INSTITUTO FEDERAL GOIANO - CAMPUS MORRINHOS
CURSO SUPERIOR DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

LUCAS DANIEL DA SILVA

ANÁLISE DE PADRÕES PRODUTIVOS E CLIMÁTICOS DA SOJA EM GOIÁS POR MEIO DE TÉCNICAS DE CLUSTERIZAÇÃO

MORRINHOS - GO
2026

LUCAS DANIEL DA SILVA

ANÁLISE DE PADRÕES PRODUTIVOS E CLIMÁTICOS DA SOJA EM GOIÁS POR MEIO DE TÉCNICAS DE CLUSTERIZAÇÃO

Monografia apresentada ao Curso Superior de Bacharelado em Ciência da Computação do Instituto Federal Goiano – Campus Morrinhos, como requisito parcial para obtenção de título de Bacharel em Ciência da Computação.
Área de concentração: Ciência da Computação – Inteligência Artificial.
Orientador: Prof. Dr. Jesmmer da Silveira Alves.

MORRINHOS - GO
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema Integrado de Bibliotecas do IF Goiano - SIBi

Silva, Lucas Daniel da
S586a ANÁLISE DE PADRÕES PRODUTIVOS E CLIMÁTICOS DA SOJA EM GOIÁS POR MEIO DE TÉCNICAS DE CLUSTERIZAÇÃO / Lucas Daniel da Silva. Morrinhos 2026.

72f. il.

Orientador: Prof. Dr. Jesmmer da Silveira Alves.
Tcc (Bacharel) - Instituto Federal Goiano, curso de 0419204 - [MO.GRAD] Bacharelado em Ciência da Computação - Morrinhos (Campus Morrinhos).
1. Zoneamento agroclimático. 2. Produtividade da soja em Goiás. 3. Análise espacial. 4. Aprendizado não supervisionado. 5. Reanálise climática. I. Título.

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO

PARA DISPONIBILIZAR PRODUÇÕES TÉCNICO-CIENTÍFICAS NO REPOSITÓRIO INSTITUCIONAL DO IF GOIANO

Com base no disposto na Lei Federal nº 9.610, de 19 de fevereiro de 1998, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia Goiano a disponibilizar gratuitamente o documento em formato digital no Repositório Institucional do IF Goiano (RIIF Goiano), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IF Goiano.

IDENTIFICAÇÃO DA PRODUÇÃO TÉCNICO-CIENTÍFICA

- ☐ Tese (doutorado)
- ☐ Dissertação (mestrado)
- ☐ Monografia (especialização)
- ☑ TCC (graduação)
- ☐ Artigo científico
- ☐ Capítulo de livro
- ☐ Livro
- ☐ Trabalho apresentado em evento
- ☐ Produto técnico e educacional - Tipo:

Nome completo do autor: Lucas Daniel da Silva

Matrícula: 2018104201940186

Título do trabalho:

ANÁLISE DE PADRÕES PRODUTIVOS E CLIMÁTICOS DA SOJA EM GOIÁS POR MEIO DE TÉCNICAS DE CLUSTERIZAÇÃO

RESTRIÇÕES DE ACESSO AO DOCUMENTO

Documento confidencial: ☑ Não ☐ Sim, justifique:

Informe a data que poderá ser disponibilizado no RIIF Goiano: 06 / 04 / 2026

O documento está sujeito a registro de patente? ☐ Sim ☑ Não

O documento pode vir a ser publicado como livro? ☐ Sim ☑ Não

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O(a) referido(a) autor(a) declara:

• Que o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;

• Que obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autoria, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia Goiano os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;

• Que cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia Goiano.

Morrinhos/GO — Local

06 / 04 / 2026 — Data

gov.br Documento assinado digitalmente
LUCAS DANIEL DA SILVA
Data: 06/04/2026 09:33:51-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor e/ou detentor d

Ciente e de acordo:

gov.br Documento assinado digitalmente
JESMMER DA SILVEIRA ALVES
Data: 06/04/2026 09:44:30-0300
Verifique em https://validar.iti.gov.br

Assinatura do(a) orientador(a)

SERVIÇO PÚBLICO FEDERAL
MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA GOIANO

Ata nº 19/2026 - CCEG-MO/CEG-MO/DE-MO/CMPMHOS/IFGOIANO

ATA DE DEFESA DE TRABALHO DE CURSO (TC)

No dia primeiro do mês de abril de 2026, por meio de videoconferência, realizou-se a sessão pública de defesa de Trabalho de Curso (TC) do Curso Bacharelado em Ciência da Computação, do acadêmico Lucas Daniel da Silva, sob orientação do professor Jesmmer da Silveira Alves, intitulada ANÁLISE DE PADRÕES PRODUTIVOS E CLIMÁTICOS DA SOJA EM GOIÁS POR MEIO DE TÉCNICAS DE CLUSTERIZAÇÃO. Compuseram a Banca Examinadora os professores:

Orientador

Professor Dr. Jesmmer da Silveira Alves

Membro 2

Professora Dra. Leila Roling Scariot da Silva

Membro 3

Professor Dr. Antônio Neco de Oliveira

Após a exposição oral, o candidato foi arguido pelos membros da banca, os quais reuniram-se reservadamente, e decidiram, **APROVADO**. Para constar, redigi a presente Ata, que aprovada por todos os presentes, vai assinada por mim, Orientador do TC, e pelos demais membros da banca.

_____

Prof. Orientador

Assinado eletronicamente

_____

Prof.(a)/Membro 2

Assinado eletronicamente

_____

Prof.(a)/Membro 3

Assinado eletronicamente

Documento assinado eletronicamente por:

- **Jesmmer da Silveira Alves**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 06/04/2026 09:13:56.
- **Antonio Neco de Oliveira**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 06/04/2026 09:17:22.
- **Leila Roling Scariot da Silva**, **COORDENADOR(A) DE CURSO - FUC1 - CCBCC-MO** , em 06/04/2026 09:19:16.

Este documento foi emitido pelo SUAP em 06/04/2026. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifgoiano.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 807304
Código de Autenticação: 226a5efc9f

INSTITUTO FEDERAL GOIANO
Campus Morrinhos
Rodovia BR-153, Km 633, Zona Rural, SN, Zona Rural, MORRINHOS / GO, CEP 75650-000
(64) 3413-7900

DEDICATÓRIA

Dedico este trabalho ao meu falecido tio José Gonçalves da Silva e às minhas falecidas avós Maria da Conceição e Marinita Cordeiro da Silva, por todo o apoio, incentivo e motivação em vida.

AGRADECIMENTOS

Expresso meus sinceros agradecimentos primeiramente a Deus, nosso Senhor Jesus Cristo, que em seu infinito amor e infinita sabedoria me concedeu forças para continuar quando tudo pareceu perdido.

Em seguida, agradeço ao meu orientador, Prof. Dr. Jesmmer da Silveira Alves, pelo acolhimento, apoio e excelente orientação ao longo do desenvolvimento deste trabalho. Sua motivação nos momentos em que tudo aparentava estagnado foi determinante para que eu não desistisse.

Por fim, e não menos importante, agradeço aos meus pais, José Lino da Silva Filho e Ana Maria da Silva, aos meus avôs, ao meu irmão, às minhas irmãs, aos meus sobrinhos, aos amigos e colegas que estiveram comigo durante essa caminhada, oferecendo apoio, palavras de incentivo e conforto nos momentos mais difíceis. A todos que contribuíram com minha jornada até aqui, deixo meu mais sincero agradecimento.

RESUMO

A soja consolida-se como a principal commodity de Goiás, contudo, a análise regional do setor carece de ferramentas que transcendam as médias estaduais e revelem padrões locais. Este trabalho aplicou técnicas de aprendizado de máquina para identificar agrupamentos produtivos e climáticos no estado, processando dados históricos de produtividade e meteorologia via linguagem Python. A metodologia compreendeu o pré-processamento de dados sazonais e a aplicação do algoritmo de clusterização K-Means. O modelo, configurado com cinco grupos (*k=5*), foi validado por métricas como as de Silhueta e Índice Caliński-Harabasz, apresentando segmentação consistente. Os resultados evidenciaram um grupo de alta performance tecnológica, mantendo produtividades superiores a 3.470 kg/ha mesmo sob regimes pluviométricos restritivos, o que sugere o uso intensivo de irrigação. Os resultados indicam que a aplicação algoritmos de agrupamento é eficaz para a extração de conhecimento de bases de dados complexas, permitindo a distinção automática entre zonas de aptidão natural e zonas de elevado manejo tecnológico.

Palavras-chave: Zoneamento agroclimático, Produtividade da soja em Goiás, Análise espacial, Aprendizado não supervisionado, Reanálise climática.

ABSTRACT

Soybeans are consolidating their position as the main commodity in Goiás, however, regional analysis of the sector lacks tools that transcend state averages and reveal local patterns. This work applied machine learning techniques to identify productive and climatic clusters in the state, processing historical productivity and meteorological data using the Python language. The methodology included the pre-processing of seasonal data and the application of the K-Means clustering algorithm. The model, configured with five groups (k=5), was validated by metrics such as Silhouette and Caliński-Harabasz Index, showing consistent segmentation. The results highlighted a group with high technological performance, maintaining productivities above 3,470 kg/ha even under restrictive rainfall regimes, suggesting the intensive use of irrigation. The results indicate that the application of clustering algorithms is effective for extracting knowledge from complex databases, allowing automatic distinction between zones of natural suitability and zones of high technological management.

Keywords: Agroclimatic zoning, Soybean productivity in Goiás, Spatial analysis, Unsupervised learning, Climate reanalysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE CÓDIGOS

ABREVIATURAS E SIGLAS

API — Application Programming Interface.

SIDRA — Sistema IBGE de Recuperação Automática.

PAM — Produção Agrícola Municipal.

ECMWF — European Centre for Medium-Range Weather Forecasts.

CDS — Climate Data Store.

IPEA — Instituto de Pesquisa Econômica Aplicada.

KDD — Knowledge Discovery in Databases.

ETL — Extract, Transform and Load.

URL — Uniform Resource Locator.

ZARC — Zoneamento Agrícola de Risco Climático.

SUMÁRIO

1 INTRODUÇÃO

A soja está entre as principais culturas agrícolas brasileiras, tanto em termos de volume de produção quanto de valor econômico. A partir da safra 2019/20, o Brasil se tornou o maior produtor de soja do mundo, superando os Estados Unidos (Agência FPA, 2021), como pode ser visto na Figura 1. Um dos principais motivos para essa ascensão é a qualidade do grão brasileiro em relação à sua taxa de proteína e ao seu custo de produção, o que o torna também o maior exportador do grão no mundo.

Figura 1 — Brasil x EUA na exportação de soja.

Fonte: AGÊNCIA FPA (2021).

A nível estadual, Goiás se destaca como um dos maiores produtores de soja do país, ocupando o terceiro lugar (G1, 2023). Todavia, ainda que vários municípios goianos ocupem notoriedade no ranking nacional entre os 15 maiores produtores do Brasil, com destaque para Rio Verde que tem aparecido como segundo maior produtor de soja do país (GOIÁS GOVERNO DO ESTADO, 2023), essa superioridade não é homogênea.

A produtividade e as condições climáticas variam significativamente entre as diversas regiões do estado, mas não há uma caracterização sistemática que explique a heterogeneidade dos valores de produção, as análises se limitam em constatar

quem produz mais, mas falham em correlacionar as condições que isso ocorre. Assim, há uma clara lacuna na categorização dos padrões agroclimáticos da região, quais combinações podem explicar os municípios que mais produzem no estado? E os que menos produzem? Para a otimização da produção e a expansão da cultura, é fundamental identificar os fatores que levam ao sucesso de alguns municípios e se esses padrões podem ser replicados.

Nos últimos anos, avanços em técnicas de análise de dados, como análise de agrupamento (clusterização) permitem identificar padrões entre municípios produtores. Esses agrupamentos são úteis tanto para análises exploratórias quanto para políticas públicas, como zoneamentos agrícolas, subsídios técnicos e estratégias de mitigação de risco climático.

Em trabalhos como o de Dias e Pascoal (2023), que focam em análises de recortes regionais, as análises são limitadas a aplicações pontuais, nesse caso, a avaliação de risco para seguros agrícolas, utilizando apenas chuva (CHIRPS[1]) como variável climática, sem considerar a influência complementar de temperatura média.

Diante desse contexto, este trabalho teve como objetivo identificar padrões produtivos e climáticos que influenciam o desempenho da produção de soja em Goiás, por meio da combinação de variáveis climáticas (temperatura média e precipitação) e produtivas (rendimento), visando subsidiar estratégias para otimização da cultura no estado.

A análise desses grupos permitirá não apenas compreender fatores chave da alta produtividade, mas também fornecer subsídios para o desenvolvimento de políticas agrícolas e estratégias que auxiliem outros municípios goianos a alcançar mais lugares de destaque no âmbito nacional.

[1] **CHIRPS** (Climate Hazards Group InfraRed Precipitation with Station data) é uma base de dados de precipitação em escala global, que combina dados de satélite de infravermelho com dados de estações de medição de chuva em solo.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

Identificar e caracterizar padrões produtivos e climáticos associados ao desempenho da soja nos municípios do estado de Goiás, por meio da integração de dados agroclimáticos e aplicação de técnicas de aprendizado não supervisionado.

1.1.2 Objetivos Específicos

- Estruturar uma base de dados integrada a partir de informações produtivas da soja (PAM/IBGE) e variáveis climáticas (ERA5-Land).
- Realizar o pré-processamento e harmonização temporal dos dados considerando o calendário agrícola (Ano Safra).
- Aplicar técnicas de clusterização para segmentação dos municípios segundo perfis agroclimáticos.
- Avaliar a consistência estatística dos agrupamentos por meio de métricas de validação.
- Analisar e interpretar os padrões identificados, relacionando-os ao desempenho produtivo da soja em Goiás.

1.2 JUSTIFICATIVA

A escolha de Goiás como recorte territorial tem como fundamentação sua posição como o terceiro maior produtor de soja do Brasil (G1, 2023), abrigando polos de excelência como Rio Verde. No entanto, a alta produtividade média do estado pode mascarar disparidades regionais que ainda carecem de explicação técnica. O problema é relevante pois, em um cenário de instabilidade climática, entender maneiras de manter (e expandir para outros cenários) a alta produtividade agrícola é uma questão de segurança econômica e resiliência para o estado.

A lacuna na literatura e nas análises governamentais reside na fragmentação dos dados: observa-se quem produz mais, mas exclui-se a correlação entre o rendimento e os dados climáticos de precipitação e temperatura. Enquanto trabalhos

recentes (DIAS; PASCOAL, 2023) limitam-se ao uso da variável chuva, este estudo inova ao integrar a temperatura média à análise de rendimento. Essa integração multidimensional permite identificar padrões que variáveis isoladas não conseguem capturar, oferecendo uma visão holística e tecnicamente mais robusta.

Nesse sentido, este trabalho propõe a integração de dados produtivos da soja e variáveis climáticas em uma abordagem de ciência de dados baseada em técnicas de clusterização. Essa estratégia permite identificar padrões agroclimáticos e contribuir para a compreensão da dinâmica produtiva da cultura no estado.

Além de sua contribuição científica, a abordagem adotada pode fornecer subsídios para o planejamento agrícola, apoio à tomada de decisão e formulação de políticas públicas, além de apresentar potencial de replicação em outras regiões agrícolas.

1.3 ESTRUTURA DO TRABALHO

Este trabalho está organizado em seis capítulos, são eles: **Capítulo 1**, apresenta-se a introdução, incluindo a contextualização, justificativa e objetivos da pesquisa. O **Capítulo 2** aborda a fundamentação teórica, detalhando os conceitos de clusterização, indicadores produtivos e climáticos relevantes, além das ferramentas utilizadas, como o SIDRA/IBGE e o ERA5-Land, o uso de KDD e ETL e por fim, um resumo dos trabalhos relacionados. O **Capítulo 3** descreve a metodologia da pesquisa, explicando os procedimentos adotados e o fluxo de trabalho para a coleta, processamento e análise dos dados. O **Capítulo 4** detalha o desenvolvimento do sistema, abordando a implementação de cada etapa definida, a integração e a aplicação dos algoritmos de clusterização e validadores. No **Capítulo 5**, são analisados e discutidos os resultados obtidos, destacando os padrões identificados entre os municípios goianos. Por fim, o **Capítulo 6** apresenta as considerações finais, com a síntese dos achados, limitações e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a variabilidade regional, as fontes de dados que servem de base para o desenvolvimento da pesquisa, detalhando os conjuntos de dados produtivos e climáticos utilizados, explora as técnicas de clusterização e validação que serão aplicadas para a análise dos padrões espaciais e temporais, apresenta o processo de descoberta de conhecimento e finaliza com a análise dos trabalhos relacionados.

2.1 HETEROGENEIDADE REGIONAL

A análise da produtividade da soja em Goiás revela uma grande heterogeneidade espacial, onde o desempenho do estado é altamente condicionado por polos regionais específicos. Conforme dados da Produção Agrícola Municipal (PAM) do IBGE, interpretados pelo Governo do Estado de Goiás (2023), apenas três municípios (Rio Verde, Jataí e Cristalina) concentraram 25,1% de todo o volume produzido no território goiano em 2022.

O destaque de Rio Verde, que alcançou a marca de 1,6 milhão de toneladas (segunda maior produção nacional), evidencia um cenário de especialização produtiva que contrasta com a realidade de outras regiões do estado. Essa disparidade entre os municípios demonstra que o sucesso agrícola não é um fenômeno homogêneo, mas sim o resultado da interação de aptidão climática e intensificação tecnológica localizada. Portanto, a análise desses dados na presente fundamentação serve para documentar a existência de lacunas de rendimento e zonas de eficiência distintas, o que justifica a aplicação de técnicas de *Machine Learning* para identificar os padrões agroclimáticos subjacentes que sustentam esses diferentes patamares de produtividade.

2.2 FONTES DE DADOS

A robustez da análise de clusterização depende diretamente da qualidade e da consistência dos dados de entrada. Para garantir essas características, é necessário o uso de fontes de dados oficiais e de alta confiabilidade para os

indicadores produtivos e climáticos, harmonizando bases de naturezas distintas para compor um conjunto de dados integrado.

2.2.1 Sidra, Sidra Api E Pam

O Sistema IBGE de Recuperação Automática (SIDRA) constitui a principal base de dados estatísticos do Instituto Brasileiro de Geografia e Estatística (IBGE). Ele disponibiliza informações socioeconômicas, demográficas e de produção agropecuária, sendo amplamente utilizado em estudos sobre o setor agrícola nacional (IBGE, 2023c). Para a agricultura em específico, o SIDRA concentra os resultados da PAM, fonte oficial sobre área plantada, área colhida, quantidade produzida e rendimento médio das principais culturas agrícolas brasileiras (IBGE, 2023b). Essa pesquisa é fundamental para análises regionais e temporais da produtividade da soja, permitindo identificar tendências e desigualdades entre municípios.

A confiabilidade dos modelos computacionais aplicados à agricultura depende de fatores como a abrangência e regularidade das fontes de dados utilizadas. Nesse contexto, a PAM destaca-se pela sua vasta capacidade de cobertura, uma vez que, segundo o IBGE (2023c), o inquérito é realizado anualmente em todo o território nacional, fornecendo informações escalonadas desde o nível federal até ao detalhamento municipal. Esta estruturação permite que a observação dos dados transite facilmente entre o panorama macroeconômico e as nuances locais. Por conseguinte, compreende-se que a atualização contínua e periódica destas informações atesta a consistência estrutural das tabelas fornecidas, garantindo a robustez do histórico de dados para a investigação, mesmo face a eventuais atrasos no calendário oficial de divulgação.

Segundo o IBGE (2023c), o SIDRA organiza essas informações em tabelas de consulta parametrizáveis, entre elas a tabela 5457. A Tabela 5457 (Área plantada ou destinada à colheita, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporárias e permanentes) atua como o alicerce dos dados produtivos, é através dela que é extraída a variável de rendimento médio (kg/ha), que serve como atributo quantitativo para o algoritmo diferenciar o nível de eficiência técnica de cada município.

Para ampliação das formas de acesso aos dados, o IBGE disponibiliza a API SIDRA, que permite recuperar informações diretamente por meio de requisições web[2]. Essa API fornece a automatização do processo de coleta e integração de dados, prática essencial em aplicações computacionais que necessitam de atualização periódica (IBGE, 2023a), possibilitando a coleta de séries históricas de produtividade da soja por município do estado goiano, organizando os dados de forma estruturada para análise. Assim, o uso da API não apenas simplifica a coleta, como garante maior consistência e transparência na formação da base de dados para clusterização. A API do SIDRA pode ser utilizada para obter quaisquer dos dados de pesquisas disponíveis no sistema do SIDRA, é claro, isto inclui também os dados de tabela da PAM, que serão utilizados durante o decorrer deste trabalho.

2.2.2 Dados Climáticos e a Base ERA5-Land

Fatores como a oscilação das temperaturas médias, a umidade de orvalho e os índices de precipitação acumulada podem alterar drasticamente os resultados da colheita, conforme observado em análises sobre variabilidade climática (LISBINSKI, 2025). Por essa razão, estabelecer uma relação entre o clima e a produtividade municipal requer fontes que garantam continuidade temporal e espacial, elementos que nem sempre estão presentes nos registros de estações de superfície isoladas.

O emprego de dados de reanálise justifica-se, primordialmente, pelas limitações inerentes à rede física de monitoramento meteorológico. Conforme documentado por Alvares et al. (2013), a densidade das estações no Brasil não apresenta uniformidade, o que resulta em "vazios" de informação que dificultam a caracterização climática em municípios distantes dos centros de monitoramento. Essa heterogeneidade exige o uso de equações multivariadas ou métodos de interpolação para estimar variáveis em locais sem sensores físicos (ALVARES et al., 2013). Tal cenário torna-se problemático para estudos de longa duração, pois falhas operacionais ou descontinuidades nas medições das estações municipais ao longo das décadas podem inviabilizar a construção de um painel de dados homogêneo e estatisticamente confiável.

[2] **Requisição web** é uma forma de comunicação (troca) de dados entre um cliente e servidor na internet, normalmente utilizando o protocolo HTTP, que padroniza e apresenta convenções de como realizar essa comunicação de maneira segura.

Nesse contexto, a reanálise climática surge como um método científico robusto para a reconstrução do estado passado da atmosfera. Conceitualmente, a reanálise consiste na combinação de modelos físicos de previsão numérica do tempo com uma vasta gama de observações históricas, como dados de sensores de satélite, radares e as próprias estações terrestres, processados por meio de técnicas de assimilação de dados (HERSBACH et al., 2020). O resultado desse processamento é a geração de uma grade contínua e tridimensional que recria o comportamento climático de forma consistente, eliminando as lacunas espaciais e temporais comuns em registros manuais. Portanto, a reanálise não é uma mera estimativa, mas uma síntese física que preserva a importância das séries históricas ao permitir o mapeamento de riscos climáticos e ciclos de anomalias com precisão científica.

A adoção da base ERA5-Land, desenvolvida pelo Centro Europeu de Previsões Meteorológicas a Médio Prazo (ECMWF), assegura a conformidade necessária para esse nível de análise regional. Este conjunto de dados oferece alta resolução espacial e informações detalhadas sobre variáveis complexas, como umidade do solo e radiação solar, superando as limitações técnicas de bases globais menos refinadas (MUÑOZ-SABATER et al., 2021). Dessa forma, a utilização desta base garante que os modelos de agrupamento operem sobre uma matriz de dados contínua, permitindo que as variações observadas na produtividade da soja sejam correlacionadas a variáveis climáticas fidedignas e espacialmente completas para cada unidade municipal analisada.

2.3 TÉCNICAS DE CLUSTERIZAÇÃO E VALIDAÇÃO

A utilização de técnicas de clusterização como ferramenta para análise de conjuntos de dados (*datasets*) permite ir além da análise descritiva. A aplicação desses métodos possibilita agrupar instâncias com características semelhantes, o que se aplica diretamente aos municípios goianos de acordo com as variáveis selecionadas para coleta. Segundo Jain (2010), essas técnicas são ferramentas de aprendizado não supervisionado que buscam identificar padrões ocultos na massa de dados sem a necessidade de classes predefinidas. Na prática, municípios com produtividades próximas e regimes climáticos semelhantes podem ser classificados

em um mesmo grupo, o que facilita a análise estratégica e as tomadas de decisão além das variáveis de geolocalização.

O K-Means consolida-se como uma das técnicas de clusterização mais bem estabelecidas na literatura, sendo amplamente reconhecido por sua capacidade de encontrar agrupamentos naturais de dados (LIAKOS et al., 2018). O algoritmo baseia-se na partição dos dados em número *K* de grupos, buscando minimizar a variabilidade interna de cada cluster. Ao configurar um modelo de clusterização via K-Means, deve-se definir esse número de destino *K*, que indica a quantidade de centroides (pontos representativos de cada cluster) desejados no modelo. O algoritmo atribui cada ponto de dado de entrada a um dos clusters através da minimização da soma de quadrados dentro do cluster (inércia) (MICROSOFT, 2024).

O roteiro de execução desse algoritmo é um processo iterativo que depende de critérios específicos de parada para garantir a estabilidade da solução. Segundo a Microsoft (2024), a execução é concluída sob duas condições principais: primeiro, quando os centroides se estabilizam, o que significa que as atribuições de pontos individuais aos clusters não sofrem mais alterações e o algoritmo convergiu em uma solução; segundo, quando o algoritmo atinge o número máximo de iterações predefinidas. Por ser sensível à posição inicial dos centroides, é comum que o algoritmo seja executado múltiplas vezes para evitar que o resultado fique concentrado em um mínimo local, garantindo que a partição final seja estatisticamente robusta.

Apesar de sua eficiência, *o* K-Means apresenta limitações matemáticas críticas, como a sensibilidade a valores extremos (*outliers*). Como o centroide é calculado pela média aritmética, um município com dados discrepantes exerce uma "força de atração" desproporcional, deslocando o centroide e distorcendo a coesão do grupo. Além disso, o algoritmo assume que os clusters possuem um formato esférico, ou seja, que a variância dos dados é uniforme em todas as direções.

Essa suposição é problemática para dados agroclimáticos que apresentem correlações lineares ou distribuições alongadas. Por exemplo, se a relação entre precipitação e rendimento formar um padrão diagonal e estreito no espaço de atributos, o algoritmo terá dificuldade em capturar essa estrutura, tendendo a dividir o agrupamento de forma artificial para satisfazer sua premissa geométrica de circularidade. Por fim, a sensibilidade à escala exige que os dados sejam padronizados via *StandardScaler* (PEDREGOSA et al., 2011), garantindo que

variáveis com grandes magnitudes nominais não enviesem o cálculo da distância euclidiana. Devido a essas particularidades, outras técnicas poderiam ser consideradas para aplicação de clusterização em cenários específicos, conforme o Quadro 1.

Quadro 1 — Exemplos de técnicas de clusterização adicionais.

Técnica	**Resumo**
Hierárquica	Gera dendrogramas[3] que revelam relações hierárquicas entre grupos.
DBSCAN	Adequado para detectar clusters de formatos arbitrários e ignorar ruídos.
GMMs (Modelos de Mistura Gaussiana)	Permitem clusters probabilísticos, úteis quando há sobreposição de dados.

Fonte: Adaptado de MURTAGH; CONTRERAS (2012), ESTER et al. (1996) e BISHOP (2006).

A validação dos clusters é essencial para garantir que a segmentação reflita diferenças reais na estrutura dos dados e não seja apenas um método algorítmico. Dado que a definição do valor de *k* é externa ao algoritmo, o uso de métricas de desempenho torna-se indispensável para comprovação de que o número adotado é o ideal. No Quadro 2, apresentam-se três das métricas mais utilizadas para validação dos clusters.

[3] **Dendrograma** é um diagrama de árvore que exibe os grupos formados por agrupamento de observações em cada passo e em seus níveis de similaridade.

Quadro 2 — Métodos de validação de cluster e uma breve definição.

Método	**Descrição**
Coeficiente da Silhueta	Combina proximidade intra-cluster[4] e separação inter-cluster[5], gerando um índice que varia entre -1 e 1, onde valores próximos de 1 indicam boa definição dos agrupamentos.
Método do Cotovelo	Analisa a soma dos quadrados intra-cluster em função do número de clusters. O ponto de inflexão do gráfico (“cotovelo”) sugere um número adequado de grupos ao equilibrar a redução da variabilidade interna com o aumento da complexidade do modelo.
Índice Caliński-Harabasz	Avalia a razão entre a dispersão entre clusters e a dispersão dentro dos clusters. Valores mais elevados indicam maior separação entre os grupos e maior coesão interna.

Fonte: Adaptado de ROUSSEEUW (1987), KODINARIYA et al (2013) e CALIŃSKI; HARABASZ (1974).

Os métodos de validação apresentados no Quadro 2 representam métricas utilizadas para verificar se a segmentação obtida reflete a estrutura dos dados analisados, conferindo maior confiabilidade à interpretação dos resultados. Para dados agroclimáticos do estado de Goiás, onde a variabilidade interanual de precipitação e temperatura pode ser acentuada, a validação estatística dos clusters torna-se especialmente relevante. Esse processo de validação é empregado para avaliar se os grupos identificados pelo algoritmo correspondem a padrões consistentes de comportamento da soja em relação às condições climáticas, contribuindo para a construção de uma base analítica voltada ao planejamento rural e à formulação de estratégias de mitigação de riscos agrícolas em escala municipal.

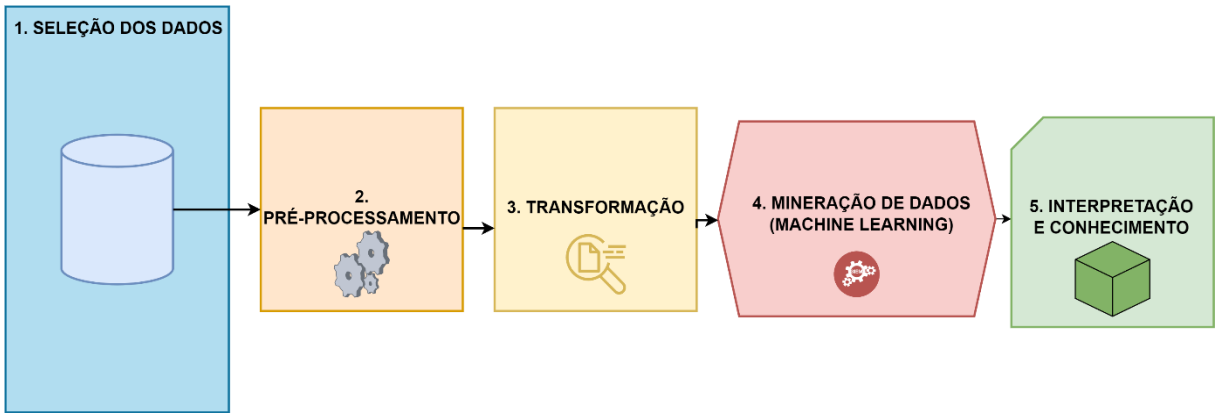
[4] **Intra-cluster** representa a distância entre elementos de um mesmo cluster e o quão distantes estão de seu centroide.

[5] **Inter-cluster** representa a distância dos clusters em si, o que normalmente indica o isolamento dos grupos e uma classificação provavelmente mais precisa.

2.4 DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS E ETL

Segundo Fayyad et al. (1996), a Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* — KDD) constitui um processo metodológico, iterativo e interativo voltado à transformação de dados brutos em conhecimento útil. Esse processo é tradicionalmente descrito em cinco etapas principais: seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados. Essa metodologia é dividida em 5 principais etapas, apresentadas na Figura 2.

Figura 2 — Etapas do processo de KDD.



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

1. **Seleção:** Definição dos subconjuntos de dados relevantes para análise.
2. **Pré-processamento:** Etapa de limpeza para remoção de ruídos e tratamento de dados ausentes (nulos), garantindo que a base seja íntegra.
3. **Transformação:** Adequação dos dados para o formato analítico, incluindo normalizar e reduzir a dimensionalidade, para facilitar a ação dos algoritmos.
4. **Mineração de dados (*data mining*):** Aplicação de algoritmos de *machine learning* (como as técnicas de clusterização) para extrair padrões para análise.
5. **Interpretação e Conhecimento:** Análise dos resultados obtidos para verificar se pode ser extraído conhecimento útil para tomadas de decisões.

No contexto da preparação e organização dos dados, podem ser empregados processos de engenharia de dados que operacionalizam parte dessas etapas. Conforme Ferreira et al. (2010), o ETL *(Extract, Transform, Load)* corresponde ao processo responsável pela extração de dados provenientes de fontes heterogêneas, sua transformação e adequação para análise e, por fim, sua consolidação em uma base estruturada. Nesse sentido, enquanto o KDD representa o processo metodológico voltado à descoberta de conhecimento, o ETL atua como mecanismo técnico de preparação e organização dos dados, contribuindo para a implementação das etapas iniciais do processo de KDD. O processo de ETL não substitui o KDD, mas atua como mecanismo técnico de implementação das etapas de preparação de dados, especialmente seleção, pré-processamento e transformação

2.5 TRABALHOS RELACIONADOS

A investigação sobre a interdependência entre fatores climáticos, desempenho agronômico e a aplicação de técnicas de inteligência de dados tem se consolidado como uma vertente estratégica na literatura científica contemporânea (LIAKOS et al., 2018). Nesse cenário, a utilização de algoritmos de agrupamento, ou clusterização, destaca-se como uma ferramenta robusta para a decodificação de padrões complexos em grandes volumes de dados, permitindo a otimização de processos decisórios no agronegócio. A literatura indica que o sucesso da gestão agrícola moderna depende da capacidade de transformar grandes volumes de dados em zonas de manejo ou territórios produtivos com características homogêneas, permitindo identificar padrões produtivos e climáticos relevantes para a tomada de decisão no setor agrícola (ARAÚJO et al., 2013; ARSEGO et al., 2019).

O Instituto de Pesquisa Econômica Aplicada (IPEA), por exemplo, utilizou a clusterização para agrupar municípios com características produtivas semelhantes, identificando um cluster de alta produtividade que incluía importantes polos de Mato Grosso, Mato Grosso do Sul, Minas Gerais, Bahia e Goiás, como o município de Rio Verde (IPEA, 2017). O estudo utilizou 118 variáveis de análise inicialmente, finalizando com 83 variáveis finais após uma redução. Apesar do elevado número de variáveis, a proposta foi analisar a nível estadual as produções multiculturais,

apresentando os resultados de maneira generalista, sem entendimento específico dos múltiplos cenários coletados.

A aplicação dessa metodologia em estudos focados na relação entre produtividade e variáveis climáticas tem sido explorada em outras regiões. Como resultado da pesquisa de Araújo et al. (2013) na porção oeste do Paraná, a análise de agrupamento espacial foi empregada para investigar a influência da precipitação pluvial e da temperatura média do ar no rendimento da cultura. De forma similar, o estudo de Arsego et al. (2019) no Rio Grande do Sul aplicou a técnica para agrupar séries de produtividade da soja em grupos homogêneos e correlacioná-los com indicadores climáticos de larga escala.

Apesar das evidências de que o clima é um dos principais fatores determinantes da produção, observa-se uma lacuna metodológica com relação ao detalhamento térmico em estudos de regionalização no estado de Goiás. Trabalhos que investigaram a distribuição de clusters espaciais da soja no estado, como o de Peixoto e De Queiroz (2021), priorizaram variáveis de desenvolvimento econômico e indicadores socioeconômicos para identificar os chamados “territórios da soja”. No entanto, ao negligenciar a dimensão climática como o vetor primário de agrupamento, tais pesquisas deixam de explicar as variações de rendimento que ocorrem entre municípios com níveis socioeconômicos similares, mas sob regimes térmicos distintos.

A necessidade de uma análise granular das temperaturas é sustentada por evidências biológicas sobre o desenvolvimento da cultura. Conforme constatado por Farias (2019), o estresse térmico severo, caracterizado pela exposição prolongada a temperaturas elevadas, compromete o crescimento vegetativo e a retenção de vagens, tendo seus efeitos potencializados quando associado ao déficit hídrico. Portanto, a inclusão de médias térmicas no modelo de clusterização é crucial para capturar esses efeitos combinados que as análises puramente pluviométricas ou econômicas acabam por omitir.

A aplicação da clusterização a esse conjunto de dados pode permitir a identificação de grupos de municípios goianos que exibem respostas produtivas similares a regimes climáticos específicos. Essa classificação agroclimática territorial pode contribuir para o conhecimento técnico aplicado ao planejamento rural, orientando a formulação de políticas agrícolas e a adoção de estratégias de otimização da produção.

Diante das limitações observadas nos estudos anteriores, este trabalho propõe integrar a análise de padrões produtivos e climáticos em um mesmo contexto regional, onde o principal diferencial é a inclusão da temperatura média como variável de agrupamento em adição aos dados pluviais. Enquanto os estudos anteriores focaram em aspectos socioeconômicos (PEIXOTO E DE QUEIROZ, 2021) ou priorizaram a precipitação como único fator ambiental (ARSEGO ET AL., 2019), este trabalho amplia a análise ao tratar a temperatura como fator biológico determinante. Esta abordagem possibilita também identificar zonas de risco térmico ignoradas por modelos generalistas, preenchendo uma lacuna crítica para o planejamento rural e para a formulação de estratégias de adaptação agrícola no território goiano.

3 METODOLOGIA

Esta pesquisa classifica-se como aplicada, de natureza quantitativa, com abordagem descritiva e exploratória para a identificação de padrões agroclimáticos. A unidade de análise compreende o território do estado de Goiás, segmentado em nível municipal, em formato município e ano de safra utilizando uma arquitetura de software modular desenvolvida em linguagem Python (versão 3.12). Essa abordagem permite a replicabilidade do modelo para outras regiões agrícolas mediante o ajuste dos parâmetros de entrada, consolidando um método estatístico baseado em aprendizado de máquina para a caracterização de grandes volumes de dados.

As variáveis utilizadas no estudo integram o indicador de rendimento médio de soja (kg/ha), extraído via API PAM do Sistema IBGE de Recuperação Automática (SIDRA), e indicadores meteorológicos provenientes da base de reanálise ERA5-Land (ECMWF). De forma automatizada, foram recuperados dados de temperatura média, precipitação acumulada e umidade do solo através da biblioteca *Earthkit*. Para o tratamento desses dados, utilizou-se o ecossistema de bibliotecas *Pandas* e *GeoPandas* para limpeza e estruturação geoespacial, seguidos pela aplicação do algoritmo *StandardScaler* (Scikit-learn) para a padronização das escalas, garantindo que as variáveis climáticas e de rendimento tivessem o mesmo peso na análise estatística.

A etapa de mineração de dados foi executada por meio do algoritmo de agrupamento não supervisionado K-Means. Para determinar o número ideal de grupos (*k*), aplicou-se uma estratégia de validação multifatorial que combinou métricas matemáticas e visuais. O modelo foi validado pelo Método do Cotovelo (*Elbow Method*), para identificar o ponto de inflexão da inércia, e pelo Coeficiente de Silhueta (*Silhouette Score*), que avaliou a qualidade da separação entre os clusters. Complementarmente, utilizou-se o Índice Caliński-Harabasz e a análise de agrupamento hierárquico via dendrograma (Método de Ward) para confirmar a estabilidade das divisões propostas.

Os critérios finais para a definição dos resultados basearam-se na convergência entre as métricas de validação e a interpretabilidade agronômica dos agrupamentos. A análise comparativa por meio de um diagrama de caixa (*boxplot*) permitiu observar a distinção das variâncias de rendimento e clima entre os grupos,

confirmando que a configuração com cinco clusters (*k*=*5*) apresentava a melhor separação estatística e relevância prática para o planejamento rural. Assim, a regionalização agroclimática final foi consolidada em mapas temáticos e gráficos gerados via *Matplotlib* e *Seaborn*, fundamentando as conclusões deste estudo.

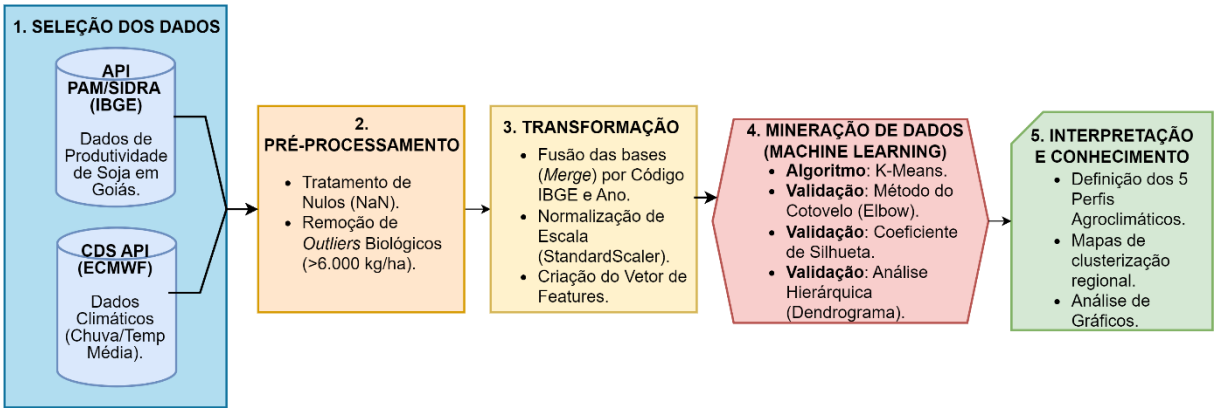
4 PIPELINE DE PROCESSAMENTO E ANÁLISE DE DADOS

Este capítulo descreve a síntese computacional da pesquisa, detalhando a arquitetura de sistema desenvolvida e a implementação das etapas do processo de KDD adaptado à pesquisa. A solução foi estruturada como um *pipeline*[6] de engenharia de dados modularizada, capaz de orquestrar a extração, transformação e modelagem de grandes volumes de dados produtivos e geoclimáticos.

4.1 ARQUITETURA DO SISTEMA E AMBIENTE EXPERIMENTAL

Para iniciar o desenvolvimento da aplicação, foi adaptado o modelo do KDD aplicado ao presente projeto, como visto na Figura 3.

Figura 3 — KDD adaptado ao contexto da pesquisa.



Fonte: Elaborado pelo Autor (2026).

Para garantir a reprodutibilidade dos experimentos e escalabilidade do código, a solução foi desenvolvida sob o paradigma da modularização, utilizando a linguagem Python. Diferente de abordagens que utilizam *scripts* monolíticos (implementação de toda lógica em um único script sequencial), o projeto foi segmentado em módulos específicos: extração (*extract*), transformação (*transform*) e carga (*load*), seguindo o padrão ETL.

[6] **Pipeline** é o processo de um sistema conectar várias tarefas de maneira sequencial ou paralela de modo: a processar dados, compilar códigos ou qualquer outro tipo de processamento complexo de maneira automatizada e estruturada.

A gestão de configurações foi centralizada em um módulo específico. Esta decisão arquitetural permite que parâmetros críticos, como o intervalo temporal da análise (2004-2024), os meses que compõem a safra da soja e as URLs (*Uniform Resource Locator*, endereço de um recurso na internet), sejam alterados sem a necessidade de refatoração do núcleo lógico do código. O Código 1 demonstra a definição dessas constantes que regem todo o fluxo de dados, pode-se observar, por exemplo, na linha 7 é definida a variável de *MESES_SAFRA* com base no Calendário Agrícola oficial (CONAB, 2022), a safra se concentra entre outubro (início do plantio) a abril (término da colheita), totalizando 7 meses de safra. Assim, o algoritmo foi configurado para descartar dados climáticos dos meses de maio à setembro, evitando ruídos na modelagem.

Código 1 — Centralização de Parâmetros de Configuração.

```
1   # Módulo config.py – Definição das fontes de dados oficiais
2   URL_PAM = "https://apisidra.ibge.gov.br/values"
3   URL_MALHA_MUNICIPAL_GOIAS_IBGE =
4   "https://geoftp.ibge.gov.br/.../GO_Municipios_2024.zip"
5   # Parametrização temporal do ciclo fenológico da soja
6   # Meses selecionados para garantir a correlação agroclimática correta
7   MESES_SAFRA = [
8   '01', '02', '03', '04', # Janeiro a Abril (Enchimento/Colheita)
9   '10', '11', '12']       # Outubro a Dezembro
10  (Plantio/Desenvolvimento)
```

Fonte: Elaborado pelo Autor (2026).

Adicionalmente, para lidar com a autenticação segura junto ao CDS, utilizou-se a biblioteca *earthkit*[7]. O sistema foi configurado para gerenciar automaticamente o cache das requisições realizadas, evitando *downloads* redundantes de arquivos pesados e otimizando o tempo de processamento de múltiplas execuções.

[7] **Earthkit:** Biblioteca Python que facilita integração aos dados climáticos disponibilizados pelo CDS.

4.2 SELEÇÃO E EXTRAÇÃO

A primeira etapa do KDD consiste na seleção e captura dos dados brutos. O desafio técnico nesta fase reside na heterogeneidade das fontes: a integração de uma API RESTful (PAM/SIDRA IBGE) com um serviço de dados de matrizes assíncrono (ECMWF/ERA5).

Antes de iniciar a extração climática, foi necessário definir matematicamente a área de interesse. Para evitar a inserção manual de coordenadas (o que tornaria o código frágil a mudanças territoriais), implementou-se em um módulo de código uma função de cálculo de geometria dinâmica.

O algoritmo consome o arquivo vetorial da malha municipal (.*shp* — *shapefile*) oficial do IBGE e calcula os limites extremos (*Total Bounds*) do estado de Goiás. O Código 2 apresenta essa implementação, destacando a aplicação de uma margem de segurança (*buffer*) de 0,25 graus na variável da linha 17, para garantir a cobertura dos pixels de borda, sem que haja exclusão de municípios, os valores de coordenadas derivam da leitura do arquivo de malha municipal do IBGE.

Código 2 — Cálculo do Bounding Box Geográfico.

```python
# Módulo geo_go_utils.py – Calcula a área de Goiás em Bouding Box
def calcular_bounding_box_goias():
    """
    Calcula os limites espaciais (Lat/Lon) baseados na malha do IBGE.
    """
    url = cfg.URL_MALHA_MUNICIPAL_GOIAS_IBGE
    try:
        # Leitura geoespacial direta da fonte oficial
        gdf = gpd.read_file(url)

        # Extração dos limites extremos: [Oeste, Sul, Leste, Norte]
        west, south, east, north = gdf.total_bounds

        # Margem de segurança para evitar cortes no grid do satélite
        margin = 0.25

        # Formatação para o padrão exigido pela API do CDS
        bbox_cds = [
            round(north + margin, 3),
            round(west - margin, 3),
            round(south - margin, 3),
            round(east + margin, 3)
        ]
        return bbox_cds
    except Exception as e:
        raise Exception(f"Falha crítica na geometria: {e}")
```

Fonte: Elaborado pelo Autor (2026).

Com as coordenadas definidas, o módulo *CdsExtract* executa a extração dos dados de reanálise. A comunicação com o servidor do ECMWF é realizada de forma programática, solicitando as variáveis de Temperatura do Ar a 2 metros e Precipitação Total. O Código 3 detalha a estrutura da requisição, por exemplo, a utilização parâmetro *product_type* apontando para o modelo *monthly_averaged_reanalysis* na linha 3, modelo que pré-processa os dados e disponibiliza as médias identificadas na linha seguinte, a parametrização correta reduz significativamente o volume de tráfego de rede e armazenamento local.

Código 3 — Requisição Assíncrona de Dados Climáticos.

```
# Módulo cds_extract.py – Configuração do Payload para a API do Climate Data Store
request_params = {
  'product_type': ['monthly_averaged_reanalysis'],
  'variable': ['2m_temperature', 'total_precipitation'],
  'year': self.years_list,   # Lista dinâmica de anos (2004-2024)
  'month': self.months_list, # Apenas meses da safra (filtragem temporal)
  'time': ['00:00'],
  'area': self.bbox_goias,   # Coordenadas calculadas no Código 2
  'data_format': 'grib',
  'download_format': 'unarchived'
}

# Execução da extração via Earthkit
data_set = ek.from_source('cds', 'reanalysis-era5-land-monthly-means', request_params)
```

Fonte: Elaborado pelo Autor (2026).

Simultaneamente, o módulo *PamExtract* é responsável pela obtenção dos dados de produtividade da soja (Tabela 5457 do PAM). A complexidade desta etapa envolve a construção dinâmica da URL de consulta para a API PAM/SIDRA, garantindo que a série histórica completa seja recuperada em uma única chamada.

O Código 4 demonstra a montagem da consulta, onde parâmetros como o código da variável *Rendimento médio da produção (Quilogramas por Hectare)* (/v/112), o período (/p/last 20) e a unidade territorial (/n6/in n3 52) são injetados programaticamente entre a linha 2 e linha 7, esses valores foram obtidos através da documentação de referência do próprio SIDRA.

Código 4 — Construção da Consulta à API PAM.

```
# Módulo pam_extract.py – Extrator dos dados da API PAM.
url = (
f'{cfg.URL_PAM}/t/{table}' # Tabela 5457
f'/p/last {qt_years}' # Série histórica (20 anos)
f'/v/{variables}' # Variáveis: Rendimento médio (112)
f'/n6/in n3 {uf_country_code}' # Nível: Municípios de Goiás (52)
f'/c782/{classification_id}' # Filtro: Cultura Soja - 40124
f'/h/n?formato=json'
)

# Execução da requisição HTTP e conversão para DataFrame
response = requests.get(url)
data = response.json()
return pd.DataFrame(data=data[1:], columns=data[0])
```

Fonte: Elaborado pelo Autor (2026).

4.3 PRÉ-PROCESSAMENTO, LIMPEZA E TRANSFORMAÇÃO

Após a extração, os dados brutos apresentavam inconsistências de formato e escala que inviabilizavam a aplicação direta de algoritmos de aprendizado de máquina. Nesta fase, o *pipeline* executa rotinas de harmonização espacial[8] para os dados climáticos e limpeza estatística para os dados produtivos.

4.3.1 Harmonização Espacial e Estatística Zonal

O objetivo técnico desta etapa foi compatibilizar os formatos dos dados. Enquanto o clima (ERA5-Land) é fornecido como uma grade de imagens de satélite (matriz de pixels), a produção agrícola (PAM) é uma tabela com um valor único por município. Para unir essas duas fontes, o sistema precisou transformar o código do município em sua respectiva área geográfica. Utilizando a biblioteca *GeoPandas*[9], o algoritmo carrega o mapa digital de Goiás e usa o contorno (coordenadas) de cada cidade para 'recortar' os dados climáticos correspondentes.

[8] **Harmonização Espacial** é a aplicação de um processo de padronização das informações geográficas para que dados coletados possam ser utilizados em diferentes conjuntos e contextos.
[9] **GeoPandas:** Biblioteca de código em Python que facilita o trabalho com dados geoespaciais, documentação disponível em https://geopandas.org/en/stable.

Para resolver este problema, foi desenvolvido o módulo que implementa a técnica de Estatística Zonal. O algoritmo percorre sobre a geometria de cada um dos 246 municípios goianos e realiza um “recorte” (*clip*) digital sobre o grid climático. A partir dos pixels interceptados pela máscara do município, calcula-se a média espacial das variáveis.

Um ponto de atenção fundamental na implementação foi a conversão temporal da precipitação. O ERA5 fornece a chuva em taxa diária (metros/dia). Para obter o volume acumulado mensal correto, foi necessário implementar uma rotina que identifica quantos dias possui cada mês processado (28, 30 ou 31), conforme demonstrado no Código 5.

Código 5 — Estatística Zonal e Conversão de Taxas.

```
# Módulo cds_transform.py – Transformação dos dados climáticos.
for index, row in tqdm(gdf.iterrows(), total=len(gdf)):
    try:
    # 1. Recorte Espacial (Clip): Isola os pixels dentro do município
    recorte = ds.rio.clip([row['geometry']], ds.rio.crs, drop=True,
all_touched=True)
    # 2. Redução Espacial: Calcula a média de todos os pixels da
cidade
    medias = recorte.mean(dim=[“latitude”, “longitude”])

    # Extração dos vetores de valores
    times = medias.time.values
    v_temp = medias['t2m'].values
    v_chuva = medias['tp'].values

    # 3. Loop Temporal para Conversão Física
    for i, t in enumerate(times):
        # Identificação da duração do mês
        ts = pd.to_datetime(t)
        dias_no_mes = ts.days_in_month # Ex: 28, 30 ou 31

        # Temperatura: Kelvin -> Celsius
        temp_celsius = float(v_temp[i]) - 273.15

        # Chuva: Taxa diária (m) -> Acumulado Mensal (mm)
        # Fórmula: Valor * 1000 (mm) * Dias do Mês
        precipitacao_mm_mes = float(v_chuva[i]) * 1000 * dias_no_mes
        # Retorno ao DataFrame
    except Exception:
       continue
```

Fonte: Elaborado pelo Autor (2026).

Nessa estrutura, existem 2 iterações principais, na primeira iteração (linha 1 à linha 14) é realizada a identificação dos dados climáticos de acordo com as coordenadas de municípios obtidas anteriormente e armazenadas em variáveis. Em seguida, a partir da linha 17 é realizada a conversão dos valores, detalhada na linha 23, onde é realizada a conversão da temperatura média de Kelvin (métrica recebida do CDS) para Celsius (padrão brasileiro). Na linha 27 é realizada a conversão dos valores de chuva de origem dia para transformação mensal (de acordo com a quantidade de dias do mês analisado), em seguida converte-se o valor pluviométrico de metros para milímetros, finalizando a adequação para métricas padronizadas para análise posterior.

4.3.2 Limpeza e Consistência Biológica (PAM)

Os dados de produtividade processados pelo algoritmo do Código 6 passaram por um rigoroso filtro de qualidade. Identificou-se que a base bruta do IBGE continha registros nulos (representados por “-”) ou valores exorbitantes decorrentes de erros de digitação ou coleta, removidos na linha 3.

Foi implementado um filtro de Consistência Biológica na linha 6, definindo um teto máximo de produtividade de 6.000 kg/ha (100 sacas por hectare). Valores acima deste limiar, foram considerados improváveis para a média municipal de soja em condições de sequeiro[10] na região goiana, fugindo da média e máxima da região, esses foram tratados como ruído e removidos para não enviesar os centroides dos clusters.

[10] **Sequeiro:** Cultivo que depende somente da água da chuva para desenvolvimento, sem que haja uso de irrigação artificial.

Código 6 — Limpeza de Dados e Remoção de Anomalias.

```
1   # Módulo pam_transform.py – Conversão de tipos e tratamento de nulos
2   self.df['valor'] = pd.to_numeric(self.df['valor'], errors='coerce')
3   self.df.dropna(subset=['valor'], inplace=True)
4   # Filtro de Consistência Biológica (Teto produtivo da soja)
5   qtd_erro = len(self.df[self.df['valor'] > 6000])
6   if qtd_erro > 0:
7      print(f"[CORREÇÃO] Removendo {qtd_erro} registros inconsistentes
8   (> 6.000 kg/ha).")
9      self.df = self.df[self.df['valor'] <= 6000]
10
```

Fonte: Elaborado pelo Autor (2026).

4.3.3 Consolidação da Base de Dados e Carga

A etapa final do pré-processamento teve como objetivo unificar as granularidades temporais distintas. Como os dados produtivos são anuais e os climáticos são mensais, foi necessário realizar uma Agregação Temporal das variáveis meteorológicas.

Aplicou-se uma lógica de "Ano Safra", onde os registros climáticos de outubro, novembro e dezembro do ano *t* foram contabilizados para a safra do ano *t+1*, alinhando-se com a metodologia do IBGE/PAM.

O algoritmo agrupa os registros mensais por município e ano, através do cálculo da média aritmética para a temperatura e o somatório para a precipitação, que representa o acumulado do período de safra. Em seguida, foi executado o cruzamento (*merge*) entre as bases utilizando a chave composta (*id_municipio, ano*).

O Código 7 apresenta a lógica de correção de ano safra, como visto a partir da linha 11, onde todo mês a partir de Outubro à Dezembro é considerado mês da safra do ano t+1. Ou seja, em uma safra de soja 2022, os meses de plantio são, respectivamente, outubro, novembro e dezembro de 2021, com a coleta seguindo a lógica dos 4 meses iniciais do ano t. Em seguida, a partir da linha 33, conclui-se a junção e a etapa final (*load*) do ETL para os dados de produtividade e clima, com o padrão Munícipio (id munícipio da malha municipal IBGE) + Ano Safra, ou seja, cada linha exportada é identificada por exemplo em: Rio Verde 2022, Morrinhos 2018 etc. Todo conteúdo processado é exportado em um conjunto de dados unificado em formato CSV, prontificado para etapa de mineração.

Código 7 — Agregação Temporal e Cruzamento de Dados.

```
# Módulo etl_pipeline.py – Adequações finais para geração do dataset.
# 1. Definição do Ano Safra (Correção Temporal)
# Converte a coluna de data para extrair mês e ano civil
df_mensal['data'] = pd.to_datetime(df_mensal['data'])
df_mensal['mes'] = df_mensal['data'].dt.month
df_mensal['ano_civil'] = df_mensal['data'].dt.year

# Lógica de Deslocamento:
# Meses >= 10 (Out, Nov, Dez) pertencem à safra do ano seguinte (+1)
# Meses < 10 (Jan a Set) pertencem à safra do próprio ano civil
df_mensal['ano_safra'] = df_mensal.apply(lambda x: x['ano_civil'] + 1
if x['mes'] >= 10 else x['ano_civil'], axis=1)

# 2. Agregação Temporal baseada no Ano Safra
# Temperatura -> Média do ciclo | Chuva -> Soma acumulada do ciclo
df_clima_anual = df_mensal.groupby(['codigo_ibge', 'ano_safra'])
.agg({
      'temperatura_media': 'mean',
     'precipitacao_total': 'sum'
}).reset_index()

# Renomeia 'ano_safra' para 'ano' para permitir o cruzamento com o
PAM
df_clima_anual.rename(columns={
'codigo_ibge': 'id_municipio',
'ano_safra': 'ano',
'temperatura_media': 'temp_media',
'precipitacao_total': 'chuva_total'
}, inplace=True)

# 3. Cruzamento (Merge): Une Produtividade (PAM) + Clima (ERA5)
# Tipo 'inner': Garante que apenas anos com dados completos sejam
mantidos
df_final = pd.merge(
df_pam_clean,
df_clima_anual,
on=['id_municipio', 'ano'],
how='inner'
)

# 4. Persistência
loader = Climate_Pam_Loader(self.path_final)
loader.save_data(df_final)
```

Fonte: Elaborado pelo Autor (2026).

4.4 MINERAÇÃO E CLUSTERIZAÇÃO DE DADOS

Com a base de dados consolidada e limpa, iniciou-se a fase de Mineração de Dados. O objetivo desta etapa foi aplicar algoritmos de aprendizado de máquina não supervisionado para identificar grupos de municípios com comportamentos agrícolas e climáticos similares. Todo o processo de mineração foi implementado utilizando a biblioteca Scikit-Learn.

4.4.1 Padronização de Escalas

Um pré-requisito fundamental para algoritmos baseados em distância Euclidiana, como o K-Means, é a padronização das variáveis. O conjunto de dados apresenta grandezas com magnitudes e unidades díspares: a produtividade oscila em milhares (kg/ha), a precipitação em centenas (mm) e a temperatura em dezenas (°C).

Sem a normalização, a variável de maior magnitude (produtividade) dominaria o cálculo da distância, fazendo com que o algoritmo ignorasse as variações climáticas. Como os *outliers* extremos já foram tratados na etapa de limpeza, optou-se pelo método *StandardScaler* (Z-Score), implementado através da biblioteca Scikit-Learn (PEDREGOSA et al., 2011). Este método centraliza a distribuição na média 0 com desvio padrão 1, preservando a variância dos dados, conforme a equação:

$$z = \frac{x - \mu}{\sigma}$$

onde:

- $\boldsymbol{x}$ é o valor original.
- $\boldsymbol{\mu}$ é a média da coluna.
- $\boldsymbol{\sigma}$ é o desvio padrão.

Como exemplo de utilização deste método, considere duas cidades hipotéticas com produtividades distintas em um cenário onde a média estadual (μ) é 3.500 kg/ha e o desvio padrão (σ) é 500 kg/ha:

1. Cidade A (3.000 kg/ha):

$$z = \frac{3000 - 3500}{500} = -1{,}0$$

O algoritmo entende que esta cidade está 1 desvio padrão abaixo da média.

2. Cidade B (4.500 kg/ha):

$$z = \frac{4500 - 3500}{500} = +2{,}0$$

O algoritmo entende que esta cidade está 2 desvios padrão acima da média.

Conforme a Regra Empírica da estatística clássica (BUSSAB; MORETTIN, 2017), em distribuições que se aproximam da normalidade, aproximadamente 99,7% das observações concentram-se a até três desvios padrões da média. Desta forma, todas as variáveis (Produtividade, Chuva e Temperatura Média) passam a operar predominantemente no mesmo intervalo numérico (entre -3 e +3). O Código 8 demonstra a implementação desta etapa na linha 10 com auxílio da biblioteca *sklearn,* necessário apenas instanciar o objeto *scaler* e em seguida efetuar a transformação.

Código 8 — Seleção de Variáveis e Padronização.

```
1   # Módulo analysis_cluster.py – Aplicação do StandardScaler
2   from sklearn.preprocessing import StandardScaler
3
4   # Seleção das features para o modelo
5   features = ['produtividade', 'chuva_total', 'temp_media']
6   X = df[features].copy()
7
8   # Padronização (Z-Score)
9   # Transforma os dados para Média=0 e Desvio Padrão=1
10  scaler = StandardScaler()
11  X_scaled = scaler.fit_transform(X)
12
    print("Dados padronizados. Média = 0 e Desvio Padrão = 1.")
```

Fonte: Elaborado pelo Autor (2026).

4.4.2 Determinação do Número de Clusters (Hiperparâmetros)

Um dos maiores desafios em algoritmos não supervisionados é a definição do hiperparâmetro *K* (número de grupos), uma vez que não existe uma “resposta correta” prévia. Para evitar escolha arbitrária, adotou-se uma estratégia de validação cruzada baseada em quatro métodos complementares: Inércia (Método do Cotovelo), Coeficiente de Silhueta, Agrupamento Hierárquico (Dendrograma) e Índice Caliński-Harabasz.

O Código 9 inicia a implementação das métricas de validação. Para a implementação do dendrograma, foi utilizado o auxílio dos métodos *dendrogram* e *linkage*, implementados na biblioteca *scipy* (VIRTANEN et al., 2020). A partir da linha 3 tem-se a definição das variáveis do cálculo do método de cotovelo, coeficiente da silhueta e a geração do agrupamento hierárquico com dendrograma com método de Ward (método da variância mínima). Observa-se a plotagem de cada gráfico de validação pós definição de cada um de seus valores, respectivamente nas linhas 18, 25 e 35.

Código 9 — Métricas de Validação e Dendrograma.

```
# Módulo analysis_cluster.py – Métricas de validação e dendrograma
from scipy.cluster.hierarchy import dendrogram, linkage
# 1. Cálculo da Inércia (Elbow) e Silhueta
inertia = []
sil_scores = []
K_range = range(2, 11)

for k in K_range:
    km = KMeans(n_clusters=k, random_state=42, n_init=10)
    km.fit(X_scaled)
    inertia.append(km.inertia_)
    sil_scores.append(silhouette_score(X_scaled, km.labels_))

# Gráfico 1: Cotovelo
plt.plot(K_range, inertia, 'bo-', markersize=8)
plt.title('Método do Cotovelo (Elbow Method)')
plt.xlabel('Número de Clusters (k)')
plt.ylabel('Inércia (Soma dos Quadrados)')
plt.savefig(output_dir / "validacao_1_cotovelo.png", dpi=300)

# Gráfico 2: Silhueta
plt.plot(K_range, sil_scores, 'rx-', markersize=8)
plt.title('Análise de Silhueta')
plt.xlabel('Número de Clusters (k)')
plt.ylabel('Silhueta Média')
plt.savefig(output_dir / "validacao_2_silhueta.png", dpi=300)

# 4. Geração do Dendrograma (Validação Hierárquica)
# Utiliza o método 'ward' para minimizar a variância
linked = linkage(X_scaled, method='ward')

# Gráfico 3: Dendrograma
plt.figure(figsize=(10, 7))
dendrogram(linked, truncate_mode='lastp', p=30)
plt.title("Dendrograma Hierárquico para Validação de K")
plt.savefig("src/results/dendrograma_validacao.png")
```

Fonte: Elaborado pelo Autor (2026).

4.4.3 Execução do Algoritmo K-Means

A definição do número ideal de grupos (*K=5*) foi determinada empiricamente por meio da execução prévia das métricas de validação implementadas nos códigos anteriores, observando os resultados obtidos pelo método do cotovelo e do método

da silhueta. Cabe ressaltar que a análise visual dos gráficos e a fundamentação estatística que justificam a escolha desse hiperparâmetro são detalhadas no capítulo seguinte (Resultados e Discussão). Inicialmente, o algoritmo K-Means atribui rótulos indexados em zero (0 a 4), para facilitar a interpretação agronômica e a visualização nos relatórios finais, realizou-se uma transformação simples nos rótulos, convertendo o intervalo para 1 a 5. O Código 10 apresenta a execução final e essa etapa de pós-processamento antes da persistência em arquivo (linha 27). Ademais, a estrutura final de processamento foi aproveitada para gerar métricas resumidas em um arquivo secundário (linha 21) e realizar a extração do Índice Caliński-Harabasz (linha 19), já considerando *K=5* para posterior análise.

Código 10 — Execução do Modelo e Ajuste de Rótulos.

```
1   # Módulo analysis_cluster.py – Execução da clusterização
2   # Definição do Hiperparâmetro final
3   k_final = 5
4
5   # Instanciação do Modelo
6   kmeans_final = KMeans(n_clusters=k_final, random_state=42, n_init=10)
7
8   # Predição dos grupos (Gera array [0, 1, 4, 2...])
9   labels_raw = kmeans_final.fit_predict(X_scaled)
10
11  # Ajuste de Legibilidade: Transforma 0-4 para 1-5
12  df['cluster'] = labels_raw + 1
13
14  # Persistência dos Resultados
15  # O caminho 'output_path' refere-se ao arquivo
16  “resultado_clusters.csv”
17  output_path = “src/results/resultado_clusters.csv”
18  # Métricas Finais
19  sil_final = silhouette_score(X_scaled, labels)
20  ch_final = calinski_harabasz_score(X_scaled, labels)
21
22  with open(output_dir / “metricas_finais.txt”, “w”) as f:
23       f.write(f”--- RESULTADOS FINAIS ---\n”)
24       f.write(f”K escolhido: {k_final}\n”)
25       f.write(f”Coeficiente de Silhueta: {sil_final:.4f}\n”)
26       f.write(f”Índice Calinski-Harabasz: {ch_final:.4f}\n”)
27
28  df.to_csv(output_path, index=False, sep=';', encoding='utf-8-sig')
29  print(f”Clusterização concluída. Dados salvos com IDs 1-{k_final}.”)
```

Fonte: Elaborado pelo Autor (2026).

4.4.4 Geração de Artefatos Visuais e Validação Espacial

Além da persistência dos dados tabulares, foi implementada uma função para gerar, em tempo de execução, visualizações que permitam validar a coerência física dos grupos formados. São elas e seus objetivos no contexto:

- **Gráficos de Validações:** Gerado nos passos anteriores junto da definição das variáveis de validação.
- **Gráficos de Dispersão (*Scatter Plots*):** Cruzar a Produtividade contra variáveis climáticas para visualizar as fronteiras de decisão.
- **Diagramas de Caixa (*Boxplots*):** Fundamental para analisar a variância interna, a simetria dos dados e a existência de outliers dentro de cada agrupamento, validando a compactação dos clusters.

O Código 11 demonstra a implementação das funções de visualização estatística. Definida em 2 principais funções, na linha 2 à linha 10 é implementado a função para geração de gráficos de dispersão, em seguida entre a linha 13 a linha 25 foi implementado a função para geração de gráficos de distribuição e variância, ou diagramas de caixa.

Código 11 — Geração de Scatter Plots e Boxplots.

```
# Módulo analysis_cluster.py – Funções para geração de gráficos.
# 1. Função para Scatter Plot (Dispersão)
def plot_scatter(df, x_col, y_col, output_path):
    plt.figure(figsize=(10, 6))
    sns.scatterplot(
        data=df, x=x_col, y=y_col, hue='cluster',
        palette='viridis', s=60, alpha=0.7
    )
    plt.title(f'Dispersão: {x_col} vs {y_col}')
    plt.savefig(output_path)
    plt.close()

# 2. Função para Boxplot (Distribuição e Variância)
def plot_boxplot(df, y_col, output_path):
    plt.figure(figsize=(12, 6))

    # Plota a distribuição da variável para cada Cluster (1-5)
    sns.boxplot(
        data=df, x='cluster', y=y_col,
        palette='viridis', showfliers=True
    )

    plt.title(f'Distribuição de {y_col.title()} por Cluster')
    plt.grid(True, axis='y', linestyle='--', alpha=0.7)
    plt.savefig(output_path)
    plt.close()

# Chamada das funções
plot_boxplot(df, 'produtividade', output_dir /
“boxplot_produtividade.png”)
```

Fonte: Elaborado pelo Autor (2026).

4.5 INTERPRETAÇÃO E CONHECIMENTO

Finalizada a mineração dos dados e geração de validações, inicia-se a etapa final do KDD, a interpretação e conhecimento. Como a base de dados abrange uma série histórica de 20 anos, um mesmo município pode ter sido classificado em clusters diferentes em safras distintas (devido a variações climáticas anuais).

Para espacializar os resultados e permitir identificação de predominância nos padrões regionais, foi implementada uma função em um algoritmo de geoprocessamento (com auxílio da biblioteca *GeoPandas*). No Código 12, realiza-se

o cálculo para definição do cluster predominante (moda), isto é, em qual cluster o munícipio mais apareceu para cada munícipio, desenvolvido na linha 6, em seguida a união dos dados organizados e classificados (*dataframe*) e da malha digital coletada anteriormente (*shapefile*) na linha 11. A partir da linha 23 foi realizada a plotagem gráfica de um mapa temático (mapa coroplético) adicionando também os polos regionais de Goiás identificados no próprio arquivo de malha municipal.

Código 12 — Geração do Mapa Geoespacial.

```
1   # Módulo analysis_cluster.py – Geração do Mapa Coroplético
2   # Carrega a malha geográfica (Shapefile do IBGE)
3   gdf = gpd.read_file(shapefile_path)
4
5   # Agrupamento para definir o cluster predominante por município
6   (Moda)
7   moda_clusters = df.groupby('id_municipio')['cluster'].agg(
8       lambda x: pd.Series.mode(x)[0]
9   ).reset_index()
10
11  # Merge: Une os dados agronômicos com a geometria do mapa
12  gdf_final = gdf.merge(moda_clusters, left_on='CD_MUN',
13  right_on='id_municipio')
14
15  # Geração do Mapa Coroplético
16  gdf_final.plot(
17      column='cluster', cmap=map_cmap, legend=True, edgecolor='black',
18  linewidth=0.3
19  )
20  plt.title(f'Distribuição Espacial dos Clusters de Soja em GO
21  (K={k_final}) e Polos Regionais')
22  plt.savefig(output_dir / "mapa_clusters_goias.png", dpi=300)
23  map_path = output_dir / "mapa_clusters_goias.png"
24     for cidade, (x, y) in cidades_polo.items():
25         # Desenha um ponto preto na localização da cidade
26         plt.plot(x, y, marker='o', color='black', markersize=3)
26
27         # Escreve o nome da região
28         plt.text(
29             x, y + 0.15, cidade, fontsize=7, fontweight='bold',
30             color='black', ha='center',
31             bbox=dict(facecolor='white', alpha=0.7, edgecolor='gray',
32  boxstyle='round,pad=0.2')
33     )
34     plt.savefig(map_path, dpi=300, bbox_inches='tight')
```

Fonte: Elaborado pelo Autor (2026).

Enquanto o módulo anterior foi responsável pela modelagem matemática, validação estatística e geração dos principais gráficos, a interpretação detalhada exigiu uma visão granular ao nível de município. Para isso, foi desenvolvido um módulo complementar. Este *script* atua na camada de inteligência de negócio, processando os dados classificados para gerar dois tipos de artefatos específicos que não são cobertos pela visão estatística global, são eles:

1. **Rankeamento Individual Combinado:** Identificação e plotagem dos 5 municípios mais representativos (maiores produtividades) dentro de cada cluster combinados em uma imagem.
2. **Síntese de Centroides:** Cálculo exato das médias de produtividade, chuva e temperatura para compor a tabela final de discussão.

O Código 13 ilustra a lógica de filtragem utilizada para gerar os gráficos de barras individuais (“Top 5”) de cada cluster em uma única imagem, permitindo a análise qualitativa dos resultados. A escolha dos 5 maiores produtores de cada cluster é realizada na linha 6 através de uma função de ordenação e agrupamento, posteriormente, na linha 11 em diante os resultados são inseridos no gráfico iterativamente para cada um dos 5 clusters.

Código 13 — Geração dos Gráficos de Perfil Municipal.

```
# Módulo analysis_resume.py – Geração de gráficos complementares.
for i, cluster in enumerate(clusters):
    ax = axes[i]
    cor_atual = hex_colors[i % len(hex_colors)]
    grupo = df[df['cluster'] == cluster]
    # Seleciona os 5 maiores produtores para detalhamento
    top5 = grupo.sort_values(by='produtividade',
ascending=False).head(5).copy()
    top5['rotulo'] = top5['nome_municipio'] + " (" +
top5['ano'].astype(str) + ")"

    sns.barplot(data=top5, x='produtividade', y='rotulo', ax=ax,
color=cor_atual, edgecolor='black', alpha=0.9)

    ax.set_title(f"Cluster {cluster}", fontsize=14, weight='bold',
color=cor_atual)
    ax.set_xlabel("Kg/ha")
    ax.set_ylabel("")
    sns.despine(left=True, bottom=True, ax=ax)

    # Valores nas barras
    for index, row in enumerate(top5.itertuples()):
        ax.text(row.produtividade - (row.produtividade * 0.05), index,
f'{int(row.produtividade)}',
        va='center', ha='right', color='white', weight='bold',
fontsize=10)
    # Salva a imagem para o capítulo de Resultados
    plt.savefig(base_dir / "grafico_top5_resumo_geral.png", dpi=300,
bbox_inches='tight')
```

Fonte: Elaborado pelo Autor (2026).

Com a execução deste módulo, a engenharia do trabalho foi finalizada, fornecendo tanto a visão macro (validações, dispersão, *Boxplot* e mapa) quanto a visão micro (Rankings municipais).

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Neste capítulo, são apresentados os resultados obtidos a partir da execução do *pipeline* de processamento de dados e da aplicação do algoritmo K-Means. A análise é estruturada em três etapas lógicas: primeiramente, discute-se a validação técnica para a escolha do número de grupos (*K*) e a verificação da separabilidade estatística; em seguida, avalia-se a distribuição espacial (mapas) desses grupos no estado de Goiás; e, por fim, realiza-se a síntese dos padrões encontrados.

A aplicação do algoritmo K-Means resultou na segmentação dos 4.015 registros válidos (safra 2004-2024) em cinco grupos distintos. A Tabela 1 detalha os centroides (médias) de cada cluster e a volumetria de dados, permitindo a identificação dos perfis predominantes.

Tabela 1 — Centroides Finais e Distribuição dos Clusters (K=5).

Cluster	N° de Registros	% do Total	Produtividade (kg/ha)	Chuva Safra (mm)	Temp. Média (°C)
1	1.304	32,5%	2.878,78	1.338,53	23,54
2	552	13,7%	2.237,84	1.169,89	24,79
3	744	18,5%	3.491,86	1.143,48	24,35
4	904	22,5%	3.022,67	1.301,21	25,81
5	511	12,7%	3.479,26	593,67	25,25
Total	**4.015**	**100**	**—**	**—**	**—**

Fonte: Elaborado pelo Autor com base nos dados dos resultados da Clusterização (2026).

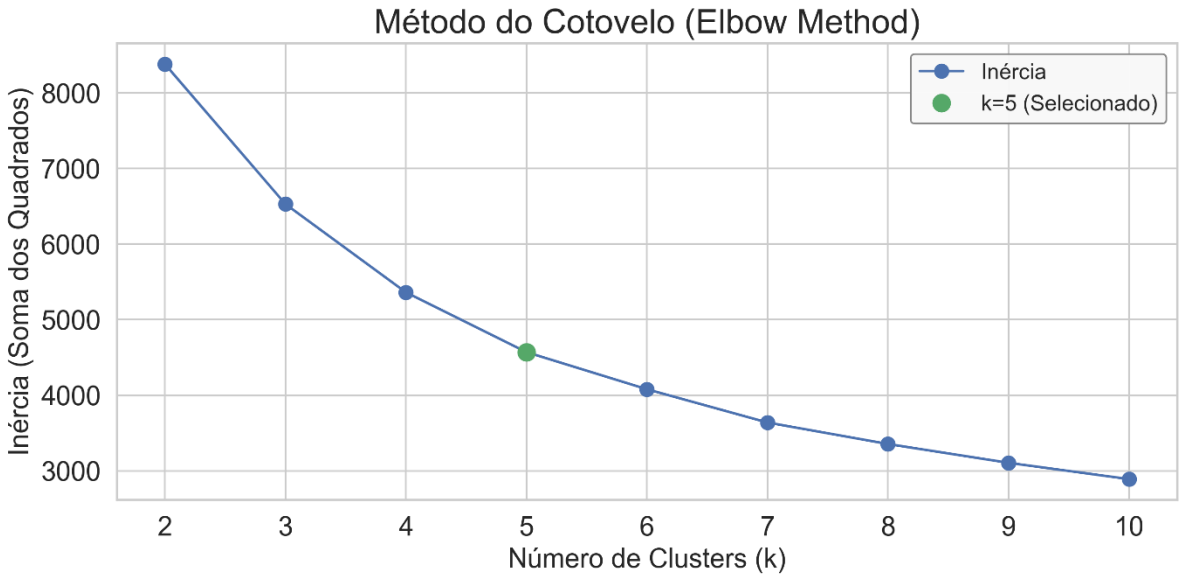
A análise preliminar dos centroides revela uma assimetria rígida no desempenho agronômico estadual. É possível observar, por exemplo, contrastes expressivos nas médias de produtividade, que variam desde um teto de aproximadamente 3.491kg/ha (Cluster 3) até pisos próximos de 2.237kg/ha (Cluster 2). Destaca-se também a importância da observação na variação climática, como no caso do Cluster 5, que obteve alta produtividade mesmo com um volume médio de precipitação baixa (cerca de 593mm). Essa disposição inicial dos dados corrobora a premissa de que a soja em Goiás não obedece a um único comportamento.

5.1 VALIDAÇÃO DO MODELO E DEFINIÇÃO DE K

A etapa inicial da clusterização consiste na determinação do hiperparâmetro *K* (número de grupos). Diferente da classificação supervisionada, onde as classes são pré-definidas, na aprendizagem não supervisionada é necessário encontrar um equilíbrio entre a compactação dos dados (homogeneidade interna) e a separação entre os grupos. Para fundamentar essa decisão com rigor científico, utilizou-se uma abordagem combinada de quatro métricas: a Inércia (Método do Cotovelo), o Coeficiente de Silhueta, a Análise Hierárquica (Dendrograma) e a Análise de Variância (*Boxplot*).

Primeiramente, a Inércia calcula a soma dos quadrados das distâncias de cada ponto até o centro do seu cluster. O objetivo é minimizar esse valor, evitando o *overfitting*, a Figura 4 apresenta a curva de decaimento da inércia. Através da análise visual, observa-se que a inércia decresce acentuadamente até K=3. Entre K=4 e K=5, a curva estabiliza, formando o “cotovelo”. A partir deste ponto, o ganho na redução da inércia torna-se marginal, concluindo que K=5 se comporta como a quantidade ideal de clusters.

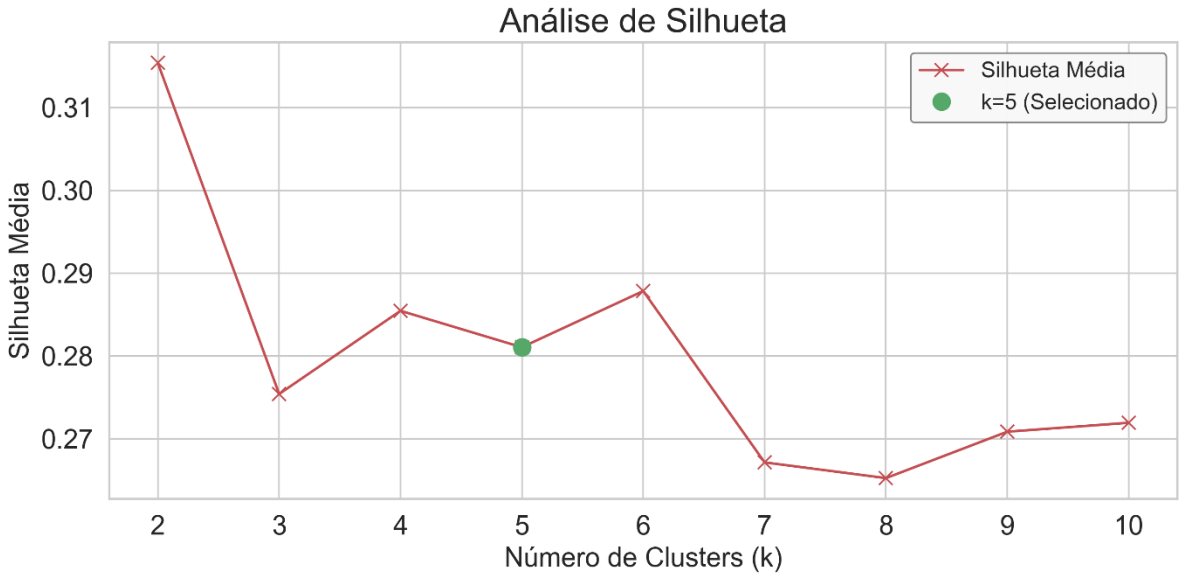
Figura 4 — Análise da Inércia para diferentes valores de K (Método do Cotovelo).



Fonte: Elaborado pelo Autor (2026).

Em seguida, para a análise do Coeficiente da Silhueta, métrica que avalia a qualidade da separação (coesão vs. separação), foi identificado na Figura 5 que para *K*=5 o coeficiente de Silhueta médio obtido foi de 0,2811. Segundo Kaufman e Rousseeuw (1990), valores do coeficiente de silhueta entre 0,26 e 0,50 caracterizam uma estrutura de agrupamento fraca, na qual os grupos são identificáveis, porém apresentam limites menos definidos entre si. O valor obtido neste trabalho (0,2811) reflete as características dos dados analisados, compostos por variáveis climáticas e produtivas de natureza contínua. Em contextos geográficos e agroclimáticos, tais variáveis tendem a apresentar gradientes ambientais e produtivos, resultando em transições graduais entre perfis semelhantes, em vez de separações abruptas entre grupos. O coeficiente positivo indica que, em média, os municípios apresentam maior similaridade com os elementos de seu próprio agrupamento (intra-cluster) do que com os demais (inter-cluster), evidenciando a presença de padrões estruturais nos dados. Assim, o particionamento em cinco clusters representa uma organização consistente dos municípios em perfis agroclimáticos comparáveis.

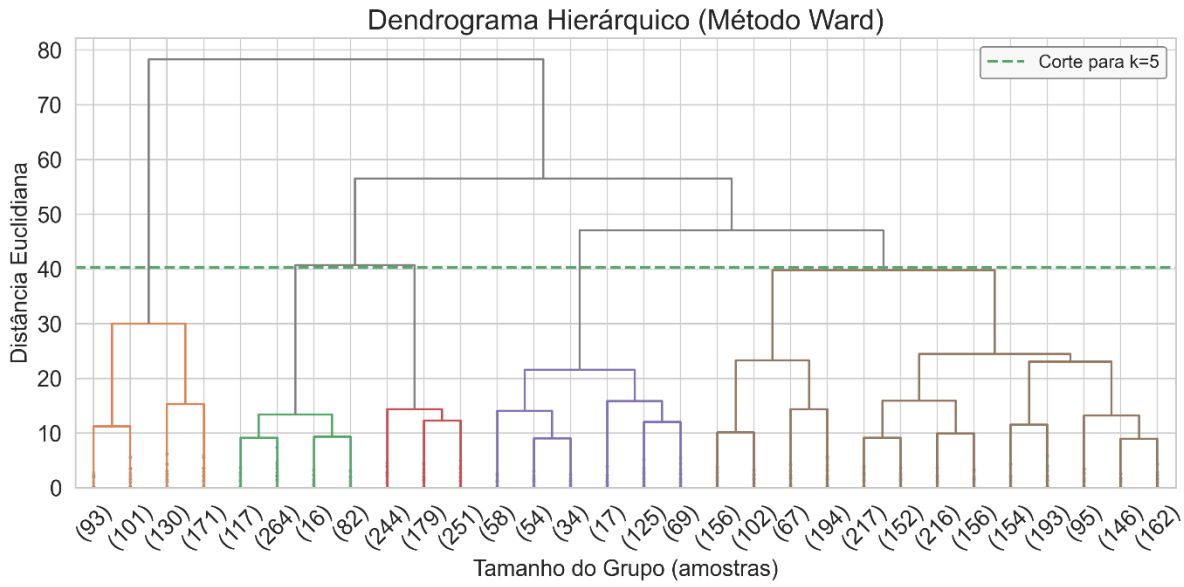
Figura 5 — Análise do Coeficiente de Silhueta.



Fonte: Elaborado pelo Autor (2026).

Para a validação hierárquica (dendrograma), aplicou-se o algoritmo de *Agglomerative Hierarchical Clustering* utilizando o método de Ward. O dendrograma resultante (Figura 6) permite visualizar a dissimilaridade entre os municípios com base na distância Euclidiana. A decisão de particionar os dados em *K=5* fica evidente ao observar as maiores hastes verticais da árvore, ao traçar uma linha de corte horizontal cruzando essa região de maior salto de distância (onde a variância intergrupos é maximizada), a árvore é seccionada em exatamente 5 ramificações principais. Esse comportamento estrutural demonstra que a separação em cinco agrupamentos não é arbitrária, mas sim uma divisão natural dos dados que corrobora a parametrização já apontada nos testes de K-Means.

Figura 6 — Dendrograma de Agrupamento Hierárquico.



Fonte: Elaborado pelo Autor (2026).

Considerando o arquivo de métricas gerado, descreve-se a Tabela 2 para representar os resultados obtidos em resumo.

Tabela 2 — Métricas de Avaliação de Desempenho do Modelo (K-Means).

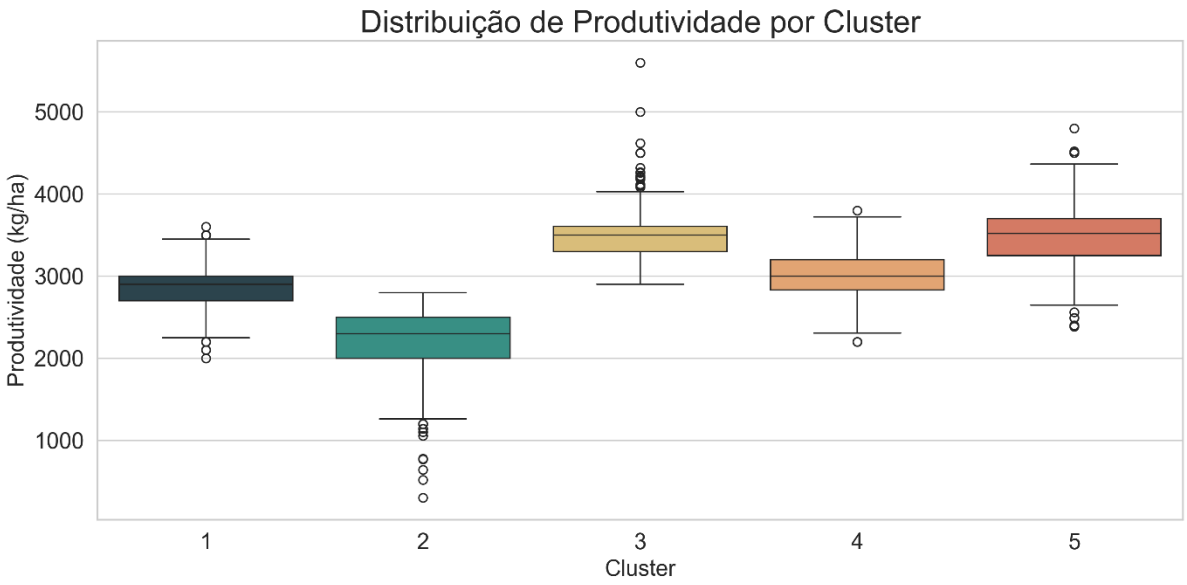
Métrica	**Valor Obtido**	**Interpretação**
Número de Clusters (K)	5	Definido pelo Método do Cotovelo (Elbow Method).
Coeficiente de Silhueta	0,2811	Indica uma separação razoável dos grupos, típica de dados climáticos complexos onde as fronteiras não são lineares.
Índice Caliński-Harabasz	1.639,33	Valor elevado indica que os grupos são densos e bem definidos em relação à dispersão geral dos dados.

Fonte: Elaborado pelo Autor (2026).

Por fim, para validar a distinção estatística dos grupos, analisou-se a distribuição da variável alvo (Produtividade) através do diagrama de caixa (Boxplot). A Figura 7 revela que a segmentação não ocorreu apenas em uma escala linear, mas em patamares de desempenho:

1. **Patamar de Excelência:** Os Clusters 3 e 5 destacam-se no topo do gráfico, apresentando medianas quase idênticas e superiores a 3.400 kg/ha. Isso indica que o algoritmo identificou dois grupos distintos que atingem o teto produtivo do estado, embora, como visto nas análises anteriores, o façam sob condições climáticas opostas.
2. **Patamar Intermediário:** Os Clusters 4 e 1 ocupam a faixa central, com produtividades oscilando entre 2.800 e 3.000 kg/ha, representando zonas de transição ou de limitações pontuais (térmicas e hídricas).
3. **Patamar Inferior:** O Cluster 2 isola-se na base do gráfico, com a mediana mais baixa e maior dispersão negativa, confirmando sua vocação como grupo de baixa aptidão ou alto risco.

Figura 7 — Dispersão da Produtividade para cada Cluster.



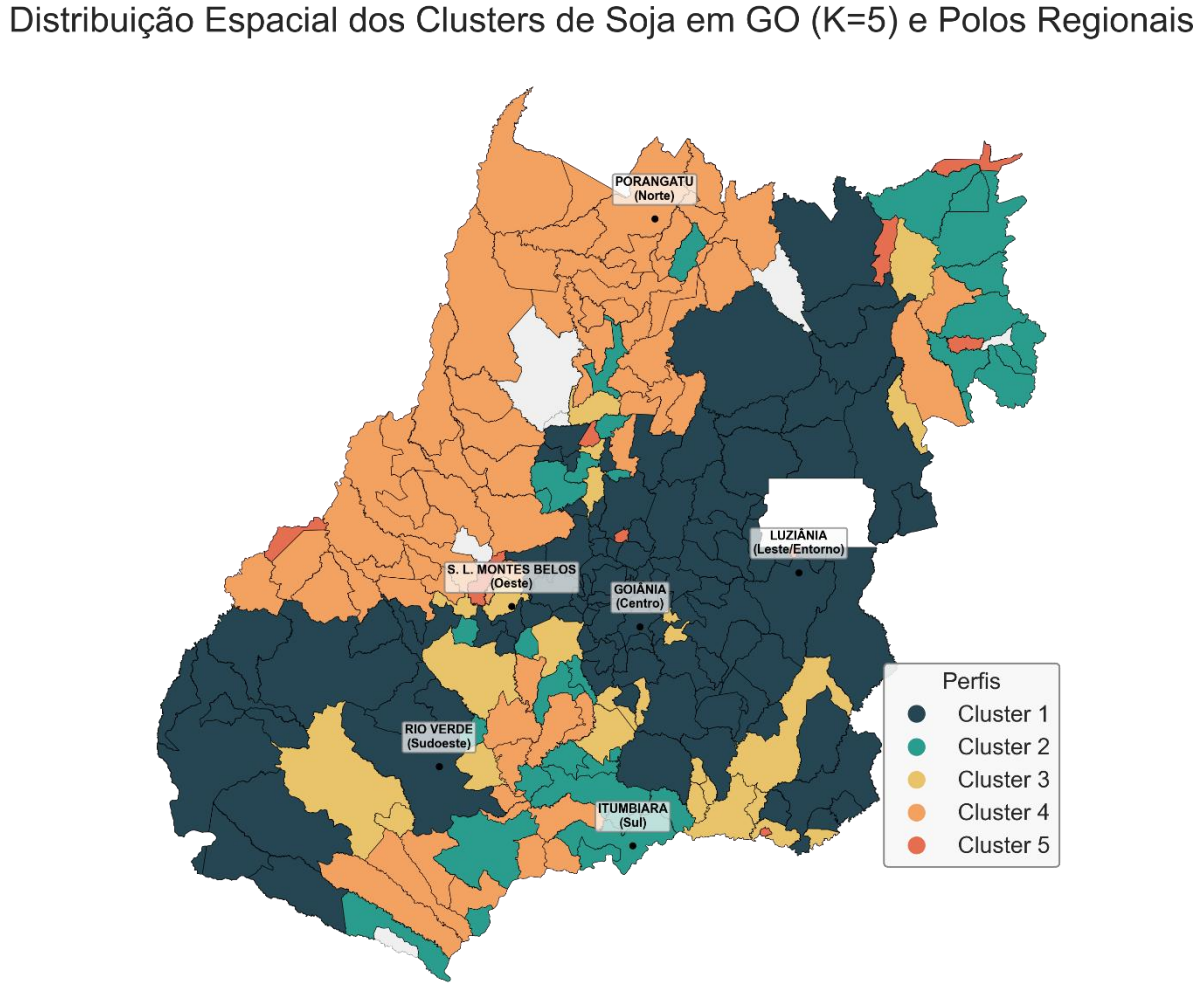
Fonte: Elaborado pelo Autor (2026).

5.2 DISTRIBUIÇÃO ESPACIAL E PADRÕES REGIONAIS

A Figura 8 apresenta a distribuição geográfica dos perfis predominantes identificados pelo algoritmo K-Means (*K*=5). Para contextualizar a análise, foram destacados os principais polos das Regiões Geográficas Intermediárias (IBGE, 2024), servindo como pontos de referência territorial, isto é, culminando no resultado final abaixo em um mapa coroplético.

Inicialmente, cabe esclarecer que as áreas não preenchidas (“espaços em branco”) no mapa correspondem a municípios que foram filtrados durante a Etapa 2 de Pré-Processamento por não apresentarem séries históricas completas ou consistentes no período Safra (Outubro a Abril) analisado (2004-2024). Essa exclusão é intencional e visa garantir que os padrões identificados sejam baseados apenas em dados robustos, evitando que ruídos estatísticos distorçam a caracterização dos perfis.

Figura 8 — Distribuição De Perfis de Domínio.



Fonte: Elaborado pelo Autor (2026).

A espacialização dos clusters revela dinâmicas regionais distintas, que podem ser interpretadas da seguinte forma:

1. Diferentemente dos demais grupos, que formam manchas espaciais mais contínuas, o **Cluster 5 (Rosa-alaranjado)** apresenta uma distribuição mais pontual e dispersa. Com 511 registros, esse grupo sugere a presença de sistemas produtivos associados ao uso de tecnologias de irrigação. Sua ocorrência não segue estritamente a climatologia natural, mas está relacionada ao nível de tecnificação agrícola. **Análise:** Ao relacionar uma produtividade elevada (**3.479 kg/ha**) com uma condição de baixa disponibilidade hídrica (**593 mm**, a menor do estudo), esse cluster evidencia municípios que apresentam menor dependência direta da

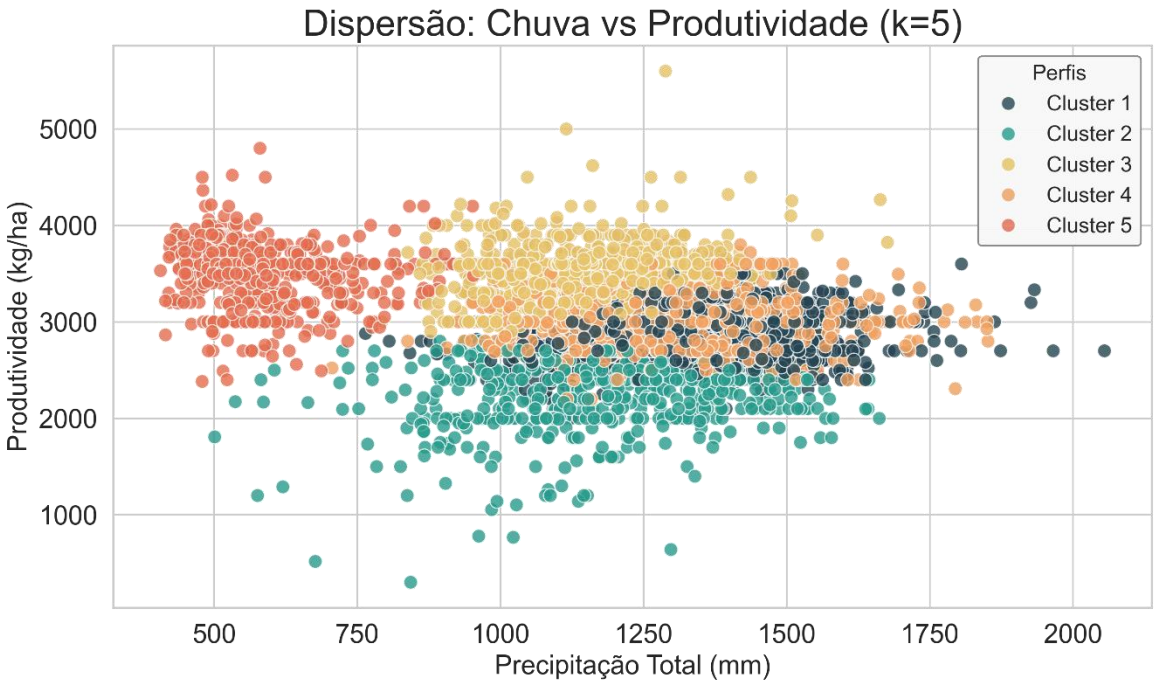
precipitação. Embora a literatura agronômica estabeleça que a cultura da soja necessite de aproximadamente 450 a 800 mm por ciclo para atingir seu potencial produtivo (FARIAS, 2019), esse valor refere-se à demanda hídrica da planta e não ao total de precipitação efetivamente aproveitado. No bioma Cerrado, em função das elevadas taxas de evapotranspiração e das perdas associadas ao escoamento superficial e à infiltração, nem toda a água precipitada permanece disponível no sistema solo-planta-atmosfera. Dessa forma, a precipitação observada pode não ser suficiente para suprir integralmente a demanda da cultura em condições de sequeiro. Esse comportamento indica que o **Cluster 5** pode estar associado tanto a anos com ocorrência de déficit hídrico quanto a regiões caracterizadas por maior nível de tecnificação agrícola, nas quais o uso de tecnologia (como a irrigação) contribui para a estabilidade produtiva.

2. O **Cluster 3 (Amarelo)** consolida-se como a representação da aptidão agrícola natural do estado de Goiás, com predominância na região **Sudoeste**, historicamente a principal área produtora. **Análise:** Apresentando a maior produtividade média do estudo (**3.491 kg/ha**) associada a um regime pluviométrico equilibrado (**1.143 mm**), esse grupo representa condições próximas ao ideal para o cultivo da soja em sistema de sequeiro. Nesses cenários, a combinação entre disponibilidade hídrica adequada e ausência de estresses climáticos significativos favorece a expressão do potencial produtivo da cultura. A literatura indica que a soja apresenta exigência hídrica de até 800 mm por ciclo (FARIAS, 2019), sendo que o atendimento dessa demanda depende não apenas do volume total de precipitação, mas também da sua distribuição ao longo do ciclo produtivo e das perdas naturais do sistema. Além do volume, o desempenho produtivo observado sugere uma distribuição pluviométrica adequada, com maior concentração de chuvas nas fases críticas da cultura, como florescimento e enchimento de grãos. Esse comportamento está em consonância com as recomendações do Zoneamento Agrícola de Risco Climático (ZARC), que orienta janelas de plantio compatíveis com as condições climáticas ideais para o desenvolvimento da cultura (BRASIL, 2023), garantindo também condições favoráveis para a colheita em períodos de menor umidade.

3. O **Cluster 4 (Laranja)** assume o papel de “Fronteira Térmica”. Sua concentração ocorre nas regiões **Norte e Oeste**. **Análise:** Apesar de receber chuvas adequadas (**1.301 mm**), este grupo enfrenta a maior temperatura média do estado (**25,8°C**). O calor excessivo acelera o ciclo metabólico da planta e provoca abortamento floral, limitando a produtividade a um patamar intermediário (**3.022 kg/ha**), inferior ao potencial do **Cluster 3**.
4. Sendo o grupo mais frequente (n=1.304) e abrangente, o **Cluster 1 (Azul)** representa o cenário climático comum de muitas safras em Goiás: muita chuva e temperatura amena. **Análise:** Embora a água seja abundante (**1.338 mm**, a maior do estudo), a produtividade não explode (**2.878 kg/ha**). Isso sugere que o excesso de nebulosidade e a pressão fitossanitária em ambientes muito úmidos atuam como fatores limitantes, impedindo que essas áreas atinjam o teto produtivo do **Cluster 3**.
5. Por fim, o **Cluster 2 (Verde)** delimita as áreas ou anos de baixa aptidão. Fortemente concentrado no **Nordeste Goiano** (região de transição para o semiárido/Matopiba) ou em solos arenosos degradados. **Análise:** Com a menor produtividade média (**2.237 kg/ha**), este grupo reflete a vulnerabilidade da cultura em condições de instabilidade, servindo como um alerta de risco agrícola para expansão de novas áreas.

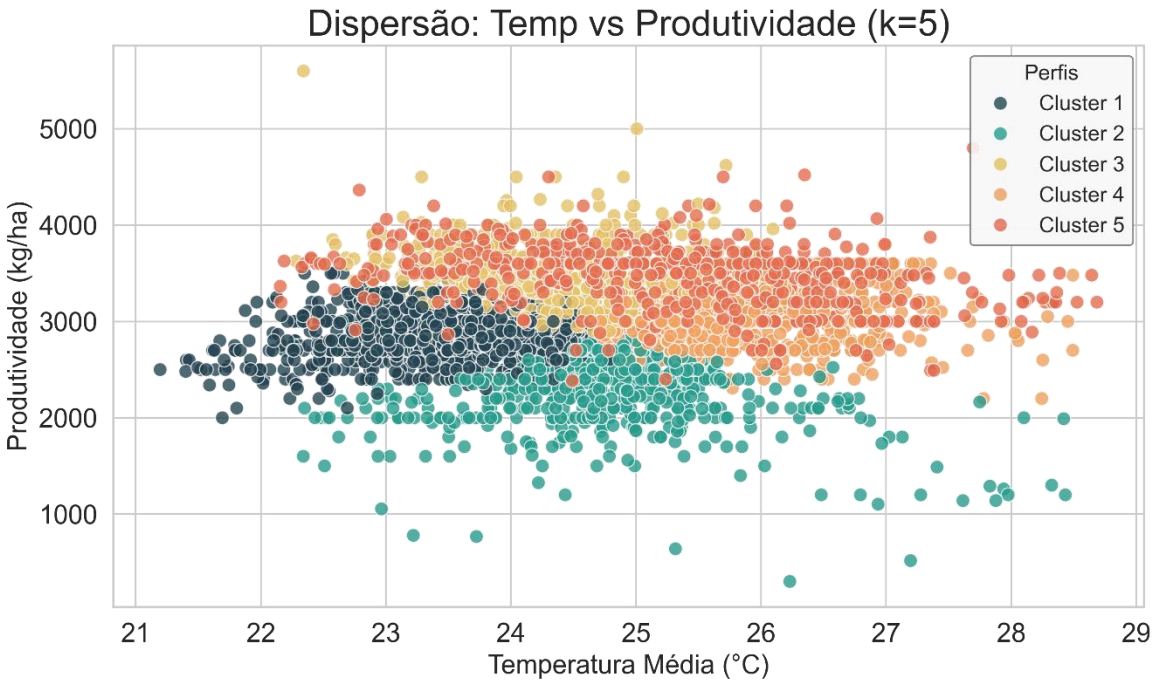
As Figuras 9 e 10 correlacionam a produtividade final com as variáveis climáticas acumuladas.

Figura 9 — Dispersão: Precipitação no Ciclo vs. Produtividade.



Fonte: Elaborado pelo Autor (2026).

Figura 10 — Dispersão: Temperatura Média no Ciclo vs. Produtividade.



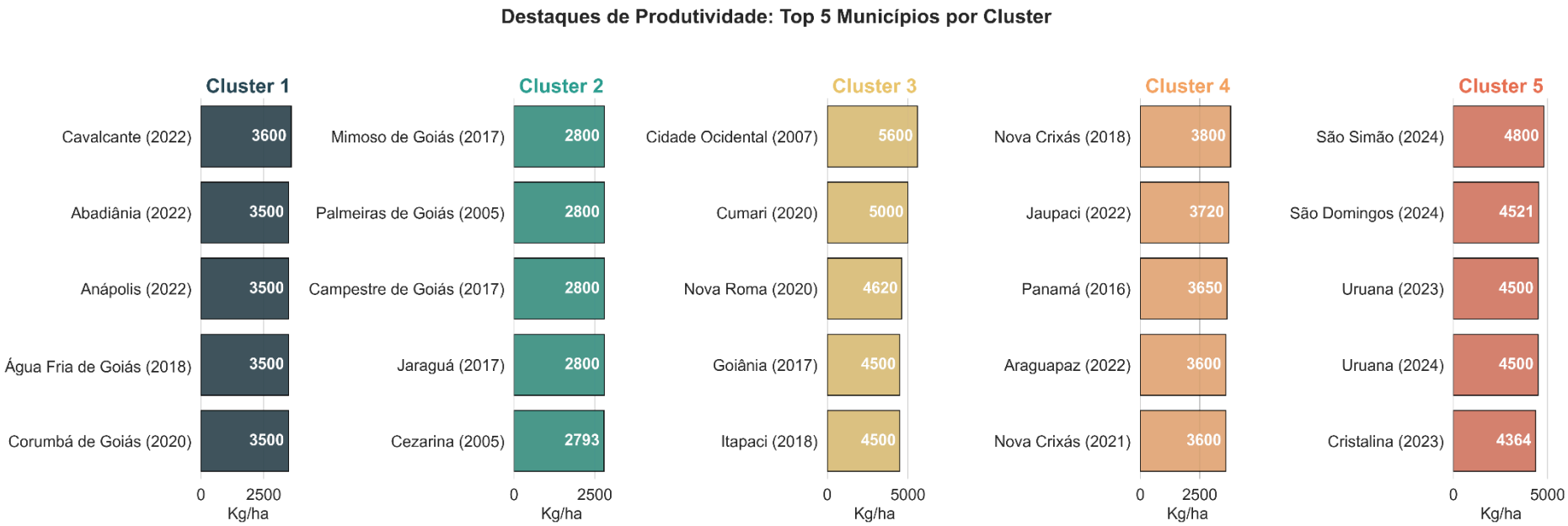
Fonte: Elaborado pelo Autor (2026).

Assim, é possível validar a consistência agronômica dos agrupamentos através de dois padrões comportamentais distintos:

- **A Assinatura da Irrigação (Chuva):** O gráfico de dispersão da precipitação (Figura 9) isola o **Cluster 5** (Rosa-alaranjado) na extremidade esquerda (escassez hídrica, < 700 mm), porém no topo da produtividade (> 3.400 kg/ha). Essa dissociação entre baixo volume de chuva e alto rendimento comprova estatisticamente o uso de tecnologia (pivô central), contrastando com o **Cluster 3** (Amarelo), que atinge o mesmo teto produtivo, mas dependente da faixa ideal de chuvas (~1.140 mm).
- **Resiliência Térmica (Temperatura):** Enquanto o calor excessivo (> 25,5°C, Figura 10) atuou como fator limitante para a maioria dos grupos (especialmente no **Cluster 4**, estagnado em patamares médios), o **Cluster 5** demonstrou alto desempenho. O grupo manteve médias de elite mesmo em zonas de alta demanda térmica, sugerindo que o volume hídrico controlado mitiga os danos fisiológicos causados pelas altas temperaturas.

Para finalizar a análise espacial, a Figura 11 detalha os picos de produtividade, destacando os municípios com os maiores registros absolutos ao longo da série histórica dentro de cada grupo. O ranqueamento confirma visualmente a disparidade de desempenho e a superioridade dos polos identificados nos Clusters 3 e 5. Ao analisar o gráfico, nota-se que o Cluster 3 alcança marcas de extrema eficiência, liderado por Cidade Ocidental (5.600 kg/ha no ano de 2007) e Cumari (5.000 kg/ha em 2020). De forma paralela, o Cluster 5 atesta seu alto rigor tecnológico com tetos produtivos expressivos, como os 4.800 kg/ha registrados em São Simão em 2024. A discrepância regional fica ainda mais evidente ao se observar o Cluster 2: os maiores desempenhos históricos desse grupo (a exemplo de Mimoso de Goiás e Palmeiras de Goiás) estagnaram na faixa de 2.800 kg/há, um teto produtivo inferior até mesmo às médias gerais dos clusters de elite. Observa-se, portanto, que os municípios líderes dos Clusters 3 e 5 não apenas trazem as estatísticas estaduais para cima, mas servem como referências de viabilidade tecnológica e manejo avançado para suas respectivas microrregiões.

Figura 11 — Destaques de cada cluster para produtividade.



Fonte: Elaborado pelo Autor (2026).

5.3 SÍNTESE DOS PADRÕES

A integração entre as análises estatísticas e espaciais permitiu consolidar a caracterização dos cinco perfis produtivos da soja em Goiás. O Quadro 3 faz um resumo de perfil das métricas centrais e a interpretação agronômica de cada agrupamento, servindo como guia definitivo para a compreensão das dinâmicas identificadas no período 2004-2024.

Quadro 3 — Resumo Categórico dos Padrões Agroclimáticos Identificados.

Cluster	**Perfil**	**Prod. (kg/ha)**	**Chuva acumulada (mm)**	**Temp. Média (°C)**	**Categorização**
3	Alta aptidão natural	3.491,86	1.143,48	24,35	**Ideal.** Clima equilibrado permitiu teto produtivo em sequeiro.
5	Alta tecnologia de irrigação	3.479,26	593,67	25,25	**Ótimo.** Alta produção com chuva mínima. Indica possível uso de pivô de irrigação.
4	Limitado pelo calor	3.022,67	1.301,21	25,81	**Estresse Térmico.** Temp. média 25,8°C limitou o potencial genético.
1	Limitado por excesso	2.878,78	1.338,53	23,54	**Excesso Hídrico.** Nebulosidade e doenças podem ter reduzido o rendimento.
2	Baixa aptidão	2.237,84	1.169,89	24,79	**Risco.** Baixa produtividade recorrente (solo ou manejo).

Fonte: Elaborado pelo Autor (2026).

6 CONSIDERAÇÕES FINAIS

Este trabalho alcançou seu objetivo principal ao desenvolver e aplicar um *pipeline* de Ciência de Dados para identificar padrões produtivos e climáticos da soja em municípios do estado de Goiás por meio de técnicas de clusterização. A integração de bases heterogêneas (dados climáticos ERA5 e agronômicos IBGE/PAM) e a aplicação do algoritmo de clusterização permitiram segmentar o estado em perfis comportamentais complexos. A robustez do modelo computacional foi validada pelas métricas de desempenho, obtendo-se um **Coeficiente de Silhueta de 0,28** e um **Índice Caliński-Harabasz de 1.639,33**, valores que atestam a consistência matemática e a alta densidade dos agrupamentos formados, demonstrando a veracidade dos dados coletados, sem inclusão de dados sintéticos.

Com relação aos objetivos específicos, a implementação da lógica de “Ano Safra” no pré-processamento dos dados provou-se crucial. A correção algorítmica do calendário, deslocando os registros para o ciclo fenológico real da cultura, eliminou ruídos temporais que poderiam enviesar o treinamento do modelo.

O principal resultado da mineração de dados foi a identificação estatística da dissociação entre chuva e produtividade. A análise dos centroides revelou que o algoritmo foi capaz de distinguir, sem supervisão humana prévia, a diferença entre aptidão natural e intervenção tecnológica:

- O **Cluster 3** confirmou a vocação natural do Sudoeste Goiano, onde a alta produtividade (3.491 kg/ha) é sustentada por um regime de chuvas ideal (~1.143 mm).
- O **Cluster 5** apresentou produtividade estatisticamente equivalente (3.479 kg/ha) mesmo sob o regime de menor pluviosidade do estado (~593 mm). Este padrão, geograficamente disperso e climaticamente atípico, evidencia estatisticamente a ação da tecnologia (irrigação), capaz de mitigar o déficit hídrico severo e garantir tetos produtivos elevados mesmo em condições adversas.

Adicionalmente, o estudo mapeou os fatores limitantes da produção no estado. Identificou-se que o estresse térmico (temperaturas médias > 25,8°C no

Cluster 4) e o excesso hídrico associado à nebulosidade (no **Cluster 1**) atuam como barreiras naturais, impedindo que certas regiões atinjam o potencial genético máximo da cultura, independentemente do volume total de chuvas.

Conclui-se que a abordagem computacional adotada superou as limitações das análises estatísticas tradicionais, frequentemente baseadas apenas em médias simples, uma vez que o modelo não supervisionado foi capaz de capturar a complexidade multidimensional dos dados. A integração simultânea da produtividade histórica com duas variáveis climáticas fundamentais (precipitação e temperatura) permitiu ir além da interpretação óbvia dos dados e classificar os dados em perfis de eficiência agroclimáticos para o estado de Goiás. Dessa forma, o trabalho demonstra como técnicas de aprendizado de máquina podem transformar dados públicos brutos em inteligência estratégica para o agronegócio. Além disso, a arquitetura desenvolvida com base em KDD com ETL garante total reprodutibilidade metodológica, viabilizando aplicações em outras regiões, culturas ou recortes temporais.

6.1 DELIMITAÇÕES DO ESTUDO

Embora a metodologia tenha tratado corretamente a sazonalidade da cultura através do ajuste para o "Ano Safra" (outubro a abril), o estudo encontrou limitações inerentes à disponibilidade de dados públicos:

- **Ausência de Variáveis:** O modelo baseou-se exclusivamente em variáveis climáticas (ERA5) e de produção (IBGE). A não inclusão de dados sobre as propriedades físicas do solo (como teor de argila e saturação de bases) limitou a capacidade do algoritmo de distinguir municípios que, embora climaticamente idênticos, possuem potenciais produtivos diferentes devido à fertilidade natural da terra.
- **Generalização Espacial:** A utilização da malha municipal como unidade mínima de análise impõe uma generalização dos dados. Em municípios de grande extensão territorial, a média pluviométrica pode mascarar microclimas locais ou a variabilidade espacial das lavouras dentro do mesmo território.

6.2 TRABALHOS FUTUROS

Com base nas limitações expostas e nos resultados gerados e analisados, sugerem-se as seguintes direções para a continuidade desta pesquisa:

- **Novas Variáveis:** A sobreposição dos clusters sugere que Clima e Produtividade, sozinhos, não explicam toda a variância. A inclusão de dados de solo (como teor de argila, pH e saturação de bases) é essencial para distinguir municípios que, mesmo sob o mesmo clima, apresentam produtividades distintas devido à fertilidade natural ou manejo do solo, por exemplo. A inclusão de variáveis climáticas adicionais, como a umidade relativa do ar, pode contribuir para o aprofundamento da análise, desde que fundamentada em estudos que evidenciem sua relevância para a produtividade da cultura. Essas adições podem fortalecer a comprovação do uso de técnicas de irrigação artificial no **Cluster 5**.
- **Teste de Outros Algoritmos:** Dado o desafio das fronteiras não lineares, sugere-se a aplicação de algoritmos baseados em densidade, como o **DBSCAN**, que pode identificar *outliers* climáticos com maior precisão e lidar melhor com a continuidade espacial dos dados agrícolas do que o K-Means. Sugere-se também a aplicação de algoritmos de aprendizado supervisionado, como os de classificação.
- **Ferramenta para uso final:** O desenvolvimento de uma aplicação computacional (tal como um sistema *web*) para automatizar o processo de análise e visualização dos clusters para clientes finais, por exemplo, produtores rurais e profissionais agrônomos.

REFERÊNCIAS

AGÊNCIA FPA. **Entenda como o Brasil se tornou o maior produtor de soja do mundo**. FPA Agropecuária, 29 jan. 2021. Disponível em: https://agencia.fpagropecuaria.org.br/2021/01/29/entenda-como-o-brasil-se-tornou-o-maior-produtor-de-soja-do-mundo/.

ALVARES, Clayton Alcarde et al. Köppen's climate classification map for Brazil. **Meteorologische zeitschrift**, v. 22, n. 6, p. 711-728, 2013.

ARAÚJO, E. C. de; URIBE-OPAZO, M. A.; JOHANN, J. A. Análise de agrupamento da variabilidade espacial da produtividade da soja e variáveis agrometeorológicas na região oeste do Paraná. **Engenharia Agrícola**, v. 33, n. 4, p. 782-795, 2013. Disponível em: https://www.scielo.br/j/eagri/a/JfHPwtdbpPJc8CgCmfMKVXv/?lang=pt. Acesso em: 22 set. 2025.

ARSEGO, Diogo Alessandro et al. Indicadores climáticos e a produtividade de soja no Rio Grande do Sul. **Revista Brasileira de Meteorologia**, v. 34, n. 2, p. 191-200, 2019. Disponível em: https://www.scielo.br/j/rbmet/a/M8sMTCr4JfCqxM9BQTkGYBv/?format=pdf&lang=pt. Acesso em: 22 set. 2025

BISHOP, Christopher M.; NASRABADI, Nasser M. Pattern recognition and machine learning. **New York: springer**, 2006.

BUSSAB, W. O.; MORETTIN, P. A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

BRASIL. Ministério da Agricultura e Pecuária. **Zoneamento Agrícola de Risco Climático (ZARC): Soja**. Brasília, DF: MAPA, 2023. Disponível em: https://mapa-indicadores.agricultura.gov.br/publico/extensions/Zarc/Zarc.html. Acesso em: 22 set. 2025.

CALIŃSKI, Tadeusz; HARABASZ, Jerzy. A dendrite method for cluster analysis. **Communications in Statistics-theory and Methods**, v. 3, n. 1, p. 1-27, 1974.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Calendário de plantio e colheita de grãos no Brasil 2022**. Brasília: Conab, 2022. Disponível em: https://www.conab.gov.br. Acesso em: 22 nov. 2025.

DIAS, T. D. O. P. C.; PASCHOAL JÚNIOR, F**. Aprendizado de máquina aplicado à dispersão geográfica de carteira de seguro agrícola para a cultura de soja das seguradoras.** Trabalho de Conclusão de Curso (Graduação em Engenharia de Produção) – Escola Politécnica, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2023. Disponível em: https://lseufrj.com.br/wp-content/uploads/2023/09/Projeto-Final-Thamirys-e-Paschoal.pdf. Acesso em: 26 jun. 2025.

ESTER, Martin et al. **A density-based algorithm for discovering clusters in large spatial databases with noise.** In: kdd. 1996. p. 226-231.

FARIAS, J. R. B. Exigências climáticas. In: OLIVEIRA, S. et al. **Tecnologias de produção de soja: região central do Brasil 2020/2021.** Londrina: Embrapa, 2019. Disponível em: https://www.embrapa.br/agencia-de-informacao-tecnologica/cultivos/soja/pre-producao/caracteristicas-da-especie-e-relacoes-com-o-ambiente/exigencias-climaticas. Acesso em: 22 set. 2025.

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.

FERREIRA, João et al. O processo etl em sistemas data warehouse. In: **INForum**. sn, 2010. p. 2010.

G1. **Goiás passa a ser o terceiro maior produtor de grãos do país, diz IBGE.** 11 set. 2023. Disponível em: https://g1.globo.com/go/goias/noticia/2023/09/11/goias-passa-a-ser-o-terceiro-maior-produtor-de-graos-do-pais-diz-ibge.ghtml. Acesso em: 26 ago. 2025.

GOIÁS GOVERNO DO ESTADO. **Rio Verde é o segundo maior produtor de soja do Brasil.** 23 nov. 2023. Disponível em: https://goias.gov.br/agricultura/rio-verde-e-o-segundo-maior-produtor-de-soja-do-brasil/. Acesso em: 26 ago. 2025.

GOIÁS. Agência Cora de Notícias. **Goiás bate recorde em produção e produtividade de soja na safra 2024/25.** 2025. Disponível em: https://agenciacoradenoticias.go.gov.br/164623-goias-bate-recorde-em-producao-e-produtividade-de-soja-na-safra-2024-25. Acesso em: 18 set. 2025.

IBGE. **API SIDRA: Interface de programação para acesso aos dados.** Rio de Janeiro: IBGE, 2023a. Disponível em: https://apisidra.ibge.gov.br. Acesso em: 20 ago. 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Malha Municipal Digital**. 2024. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais/15774-malhas.html. Acesso em: 20 ago. 2025.

IBGE. **Pesquisa Agrícola Municipal: PAM**. Instituto Brasileiro de Geografia e Estatística, 2023b. Disponível em: https://sidra.ibge.gov.br/pesquisa/pam. Acesso em: 20 ago. 2025.

IBGE. **Sistema IBGE de Recuperação Automática – SIDRA**. Instituto Brasileiro de Geografia e Estatística, 2023c. Disponível em: https://sidra.ibge.gov.br. Acesso em: 20 ago. 2025.

IPEA. **Clusterização Espacial e não Espacial: um estudo aplicado à agropecuária brasileira**. Brasília: Ipea, 2017. (Texto para Discussão, n. 2279). Disponível em: https://repositorio.ipea.gov.br/bitstreams/b4ce09bc-5a30-43b9-a573-7197c781f52c/download. Acesso em: 22 set. 2025.

JAIN, Anil K. Data clustering: 50 years beyond K-means. **Pattern recognition letters**, v. 31, n. 8, p. 651-666, 2010.

KODINARIYA, Trupti M. et al. Review on determining number of Cluster in K-Means Clustering. **International Journal**, v. 1, n. 6, p. 90-95, 2013.

LIAKOS, Konstantinos G. et al. Machine learning in agriculture: A review. **Sensors**, v. 18, n. 8, p. 2674, 2018.

LISBINSKI, Fernanda Cigainski. Variabilidade climática na produção de milho, trigo e soja. **Revista de Política Agrícola.** v. 33, p. e01974, 2025. Disponível em: https://rpa.sede.embrapa.br/RPA/article/view/1974. Acesso em: 18 set. 2025.

MICROSOFT. **Clustering K-Means: Referência de Componente.** 1 set. 2024. Disponível em: https://learn.microsoft.com/pt-br/azure/machine-learning/component-reference/k-means-clustering?view=azureml-api-2. Acesso em: 22 set. 2025.

MUÑOZ-SABATER, Joaquín et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. **Earth System Science Data**, v. 13, p. 4349–4383, 2021. DOI: 10.5194/essd-13-4349-2021.

MURTAGH, Fionn; CONTRERAS, Pedro. Algorithms for hierarchical clustering: an overview. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, v. 2, n. 1, p. 86-97, 2012.

PEDREGOSA, Fabian et al. Scikit-learn: Machine learning in Python. **the Journal of machine Learning research**, v. 12, p. 2825-2830, 2011.

PEIXOTO, Ingrid Cristina Andrade; DE QUEIROZ, Antônio Marcos. **Identificação e Distribuição dos Clusters Espaciais da Sojicultura em Goiás, 2000 e 2016: uma análise de dados exploratórios municipais**. Curso de Ciencias Economicas da Universidade Federal de Goias-FACE, 2021. Disponível em: https://ideas.repec.org/p/ufb/wpaper/091.html. Acesso em: 22 set. 2025.

ROUSSEEUW, Peter J.; KAUFMAN, L. Finding groups in data. **Hoboken: Wiley Online Library**, v. 1, p. 371, 1990.

ROUSSEEUW, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, v. 20, p. 53-65, 1987.

U.S. DEPARTMENT OF AGRICULTURE. Foreign Agricultural Service. **Crop Explorer**. 2022. Disponível em: https://www.fas.usda.gov/data/production/2222000. Acesso em: 18 fev. 2026.

VIRTANEN, Pauli et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature methods**, v. 17, n. 3, p. 261-272, 2020.