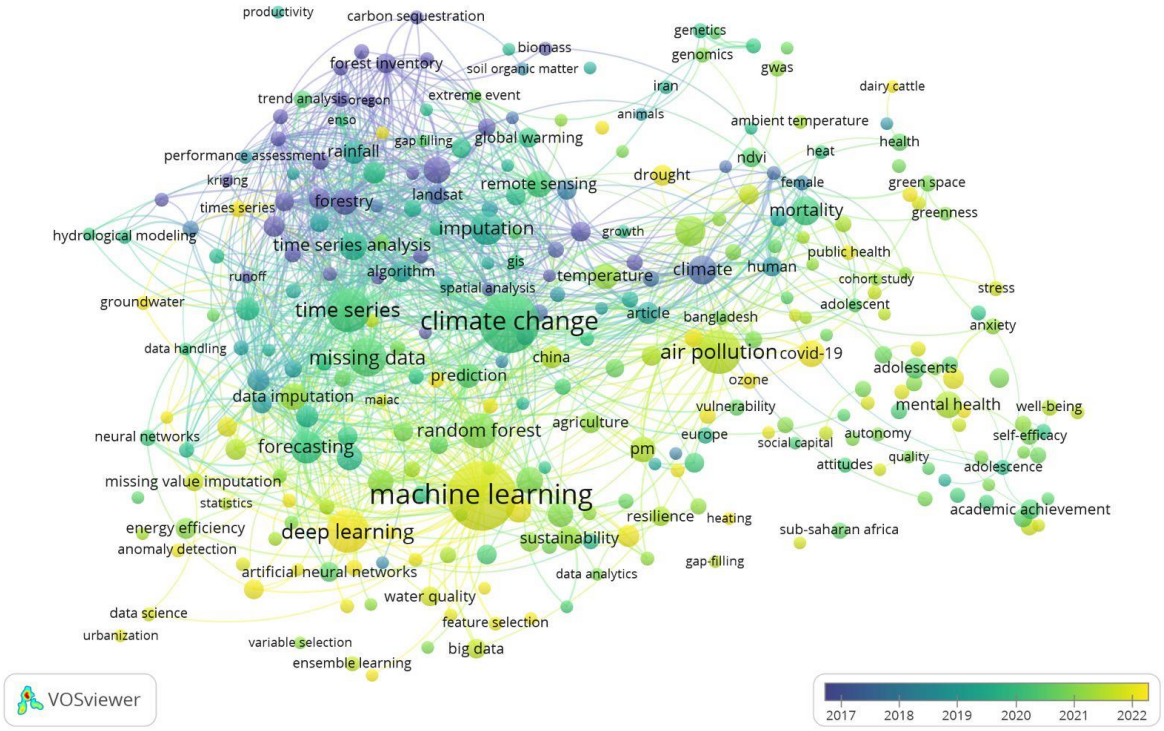
Desenvolver um plano para lidar com os dados ausentes de maneira consistente e informada, e documentar todas as etapas tomadas.

A avaliação do grau de comprometimento da base de dados devido a dados ausentes é essencial para garantir a integridade e a validade de suas análises e modelos climáticos. A abordagem específica dependerá da natureza dos dados e dos objetivos da sua pesquisa ou análise.

3.3 PESQUISA EXPLORATÓRIA NA BASE CIENTÍFICA WEB OF SCIENCE

Inicialmente pesquisou-se sobre alterações climáticas (*climate change*) e (*machine Learning*) para ver temas importantes na área. Pode ser observado na Figura 2 que dados faltantes e imputação aparecem com algum destaquem na nuvem de palavras.

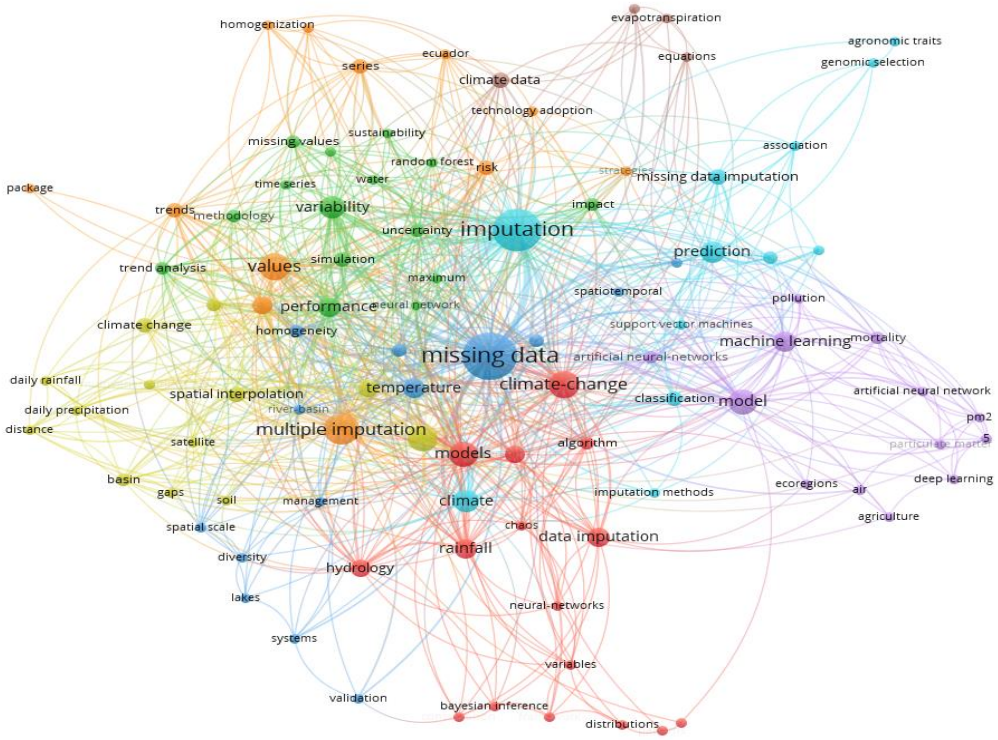
Figura 2 - Mapa de Tópicos Relevantes em Aprendizado de Máquina e Mudanças Climáticas



Fonte: autoria própria

Outra análise foi conduzida com as palavras chaves *missing data*, *imputation*, *machine learning*. Essa análise contribuiu para identificar os principais termos e conceitos relacionados ao tema, orientando a compreensão e a delimitação da pesquisa. A Figura 3 mostra o foco em completamento de bases climáticas.

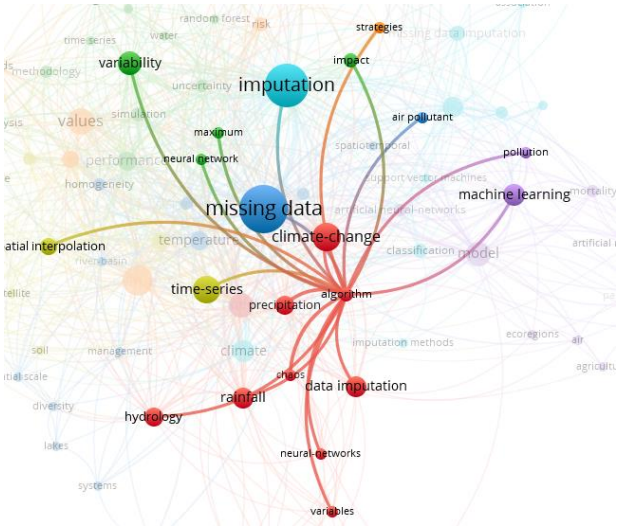
Figura 3 - Nuvem de palavras-chave da pesquisa



Fonte: autoria própria

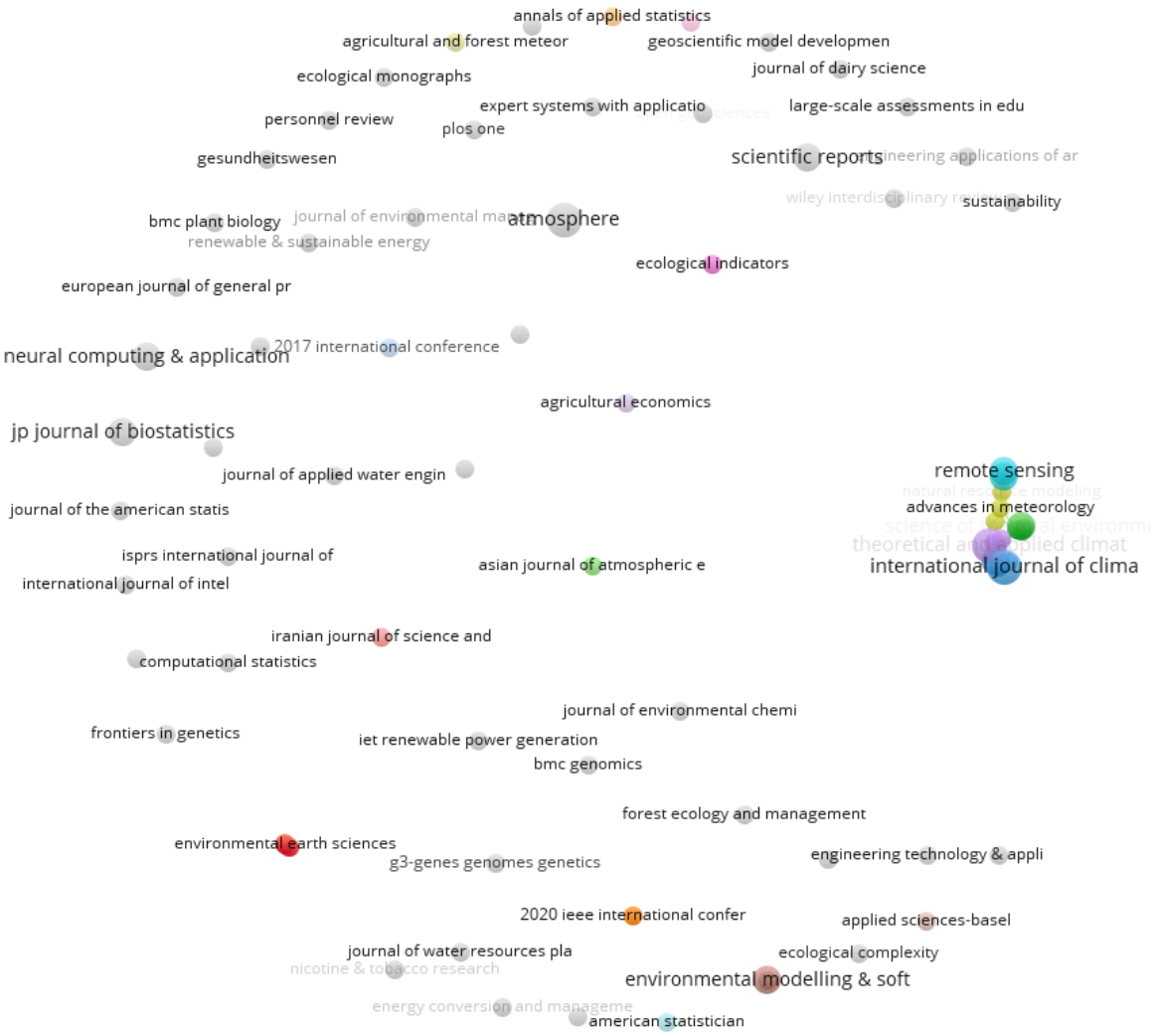
Na Figura 4, realizou-se um zoom para ver a palavra-chave *missing* data e suas relações na nuvem de palavras. Na Figura 5 o objetivo e mostrar as fontes de artigos na área.

Figura 4 - Zoom dos algoritmos relacionados com a imputação de dados perdidos em bases climáticas.



Fonte: autoria própria

Figura 5 - Fontes dos artigos – avaliação da qualidade das fontes



Fonte: autoria própria

3.4 ANÁLISE DA REVISÃO EXPLORATÓRIA DA LITERATURA: MÉTODOS DE PRÉ-PROCESSAMENTO DE DADOS E MÉTODOS DE IMPUTAÇÃO.

O Objetivo da análise foi avaliar os métodos presentes na literatura para a proposta de completamento de dados faltantes. A Tabela 1 mostra uma lista de artigos consultados para esta finalidade. A Tabela compila informações para demonstrar a qualidade dos trabalhos utilizados.

Tabela 1 - Trecho da Tabela de Portifólio de artigos selecionados na base Web of Science

TÍTULO	Resumo	Palavras-chave do Autor	Palavras-chave da Base	Citações	Ano	Valor do JCR	InOrdinatio	Autor(es)	Revista	Referências
Missing data imputation of high-resolution temporal climate time series data	Analys	high-r	VALU	38	2.020	2,451	110,451	Afrifa-Yamoah,	METEOROLOGICAL APPLICATIONS	40
Comparative assessment of univariate and multivariate imputation models for varying lengths of missing rainfall data in a humid tropical region: a case study of Kozhikode, Kerala, India	Accur	Rainfa	MULT	0	2.023	2,293	102,293	Kannegowda, N	ACTA GEOPHYSICA	39
Frequency based imputation of precipitation	Chang	Frequ	DRIVE	3	2.017	3,821	10,821	Dikbas, Fatih	STOCHASTIC ENVIRONMENTAL RESEARCH AND RISK ASSESSMENT	24
Improving Groundwater Imputation through Iterative Refinement Using Spatial and Temporal Correlations from In Situ Data with Machine Learning	Obtair	groun	STOR	1	2.023	3,530	104,530	Ramirez, Saul G.	WATER	40
Analysis of Preprocessing Techniques for Missing Data in the Prediction of Sunflower Yield in Response to the Effects of Climate Change	Machi	imput		0	2.023	2,838	102,838	Calin, Alina Deli	APPLIED SCIENCES-BASEL	38
Incremental Missing-Data Imputation for Evolving Fuzzy Granular Prediction	Missir	Adapt	DATA	44	2.020	12,253	126,253	Garcia, Cristianc	IEEE TRANSACTIONS ON FUZZY SYSTEMS	48
Data-Driven Modeling of Flows of Antalya Basin and Reconstruction of Missing Data	The le	Antaly	SUPP	1	2.020	1,461	72,461	Dikbas, Fatih; Ya	IRANIAN JOURNAL OF SCIENCE AND TECHNOLOGY-TRANSACTIONS OF CIVIL ENGINEERING	21
Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model	Havin	Differ	PERF	3	2.023	11,533	114,533	Basakin, Eyyup E	ENERGY CONVERSION AND MANAGEMENT	87
Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder	A key	Missir	DENO	5	2.022	5,102	100,102	Wardana, I. Nyo	NEURAL COMPUTING & APPLICATIONS	66

Fonte: autoria própria

3.5 A ETAPA DE PRÉ-PROCESSAMENTO

O primeiro passo na análise de dados faltantes é a identificação do padrão e do mecanismo desses dados, um processo que aumenta significativamente a complexidade da análise. Os mecanismos de dados faltantes são principalmente baseados em três categorias: faltando completamente ao acaso (MCAR), faltando não ao acaso (MNAR) e faltando ao acaso (MAR). No caso do MCAR, os dados faltantes ocorrem completamente ao acaso, independentemente de qualquer parâmetro, enquanto nos mecanismos MNAR e MAR, os dados são reduzidos pelo efeito de uma ou mais variáveis independentes. Além disso, existem diversos padrões que indicam a distribuição dos dados faltantes, como os que abordam a aleatoriedade dos dados ou padrões que seguem uma cadeia específica. Estes padrões são chamados de univariados, multivariados, monótonos, não monótonos e padrões gerais (T. Aljuaid and S. Sasi, 2016)

A segunda parte crucial da análise de dados faltantes é decidir se os dados devem ser removidos do conjunto de dados ou preservados. A literatura pertinente argumenta que os dados faltantes podem ser removidos do conjunto de dados principal se representarem 5% ou menos do total. Por outro lado, estratégias de

imputação de dados faltantes podem ser realizadas para preencher as lacunas existentes, reforçando a natureza do conjunto de dados original. Os métodos tradicionais incluem estratégias baseadas em estatísticas descritivas do conjunto de dados original, como média, moda e mediana. No entanto, a imputação de dados faltantes através de parâmetros estatísticos pode levar à monotonia dos espaços preenchidos, resultando em atribuições imprecisas. (LOPEZ et al, 2021)

Recentemente, métodos modernos, como algoritmos de aprendizado de máquina, têm sido explorados por serem capazes de produzir estimativas precisas sem restrições ou suposições estatísticas. Esses algoritmos são classificados em quatro grupos: supervisionados, semissupervisionados, não supervisionados e de reforço. Entre eles, os algoritmos de aprendizado supervisionado, que estimam variáveis dependentes por meio de variáveis independentes, são as ferramentas mais utilizadas. Esses algoritmos operam através da introdução de amostras pré-definidas para treinamento, seguida pela imputação dos valores faltantes. Estudos nas últimas duas décadas forneceram insights vitais de que técnicas modernas de modelagem produzem estimativas mais precisas em comparação com métodos tradicionais de imputação (GAD & MANJUNATHA, 2017).

Tabela 2 - Nome do artigo x pré-processamento realizado para a aplicação das técnicas.

Artigo	Pré-processamento dos Dados Faltantes
An Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach	Utilização do algoritmo Random Forest na etapa de pré-processamento para imputar dados faltantes e seleção de características.
Autoencoder model with spatiotemporal considerations for air quality data	Duas fases de pré-processamento focadas em correlações de poluentes e preparação de características espaço-temporais para o modelo de aprendizado profundo.
Optimization of regression methods for imputation of missing data in climatology	Pré-tratamento do conjunto de dados, testando semelhanças entre estações e recalculando a porcentagem de dados faltantes após cada etapa de imputação.

Artigo	Pré-processamento dos Dados Faltantes
"O efeito de imputações simples baseadas em quatro variantes de métodos PCA nos quantis de dados anuais de precipitação"	Geração de lacunas de dados aleatórias (MCAR) em diferentes porcentagens (10, 20, 30, 40%) a partir de dados observados, e avaliação da imputação usando RMSE, MAE, EQR e CC.
"Melhorando a Imputação de Águas Subterrâneas Através de Refinamento Iterativo Usando Correlações Espaciais e Temporais de Dados In Situ com Aprendizado de Máquina"	Inicialmente, foi realizada uma imputação inicial para gerar uma série temporal completa para cada poço, substituindo qualquer valor imputado que possuía uma medição observada pelo dado original. Após a imputação inicial, foram selecionados o número de iterações e utilizou-se um filtro Hampel para suavizar picos de dados sintéticos, removendo outliers do conjunto de dados inicialmente imputado.

Fonte: autoria própria

A Tabela 3 apresenta a lista de artigos e o tipo de método de completamento utilizado.

Tabela: Nome do artigo x método de imputação utilizado. Em todas as propostas observa-se que são técnicas de aprendizado de máquina. Os métodos relacionados incorporam metaheurísticas, floresta aleatória, lógica *Fuzzy*, regressão linear múltipla, gradiente *boosting*, modelo de *autoencoder*, entre outros. Para cada artigo foi descrito um pequeno resumo sobre as propostas.

Tabela 3 - Nome do artigo x Métodos de Completamento

Artigo	Resumo
Developing a novel approach for missing data imputation of solar radiation	Este estudo introduz uma técnica nova para imputação de dados faltantes em medições de radiação solar, combinando algoritmos de aprendizado de máquina com um algoritmo de otimização meta-heurístico.

Artigo	Resumo
An Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach	Propõe um método para previsão de índices de qualidade do ar utilizando redes neurais artificiais e a técnica de imputação de dados missForest baseada em florestas aleatórias.
Incremental Missing-Data Imputation for Evolving Fuzzy Granular Prediction	Aborda a imputação de dados faltantes em contextos de fluxos de dados do mundo real, destacando a natureza comum deste problema.
Missing data imputation of high-resolution temporal climate time series	Avalia múltiplas abordagens para a imputação de valores faltantes em séries temporais climáticas de alta resolução, incluindo modelos de séries temporais estruturais e regressão linear múltipla.
Introducing Gradient Boosting as a universal gap filling tool for meteorological time series	Introduz o Gradient Boosting como um método eficaz para preencher lacunas causadas por dados faltantes ou errôneos em séries temporais meteorológicas.
Machine Learning Techniques for Missing Climatic Values	Descreve um estudo sobre a combinação de várias técnicas de aprendizado de máquina para determinar valores climáticos faltantes, incluindo redes neurais, árvores de regressão e florestas aleatórias.
Optimization of regression methods for imputation of missing data in climatology	Propõe um algoritmo de imputação baseado na otimização de métodos de regressão para dados faltantes em climatologia.
Two-dimensional data-driven model of precipitation: The FBI method	Implementa um modelo bidimensional orientado a dados para a estimação de precipitação, utilizando o método de imputação baseado em frequência (FBI).
Autoencoder model with spatiotemporal considerations for air quality data	Estuda um modelo de autoencoder com considerações espaço-temporais para estimar valores faltantes em dados de qualidade do ar.

Artigo	Resumo
Comparison and selection criterion of missing imputation methods in Ethiopia	Avalia o desempenho de oito métodos de estimação de dados faltantes e usa um método de decisão multicritério para identificar o melhor método de imputação.
"O efeito de imputações simples baseadas em quatro variantes de métodos PCA nos quantis de dados anuais de precipitação"	1. Métodos simples de imputação e imputações múltiplas (Presti et al. 2010; Audigier et al. 2016). 2. Quatro variantes de imputação simples baseadas em Análise de Componentes Principais (PCA): PCA Probabilístico (PPCA), PCA de Expectativa-Maximização (EMPCA), PCA Regularizado (RPCA) e PCA de Decomposição em Valores Singulares (SVDPCA).
"Modelagem de Dados Orientada por Dados de Fluxos da Bacia de Antalya e Reconstrução de Dados Ausentes"	Método de imputação baseado em frequência, que estima associações quantitativas entre elementos de uma matriz nas direções horizontal, vertical e diagonal.
"Melhorando a Imputação de Águas Subterrâneas Através de Refinamento Iterativo Usando Correlações Espaciais e Temporais de Dados In Situ com Aprendizado de Máquina"	1. Método de Modelo de Refinamento Iterativo (IRM), incluindo etapas para selecionar o número de iterações e usar um filtro Hampel para suavizar picos de dados sintéticos 2. Um método para melhorar as previsões de imputação através de uma abordagem IRM, aplicável a conjuntos de dados de aquíferos com uma mistura de valores medidos e imputados.

Fonte: autoria própria

3.6 COLETA DE DADOS

Inicialmente, o projeto partiu de uma Iniciação Científica que culminou na criação de um sistema para a coleta e armazenamento automática de dados

meteorológicos do Instituto Nacional de Meteorologia (INMET). Este sistema, operando de maneira eficiente, realiza a importação de dados de hora em hora através de uma API, garantindo assim a atualização contínua e a confiabilidade das informações coletadas.

O escopo atual do sistema abrange um total de 567 estações meteorológicas automáticas, das quais 509 estão operantes e as demais estão inativas devido a falhas técnicas. Os registros dessas estações têm um histórico abrangente, iniciando-se no ano 2000 até a presente data desta pesquisa (2023), o que oferece uma base de dados robusta e extensa para análises e estudos climáticos.

3.7 MÉTODOS DE COMPLETAMENTO

Neste trabalho, três métodos principais foram utilizados para o completamento de dados faltantes: o Método do Vizinho Mais Próximo (*Nearest Neighbor*), Regressão Linear e Redes Neurais de Camadas Densas. A escolha destes métodos deve-se à sua eficácia comprovada em diferentes contextos e à capacidade de lidar com a diversidade e complexidade dos dados climáticos.

3.7.1 Método do Vizinho Mais Próximo (*Nearest Neighbor*)

O Método do Vizinho Mais Próximo é uma técnica simples e intuitiva de imputação, onde os valores faltantes são substituídos pelos valores mais próximos de outras amostras no conjunto de dados. Este método se baseia na premissa de que pontos próximos no espaço de características terão valores de atributos similares. No contexto deste trabalho, este método foi aplicado utilizando a métrica de distância euclidiana para determinar os vizinhos mais próximos. Este método é particularmente útil quando os dados apresentam uma estrutura espacial clara ou quando as variáveis apresentam correlações fortes com seus vizinhos mais próximos.

3.7.2 Regressão Linear

A Regressão Linear é uma técnica estatística amplamente utilizada para modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. Para a imputação de dados faltantes, a regressão linear foi empregada para prever os valores ausentes com base nas outras variáveis disponíveis. Este método é eficaz para dados que possuem uma relação linear entre as variáveis e é amplamente utilizado devido à sua simplicidade e interpretabilidade.

3.7.3 Redes Neurais de Camadas Densas

As Redes Neurais com Camadas Densas são um tipo de rede neural onde cada neurônio de uma camada está conectado a todos os neurônios da camada anterior. Este método é capaz de capturar relações não lineares complexas entre as variáveis, tornando-o particularmente poderoso para dados climáticos que frequentemente exibem tais relações. No contexto deste trabalho, uma rede neural foi treinada para prever os valores faltantes utilizando as variáveis disponíveis como entradas. A arquitetura da rede foi otimizada através de múltiplas camadas densas e a função de ativação ReLU (Rectified Linear Unit).

3.7.4 Complemento sobre Implementação

Os métodos foram aplicados aos dados climáticos coletados, envolvendo várias etapas cruciais para garantir a integridade e a precisão dos dados imputados:

Pré-processamento dos Dados: Incluiu a normalização dos dados, tratamento de outliers e transformação das variáveis para assegurar que os dados estivessem em um formato adequado para a aplicação dos métodos de completamento

Divisão dos Dados: O conjunto de dados foi dividido em 80% para treinamento e validação e 20% para teste. A divisão foi feita dessa forma para garantir que o modelo tivesse dados suficientes para aprender e ajustar seus parâmetros durante o treinamento e a validação, ao mesmo tempo que mantinha um conjunto separado para avaliar o desempenho final do modelo. Essa abordagem é amplamente utilizada para evitar *overfitting* e assegurar que os modelos tenham capacidade de generalização para dados não vistos, para uma comparação mais justa, o mesmo conjunto de treino e teste usado na regressão linear foi utilizado na camada densa.

Treinamento dos Modelos: Cada método foi treinado utilizando o conjunto de treinamento, ajustando os parâmetros para minimizar o erro de predição.

Avaliação do Desempenho: O desempenho dos métodos foi avaliado utilizando o RMSE, comparando os valores imputados com os valores reais presentes no conjunto de teste. Este processo permitiu identificar o método mais eficaz para cada variável climática.

4. RESULTADOS E DISCUSSÃO

4.1 DATASET

Utilizou-se uma base de dados importada do Instituto Nacional de Meteorologia (INMET) utilizando a API desenvolvida anteriormente, o período de abrangência e de 2000 a 2023. Para a análise específica dos dados, focou-se em uma estação meteorológica determinada. No entanto, ao analisar os métodos, considerou-se registros de todas as estações meteorológicas disponíveis no Brasil, no mesmo intervalo de tempo. Isso permitirá um preenchimento completo dos dados, levando em conta a diversidade de biomas e climas presentes no país.

4.2 ANÁLISE E ESTUDO DOS DADOS

Para esta análise de dados, o foco recai sobre as estações meteorológicas localizadas no estado de Goiás. Este estado possui 26 estações, das quais duas estão atualmente inoperantes. Concentrar-se em Goiás permite uma análise mais detalhada e específica, contribuindo para a compreensão das peculiaridades climáticas regionais.

Cada estação realiza registros horários, resultando em 24 medições diárias. Estas medições abrangem variáveis cruciais para a análise climática, como temperatura (°C), umidade relativa (%), pressão atmosférica (hPa), velocidade do vento (m/s), precipitação (mm), ponto de orvalho (°C) e radiação solar (KJ/m²). A diversidade e precisão dessas variáveis são fundamentais para um estudo climático abrangente e detalhado.

Os dados fornecidos pelo INMET são classificados como "Dados disponíveis em tempo real (sem controle de qualidade)", o que desempenha um papel crucial na análise e tratamento dessas informações. Esta classificação indica que, embora sejam atualizados e disponibilizados de maneira rápida e constante, os dados podem conter inconsistências ou erros, não tendo passado por um processo rigoroso de validação. Isso ressalta a necessidade de uma análise cuidadosa e detalhada dos dados coletados. A identificação de anomalias, como *outliers* ou dados faltantes, é essencial para garantir a confiabilidade e precisão das informações utilizadas.

Para uma análise prévia, foi selecionada uma amostra de dados da estação meteorológica de Goiânia, abrangendo o período de 2001 a 2023, totalizando 197.456 registros. Deste total, 12.000 registros apresentavam valores nulos em pelo menos

uma variável, e 5.614 registros eram totalmente nulos. Após a remoção desses dados faltantes, se obteve um *dataset* com 181.855 registros. Estas informações são detalhadas nas Figuras 6 e 7.

Figura 6 - Resumo conciso do DataFrame

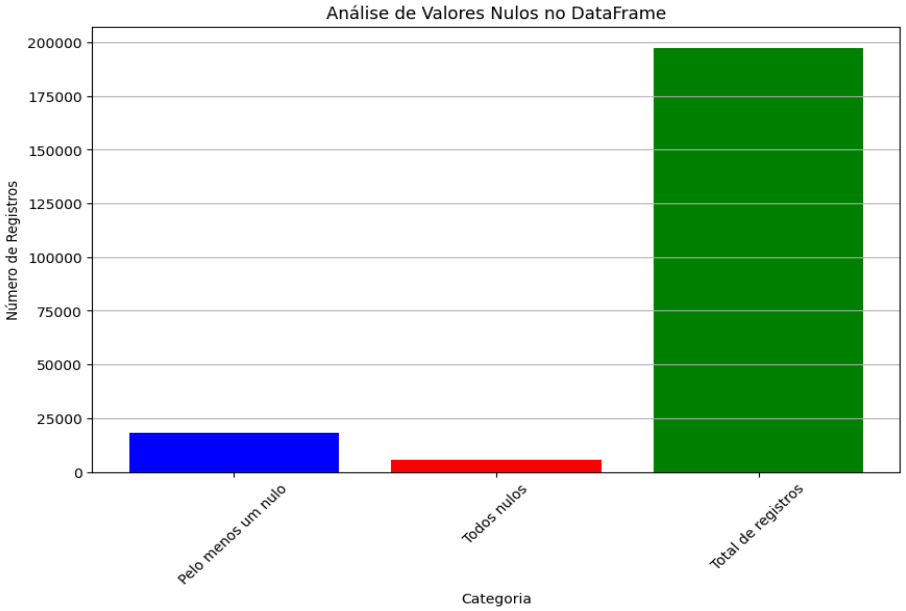
```
<bound method DataFrame.info of         pre_ins  tem_sem  rad_glo  pto_ins  ven_dir  chuva  ven_vel  ven_raj  \
0           NaN      NaN      NaN      NaN      NaN    NaN      NaN      NaN
1           NaN      NaN      NaN      NaN      NaN    NaN      NaN      NaN
2           NaN      NaN      NaN      NaN      NaN    NaN      NaN      NaN
3           NaN      NaN      NaN      NaN      NaN    NaN      NaN      NaN
4           NaN      NaN      NaN      NaN      NaN    NaN      NaN      NaN
...         ...      ...      ...      ...      ...    ...      ...      ...
197451    936.8     21.3   -3.500     11.8    105.0    0.0      0.3      2.3
197452    932.8      NaN   -3.153     18.9    140.0    0.0      0.7      2.3
197453    932.7      NaN   -3.360     18.2    213.0    0.0      0.7      3.3
197454    931.8      NaN   -3.473     18.2    184.0    0.0      0.3      2.3
197455    931.8      NaN   -2.958     18.4    122.0    0.0      1.6      4.6

        tem_ins  umd_ins                 hr_medicao
0           NaN      NaN  2001-05-28 00:00:00.000000
1           NaN      NaN  2001-05-28 01:00:00.000000
2           NaN      NaN  2001-05-28 02:00:00.000000
3           NaN      NaN  2001-05-28 03:00:00.000000
4           NaN      NaN  2001-05-28 04:00:00.000000
...         ...      ...                         ...
197451     21.2     55.0  2018-06-15 23:00:00.000000
197452     20.5     91.0  2023-12-06 03:00:00.000000
197453     19.9     90.0  2023-12-06 04:00:00.000000
197454     19.7     91.0  2023-12-06 05:00:00.000000
197455     19.7     92.0  2023-12-06 06:00:00.000000

[197456 rows x 11 columns]>
```

Fonte: Autoria própria

Figura 7 - Análise de dados faltantes com total de registros



Fonte: Autoria própria

Figura 8 - Análise estatística descritiva dos dados sem tratamento

	pre_ins	tem_sem	rad_glo	pto_ins	\
count	181855.000000	181855.000000	181855.000000	181855.000000	
mean	932.384564	23.853307	757.812752	15.474612	
std	3.659552	4.451359	1055.150869	4.641281	
min	889.100000	3.000000	-7.100000	-4.600000	
25%	930.500000	21.300000	-3.500000	12.200000	
50%	932.400000	23.900000	26.300000	17.000000	
75%	934.300000	26.800000	1493.550000	19.200000	
max	977.900000	62.800000	5533.500000	36.500000	

	ven_dir	chuva	ven_vel	ven_raj	\
count	181855.000000	181855.000000	181855.000000	181855.000000	
mean	175.015545	0.157002	1.224219	3.752546	
std	103.952854	1.340413	1.094365	2.339079	
min	1.000000	0.000000	0.000000	0.000000	
25%	96.000000	0.000000	0.300000	2.000000	
50%	140.000000	0.000000	1.000000	3.200000	
75%	280.000000	0.000000	1.800000	5.200000	
max	360.000000	72.000000	13.000000	23.200000	

	tem_ins	umd_ins
count	181855.000000	181855.000000
mean	23.720266	64.715906
std	4.892659	21.938784
min	5.700000	9.000000
25%	20.600000	48.000000
50%	23.100000	69.000000
75%	27.400000	84.000000
max	41.300000	100.000000

Fonte: Autoria própria

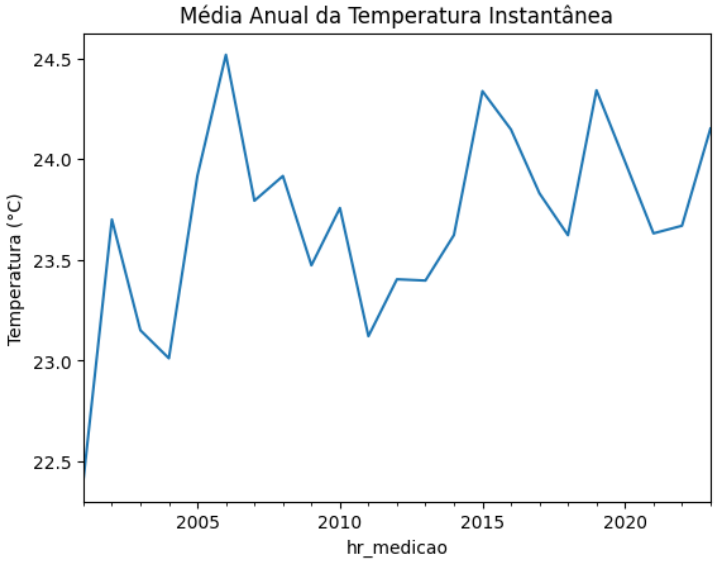
A análise estatística descritiva do conjunto de dados de Goiânia apresentada na Figura 8 revelou informações detalhadas sobre as condições climáticas da região entre 2001 e 2023. A pressão atmosférica (pre_ins) teve uma média de 932,38 hPa, refletindo as condições estáveis de pressão atmosférica na região. A temperatura de sensoriamento (tem_sem) registrou uma média de 23,85 °C, com um desvio padrão de 4,45 °C, indicando variações moderadas ao longo do período observado. Da mesma forma, a temperatura instantânea (tem_ins) teve uma média de 23,72 °C.

A radiação global (rad_glo) apresentou uma média de 757,81 Kj/m², ilustrando a intensa exposição solar característica da área. O ponto de orvalho (pto_ins), um indicador chave da umidade ambiental, teve uma média de 15,47 °C. A velocidade média do vento (ven_vel) foi de 1,22 m/s, com rajadas de vento (ven_raj) atingindo uma média de 3,75 m/s.

A precipitação (chuva) teve uma média de apenas 0,16 mm por registro, com variações significativas, sublinhando o padrão irregular de chuvas na região. A umidade instantânea (umd_ins) teve uma média de 64,71%, o que é crucial para entender o conforto e o clima na região. Esses *insights* são fundamentais para

compreender as condições climáticas em Goiânia, oferecendo uma base para análises mais aprofundadas e estudos futuros.

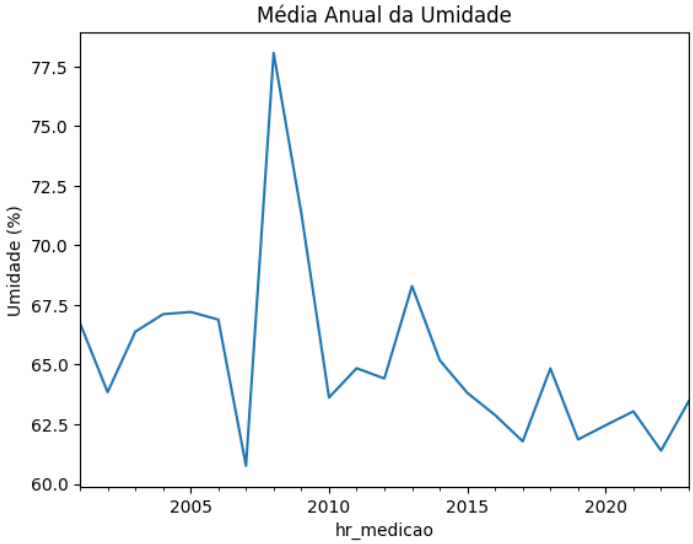
Figura 9 - Média anual da temperatura (2000-2023)



Fonte: Autoria própria

Quando se observa o gráfico da Figura 9 (média anual da temperatura instantânea), nota-se flutuações anuais que são típicas do clima dinâmico. Apesar dessas oscilações naturais, uma inspeção mais detalhada sugere uma possível tendência ascendente de leve aquecimento ao longo do período de 22 anos. Especificamente, os primeiros anos do intervalo mostram temperaturas médias anuais um pouco mais baixas quando comparadas aos últimos anos. Essa percepção visual preliminar pode indicar um aumento gradual na temperatura média, sugerindo que Goiânia pode estar experimentando uma tendência de aquecimento ao longo das últimas duas décadas. Contudo, para uma afirmação definitiva, seria necessário realizar uma análise de regressão linear para quantificar e verificar a significância estatística dessa tendência.

Figura 10 - Média anual da Umidade (2000-2023)



Fonte: Autoria própria

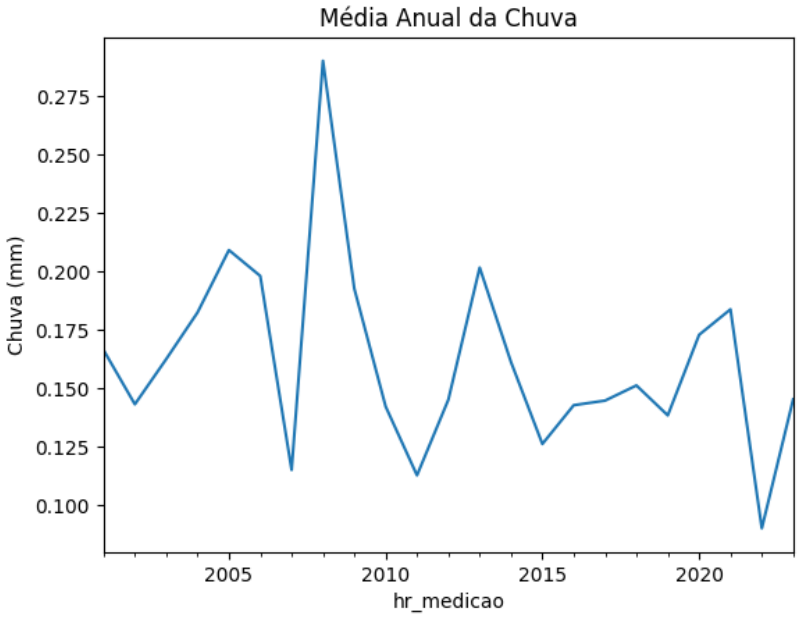
Ao observar a tendência da umidade instantânea (Figura 10) ao longo de duas décadas, ocorre um padrão que sugere uma diminuição gradual da umidade anual média em Goiânia. A partir de uma média de 66,74% em 2001, a umidade exibe uma tendência de declínio, intercalada por anos de aumento esporádico, como evidenciado pelo pico em 2008. Esta variação pode ser atribuída a fatores como variações sazonais, alterações nos padrões climáticos locais ou mesmo mudanças mais amplas associadas a fenômenos globais como o El Niño ou La Niña.

Notavelmente, os anos mais recentes da série, particularmente 2017 e 2022, apresentam valores mais baixos de umidade, com médias de 61,78% e 61,39%, respectivamente, indicando que as condições podem ter se tornado progressivamente mais secas. Esta observação é especialmente relevante quando considera-se a relação inversa com a temperatura, sugerindo que os aumentos na temperatura poderiam estar correlacionados com a diminuição da umidade. É uma relação comum em climatologia onde o ar mais quente tende a reter menos umidade, resultando em uma menor umidade relativa.

A análise da média anual de precipitação na Figura 11 revela uma variação significativa ao longo dos anos, refletindo a complexidade dos padrões de chuva na região. Iniciando com uma média de 0,166 mm em 2001, os dados mostram uma tendência de flutuação que alcança um pico expressivo de 0,290 mm em 2008. Este valor pode ser atribuído a eventos climáticos específicos que resultaram em um aumento acentuado das chuvas naquele ano.

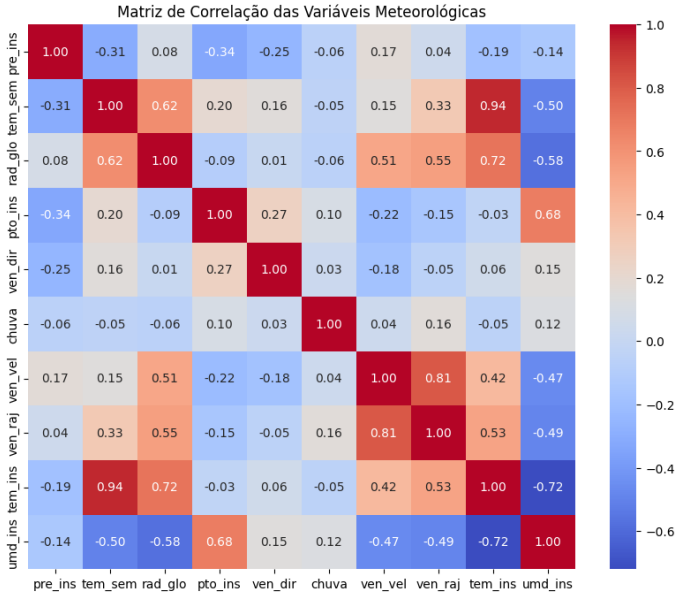
Após o pico de 2008, observa-se uma tendência de diminuição com anos subsequentes apresentando menores médias anuais, culminando em um mínimo de 0,090 mm em 2022. Este decréscimo na precipitação ao longo dos anos mais recentes pode ser um indicador de mudanças nos padrões climáticos ou variações interanuais naturais.

Figura 11 - Média anual da chuva (2000-2023)



Fonte: Autoria própria

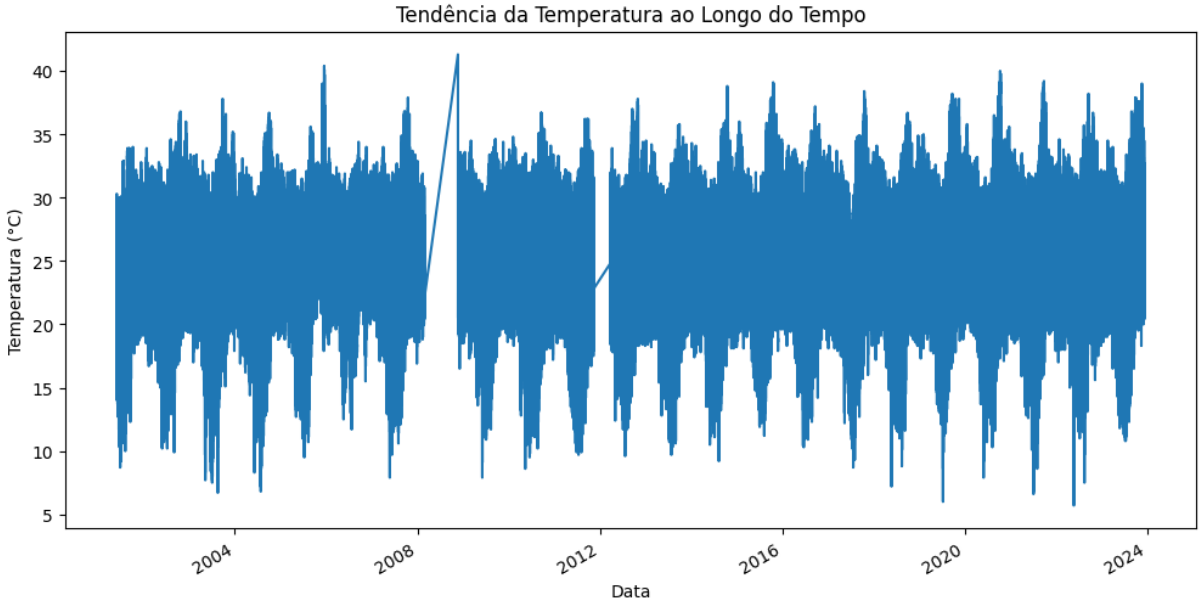
Figura 12 - Correlação entre variáveis do DataSet



Fonte: Autoria própria

Com base nessas observações da Figura 12, conclui-se que as variáveis de temperatura estão inter-relacionadas e inversamente relacionadas à umidade, como é comumente compreendido na meteorologia. A falta de correlações fortes entre precipitação e outras variáveis meteorológicas sugere que os padrões de chuva na região podem ser determinados por uma combinação complexa de fatores locais e possivelmente por eventos de larga escala que não são capturados apenas por estas variáveis.

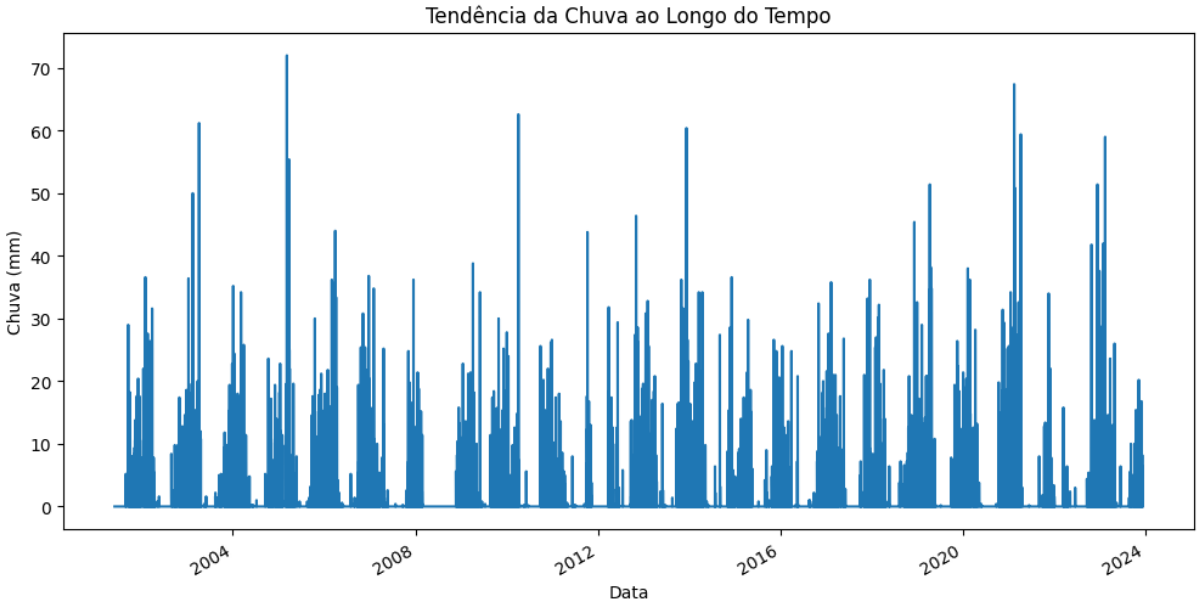
Figura 13 - Tendencia da temperatura ao longo do tempo (2000-2023)



Fonte: Autoria própria

O gráfico de tendências da Figura 13 da temperatura ao longo do tempo exibe uma variação sazonal pronunciada, refletindo o ciclo de altas e baixas temperaturas que correspondem às mudanças sazonais esperadas. Os padrões observados no gráfico indicam uma consistência nos dados, com temperaturas retornando a faixas semelhantes em intervalos regulares, o que sugere uma previsibilidade no comportamento climático da região. No entanto, alguns picos e vales extremos são notáveis e podem ser indicativos de eventos climáticos atípicos, como ondas de calor ou períodos de frio intenso. Não obstante as oscilações anuais evidentes, não é possível discernir uma tendência de longo prazo de aumento ou diminuição na temperatura sem uma análise estatística mais detalhada. Para confirmar a existência de tendências sustentadas ao longo das duas décadas representadas, seria necessário empregar técnicas analíticas adicionais, como a aplicação de linhas de tendência ou modelos de médias móveis, que poderiam oferecer uma compreensão mais clara das mudanças climáticas ao longo do tempo.

Figura 14 - Tendencia da chuva ao longo do tempo (2000-2023)

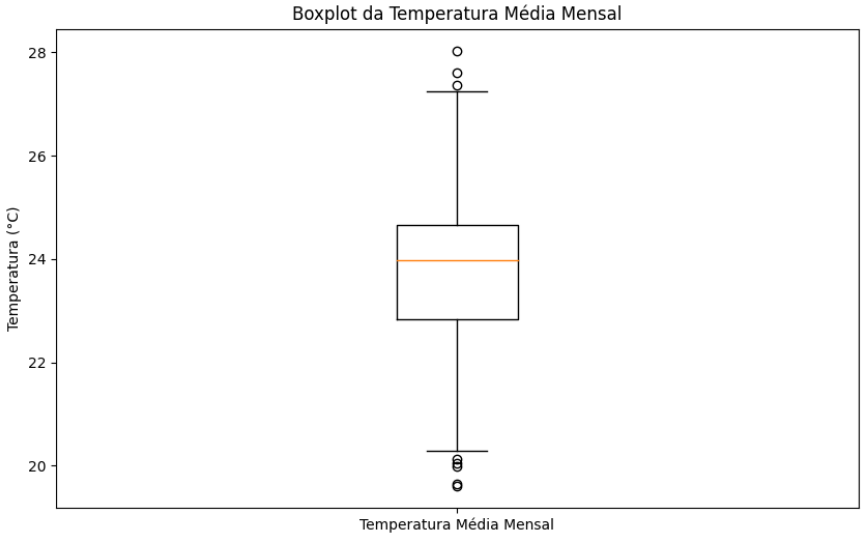


Fonte: Autoria própria

O gráfico da tendência da chuva ao longo do tempo em Goiânia na Figura 14 destaca a natureza intermitente e variável da precipitação na área. As barras altas, representando picos de precipitação, são indicativos de eventos de chuva mais intensos, que podem estar associados a temporadas de chuvas ou a eventos climáticos específicos, como tempestades. Por outro lado, os períodos com barras mais baixas sugerem temporadas mais secas. Uma análise mais aprofundada, talvez

utilizando técnicas de suavização como médias móveis, poderia oferecer uma compreensão mais clara da variação da precipitação e revelar tendências que não são imediatamente aparentes em um gráfico de barras.

Figura 15 - Boxplot da temperatura Media Mensal

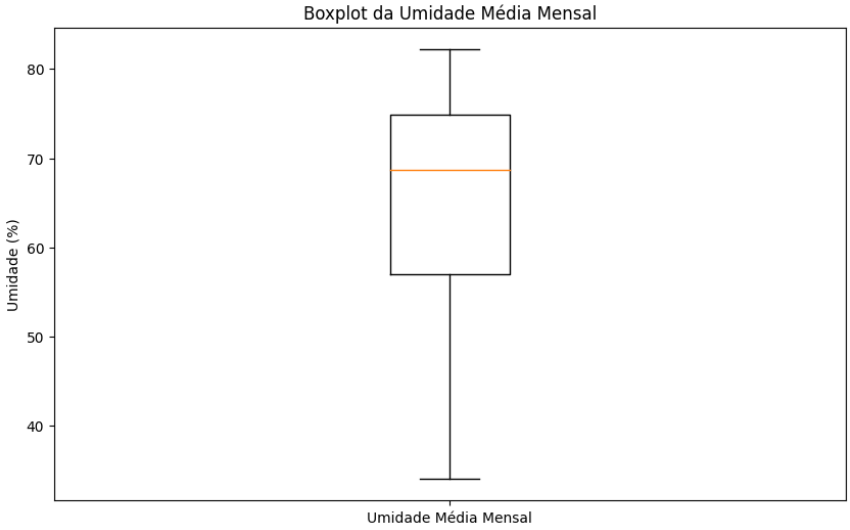


Fonte: Autoria própria

O *boxplot* (Figura 15) da temperatura média mensal fornece uma visão clara das variações climáticas em Goiânia, suavizando as flutuações diárias e destacando as tendências mais relevantes para uma análise climática. Observa-se que a mediana das temperaturas mensais, indicada pela linha laranja, está em torno dos 24°C, o que pode ser considerado típico para a região. A caixa do *boxplot*, representando o intervalo interquartílico, mostra a maioria das médias mensais de temperatura concentradas entre aproximadamente 22°C e 26°C. Isso reflete a variação esperada de temperatura ao longo dos meses, levando em conta a mudança das estações.

Os "bigodes" do *boxplot* estendem-se de cerca de 20°C a quase 28°C, abrangendo a gama total de variações mensais típicas sem considerar valores extremos. No entanto, pontos fora dos bigodes, os chamados outliers, são notáveis tanto na extremidade inferior quanto na superior do gráfico. Estes representam meses com médias de temperatura excepcionalmente altas ou baixas, que desviam significativamente do padrão geral. Esses outliers podem ser indicativos de condições climáticas atípicas, como ondas de calor ou frio inesperado para o período.

Figura 16 – Boxplot da temperatura Media Mensal



Fonte: Autoria própria

O *boxplot* (Figura16) da umidade média mensal apresenta uma visão consolidada da variação da umidade em Goiânia, com a mediana situada aproximadamente nos 70%, indicando que a maior parte do tempo a região experimenta um nível de umidade relativa considerável. A caixa em si, que representa o intervalo interquartílico, estende-se de cerca de 60% a 80%, refletindo uma distribuição concentrada da umidade ao longo do ano, com 50% das observações mensais situando-se dentro desta faixa.

Os bigodes do *boxplot*, que se estendem de aproximadamente 50% a um pouco acima de 80%, indicam a variação total da umidade dentro do alcance típico, excluindo valores extremos. O gráfico não mostra pontos acima ou abaixo dos bigodes, o que sugere que não há meses com umidade extremamente alta ou baixa que sejam considerados outliers no contexto do conjunto de dados.

4.3 MÉTODO DO VIZINHO MAIS PRÓXIMO (NEAREST NEIGHBOR METHOD)

O Método do Vizinho Mais Próximo (*Nearest Neighbor Method*) é implementado considerando a menor distância entre duas estações meteorológicas. Este método imputa dados faltantes na série temporal observada a partir da estação mais próxima, calculando a distância com a Fórmula de *Vincenty*, que leva em consideração a forma elipsoidal da Terra para obter medições precisas. Após a convergência dos cálculos, os dados do vizinho mais próximo são sincronizados pela hora e dia de medição para garantir consistência.

A Fórmula de *Vincenty* leva em consideração a forma elipsoidal da Terra, ao contrário dos métodos que assumem uma Terra esférica. A Terra é mais achatada nos polos e mais larga no equador, e essa forma precisa ser considerada para obter medições precisas. A equação utiliza uma série de iterações para chegar à distância correta entre dois pontos definidos por suas latitudes e longitudes. Sendo definido pela equação matemática:

1. Inicialize $\lambda = L$, onde

$$L = \lambda_2 - \lambda_1.$$

2. Itere até a convergência de λ:

$$\begin{aligned}
U_1 &= \arctan\left((1-f)\tan\phi_1\right) \\
U_2 &= \arctan\left((1-f)\tan\phi_2\right) \\
\sin\sigma &= \sqrt{(\cos U_2 \sin\lambda)^2 + (\cos U_1 \sin U_2 - \sin U_1 \cos U_2 \cos\lambda)^2} \\
\cos\sigma &= \sin U_1 \sin U_2 + \cos U_1 \cos U_2 \cos\lambda \\
\sigma &= \arctan\left(\frac{\sin\sigma}{\cos\sigma}\right) \\
\sin\alpha &= \frac{\cos U_1 \cos U_2 \sin\lambda}{\sin\sigma} \\
\cos^2\alpha &= 1 - \sin^2\alpha \\
\cos 2\sigma_m &= \cos\sigma - \frac{2\sin U_1 \sin U_2}{\cos^2\alpha} \\
C &= \frac{f}{16}\cos^2\alpha\left(4 + f\left(4 - 3\cos^2\alpha\right)\right) \\
\lambda &= L + (1-C) f \sin\alpha \left(\sigma + C\sin\sigma\left(\cos 2\sigma_m + C\cos\sigma\left(-1 + 2\cos 2\sigma_m^2\right)\right)\right)
\end{aligned}$$

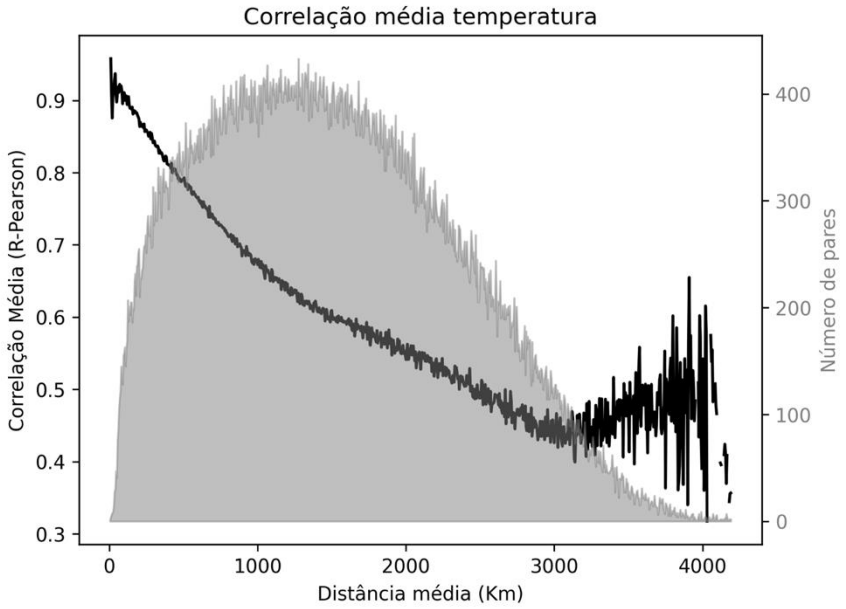
3. Após a convergência de λ, calcule a distância s:

$$\begin{aligned}
u^2 &= \cos^2\alpha\left(\frac{a^2 - b^2}{b^2}\right) \\
A &= 1 + \frac{u^2}{16384}\left(4096 + u^2\left(-768 + u^2\left(320 - 175u^2\right)\right)\right) \\
B &= \frac{u^2}{1024}\left(256 + u^2\left(-128 + u^2\left(74 - 47u^2\right)\right)\right) \\
\Delta\sigma &= B\sin\sigma\left(\cos 2\sigma_m + \frac{B}{4}\left(\cos\sigma\left(-1 + 2\cos 2\sigma_m^2\right) - \frac{B}{6}\cos 2\sigma_m\left(-3 + 4\sin\sigma^2\right)\right.\right. \\
s &= bA(\sigma - \Delta\sigma)
\end{aligned}$$

As estações meteorológicas estão distribuídas por todo o território brasileiro de maneira não uniforme, resultando em variações significativas nas distâncias entre elas. Realizou-se uma avaliação de correlação entre os dados de cada estação para

determinar até que distância (em quilômetros) se pode utilizar os dados sem comprometer significativamente a qualidade da análise.

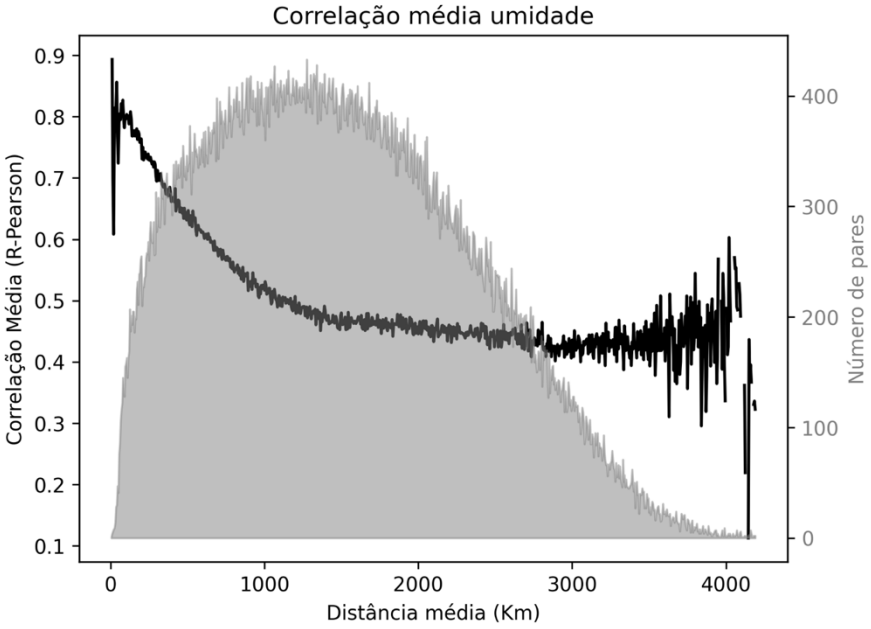
Figura 17 – Correlação média temperatura



Fonte: Autoria própria

O gráfico da Figura 17 apresentado ilustra a correlação média da temperatura entre estações meteorológicas em função da distância média entre elas. Nota-se que, para distâncias curtas, a correlação é alta, próxima de 0.9, indicando que estações próximas possuem dados de temperatura muito semelhantes. Com o aumento da distância, a correlação diminui gradualmente, evidenciando a redução da influência de fatores locais específicos. Após cerca de 2000 km, a correlação começa a cair abaixo de 0.6, sugerindo que além dessa distância, a similaridade nos dados de temperatura entre as estações se torna menos significativa. Este padrão reflete a importância das estações próximas para a obtenção de dados de temperatura confiáveis, indicando que a utilidade das estações vizinhas para a temperatura diminui significativamente além de aproximadamente 1000-1500 km.

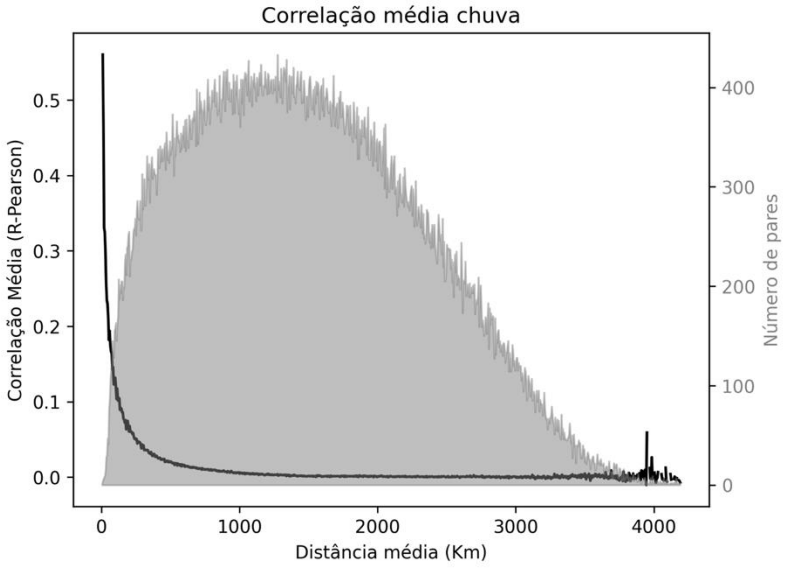
Figura 18 – Correlação média umidade



Fonte: Autoria própria

O gráfico da Figura 18 apresentado ilustra a correlação média da umidade entre estações meteorológicas em função da distância média entre elas. Observa-se que, para distâncias curtas, a correlação é alta, próxima de 0,8, indicando que estações próximas possuem dados de umidade bastante semelhantes. À medida que a distância aumenta, a correlação diminui de forma gradual, evidenciando a redução da influência de fatores locais específicos. Em distâncias superiores a aproximadamente 2000 km, a correlação começa a cair abaixo de 0.6, sugerindo que além dessa distância, a similaridade nos dados de umidade entre as estações se torna menos significativa. Assim, para a umidade, a utilidade dos dados das estações vizinhas diminui significativamente além de aproximadamente 1000-1500 km, onde a correlação cai abaixo de 0,6, indicando uma perda notável de similaridade nos dados.

Figura 19 - Correlação média chuva

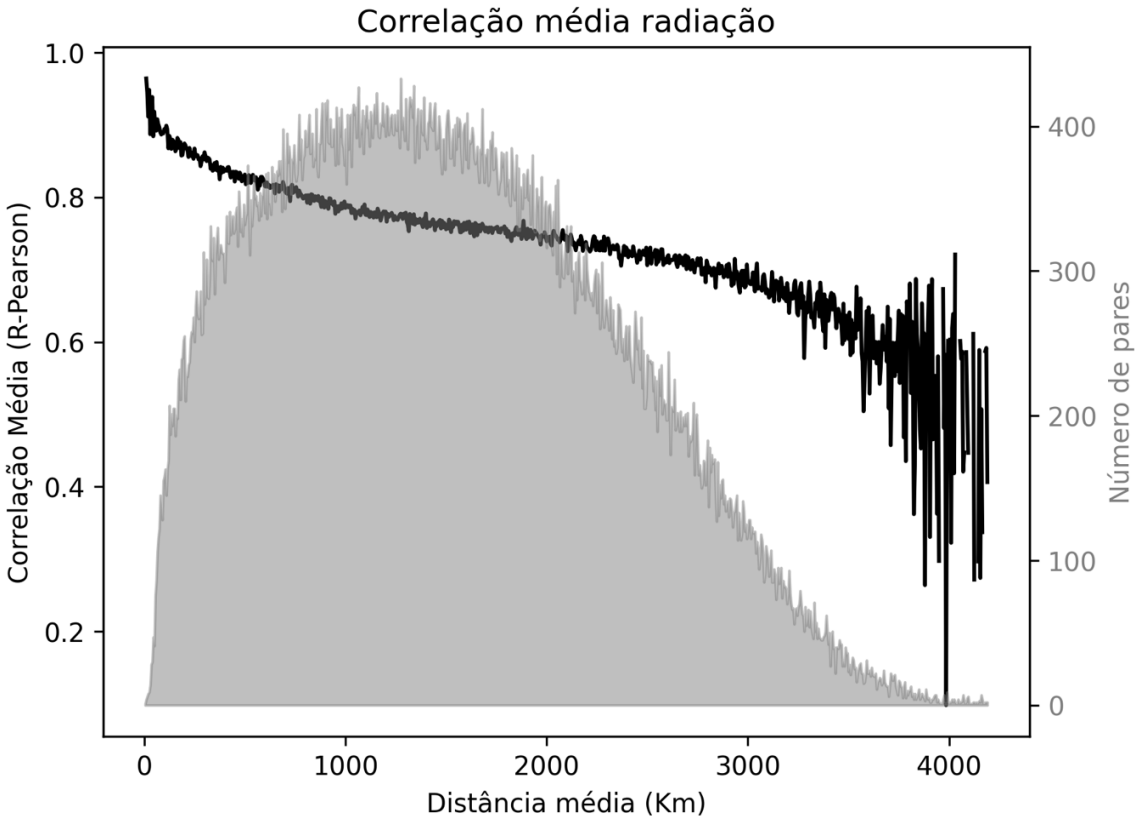


Fonte: autoria própria

O gráfico da Figura 19 mostra a correlação média da chuva entre estações meteorológicas em função da distância média entre elas. Observa-se que, para distâncias muito curtas, a correlação é moderada, próxima de 0.5, mas rapidamente cai à medida que a distância aumenta, atingindo valores próximos de 0 para distâncias superiores a 500 km. Este comportamento indica que os dados de precipitação são altamente localizados e que a similaridade entre os dados de estações distantes é muito baixa.

Essa rápida queda na correlação pode ser explicada pelo fato de que a precipitação não é um fenômeno geoestacionário. Diferente de variáveis como pressão e temperatura, que podem apresentar padrões mais amplos e contínuos sobre grandes áreas, a chuva é um evento altamente variável e localizado. Tempestades, frentes frias e outros eventos meteorológicos que causam chuva podem ocorrer de forma isolada e com grande variabilidade espacial. Portanto, a utilidade dos dados de precipitação de uma estação vizinha é muito limitada, geralmente se restringindo a distâncias muito curtas, tipicamente abaixo de 500 km. Essa característica torna essencial a instalação de uma rede densa de estações meteorológicas para monitorar a chuva com precisão em diferentes regiões.

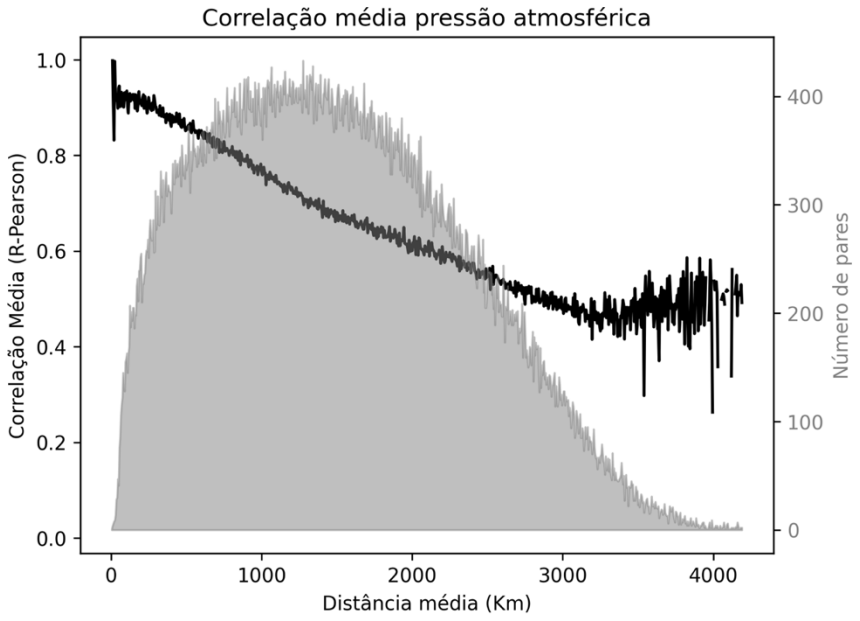
Figura 20 – Correlação média radiação



Fonte: Autoria própria

O gráfico da Figura 20 mostra a correlação média da radiação entre estações meteorológicas em função da distância média entre elas. Observa-se que, para distâncias curtas, a correlação é muito alta, próxima de 1, indicando que estações próximas têm medições de radiação solar muito semelhantes. À medida que a distância aumenta, a correlação diminui gradualmente, mantendo-se acima de 0.8 até cerca de 2000 km. Isso sugere que a radiação solar é uma variável relativamente consistente em médias distâncias, possivelmente devido a fatores climáticos regionais que afetam amplamente a quantidade de radiação recebida.

Figura 21 - Correlação média pressão atmosférica



Fonte: Autoria própria

O gráfico da Figura 21 apresenta a correlação média da pressão atmosférica entre estações meteorológicas em função da distância média entre elas. Observa-se que, para distâncias muito curtas, a correlação é alta, próxima de 1, indicando que as estações próximas têm medições de pressão muito semelhantes. À medida que a distância aumenta, a correlação diminui gradualmente.

A correlação média se mantém relativamente alta até cerca de 1000 km, após o que começa a cair de forma mais acentuada. Em distâncias acima de aproximadamente 3000 km, a correlação estabiliza em valores mais baixos, sugerindo que além dessa distância, as estações meteorológicas apresentam comportamentos de pressão atmosférica mais independentes entre si.

Essa tendência indica que a utilidade dos dados de pressão atmosférica de estações vizinhas diminui significativamente além de cerca de 1000-1500 km. A pressão atmosférica tende a ser mais influenciada por padrões climáticos regionais que podem se estender por grandes áreas, mas que eventualmente se tornam menos correlacionados em distâncias muito grandes devido às variações geográficas e climáticas específicas de cada região.

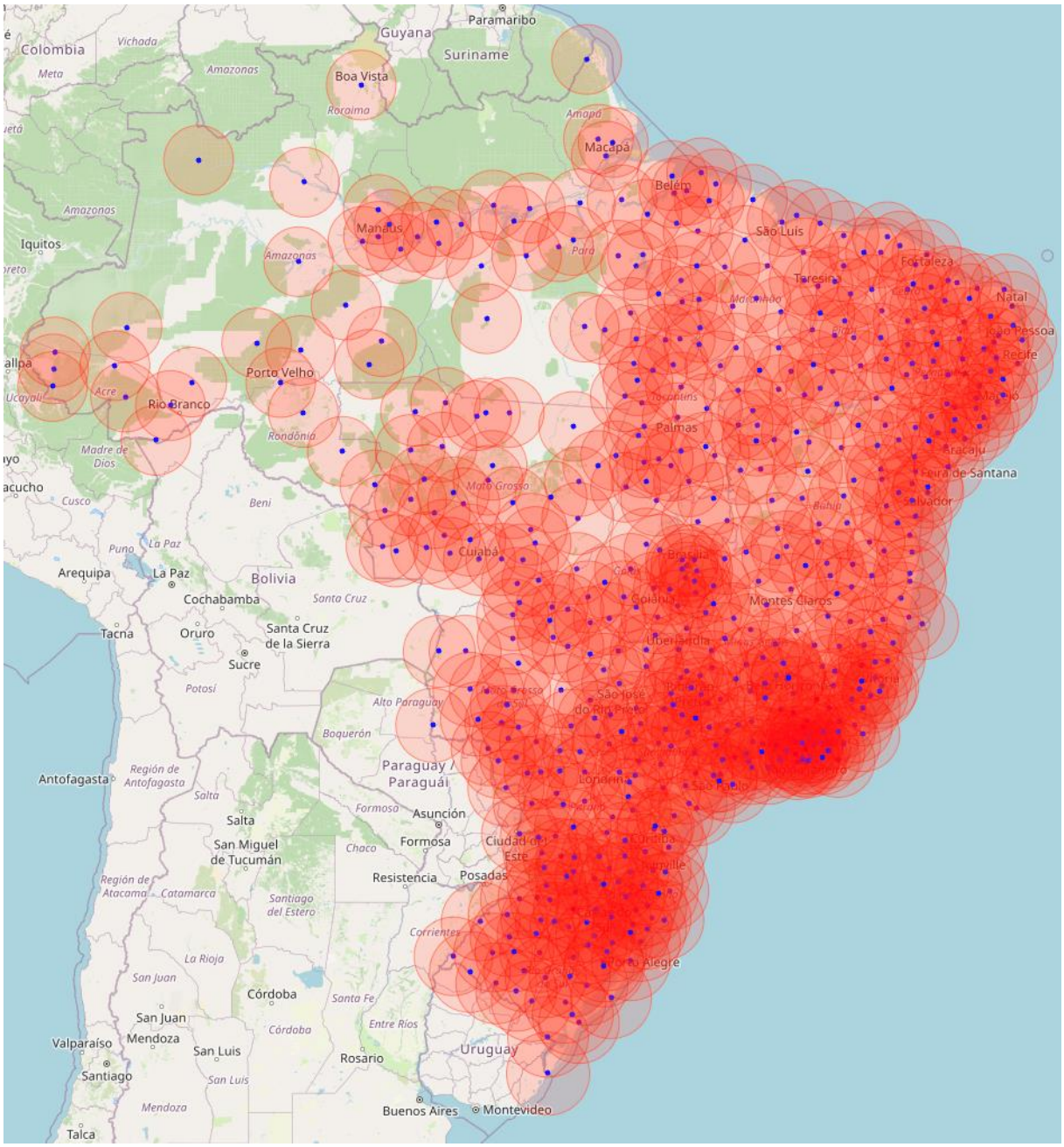
A análise dos gráficos de correlação média para pressão atmosférica, temperatura, umidade, chuva e radiação entre estações meteorológicas em função da

distância revela um padrão consistente: a correlação entre os dados diminui à medida que a distância entre as estações aumenta. Observa-se que, para a maioria das variáveis, a correlação é alta para distâncias curtas (até cerca de 150 km), e começa a diminuir de forma significativa conforme a distância aumenta. Este comportamento indica que a proximidade das estações é crucial para garantir a precisão e a similaridade dos dados meteorológicos.

Com base na análise de correlação, decidiu-se utilizar apenas dados de estações meteorológicas situadas a até 150 km de distância umas das outras, onde a correlação média para todas as variáveis analisadas se mantém em níveis altos (aproximadamente 0.76). Esta escolha é justificada pelo fato de que, dentro desse intervalo, a correlação média para todas as variáveis analisadas se mantém em níveis altos, assegurando a consistência e a comparabilidade dos dados. Além dessa distância, a correlação cai significativamente, o que poderia comprometer a qualidade das análises devido à variabilidade regional e local dos fenômenos meteorológicos. Portanto, restringir a análise a estações próximas garantirá dados mais confiáveis e representativos para o estudo.

O gráfico da Figura 22 ilustra a localização das estações meteorológicas no Brasil, cada uma com um raio de cobertura de 150 km destacado em vermelho. Este raio representa a área dentro da qual se consideram as estações vizinhas para análises de correlação. Observa-se que, apesar deste raio de 150 km, algumas estações não possuem nenhuma outra estação vizinha dentro de sua área de cobertura totalizando um total de 48 estações sem vizinhos. Devido à falta de estações vizinhas para essas áreas, optou-se por remover essas estações do conjunto de dados restando então um total de 519 estações, garantindo assim que todas as estações restantes tenham vizinhas suficientes para análises confiáveis e robustas.

Figura 22 - Distribuição das Estações Meteorológicas com Raio de Cobertura de 150 km



Fonte: Autoria própria

Com o cálculo da distância entre cada estação e a distância máxima permitida segundo a análise de correlação entre as estações para as variáveis observadas, pôde-se implementar o *Nearest Neighbor Method* (Método do Vizinho Mais Próximo). Considerou-se todo o intervalo de dados dos últimos 23 anos, onde, para cada estação, foi capturado os dados do vizinho mais próximo, sincronizando os dados pela

hora e dia de medição de ambas as estações. Este método pressupõe que não é possível prever um dado sem que ele exista no vizinho. Assim, para cada estação x, buscou-se a estação *y* até 150 km de distância e com a menor diferença de altitude entre elas, considerando que a altitude em relação ao nível do mar pode influenciar significativamente o clima observado. Este método não requer processos adicionais, bastando utilizar um método para avaliar o erro. Para comparações, utilizou-se o RMSE (*Root Mean Square Error*), obtendo os seguintes resultados.

Tabela 4 - Desempenho do Método do Vizinho Mais Próximo

Variáveis observadas	RMSE
Temperatura	0.347
Umidade	0.522
Chuva	1.298
Pressão atmosferica	0.233
Radiação Solar	0.756

Fonte: Autoria propria

A Tabela 4 apresenta os valores de RMSE (*Root Mean Square Error*) para diferentes variáveis meteorológicas: Temperatura, Umidade, Chuva, Pressão Atmosférica e Radiação Solar. Os resultados mostram que a menor margem de erro foi observada para a pressão atmosférica (0.233), enquanto a maior margem de erro foi para a chuva (1.298). Estes dados indicam a precisão do método NN (Método do Vizinho Mais Próximo) para cada variável, destacando variações na eficácia dependendo do tipo de dado meteorológico.

No entanto, é importante ressaltar que este método só pode ser utilizado quando existem registros no período desejado para o completamento dos dados. Além disso, deve-se considerar que as estações meteorológicas foram sendo instaladas ao longo do tempo. A estação mais recente iniciou suas atividades em 7 de dezembro de 2022, localizada em Porto Alegre - Belém Novo, no estado do Rio Grande do Sul. Esta evolução temporal das estações pode impactar a disponibilidade e a continuidade dos dados ao longo dos anos. Consequentemente, embora o método NN forneça bons resultados, ele só é capaz de completar os dados ausentes para períodos em que há registros existentes nas estações vizinhas. Portanto, não é possível montar uma série

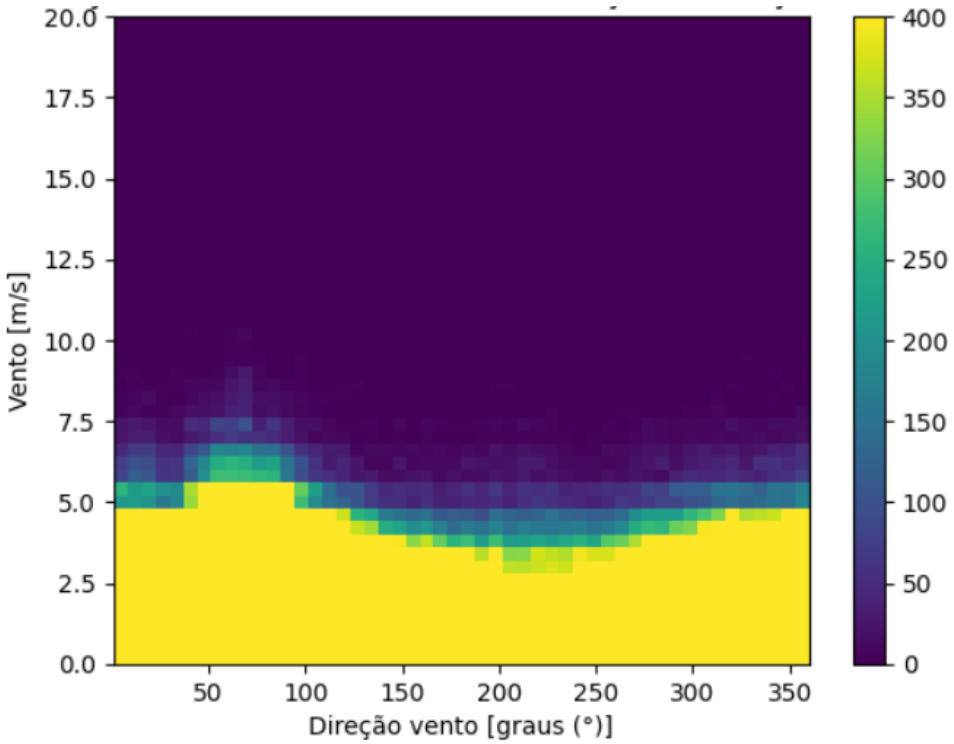
temporal histórica completa se não houver dados disponíveis para todos os períodos analisados.

4.4 TRATAMENTO DE VARIÁVEIS

4.4.1 Vento e direção do vento

Para as próximas análises deve-se ter um pouco de atenção sobre algumas variáveis da base, em especifico a velocidade e a direção do vento. É fundamental considerar como esses dados são representados e utilizados em modelos de aprendizado de máquina, como regressão linear e redes neurais de camada densas.

Figura 23 – Distribuição da Velocidade do Vento em Função da Direção do Vento



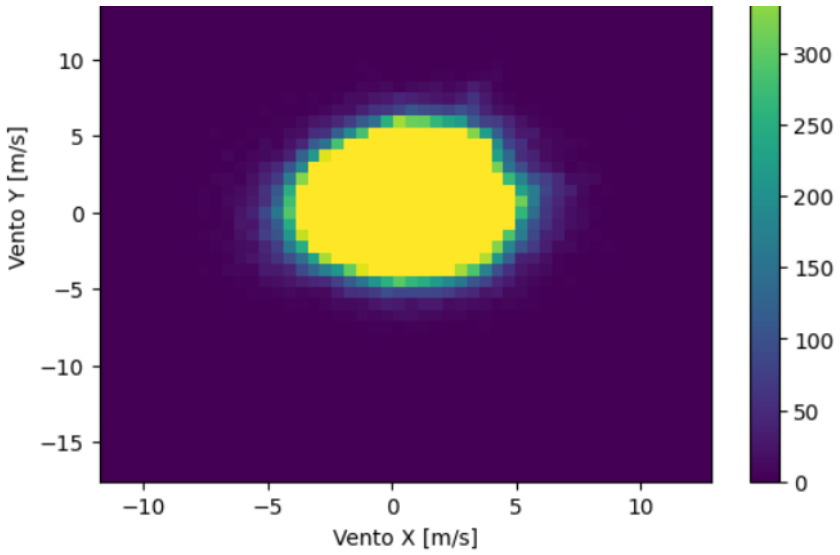
Fonte: Autoria própria

Na Figura 23 se pode notar que quando o vento não está soprando (velocidade de 0 m/s), a direção medida é essencialmente sem significado. No gráfico, isso pode ser observado na linha horizontal inferior, onde a direção do vento apresenta variação, mas a velocidade permanece nula. Esta característica destaca a importância de considerar a velocidade do vento antes de analisar a direção, pois uma direção de

vento com velocidade zero não contribui para a dinâmica do vento e pode ser considerada ruído nos dados.

Essa inconsistência pode confundir o modelo e afetar negativamente a qualidade das previsões. Transformar a velocidade e a direção do vento em componentes cartesianas (`Wx` e `Wy`) resolve esses problemas ao linearizar as relações e eliminar a periodicidade, proporcionando dados mais consistentes e interpretáveis para o modelo.

Figura 24 – Distribuição dos Componentes do Vento após Transformação para Coordenadas Cartesianas



Fonte: Autoria própria

A transformação dos dados de vento em componentes cartesianas representa uma melhoria significativa na análise e modelagem dos dados. O gráfico resultante oferece uma visualização clara e direta das distribuições dos componentes do vento, eliminando a complexidade associada à direção cíclica e à interação não linear entre velocidade e direção. Este aprimoramento não apenas facilita a interpretação dos dados, mas também melhora o desempenho de modelos de regressão linear e redes neurais de camada densas, tornando a análise mais robusta e eficiente.

4.4.2 Data e hora

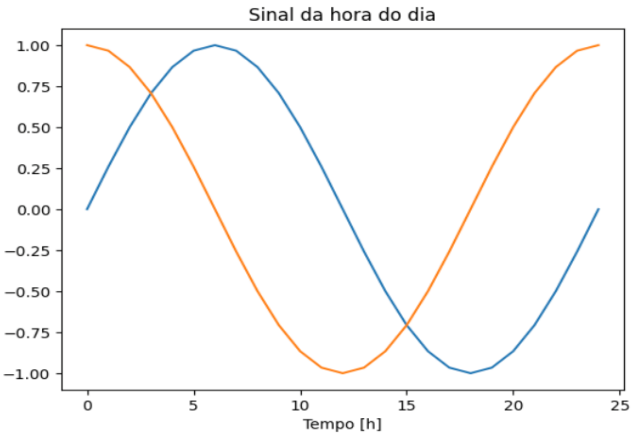
Assim como a direção do vento, a coluna *`Date Time`* contém informações valiosas que influenciam variáveis como radiação, temperatura e umidade. No Brasil, onde as estações climáticas são bem definidas, o impacto da hora e do dia é

significativo. No entanto, a informação de `Date Time` no formato de *string* não é diretamente útil para os modelos de aprendizado de máquina. Portanto, o primeiro passo é converter esses valores de data e hora para segundos, transformando-os em uma escala numérica contínua. Mesmo assim, utilizar o tempo em segundos como entrada para o modelo pode não ser a melhor abordagem, já que os dados meteorológicos exibem periodicidades claras, tanto diárias quanto anuais.

Para lidar com essa periodicidade, uma técnica eficaz é usar transformações de seno e cosseno. Essas transformações permitem que o modelo capture as variações periódicas ao longo do dia e do ano de maneira mais eficiente. Especificamente, os sinais de seno e cosseno para "Hora do dia" são calculados para refletir a variação dentro de um dia completo (24 horas), enquanto os sinais de "Hora do ano" são calculados para refletir a variação dentro de um ano completo. Isso cria novas colunas no conjunto de dados que representam a "Hora do dia" e a "Hora do ano" de forma cíclica obtendo *'Day sin', 'Day cos', 'Year sin', 'Year cos'*.

Transformar o tempo dessa maneira permite que o modelo identifique padrões sazonais e diurnos, melhorando significativamente sua capacidade de fazer previsões precisas. Esta abordagem é particularmente poderosa porque incorpora conhecimentos pré-existentes sobre a natureza periódica dos dados meteorológicos, garantindo que o modelo tenha acesso aos sinais de frequência mais importantes desde o início.

Figura 25 - Sinais transformados



Fonte: Autoria própria

Ao visualizar esses sinais transformados, na Figura 25, fica evidente que eles fornecem uma representação clara e contínua das variações temporais, facilitando a interpretação e a análise pelo modelo. Em resumo, essa transformação permite que o modelo de aprendizado de máquina capture de forma eficiente e precisa as variações

temporais essenciais presentes nos dados meteorológicos, resultando em previsões mais robustas e confiáveis.

4.4.3 Normalização

A normalização dos *datasets* é um passo essencial no pré-processamento de dados para modelos de aprendizado de máquina. A normalização é o processo de ajustar os valores das variáveis em uma escala comum, o que é particularmente importante quando os dados possuem diferentes unidades de medida e magnitudes. A falta de normalização pode levar a modelos enviesados ou ineficientes, pois variáveis com valores maiores podem dominar a função de perda e afetar desproporcionalmente o processo de treinamento do modelo.

A normalização transforma os dados para que todas as variáveis contribuam de forma equitativa na análise, melhorando a convergência durante o treinamento de modelos como redes neurais e regressão linear. Existem várias técnicas de normalização, como a normalização min-max, que ajusta os valores para um intervalo específico (geralmente 0 a 1), e a padronização *z-score*, que transforma os dados para que tenham média zero e desvio padrão um. Essas transformações garantem que todas as variáveis sejam tratadas de forma uniforme pelo modelo, facilitando a comparação e a integração de diferentes tipos de dados.

No contexto de análise climática, aplicou-se a padronização *z-score* às variáveis temperatura, umidade, chuva, pressão e radiação. Estas variáveis possuem escalas diferentes e unidades de medida variadas. Por exemplo, enquanto a temperatura é medida em graus Celsius, a precipitação pode ser medida em milímetros, e a pressão em hectopascais. Sem a normalização, a diferença nas escalas poderia fazer com que o modelo atribuísse mais importância às variáveis com magnitudes maiores, como a pressão, em detrimento das variáveis com magnitudes menores, como a radiação.

Normalizar essas variáveis permite que o modelo aprenda de forma mais eficiente e equilibrada. A normalização também facilita a interpretação dos coeficientes e pesos do modelo, uma vez que todos os parâmetros estão na mesma escala. Isso é crucial para garantir que o modelo não seja enviesado por variáveis com valores mais altos e que todas as variáveis possam influenciar o modelo de forma justa.

4.4.4 Separação dos dados

A separação dos dados em conjuntos de treino e teste é um passo crítico no desenvolvimento e avaliação de modelos de aprendizado de máquina. Para garantir que o modelo seja capaz de generalizar bem para novos dados e evitar o *overfitting*, é essencial avaliar seu desempenho em um conjunto de dados que não foi utilizado durante o treinamento.

No caso especificamente, dividiu-se o *dataset* em variáveis preditoras (*X*) e a variável alvo (*Y*). A proporção de 20% dos dados foi reservada para o conjunto de teste. Esta escolha é bastante comum e balanceia bem a necessidade de ter um conjunto de dados suficiente para treinar o modelo e um conjunto separado suficiente para avaliar o seu desempenho. Reservar 20% dos dados para o teste garante que se tem uma amostra representativa do *dataset* original para validar o modelo.

Separar os dados em conjuntos de treino e teste é fundamental para avaliar o desempenho real do modelo. Treinar o modelo apenas com o conjunto de treino permite que ele ajuste seus parâmetros para melhor prever os dados de treino. Em seguida, ao avaliar o modelo com o conjunto de teste pode medir o desempenho do modelo em dados não vistos anteriormente. Isso fornece uma estimativa precisa de como o modelo se comportará em dados reais futuros.

4.5 REGRESSÃO LINEAR

Para implementar a regressão linear nesta pesquisa, iniciou-se com o treinamento individual de cada variável climática (temperatura, umidade, chuva, pressão, radiação), considerando que as falhas no *dataset* são aleatórias entre as colunas. O objetivo foi utilizar as variáveis da estação A (vizinha) para predizer os valores faltantes na estação B (desejada). Para avaliação do método usou-se os dados de teste juntamente com a métrica RMSE para avaliar a magnitude dos erros de previsão do modelo.

A abordagem inicial envolveu o uso das seguintes variáveis preditoras da estação A para completar cada uma das variáveis climáticas alvo na estação B:

- Para chuva: ('chuva', 'Wx', 'Wy', 'distância', 'dif_altura' da estação A)

- Para temperatura: ('temperatura', 'Wx', 'Wy', 'distância', 'dif_altura' da estação A)
- Para umidade: ('umidade', 'Wx', 'Wy', 'distância', 'dif_altura' da estação A)
- Para pressão: ('presao', 'Wx', 'Wy', 'distância', 'dif_altura' da estação A)
- Para radiação: ('radiacao', 'Wx', 'Wy', 'distância', 'dif_altura' da estação A)

Após o treinamento inicial, os resultados de RMSE foram:

Tabela 5 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.336
Umidade	0.504
Chuva	0.987
Pressão atmosferica	0.095
Radiação Solar	0.699

Fonte: Autoria propria

Em seguida, acrescentou-se as variáveis de periodicidade (*'Day sin', 'Day cos', 'Year sin', 'Year cos'*) ao conjunto de preditoras e reexecutou-se a regressão linear para todas as variáveis. Os resultados obtidos foram:

Tabela 6 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.332
Umidade	0.492
Chuva	0.985
Pressão atmosferica	0.095
Radiação Solar	0.664

Fonte: Autoria propria

Observou-se uma leve melhora na precisão para algumas variáveis, indicando que a inclusão das variáveis de periodicidade ajudou a capturar padrões sazonais e diários importantes.

Prosseguiu-se adicionando a variável temperatura da estação A como preditora para todas as variáveis-alvo na estação B e obteve-se:

Tabela 7 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.332
Umidade	0.492
Chuva	0.985
Pressão atmosferica	0.095
Radiação Solar	0.661

Fonte: Autoria propria

Prosseguiu-se adicionando a variável chuva da estação A como preditora para todas as variáveis-alvo na estação B e obteve-se:

Tabela 8 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.332
Umidade	0.491
Chuva	0.985
Pressão atmosferica	0.095
Radiação Solar	0.661

Fonte: Autoria propria

Prosseguiu-se adicionando a variável umidade da estação A como preditora para todas as variáveis-alvo na estação B e obteve-se:

Tabela 9 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.328
Umidade	0.491
Chuva	0.983
Pressão atmosferica	0.095
Radiação Solar	0.660

Fonte: Autoria propria

Prosseguiu-se adicionando a variável radiação da estação A como preditora para todas as variáveis-alvo na estação B e obtive-se:

Tabela 10 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.328
Umidade	0.491
Chuva	0.983
Pressão atmosferica	0.095
Radiação Solar	0.660

Fonte: Autoria propria

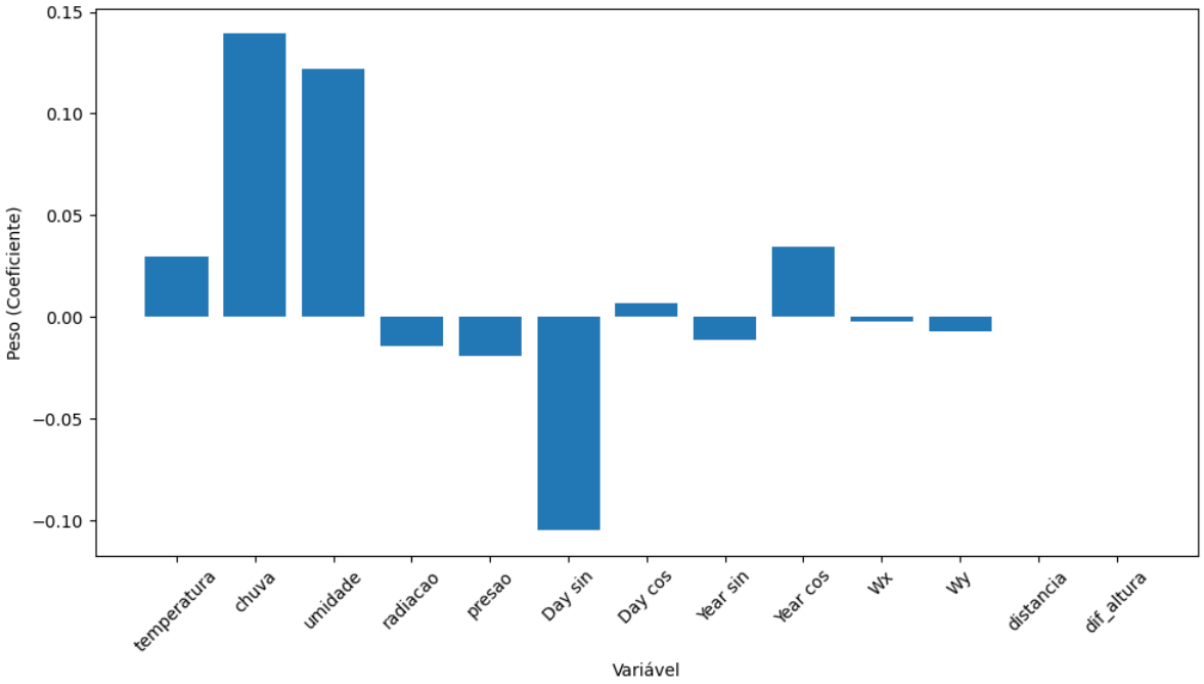
Prosseguiu-se adicionando a variável pressão atmosférica da estação A como preditora para todas as variáveis-alvo na estação B e obteve-se:

Tabela 11 - Desempenho do Método Regressão Linear

Variáveis observadas	RMSE
Temperatura	0.328
Umidade	0.490
Chuva	0.982
Pressão atmosferica	0.095
Radiação Solar	0.660

Fonte: Autoria propria

Figura 26 - Pesos das Variáveis na Regressão Linear para Chuva



Fonte: Autoria própria

A análise dos pesos das variáveis na Figura 26 confirma que a inclusão de múltiplas variáveis preditoras da estação A melhora significativamente a precisão do modelo para a estação B. Variáveis como umidade e chuva têm uma influência predominante, enquanto a pressão, periodicidade, vento e características geográficas também desempenham papéis importantes. Este modelo de regressão linear robusto demonstra a eficácia de utilizar uma abordagem abrangente, capturando as complexas interações climáticas para prever com maior precisão os dados ausentes.

Utilizando as informações das estações vizinhas, conseguiu-se criar um modelo que reflete de forma mais precisa a realidade climática das áreas onde os dados estavam incompletos. Essa abordagem permite capturar de maneira mais abrangente as relações complexas entre as variáveis climáticas, proporcionando um modelo mais robusto e confiável para o completamento de dados.

4.6 REDE DE CAMADAS DENSAS

Para a modelagem dos dados de precipitação, utilizei uma rede neural com camadas densas. A arquitetura do modelo é composta por duas camadas densas. A

primeira camada contém 32 neurônios e utiliza a função de ativação *ReLU (Rectified Linear Unit)*, que é amplamente utilizada devido à sua capacidade de introduzir não linearidades no modelo, permitindo que ele capture relações complexas nos dados. A segunda camada é a camada de saída, que contém um único neurônio, adequado para problemas de regressão.

A compilação do modelo foi feita utilizando o otimizador `adam`, que é um método de descida do gradiente estocástica baseado na estimativa adaptativa de momentos de primeira e segunda ordem. Este otimizador é popular devido à sua eficiência computacional e baixos requisitos de memória. O modelo foi treinado para minimizar o erro quadrático médio (MSE) e utilizou o erro absoluto médio (MAE) como métrica adicional de desempenho. O MSE é uma escolha comum para problemas de regressão, pois penaliza erros grandes de forma mais severa, enquanto o MAE fornece uma interpretação mais direta do erro médio em termos das unidades dos dados.

Para melhorar a robustez do treinamento e evitar *overfitting*, foram definidos dois *callbacks* importantes: `ModelCheckpoint` e `EarlyStopping`. O `ModelCheckpoint` foi configurado para monitorar a perda no conjunto de validação (`val_loss`) e salvar os melhores pesos do modelo em um arquivo. Esta configuração garante que apenas os pesos correspondentes ao menor valor de `val_loss` sejam armazenados, o que é essencial para preservar o melhor estado do modelo durante o treinamento

O callback de `EarlyStopping` foi configurado para monitorar a mesma métrica (`val_loss`) e interromper o treinamento se não houver melhoria após 4 *epochs* consecutivas, restaurando os melhores pesos encontrados até então. Esta técnica é útil para evitar o sobreajuste, garantindo que o modelo não continue a treinar além do necessário, o que poderia levar a uma piora no desempenho em dados não vistos.

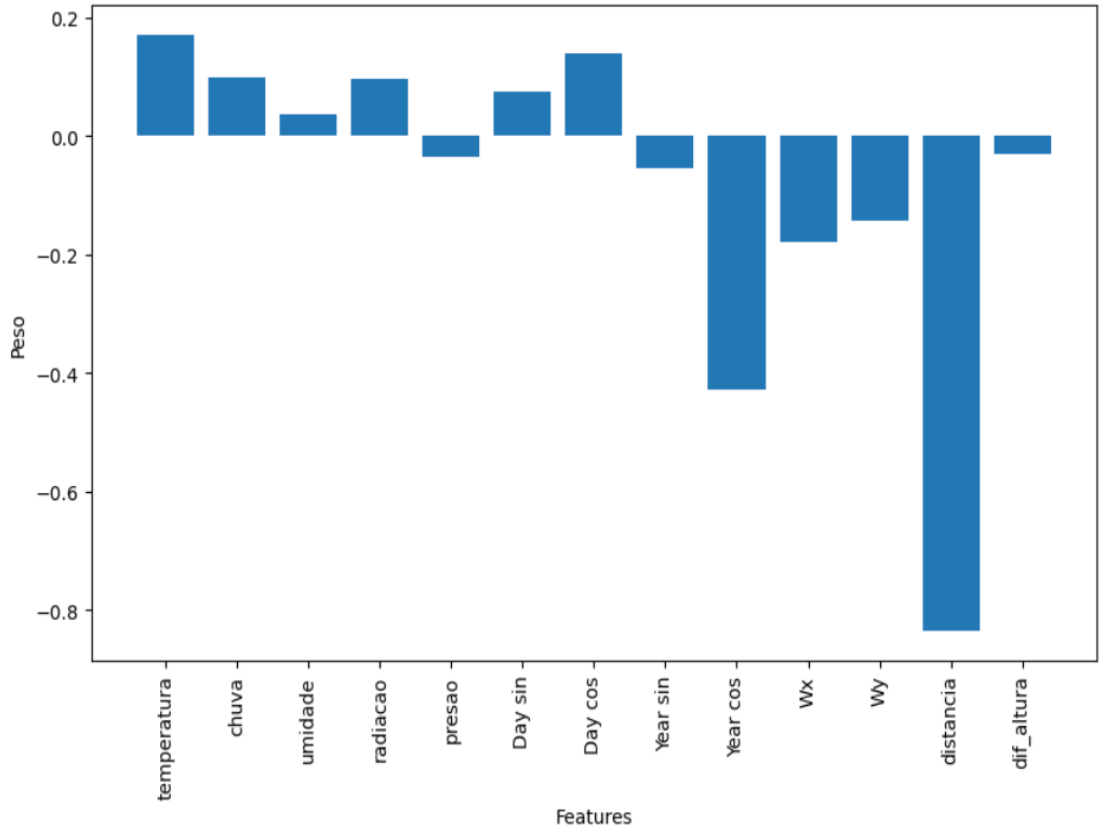
O treinamento foi realizado usando os mesmos dados de treino e formato da regressão linear treinando uma rede neural para cada variável, para ter-se uma comparação mais precisa. Foi executado por até 20 *epochs* com um tamanho de lote de 32, utilizando 20% dos dados de treino para validação. Durante o treinamento, os *callbacks* definidos foram aplicados para monitorar e otimizar o processo. Após o treinamento validou-se cada rede com os mesmos dados utilizados na regressão linear, além da métrica RMSE para validação dos resultados, com isto obteve-se os seguintes resultados:

Tabela 12 - Desempenho do Método Camada Densa

Variáveis observadas	RMSE
Temperatura	0.314
Umidade	0.468
Chuva	0.967
Pressão atmosferica	0.094
Radiação Solar	0.570

Fonte: Autoria propria

Figura 27 - Pesos do Primeiro Neurônio da Primeira Camada



Fonte: Autoria própria

A Figura 27 apresentado ilustra os pesos associados ao primeiro neurônio da primeira camada densa da rede neural. Cada barra representa o peso de uma das variáveis de entrada no modelo. Observa-se que variáveis como temperatura, umidade, radiação, pressão, *Day sin, Day cos e Year sin* possuem pesos positivos, indicando que, conforme os valores dessas variáveis aumentam, a influência no

neurônio da camada densa é positiva, contribuindo para um aumento na saída do neurônio.

Por outro lado, variáveis como chuvas, *Wx*, *Wy*, Year cos e distância possuem pesos negativos, o que significa que, à medida que os valores dessas variáveis aumentam, elas têm um efeito de diminuição na saída do neurônio. Notavelmente, a variável distância apresenta um peso negativo muito significativo, sugerindo que a distância entre as estações tem uma forte influência inversa sobre a saída do neurônio. Isso possivelmente reflete que estações mais distantes têm menor correlação nos dados.

Além disso, a variável dif_altura (diferença entre alturas das estações) mostra um peso praticamente neutro, indicando que sua influência no primeiro neurônio é mínima. Este tipo de análise de pesos é fundamental para entender quais variáveis têm maior impacto nas previsões do modelo e como cada uma contribui para a saída do modelo, ajudando a interpretar a importância relativa das características no contexto da modelagem climática.

4.7 ANÁLISE E COMPARAÇÃO DOS DADOS

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da aplicação de diferentes métodos de completamento de dados faltantes em bases climáticas. Foram utilizadas três principais abordagens: o Método do Vizinho Mais Próximo (NN), a Regressão Linear, e Redes Neurais com Camadas Densas. O objetivo é comparar a eficácia desses métodos na imputação de dados para variáveis climáticas chave, como temperatura, umidade, chuva, pressão atmosférica e radiação solar. A comparação é feita com base no Erro Médio Quadrático (RMSE), que é uma métrica comum para avaliar a precisão dos modelos. Adicionalmente, são discutidas as limitações de cada método, o consumo de recursos computacionais, e as implicações práticas dos resultados obtidos.

4.7.1 Método do Vizinho Mais Próximo (NN)

O Método do Vizinho Mais Próximo (NN) apresenta limitações significativas em comparação aos métodos de regressão linear e camadas densas. Primeiramente, o método NN depende fortemente da disponibilidade de dados da estação vizinha mais

próxima. Se não houver uma estação vizinha suficientemente próxima, ou se os dados da estação vizinha também estiverem incompletos, a imputação pode ser imprecisa ou até impossível. Além disso, a precisão do método NN diminui à medida que a distância entre as estações aumenta, o que pode ser problemático em áreas com baixa densidade de estações meteorológicas.

4.7.2 Regressão Linear

A regressão linear, embora mais robusta do que o método NN, também tem suas limitações. Este método assume uma relação linear entre as variáveis, o que pode não capturar todas as complexidades dos dados meteorológicos, especialmente para variáveis com comportamentos não lineares, como a precipitação. No entanto, a regressão linear não depende da proximidade geográfica das estações e pode utilizar dados históricos mais profundos para fazer previsões.

4.7.3 Redes de Camadas Densas

As redes neurais com camadas densas superam as limitações dos métodos NN e de regressão linear ao capturar relações não lineares complexas nos dados. Este método não está limitado pela disponibilidade de dados de estações vizinhas e pode utilizar dados históricos mais extensos para treinar o modelo, permitindo a construção de séries temporais mais profundas e precisas. A capacidade das redes neurais de aprender padrões complexos resulta em uma maior precisão na imputação de dados faltantes e na previsão de variáveis meteorológicas.

4.7.4 Comparação dos Resultados

A tabela a seguir (Tabela 13) apresenta os resultados obtidos para as variáveis climáticas analisadas, utilizando os três métodos de imputação de dados faltantes: Método do Vizinho Mais Próximo (NN), Regressão Linear, Regressão Linear com periodicidade e Redes Neurais com Camadas Densas.

Tabela 13 – Comparação de Métodos de Completamento

Variável	(NN)	Regressão Linear	Regressão Linear (com periodicidade)	Camadas Densas
Temperatura	0.347	0.336	0.332	**0.314**
Umidade	0.522	0.504	0.492	**0.468**
Chuva	1.298	0.987	0.985	**0.967**
Pressão At.	0.233	0.095	0.095	**0.094**
Radiação Solar	0.756	0.699	0.664	**0.570**

Fonte: Autoria propria

4.7.5 Temperatura

Método do Vizinho Mais Próximo (NN): Apresentou um RMSE de 0.347, sendo o menos preciso entre os métodos analisados. Este método depende da proximidade e disponibilidade dos dados das estações vizinhas, o que pode limitar sua eficácia em regiões com baixa densidade de estações.

Regressão Linear: Melhorou ligeiramente a precisão com um RMSE de 0.336. Este método, embora simples, se beneficia da utilização de dados históricos mais profundos.

Regressão Linear com periodicidade: Houve uma pequena melhora adicional, resultando em um RMSE de 0.332, indicando que a inclusão de variáveis de periodicidade ajuda a capturar padrões sazonais e diurnos.

Redes Neurais com Camadas Densas: Apresentou o menor RMSE (0.314), mostrando maior capacidade de capturar relações complexas nos dados.

4.7.6 Umidade

Método do Vizinho Mais Próximo (NN): Apresentou um RMSE de 0.522, sendo novamente o menos preciso, principalmente devido à variabilidade espacial da umidade.

Regressão Linear: Teve um desempenho melhor com um RMSE de 0.504.

Regressão Linear com periodicidade: Melhorou ainda mais, resultando em um RMSE de 0.492.

Redes Neurais com Camadas Densas: Demonstrou a maior precisão com um RMSE de 0.468.

4.7.7 Chuva

Método do Vizinho Mais Próximo (NN): Teve o pior desempenho com um RMSE de 1.298, devido à alta variabilidade espacial e temporal da precipitação.

Regressão Linear: Reduziu significativamente o erro, com um RMSE de 0.987.

Regressão Linear com periodicidade: Houve uma pequena melhora, com um RMSE de 0.985.

Redes Neurais com Camadas Densas: Apresentou o menor RMSE de 0.967, mostrando melhor capacidade de lidar com a variabilidade da chuva.

4.7.8 Pressão Atmosférica

Método do Vizinho Mais Próximo (NN): Apresentou um RMSE de 0.233.

Regressão Linear: Teve um RMSE significativamente menor de 0.095.

Regressão Linear com periodicidade: Manteve o RMSE de 0.095.

Redes Neurais com Camadas Densas: Teve um desempenho ligeiramente melhor com um RMSE de 0.094.

4.7.9 Radiação Solar

Método do Vizinho Mais Próximo (NN): Teve um RMSE de 0.756.

Regressão Linear: Melhorou para um RMSE de 0.699.

Regressão Linear com periodicidade: Apresentou um RMSE de 0.664.

Redes Neurais com Camadas Densas: Apresentou o melhor desempenho com um RMSE de 0.570.

Os resultados mostram que as Redes Neurais com Camadas Densas geralmente apresentaram os menores valores de RMSE para todas as variáveis, indicando a maior precisão na imputação de dados faltantes. Embora a Regressão Linear com periodicidade tenha melhorado em relação à regressão linear simples, a

diferença de desempenho em relação às Redes Neurais foi pequena. No entanto, essa pequena melhora nas Redes Neurais deve ser ponderada contra o consumo significativo de recursos computacionais necessários para treinar tais modelos, especialmente em comparação com os métodos de Regressão Linear, que são muito menos intensivos em termos de recursos.

Por outro lado, o Método do Vizinho Mais Próximo (NN) mostrou ser o menos eficaz, principalmente devido à sua dependência da proximidade e disponibilidade de dados das estações vizinhas, o que limita sua aplicabilidade em áreas com baixa densidade de estações meteorológicas.

É importante considerar o custo-benefício desta melhoria. A implementação de redes neurais com camadas densas, apesar de mais precisa, requer um consumo significativo de recursos computacionais. O processo de treinamento de redes neurais é intensivo em termos de tempo e poder de processamento, necessitando de hardware especializado, como GPUs, para alcançar tempos de treinamento viáveis. Além disso, as redes neurais demandam um maior volume de dados e ajustes de hiperparâmetros para otimizar seu desempenho, o que pode ser um desafio adicional em termos de tempo e recursos.

Os métodos de regressão linear e camadas densas têm a vantagem adicional de poderem construir dados históricos mais profundos, independentemente da proximidade geográfica das estações meteorológicas. Isso os torna mais versáteis e aplicáveis em áreas com baixa densidade de estações ou com lacunas significativas nos dados.

Por outro lado, o método do Vizinho Mais Próximo (NN) está sujeito à disponibilidade de dados das estações vizinhas, o que limita sua aplicabilidade em áreas menos monitoradas. Além disso, a precisão deste método diminui significativamente com o aumento da distância entre as estações, especialmente para variáveis como a precipitação, que têm alta variabilidade espacial.

A análise comparativa dos métodos destaca a superioridade camadas densas em termos de precisão e versatilidade. Embora o método NN possa ser útil em cenários específicos, sua dependência de dados de estações vizinhas e a diminuição da precisão com a distância o tornam menos ideal para aplicações generalizadas. Os métodos de regressão linear e redes neurais, por sua vez, oferecem soluções mais robustas e precisas para a imputação de dados faltantes e previsão de variáveis

meteorológicas, especialmente em contextos em que a profundidade histórica dos dados é crucial.

5. CONCLUSÃO

A presente pesquisa teve como objetivo desenvolver e avaliar métodos de completamento de dados perdidos em bases climáticas. A abordagem adotada se mostrou eficaz ao integrar métodos de regressão linear, redes neurais com camadas densas e o método do vizinho mais próximo. Os resultados indicaram que as redes neurais com camadas densas apresentaram o melhor desempenho em termos de precisão, demonstrando a capacidade superior desse método em capturar padrões complexos nos dados climáticos. No entanto, é importante considerar o alto custo computacional associado ao treinamento dessas redes, que pode ser um desafio em aplicações práticas com recursos limitados.

Os métodos de regressão linear, especialmente com a inclusão de variáveis de periodicidade, também apresentaram bons resultados, oferecendo uma alternativa eficiente e menos custosa em termos de recursos computacionais. O método do vizinho mais próximo, embora menos preciso, mostrou-se útil em situações em que a proximidade geográfica das estações meteorológicas é um fator relevante. Essa diversidade de métodos permite uma flexibilidade maior na escolha da abordagem mais adequada para diferentes cenários e tipos de dados faltantes.

Uma das principais contribuições deste trabalho foi a implementação de uma API que coleta e processa dados climáticos em tempo real, e a comparação e análise dos métodos de completamento citados, futuramente com a integração da API com o método de imputação de dados, se tornara uma ferramenta poderosa tendo um vasto período de dados climáticos já verificado e tratado. Essa ferramenta facilita o acesso a dados climáticos mais completos e precisos, sendo um recurso valioso para pesquisadores, meteorologistas e profissionais de diversas áreas que dependem de dados climáticos confiáveis. A aplicação prática dessa API pode contribuir significativamente para a melhoria das previsões meteorológicas e análises climáticas, auxiliando na tomada de decisões em setores críticos como agricultura, gestão de recursos hídricos e planejamento urbano.

5.1 PERSPECTIVAS PARA ESTUDOS FUTUROS

Para dar continuidade a este estudo, é proposto o teste de outras arquiteturas avançadas de redes neurais, como redes *transformers* e redes adversariais generativas (GANs). As redes *transformers*, conhecidas por seu desempenho

excepcional em processamento de linguagem natural, podem ser adaptadas para o completamento de dados meteorológicos, oferecendo potencial para capturar dependências temporais de longo alcance com maior eficiência. As *GANs*, por sua vez, podem ser exploradas para gerar dados sintéticos de alta qualidade, auxiliando no preenchimento de lacunas em séries temporais meteorológicas.

Além disso, é essencial comparar os resultados de predições utilizando dados completados e dados originais sem completamento, considerando séries históricas mais longas. Esta comparação permitirá avaliar o impacto do completamento de dados na precisão das predições meteorológicas e a utilidade de técnicas avançadas de imputação em cenários reais. A análise poderá demonstrar o valor agregado de utilizar dados completados para melhorar modelos de previsão e fornecer insights mais robustos e confiáveis para a tomada de decisão em contextos meteorológicos e ambientais.

Outra abordagem promissora é a triangulação de dados, que envolve o uso de dados de mais de duas estações meteorológicas próximas para analisar e completar os dados da estação observada. Essa técnica pode proporcionar um completamento mais detalhado e preciso, além de potencialmente melhorar a predição de condições meteorológicas futuras. A triangulação pode ajudar a capturar variações locais e regionais com maior precisão, resultando em um modelo mais robusto e confiável. Implementar a triangulação de dados pode ser particularmente útil em regiões com alta variabilidade climática, onde a influência de múltiplas estações vizinhas pode fornecer uma visão mais completa das condições meteorológicas.

6. AGRADECIMENTOS

Esta pesquisa foi financiada com uma bolsa de graduação pelo Projeto de Pesquisa, Desenvolvimento e Inovação (PD+I) da ANEEL, desenvolvido em parceria com a Pontifícia Universidade Católica (PUC) Goiás, Universidade de Brasília (UnB), Universidade Federal de Goiás (UFG) e a Eletrobras Furnas, intitulado: “Modelagem em Diversas Escalas da Geração de Sedimentos em Erosões e o Aporte em Reservatórios de UHEs” - PD.0394-1705/2017.

REFERÊNCIAS

Afrifa-yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications, 27(1). doi:10.1002/met.1873

Alkabbani, H.; Ramadan, A.; Zhu, Q.; Elkamel, A. Na Improved Air Quality Index Machine Learning-Based Forecasting with Multivariate Data Imputation Approach. Atmosphere. 2022, 13, 1144. https://doi.org/10.3390/ atmos13071144

Avila-Diaz, A., Benezoli, V., Justino, F., Torres, R., & Wilson, A. (2020). Assessing current and future trends of climate extremes across Brazil based on reanalyses and earth system model projections. Climate Dynamics, 55(5-6), 1403-1426. https://link.springer.com/article/10.1007/s00382-020-05333-z

Başakın, Ensar; Ekmekcioğlu, Ömer Özger, Mehmet Özger. Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model. Energy Conversion and Management. Volume 280, 15 March 2023, 116780. https://doi.org/10.1016/j.enconman.2023.116780

Bayma, L. O.; Pereira, M. A. Identifying Finest Machine Learning Algorithm for Climate Data Imputation in the State of Minas Gerais, Brazil. Journal of Information and Data Management, Vol. 9, No. 3, December 2018, Pages 259–274.

Breve, M. M., Balsa, C., Rufino, J., & Martins, F. (2023). Enhancing weather data reconstruction through hybrid methods with dimensionality reduction. Computation, 11(5), 98. https://doi.org/10.3390/computation11050098

Chakraborty, H., Samanta, P., & Zhao, L. (2021). Sequential Data Imputation with Evolving Generative Adversarial Networks. 2021 International Joint Conference on Neural Networks (IJCNN), 1-8. https://doi.org/10.1109/IJCNN52387.2021.9534108.

Chen, K., Liang, X., Zhang, Z., & Ma, Z. (2022). GEDI: A Graph-based End-to-end Data Imputation Framework. ArXiv, abs/2208.06573. https://doi.org/10.48550/arXiv.2208.06573.

Deng, G., Han, C., & Matteson, D. (2020). Learning to Rank with Missing Data via Generative Adversarial Networks. ArXiv, abs/2011.02089.

Desai, M.; Shah, M. An anatomization on Breast Cancer Detection and Diagnosis employing Multi-layer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN). Clinical eHealth, nov. 2020. Doi: https://doi.org/10.1016/j.ceh.2020.11.002

Gao, F. (2023). Ae2I: A Double Autoencoder for Imputation of Missing Values. ArXiv, abs/2301.06633. https://doi.org/10.48550/arXiv.2301.06633.

Garcia, C.; Leite, D. and Škrjanc, I., "Incremental Missing-Data Imputation for Evolving Fuzzy Granular Prediction," in IEEE Transactions on Fuzzy Systems, vol. 28, no. 10, pp. 2348-2362, Oct. 2020, doi: 10.1109/TFUZZ.2019.2935688.

Gad, D.; I., & Manjunatha, B., 2017. Avaliação de desempenho de modelos preditivos para imputação de dados faltantes em dados meteorológicos. Conferência Internacional de 2017 sobre Avanços em Computação, Comunicações e Informática (ICACCI), pp. https://doi.org/10.1109/ICACCI.2017.8126025.

GIL, Antônio Carlos Como Elaborar Projetos de Pesquisa. 6. ed. São Paulo: Editora Atlas Ltda., 2017.

Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning. Cambridge, Massachusetts: The Mit Press, ISBN: 978-0262035613, 2016.

Haykin, S. (2000). Redes Neurais: Princípios e Prática (P. M. Engel, Trad.). (2ª ed.). Bookman.

Intergovernmental Panel on Climate Change (IPCC). (2023). Climate Change 2023: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [AR6 Synthesis Report].

Acedido de https://www.ipcc.ch/report/ar6/syr/downloads/report/IPCC_AR6_SYR_FullVolume.pdf

Kadow, C., Hall, D. M., & Ulbrich, U. (2020). "Artificial intelligence reconstructs missing climate information." Nature Geoscience, 13, 408–413. https://www.nature.com/articles/s41561-020-0582-5.

Khodir Madani, Marco Scarpa, Brunella Bonaccorso, Khalef Lefsih. A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. Heliyon 5 (2019) e01247. doi: 10.1016/j.heliyon. 2019.e01247

Körner, Philipp; Kronenberg, Rico; Genzel, Sandra; Bernhofer, Christian (2018). Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. Meteorologische Zeitschrift, 27(5), 369–376. doi:10.1127/metz/2018/0908

Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. Communications of the ACM, v. 60, n. 6, p. 84–90, 24 maio 2012. Doi: https://doi.org/10.1145/3065386

Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. Nature, v. 521, n. 7553, p. 436–444, maio 2015. Doi: https://doi.org/10.1038/nature14539

Li, S., Jiang, B., & Marlin, B. (2019). MisGAN: Learning from Incomplete Data with Generative Adversarial Networks. ArXiv, abs/1902.09599.

Lopez, J., Hernández, S., Urrutia, A., López-Cortés, X., Araya, H., & Morales-Salinas, L., 2021. Efeito dos dados faltantes em séries temporais curtas e sua aplicação na caracterização da temperatura superficial por análise de flutuação detendida. Computação. Geociências., 153, pp.

Lovejoy, S., & Schertzer, D. (2022). Maximum rates of climate change are systematically underestimated in the geological record. Nature Communications, 13(1). DOI: 10.1038/s41467-022-28350-y

Lucas O. Bayma, Marconi A. Pereira. Identifying Finest Machine Learning Algorithm for Climate Data Imputation in the State of Minas Gerais, Brazil. Journal of Information and Data Management, Vol. 9, No. 3, December 2018, Pages 259–274.

Machado, R. D.; Fortes, Bravo, G.; Starke, A.; Lemos. L.; Colle, S. Generation of 441 typical meteorological year from INMET stations – Brazil. IEA SHC International Conference on Solar Heating and Cooling for Buildings and Industry 2019.

Marivate, V., Nelwamondo, F., & Marwala, T. (2007). Autoencoder, Principal Component Analysis and Support Vector Regression for Data Imputation. ArXiv, abs/0709.2506.

Mcculloch, W. S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. The Bulletin of Mathematical Biophysics, v. 5, n. 4, p. 115–133, dez. 1943. Doi: https://doi.org/10.1007/BF02478259

Mitchell, T. M. Machine learning. New York: Mcgraw-Hill, ISBN: 0070428077, 1997.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

Moura, A.; Fortes, L., The Brazilian National Institute of Meteorology (INMET) and its contributions to agrometeorology. AGROMETEOROS, NOV 2016 10.31062/argon. v24i1.24878.

Ramchoun, H. et al. Multilayer Perceptron: Architecture Optimization and Training. International Journal of Interactive Multimedia and Artificial Intelligence, v. 4, n. 1, p. 26, 2016. Doi: https://doi.org/10.9781/ijimai.2016.415

Ref T. Aljuaid and S. Sasi, "Proper imputation techniques for missing values in data sets," 2016 International Conference on Data Science and Engineering (ICDSE), Cochin, India, 2016, pp. 1-5, doi: 10.1109/ICDSE.2016.7823957.

RusseL, S.; Norvig, P. Artificial Intelligence: A Modern approach. 4. ed. [s.l.] Prentice Hall, ISBN: 9780134610993, 2020.

Sainath, T. N. et al. Deep convolutional neural networks for LVCSR. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), p. 8614-8618, maio 2013. Doi: https://doi.org/10.1109/ICASSP.2013.6639347

Sanches, F.H.C., Martins, F.R., Conti, W.R.P. et al. The increase in intensity and frequency of surface air temperature extremes throughout the western South Atlantic coast. Sci Rep 13, 6293 (2023). https://doi.org/10.1038/s41598-023-32722-1

Sattari, M.-T., Apel, H., & Koch, M. (2017). Assessment of different methods for estimation of missing data in precipitation studies. Hydrology Research, 48(4), 1032-1046. https://doi.org/10.2166/nh.2016.364

Silva Junior, C.A., Teodoro, P.E., Delgado, R.C. et al. Persistent fire foci in all biomes undermine the Paris Agreement in Brazil. Sci Rep 10, 16246 (2020). https://doi.org/10.1038/s41598-020-72571-w

Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. Proceedings of the 38th International Conference on Machine Learning. v. 139, p. 10096–10106, Jul. 2021. Url: https://proceedings.mlr.press/v139/tan21a.html

Wazlawick, R. S. Metodologia da Pesquisa para Ciência da Computação. 2a. ed. [S.l.]: Campus, 2014.

Xi, N., & Li, J. (2023). Benchmarking the Autoencoder Design for Imputing Single-Cell RNA Sequencing Data. bioRxiv. https://doi.org/10.1101/2023.02.16.528866.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
GABINETE DO REITOR

Av. Universitária, 1069 ● Setor Universitário
Caixa Postal 86 ● CEP 74605-010
Goiânia ● Goiás ● Brasil
Fone: (62) 3946.1000
www.pucgoias.edu.br ● reitoria@pucgoias.edu.br

RESOLUÇÃO n° 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O(A) estudante Norton Pereira Ricardo do Curso de Ciência da Computação, matrícula 20182002800500, telefone: (62) 99201-6959 e-mail nortonricardo@live.com, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado INOVAÇÃO EM COMPLETAÇÃO DE DADOS CLIMÁTICOS: MÉTODOS ~~BASEADOS EM VIZINHOS, REGRESSÃO LINEAR E REDES NEURAIS DE CAMADAS~~ DENSAS, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 05 de junho de 2024.

gov.br Documento assinado digitalmente
NORTON PEREIRA RICARDO
Data: 26/06/2024 18:50:56-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor: _____

Nome completo do autor: Norton Pereira Ricardo

Assinatura do professor–orientador: _____

Nome completo do professor-orientador: Maria José Pereira Dantas