UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

André Carneiro da Silva

Exploração de dados para aprendizagem de máquina: o caso de uma operação agrícola

São Carlos

2024

André Carneiro da Silva

Exploração de dados para aprendizagem de máquina: o caso de uma operação agrícola

Monografia apresentada ao Curso de Engenharia de Produção, da Escola de Engenharia de São Carlos da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Engenheiro de Produção.

Orientador: Prof. Dr. Lucas Gabriel Zanon

São Carlos

2024

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Silva, André Carneiro da
S289e Exploração de dados para aprendizagem de máquina: o caso de uma operação agrícola / André Carneiro da Silva; orientador Lucas Gabriel Zanon. São Carlos, 2024.

Monografia (Graduação em Engenharia de Produção) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2024.

1. KDD. 2. AgTech. 3. EDA. 4. IoT. 5. Agricultura Inteligente. 6. Agricultura. 7. Estações meteorológicas. 8. ML. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

FOLHA DE APROVAÇÃO

Candidato: André Carneiro da Silva
Título do TCC: Exploração de dados para aprendizagem de máquina: o caso de uma operação agrícola
Data de defesa: 13/12/2024

Comissão Julgadora	**Resultado**
Professor Doutor Lucas Gabriel Zanon (orientador)	APROVADO
Instituição: EESC **-** SEP	
Professor Doutor Rafael Ferro Munhoz Arantes	APROVADO
Instituição: EESC **-** SEP	
Professor Titular Luiz Cesar Ribeiro Carpinetti	APROVADO
Instituição: EESC - SEP	

Presidente da Banca: **Professor Doutor Lucas Gabriel Zanon**

RESUMO

SILVA, A. C. **Exploração de dados para aprendizagem de máquina: o caso de uma operação agrícola**. 2024. 75p. Monografia (Trabalho de Conclusão de Curso) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2024.

O crescimento da população mundial intensifica a demanda por alimentos, tornando indispensável a adoção de soluções mais eficientes e sustentáveis para as operações agrícolas. Nesse cenário, o avanço da Agricultura Inteligente e o surgimento de empresas de cunho tecnologico no contexto agrícola, as *AgTechs*, aliado às tecnologias de IoT e às estações meteorológicas, possibilita o monitoramento em tempo real das condições climáticas das plantações. Essas tecnologias geram grandes volumes de dados, que precisam ser analisados e traduzidos em *insights* aplicáveis à operação, com o objetivo de agregar valor e aumentar a produtividade nas fazendas, o que é uma forma de suprir a demanda crescente de alimentos. Uma das formas de traduzir grandes volumes de dados em *insights* é por meio de modelos de *machine learning* (ML). Entretanto para uma boa assertividade desses modelos, é necessário o seu treinamento com dados confiáveis e validados. Nesse contexto, se faz necessário a existência de uma metodologia robusta para exploração e tratamento desses dados a fim de garantir essa confiabilidade. O processo de KDD (*Knowledge Discovery in Databases*) visa identificar e validar padrões úteis e compreensíveis a partir de dados, enquanto a EDA (*Exploratory Data Analysis*) é uma abordagem preliminar focada em explorar o que os dados revelam sem a aplicação de testes formais, como testes de hipóteses. Juntas, essas abordagens fornecem uma base sólida para análise e interpretação dos dados, facilitando a descoberta de insights relevantes para a operação agrícola. Dessa forma, o presente trabalho propõe aplicar as etapas iniciais do KDD, aliadas às técnicas de EDA em um *dataset* agrícola real de forma estruturada para viabilizar um futuro desenvolvimento de um *framework* de gestão de desempenho baseado em modelos de ML, com o objetivo de prever os impactos das variáveis analisadas na produtividade para basear a tomada de decisão. Como resultado dessas etapas iniciais, objetiva-se buscar um melhor entendimento dos dados e transformar o *dataset* fornecido pela empresa parceira em uma base de dados pronta para a aplicação de modelos de ML por meio da aplicação de filtros, redução do número de colunas, *data cleaning*, tratamento de *outliers* e valores faltantes, adição de novas colunas de interesse e *data completion*. Esse *dataset* contém medições reais de variáveis meteorológicas coletadas por uma estação de uma empresa parceira, contendo informações como pressão atmosférica, temperatura, temperatura do solo, umidade relativa do ar e do solo, velocidade do vento, direção do vento, quantidade de chuva e radiação solar

Palavras-chave: KDD. *Agtech*. EDA. IoT. Agricultura Inteligente. Agricultura. Estações meteorológicas. ML.

ABSTRACT

SILVA, A. C. **Data Exploration for Machine Learning: The Case of an Agricultural Operation** 2024. 75p. Monograph (Conclusion Course Paper) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2024.

The growth of the world's population intensifies the demand for food, making it essential to adopt more efficient and sustainable solutions for agricultural operations. In this context, the advancement of Smart Agriculture and the emergence of technology-based companies in the agricultural sector, the AgTechs, along with IoT technologies and meteorological stations, enable real-time monitoring of climatic conditions in crops. These technologies generate large volumes of data, which need to be analyzed and translated into insights applicable to the operation, aiming to add value and increase farm productivity, which is one way to meet the growing food demand. One way to translate large volumes of data into insights is through machine learning models; however, for these models to be effective, they need to be trained with reliable and validated data. In this context, the existence of a robust methodology for exploring and processing this data is necessary to ensure its reliability. The KDD (Knowledge Discovery in Databases) process aims to identify and validate useful and understandable patterns from data, while EDA (Exploratory Data Analysis) is a preliminary approach focused on exploring what the data reveals without applying formal tests, such as hypothesis testing. Together, these approaches provide a solid foundation for data analysis and interpretation, facilitating the discovery of relevant insights for agricultural operations. Thus, this work proposes to apply the initial steps of KDD (initial understanding of the dataset, selection, preprocessing, and data transformation), combined with EDA techniques, to a real agricultural dataset in a structured manner, to enable the future development of a performance management framework based on ML models, with the goal of predicting the impacts of the analyzed variables on productivity to inform decision-making. As a result of these initial steps, the aim is to gain a better understanding of the data and transform the dataset provided by the partner company into a database ready for applying ML models. This process involves applying filters, reducing the number of columns, performing data cleaning, handling outliers and missing values, adding new columns of interest, and completing missing data. This dataset contains real measurements of meteorological variables collected by a station from a partner company, including information such as atmospheric pressure, temperature, soil temperature, relative humidity of the air and soil, wind speed, wind direction, rainfall, and solar radiation.

Keywords: KDD. Agtech. EDA. IoT. Smart Farming. Agriculture. Weather stations. ML.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	*Knowledge Discovery in Databases*
EDA	*Exploratory Data Analysis*
ML	*Machine Learning*
IA	*Inteligência Artificial*
IoT	*Intenet of Things*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta a contextualização para o tema do trabalho (Seção 1.1), a justificativa e as lacunas de pesquisa (Seção 1.2), bem como seus objetivos e a questão de pesquisa correspondente (Seção 1.3).

1.1 Contextualização

O crescimento populacional mundial é um tema constantemente discutido pelas Nações Unidas, que projetam que a população global alcance 9,7 bilhões de pessoas em meados de 2050 (NAÇÕES UNIDAS, 2023). Esse aumento, partindo dos atuais 8,2 bilhões (NAÇÕES UNIDAS, 2024), representa uma elevação de 18,29% em 36 anos, criando uma pressão adicional sobre os sistemas alimentares. Estima-se que, entre 2019 e 2050, a demanda por alimentos aumente em cerca de 57%, exigindo um avanço expressivo na produtividade agrícola (FALCON; NAYLOR; SHANKAR, 2022). Para suprir essa demanda crescente, as operações agrícolas precisam se adaptar a uma realidade em que os recursos naturais são mais escassos e a eficiência produtiva se torna essencial (SHAIKH; RASOOL; LONE, 2022).

Uma forma de lidar com esse aumento crescente de demanda por alimentos é buscar a automação de operações agrícolas e a otimização da produção com base em decisões informadas por dados provenientes da agricultura 4.0 ou *smart farming* (MENDES *et al.*, 2024). Nesse sentido, o conceito de *Agricultural Technology* (Tecnologia Agrícola), ou *AgTech*, representa o ecossistema de inovação tecnológica que engloba startups e soluções avançadas voltadas para o setor agrícola (BAMBINI; BONACELLI, 2019). Esse ecossistema se beneficia do uso de tecnologias como a Internet das Coisas (IoT) e estações meteorológicas, que permitem a coleta em tempo real de dados essenciais para a produtividade das culturas, como pressão atmosférica, umidade do ar e do solo, temperatura, direção e velocidade do vento, precipitação e radiação solar (SHAIKH; RASOOL; LONE, 2022).

Nesse contexto, o aprendizado de máquina surge como uma ferramenta promissora para otimizar processos agrícolas, com a capacidade de aprender a partir de grandes volumes de dados e gerar previsões úteis. Modelos de *Machine Learning* (ML) têm sido amplamente empregados para prever o rendimento das colheitas (PALLATHADKA et al., 2023), permitindo aos agricultores tomar decisões informadas sobre o que plantar e quando plantar (VAN KLOMPENBURG; KASSAHUN; CATAL, 2020). Contudo, para que esses modelos sejam eficazes, é essencial que o grande volume de dados rurais passem por um pré-processamento cuidadoso, devido às frequentes inconsistências e lacunas, especialmente em áreas com infraestrutura limitada. A preparação adequada dos dados é, portanto, um passo crucial para garantir que as previsões forneçam informações confiáveis para a gestão agrícola (SHAIKH;

RASOOL; LONE, 2022).

Dessa forma, é necessário o uso de uma metodologia cientificamente comprovada para processar o alto volume de dados gerados, possibilitando a extração de conhecimento. Uma metodologia amplamente reconhecida na literatura científica para esse propósito é o *Knowledge Discovery in Databases* (KDD). O KDD é composto por uma série de etapas estruturadas, que incluem: seleção dos dados, pré-processamento, transformação dos dados, mineração de dados, interpretação e avaliação (FAYYAD *et al.*, 1996). Por fim, o presente trabalho concentrou-se nas etapas de seleção, pré-processamento e transformação dos dados, integrando técnicas de Análise Exploratória de Dados (EDA) como suporte complementar a cada uma dessas fases.

1.2 Justificativa e lacunas de pesquisa

Considerando que os dados agrícolas são obtidos de diversas bases de dados, modelos, além de serem coletados em estações e dispositivos IoT na zona rural, é comum que apresentem desorganização e diversas lacunas ou dados faltantes. Essas fontes frequentemente geram dados com inconsistências e problemas variados, o que torna indispensável um processo rigoroso de preparação antes da aplicação de técnicas de mineração de dados e aprendizado de máquina. Assim, para que esses métodos sejam eficazes em datasets agrícolas, é fundamental assegurar a consistência e qualidade dos dados, evitando que os problemas citados acima comprometam da análise desejada (SHAIKH; RASOOL; LONE, 2022).

Diversas aplicações do KDD tem sido feitas na literatura em setores como saúde pública (SANTOS; STEINER; LIMA, 2022), meteorologia (ANDRIYANA *et al*., 2024) (PENG; LI; TSUNG, 2024), construção civil (LLATAS *et al.*, 2024). Entretanto, aplicações para agricultura e mais especificamente para a agricultura 4.0 ainda são escassas com sua maioria voltada para cibersegurança nas fazendas (PEPPES *et al.*, 2021; PRAMILARANI; KUMARI, 2024) e, com o conhecimento adquirido pelo autor até o presente, nenhuma buscando prever produtividade baseado em dados de estações meteorológicas.

Além disso, foram encontradas apenas 9 resultados no *Scopus* para a *query* ("*agtech**" OR "*agritech**" OR "*agricultural startup**" OR "*agrotech**" OR "*agriculture startup**" OR "*agricultural technolog* startup**" AND "KDD" OR "*Knowledge Discovery in Database**"). Por fim, nenhum dos resultados mencionados combinam KDD com *agritech* ou *agtech*, o que indica uma possível lacuna de pesquisa.

Por fim, segundo Gupta *et al.* (2023) uma das principais oportunidades de pesquisa é a previsão do rendimento de culturas utilizando os dados gerados pelas tecnologias emergentes. Nesse contexto, a justificativa para este trabalho reside na aplicação do KDD em um conjunto de dados agrícolas reais, fornecido por uma *AgTech* parceira, com o objetivo de prever a produtividade. Esta aplicação contribui para o avanço do conhecimento científico ao integrar

KDD e *AgTech*, oferecendo uma metodologia robusta de entendimento, seleção, pré-processamento e transformação de dados, com o objetivo de aumentar a confiabilidade e qualidade dos dados agrícolas, atendendo a uma das principais preocupações relacionadas aos dados gerados por dispositivos IoT (SHAIKH; RASOOL; LONE, 2022). Além disso, serão utilizadas técnicas de EDA para complementar as etapas do KDD.

1.3 Questão de pesquisa e objetivos

Nessa seção explicita-se a questão da pesquisa (seção 1.3.1), o objetivo geral (seção 1.3.2) e desdobra-se esse objetivo em objetivos específicos (seção 1.3.3).

1.3.1 Questão de pesquisa

A questão de pesquisa que motiva a elaboração deste trabalho é: como estruturar uma abordagem sistemática de exploração de dados de dados para viabilizar a aplicação de algorítmos de aprendizagem de máquina em *dataset* de uma operação agrícola?

1.3.2 Objetivo geral

O objetivo do presente Trabalho de Conclusão de Curso é estruturar uma abordagem sistemática de exploração e pré-processamento de dados para viabilizar a aplicação de algoritmos de aprendizagem de máquina em *dataset* de uma operação agrícola.

Assim, esse trabalho tem o intuito de estruturar uma metodologia robusta de exploração e pré-processamento de dados em bases agrícolas para possibilitar a extração de conhecimento por meio do uso posterior de técnicas de aprendizado de máquina para previsão de produtividade e apoio à decisão e aplicá-la nos dados da empresa parceira.

1.3.3 Objetivos específicos

O objetivo geral se desdobra nos seguintes objetivos específicos:

(I) Realizar uma revisão bibliográfica sobre o uso do KDD e as etapas de exploração e pré-processamento de dados em aprendizagem de máquina;

(II) Desenvolver uma abordagem sistemática que apresente um passo-a-passo para exploração e pré-processamento de dados;

(III) Aplicar a abordagem desenvolvida em caso real com foco no delineamento do problema de aprendizagem a partir dos dados, descobrimento dos dados, mapeamento da estrutura dos dados, identificação de principais atributos e sua natureza, visualização de dados, análise de correlação entre atributos, pré-processamento e data-completion;

(IV) Comparar os métodos de data-completion com base em estatísticas descritivas;

(V) Extrair *insights* dos dados.

2 METODOLOGIA

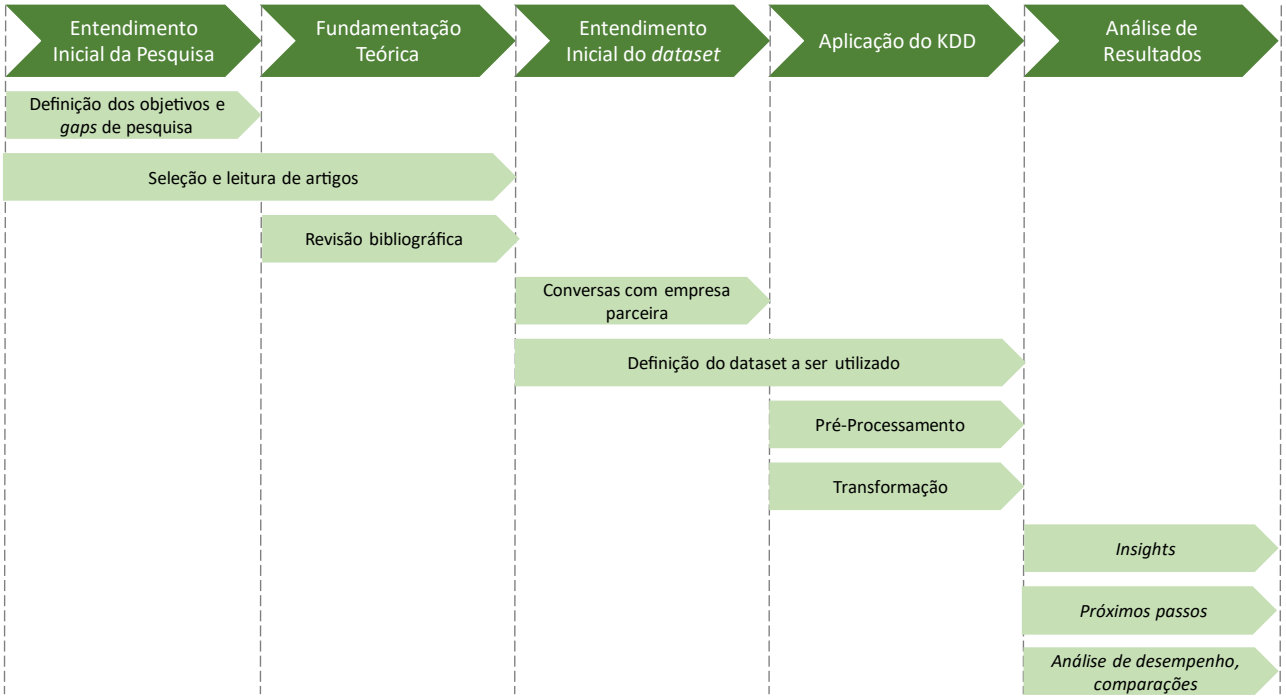
Este trabalho utiliza da metodologia de modelagem quantitativa, axiomática e normativa (BERTRAND; FRANSOO, 2009) e é caracterizado por sua ênfase em propor soluções específicas para problemas delimitados. Na modelagem quantitativa, utiliza-se um conjunto de variáveis relacionadas a um domínio específico, estabelecendo conexões e dependências entre elas. A abordagem axiomática baseia-se na definição de premissas e suposições fundamentais que orientam a construção do modelo, permitindo explorar a estrutura do problema e compreender como diferentes variáveis se comportam ou interagem. Já o aspecto normativo está voltado para a aplicação prática, com o objetivo de fornecer recomendações ou diretrizes que auxiliem na resolução de desafios ou na tomada de decisões baseadas nas conclusões obtidas a partir do modelo.

A escolha de metodologia para este trabalho objetiva permitir a aplicação das etapas iniciais do KDD no *dataset* da empresa parceira, que são: seleção, pré-processamento e transformação de dados. Na figura 1, ilustra-se as etapas do trabalho: entendimento inicial da pesquisa, fundamentação teórica, entendimento inicial do *dataset*, aplicação do KDD e análise de resultados.

É importante ressaltar que este trabalho conta com a parceria de uma empresa agrícola de base tecnológica, que concedeu acesso ao autor às bases de dados de algumas de suas operações devido à necessidades e motivações semelhantes, tais como a necessidade de uma metodologia de pré-processamento de dados robusta para lidar com dados incompletos devido aos problemas de equipamentos IoT e/ou falhas de conexão, rede ou transmissão de dados.

A empresa parceira atua no setor de coleta de dados meteorológicos em campo por meio de dispositivos IoT, oferecendo serviços de recomendações para apoio à tomada de decisões em tempo real pelos produtores. Além disso, integra um *hub* de inovação composto por empresas agrícolas de base tecnológica, localizado em Piracicaba, no interior do estado de São Paulo.

Figura 1 – Sequência de etapas compondo a metodologia do trabalho



Fonte: Elaboração própria (2024)

2.1 Entendimento inicial da pesquisa

Inicialmente, foram definidos a questão de pesquisa, os objetivos gerais e específicos, bem como as lacunas de pesquisa identificadas. Posteriormente, foi realizada uma breve revisão bibliográfica para identificar para identificar os principais constructos teóricos relacionados a esse trabalho. Assim, selecionou-se os artigos relevantes para o desenvolvimento do trabalho por meio de recomendações do orientador, bem como pesquisas em renomadas bases de dados, tais como *Scopus* e *Web of Science*. No primeiro momento, foi realizada uma leitura dos resumos dos artigos para iniciar a contextualização, entendimento das metodologias e conceitos fundamentais de forma gradual e selecionar quais mais se alinham com a pesquisa e fortaleceriam a base teórica com uma leitura integral.

2.2 Fundamentação teórica

Posteriormente, identificando-se os principais constructos, foram selecionados estudos para estruturação da revisão bibliográfica. Assim, ela foi feita sobre os seguintes temas: Transformação digital na Agricultura, KDD, *Python* para EDA. O principal objetivo dessa etapa foi consolidar uma base teórica robusta sobre os temas relevantes para possibilitar o desenvolvimento do trabalho.

2.3 Entendimento inicial do dataset

Nesta etapa, deu-se início à aplicação da teoria revisada e a um maior contato com a empresa parceira. Receberam-se os dados brutos das estações coletoras de dados meteorológicos da empresa parceira em um arquivo de formato csv. Em seguida, foram marcadas duas chamadas via *Google Meets*, além de troca de *e-mails* para dúvidas pontuais, visando um maior entendimento do *dataset* e embasar com o conhecimento da operação agrícola responsável pelos dados o início da aplicação do KDD.

2.4 Aplicação do KDD

Finalmente, após compreender a operação da empresa parceira e as particularidades da base de dados a ser analisada, torna-se possível iniciar a aplicação do KDD. As etapas abordadas neste trabalho incluem: entendimento inicial, seleção, pré-processamento e transformação dos dados. As demais etapas de *data mining* e a interpretação para a geração de conhecimento serão deixadas para pesquisas futuras.

2.5 Análise dos resultados

Por fim, após a aplicação do KDD buscou-se consolidar as informações geradas, os gráficos construídos e analisar os principais *insights* e conhecimentos gerados sobre o *dataset* especificamente e sobre a aplicação da metodologia de maneira geral para viabilizar a aplicação de um modelo de ML em operações agrícolas e as demais etapas do KDD. Além disso, buscou-se também elencar os próximos passos e fazer uma análise de desempenho.

3 REVISÃO BIBLIOGRÁFICA

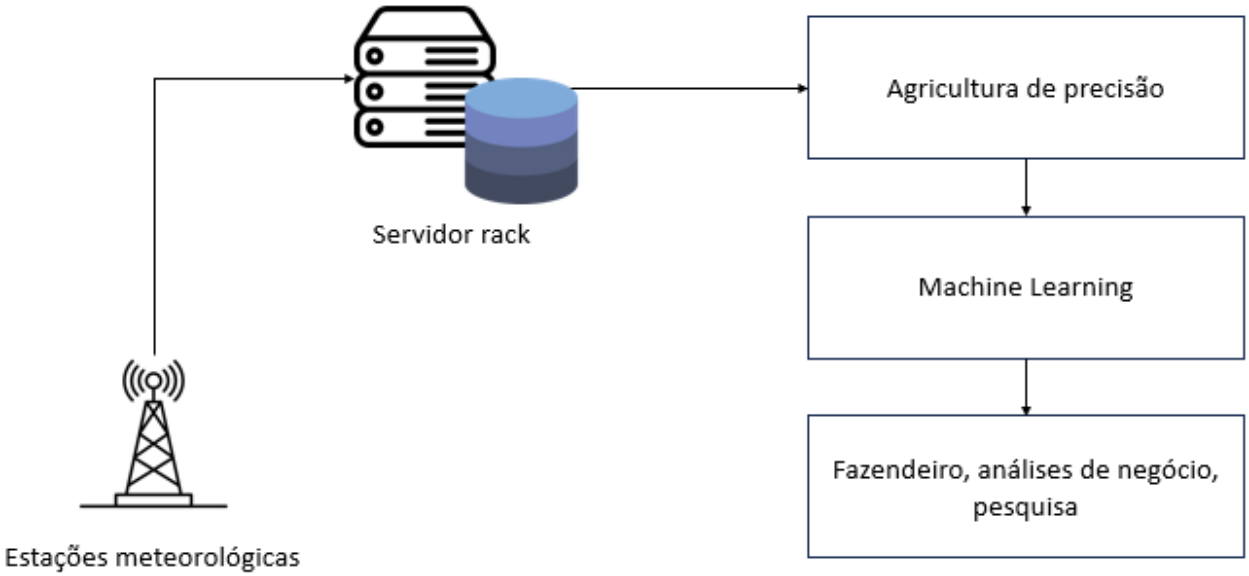
Esta seção contém as bases teóricas necessárias para fundamentar este trabalho, abordando os temas: transformação digital na agricultura, Knowledge Discovery in Databases (KDD) e *Python* para análise exploratória de bases de dados (EDA).

3.1 Transformação digital na agricultura

Devido à crescente demanda por alimentos (MENDES *et al.*, 2022), é essencial que os fazendeiros e setor agrícola como um todo adotem as tecnologias mais recentes de digitalização da agricultura (KÄRNER, 2017). Segundo Shaikh, Rasool, Lone (2022) a agricultura 4.0 está alinhada com essa necessidade, uma vez que utiliza Inteligência Artificial (IA) e IoT para a gestão integrada das operações agrícolas e busca um aumento de produtividade de forma sustentável.

Assim, o sistema de Agricultura Inteligente é sustentado por automação digital, coleta e transmissão de dados, processamento e análise por meio de tecnologias como sensores, GPS, câmeras e etiquetas de identificação por radiofrequência (RFID) (SHAIKH; RASOOL; LONE, 2022). Essas informações são então gerenciadas em plataformas baseadas na nuvem, permitindo uma operação eficiente e conectada, conforme a figura 2.

Figura 2 – Sistema de Agricultura Inteligente esquemático



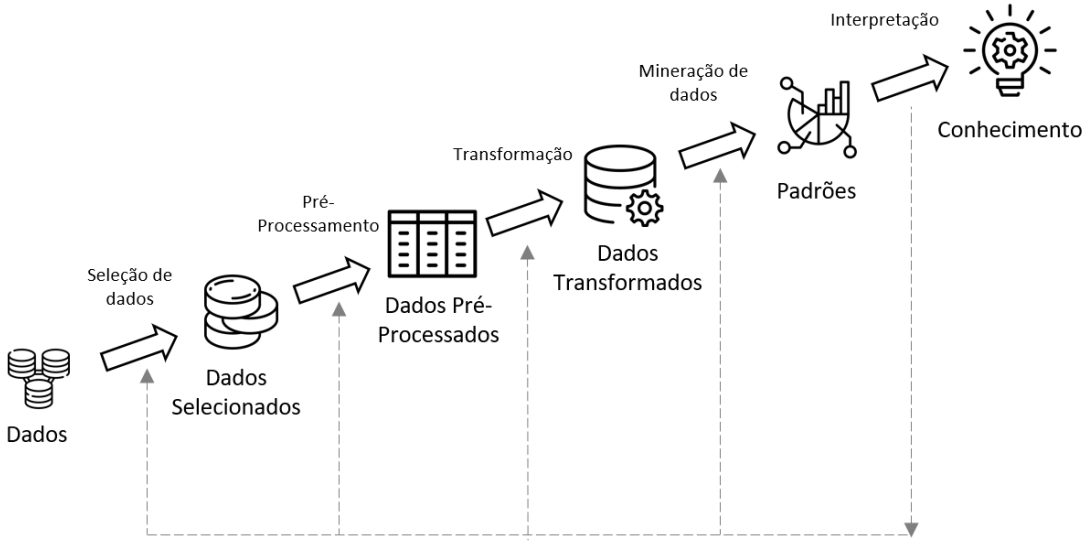
Fonte: Adaptado de Shaikh, Rasool, Lone (2022)

Dessa forma, à medida que a geração de dados agrícolas cresce exponencialmente com as tecnologias da Agricultura Inteligente, torna-se cada vez mais importante a extração de conhecimento e aplicação de modelos de ML para suportar a tomada de decisão, já que o volume de dados ultrapassa a capacidade de análise humana (KÄRNER, 2017).

3.2 Knowledge Discovery in Databases (KDD)

Existem diversos termos similares para o processo de extrair conhecimento de base de dados. O autor Sarker (2021), afirma que alguns deles são *data mining, knowledge mining from data, knowledge extraction, data/pattern analysis, knowledge discovery from data (KDD)*. De acordo com Han *et al.*, 2011 um nome mais apropriado seria *knowledge mining from data*. Por outro lado, Fayyad *et al.* apresentam o KDD como o processo da figura 3.

Figura 3 – Processo KDD



Fonte: Adaptado de Fayyad *et al.* (1996)

Com os diversos termos relacionados citados, o processo de descoberta de conhecimento em base de dados (KDD) é definido como um processo para identificar e validar padrões úteis e compreensíveis em dados (FAYYAD; SHAPIRO; SMYTH, 1996). Além disso, o KDD é um processo interativo e iterativo, com várias decisões realizadas pelo usuário em diferentes etapas e pode ser repetido diversas vezes até atender a necessidade definida pelo usuário, tornando-se dinâmico e adaptável às necessidades específicas da aplicação (FOTE *et al.*, 2020). Por fim, esse processo envolve múltiplas áreas do conhecimento, sendo um processo multidisciplinar envolvendo temas de estatística, banco de dados, IA, ML, dentre outros (SCHEMBERGER *et al.*, 2017).

Segundo Fayyad *et al.* (1996) o processo do KDD se divide nas seguintes etapas: (i) entendimento do domínio da aplicação e identificação do objetivo do processo; (ii) criação do

dataset alvo (seleção); (iii) limpeza e pré-processamento de dados; (iv) transformação dos dados; (v) *data mining* para a busca de padrões; (vi) interpretação dos dados minerados e geração do conhecimento.

Além disso, desde 1996 diversas pesquisas validaram o KDD e suas aplicações relacionando-o como uma importante forma de extrair conhecimento de dados e garantir uma maior taxa de acerto em modelos de ML de diversos campos do conhecimento (GUPTA *et al.*, 2018).

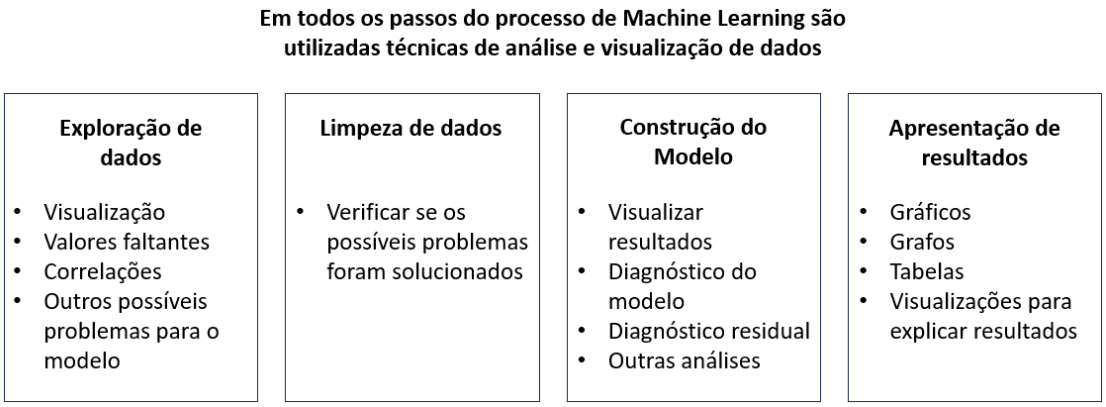
3.3 Análise exploratória de bases de dados (EDA) para ML com *Python*

Segundo STANČIN e JOVIĆ (2019), recomenda-se o uso da ferramenta *Python* para pré-processamento e manipulação de dados, devido à sua comunidade e à ampla gama de bibliotecas, tais como *Matplotlib*, *Seaborn* e *Plotly* para visualização de dados, *pandas*, *numpy* e *scipy* pela capacidade de manipulação de dados e *scikit-learn* e *TensorFlow* para aplicação de modelos de ML.

Além disso, outra biblioteca útil do *Python* no contexto da EDA é a *AutoViz*. Essa ferramenta simplifica e acelera a exploração de dados ao gerar automaticamente diversas visualizações adaptadas às características do conjunto de dados. Com apenas uma linha de código e em poucos segundos, ela permite que o usuário compreenda as nuances iniciais dos dados, eliminando a necessidade de configurações manuais extensas (PRASAD *et al*. 2024).

A EDA é uma abordagem voltada para descobrir o que os dados comunicam sem usar tarefas formais como testes de hipóteses, que permite resumir características estatísticas dos dados, focando nos seguintes aspectos principais: medidas de tendência central (média, moda e mediana), medidas de dispersão (desvio padrão e variância), a forma da distribuição, a identificação de *outliers* e a correlação entre as diferentes variáveis (SAHOO *et al.*, 2019). Na figura 4, destaca-se que componentes chaves da EDA, como análise e visualização de dados, fazem parte do processo de ML desde o seu início.

Figura 4 – EDA para ML



Fonte: Sahoo *et al.* (2019)

Ainda segundo SAHOO *et al.*, 2019 a EDA também possui sua contraparte visual (GEDA, do inglês *graphical exploratory data analysis*) que foca nos mesmos quatro aspectos. Ela pode ser categorizada como univariada quando explora uma variável por vez, utilizando ferramentas como histogramas, gráficos de densidade, *boxplots* e gráficos de caule e folhas; bivariada quando analisa relações entre duas variáveis, usando *boxplots* e *violinplot*; e multivariada quando examina relações entre múltiplas variáveis, recorrendo a gráficos de dispersão 3D, *heatmaps* e *pair plots*.

4 Resultados e discussões

Esta seção contém os resultados obtidos e as discussões de cada uma das etapas do KDD aplicadas no trabalho, além de uma etapa de próximos passos: (i) Dados; (ii) Seleção de Dados; (iii) Pré-Processamento; (iv) Transformação dos Dados; (v) Próximos Passos. Para cada uma dessas etapas adaptou-se em um passo a passo mais específico para o contexto do presente trabalho.

4.1 Dados

A etapa inicial do trabalho contou com o recebimento de dados das estações meteorológicas em formato CSV, fornecidos pela empresa parceira. Com esses dados em mãos, busca-se aprofundar o entendimento do *dataset* para embasar as próximas etapas e viabilizar a aplicação de modelos de ML no futuro. Para isso, dividiu-se essa seção em quatro passos bem definidos que serão explicitados a seguir: entendimento inicial dos dados, EDA inicial com a biblioteca *AutoViz*, EDA individualizada para cada variável do conjunto de dados e definição do objetivo KDD.

4.1.1 Entendimento inicial dos dados

Nessa etapa busca-se adquirir os conhecimentos prévios relevantes para análise. Para isso, buscou-se, inicialmente, extrair o conhecimento da empresa parceira para entender a natureza da base compartilhada. Posteriormente, buscou-se explorar estatisticamente os dados por meio das principais métricas da estatística descritiva.

Assim, foram marcadas conversas com a responsável pelas análises meteorológicas e com o responsável pelo tratamento dos dados da empresa parceira para entender melhor os dados compartilhados. Com isso, identificou-se que a base fornecida já passou por um processamento prévio em que as linhas da base foram agregadas em janelas de 15 minutos. Além disso, foram realizadas correções específicas relacionadas à operação, como o tratamento de medições provenientes de sensores quebrados, eliminação de *outliers* fora do intervalo esperado para os sensores e ajustes de dados inconsistentes causados por manutenções nas estações. Dessa forma, a etapa de pré-processamento, que será feita mais adiante, tende a ser reduzida, já que os dados já receberam um tratamento prévio.

Dado o histórico e a expertise da empresa parceira no setor, decidiu-se aceitar os dados processados com esses tratamentos prévios, mesmo reconhecendo o risco inerente de potenciais erros decorrentes dessas intervenções. Essa abordagem foi adotada com base na confiança no conhecimento técnico da empresa e na relevância dos dados para o projeto em questão.

Além disso, a base de dados recebida possui 375.936 linhas com as variáveis coletadas

em duas fazendas distintas. Para a primeira, apenas uma estação (station_id 17) é responsável pela coleta dos dados, já para a segunda existem seis estações responsáveis. Vale ressaltar também que, segundo a empresa parceira, a primeira fazenda é mais madura (mais tempo de dados consolidados coletados), tem condições meteorológicas e de solo melhores que a primeira. Por fim, observa-se na Tabela 1 quais são as variáveis auferidas pelas estações, sua descrição, qual a unidade de medida, bem como a amplitude de medição dos sensores.

Tabela 1 – Descrição de colunas do dataset

Coluna	Descrição	Unidade de Medida	Amplitude dos sensores	Porcentagem de valores nulos
id	Id			0.00%
station_id	Id da estação metereológica			0.00%
received_at	Janela de medição			0.00%
press	Pressão atmosférica	kPa	30 a 110 kPa	4.24%
mCnpD	Umidade da copa media (medição no topo da estação)	%	0 a 100%	26.59%
mCnpN	Umidade da copa mínima (medição no topo da estação)	%	0 a 100%	26.59%
mCnpX	Umidade da copa máxima (medição no topo da estação)	%	0 a 100%	26.59%
tCnpD	Temperatura da copa media (medição no topo da estação)	°C	-40 a 125°C	26.59%
tCnpN	Temperatura da copa minima (medição no topo da estação)	°C	-40 a 125°C	26.59%
tCnpX	Temperatura da copa máxima (medição no topo da estação)	°C	-40 a 125°C	26.59%
mTopD	Umidade Relativa media	%	0 a 100%	5.43%
mTopN	Umidade Relativa minima	%	0 a 100%	5.43%
mTopX	Umidade Relativa máxima	%	0 a 100%	5.43%
tTopD	Temperatura média	°C	-40 a 125°C	5.43%
tTopN	Temperatura minima	°C	-40 a 125°C	5.43%
tTopX	Temperatura máxima	°C	-40 a 125°C	5.43%
mSoilA	Umidade Relativa do solo com o sensor à uma distância de 10 a 20 cm da superfície do solo.	%	0 a 100%	5.59%
mSoilB	Umidade Relativa do solo com o sensor à uma distância de 20 a 30 cm da superfície do solo.	%	0 a 100%	60.18%
tSoil	Temperatura do solo	°C	-40 a 125°C	7.68%
wDirL	Direção do vento	°	360°	4.26%

wDirX	Direção da rajada do vento (medições máximas para o intervalo de quinze minutos)	°	360°	4.26%
wSpdD	Velocidade do vento	m/s graus	0.5 a 89 m/s	4.26%
wSpdX	Velocidade da rajada do vento (medições máximas para o intervalo de quinze minutos)	m/s graus	0.5 a 89 m/s	4.26%
rainFall	Chuva	mm	250 mm/h	4.55%
solarRad	Radiação Solar	W/m²	0 a 1800 W/m²	4.22%
is_waiting_for_data	Controle interno da empresa parceira			0.00%
estimated_data	Controle interno da empresa parceira			99.86%
sensor_vars_failing	Falha nos sensores			94.80%

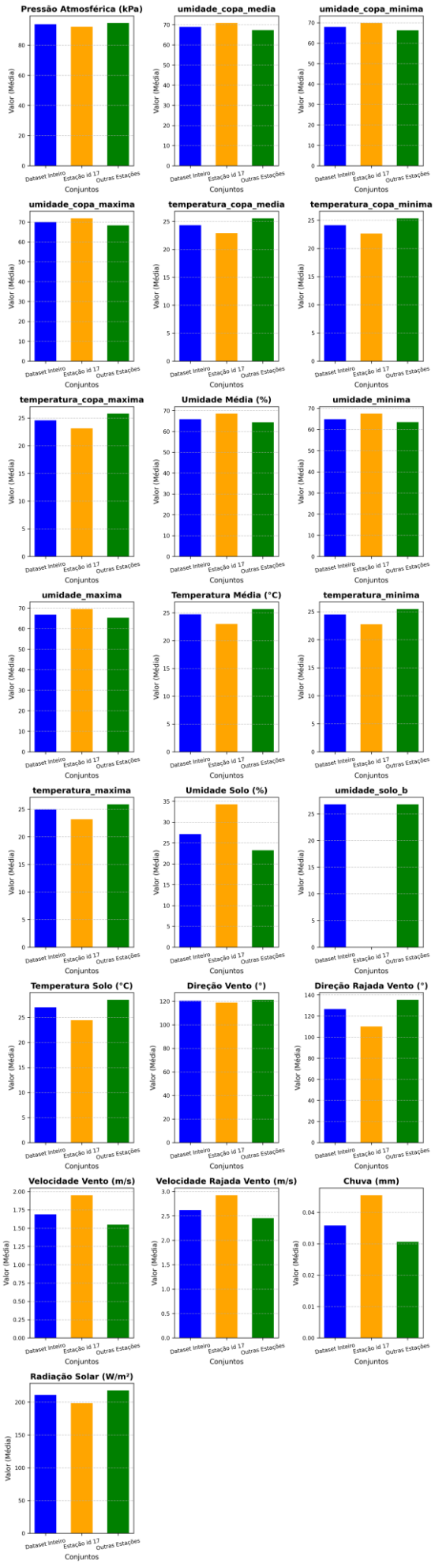
Fonte: Empresa parceira (2024)

Em seguida, busca-se comparar estatisticamente as medições do *dataset* inteiro (375.936 linhas), bem como de cada fazenda, a de estação de id 17 mais madura (127.872 linhas) e a fazenda das demais estações (248.064 linhas). Com isso, embasa-se a seleção de dados, que é a próxima etapa do KDD. Para isso, utilizou-se o método *describe* dos *DataFrames* de cada amostra, adicionando uma coluna de amplitude (valor máximo subtraído do valor mínimo). Vale ressaltar que a umidade do solo B não é medida para a estação de id 17.

Conforme a Figura 5, as médias não variam consideravelmente em cada *dataset*, entretanto, observa-se que para as variáveis de umidade do solo e chuva a amostra da estação de id 17 tem uma média consideravelmente mais alta, o que está em linha com o que foi conversado com a empresa parceira. Já para a figura 6, observa-se os desvios padrões de cada amostra e percebe-se que a estação 17 tem uma menor variação de pressão atmosférica e uma menor variação de umidade do solo e uma maior variação na quantidade de chuvas. Por fim, na figura 7, observa-se a amplitude de cada variável para cada amostra e percebe-se que a amplitude de pressão atmosférica e umidade do solo é menor para a estação de id 17.

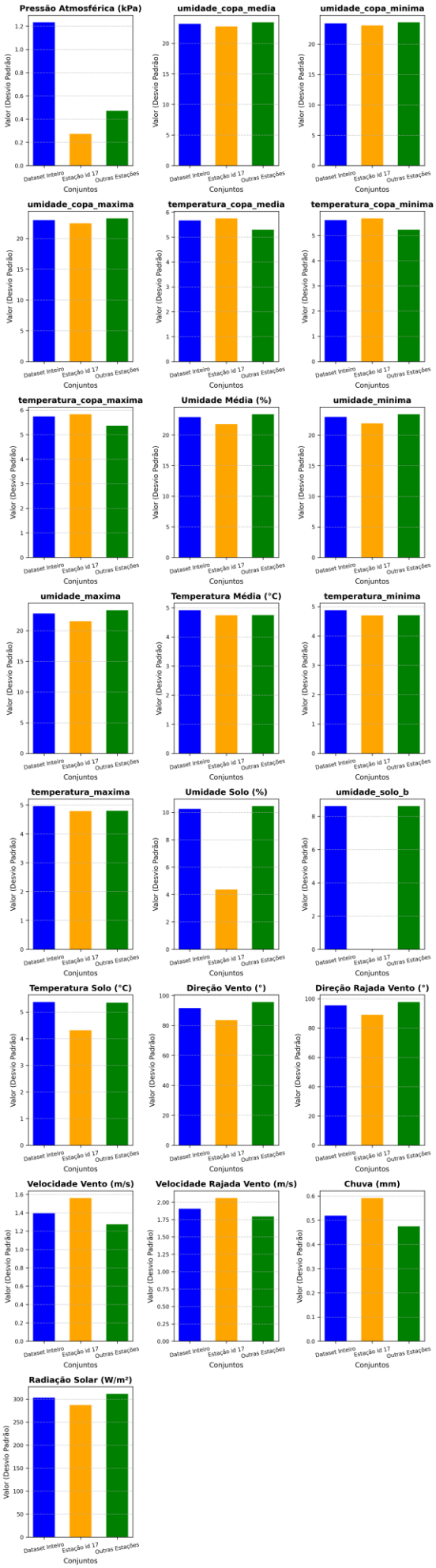
Portanto, a análise preliminar do *dataset* permitiu uma compreensão inicial da base de dados fornecida pela empresa parceira, evidenciando as principais características das medições. Com o entendimento das correções realizadas previamente e a diferenciação das condições das fazendas e das estações de coleta, foi possível realizar uma primeira comparação estatística das variáveis, destacando as diferenças nas médias, desvios padrões e amplitudes das medições entre as diferentes amostras. Esses resultados são fundamentais para embasar a próxima fase do processo de KDD, que será a seleção dos dados, permitindo que as análises subsequentes sejam realizadas com uma base mais precisa e relevante para o trabalho.

Figura 5 – Comparação de métricas: Média



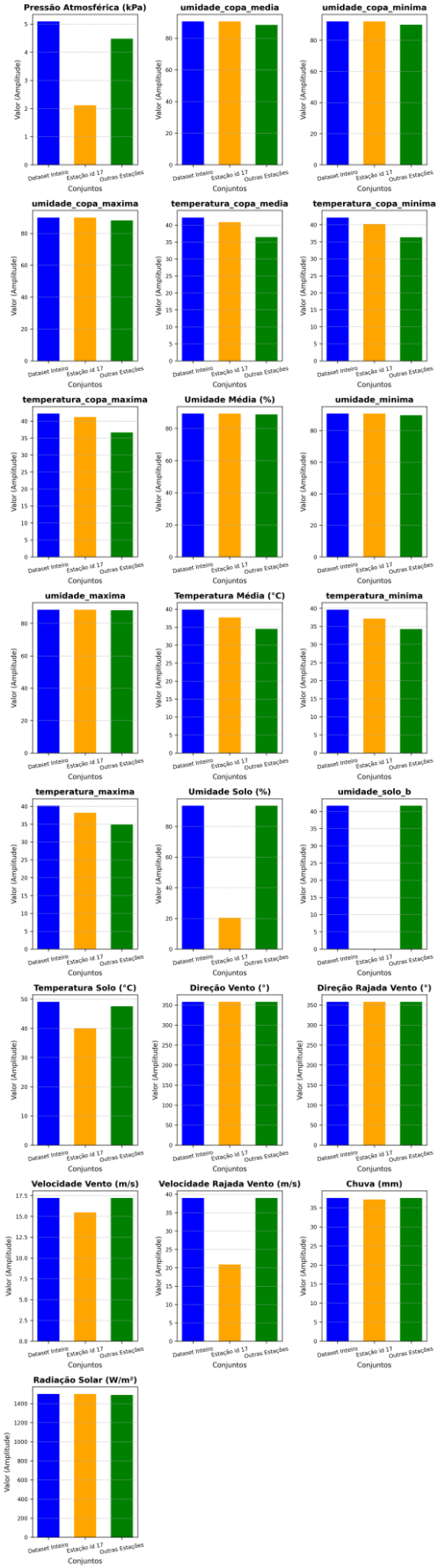
Fonte: Elaboração Própria (2024)

Figura 6 – Comparação de métricas: Desvio Padrão



Fonte: Elaboração Própria (2024)

Figura 7 – Comparação de métricas: Amplitude



Fonte: Elaboração Própria (2024)

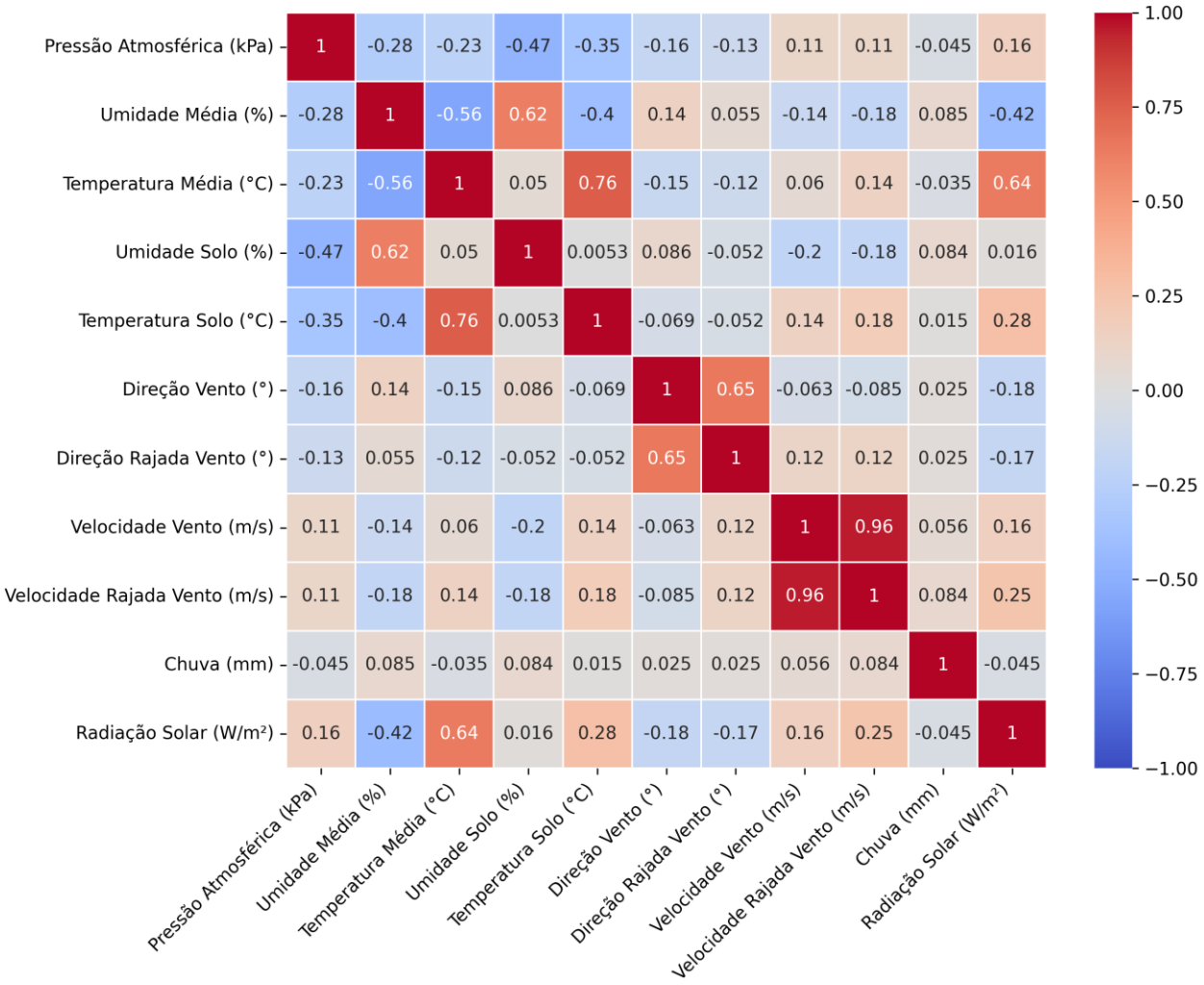
4.1.2 EDA inicial com a biblioteca *AutoViz*

Após um entendimento das principais métricas estatísticas dos dados, segue-se para uma análise exploratória visual, dessa vez buscando entender a correlação entre as variáveis. Para isso, utiliza-se a biblioteca *AutoViz* para gerar diversos gráficos de maneira rápida a fim de uma exploração extensiva dos dados, entretanto, utilizou-se apenas o conjunto de *scatter plots* gerado por ela, já que o *heatmap* foi plotado usando a biblioteca *matplotlib* para maior controle das propriedades do gráfico. Vale destacar que a partir desse momento, todas as análises serão feitas com o conjunto de dados filtrados para a estação de id 17, que será o foco do trabalho. Para maiores explicações, consultar a etapa de seleção dos dados e a discussão da seção 4.1.2.

Como primeira abordagem para entender a correlação entre as variáveis, foram utilizados mapas de calor baseados em dois métodos distintos de correlação: o de *Pearson*, que avalia relações lineares entre as variáveis, e o de *Spearman*, que não depende de linearidade. Essa análise exploratória busca comparar os resultados obtidos por ambos os métodos, permitindo identificar aquele que melhor se adapta ao conjunto de dados analisado.

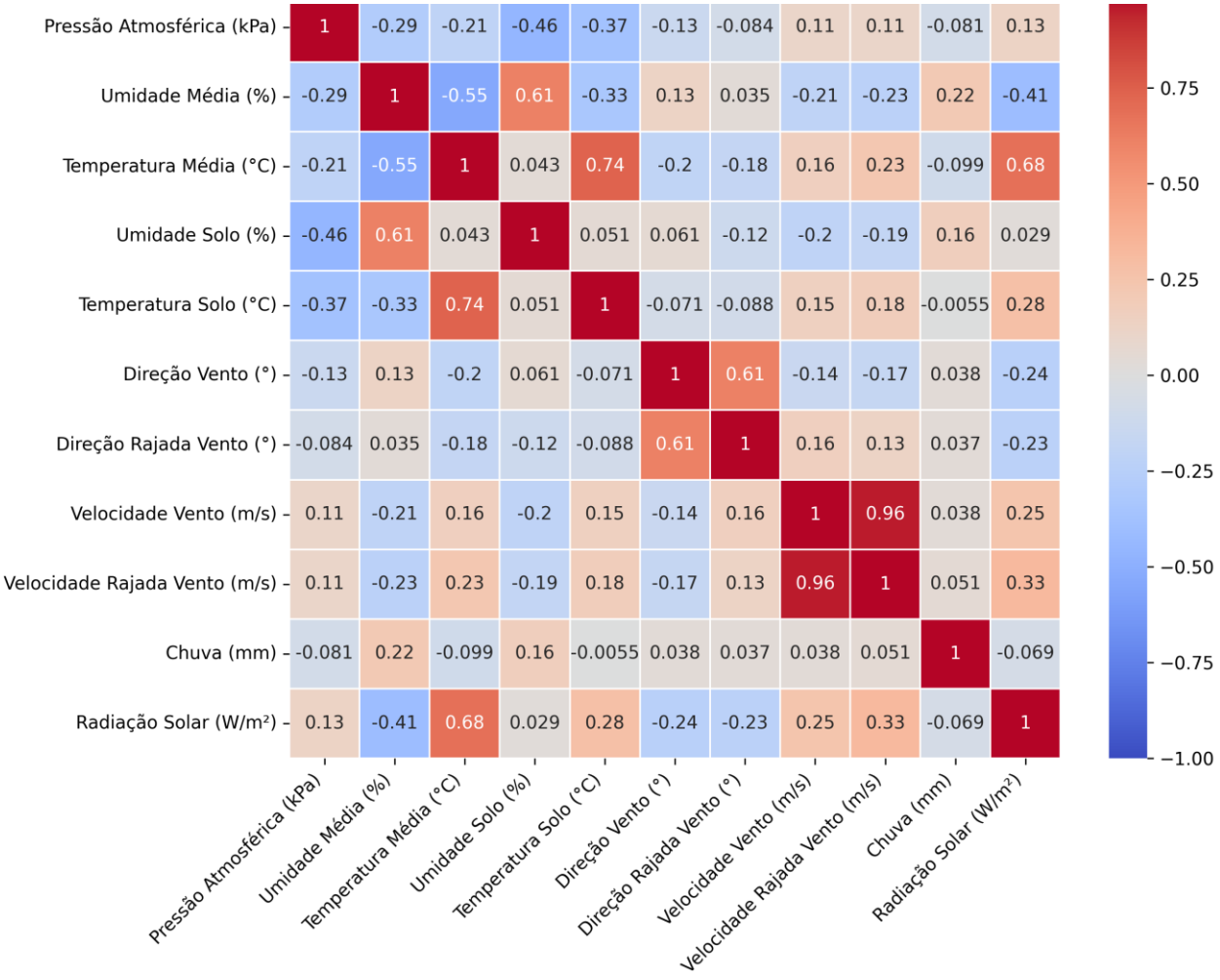
Inicialmente, foi utilizado o método de correlação de Pearson para identificar as relações entre as variáveis. Os resultados, apresentados na Figura 8, mostram que as variáveis mais correlacionadas são: Umidade do Solo (%) com Umidade Média (%), Temperatura do Solo (°C) com Temperatura Média (°C), e Temperatura Média (°C) com Radiação Solar (W/m²). Por outro lado, algumas variáveis apresentaram correlação média negativa, como: Temperatura Média (°C) com Umidade Média (%), Radiação Solar (W/m²) com Umidade Média (%), e Umidade do Solo (%) com Pressão Atmosférica (kPa). Em seguida, o método de correlação de Spearman foi aplicado, gerando os resultados mostrados na Figura 9. Os resultados foram bastante similares aos obtidos com o método de Pearson, com algumas pequenas variações que não geram *insights* diferentes dos citados acima.

Figura 8 – Mapa de Calor pelo método de *Pearson*



Fonte: Elaboração própria (2024)

Figura 9 – Mapa de calor para o método de *Spearman*



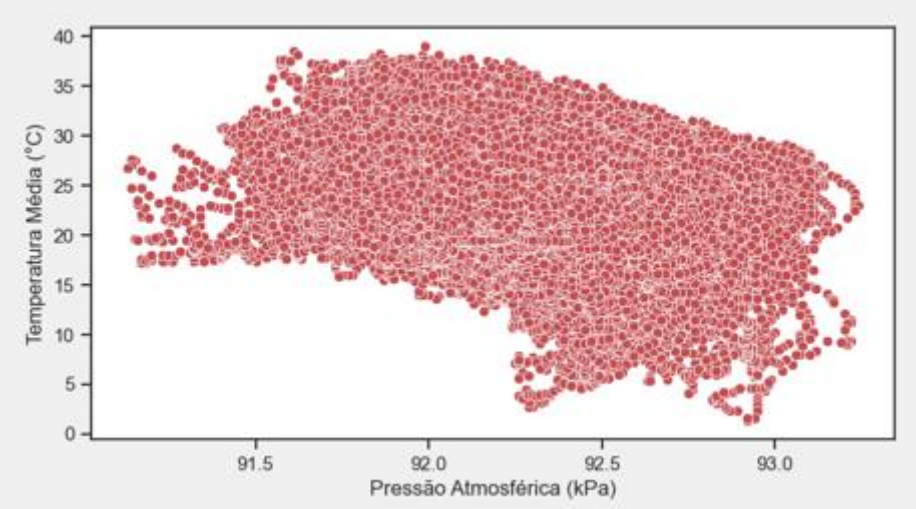
Fonte: Elaboração própria (2024)

Além disso, destaca-se que correlação não implica causalidade, e coeficientes baixos para ambos não significa necessariamente ausência de relação. Algumas das variáveis podem ter relações que não são bem representadas pelas correlações de *Pearson* e *Spearman.* Por isso, a segunda abordagem para compreender a correlação das variáveis foi por meio de *scatter plots* gerados pela biblioteca *AutoViz*, que comparam duas variáveis por vez, com uma sendo representada no eixo x e a outra no eixo y. A análise desses gráficos permitiu identificar os seguintes padrões não lineares:

- Pressão atmosférica e temperatura média: A pressão atmosférica tende a reduzir à medida que a temperatura média diminui (Figura 10).
- Pressão atmosférica e chuva: Valores centrais na distribuição de pressão atmosférica estão associados a maior incidência de chuva (Figura 11).
- Temperatura média e umidade relativa do ar: A umidade média relativa do ar diminui com o aumento da temperatura média (Figura 12).

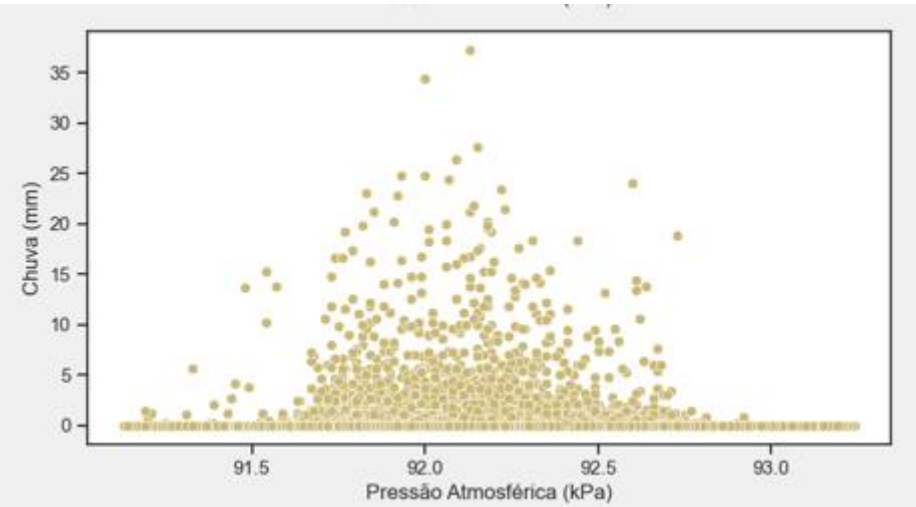
- Chuva e umidade relativa do ar: A umidade média relativa do ar aumenta com o volume de chuva (Figura 13).
- Temperatura média do solo e do ar: A temperatura média do solo eleva-se conforme a temperatura média do ar aumenta (Figura 14).
- Temperatura média do ar e chuva: Os valores medianos de temperatura média do ar apresentam maior ocorrência de chuva (Figura 15).
- Radiação solar e temperatura média do ar: A temperatura média do ar cresce com o aumento da radiação solar (Figura 16).
- Radiação solar e chuva: A alta radiação solar está associada a menores volumes de chuva (Figura 17).

Figura 10 – *Scatter Plot* Pressão x Temperatura



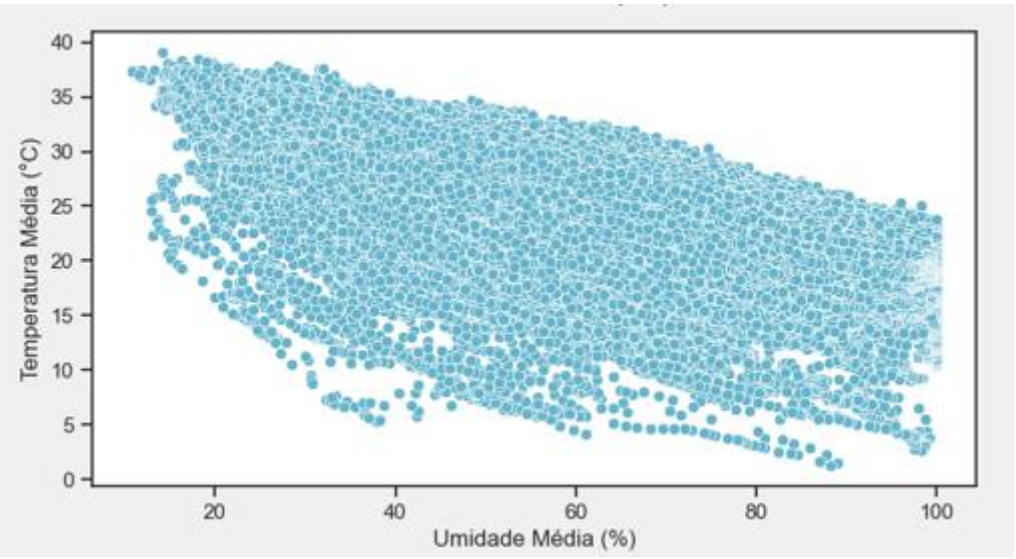
Fonte: Elaboração própria (2024)

Figura 11 – *Scatter Plot* Pressão x Chuva



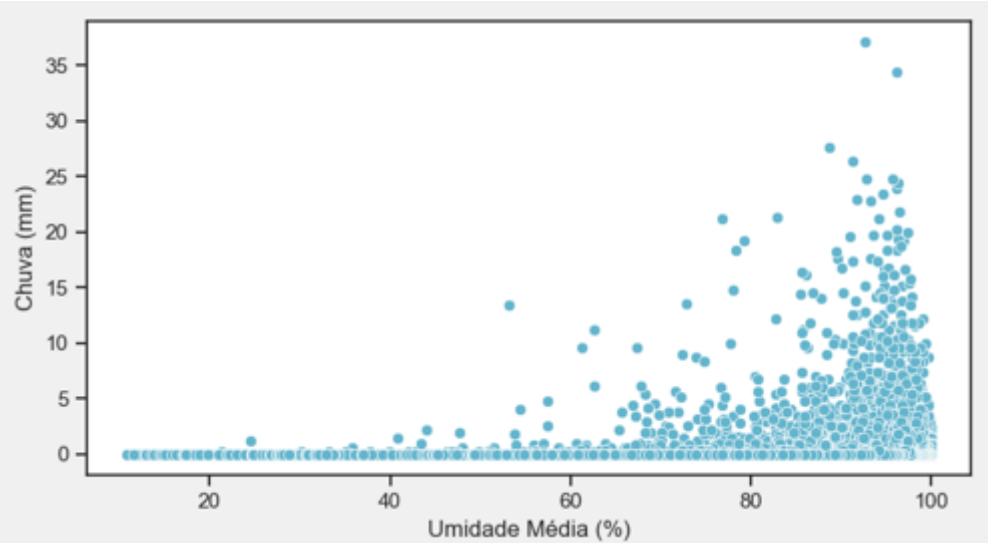
Fonte: Elaboração própria (2024)

Figura 12 – *Scatter Plot* Umidade do ar x Temperatura



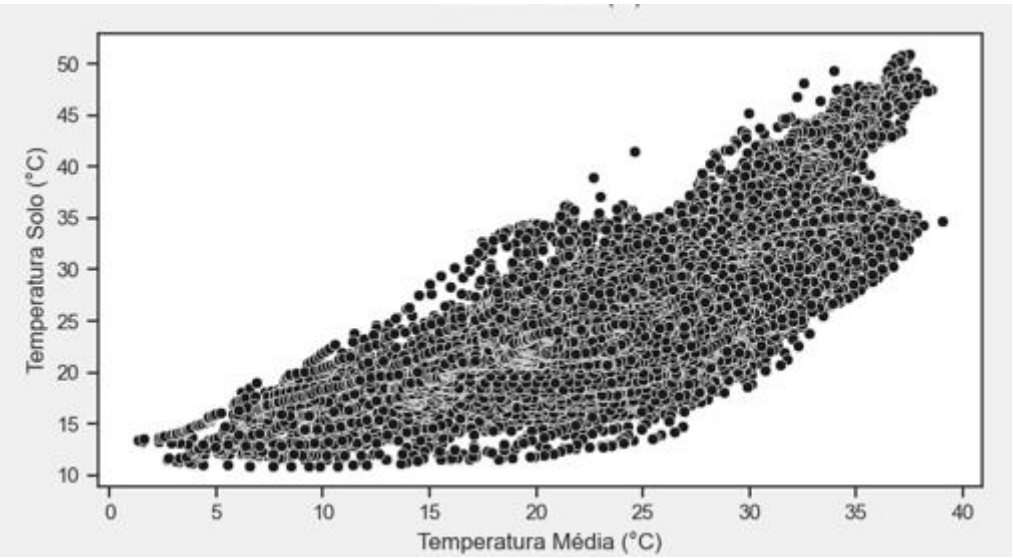
Fonte: Elaboração própria (2024)

Figura 13 – *Scatter Plot* Umidade do ar x Chuva



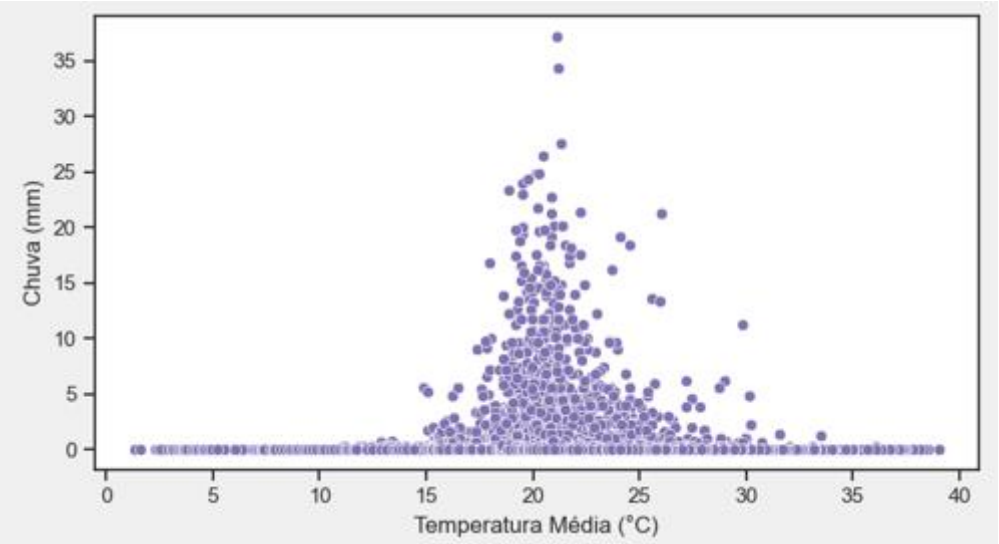
Fonte: Elaboração própria (2024)

Figura 14 – *Scatter Plot* Temperatura x Temperatura do Solo



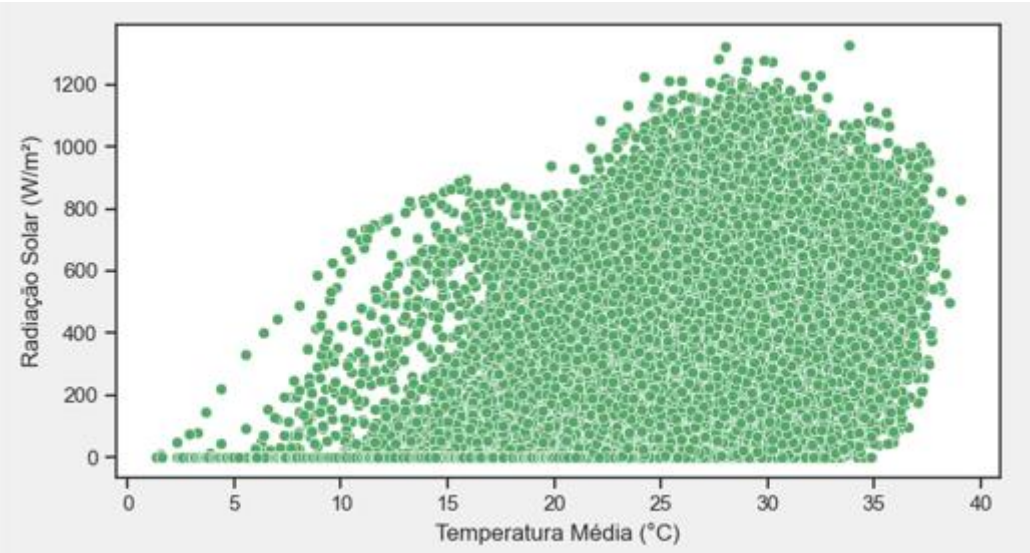
Fonte: Elaboração própria (2024)

Figura 15 – *Scatter Plot* Temperatura x Chuva



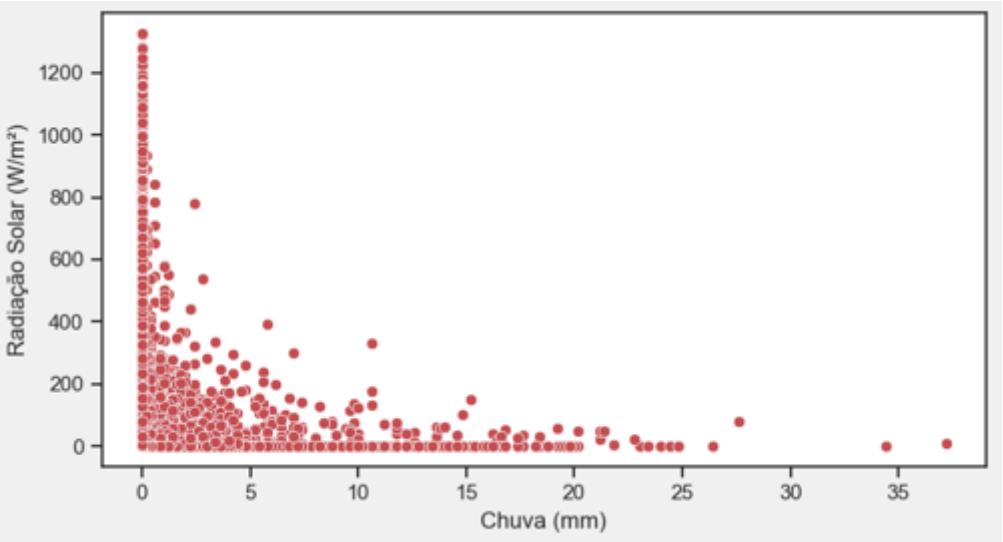
Fonte: Elaboração própria (2024)

Figura 16 – *Scatter Plot* Temperatura x Radiação Solar



Fonte: Elaboração própria (2024)

Figura 17 – *Scatter Plot* Chuva x Radiação Solar



Fonte: Elaboração própria (2024)

Nesta etapa inicial, foi realizada uma EDA com o objetivo de entender as correlações entre as variáveis, utilizando os métodos de correlação de *Pearson* e *Spearman*, além de gráficos de dispersão para identificar padrões não lineares. A relevância dessas visualizações está em proporcionar um entendimento mais profundo sobre as relações entre as variáveis, o que permite uma exploração mais informada do *dataset*. Essa análise prepara o caminho para as etapas subsequentes do KDD, destacando como as variáveis se influenciam e orientando as análises futuras.

4.1.3 EDA individualizada para cada variável

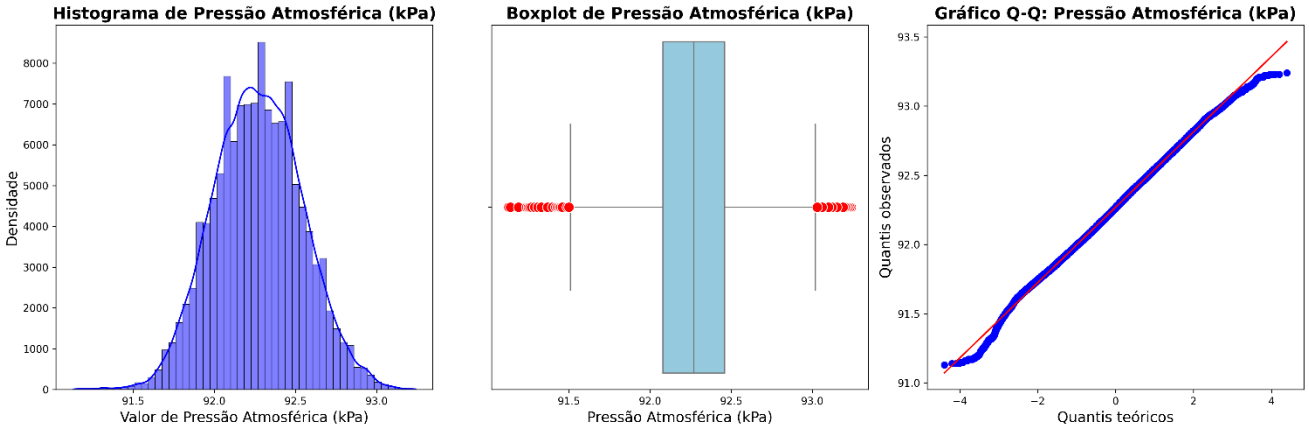
Após compreender a relação entre as variáveis, busca-se compreender a distribuição dos dados por meio de gráficos como histogramas, *boxplots* e gráficos Q-Q para cada variável. Adicionalmente, a sazonalidade dos dados é analisada utilizando um gráfico de linha mensal.

Iniciando pela pressão atmosférica, embora a Figura 18 indique uma boa aproximação visual à normalidade, os testes formais fornecem evidências estatísticas para rejeitar a hipótese de normalidade, conforme Tabela 2. Esse resultado pode ser explicado pela alta sensibilidade dos testes a pequenos desvios, especialmente em grandes amostras. No entanto, os valores de assimetria e curtose estão dentro da faixa considerada aceitável para uma distribuição normal (são menores que 0,5), conforme Tabela 3. Assim, embora a distribuição não seja perfeitamente normal, pode-se assumir normalidade para as análises dessa variável dada sua proximidade à distribuição normal.

Uma distribuição próxima da normalidade pode sugerir que os padrões climáticos são

relativamente estáveis. Isso ocorre porque a pressão atmosférica é influenciada por grandes sistemas meteorológicos, como massas de ar e frentes frias. Na ausência de variações bruscas ou frequentes nesses sistemas, é esperado que a pressão atmosférica se concentre ao redor de um valor médio local, aproximando-se de uma distribuição normal. Com base nisso, uma análise interessante seria investigar os desvios da normalidade, pois esses poderiam indicar influências sazonais, eventos climáticos extremos ou até mudanças climáticas pontuais.

Figura 18 – Gráficos de Pressão atmosférica



Fonte: Elaboração própria (2024)

Tabela 2 – Resultados dos testes de normalidade para pressão atmosférica

Teste	Estatística	Valor-P	Rejeita H0
Shapiro-Wilk	0,9993	0,0000	Sim
Kolmogorov-Smirnov	0,0155	0,0000	Sim
D'Agostino-Pearson	45,9083	0,0000	Sim

Fonte: Elaboração própria (2024)

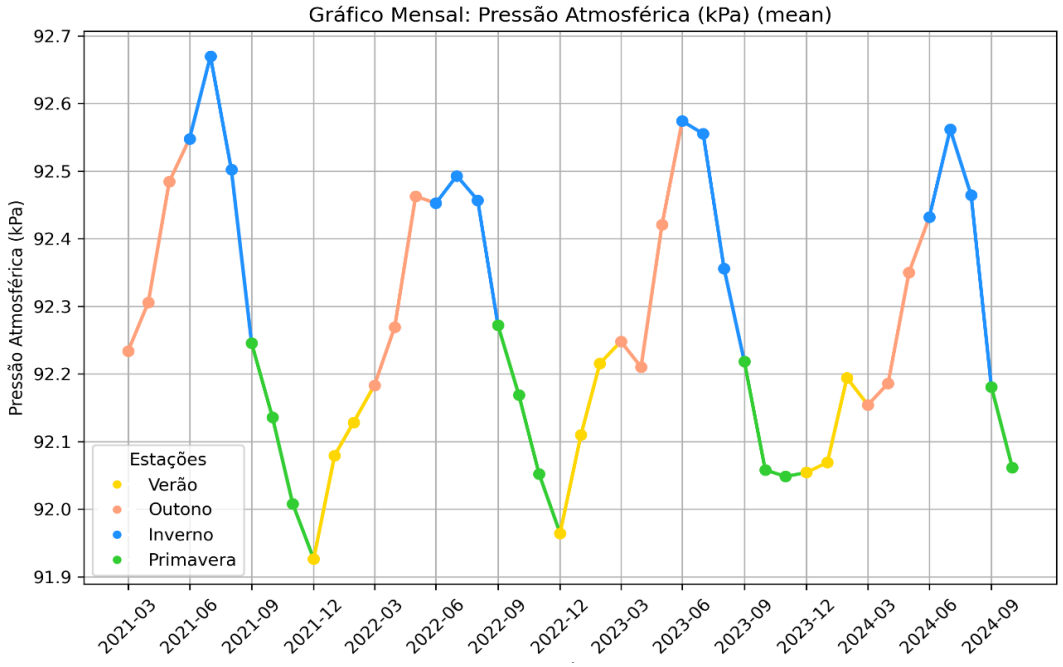
Tabela 3 – *Skewness* e *curtosis* para distribuição de pressão atmosférica

Skewness	0,0244
Curtosis	0,0770

Fonte: Elaboração própria

Assim, para entender a sazonalidade da pressão atmosférica, plotou-se a Figura 19, tirando a média mensal das pressões auferidas. Percebe-se um padrão que em dezembro durante todo o período do *dataset* a pressão atmosférica alcança as mínimas do ano, e por outro lado, entre o mês de junho e julho alcança as máximas, o que está relacionado às estações do ano verão e inverno no Brasil.

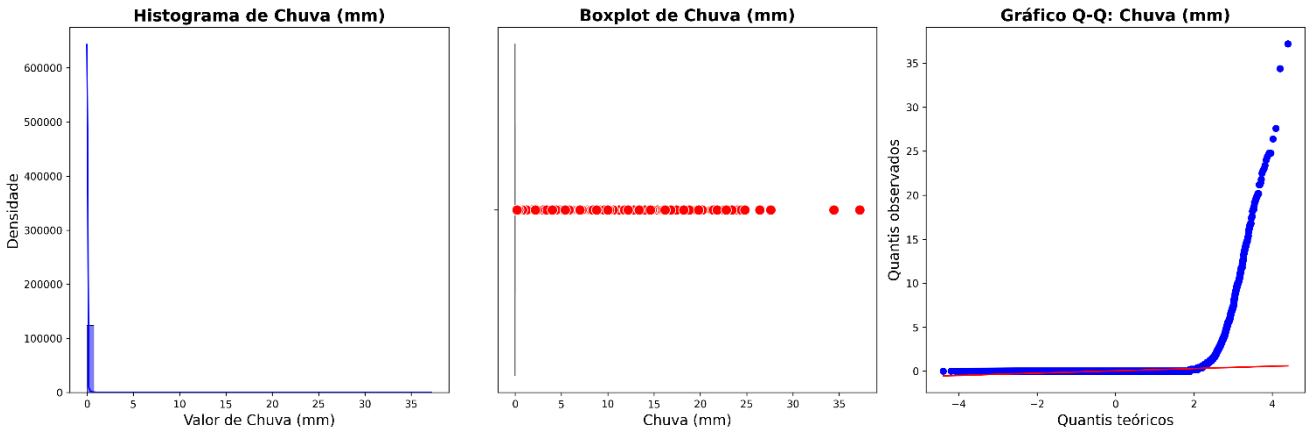
Figura 19 – Gráfico mensal de pressão atmosférica



Fonte: Elaboração própria (2024)

Para a chuva, observa-se, na Figura 20, que o valor mais frequente é zero e que foram identificados 3.600 valores extremos que o *boxplot* classificou como *outliers* estatísticos. Nesses casos, esses valores não devem ser considerados *outliers*, pois estão dentro da amplitude de medição dos sensores, e existem dias com índices de precipitação maiores do que a média dos outros dias (conferir seção 4.3.2). Além disso, é importante destacar que a distribuição dos dados não segue um padrão normal.

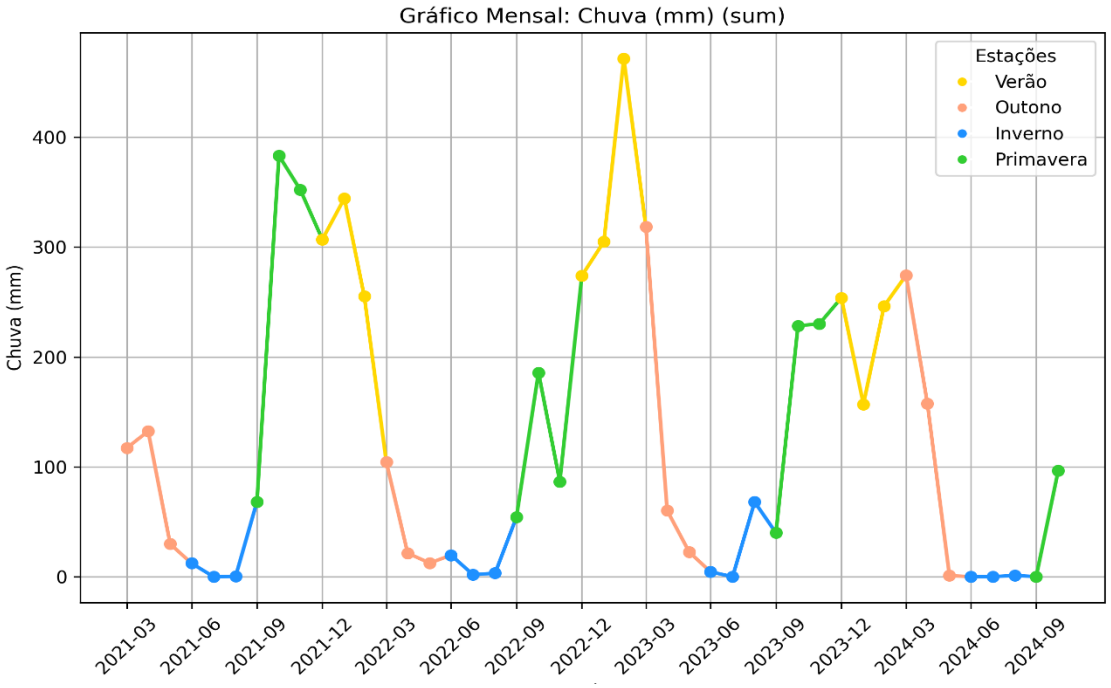
Figura 20 – Gráficos de chuva



Fonte: Elaboração própria (2024)

Para uma melhor compreensão da sazonalidade da chuva, que impacta a distribuição dos dados, foi construído um gráfico mostrando a quantidade de chuva mensal, conforme Figura 21. Os dados foram agrupados por mês e agregados pela soma total de precipitação. Observou-se que os períodos de maior volume de chuva ocorreram no verão e na primavera, enquanto os de menor volume ocorreram no inverno e no outono, chegando a valores próximos de zero.

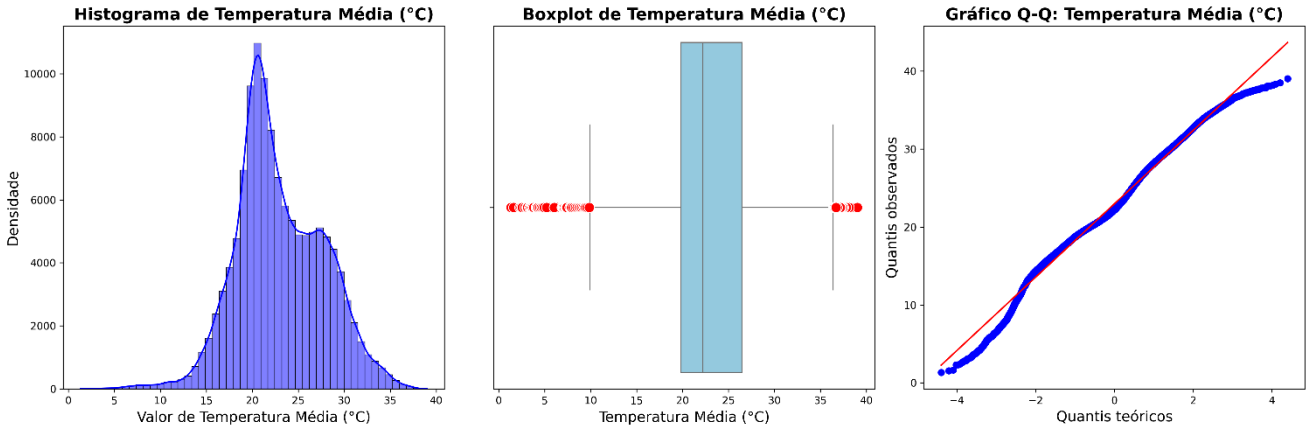
Figura 21 – Gráfico mensal de chuva



Fonte: Elaboração própria (2024)

Para a temperatura média, construiu-se os gráficos da Figura 22. Observa-se uma assimetria, com uma cauda maior à direita, e a maior densidade de valores está em torno de 20°C. Além disso, foram identificados 854 outliers, com valores superiores a 35°C e inferiores a 10°C. No entanto, esses valores estão dentro da amplitude de medição dos sensores e podem ser legítimos, portanto, devem ser considerados na análise (conferir seção 4.3.2). Dessa forma, pode-se concluir que a distribuição não é normal.

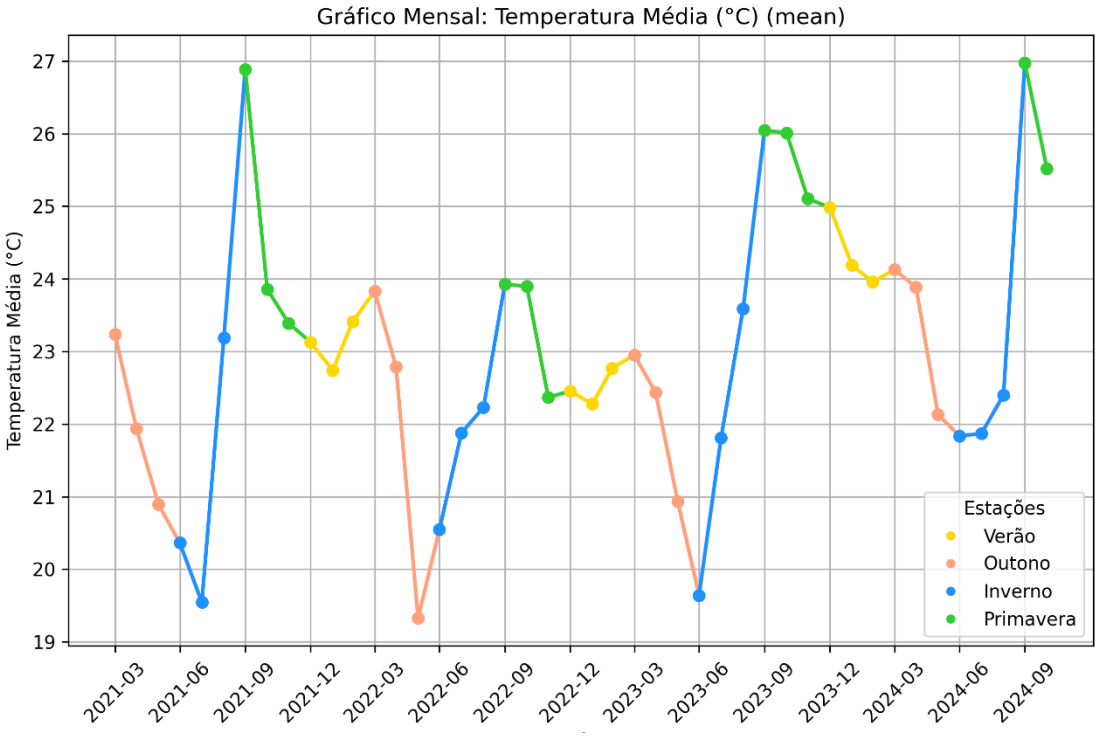
Figura 22 – Gráficos de temperatura média



Fonte: Elaboração própria (2024)

Além disso, os dados foram agrupados em intervalos mensais e calculada a média para cada intervalo, conforme mostrado no gráfico temporal, conforme Figura 23. Assim como a pressão atmosférica, esse gráfico apresenta uma alta correlação com as estações do ano, já que entre setembro e março as temperaturas são mais elevadas, enquanto entre março e julho as temperaturas são mais baixas.

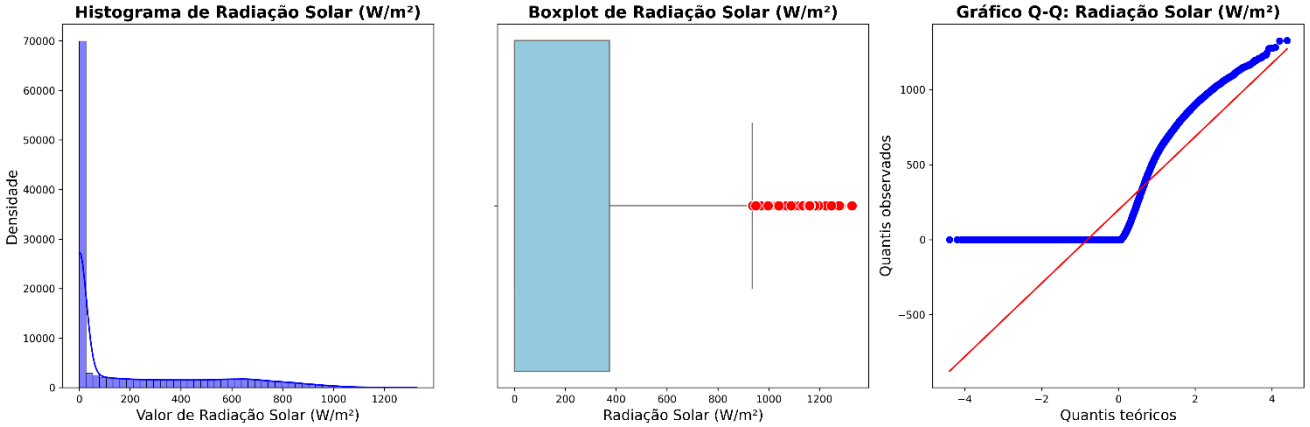
Figura 23 – Gráfico mensal de temperatura média



Fonte: Elaboração própria (2024)

Para a radiação solar, conforme a Figura 24, observa-se que o valor mais frequente é zero. Além disso, a distribuição dos dados não segue um padrão normal. O *boxplot* foi gerado com fins exploratórios, já que valores extremos, como valores muito baixos ou muito altos de radiação, não são considerados *outliers*, pois estão dentro da amplitude de medição dos sensores (conferir seção 4.3.2). Esses valores podem ocorrer em períodos de transição entre o dia e a noite ou em dia com nuvens muito espessas, e devem ser considerados na análise, com ressalva das medições realizadas durante a noite. Dessa forma, recomenda-se agregar a soma das radiações solares para o período de análise e evitar assim a influência das medições noturnas com o consolidado diário.

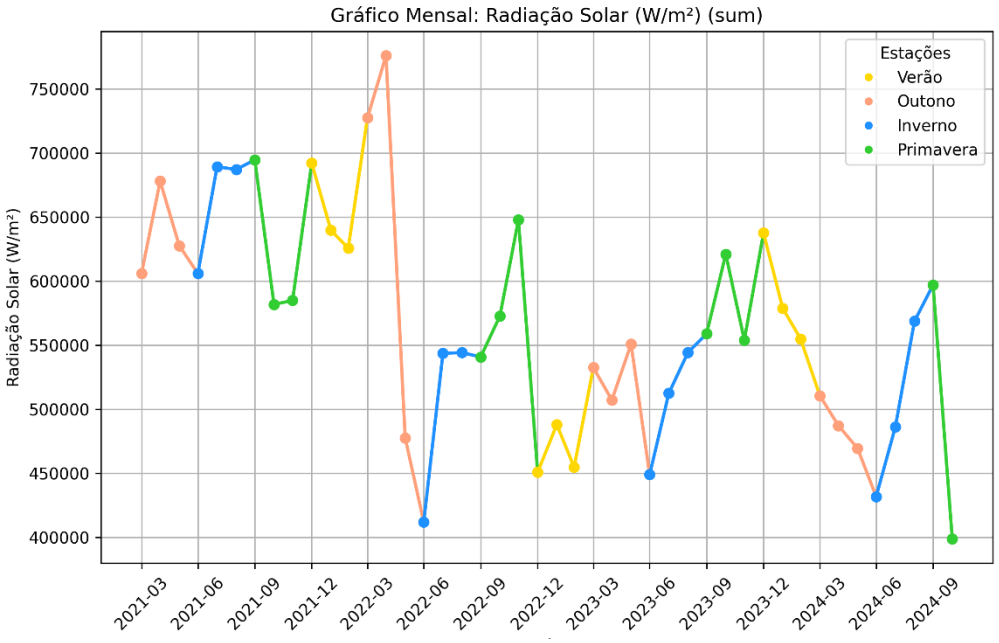
Figura 24 – Gráficos de radiação solar



Fonte: Elaboração própria (2024)

Para o gráfico temporal, os dados foram agrupados mensalmente pela soma e o gráfico da Figura 25 foi construído. No caso da radiação solar, não foi encontrada uma relação direta entre os valores e as estações do ano. No entanto, como na visualização a radiação solar é somada mensalmente, as medições durante a noite não prejudicam a visualização da distribuição dos dados, o que não impacta negativamente a análise.

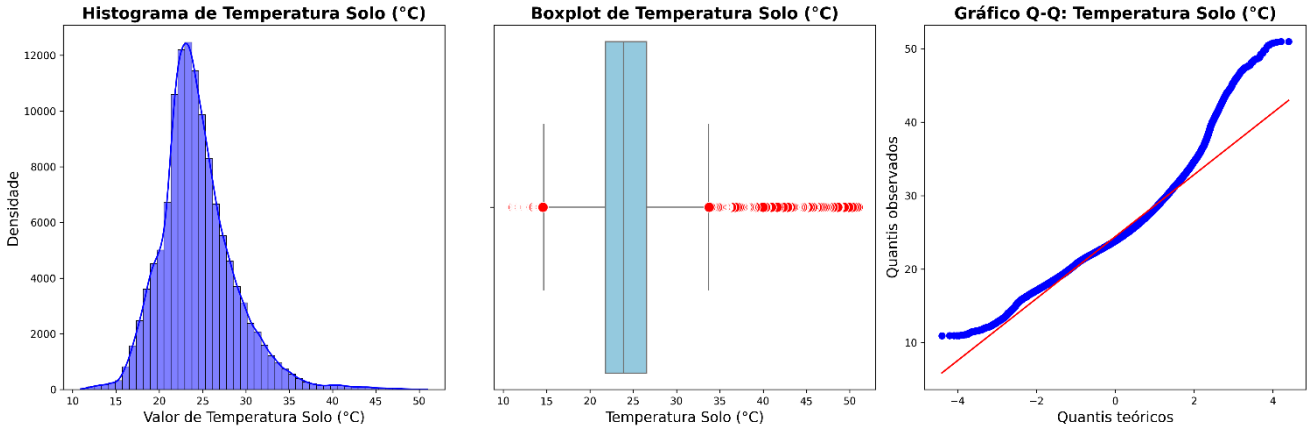
Figura 25 – Gráfico mensal de radiação solar



Fonte: Elaboração própria

Para a temperatura do solo, observa-se na Figura 26 uma distribuição assimétrica, com o pico centrado entre 20°C e 25°C. A distribuição apresenta uma cauda à direita, sugerindo que, embora a maior parte dos dados estejam concentrados em torno de valores mais baixos, existem valores mais altos, com temperaturas superiores a 30°C, associadas a dias mais quentes. Além disso, os dados não seguem uma distribuição normal.

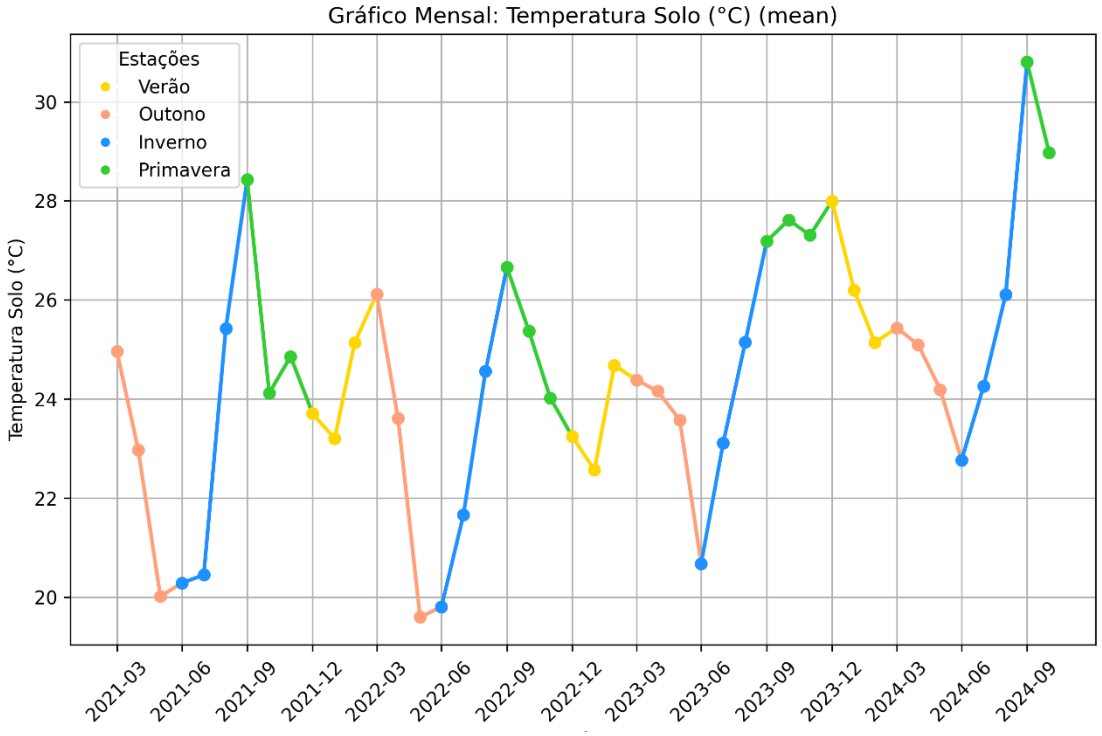
Figura 26 – Gráficos de temperatura do solo



Fonte: Elaboração própria (2024)

Já a distribuição temporal das médias das temperaturas do solo da Figura 27 indica que a variável acompanha as estações conforme a temperatura média, com picos no verão e mínimas no inverno.

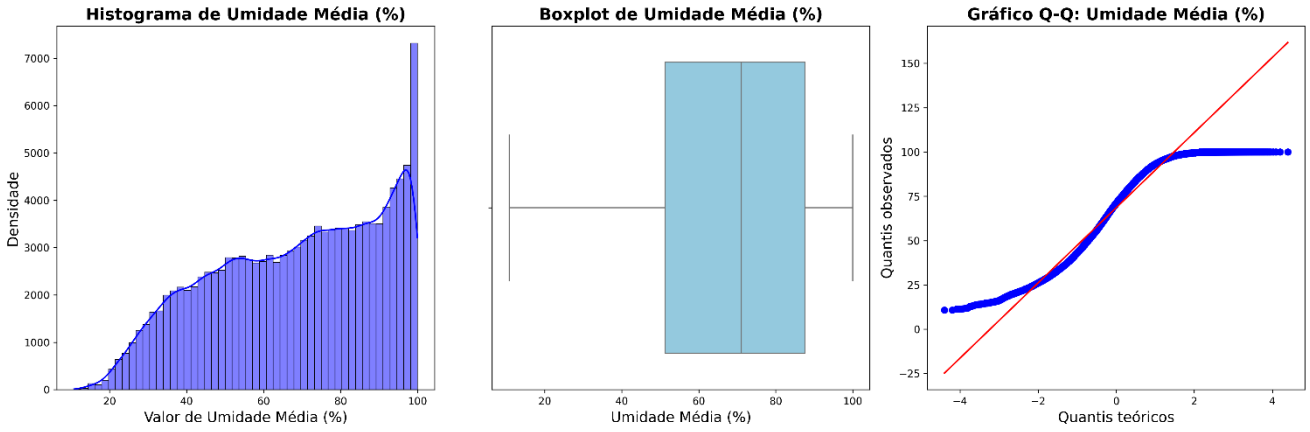
Figura 27 – Gráfico mensal de temperatura do solo



Fonte: Elaboração própria (2024)

Para a umidade média relativa do ar observa-se na Figura 28 uma assimetria com uma cauda consideravelmente maior à direita, o que indica que a estação se localiza em uma região com maiores períodos de umidade relativa do ar elevada. Não se trata de uma distribuição normal e não possui *outliers*, segundo o *boxplot*.

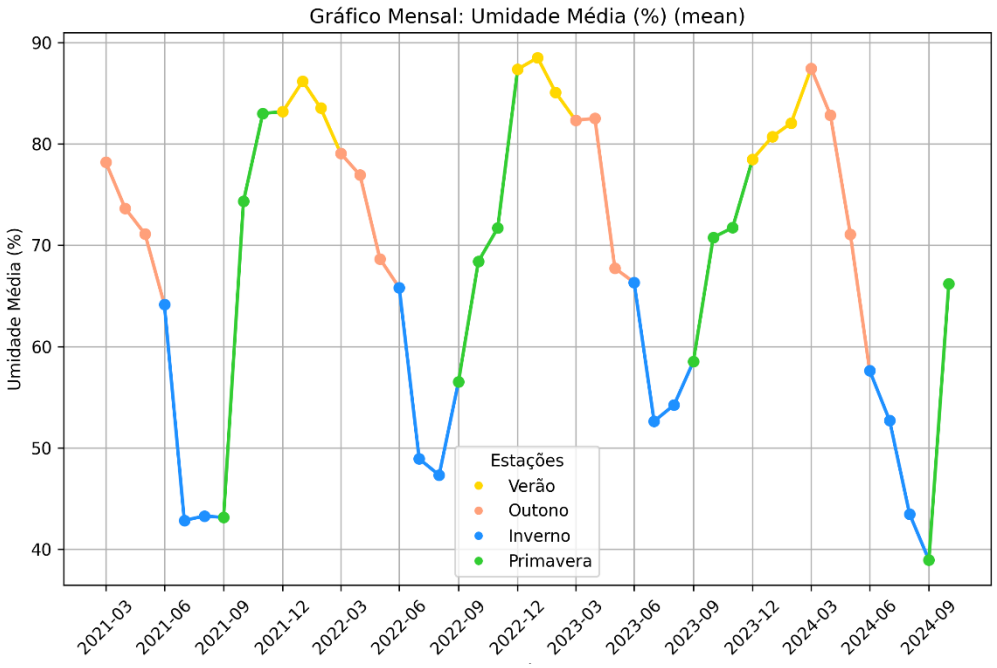
Figura 28 – Gráfico mensal umidade média relativa do ar



Fonte: Elaboração própria (2024)

Construiu-se também um gráfico temporal na Figura 29 para entender a distribuição da variável ao longo dos meses. A umidade relativa do ar, assim como as outras variáveis, tem picos durante o verão e primavera e mínimas durante o inverno.

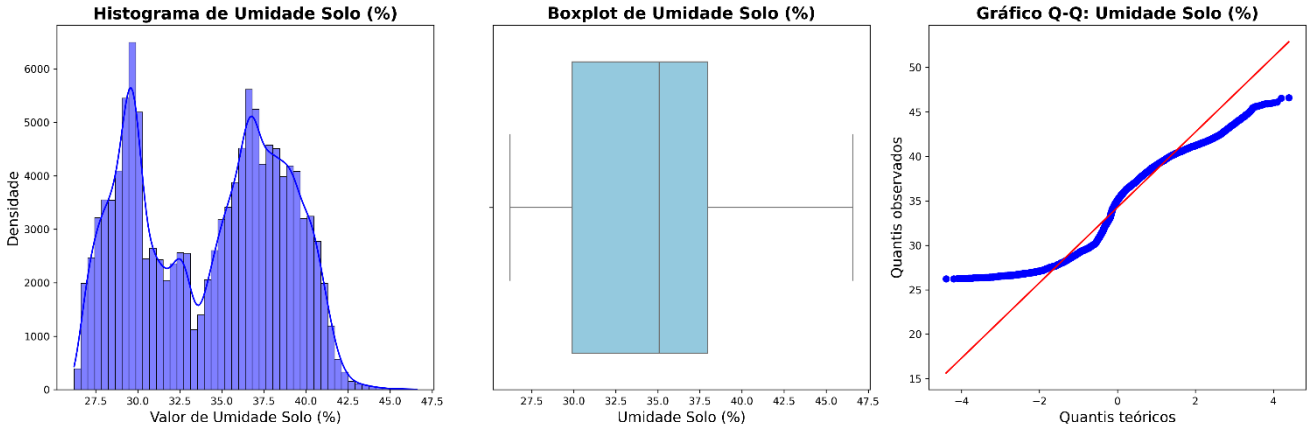
Figura 29 – Gráfico mensal da umidade média relativa do ar



Fonte: Elaboração própria (2024)

Para a umidade do solo, observa-se na Figura 30 uma distribuição bimodal, com dois principais picos em torno de 29% e 37%. Isso sugere que os dados podem refletir dois grupos distintos de variação, como, por exemplo, irrigação e chuva. O pico de 29% pode estar associado a condições de seca natural, períodos entre chuvas ou irrigação. Já o pico de 37% pode indicar períodos pós-chuva ou pós-irrigação. Além disso não foram detectados *outliers* pelo *boxplot*.

Figura 30 – Gráficos umidade do solo



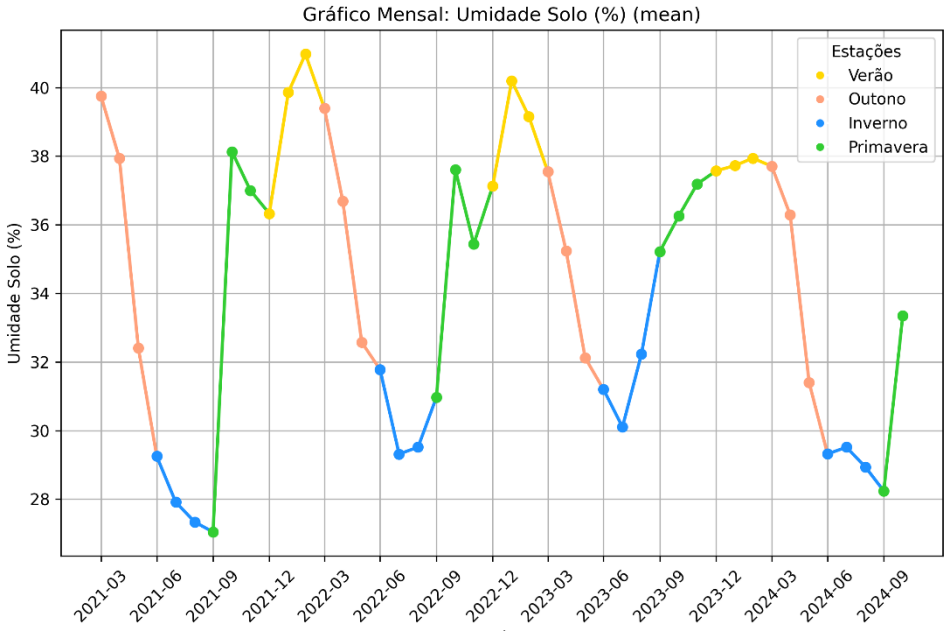
Fonte: Elaboração própria (2024)

Para entender melhor se a irrigação poderia ser uma das causas da distribuição bimodal observada, foi construído um gráfico mensal na Figura 31, agregando os dados pelas médias.

Nesse gráfico, é possível perceber que a variável segue o padrão das estações do ano e, consequentemente, das chuvas. Se a irrigação fosse um fator determinante, a amplitude observada no gráfico não seria tão elevada. Assim, as estações do ano têm forte influência na umidade do solo no *dataset* estudado

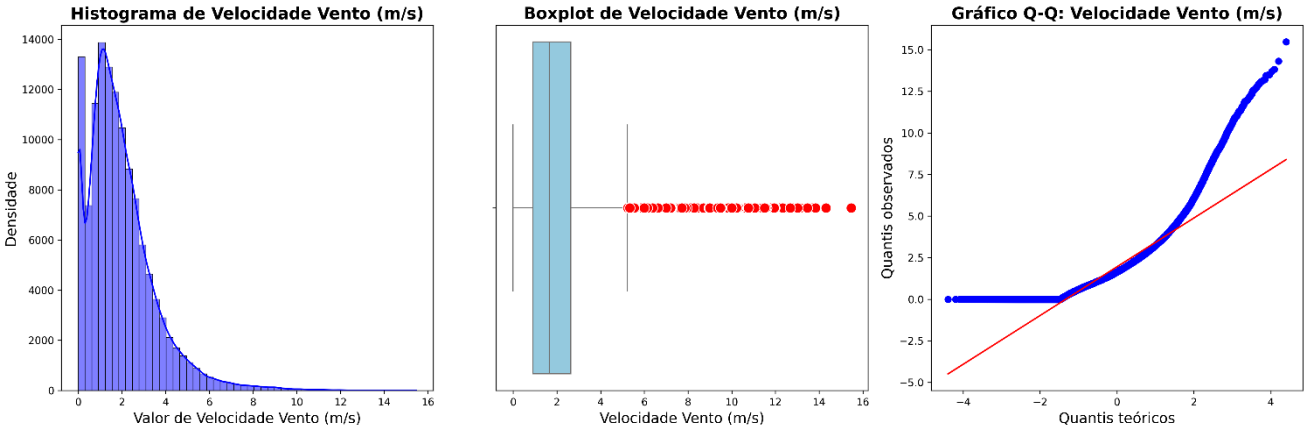
Fonte: Elaboração própria (2024)

Figura 31 – Gráfico mensal de umidade do solo



Para a velocidade do vento nota-se, na Figura 32, uma distribuição assimétrica com concentração de valores próximos a 0 m/s, sugerindo uma predominância de dias com vento moderado. A presença de outliers à direita, com ventos superiores a 5 m/s, indica a ocorrência de tempestades ou ventanias que devem ser incluídos na análise, já que estão dentro do limite de amplitude dos sensores. A distribuição não é normal, o que é consistente com a presença de uma cauda longa, indicando que ventos fortes são menos frequentes, mas ainda assim ocorrem esporadicamente.

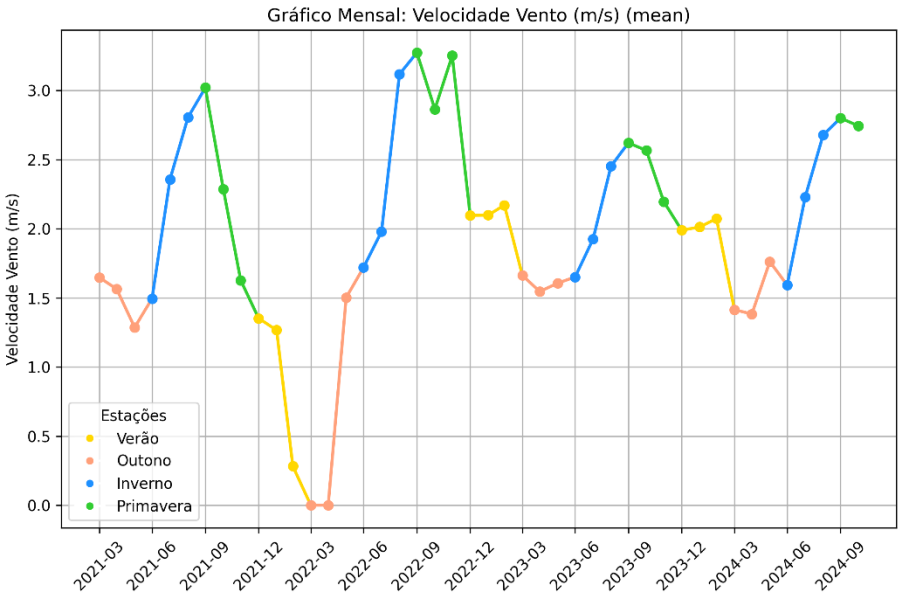
Figura 32 – Gráficos de velocidade do vento



Fonte: Elaboração própria (2024)

Para maior entendimento, construiu-se o gráfico mensal da Figura 33 agregando pelas médias. Para a velocidade do vento nota-se um comportamento um pouco diferente das outras variáveis analisadas até aqui. Os maiores valores se dão durante o inverno e a primavera, enquanto os menores valores se dão durante o verão e outono. O gráfico indica também que nos meses março e abril de 2022 todas as medições de vento foram zeradas (não apenas para essa variável, mas também para as outras relacionadas das Figuras 33, 37 e 39), o que é um forte indício de falhas nos medidores e que esses meses foram *outliers*. Posteriormente, na etapa de pré-processamento aborda-se esse tema com maior profundidade.

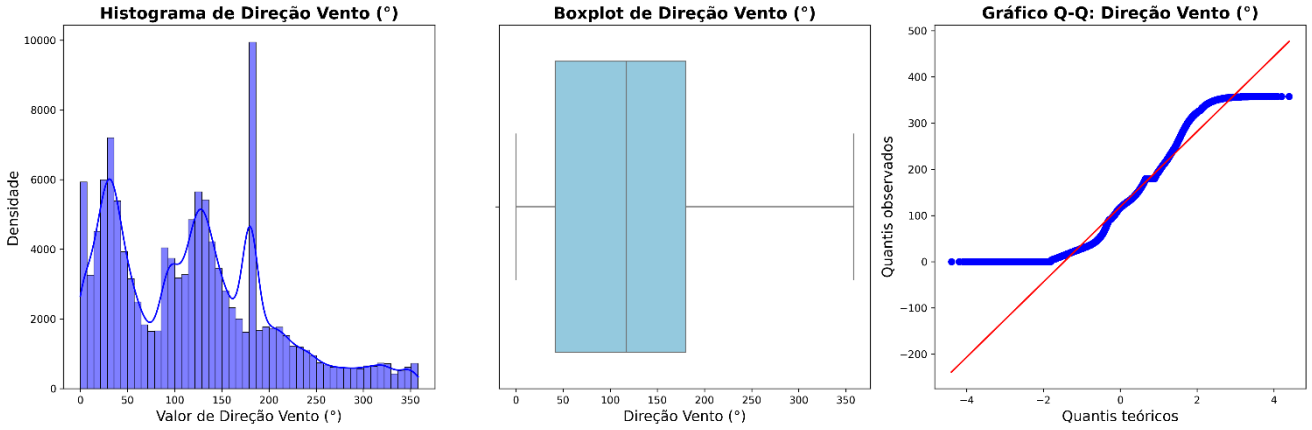
Figura 33 – Gráfico mensal de velocidade do vento



Fonte: Elaboração própria (2024)

Para a direção do vento, observa-se na Figura 34 uma distribuição multimodal, com picos em várias direções. Isso sugere que o vento tende a ser predominantemente orientado em algumas direções específicas, possivelmente refletindo padrões locais de vento ou influências geográficas, como a presença de montanhas, planícies ou massas de ar dominantes. Além disso, a distribuição não segue um padrão normal.

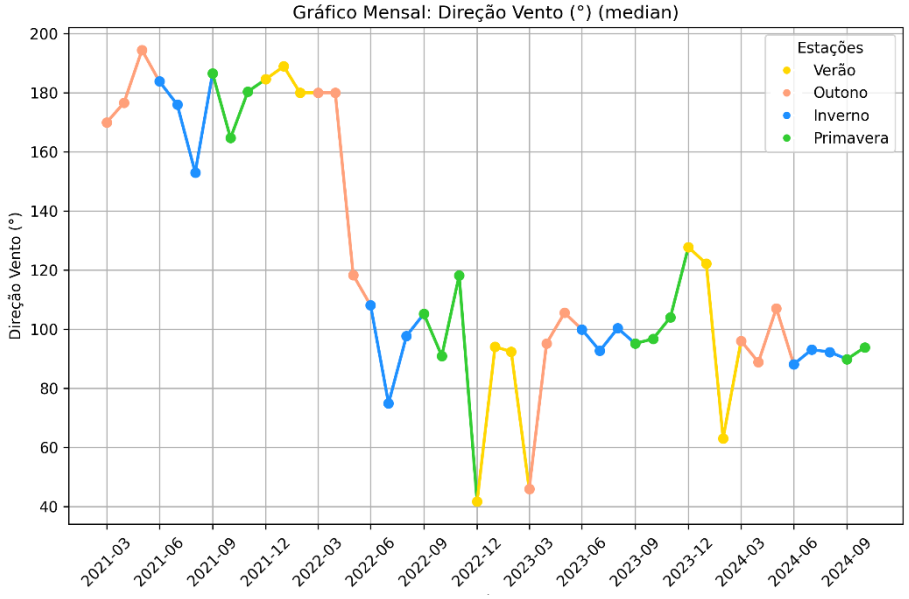
Figura 34 – Gráficos de direção do vento



Fonte: Elaboração própria (2024)

Para a direção do vento, optou-se por construir o gráfico mensal da Figura 35 agregado pela mediana, em vez da média, a fim de melhor representar o valor central da distribuição. Isso se deve ao fato de que a direção do vento é uma variável circular, e a média aritmética pode ser distorcida por valores extremos próximos de 0° ou 360°. Além disso, a mediana é mais robusta em relação a outliers, oferecendo uma análise mais precisa, especialmente em distribuições multimodais como a observada para essa variável análise.

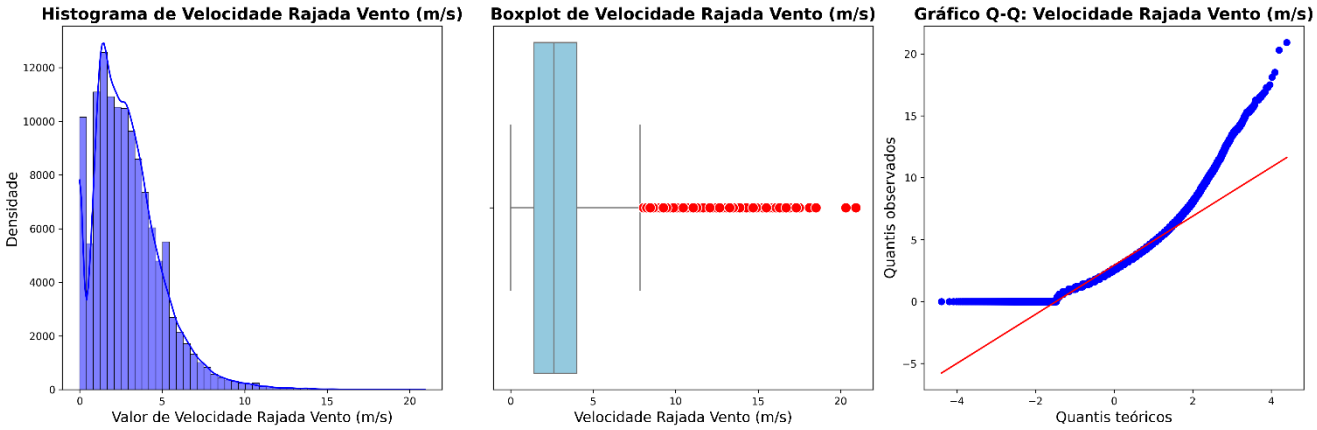
Figura 35 – Gráfico mensal de direção do vento



Fonte: Elaboração própria (2024)

Para a velocidade das rajadas de vento, observa-se, na Figura 36, que a maior parte dos dados está concentrada entre 0 e 5 m/s. A distribuição apresenta uma cauda longa à direita, indicando que, embora as rajadas de vento mais altas sejam menos frequentes, elas ainda ocorrem com menos frequência. Além disso, a distribuição não segue um padrão normal.

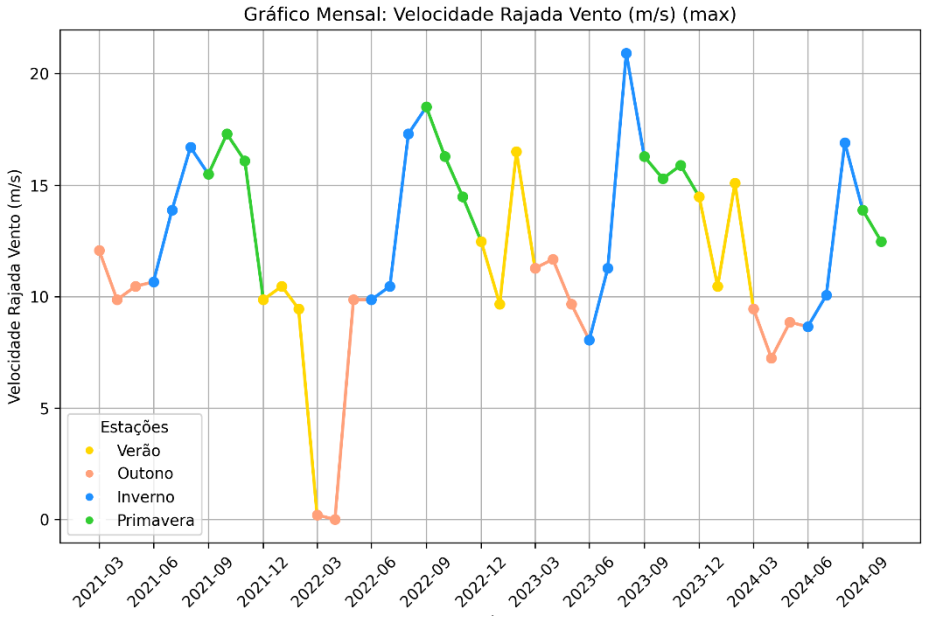
Figura 36 – Gráficos de velocidade da rajada do vento



Fonte: Elaboração própria (2024)

Para o gráfico mensal da Figura 37, escolheu-se agregar os dados das rajadas de vento pelo valor máximo para capturar os picos de velocidade, que são cruciais na análise de eventos extremos, como tempestades. O valor máximo destaca os períodos de maior intensidade, importantes para entender os impactos em condições operacionais e possíveis danos, além de facilitar a identificação de padrões de comportamento das rajadas ao longo do tempo. Assim, percebe-se que as maiores rajadas de vento costumam ocorrer durante o inverno e primavera enquanto as menores em outono.

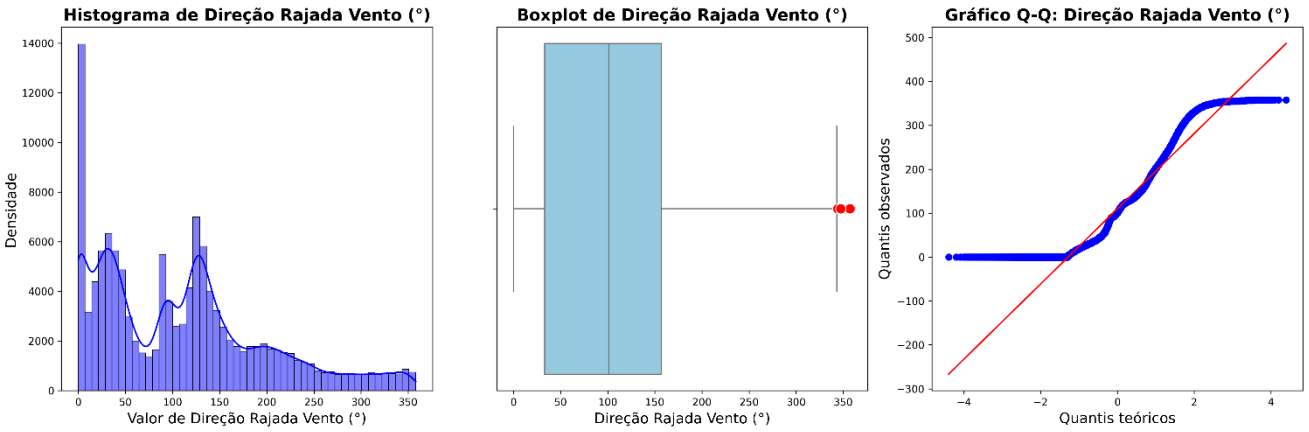
Figura 37 – Gráfico mensal de velocidade da rajada de vento



Fonte: Elaboração própria (2024)

Para a direção da rajada do vento, nota-se na Figura 38, uma distribuição similar à distribuição de ventos, sendo multimodal, além de não seguir o padrão normal.

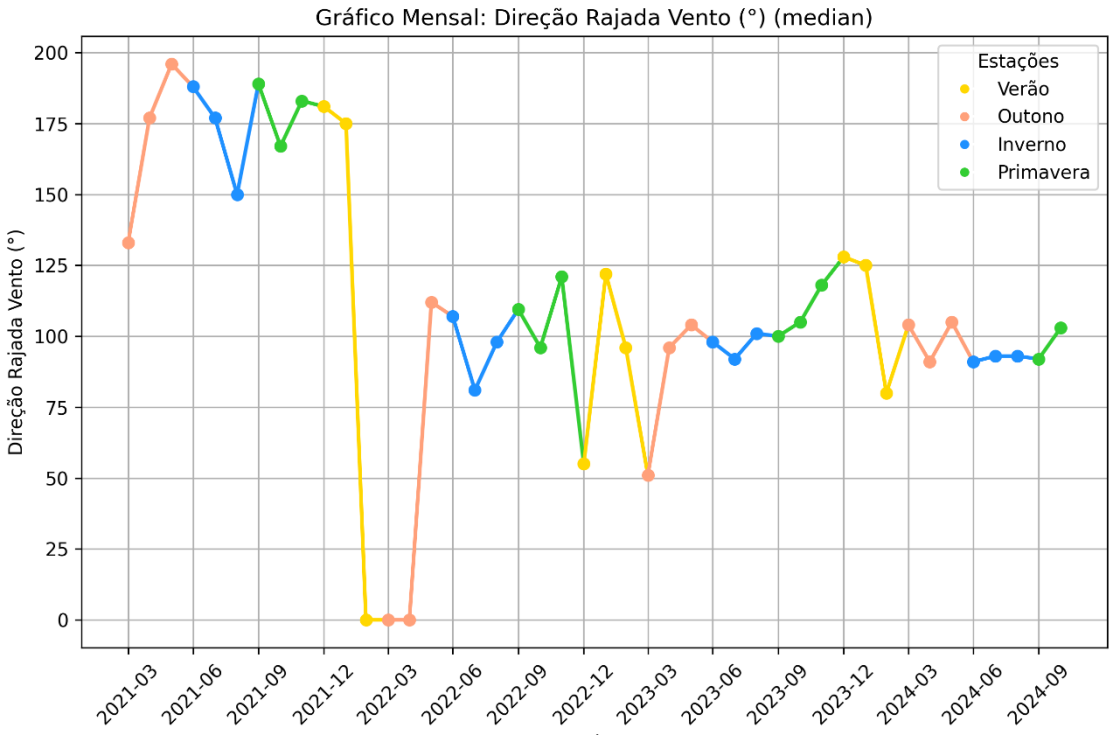
Figura 38 – Gráficos de direção da rajada do vento



Fonte: Elaboração própria (2024)

Para a direção da rajada do vento, optou-se por construir o gráfico mensal da Figura 39 agregado pela mediana, em vez da média, pelos mesmo motivos explicitados para a direção do vento.

Figura 39 – Gráfico mensal de direção da rajada do vento



Fonte: Elaboração própria (2024)

Por fim, a EDA individualizada para cada coluna complementou a etapa anterior ao oferecer uma análise mais específica para cada variável do estudo. Através de gráficos como histogramas, *boxplots* e gráficos Q-Q, foi possível identificar padrões e assimetrias nas distribuições, o que contribui para uma compreensão mais detalhada do comportamento dos dados. A análise de sazonalidade, por meio de gráficos temporais, também permitiu observar variações ao longo do tempo, destacando o impacto de fatores sazonais. Assim, essa etapa foi relevante para orientar o tratamento e a transformação dos dados nas etapas subsequentes do KDD, além de identificar possíveis outliers nas variáveis de vento, o que será crucial para o pré-processamento.

4.1.4 Definição de objetivo do KDD

O processo inicial do KDD abordado neste trabalho tem como objetivo viabilizar a criação de um modelo de ML capaz de prever a produtividade das safras com base nas variáveis analisadas e seus padrões. Essa solução busca fornecer aos agricultores uma ferramenta poderosa que não apenas antecipe a produtividade das culturas, mas também ofereça *insights* valiosos para apoiar decisões mais precisas. Ao otimizar o uso de recursos e permitir um planejamento agrícola mais eficiente, espera-se que essa abordagem contribua para a modernização da agricultura, um passo essencial para enfrentar o desafio da fome e atender à

crescente demanda por alimentos de forma sustentável.

Portanto, nessa primeira etapa do KDD obteve-se o entendimento inicial do *dataset* por meio de uma análise estatística descritiva e por meio de uma extensa EDA. Dessa forma, obteve-se uma compreensão do contexto dos dados, obtendo *insights* de cada uma das variáveis que possibilitam as demais etapas.

4.2 Seleção dos dados

Nessa etapa, objetiva-se selecionar a parte do conjunto de dados que, de fato, será utilizada até o fim do estudo.

4.2.1 Aplicação de filtros

Na etapa anterior, a análise dos dados em conjunto com a empresa parceira revelou que apenas a estação 17 era adequada para o estudo. As demais estações pertenciam a uma fazenda mais recente, caracterizada por um solo arenoso, um período de operação mais curto e dados menos consolidados, fatores que dificultariam a comparação de produtividade com base nas variáveis analisadas, especialmente pela ausência de registros de produtividade para determinadas safras. Por isso, optou-se por filtrar os dados, restringindo o estudo à estação 17, o que resultou em uma amostra de 127.872 linhas de dados coletados entre 2021 e 2024. Essa decisão garantiu uma base de dados mais confiável, essencial para associar os resultados às produtividades das safras, aspecto central para as análises desenvolvidas neste trabalho.

4.2.2 Seleção de variáveis relevantes para o estudo

A partir da tabela 4, percebe-se que a estação selecionada não mede a umidade do solo B, o que não prejudica a análise já que se tem ainda a umidade A e a única diferença entre elas é a altura do sensor. Portanto, ela será desconsiderada da análise. Além disso, percebe-se que os valores nulos para essa estação são reduzidos, em comparação com o *dataset* incluindo as demais estações, conforme tabela 1 na seção 4.1.1. Por fim, decidiu-se remover a coluna estimated_data e sensor_vars_failing já que são colunas utilizadas pela própria operação e para a amostra escolhida são majoritariamente compostas de valores nulos. Removeu-se também a coluna is_waiting_for_data que indica se os dados já foram recebidos ou se ainda serão. Essa informação é redundante, já que para os dados não recebidos as demais colunas estarão nulas.

Tabela 4 – Valores nulos para as colunas renomeadas

Coluna original	Coluna renomeada	Valores nulos
id		0%
station_id		0%
received_at		0%
press	Pressão Atmosférica (kPa)	2%
mCnpD	Umidade Copa Média (%)	1%
mCnpN	Umidade Copa Mínima (%)	1%
mCnpX	Umidade Copa Máxima (%)	1%
tCnpD	Temperatura Copa Média (°C)	1%
tCnpN	Temperatura Copa Mínima (°C)	1%
tCnpX	Temperatura Copa Máxima (°C)	1%
mTopD	Umidade Média (%)	1%
mTopN	Umidade Mínima (%)	1%
mTopX	Umidade Máxima (%)	1%
tTopD	Temperatura Média (°C)	1%
tTopN	Temperatura Mínima (°C)	1%
tTopX	Temperatura Máxima (°C)	1%
mSoilA	Umidade Solo 1 (%)	1%
mSoilB	Umidade Solo 2 (%)	100%
tSoil	Temperatura Solo (°C)	1%
wDirL	Direção Vento (°)	2%
wDirX	Direção Rajada Vento (°)	1%
wSpdD	Velocidade Vento (m/s)	1%
wSpdX	Velocidade Rajada Vento (m/s)	1%
rainFall	Chuva (mm^3)	1%
solarRad	Radiação Solar (W/m^2)	1%
is_waiting_for_data		0%
estimated_data		100%
sensor_vars_failing		100%

Fonte: Elaboração própria (2024)

Além disso, nota-se que algumas colunas possuem mesma natureza com pequenas diferenças de medição. São elas: Umidade Copa Mínima (%), Umidade Copa Média (%), Umidade Copa Máxima (%), Temperatura Copa Mínima (°C), Temperatura Copa Média (°C), Temperatura Copa Máxima (°C), Umidade Mínima (%), Umidade Média (%), Umidade Máxima (%), Temperatura Copa Mínima (°C), Temperatura Copa Média (°C), Temperatura Copa Máxima (°C). Para simplificar a análise e reduzir a dimensionalidade dos dados, busca-se identificar se todas essas variáveis são realmente relevantes ou se elas são significativamente parecidas. Para isso, foi feita uma análise de correlação de *Pearson*, representada por mapas de calor conforme as figuras 40 e 41. Percebe-se que as variáveis possuem correlações elevadas, já que só mudam a posição do sensor que as coleta. Portanto, concluiu-se que as variáveis

analisadas se tratam de informações similares com altas correlações entre elas. Diante disso, optou-se por usar apenas os valores médios para o trabalho (Temperatura Média (°C) e Umidade Média (%)), removendo as demais colunas de umidade e temperatura que foram consideradas irrelevantes. Isso permite reduzir significativamente o número de variáveis sem comprometer a precisão do modelo.

Figura 40 – Mapa de calor para variáveis de temperatura

Fonte: Elaboração própria (2024)

Figura 41 – Mapa de calor para variáveis de umidade

Fonte: Elaboração própria (2024)

Portanto, a etapa de seleção de dados do processo de KDD foi concluída com a definição do subconjunto de dados e variáveis que serão utilizados no restante do estudo. Inicialmente, selecionou-se os dados proveniente da estação 17 apenas. Em seguida, foram aplicados filtros para remover variáveis redundantes ou irrelevantes, como colunas com alta proporção de valores nulos ou informações redundantes sobre o estado dos dados. Adicionalmente, uma análise de correlação de Pearson permitiu identificar variáveis altamente correlacionadas relacionadas à medição de umidade e temperatura. Para simplificar a análise e reduzir a dimensionalidade dos dados, optou-se por manter apenas os valores médios dessas variáveis. Essas decisões asseguram um conjunto de dados mais conciso, consistente e apropriado para as próximas etapas do trabalho, mantendo a robustez necessária para atender aos objetivos propostos e reduzindo a complexidade do *dataset*.

4.3 Pré-Processamento

Com a amostra dos dados a ser utilizada selecionada, procede-se para a etapa de pré-processamento. Vale ressaltar que, como já explicitado anteriormente, o *dataset* já passou por um tratamento pela empresa parceira, então essa etapa seria mais extensa para os dados brutos da estação. De qualquer forma, a etapa divide-se em duas: *Data Cleaning* e identificação de

outliers.

4.3.1 Data Cleaning

Primeiramente, foram encontradas 29 linhas duplicadas que foram removidas e modificou-se o nome das colunas para facilitar a interpretação e manuseio dos dados, conforme a coluna Coluna renomeada da tabela 4 na seção 4.2.2. Em seguida, analisou-se a quantidade de valores nulos. Não foram encontradas linhas inteiramente nulas; portanto, contou-se o número de valores nulos presentes nas variáveis numéricas relevantes para a análise (cada célula do *DataFrame* no Pandas). No total, identificaram-se 24.426 campos nulos, o que corresponde a 1,27% do número total de campos. Posteriormente na etapa de transformação, objetiva-se analisar o impacto de diferentes formas de lidar com esses valores nulos.

4.3.2 *Outliers*

A forma mais simples de encontrar *outliers* para os dados do trabalho foi comparando os valores máximos e mínimos de cada uma das variáveis com a amplitude de seus sensores (tabela 5). Dessa forma, se existirem dados fora dessa amplitude conclui-se que se trata de um *outlier*. Entretanto, conforme observado na tabela 5 e na figura 42 percebe-se que nenhuma das variáveis está fora da amplitude dos sensores, o que indica que as medições podem ser reais mesmo que sejam extremas em relação à média dos dados.

Tabela 5 – Comparação dos valores minímos e máximos com a amplitude dos sensores

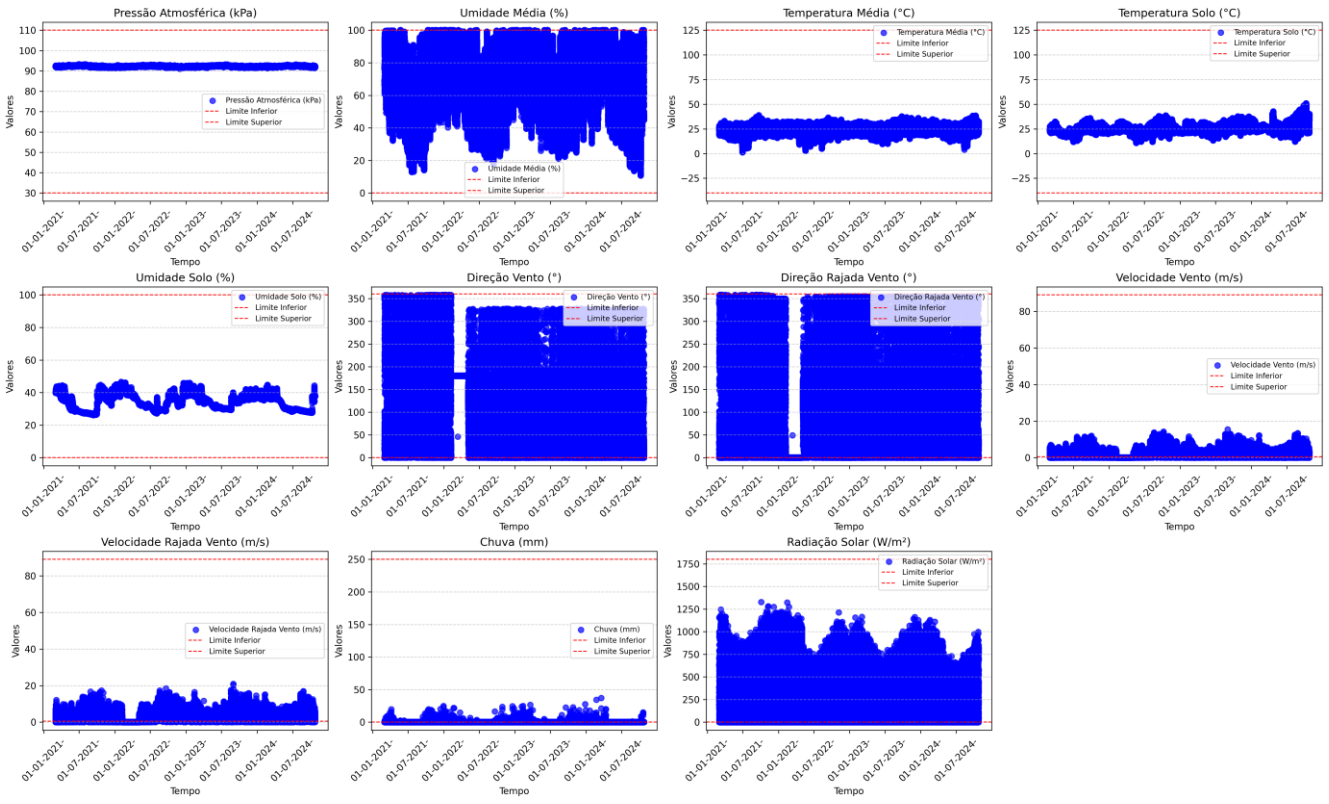
Coluna	Descrição	Min	Max	Unidade de Medida	Amplitude dos sensores
Pressão Atmosférica (kPa)	Pressão atmosférica	91.13	93.24	kPa	30 a 110 kPa
Umidade Média (%)	Umidade Relativa media	10.78	99.99	%	0 a 100%
Temperatura Média (°C)	Temperatura média	1.34	39.02	°C	-40 a 125°C
Umidade Solo (%)	Umidade Relativa do solo com o sensor à uma distância de 10 a 20 cm da superfície do solo.	26.22	46.59	%	0 a 100%
Temperatura Solo (°C)	Temperatura do solo	10.93	50.96	°C	-40 a 125°C
Direção Vento (°)	Direção do vento	0	358	°	360°
Direção Rajada Vento (°)	Direção da rajada do vento (medições máximas para o intervalo de quinze minutos)	0	358	°	360°
Velocidade Vento (m/s)	Velocidade do vento	0	15.47	m/s graus	0.5 a 89 m/s

Velocidade Rajada Vento (m/s)	Velocidade da rajada do vento (medições máximas para o intervalo de quinze minutos)	0	20.92	m/s graus	0.5 a 89 m/s
Chuva (mm)	Chuva	0	37.2	mm	250 mm/h
Radiação Solar (W/m²)	Radiação Solar	0	1326.69	W/m²	0 a 1800 W/m²

Fonte: Elaboração própria

Além da visualização das tabelas, construiu-se também os gráficos da Figura 42. Neles o eixo X representa o período das medições, o eixo Y os valores das medições e as linhas vermelhas os limites dos sensores. Percebe-se visualmente que todos os valores estão dentro dos limites. Entretanto, todas as quatro variáveis de vento (direção, sentido da rajada e do vento) estão com uma grande concentração de valores em sua maioria zerados para o período entre março e abril de 2022, conforme verificado anteriormente na seção 4.1.3.

Figura 42 – Scatter Plots com amplitude dos sensores para cada variável

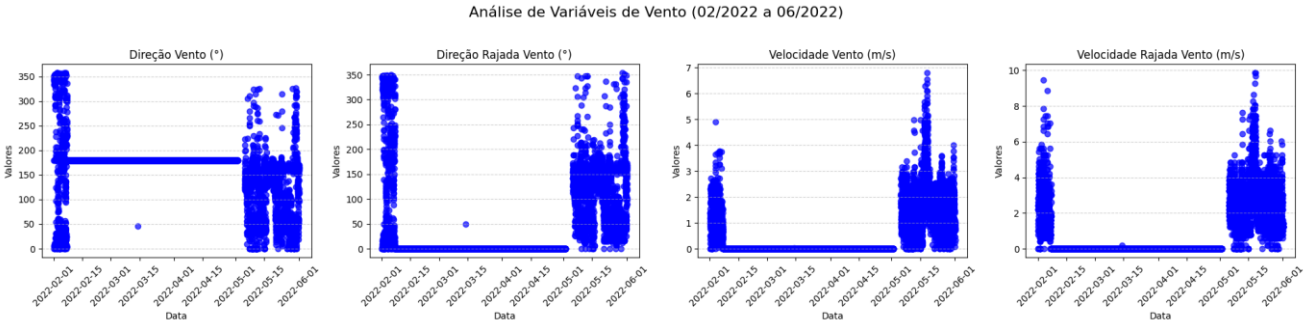


Fonte: Elaboração própria (2024)

A fim de investigar esses valores zerados com uma melhor visualização construiu-se gráficos similares, mas para os períodos com suspeita de falha nos sensores, conforme a Figura 43. De fato, nota-se que os ventos medidos para esse período fogem dos padrões apresentados pelo *dataset*, sendo para direção do vento valores fixos entre 150 e 200 e os demais valores zerados. Assim, mesmo esses valores mencionados estando dentro da amplitude dos sensores, é

altamente improvável que todas as medições do período analisado sejam distribuídas dessa forma. Por isso, considerou-se essa janela como *outlier*.

Figura 43 – Scatter Plots para as variáveis de vento com outliers



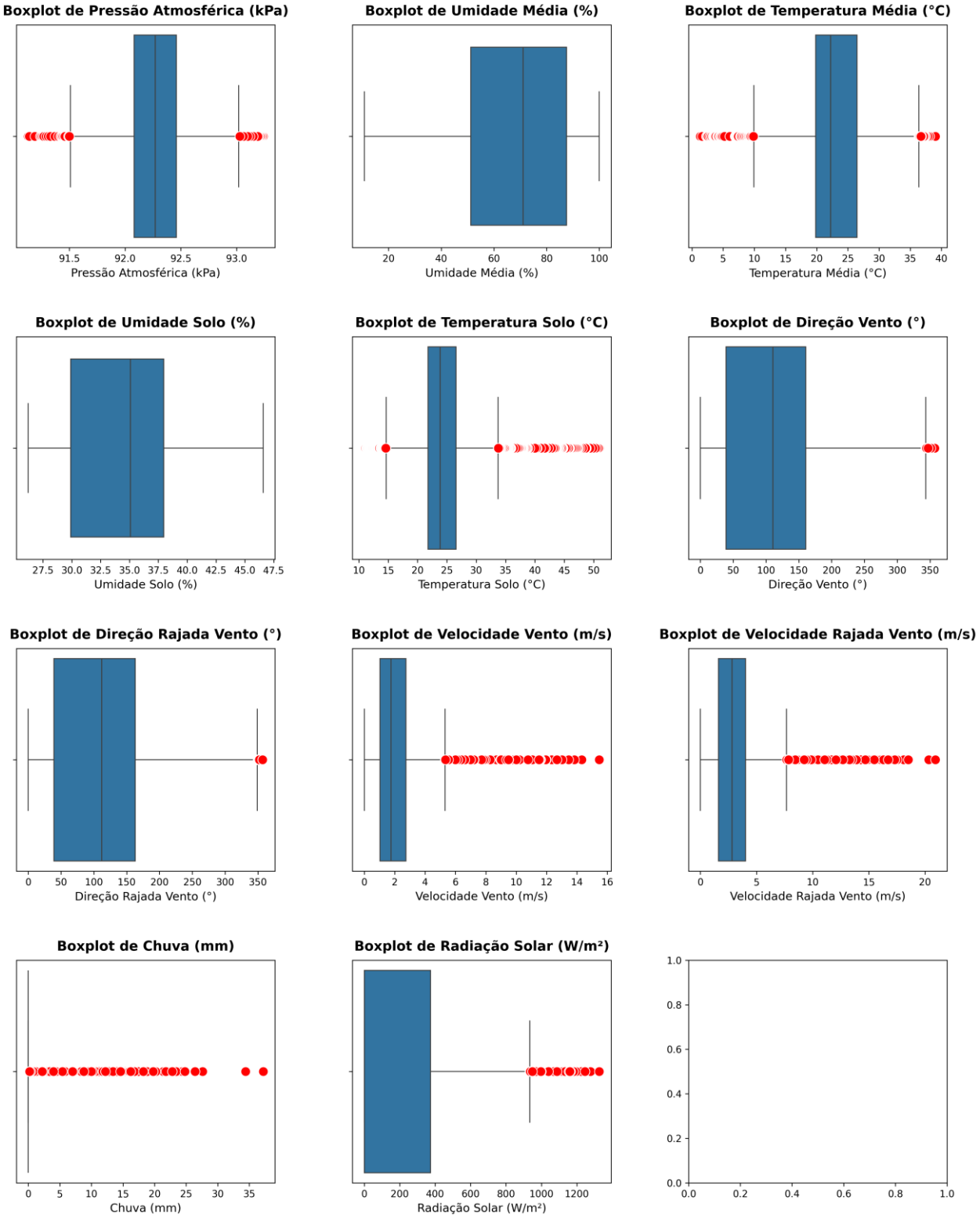
Fonte: Elaboração própria (2024)

Além disso, buscando um entendimento de possíveis *outliers* mesmo dentro da amplitude dos sensores os gráficos *boxplot* apresentados na seção 4.1.2 na primeira etapa do KDD dão um breve indicativo. Assim, retoma-se essa análise com foco em identificar outliers.

Para cada coluna, busca-se identificar possíveis outliers com base na análise gráfica. Conforme a Figura 44, as variáveis de umidade do ar e do solo não apresentam valores considerados *outliers* estatísticos. Por outro lado, para as variáveis de direção do vento, observa-se que a metodologia do *boxplot* identificou valores acima de 350° como *outliers*. No *boxplot*, valores são considerados outliers se estiverem muito abaixo ou muito acima dos valores típicos de um conjunto de dados, com base em uma faixa calculada a partir dos quartis. Entretanto, para variáveis de direção do vento, que possuem valores restritos ao intervalo de 0° a 360° e cuja interpretação é cíclica, em que 0° e 360° representam a mesma direção, essa abordagem se torna incoerente. Valores próximos de 360° são parte natural da escala e não indicam anomalias. Por essa razão, esses valores foram tratados como válidos e não foram considerados outliers.

Para as demais variáveis, dado que os valores extremos estão dentro da amplitude dos sensores eles podem ser considerados *outliers* estatísticos, porém entende-se que são medições que podem ocorrer durante eventos climáticos mais extremos durante o ano e devem ser considerados para a análise.

Figura 44 – Boxplots de todas as variáveis do estudo



Fonte: Elaboração própria (2024)

Por fim, a etapa de pré-processamento de dados do processo de KDD foi finalizada com a aplicação de técnicas de limpeza e identificação de *outliers*, garantindo a qualidade e consistência do *dataset* para as próximas etapas do estudo. Além disso, os *outliers* das variáveis de vento serão tratados na próxima etapa, a de transformação.

4.4 Transformação

Nesta etapa objetiva-se transformar-se os dados para adaptá-los à necessidade de uso para os próximos passos. Dessa forma, dividiu-se em duas transformações principais para o contexto desse trabalho: adição de colunas de interesse e *Data Completion*.

4.4.1 Adição de colunas de interesse

Primeiramente, adicionou-se colunas para identificar a qual safra cada linha do *dataset* pertence, e com isso, associar as variáveis à produtividade daquele período. Para isso, foram utilizados os períodos aproximados das safras e a sua produtividade da Tabela 3. A partir da coluna received_at (coluna com as datas de medição) classifica-se qual safra cada medição das linhas faz parte, conforme a tabela 3, desconsiderando as entressafras e as safras de 2025 a qual ainda não existe dado de produtividade (restaram 103.462 linhas). Para cada safra tem-se a produtividade do período em sacas por hectare. Dessa forma, separamos as variáveis por safra com sua respectiva produtividade.

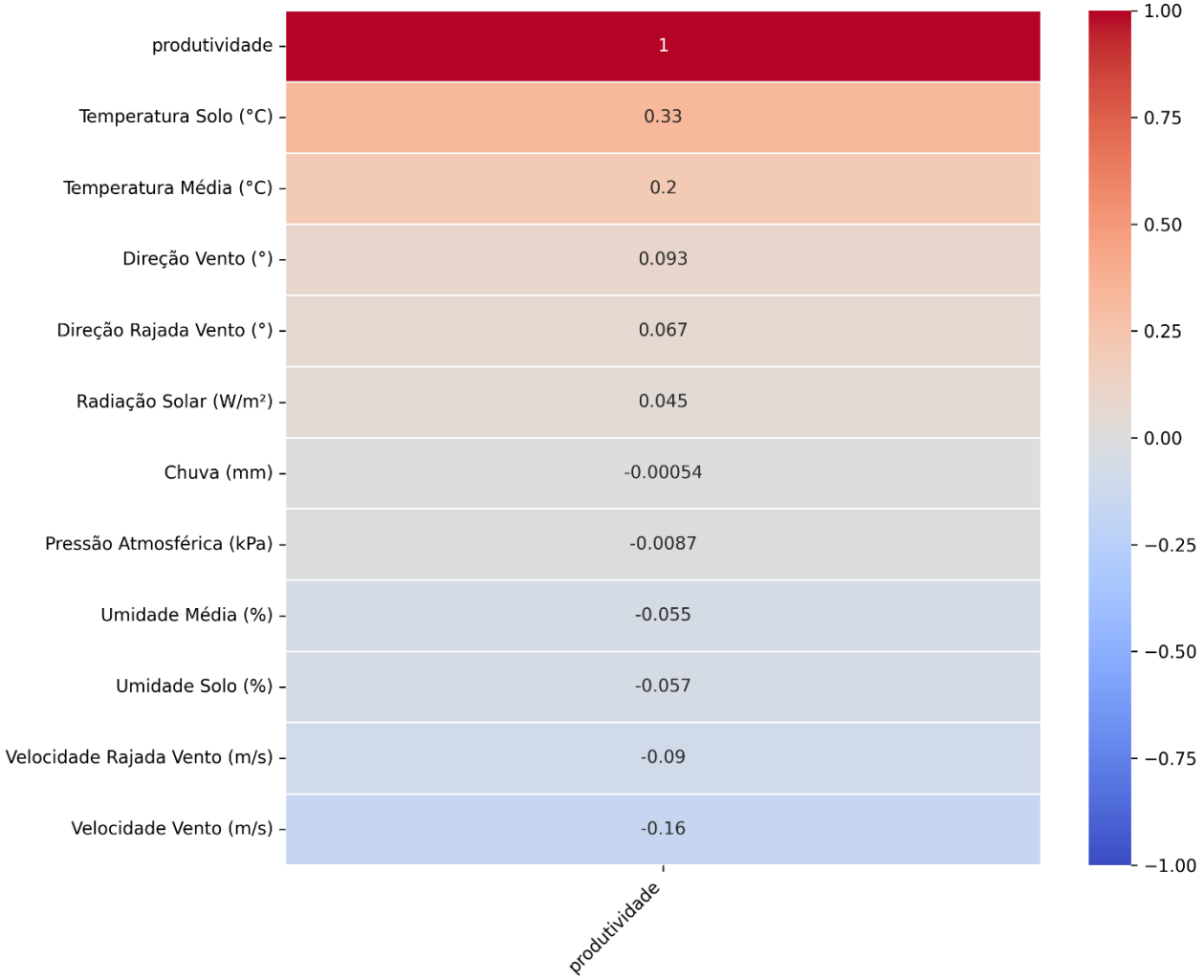
Tabela 3 – Período das safras e suas produtividades

Título	Início	Fim	Produtividade (sacas/ha)
Safra Soja 20/21	10/10/2020	14/02/2021	
Safra Soja 21/22	10/10/2021	14/02/2022	62
Safra Soja 22/23	10/10/2022	14/02/2023	55
Safra Soja 23/24	10/10/2023	14/02/2024	65
Safra Soja 24/25	10/10/2024	14/02/2025	
Segunda Safra Milho 21	15/02/2021	15/08/2021	96
Segunda Safra Milho 22	15/02/2022	15/08/2022	82
Segunda Safra Milho 23	15/02/2023	15/08/2023	94
Segunda Safra Milho 24	15/02/2024	15/08/2024	157
Segunda Safra Milho 25	15/02/2025	15/08/2025	
Entressafra 21	16/08/2021	9/10/2021	0
Entressafra 22	16/08/2022	9/10/2022	0
Entressafra 23	16/08/2023	9/10/2023	0
Entressafra 24	16/08/2024	9/10/2024	0
Entressafra 25	16/08/2025	9/10/2025	0

Fonte: Empresa Parceira

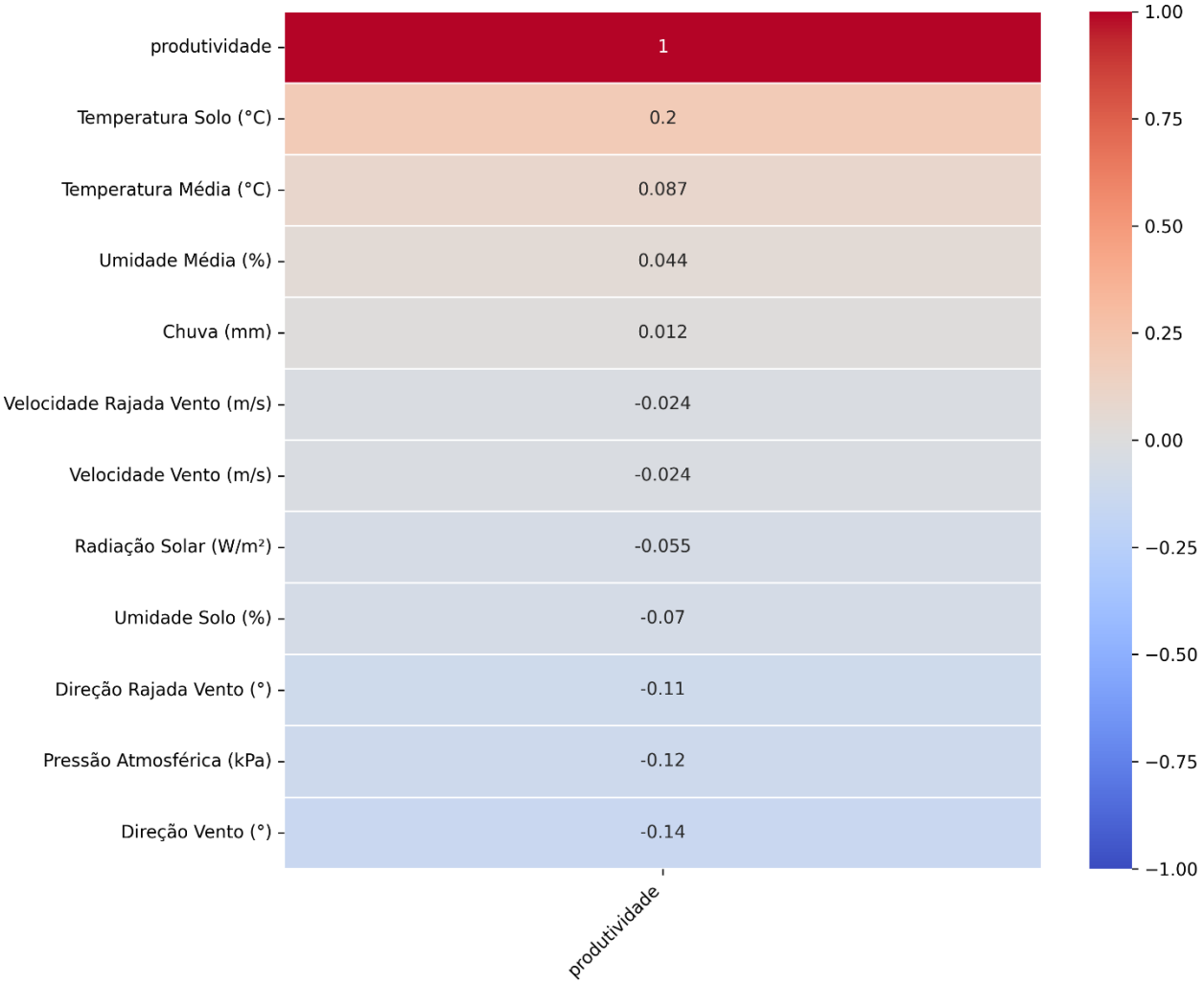
Em seguida, o *dataset* foi segmentado em safras de milho e soja, com o objetivo de compreender o impacto das variáveis em cada cultura. Para isso, foram construídos mapas de calor que comparam as variáveis com a produtividade: o mapa referente ao *dataset* de soja é apresentado na Figura 45, enquanto o mapa referente ao *dataset* de milho está na Figura 46. Para essas figuras, foi utilizado o *dataset* contendo medições realizadas a cada 15 minutos, com a produtividade atribuída à linha correspondente. No entanto, observou-se que os valores obtidos não refletem a realidade, uma vez que nenhuma das variáveis analisadas apresentou correlação significativa com a produtividade. Esse resultado deve-se ao fato de que as medições das variáveis e os valores de produtividade foram considerados em períodos distintos, comprometendo a análise.

Figura 45 – Correlações das variáveis com a produtividade de soja



Fonte: Elaboração própria (2024)

Figura 46 – Correlações das variáveis com a produtividade de milho

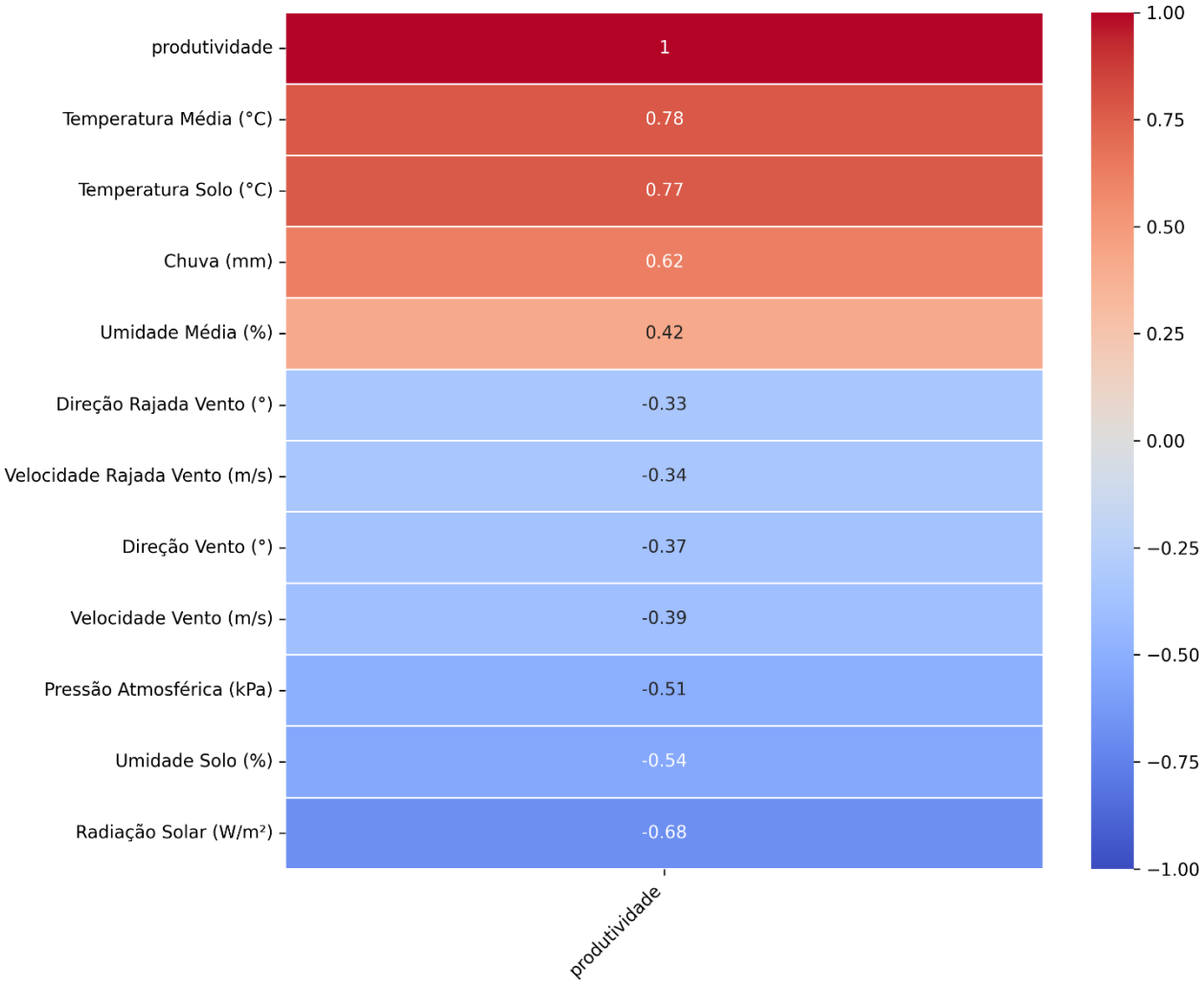


Fonte: Elaboração própria (2024)

Para resolver o problema identificado anteriormente, as variáveis foram agrupadas pela média, de modo que elas e a produtividade estivessem associadas ao mesmo período de tempo. Em seguida, foi gerado novos mapas de calor para cada cultura. Dessa vez, os resultados mostraram-se mais esclarecedores. No caso da soja, as variáveis com maiores correlações positivas foram temperatura média, temperatura do solo, chuva e umidade média, enquanto umidade do solo e radiação solar apresentaram correlações negativas (Figura 47). Por outro lado, para o milho, as variáveis com maiores correlações positivas foram radiação solar, temperatura média e temperatura do solo, enquanto a correlação da chuva foi próxima de zero e a umidade apresentou uma relação consideravelmente negativa (Figura 48). Embora os resultados sejam mais elucidativos, não é possível tirar conclusões definitivas apenas com essa análise, pois, após

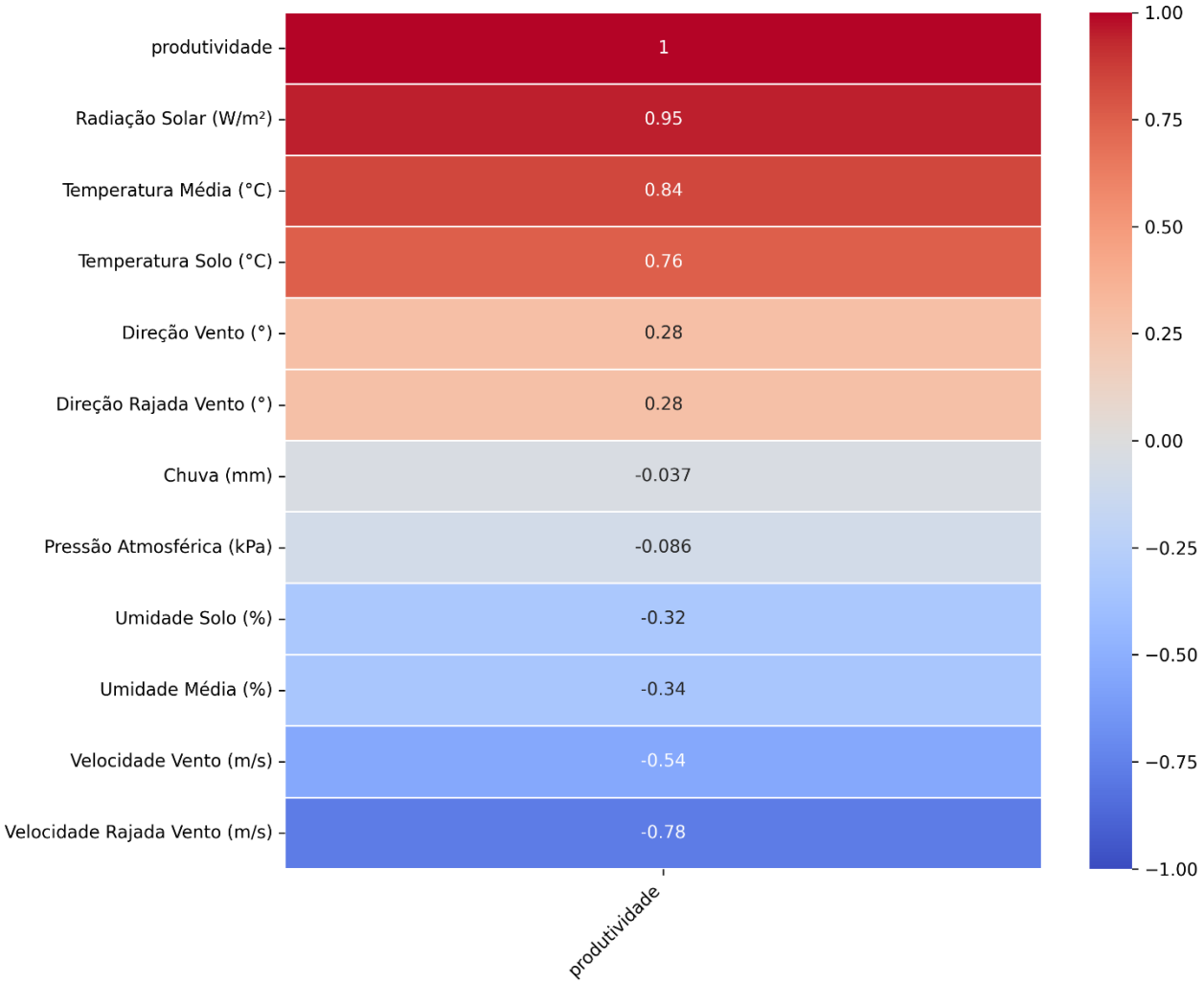
o agrupamento, obteve-se uma amostra reduzida, com apenas 3 linhas para soja e 4 para milho. Para conclusões mais robustas, seria necessário um conjunto de dados maior, abrangendo um período com mais safras.

Figura 47 – Correlações das variáveis com a produtividade de soja (base agrupada por safra)



Fonte: Elaboração própria (2024)

Figura 48 – Correlações das variáveis com a produtividade de milho (base agrupada por safra)



Fonte: Elaboração própria (2024)

4.1.2 Data Completion para valores nulos

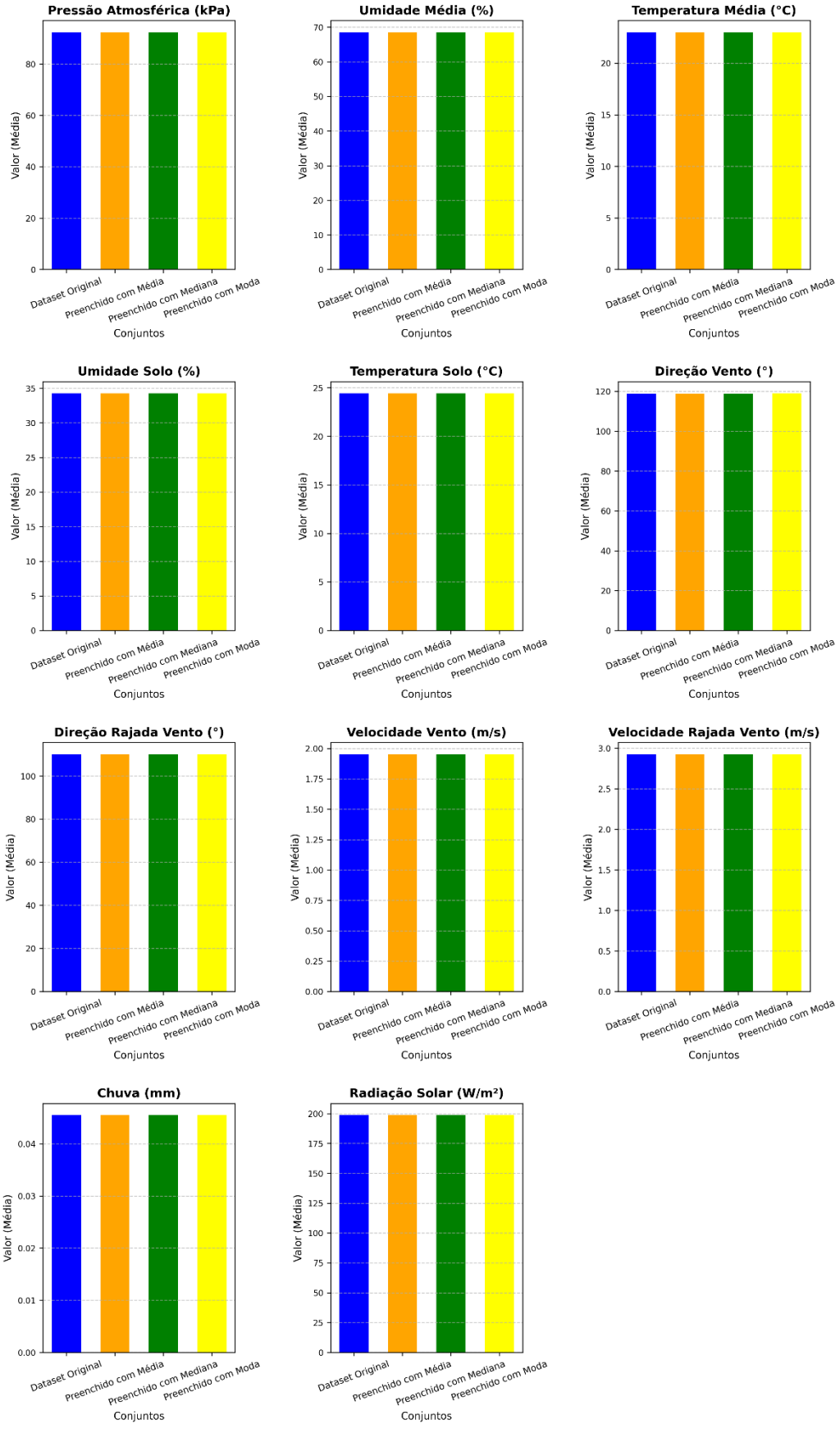
Posteriormente, separou-se a amostra em quatro: dados nulos removidos, dados nulos preenchidos com a média, dados nulos preenchidos com a mediana e dados nulos preenchidos com a moda. Para todas essas amostras, serão consideradas apenas as colunas de valores numéricos onde podem ser calculadas médias, medianas e modas. As demais serão removidas, para as análises seguintes. Após esse preenchimento, as três amostras não possuem nenhum valor nulo.

A análise das imagens 49, 50, 51 e 52 revelam que, devido à baixa quantidade de valores nulos no conjunto de dados (1,27% do total de campos) e ao preenchimento com valores representativos, não houve alterações significativas nas principais medidas estatísticas, como média, desvio padrão, mediana e moda. Isso sugere que os métodos de preenchimento utilizados

preservam a distribuição original dos dados e podem ser utilizados para tratar os dados faltantes.

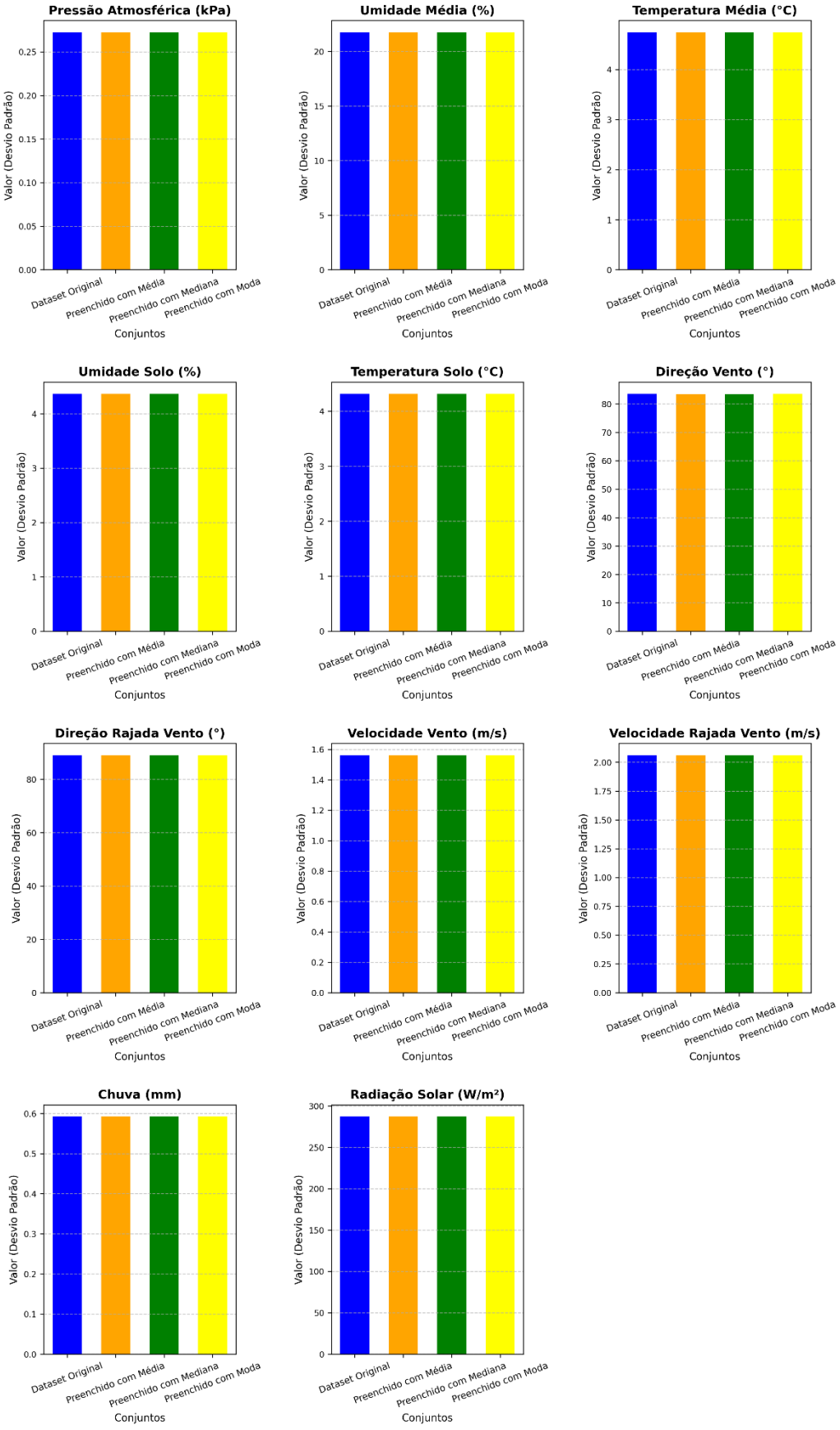
No entanto, para as variáveis relacionadas ao vento, chuva e radiação solar, recomenda-se evitar o uso de preenchimento com mediana ou moda. Isso ocorre porque essas variáveis apresentam uma alta concentração de valores zerados, tornando a mediana e a moda iguais a 0, o que poderia introduzir mais inconsistências ao conjunto de dados ao invés de resolver os problemas associados aos valores nulos.

Figura 49 – Comparação de métricas: Média



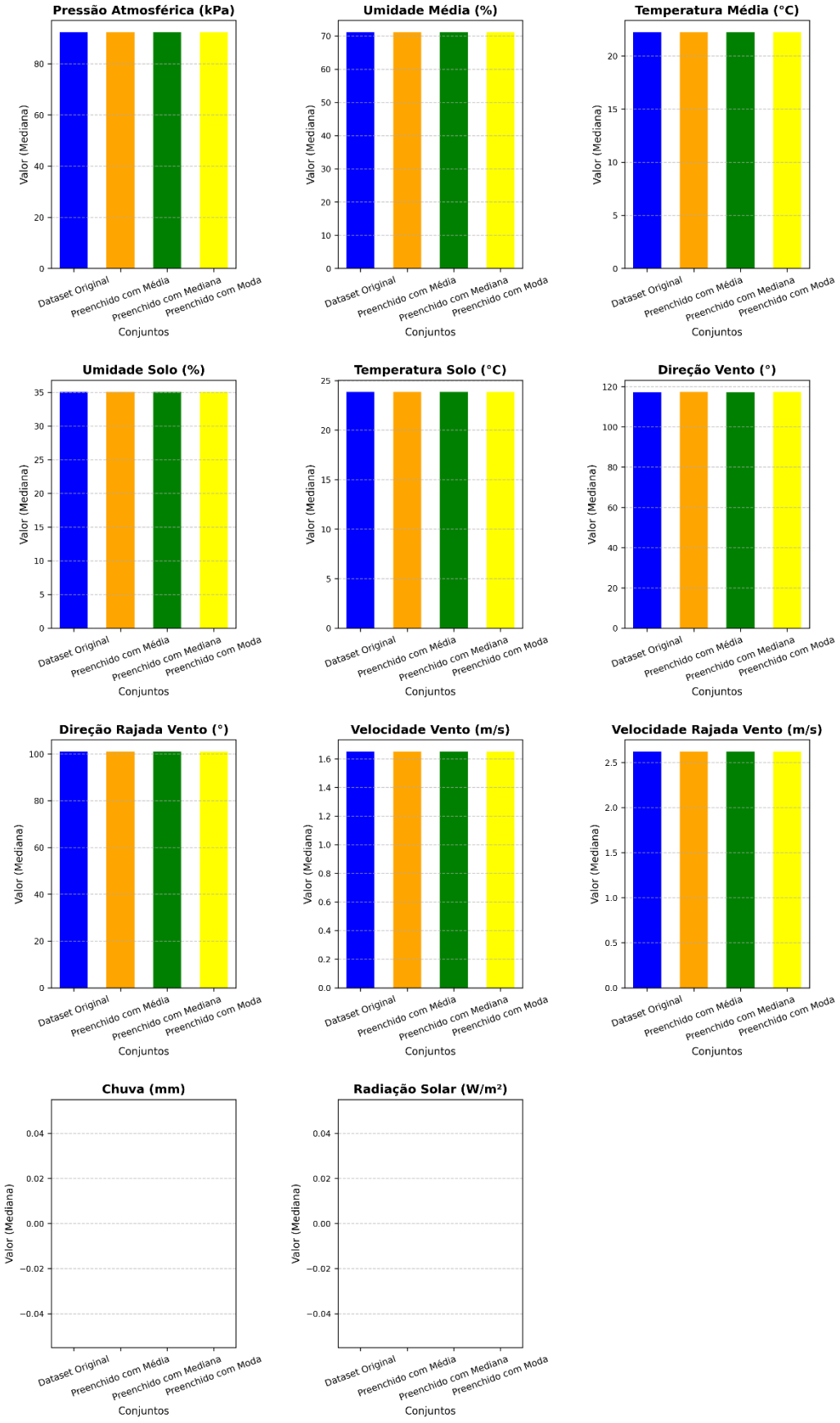
Fonte: Elaboração própria (2024)

Figura 50 – Comparação de métricas: Desvio Padrão



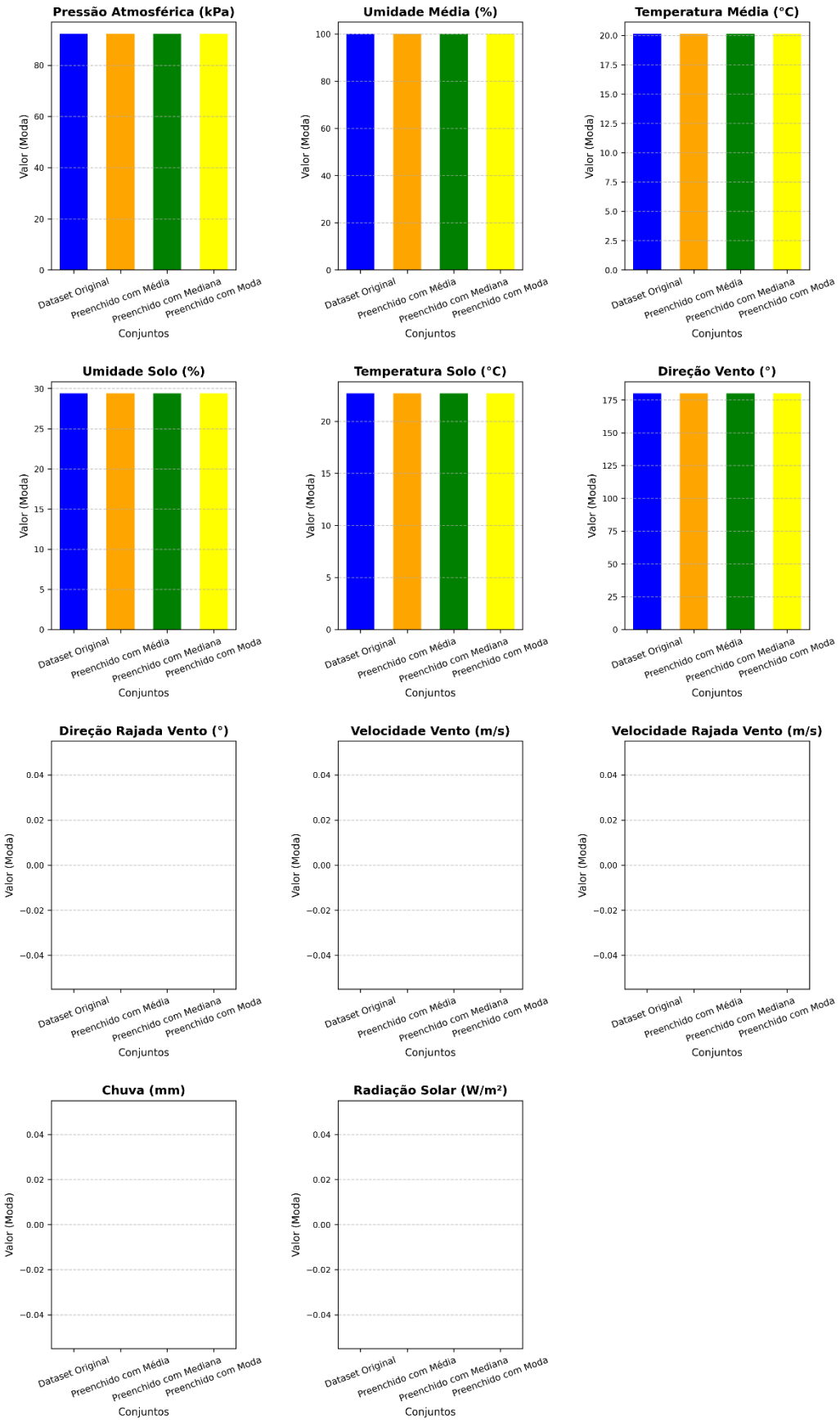
Fonte: Elaboração própria (2024)

Figura 51 – Comparação de métricas: Média



Fonte: Elaboração própria (2024)

Figura 52 – Comparação de métricas: Moda



Fonte: Elaboração própria (2024)

4.1.3 *Data Completion* para *outliers*

Para lidar com os dados faltantes de direção do vento, cuja natureza apresenta maior aleatoriedade, optou-se por preenchê-los com valores gerados aleatoriamente, seguindo a distribuição histórica do conjunto de dados, como apresentado na Figura 53. Já no caso da velocidade, priorizou-se a preservação das características estatísticas, como a média e a dispersão. Assim, os valores ausentes foram preenchidos com números aleatórios calculados a partir da média e do desvio padrão do *dataset*, conforme demonstrado na Figura 54. Os resultados dos métodos de preenchimento estão apresentados nas Figuras 55 e 56, que ilustram as distribuições das variáveis após o preenchimento, tanto no período correspondente aos *outliers* quanto no período completo do conjunto de dados. Observa-se que as distribuições obtidas são consistentes e compatíveis com o comportamento geral do *dataset*, de forma a minimizar o impacto negativo da presença dos *outliers*.

Figura 53 – Código para preenchimento dos *outliers* de direção de vento

```
# Definir o período com valores faltantes ou zerados
start_date = "2022-02-08"
end_date = "2022-05-05"

# Filtrar as linhas no período identificado
filtered_df = df_not_null[(df_not_null['received_at'] >= start_date) & (df_not_null['received_at'] <= end_date)]

# Filtrar apenas os casos onde os valores de "Direção Vento (°)" estão zerados ou inválidos
mask = (df_not_null['received_at'] >= start_date) & (df_not_null['received_at'] <= end_date)

# Gerar valores aleatórios para "Direção Vento (°)" com base na distribuição histórica
prob_dist = df_not_null['Direção Vento (°)'].value_counts(normalize=True)
direcoes_aleatorias = np.random.choice(prob_dist.index, size=mask.sum(), p=prob_dist.values)

# Atualizar os valores no dataframe original
df_not_null.loc[mask, 'Direção Vento (°)'] = direcoes_aleatorias


# Gerar valores aleatórios para "Direção Vento (°)" com base na distribuição histórica
prob_dist = df_not_null['Direção Rajada Vento (°)'].value_counts(normalize=True)
direcoes_aleatorias = np.random.choice(prob_dist.index, size=mask.sum(), p=prob_dist.values)

# Atualizar os valores no dataframe original
df_not_null.loc[mask, 'Direção Rajada Vento (°)'] = direcoes_aleatorias
✓ 0.2s
```

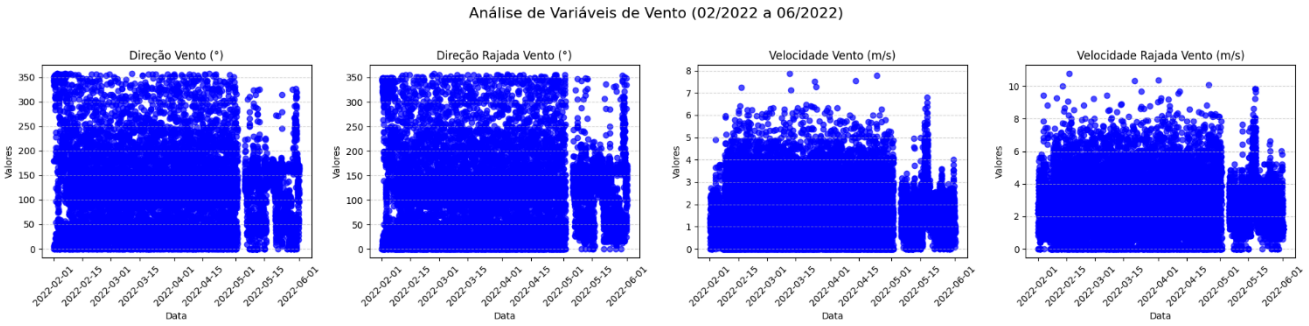
Fonte: Elaboração própria (2024)

Figura 54 – Código para preenchimento dos *outliers* de velocidade de vento

```python
# Função para imputar valores aleatórios ao redor da média do dataset inteiro com limite inferior de 0
def inputar_aleatorio_global(df, coluna, periodo_inicio, periodo_fim):
    # Filtrar o DataFrame para o período especificado
    periodo_df = df[(df['received_at'] >= periodo_inicio) & (df['received_at'] <= periodo_fim)]

    # Calcular a média e o desvio padrão do dataset inteiro
    media_global = df[coluna].mean()
    desvio_padrao_global = df[coluna].std()

    # Identificar índices dos valores ausentes (zeros) no período desejado
    indices_ausentes = periodo_df[periodo_df[coluna] == 0].index

    # Gerar valores aleatórios com base na média e desvio padrão globais
    valores_aleatorios = np.random.normal(loc=media_global, scale=desvio_padrao_global, size=len(indices_ausentes))

    # Limitar os valores gerados para que sejam maiores ou iguais a 0
    valores_aleatorios = np.clip(valores_aleatorios, a_min=0, a_max=None)

    # Atualizar o DataFrame com os valores gerados
    df.loc[indices_ausentes, coluna] = valores_aleatorios

    return df

# Exemplo de uso
df_not_null = inputar_aleatorio_global(df_not_null, 'Velocidade Vento (m/s)', "2022-02-08", "2022-05-05")
df_not_null = inputar_aleatorio_global(df_not_null, 'Velocidade Rajada Vento (m/s)', "2022-02-08", "2022-05-05")
```

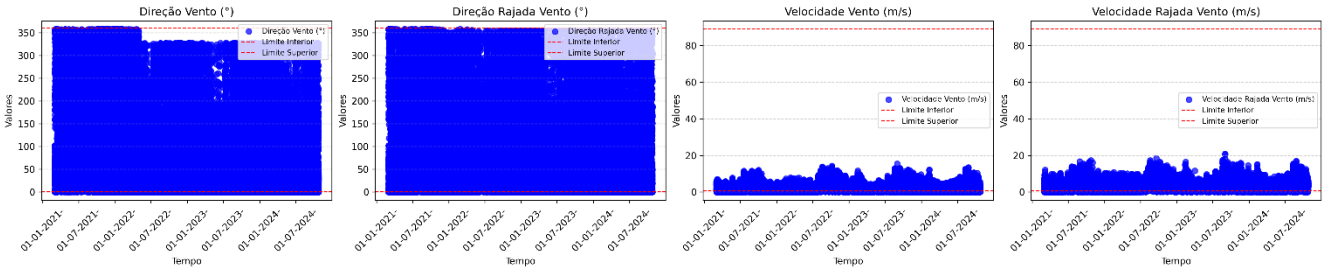
Fonte: Elaboração própria (2024)

Figura 55 – *Scatter Plots* para o período de outliers de vento após preenchimento



Fonte: Elaboração própria (2024)

Figura 56 – *Scatter Plots* para o todo o período após preenchimento de outliers de vento



Fonte: Elaboração própria (2024)

Assim, os *outliers* foram tratados e a base de dados ajustada será utilizada ao longo do trabalho. Essa estratégia visa aumentar a confiabilidade das análises, pois os *outliers* relacionados ao vento foram substituídos de forma criteriosa, preservando as demais variáveis correspondentes aos períodos em que as medições eram consistentes com a realidade. Dessa forma, ao final desta etapa, obteve-se um conhecimento mais aprofundado sobre o conjunto de dados em comparação ao início do estudo, resultando em um *dataset* devidamente tratado e pré-processado, pronto para as análises subsequentes e a aplicação de técnicas de ML.

4.5 Próximos passos

Com a conclusão dos quatro primeiros passos do processo de KDD, o presente trabalho consolidou um *dataset* explorado, limpo e tratado, oferecendo uma base consistente para as próximas etapas. A aplicação de uma metodologia estruturada permitiu compreender o contexto das variáveis, suas distribuições e comportamentos ao longo das estações do ano, além de garantir a qualidade dos dados com o tratamento de *outliers* e valores faltantes. Com essa preparação concluída, os próximos passos incluem as etapas do KDD de *data mining* e interpretação, que, nesse contexto, objetiva a criação de um *framework* de gestão de desempenho baseado em modelos de ML, com o objetivo de prever os impactos das variáveis analisadas na produtividade para basear a tomada de decisão do dono da operação e extrair conhecimento desse framework continuamente.

5 Conclusão

O presente Trabalho de Conclusão de Curso teve como objetivo estruturar uma abordagem sistemática de exploração e pré-processamento de dados para viabilizar a aplicação de algoritmos de aprendizagem de máquina em *dataset* de uma operação agrícola. Esse objetivo foi atingido por meio da elaboração de um passo a passo baseado no KDD, conforme observado na seção 4.

Inicialmente na etapa de entendimento dos dados, buscou-se compreender melhor a operação por meio de conversas com a empresa parceira. Durante essas interações, foi possível entender a natureza dos sensores e das variáveis envolvidas, bem como verificar que a base de dados fornecida já havia passado por alguns tratamentos prévios realizados pela empresa. Com esses dados em mãos, compararam-se as informações das duas fazendas disponíveis, analisando as diferenças entre as estações. Decidiu-se, então, que o foco do trabalho seria direcionado à estação 17 para o restante da análise.

Posteriormente, foi realizada uma EDA, na qual as correlações do mapa de calor entre as variáveis não apresentaram insights significativos. No entanto, os *scatter plots* revelaram algumas relações interessantes:

- Quanto maior a pressão atmosférica, menor a temperatura média.
- As chuvas se concentram em uma faixa mais central de pressão atmosférica.
- Quanto maior a temperatura média, menor a umidade relativa do ar.
- Chuvas mais intensas estão associadas a uma maior umidade relativa do ar.
- Quanto maior a temperatura média, maior a temperatura do solo.
- As chuvas tendem a ocorrer em valores centrais de temperatura média.
- Quanto maior a temperatura, maior a radiação solar.
- Chuvas mais intensas estão associadas a uma redução na radiação solar.

A análise avançou para uma EDA individualizada de cada variável, utilizando histogramas, *boxplots* e gráficos Q-Q para examinar a distribuição dos dados. Além disso, foi plotada a série histórica das variáveis, permitindo compreender seu comportamento ao longo do tempo e relacioná-las com as estações do ano. Essa etapa proporcionou maior entendimento do contexto das medições presentes no *dataset* e gerou *insights* relevantes, incluindo a identificação de *outliers* nas variáveis relacionadas ao vento e padrões específicos de comportamento em cada variável:

- Todas as distribuições não se assemelham a normalidade com exceção da pressão atmosférica, o que indica que a pressão atmosférica é a única das variáveis que se mantém mais estáveis devido a padrões climáticos mais estáveis.
- Notou-se forte dependência de quase todas as variáveis às estações do ano (com exceção das variáveis de direção do vento), cada uma com suas especificidades.

- Notou-se que para o período entre março e abril de 2022 as medições para as variáveis de vento foram inconsistentes com maioria igual a zero, o que sugere uma falha nas medições da estação.

Em seguida na etapa de seleção de dados, identificou-se que algumas variáveis adicionavam complexidade desnecessária à análise sem contribuir significativamente para os resultados. Essas variáveis foram removidas para reduzir a dimensionalidade do *dataset* e tornar as análises mais assertivas. Além disso na etapa de pré-processamento, foram eliminadas 29 linhas duplicadas, e constatou-se que 1,27% dos campos apresentavam valores nulos. Para tratar os *outliers* nas variáveis relacionadas ao vento, utilizou-se o preenchimento de dados (*data completion*) para corrigir os casos identificados durante a EDA. Para as demais variáveis, não foram encontrados outliers, pois as medições estavam coerentes e dentro da amplitude esperada dos sensores.

Além disso na etapa de transformação de dados, com base nos períodos de safra e na produtividade fornecida pela empresa parceira, foram adicionadas colunas que indicavam a qual safra cada medição pertencia e a produtividade associada. A partir disso, iniciou-se a relação entre as variáveis e a produtividade, o que revelou um grande desafio: enquanto as medições eram realizadas em intervalos de 15 minutos, a produtividade era registrada por safras de culturas diferentes. Para lidar com essa discrepância, os dados foram agregados pela média de cada safra, e um mapa de calor foi gerado para comparar as variáveis com a produtividade. Embora os resultados obtidos fossem coerentes, concluiu-se que seriam necessários mais dados para alcançar relevância estatística, já que estavam disponíveis apenas sete safras no total. Identificou-se, portanto, uma oportunidade de pesquisa futura utilizando dados meteorológicos do INMET ou outras fontes públicas com séries históricas maiores para complementar o presente trabalho.

Ainda na mesma etapa, compararam-se diferentes métodos de preenchimento de dados para os campos nulos, incluindo preenchimento por média, mediana e moda, além da remoção dos dados faltantes. O principal *insight* foi que, devido à baixa quantidade de dados nulos, qualquer um desses métodos poderia ser utilizado, com exceção do preenchimento por mediana ou moda em variáveis com muitos valores iguais a zero, pois isso poderia distorcer a análise.

Por fim, destaca-se que a metodologia pode ser replicada para as etapas iniciais de exploração e tratamento de dados em demais trabalhos no contexto de operações agrícolas, visando a preparação de *datasets* similares obtidos para a aplicação de modelos de ML. Além disso, o próximo passo dessa aplicação em específico é a continuação das etapas do KDD objetivando a criação de um *framework* de gestão de desempenho baseado em modelos de ML, com o objetivo de prever os impactos das variáveis analisadas na produtividade.

6 Referências

ANDRIYANA, Yudhie et al. Spatial Durbin Model with Expansion Using Casetti's Approach: A Case Study for Rainfall Prediction in Java Island, Indonesia. **Mathematics**, v. 12, n. 15, p. 2304, 2024.

BAMBINI, Martha Delphino; BONACELLI, Maria Beatriz Machado. Ecossistemas Agtech no Brasil: localização, caracterização e atores envolvidos. In: **Embrapa Informática Agropecuária-Artigo em canais de congresso (ALICE)**. 2019.

BERTRAND, J. W. M.; FRANSOO, J. C. Operations management research methodologies using quantitative modeling. **International Journal of Operations & Production Management**, v. 22, n. 2, p. 241–264, 2002.

DOS SANTOS, Bruno Samways; STEINER, Maria Teresinha Arns; LIMA, Rafael Henrique Palma. Proposal of a method to classify female smokers based on data mining techniques. **Computers & Industrial Engineering**, v. 170, p. 108363, 2022.

FALCON, Walter P.; NAYLOR, Rosamond L.; SHANKAR, Nikhil D. Rethinking global food demand for 2050. **Population and Development Review**, v. 48, n. 4, p. 921-957, 2022.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996.

GUPTA, Bhumika et al. Towards information discovery on large scale data: state-of-the-art. In: **2018 International Conference on Soft-Computing and Network Security (ICSNS)**. IEEE, 2018. p. 1-9.

GUPTA, Sushil et al. Opportunities in farming research from an operations management perspective. **Production and Operations Management**, v. 32, n. 6, p. 1577-1596, 2023.

HAN, Jiawei; PEI, Jian; TONG, Hanghang. **Data mining: concepts and techniques**. Morgan Kaufmann, 2022.

KÄRNER, Ene. The future of agriculture is digital: Showcasing e-Estonia. **Frontiers in Veterinary Science**, v. 4, p. 151, 2017.

LLATAS, Carmen et al. Application of Knowledge Discovery in Databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design. **Journal of Building Engineering**, v. 91, p. 109546, 2024.

PALLATHADKA, Harikumar et al. Impact of machine learning on management, healthcare and agriculture. **Materials Today: Proceedings**, v. 80, p. 2803-2806, 2023.

PEPPES, Nikolaos et al. Performance of machine learning-based multi-model voting ensemble methods for network threat detection in agriculture 4.0. Sensors, v. 21, n. 22, p. 7475, 2021.

PENG, Xinghao; LI, Yanting; TSUNG, Fugee. A graph attention network with spatio-temporal wind propagation graph for wind power ramp events prediction. **Renewable Energy**, v. 236, p. 121280, 2024.

PRAMILARANI, K.; KUMARI, P. Vasanthi. Cost-based Random Forest Classifier for

Intrusion Detection System in Internet of Things. **Applied Soft Computing**, v. 151, p. 111125, 2024.

PRASAD, Guru et al. Exploratory Data Analysis using Autoviz for Machine Learning Classification Problem. In: **2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)**. IEEE, 2024. p. 496-500.

SARKER, Iqbal H. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. **SN Computer Science**, v. 2, n. 5, p. 377, 2021.

SCHEMBERGER, Elder E. et al. Data mining for the assessment of management areas in precision agriculture. **Engenharia Agrícola**, v. 37, p. 185-193, 2017.

SHAIKH, Tawseef Ayoub; RASOOL, Tabasum; LONE, Faisal Rasheed. Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. **Computers and Electronics in Agriculture**, v. 198, p. 107119, 2022.

STANČIN, Igor; JOVIĆ, Alan. An overview and comparison of free Python libraries for data mining and big data analysis. In: **2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)**. IEEE, 2019. p. 977-982.

VAN KLOMPENBURG, Thomas; KASSAHUN, Ayalew; CATAL, Cagatay. Crop yield prediction using machine learning: A systematic literature review. **Computers and Electronics in Agriculture**, v. 177, p. 105709, 2020.

VOGEL, Elisabeth et al. The effects of climate extremes on global agricultural yields. **Environmental Research Letters**, v. 14, n. 5, p. 054010, 2019.