UNIVERSIDADE DE SÃO PAULO
ESCOLA SUPERIOR DE AGRICULTURA "LUIZ DE QUEIROZ"
DEPARTAMENTO DE ENGENHARIA DE BIOSSISTEMAS

TCC

Modelagem e Inteligência Artificial para previsão de produtividade do milho nos principais municípios produtores do Mato Grosso

Aluno: Ulysses Chaves de Menezes Netto
NºUSP: 10552381
E-mail: ulysses.netto@usp.br

Orientador: Prof. Dr. Fábio Ricardo Marin
E-mail: fabio.marin@usp.br

Prof. Dr. Roberto Fray da Silva
E-mail: roberto.fray.silva@usp.br

Pesquisador. Dr. Danilo Augusto Sarti
E-mail: danilo.stats@gmail.com

Pesquisador. Me. Izael Martins Fattori Junior
E-mail: izael.fattori@usp.br

Junho/2025

SUMÁRIO

RESUMO

Modelagem e Inteligência Artificial para previsão de produtividade do milho nos principais municípios produtores do Mato Grosso

Este projeto tem como objetivo abordar a importância da produção agrícola sustentável para alcançar a segurança alimentar global em meio a desafios como mudanças climáticas e restrições de uso da terra. O Brasil desempenha um papel crucial nesse contexto, sendo líder no saldo comercial agrícola e atendendo à crescente demanda de nações em desenvolvimento. O milho é uma cultura de extrema importância, fornecendo cerca de 60% da ingestão de energia global e sendo requisitado tanto para alimentação humana como para ração animal e produção de biocombustível. O estado de Mato Grosso é um destaque na produção de milho, principalmente da segunda safra. Dado o cenário de variações climáticas e a importância do milho para a segurança alimentar, este estudo **propõe o desenvolvimento de** um algoritmo sofisticado de inteligência artificial **para prever** a **produtividade** de milho segunda safra nas localidades de Sorriso, Nova Ubiratã e Nova Mutum, situadas no estado de Mato Grosso. O projeto utilizará dados espaço-temporais, incluindo informações climáticas e socioeconômicas, para identificar indicadores climáticos nos meses anteriores ao início da safra, permitindo antever possíveis variações na produção do milho. O trabalho buscará as melhores fontes de dados, incluindo informações meteorológicas e climáticas, dados socioeconômicos e informações sobre o fenômeno El-Niño/Oscilação Sul (ENOS). Espera-se que, ao final do estudo, sejam apresentadas as performances de diferentes algoritmos de IA para estimar a **produtividade** do milho e identificar as principais variáveis que explicam a variabilidade da produção da cultura no estado de Mato Grosso.

Palavras-chave: Segurança alimentar, Produção agrícola, Milho, Mato Grosso, Inteligência Artificial, Previsão de produtividade

ABSTRACT

Modeling and Artificial Intelligence for forecasting corn yield in the main producing municipalities of Mato Grosso

This project aims to address the importance of sustainable agricultural production in achieving global food security amidst challenges such as climate change and land use restrictions. Brazil plays a crucial role in this context, being a leader in the agricultural trade balance and meeting the growing demand from developing nations. Corn is an extremely important crop, providing about 60% of global energy intake and being in demand for human consumption, animal feed, and biofuel production. The state of Mato Grosso is a highlight in corn production, especially the second harvest (safrinha). Given the scenario of climatic variations and the importance of corn for food security, this study **develops** a sophisticated artificial intelligence algorithm **to predict** second-harvest corn **yield** in the locations of Sorriso, Nova Ubiratã, and Nova Mutum, situated in the state of Mato Grosso. The project will use spatio-temporal data, including climatic and socioeconomic information, to identify climatic indicators in the months prior to the start of the harvest, **to anticipate** possible variations in corn production. The work will seek the best data sources, including meteorological and climatic information, socioeconomic data, and information on the El Niño-Southern Oscillation (ENSO) phenomenon. It is expected that, at the end of the study, the performances of different AI algorithms for estimating corn **yield** will be presented, along with the identification of the main variables that explain the variability of the crop's production in the state of Mato Grosso.

Keywords: Food security, Agricultural production, Mato Grosso, Artificial Intelligence, Yield prediction

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CO2	Dióxido de Carbono
EFB	Exclusive Feature Bundling
ENSO	El Niño-Southern Oscillation
ESALQ	Escola Superior de Agricultura “Luiz de Queiroz”
ETo	Evapotranspiração de Referência
FAO	Food and Agriculture Organization
GOSS	Gradient-based One-Side Sampling
GS-CV	GridSearchCV
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IQR	Intervalo Interquartil
k-NN	k-Nearest Neighbors
LGBM	LightGBM
MAE	Mean Absolute Error
PIB	Produto Interno Bruto
Pr	Precipitação
R2	Coeficiente de Determinação
RF	Random Forest
RMSE	Root Mean Square Error
RS-CV	RandomizedSearchCV
SVM	Support Vector Machine
TCC	Trabalho de Conclusão de Curso
Tmax	Temperatura Máxima
Tmin	Temperatura Mínima
USP	Universidade de São Paulo
XGB	XGBoost

1 INTRODUÇÃO

Para alcançar a segurança alimentar de forma sustentável, é essencial aumentar a produção agrícola, pois o crescimento populacional exigirá um suprimento alimentar cada vez maior até que atinja uma estabilização em torno de 9 bilhões de pessoas em meados deste século. Além disso, é crucial considerar desafios como as mudanças climáticas e a restrição do uso da terra pela população humana (Charles *et al.*, 2010). O Brasil desempenha um papel crucial na segurança alimentar global durante a atual crise, graças à sua liderança no saldo comercial agrícola e ao aproveitamento da crescente demanda de nações em desenvolvimento, em especial da Ásia e do Oriente Médio (Vieira *et al.*, 2019).

Ao longo dos anos, o comércio agrícola brasileiro passou por mudanças significativas, incluindo uma redução considerável na participação relativa da Europa, que viu sua fatia nas exportações agrícolas brasileiras diminuir de 57% para 19% no mesmo período (Vieira *et al.*, 2019). Diante dessas transformações, torna-se crucial para o Brasil desenvolver estratégias que ampliem sua importância na oferta de alimentos para outros países e que o mantenham como uma fonte confiável de alimentos para toda a humanidade.

O milho tem ganhado grande destaque devido ao seu status como a segunda cultura mais cultivada no mundo. Em conjunto com o arroz e o trigo, ele fornece cerca de 60% da ingestão de energia global (Monfreda *et al.*, 2008). Além disso, o milho é uma das principais culturas de cereais, desempenhando um papel crucial como alimento essencial e exercendo influência significativa nas esferas cultural, econômica, ambiental e nutricional em escala global (Guzzon *et al.*, 2021).

Sua relevância transcende a sua importância como componente fundamental na alimentação humana. A cultura do milho é amplamente requisitada devido aos investimentos internacionais em ração animal e à produção de etanol como biocombustível (Tanumihardjo *et al.*, 2019). Essa demanda adicional impulsiona ainda mais a produção e o cultivo do milho, tornando-o uma cultura de extrema importância para diversos setores econômicos e sustentáveis.

O milho é uma planta do tipo C4, ou seja, possui adaptações que o torna mais eficiente em condições de déficit hídrico. Ele utiliza uma rota alternativa de fixação de CO_2, através da enzima PEP Carboxilase, na qual possui maior afinidade com o CO_2 do que a enzima Rubisco presente nas plantas C3. Também pode-se ressaltar que nele o CO_2 é convertido em ácidos de quatro carbonos no mesófilo e transportado para as células da bainha do feixe vascular, onde ocorre a liberação do CO_2 e sua redução a carboidratos pelo Ciclo de Calvin, resultando na produção de fotoassimilados. Além disso, ele apresenta uma reduzida fotorrespiração devido ao baixo teor de oxigênio nas células da bainha do feixe vascular, o que evita a competição com a fotossíntese e permite uma melhor utilização do carbono (Vieira Junior, 2006).

Este cereal apresenta um crescimento satisfatório com uma temperatura mínima de 8°C. A Temperatura ótima é de 30°C, enquanto a temperatura máxima tolerada é de aproximadamente 44°C. As maiores taxas de crescimento do milho são alcançadas quando a temperatura está entre 26°C e 34°C (Bergamaschi e Matzenauer, 2014). Estas características são encontradas no estado que está no topo do ranking na produção brasileira.

O estado de Mato Grosso, no Brasil, é conhecido por sua destacada produção de cereais, especialmente o milho segunda safra. Com uma participação de $37,6\%$ na produção nacional de milho, o estado é responsável por $49,3\%$ da produção de milho safrinha, contribuindo com $10,5\%$ do PIB de Mato Grosso (IMEA - Instituto Mato-Grossense de Economia Agropecuária, 2023).

A compreensão da distribuição temporal e espacial dos atributos climáticos de superfície requer a análise de diversos sistemas, entre eles o fenômeno conhecido como ENSO (El-Niño - Oscilação Sul). O ENSO desempenha um papel fundamental nesse contexto, pois é considerado um dos principais fenômenos a serem estudados (Santos, 2000), sendo que estudos revelam que durante o El Niño/La Niña, ocorre uma relação negativa com as chuvas, ou seja, está associado a períodos de chuvas acima da média ou

secas. Por outro lado, as correlações positivas indicam que o El Niño/La Niña está relacionado a períodos de secas ou chuvas acima da média (Peres e Maier, 2019). Além disso, segundo de Souza (2018) e Jones e Thornton (2003) os senários futuros indicam queda na produtividade do milho e aumento do risco climático para a maioria das regiões do Brasil.

Devido à importância do estado de Mato Grosso na produção de milho e à inerente variabilidade da produtividade da cultura, fortemente influenciada pelo clima, o acompanhamento dessas lavouras e a aplicação de métodos preditivos tornam-se essenciais. A capacidade de antever riscos para a produtividade é crucial para a formulação de políticas de segurança alimentar. Nesse contexto, a realização de estudos que correlacionam variáveis climáticas com a produtividade do milho safrinha é imperativa.

No contexto nacional, estudos anteriores abordaram a previsão de produtividade com base em dados climáticos: no Mato Grosso do Sul, foram utilizadas séries históricas de radiação solar global, temperaturas (média, máxima e mínima do ar), umidade relativa, velocidade do vento e precipitação pluvial como variáveis de entrada para estimar a produtividade do milho safrinha através de um modelo mecanístico, comprovando sua confiabilidade com índices estatísticos (Aparecido *et al.*, 2020). De forma semelhante, em São Paulo e em âmbito nacional, outros trabalhos desenvolveram modelos para a cultura do milho Pinheiro (2004) e da cana-de-açúcar Marin e Nassif (2013), visando explicar os processos bioquímicos, físicos e fisiológicos que influenciam a produtividade.

Apesar de tais trabalhos fornecerem informações relevantes sobre os riscos climáticos e a produtividade do milho, eles não apresentam uma metodologia clara para a previsão de riscos. Ou seja, ainda não há uma abordagem que busque antecipar possíveis variações na produção do milho segunda safra com tempo hábil para uma tomada de decisão eficaz. Portanto, até onde se tem conhecimento, carecem estudos que se proponham a prever com antecedência as variações na produtividade do milho, fundamentando-se nas condições observadas antes do início da safra.

Uma forma de se entender e identificar fatores que afetam uma variável resposta, como a produtividade é o uso da inteligência artificial (IA) na modelagem tradicional de produção. A inteligência artificial (IA) consiste em desenvolver algoritmos e paradigmas que permitem às máquinas realizarem tarefas cognitivas que normalmente requerem habilidades humanas. Em geral, um sistema de IA deve ser capaz de armazenar conhecimento, aplicar esse conhecimento para resolver problemas e adquirir novo conhecimento através da experiência. A IA é frequentemente dividida em três componentes fundamentais: representação, raciocínio e aprendizagem (Haykin, 2001).

A Inteligência Artificial está sendo usada por muitos pesquisadores para modelar uma ampla gama de tarefas na agricultura. Há muitos estudos em que foram analisadas variáveis espaço-temporais e relacionando-as com a previsão da produtividade da cultura utilizando-se vários modelos Rede Neural Artificial e o modelo de Máquina de Vetores Suporte (Nyéki *et al.*, 2019).

Ainda com esse mesmo objetivo, foram utilizados modelos com sistemas de Inferência Fuzzy Neuro-adaptativos (Khashei-Siuki *et al.*, 2011) e *Random Forest* (Suresh *et al.*, 2021) nos quais obtiveram resultados satisfatórios.

2 OBJETIVOS

Este projeto de estudo visa **avaliar e comparar o desempenho de diferentes algoritmos de Inteligência Artificial** (*Random Forest*, XGBoost e LightGBM) para prever a produtividade do milho segunda safra no estado de Mato Grosso. Para tanto, serão utilizados dados de variáveis espaço-temporais de **todos os municípios produtores do estado**. Adicionalmente, o estudo busca identificar os principais sinais climáticos e socioeconômicos que antecedem o início da safra e que permitem antever as variações na produtividade da cultura.

3 MATERIAL E MÉTODOS

3.1 Coleta de Dados

A coleta de dados foi crucial, envolvendo informações de duas fontes principais: dados meteorológicos foram obtidos do NASAPower e dados de produção e produtividade de soja e milho, da plataforma SIDRA do IBGE. Ambos os conjuntos foram baixados em formato CSV para garantir a uniformidade na análise. O diagrama de fluxo que ilustra este processo de coleta e unificação dos dados é apresentado na Figura 3.1.

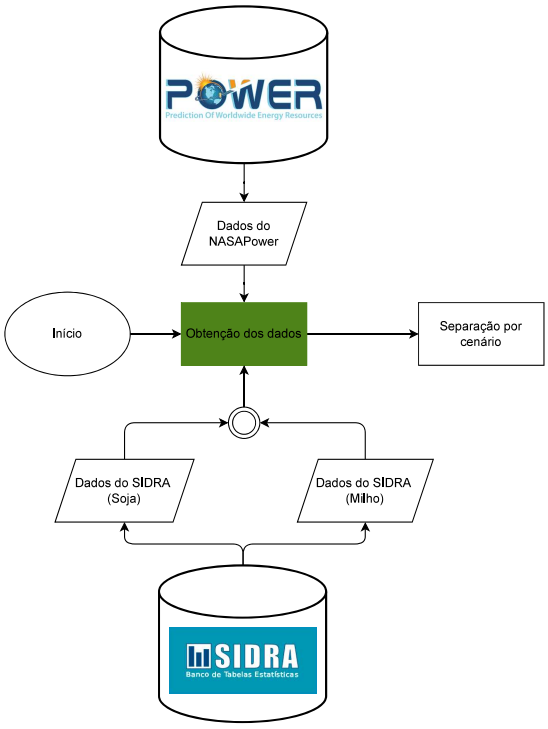


Figura 3.1. Diagrama do fluxo de obtenção dos dados.
Fonte: Elaborado pelo autor.

3.2 Exploração de Dados

A fase de exploração de dados teve como objetivo central identificar padrões e variáveis com potencial preditivo para a produtividade do milho safrinha, a partir de duas fontes de dados distintas: NASAPower e SIDRA. A premissa é que as variáveis meteorológicas (NASAPower) impactam diretamente o ciclo de desenvolvimento da cultura e sua produtividade final. Adicionalmente, investigou-se a hipótese de que características da safra de soja anterior (SIDRA) — tais como o acúmulo de matéria seca, a retenção de nitrogênio e a fertilidade residual do solo — atuam como variáveis secundárias que influenciam o rendimento do milho safrinha subsequente.

3.2.1 Dados referentes à cultura da Soja

A análise da produtividade histórica da soja no estado de Mato Grosso, ilustrada na Figura 3.2, revela uma tendência geral de crescimento, embora com flutuações anuais notáveis. O gráfico apresenta a evolução da produtividade média (em kg/ha), onde se observa um pico em torno do ano de 2017, seguido por uma queda e posterior recuperação. A faixa sombreada, representando o desvio padrão, indica a variabilidade dos dados em cada safra, sendo mais acentuada em períodos de maior instabilidade produtiva. A compreensão desta tendência histórica é fundamental para contextualizar a performance da cultura e os fatores macroclimáticos que a influenciam.

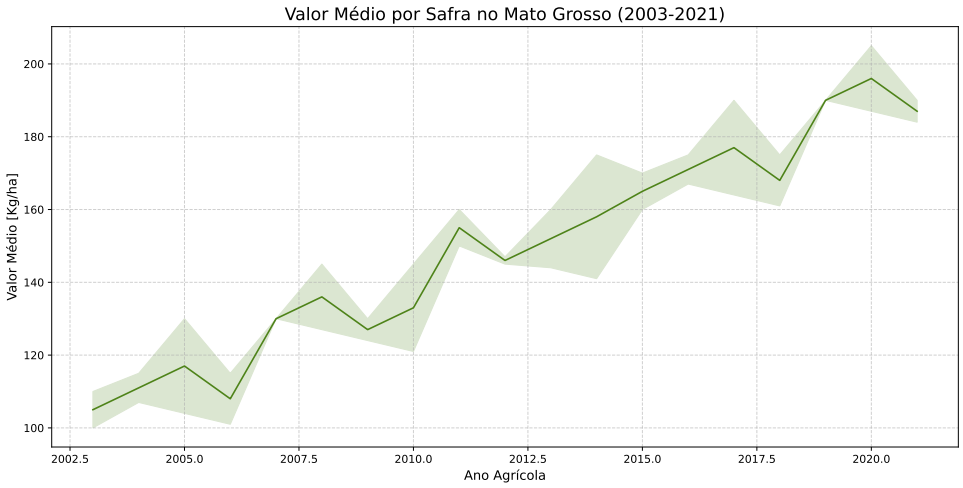


Figura 3.2. Evolução da produtividade média de soja (kg/ha) no estado de Mato Grosso (Safra 2003-2021). A linha representa a média e a área sombreada, o desvio padrão.
Fonte: Elaborado pelo autor.

3.2.2 Dados referentes à cultura do Milho Safrinha

A expansão da cultura do milho safrinha no Brasil pode ser observada através de múltiplas métricas. O número de municípios engajados no cultivo tem apresentado um crescimento constante desde 2003, com uma aceleração notável a partir de 2010, como pode ser visto no gráfico superior das Figuras 3.3, 3.4 e 3.5.

Essa expansão não se limita apenas ao número de produtores. A Figura 3.3 detalha a distribuição da área plantada por município, revelando que a mediana de hectares dedicados à cultura também tem aumentado consistentemente ao longo dos anos. Como consequência direta do aumento da área, a produção total por município também cresceu de forma significativa, conforme ilustrado na Figura 3.4. A presença de *outliers* neste gráfico indica uma concentração da produção, com alguns municípios alcançando volumes excepcionalmente altos.

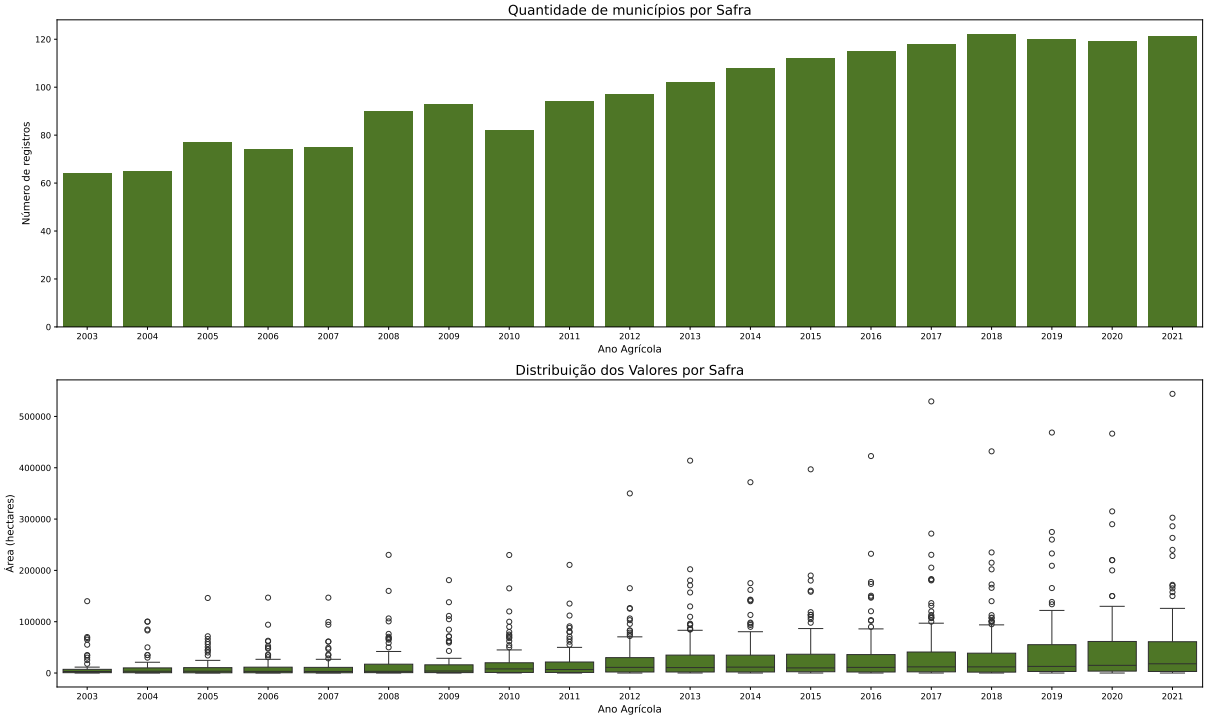


Figura 3.3. Evolução do número de municípios produtores e da distribuição da área plantada (hectares) com milho safrinha.

Fonte: Elaborado pelo autor.

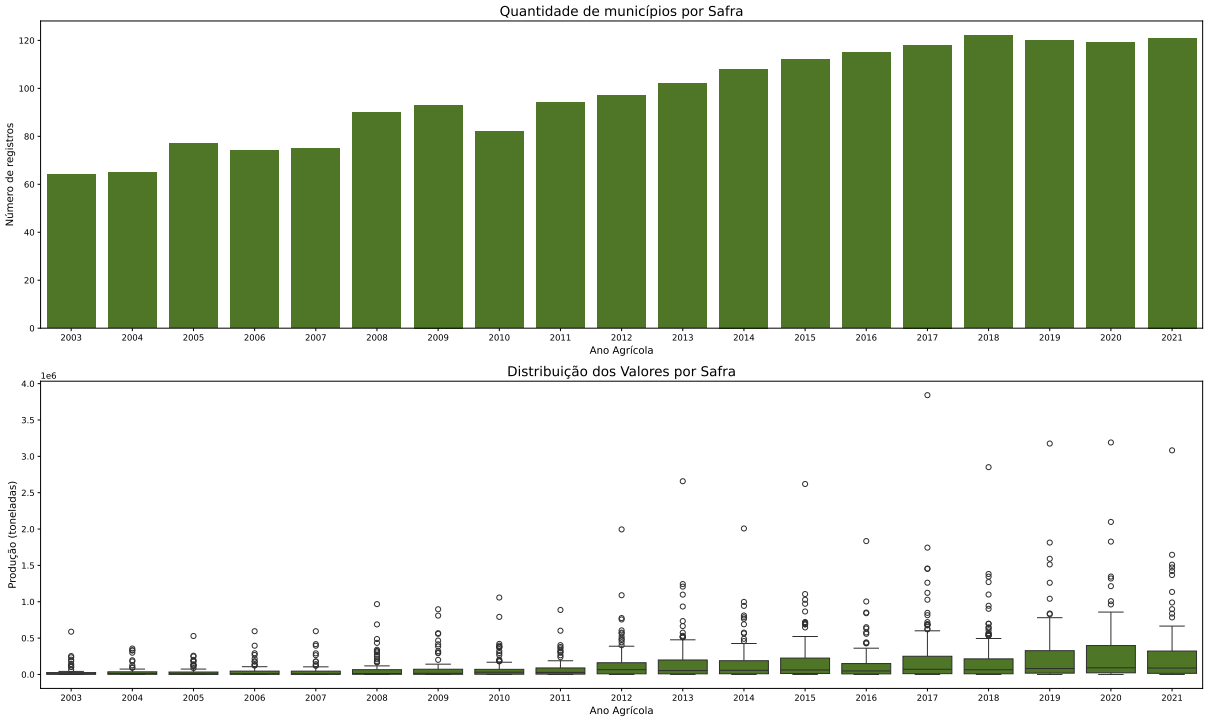


Figura 3.4. Evolução do número de municípios produtores e da distribuição da produção total (toneladas) de milho safrinha.

Fonte: Elaborado pelo autor.

Em contrapartida, ao analisar a produtividade (toneladas por hectare) na Figura 3.5, nota-se que o rendimento por área se manteve relativamente estável, embora com alta variabilidade entre os produtores em cada safra. A presença de *outliers* neste caso sugere que, enquanto a maioria dos municípios mantém

uma produtividade consistente, uma minoria alcança rendimentos muito superiores, possivelmente devido a avanços tecnológicos ou condições microclimáticas favoráveis. Esta análise conjunta sublinha que a crescente importância do milho safrinha foi impulsionada primariamente pela expansão da fronteira agrícola, mais do que por saltos de produtividade média.

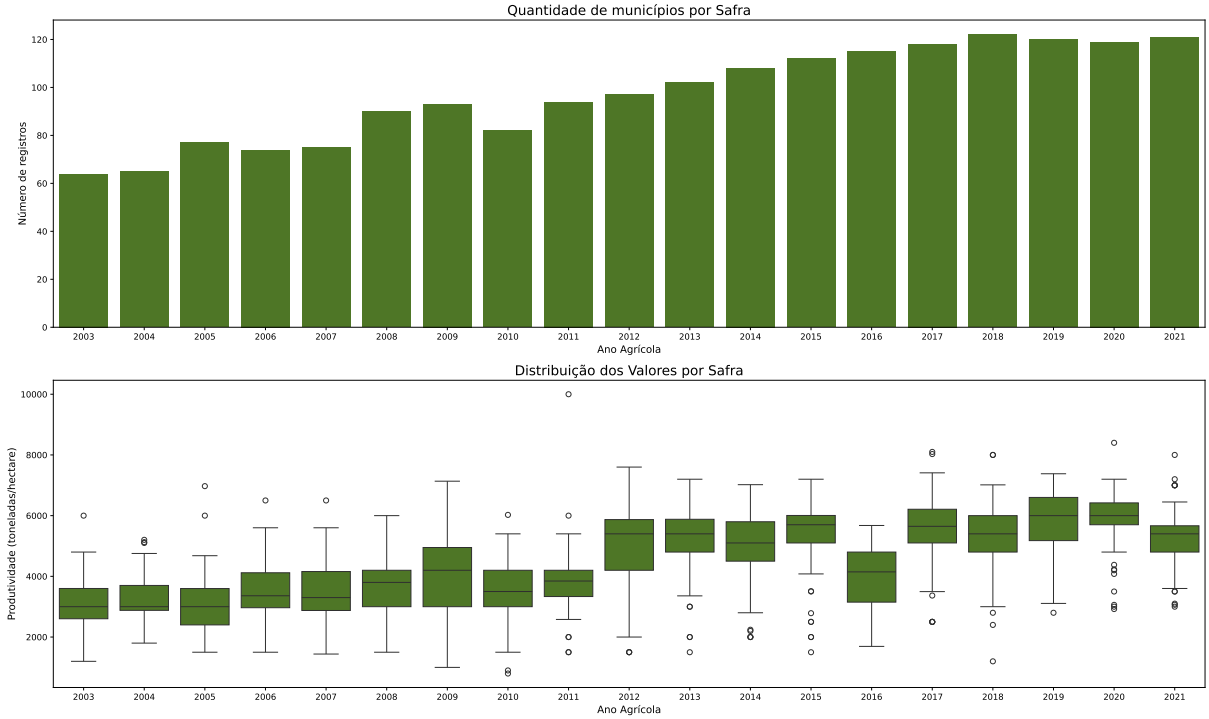


Figura 3.5. Evolução do número de municípios produtores e da distribuição da produtividade (toneladas/hectare) de milho safrinha.

Fonte: Elaborado pelo autor.

Para uma compreensão mais detalhada da dinâmica da produtividade ao longo do tempo, a Figura 3.6 apresenta a evolução da distribuição de produtividade do milho em Mato Grosso de 2003 a 2023. Esta visualização em 3D permite observar como a forma, a média e a variabilidade da produtividade se modificaram ao longo dos anos, revelando padrões e tendências que são cruciais para a modelagem preditiva.

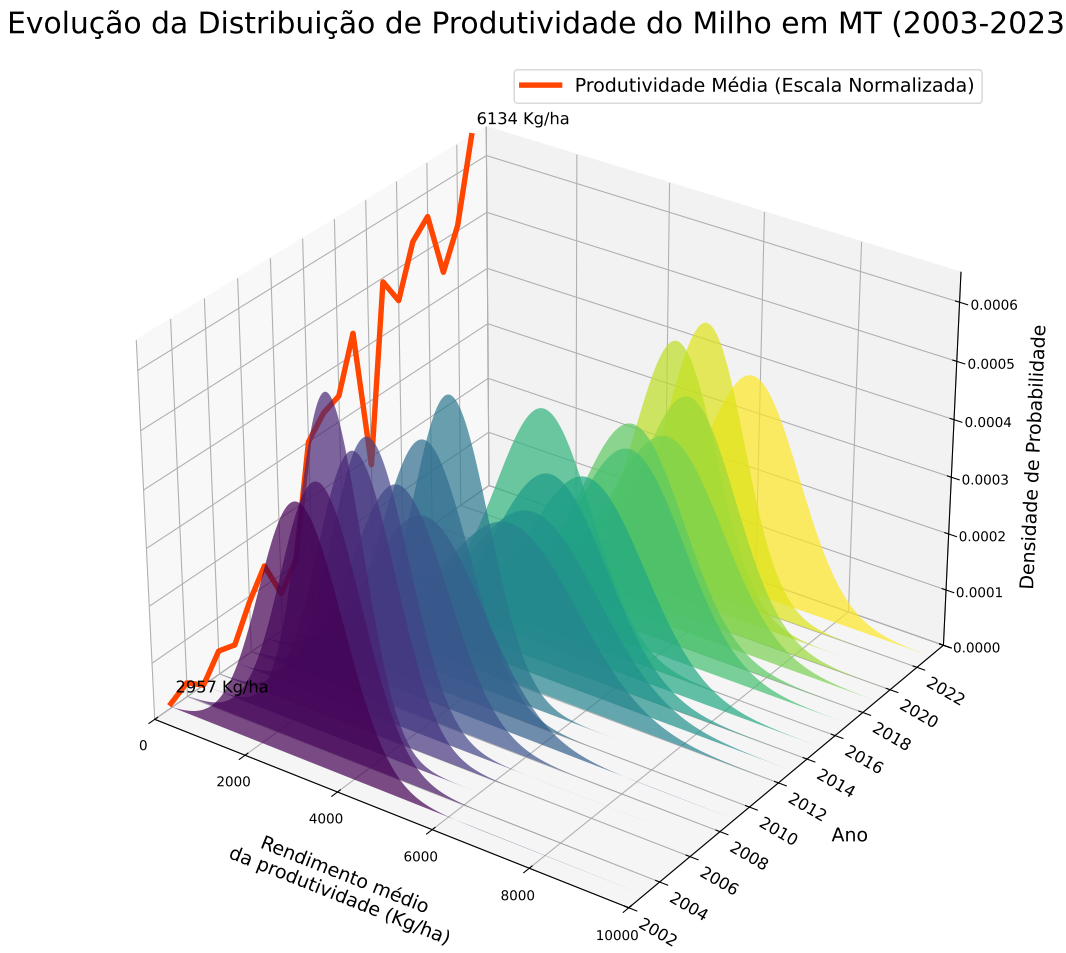


Figura 3.6. Evolução da Distribuição de Produtividade do Milho em MT (2003-2023).
Fonte: Elaborado pelo autor.

Para uma compreensão mais aprofundada da distribuição espacial dessa produtividade em um ano específico, a Figura 3.7 ilustra a distribuição de probabilidade e o mapa de produtividade média de milho safrinha por município no Mato Grosso para o ano de 2021, complementando a análise temporal com um contexto geográfico crucial.

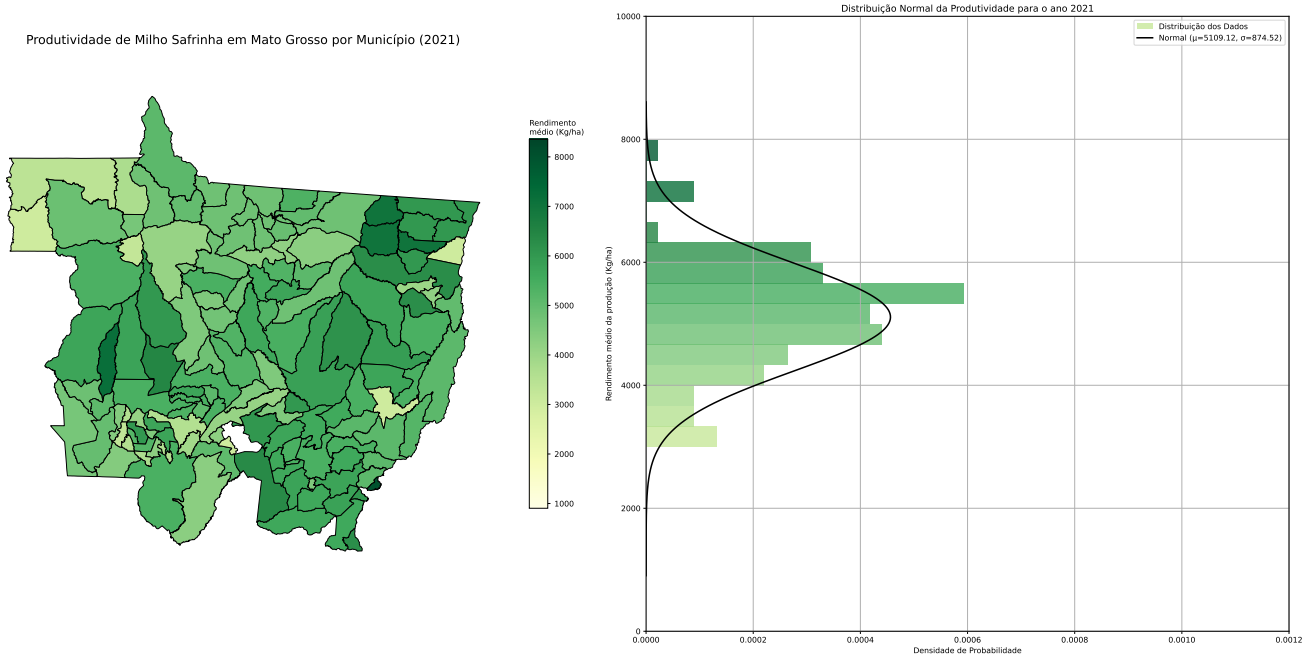


Figura 3.7. Distribuição de probabilidade e mapa de produtividade média de milho safrinha por município no Mato Grosso para o ano de 2021.

Fonte: Elaborado pelo autor.

3.2.3 Dados referentes à Evapotranspiração de referência (ETo)

A Evapotranspiração de Referência (ET_0) é uma variável agrometeorológica fundamental, pois quantifica a demanda hídrica da atmosfera, influenciando diretamente as necessidades de água da cultura. Para estimar este parâmetro crucial, foi utilizada a equação padrão FAO Penman-Monteith (Equação 3.1), recomendada por sua robustez e precisão em diversas condições climáticas (Allen *et al.*, 1998).

$$ET_0 = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} u_2 (e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)} \tag{3.1}$$

A aplicação deste método aos dados anuais dos municípios revelou uma considerável variabilidade interanual, como é apresentado na Figura 3.8. A análise do período revela uma considerável variabilidade de ano para ano, sem uma tendência clara de aumento ou diminuição a longo prazo. Destaca-se o ano de 2016, que registrou picos de ETo para Sorriso e Nova Ubiratã, indicando um ano de particular interesse para a análise da demanda hídrica. Comparativamente, o município de Nova Ubiratã consistentemente exibiu os valores mais elevados de ETo, enquanto Nova Mutum e Sorriso apresentaram comportamentos mais próximos. Essa variabilidade ressalta a importância de considerar as condições específicas de cada safra no planejamento do manejo hídrico.

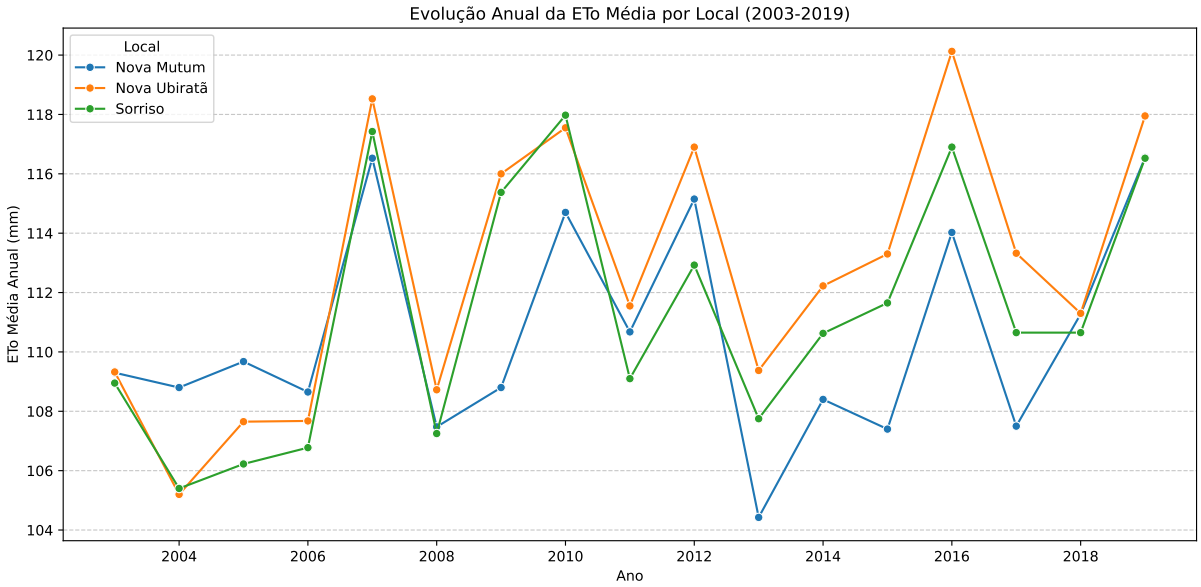


Figura 3.8. Evolução da Evapotranspiração de Referência (ETo) ao longo do período de estudo.
Fonte: Elaborado pelo autor.

3.2.4 Dados referentes à Temperatura Máxima (Tmax)

As temperaturas máximas anuais (Tmax) nos municípios estudados apresentaram uma notável variabilidade interanual no período de 2005 a 2019, como detalhado na Figura 3.9. Embora não se observe uma tendência clara de aquecimento ou resfriamento a longo prazo, picos de temperatura são evidentes em anos específicos, com destaque para 2010 e 2015. Na comparação entre as localidades, Sorriso consistentemente registrou as maiores médias de Tmax, enquanto Nova Mutum apresentou, na maioria dos anos, os valores ligeiramente mais baixos. Essa diferença no comportamento térmico entre municípios próximos é um fator fundamental para a compreensão das respostas do agroecossistema pelos modelos implementados.

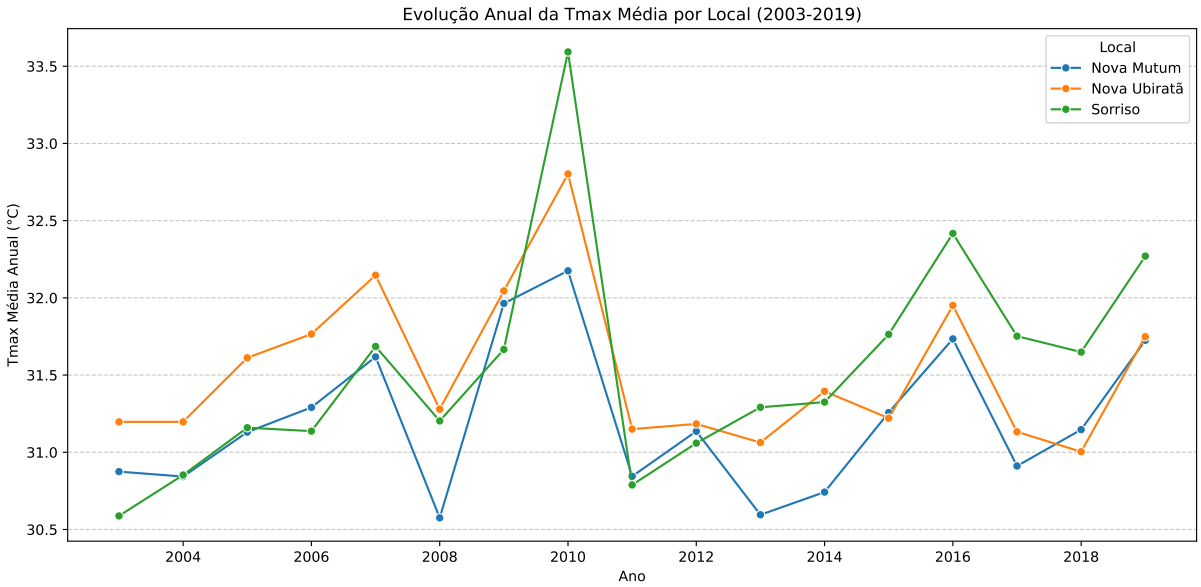


Figura 3.9. Evolução da Temperatura Máxima (Tmax) ao longo do período de estudo.
Fonte: Elaborado pelo autor.

3.2.5 Dados referentes à Temperatura Mínima (Tmin)

A análise das temperaturas mínimas anuais (Tmin), apresentada na Figura 3.10, demonstra um comportamento distinto e consistente entre os três municípios. Ao longo do período, observa-se uma forte variabilidade interanual, com destaque para uma queda acentuada em 2011, em vez de uma tendência de longo prazo. Um padrão claro que emerge do gráfico é a estratificação térmica entre as cidades: Sorriso consistentemente registra as temperaturas mínimas mais elevadas, enquanto Nova Mutum apresenta as mais baixas. O monitoramento da Tmin é vital na agricultura, pois valores extremamente baixos podem indicar risco de geadas, além de influenciar diretamente os ciclos fenológicos da cultura.

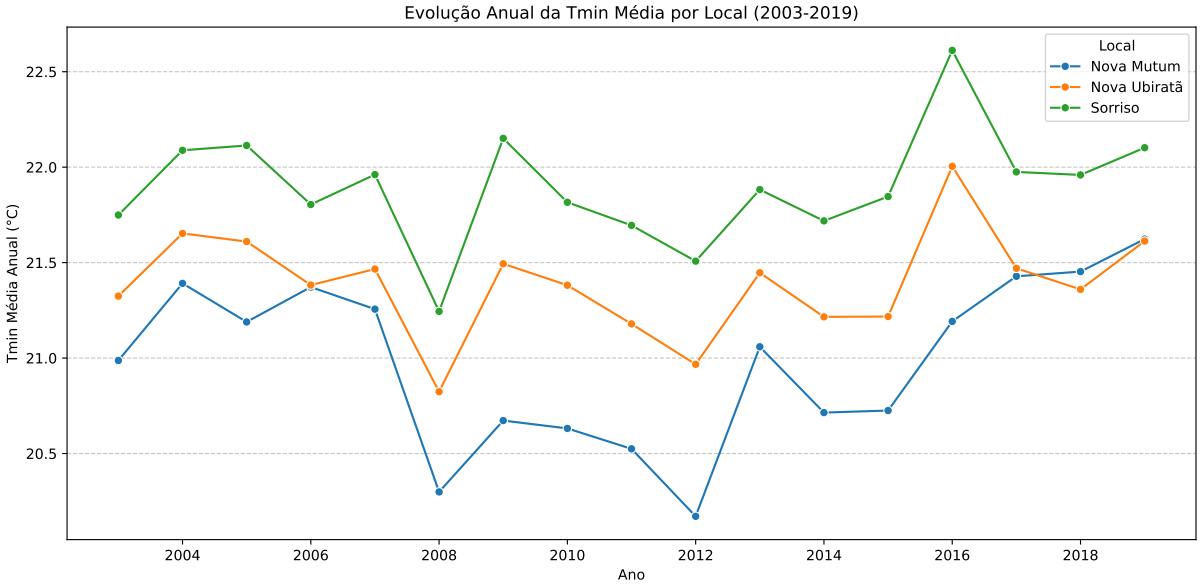


Figura 3.10. Evolução da Temperatura Mínima (Tmin) ao longo do período de estudo.
Fonte: Elaborado pelo autor.

3.2.6 Dados referentes à Precipitação (Pr)

O regime de precipitação anual (Pr), ilustrado na Figura 3.11, é caracterizado por uma elevada e errática variabilidade interanual, sendo um dos fatores mais imprevisíveis para a agricultura na região. Diferentemente das variáveis de temperatura, não há uma estratificação clara entre os municípios, cujos totais anuais de chuva frequentemente se alternam em liderança. Anos como 2016 destacam-se por picos de precipitação nas três localidades, enquanto 2015 registrou um dos valores mais baixos do período. Essa imprevisibilidade do volume de chuvas de um ano para o outro evidencia a vulnerabilidade da produção agrícola à disponibilidade hídrica e reforça a importância de modelos preditivos que possam auxiliar no planejamento e na mitigação de riscos.

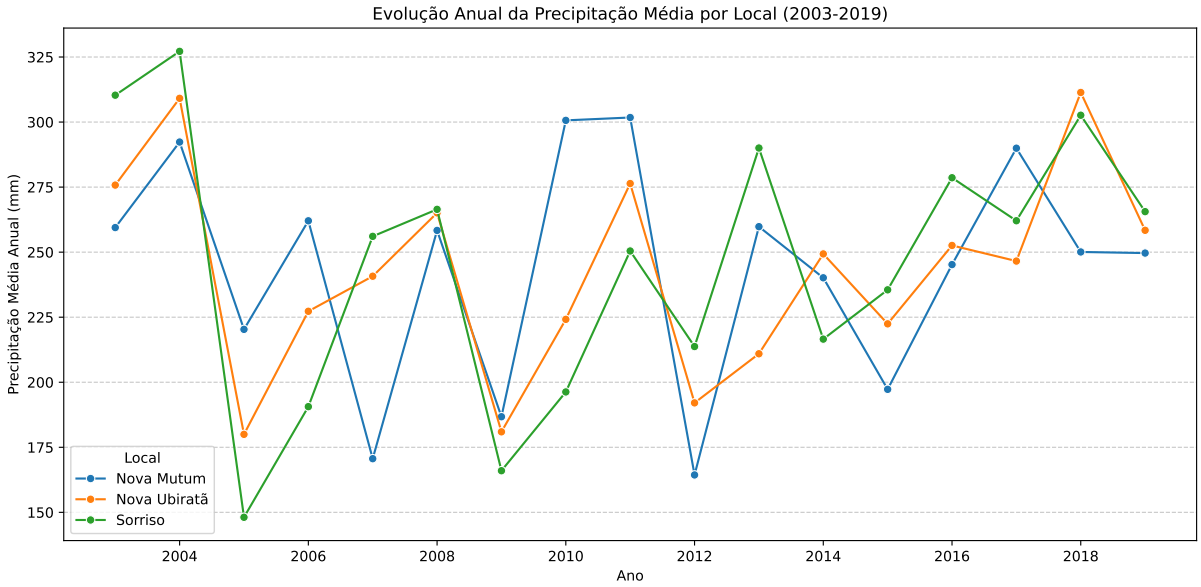


Figura 3.11. Evolução da Precipitação (Pr) ao longo do período de estudo.
Fonte: Elaborado pelo autor.

3.2.6.1 Análise de Correlação entre as Variáveis

Para investigar a relação linear entre as variáveis preditoras e a produtividade do milho, foi gerada uma matriz de correlação de Pearson, visualizada através do heatmap na Figura 3.12. A análise foca nas variáveis com maior correlação com a `produtividade_milho`.

Observa-se uma correlação negativa de moderada a forte com a temperatura mínima, especialmente nos meses de fevereiro a abril (`Tmin_FEV`, `Tmin_MAR`, `Tmin_ABR`), sugerindo que noites mais quentes nesses períodos podem ser prejudiciais à produtividade. Em contrapartida, a precipitação em março (`pr_MAR`) exibe uma correlação positiva, indicando que chuvas neste mês tendem a ser benéficas. Outras correlações notáveis incluem a forte associação positiva entre o ano e a produção total de milho, o que reflete a expansão da cultura ao longo do tempo.

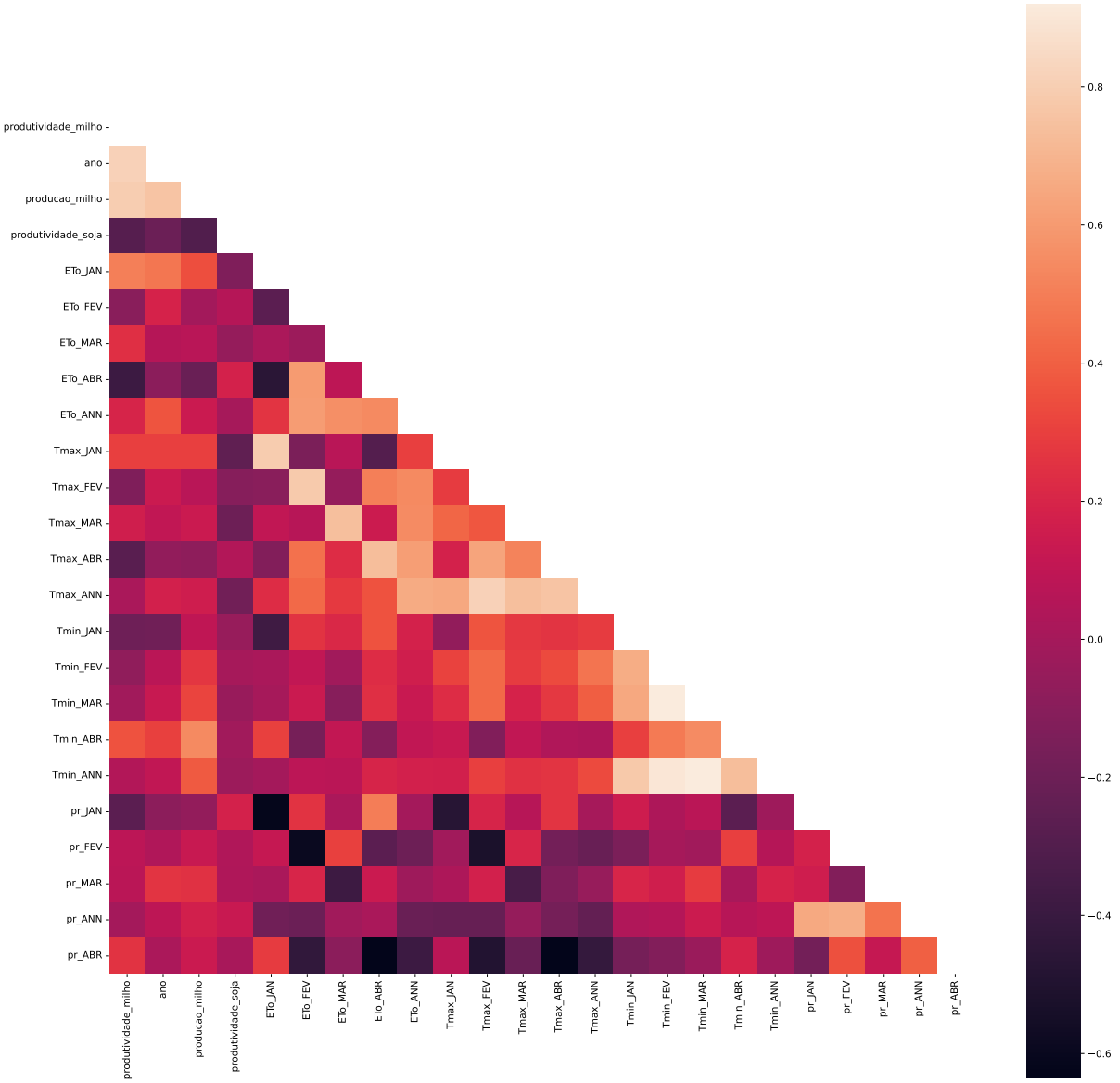


Figura 3.12. Heatmap de correlação de Pearson entre a produtividade do milho e as variáveis preditoras.
Fonte: Elaborado pelo autor.

3.3 Montagem dos Cenários

Para avaliar o impacto do tempo de antecedência na eficiência da previsão da produtividade, a modelagem foi estruturada em quatro cenários distintos, como ilustrado no esquema da Figura 3.13. Cada cenário utiliza um conjunto de dados de entrada (*features* X) com um número diferente de meses, variando de dois (Janeiro-Fevereiro) a cinco meses (Janeiro-Maio), para prever a produtividade do milho (*target* Y). Após a estruturação dos cenários, foi aplicado um processo de seleção sequencial de características para identificar o subconjunto de variáveis mais informativo para cada modelo. A Figura 3.14 exemplifica este processo para o Cenário 1 com o modelo LightGBM, onde o desempenho ótimo foi alcançado com o uso de 13 *features*, resultando em um score de 0.5711. Esta abordagem permite não só comparar a performance preditiva entre os diferentes cenários temporais, mas também otimizar cada modelo com as *features* mais relevantes.

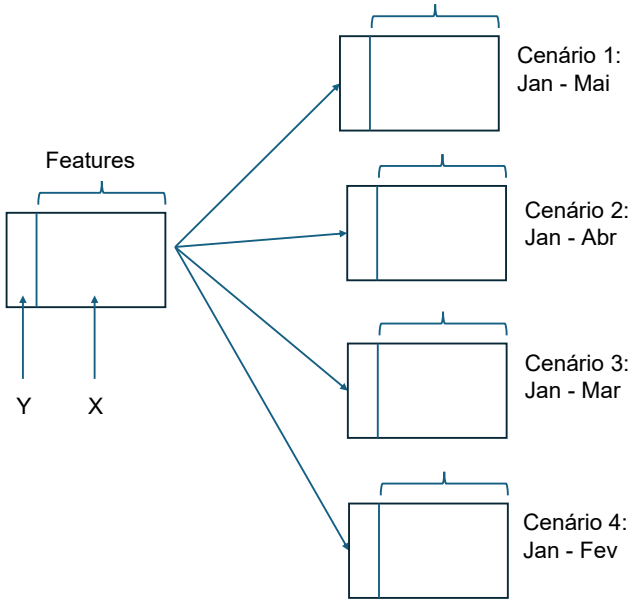


Figura 3.13. Esquema da montagem dos diferentes cenários de modelagem, variando o número de meses nos dados de entrada.

Fonte: Elaborado pelo autor.

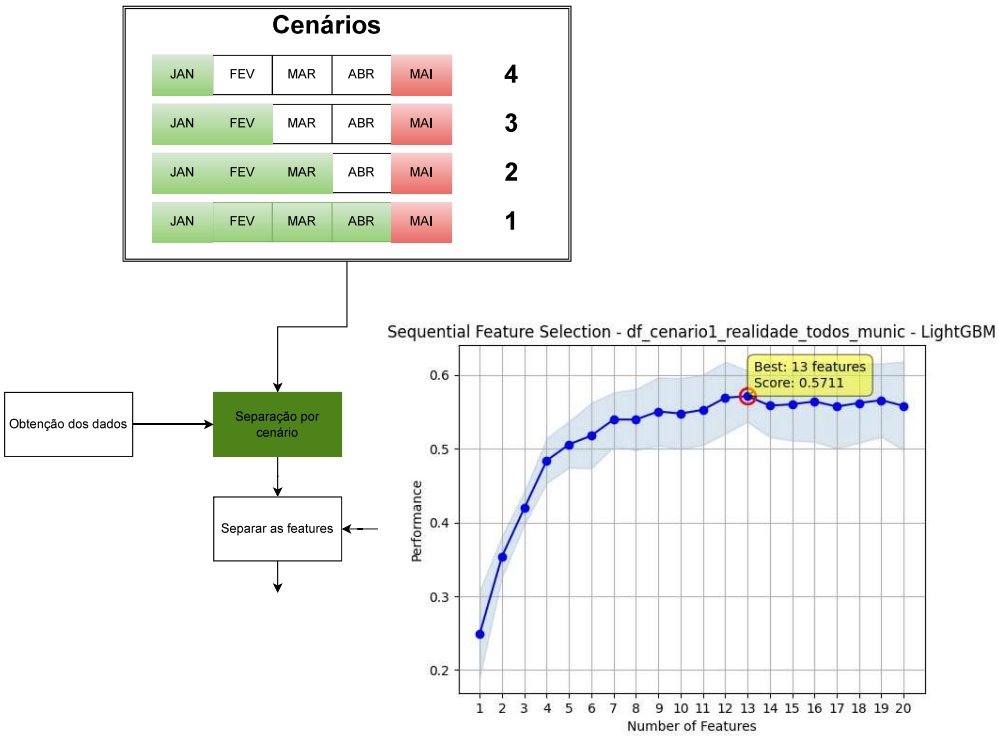


Figura 3.14. Exemplo do processo de Seleção Sequencial de *Features* para o Cenário 1 com o modelo LightGBM.

Fonte: Elaborado pelo autor.

3.3.1 Organização e Estruturação de Dados para Análise Temporal

A fase de preparação e organização dos dados é crucial para garantir que as informações estejam no formato mais adequado para as análises subsequentes e para a aplicação de modelos preditivos. Inicialmente, os dados de produção, área e produtividade, coletados para cada safra, apresentavam-se em um formato sequencial, onde cada linha correspondia às métricas de um único ano agrícola. Embora essa estrutura seja padrão para armazenamento, ela não otimizava a visualização e a quantificação direta das variações interanuais, um requisito fundamental para compreender a dinâmica da produtividade do milho ao longo do tempo.

A principal inovação na organização dos dados consiste em apresentar, para cada "Safra"(doravante referida como "Ano Presente"), as métricas correspondentes da "Safra Anterior"na mesma linha de registro. Essa abordagem visa explicitamente facilitar a identificação de padrões de evolução, variações percentuais e a influência de condições passadas no desempenho atual da cultura. A transformação dos dados seguiu as seguintes etapas:

1. **Criação do *dataset* Base:** As informações iniciais de Área, Produção e Produtividade foram consolidadas para cada ano agrícola.
2. **Geração do *dataset* de Ano Passado:** Uma cópia idêntica do *dataset* base foi criada, e suas colunas de métricas (`Área`, `Produção`, `Produtividade`) foram renomeadas para incluir o sufixo "(Ano Passado)". A coluna "Safra"neste *dataset* foi então incrementada em um ano, efetivamente deslocando temporalmente os dados para que se alinhassem com o "Ano Presente"subsequente.
3. **Combinação Horizontal:** Os dois *datasets* (o original, representando o "Ano Presente", e a cópia renomeada e deslocada, representando o "Ano Passado") foram unidos horizontalmente com base na coluna "Safra". Essa operação resultou em um conjunto de dados expandido, onde cada linha contém agora as métricas de uma Safra e as métricas da Safra imediatamente anterior.
4. **Filtragem e Ordenação Final:** O conjunto de dados combinado foi então filtrado para abranger o período específico de análise de 2004 a 2009. Além disso, as colunas foram reordenadas para otimizar a visualização e a interpretabilidade, priorizando a "Safra", seguida pelas métricas do "Ano Presente"e, subsequentemente, pelas métricas do "Ano Passado"para cada variável (`Área`, `Produção` e `Produtividade`), conforme a ordem lógica da análise.

A Tabela 3.1 ilustra o processo que foi realizado para obter a estrutura final do conjunto de dados, demonstrando como esta organização facilita a comparação direta das métricas entre o Ano Presente e o Ano Passado. Por exemplo, é possível observar que a Área do Ano Presente para a safra de 2004 (83.000) se torna a Área do Ano Passado para a safra de 2005, evidenciando a continuidade e a dependência temporal dos dados.

Comparativo de Métricas Agrícolas por Safra (2004 - 2009)

Safra	Área (Ano Presente)	Área (Ano Passado)	Produção (Ano Presente)	Produção (Ano Passado)	Produtividade (Ano Presente)	Produtividade (Ano Passado)
2004	83.000	70.000	298.800	250.000	3.600	3.571
2005	66.240	83.000	179.532	298.800	2.710	3.600
2006	52.517	66.240	217.420	179.532	4.140	2.710
2007	99.433	52.517	416.680	217.420	4.191	4.140
2008	76.250	99.433	320.250	416.680	4.200	4.191
2009	12.400	76.250	46.680	320.250	3.765	4.200

Tabela 3.1. Comparativo de Métricas Agrícolas por Safra (2004 - 2009).
Fonte: Elaborado pelo autor.

Inicialmente, o estudo foi planejado para focar nos três principais municípios produtores de milho safrinha do estado de Mato Grosso: Sorriso, Nova Ubiratã e Nova Mutum. A primeira composição de dados, descrita a seguir, reflete essa abordagem inicial. Contudo, durante a fase de modelagem exploratória, constatou-se que o volume de dados restrito a apenas três municípios era insuficiente para treinar os algoritmos de aprendizado de máquina com a robustez necessária para obter resultados generalizáveis. Diante desta limitação, tomou-se a decisão metodológica de ampliar a abrangência do estudo, passando a incluir todos os municípios produtores do estado, a fim de construir um conjunto de dados mais rico e representativo para a análise subsequente.

- **Total de Amostras:** O conjunto de dados completo contém 51 amostras, calculadas a partir de 3 municípios com 17 observações cada ($3 \times 17 = 51$). As 17 observações se referem aos anos analisados (de 2005 à 2021).
- **Conjunto de Teste:** Corresponde a 20% do total, resultando em aproximadamente 11 amostras para teste.
- **Conjunto de Treino:** Composto pelas 40 amostras restantes ($51 - 11 = 40$).
- **Validação Cruzada (k-Fold):** O conjunto de treino de 40 amostras foi dividido em 5 folds, onde cada fold contém 8 amostras para o processo de validação do modelo.

O processo de preparação dos dados para a modelagem iniciou-se com a estruturação da matriz de entrada, composta por um total de 51 amostras, calculadas a partir de 3 municípios com 17 observações cada, e 27 variáveis (Figura 3.13). O primeiro passo metodológico consistiu na separação deste conjunto de dados em uma matriz de features (variáveis preditoras, X) e um vetor target (variável alvo, Y), que representa a produtividade do milho (Figura 3.14)

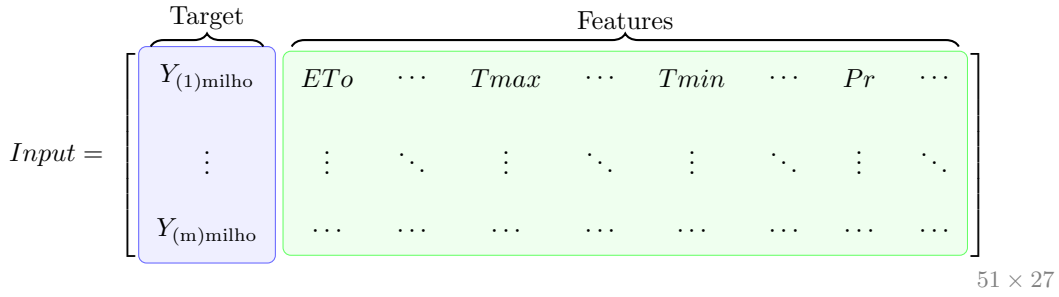


Figura 3.15. Diagrama inicial da matriz de entrada de dados.
Fonte: Elaborado pelo autor.

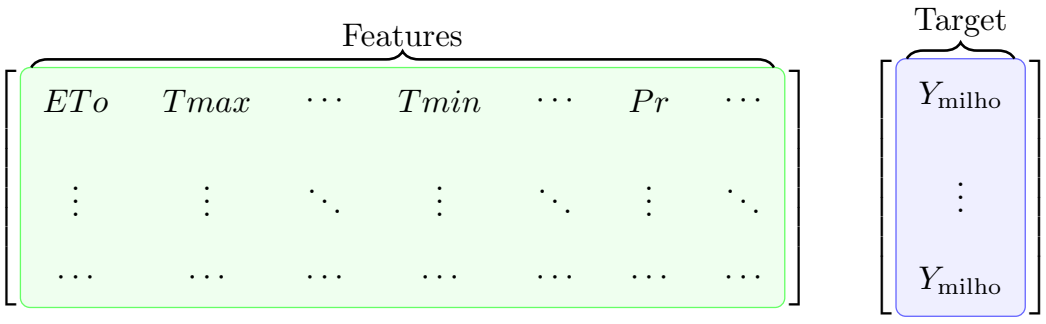


Figura 3.16. Separação da matriz em Target e Features.
Fonte: Elaborado pelo autor.

Subsequentemente, o conjunto de dados foi particionado em subconjuntos de treino e teste, uma etapa fundamental para a avaliação do modelo (Figura 3.15). O conjunto de teste foi composto por 20% do total de dados, resultando em aproximadamente 11 amostras que foram reservadas para a avaliação final do modelo. Os 80% restantes, totalizando 40 amostras, formaram o conjunto de treino, como visualizado na Figura 3.16.

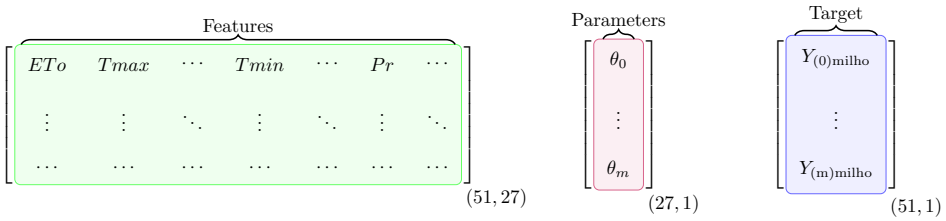


Figura 3.17. Divisão dos dados em conjuntos de treino e teste.

Figura 3.18. Visualização dos blocos de treino e teste.
Fonte: Elaborado pelo autor.

Para garantir uma avaliação robusta e evitar o superajuste (overfitting), uma estratégia de validação cruzada (k-Fold Cross-Validation) foi aplicada exclusivamente sobre o conjunto de treino. Este conjunto de 40 amostras foi dividido em 5 folds, onde cada fold contém 8 amostras (Figura 3.17). Este procedimento permite que o modelo seja treinado e validado múltiplas vezes em diferentes subconjuntos dos dados de treino, oferecendo uma estimativa mais confiável de seu desempenho.

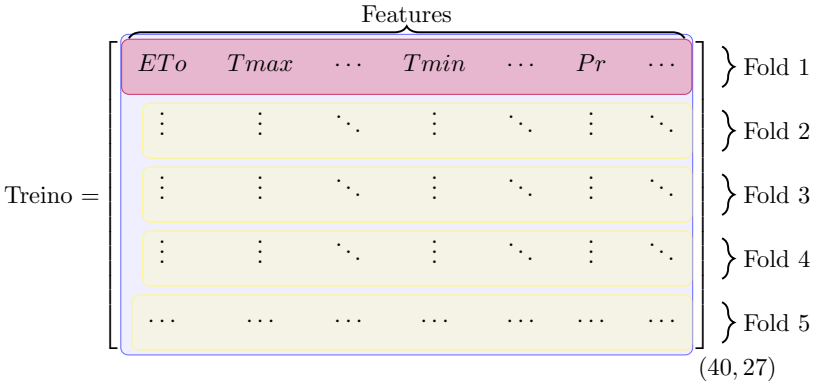


Figura 3.19. Aplicação da validação cruzada com 5 Folds.
Fonte: Elaborado pelo autor.

3.4 Base Teórica para a Necessidade de Dados: Uma Análise Sob a Ótica da Álgebra Linear e da Generalização em Machine Learning

Ao iniciar este estudo, a análise da relação entre o volume de dados disponíveis e o número de variáveis preditoras levantou uma questão fundamental sob a ótica da Álgebra Linear, que, em problemas práticos de Machine Learning, diverge da intuição inicial. O conjunto de dados preliminar, restrito a três municípios com 17 observações anuais cada, resultou em um total de 51 amostras. Para cada amostra, foram consideradas 27 variáveis preditoras, visando prever a produtividade do milho.

Em um sistema de equações lineares puramente teórico com m equações e n incógnitas, a relação entre m e n é crucial para a natureza das soluções. Um cenário com $m > n$, conhecido como sistema superdeterminado, é frequentemente abordado em Álgebra Linear. Para um modelo linear simples da forma

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n \tag{3.2}$$

onde Y é a produtividade e X_i são as variáveis preditoras, teríamos m equações (uma para cada amostra) e $n + 1$ incógnitas (os n coeficientes β_i e o intercepto β_0).

No nosso caso inicial, o conjunto de treino continha aproximadamente 40 amostras (após a separação inicial para teste), e o modelo linear correspondente teria $27 + 1 = 28$ incógnitas (features mais o intercepto). Portanto, o sistema para determinar os coeficientes de um modelo linear a partir dos dados de treino seria um sistema com $m = 40$ equações e $n = 28$ incógnitas, caracterizando um sistema superdeterminado ($m > n$).

Conforme discutido em (Anton e Rorres, 2012) (ver, por exemplo, Capítulo 1, Seção 1.6, sobre sistemas superdeterminados), um sistema linear com mais equações do que incógnitas geralmente não possui uma solução exata que satisfaça todas as equações simultaneamente, a menos que haja dependência linear específica entre as equações. No entanto, a teoria da Álgebra Linear oferece ferramentas para encontrar a "melhor" solução aproximada para tais sistemas inconsistentes, por meio do método de mínimos quadrados (ver (Anton e Rorres, 2012), Capítulo 6, Seção 6.4). Este método busca um vetor de coeficientes que minimiza o erro quadrático entre os valores preditos pelo modelo linear e os valores reais observados. A solução para um sistema $A\mathbf{x} = \mathbf{b}$ pelo método de mínimos quadrados é dada pelas equações normais:

$$A^T A\mathbf{x} = A^T \mathbf{b} \tag{3.3}$$

Se as colunas da matriz A forem linearmente independentes, a matriz $A^T A$ é invertível, garantindo uma solução única de mínimos quadrados (Teorema 6.4.4 em (Anton e Rorres, 2012)).

Essa base teórica da Álgebra Linear poderia, em uma análise superficial, sugerir que ter $m = 40$ amostras para $n = 28$ incógnitas seria suficiente para determinar um modelo (linear) de forma precisa, talvez até única no sentido de mínimos quadrados. No entanto, no contexto do Machine Learning aplicado a dados do mundo real, essa intuição da Álgebra Linear, focada em sistemas lineares exatos ou minimamente aproximados, precisa ser expandida para considerar a **natureza complexa, não-linear e ruidosa dos dados** e o objetivo principal do Machine Learning: a **generalização**.

Modelos de Machine Learning, especialmente aqueles flexíveis e não-lineares como *Random Forest*, *XGBoost* e *LightGBM* utilizados neste estudo, são capazes de aprender relações e interações complexas entre as variáveis que vão muito além de uma combinação linear simples. Para que esses modelos aprendam esses padrões complexos de forma confiável e, crucialmente, sejam capazes de fazer previsões precisas em dados *novos* que não foram vistos durante o treinamento (generalização), eles necessitam de um volume substancial de dados de treino.

Com apenas 40 amostras de treino para 27 variáveis preditoras, apesar de formalmente termos mais "equações"do que "incógnitas"em uma analogia linear simples, a quantidade de dados era **insuficiente**

para alimentar a complexidade dos modelos não-lineares e permitir que eles distinguissem os padrões preditivos reais do ruído aleatório presente no pequeno conjunto de dados. Isso aumenta drasticamente o risco de *overfitting*, onde o modelo memoriza as particularidades e o ruído do conjunto de treino em vez de aprender as relações generalizáveis.

Os resultados da análise preliminar, apresentados na Seção 4.1, confirmaram essa limitação, mostrando uma baixa capacidade preditiva mesmo após a otimização de hiperparâmetros. É importante ressaltar que essa seção estabelece a justificativa teórica e prática para a ampliação do conjunto de dados, que será detalhada na sequência da metodologia.

3.5 Pré-processamento

3.5.1 Transformação e Escalonamento de Dados

O pré-processamento dos dados é uma etapa crucial para garantir o desempenho ideal dos modelos de aprendizado de máquina. Nesta fase, foram aplicadas e avaliadas diferentes técnicas de escalonamento e transformação para ajustar as escalas das variáveis preditoras, cada uma com uma abordagem matemática distinta.

Para otimizar o desempenho dos modelos de aprendizado de máquina, foram empregadas duas técnicas principais para o ajuste de hiperparâmetros: a Busca em Grade (GridSearchCV) e a Busca Aleatória (RandomizedSearchCV).

A seguir, é apresentada a definição dos escalonadores utilizados para o pré-processamento dos dados, instanciando cada método de transformação para uso posterior no pipeline de modelagem:

```
# Pré-processamento
scalers = {
    'StandardScaler': StandardScaler(),
    'MinMaxScaler': MinMaxScaler(),
    'RobustScaler': RobustScaler(),
    'PowerTransformer': PowerTransformer()
}
```

Listing 3.1. Definição dos escalonadores para pré-processamento de dados.

Padronização (StandardScaler) A padronização transforma os dados para que tenham uma média igual a 0 e um desvio padrão igual a 1. Esta técnica é particularmente útil para algoritmos que assumem uma distribuição normal dos dados, como SVM e regressão linear (Géron, 2019). A fórmula é:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma} \tag{3.4}$$

onde μ é a média e σ é o desvio padrão da amostra.

Normalização (MinMaxScaler) A normalização ajusta os dados para um intervalo fixo, tipicamente entre 0 e 1. É uma técnica eficaz quando a distribuição dos dados não é gaussiana ou para algoritmos que não fazem suposições sobre a distribuição, como k-NN (Géron, 2019). A fórmula é:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{3.5}$$

onde $X_{\min}$ e $X_{\max}$ são os valores mínimo e máximo da característica.

Escalonamento Robusto (RobustScaler) Esta técnica é projetada para ser resistente a *outliers*. Ela utiliza a mediana e o intervalo interquartil (IQR) para escalar os dados, tornando-a ideal para conjuntos de dados com valores extremos (Géron, 2019). A fórmula é:

$$X_{\text{scaled}} = \frac{X - Q_2}{Q_3 - Q_1} \tag{3.6}$$

onde Q_1 é o primeiro quartil (25º percentil), Q_2 é a mediana (50º percentil) e Q_3 é o terceiro quartil (75º percentil).

Transformação de Potência (PowerTransformer) Diferente dos escalonadores, a transformação de potência visa estabilizar a variância e tornar a distribuição dos dados mais gaussiana. Isso é benéfico para muitos modelos que têm melhor desempenho com dados normalmente distribuídos. Uma das implementações mais conhecidas é a transformação de Box-Cox (Kukreti, 2021; Scikit-learn, 2024), definida como:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \ln(x) & \text{se } \lambda = 0 \end{cases} \tag{3.7}$$

onde o algoritmo encontra o valor ideal para o parâmetro λ que melhor normaliza os dados.

3.5.2 Melhorando os Hiperparâmetros

O **GridSearchCV** realiza uma busca exaustiva, testando sistematicamente todas as combinações de hiperparâmetros fornecidas em uma grade. Embora esta abordagem garanta que a melhor combinação dentro da grade seja encontrada, seu custo computacional pode ser muito elevado e ineficiente, especialmente quando alguns hiperparâmetros têm pouco impacto no resultado final.

Por outro lado, o **RandomizedSearchCV** testa um número fixo de combinações aleatórias dentro de um espaço de busca definido. Esta abordagem é computacionalmente mais eficiente e, conforme demonstrado por Bergstra e Bengio (2012) e ilustrado na Figura 3.20, a busca aleatória frequentemente encontra melhores hiperparâmetros com o mesmo número de iterações. Isso ocorre porque ela não desperdiça avaliações em dimensões pouco importantes do espaço de busca, aumentando a chance de explorar regiões mais promissoras.

Os dicionários a seguir definem os espaços de busca dos hiperparâmetros utilizados para otimização dos modelos. `param_grid` especifica as combinações testadas na Busca em Grade (GridSearchCV), e `param_dist` define o espaço para a Busca Aleatória (RandomizedSearchCV):

```
# Ajuste de hiperparâmetro
param_grid = {
    'n_estimators': [100, 200, 300, 400, 500],
    'max_depth': [1, 2, 3, 4, 5, 6, 7, 8, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
}

param_dist = {
    'n_estimators': [100, 200, 300, 400, 500],
    'max_depth': [1, 2, 3, 4, 5, 6, 7, 8, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
}
```

Listing 3.2. Dicionários de hiperparâmetros para GridSearchCV e RandomizedSearchCV.

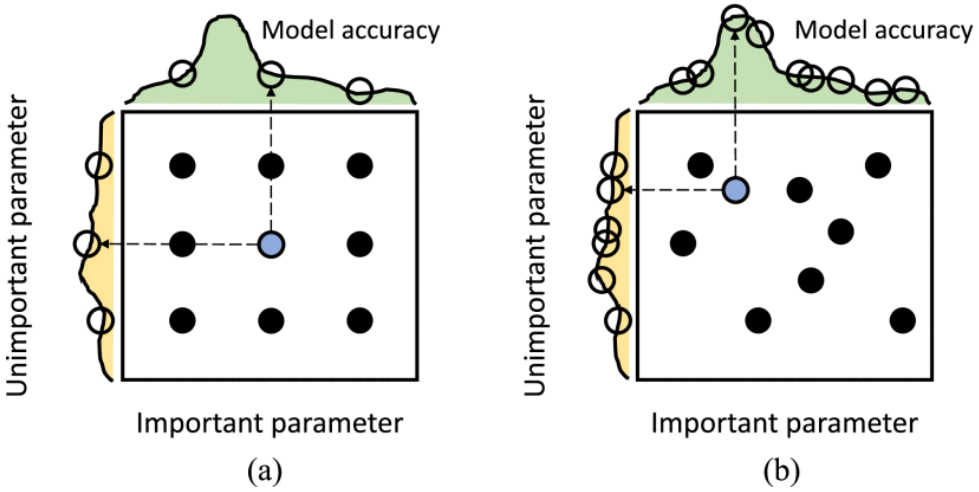


Figura 3.20. Ilustração da eficiência da Busca Aleatória (b) em comparação com a Busca em Grade (a) para a otimização de hiperparâmetros. A busca aleatória pode encontrar um modelo com maior acurácia ao não se prender a uma grade rígida. Adaptado de Pilario, Cao e Shafiee (2021) (Pilario *et al.*, 2021).

3.5.3 Resumo dos Modelos de Aprendizado de Máquina

Para a previsão da produtividade do milho, este trabalho explorou três algoritmos proeminentes baseados em árvores de decisão: *Random Forest*, *XGBoost* e *LightGBM*. A seleção desses modelos foi guiada por sua reconhecida eficácia em lidar com dados complexos e sua adequação a problemas de predição em larga escala no domínio agrícola.

- **Random Forest (Breiman, 2001):** Este é um método de *ensemble* que constrói múltiplas árvores de decisão de forma independente. Sua robustez advém da aleatoriedade introduzida na seleção de dados (*bagging*) e de *features* para a construção de cada árvore, o que minimiza a variância e o *overfitting*, tornando-o eficaz para dados com ruído e alta dimensionalidade.

- **XGBoost (Chen e Guestrin, 2016):** Uma implementação avançada de *Gradient Boosting*, o *XGBoost* se destaca pela sua performance e escalabilidade. Ele otimiza a construção de árvores sequenciais que corrigem erros passados, incorporando regularização para evitar o *overfitting* e eficientes mecanismos para lidar com esparsidade e grandes volumes de dados.

- **LightGBM (Ke *et al.*, 2017):** Representa a evolução mais recente e otimizada do *Gradient Boosting*. O *LightGBM* é notavelmente mais rápido e eficiente em termos de consumo de memória que seus predecessores, graças a técnicas inovadoras como a amostragem unilateral baseada em gradiente (*GOSS*) e o agrupamento exclusivo de características (*EFB*). Sua eficiência o torna ideal para análise de dados massivos em tempo hábil.

A análise comparativa desses modelos revelou que o *LightGBM* proporcionou o melhor desempenho para a previsão da produtividade do milho neste estudo, destacando sua adequação e eficiência para a tarefa proposta.

4 RESULTADOS

4.1 Análise Preliminar com Três Municípios

4.1.1 Seleção das melhores *features*

Para a Figura 4.1 (StandardScaler com GridSearchCV):

No modelo otimizado com `GridSearchCV`, a Evapotranspiração de Março (`ETo_MAR`) e a Precipitação de Maio (`pr_MAI`) surgem como as duas características mais influentes. Isso indica uma alta relevância das condições hídricas no final da estação para a previsão da produtividade, com o modelo aprendendo a dar peso a um balanço entre a demanda evaporativa e a oferta de chuva.

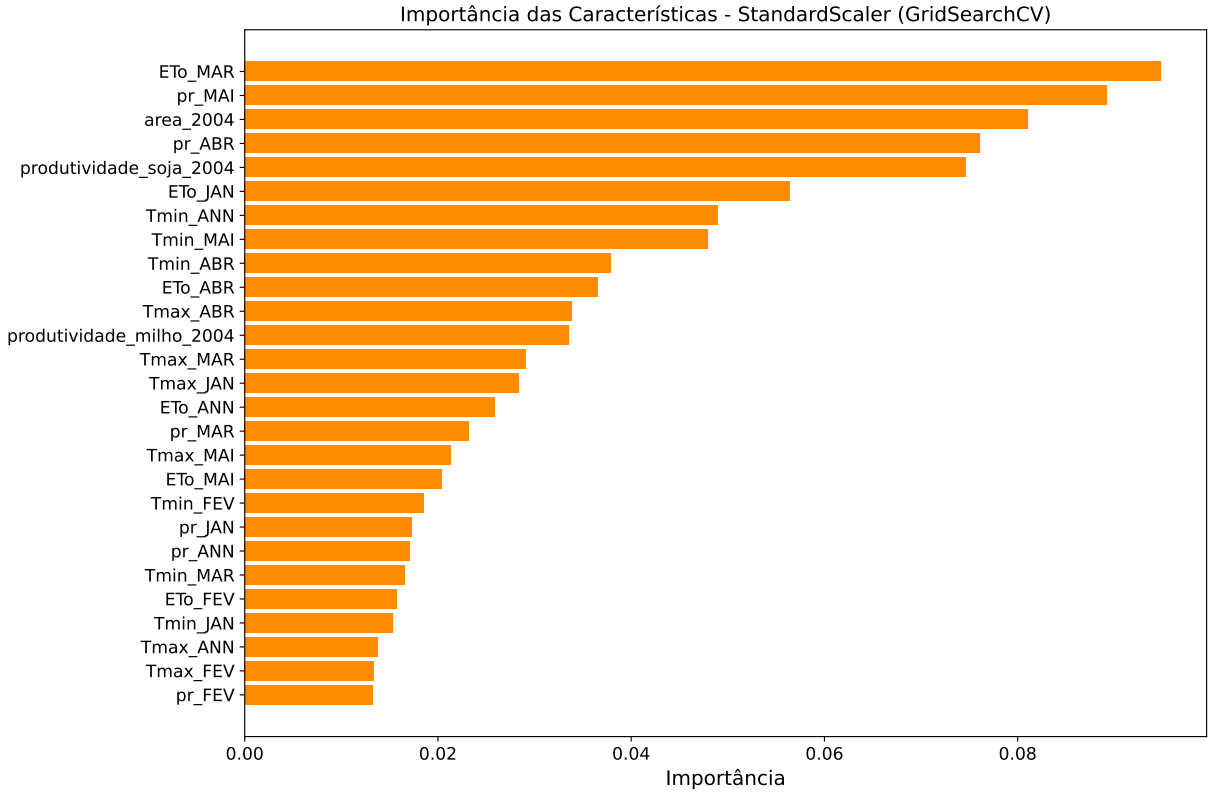


Figura 4.1. Importância das características no modelo otimizado com GridSearchCV.
Fonte: Elaborado pelo autor.

Para a Figura 4.2 (StandardScaler com RandomizedSearchCV):

Quando a otimização é feita com `RandomizedSearchCV`, a ordem de importância se altera sutilmente, com a Precipitação de Maio (`pr_MAI`) assumindo a primeira posição. Isso demonstra que, embora o mesmo grupo de variáveis climáticas seja consistentemente relevante, a técnica de otimização pode refinar a sensibilidade do modelo, neste caso, dando um peso ligeiramente maior para a chuva no final do período.

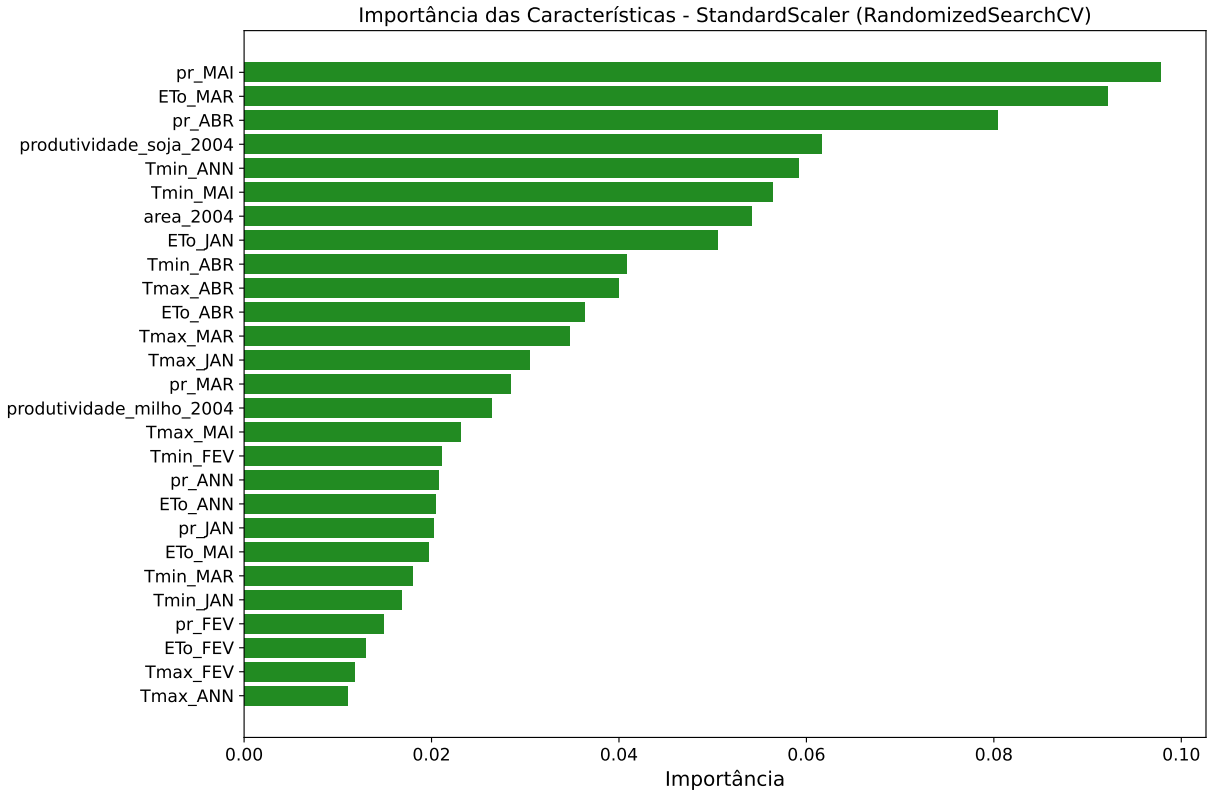


Figura 4.2. Importância das características no modelo otimizado com RandomizedSearchCV.
Fonte: Elaborado pelo autor.

Para a Figura 4.3 (StandardScaler Sem Ajuste):

Na versão sem ajuste de hiperparâmetros, a variável histórica `area_2004` torna-se desproporcionalmente importante, dominando completamente o modelo. Este resultado sugere que, sem a otimização, o modelo tende a uma solução mais simplista e possivelmente com menor capacidade de generalização, dependendo excessivamente de um único fator histórico em vez de aprender com as interações complexas das variáveis climáticas anuais.

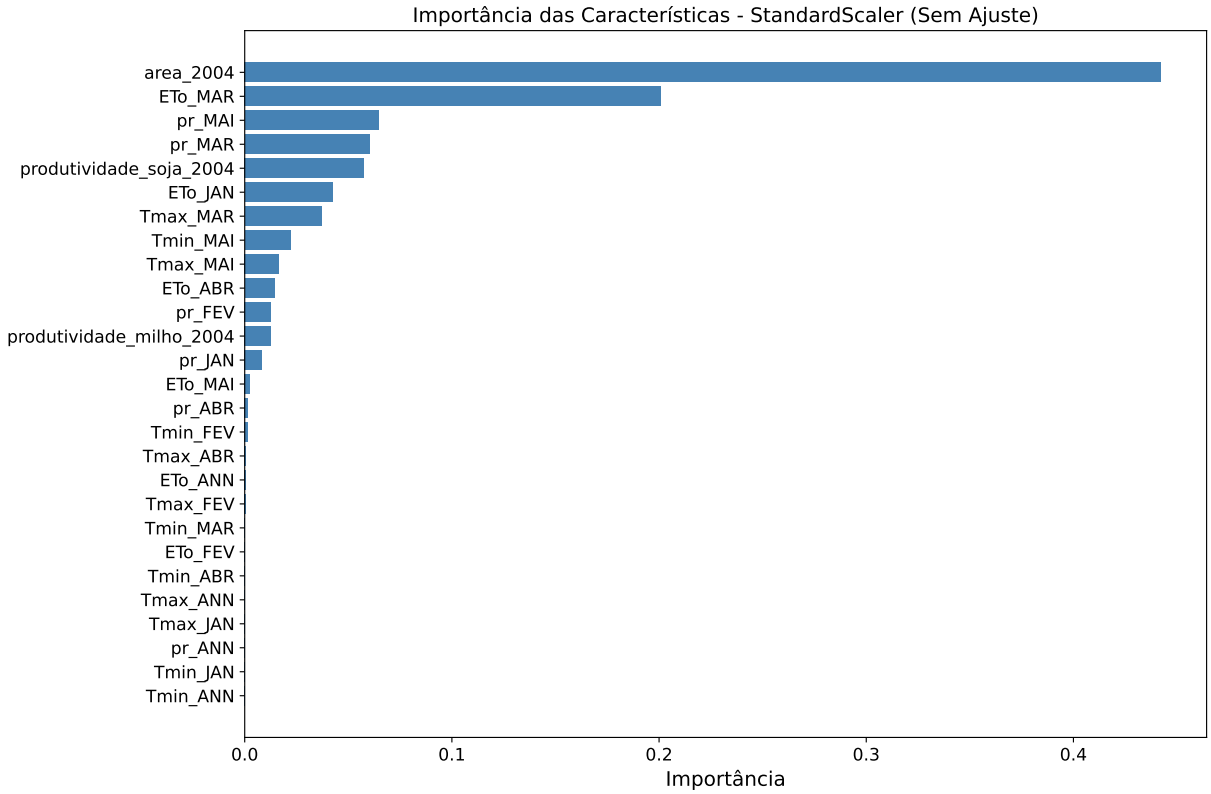


Figura 4.3. Importância das características no modelo sem ajuste de hiperparâmetros.
Fonte: Elaborado pelo autor.

Observação Geral:

Em conjunto, as figuras demonstram o papel crucial da otimização de hiperparâmetros. Os modelos otimizados (`GridSearchCV` e `RandomizedSearchCV`) conseguem identificar e balancear a importância de múltiplas variáveis climáticas, enquanto o modelo sem ajuste apresenta um comportamento mais "ingênuo", focando em uma única variável.

Em nítido contraste, na versão **sem ajuste** de hiperparâmetros (Figura 4.3), o modelo exibe um comportamento diferente. A variável histórica `area_2004` torna-se desproporcionalmente importante, dominando completamente a predição.

Essa comparação demonstra o papel crucial da otimização: os modelos ajustados aprendem a balancear as complexas interações das variáveis climáticas anuais, enquanto o modelo não otimizado recorre a uma solução mais simplista, com uma forte dependência de um único fator histórico.

A modelagem inicial, focada exclusivamente nos dados dos três principais municípios produtores, revelou limitações significativas que impediram a obtenção de um modelo preditivo robusto. A Tabela 4.1 resume as métricas de desempenho para os modelos treinados com este conjunto de dados restrito. Mesmo após a otimização de hiperparâmetros, o melhor resultado, obtido com `RandomizedSearchCV`, alcançou um coeficiente de determinação (R^2) de apenas $0,1979$, indicando que o modelo conseguia explicar menos de 20% da variabilidade na produtividade do milho.

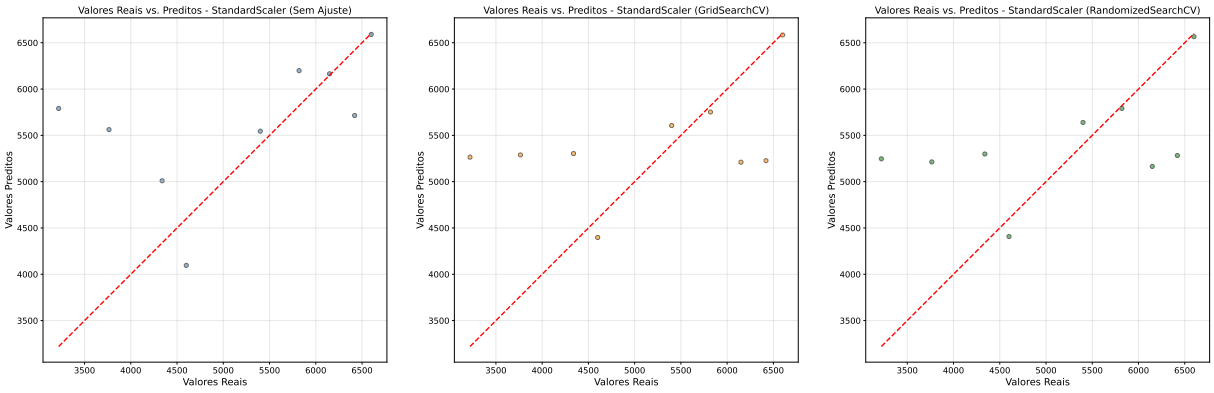
A baixa acurácia é visualmente confirmada nos gráficos de dispersão (Figura 4.4), que comparam os valores reais com os valores preditos pelo modelo. Para todas as configurações, os pontos se encontram amplamente dispersos ao redor da linha de referência ideal (onde os valores preditos seriam iguais aos reais), evidenciando uma alta margem de erro e baixa capacidade preditiva.

Concluiu-se que o limitado número de amostras (51) era o principal fator restritivo. Para superar essa limitação e construir um modelo mais preciso e generalizável, tomou-se a decisão metodológica de ampliar o escopo do estudo, utilizando os dados de todos os municípios produtores do estado de Mato

Grosso.

Tabela 4.1. Resultados dos modelos com diferentes configurações, utilizando dados de apenas 3 municípios.

Scaler	Hyperparameter Tuning	R^2	RMSE	MAE
StandardScaler	None	0,053	1115,680	755,290
StandardScaler	RandomizedSearchCV	0,198	1026,710	783,560
StandardScaler	GridSearchCV	0,169	1044,960	795,370



(a) Sem Ajuste (b) GridSearchCV (c) RandomizedSearchCV

Figura 4.4. Gráficos de dispersão entre valores reais e preditos para cada configuração de modelo, demonstrando a baixa acurácia com o conjunto de dados restrito.

Fonte: Elaborado pelo autor.

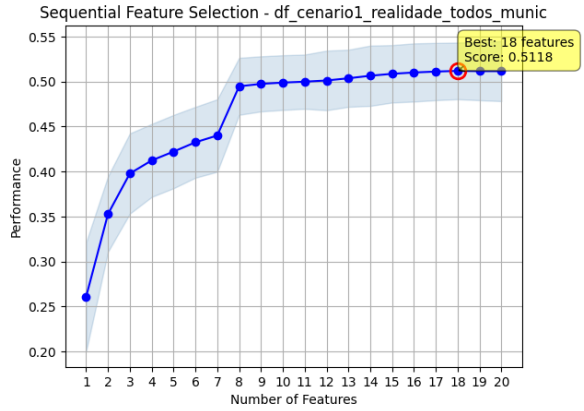
4.2 Análise Definitiva com o Conjunto de Dados Ampliado

A Figura 4.5 apresenta os resultados do processo de *Sequential Feature Selection* (SFS) para o modelo `Random Forest` em cada um dos quatro cenários de modelagem propostos. O objetivo desta análise foi determinar o número ideal de variáveis preditoras que maximiza o desempenho do modelo, evitando a inclusão de *features* redundantes ou com pouco poder preditivo.

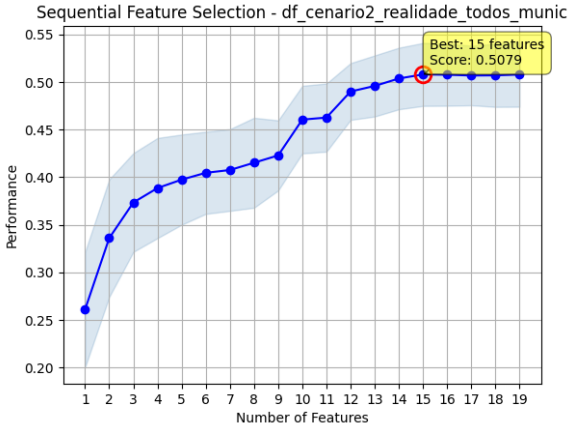
Em todos os cenários, observa-se um padrão consistente: a performance do modelo aumenta rapidamente com a adição das primeiras 5 a 7 *features* mais importantes. Após este ponto, o ganho de desempenho torna-se marginal, formando um "platô", o que indica que as variáveis adicionais contribuem pouco para a precisão do modelo.

Ao comparar os resultados máximos de cada cenário, o Cenário 1 (a) destacou-se como o mais eficaz, alcançando o maior *score* de desempenho $(0,5118)$ com um subconjunto de 18 *features*. Os cenários com menos meses de dados, Cenário 2 (b), Cenário 3 (c) e Cenário 4 (d), apresentaram uma performance inferior, com *scores* de $0,5079$, $0,4291$ e $0,3966$, respectivamente. Como era esperado, o Cenário 1 (a), que utiliza mais dados (incluindo o mês de maio), superou os demais cenários.

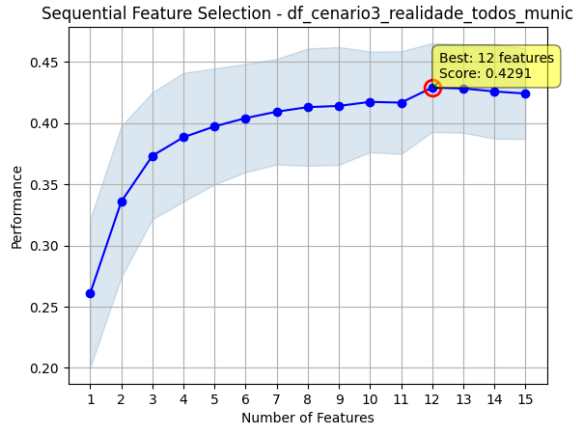
Esta análise justifica, portanto, a escolha do Cenário 1 como a janela temporal ótima e a seleção de suas 18 melhores *features* como a configuração ideal para o modelo `Random Forest`, equilibrando alta performance com maior simplicidade. É crucial destacar que este subconjunto de *features* otimizado, derivado da análise de SFS com o `Random Forest` no Cenário 1, foi subsequentemente aplicado como entrada para a avaliação de todos os demais modelos de Machine Learning (`Random Forest`, `XGBoost` e `LightGBM`) em seus respectivos cenários, visando uma base comparativa padronizada para a predição da produtividade do milho.



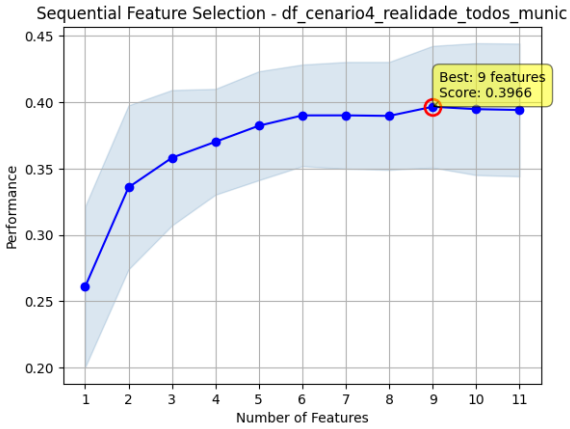
(a) Cenário 1



(b) Cenário 2



(c) Cenário 3



(d) Cenário 4

Figura 4.5. Importância das features para o modelo Random Forest nos quatro cenários de modelagem: (a) Cenário 1, (b) Cenário 2, (c) Cenário 3, e (d) Cenário 4.

Fonte: Elaborado pelo autor.

O ranking de classificação das melhores *features* para cada cenário com o Conjunto de Dados Ampliado é detalhado nas Tabelas 4.2 e 4.3, fornecendo *insights* sobre a importância relativa das variáveis ao longo do tempo.

Tabela 4.2. Ranking das melhores features por cenário (Cenários 1 e 2) para a predição da produtividade do milho com o Conjunto de Dados Ampliado.

Ranking	**Cenário 1**	**Cenário 2**
1º	produtividade_milho_2004	produtividade_milho_2004
2º	ETo_ABR	ETo_JAN
3º	ETo_JAN	Tmax_FEV
4º	area_2004	area_2004
5º	pr_ABR	Tmax_MAR
6º	ETo_MAR	Tmin_FEV
7º	Tmin_FEV	Tmin_ANN
8º	Tmin_ABR	Tmin_JAN
9º	pr_MAR	Tmin_ANN
10º	produtividade_soja_2004	Tmax_ANN

Fonte: Elaborado pelo autor.

Tabela 4.3. Ranking das melhores features por cenário (Cenários 3 e 4) para a predição da produtividade do milho com o Conjunto de Dados Ampliado.

Ranking	**Cenário 3**	**Cenário 4**
1º	produtividade_milho_2004	produtividade_milho_2004
2º	ETo_JAN	ETo_JAN
3º	produtividade_soja_2004	ETo_ANN
4º	area_2004	produtividade_soja_2004
5º	Tmax_ANN	area_2004
6º	Tmin_FEV	Tmin_JAN
7º	ETo_FEV	Tmax_ANN
8º	ETo_ANN	Tmax_JAN
9º	Tmax_JAN	pr_ANN
10º	Tmin_ANN	pr_JAN

Fonte: Elaborado pelo autor.

O desempenho do modelo final, utilizando o algoritmo `LightGBM` na configuração otimizada (Cenário 1, com *features* selecionadas, escalonamento `MinMaxScaler` e ajuste com `GridSearchCV`), é apresentado na Figura 4.6. O gráfico de dispersão compara os valores de produtividade preditos pelo modelo (eixo Y) com os valores reais observados (eixo X).

Visualmente, nota-se que as predições (pontos azuis) se agrupam em torno da linha de predição perfeita (linha tracejada preta), indicando uma forte correlação positiva e a capacidade do modelo de capturar a tendência geral dos dados. A linha de regressão (vermelha) resume essa tendência.

Quantitativamente, o modelo alcançou um coeficiente de determinação (R^2) de $0,609$, o que significa que ele consegue explicar aproximadamente 61% da variabilidade na produtividade do milho. O Erro Médio Absoluto (MAE) foi de $630,01$ kg/ha, representando o desvio médio das previsões em relação aos valores reais. Estes resultados consolidam o `LightGBM`, dentro da metodologia aplicada, como a abordagem mais robusta para a previsão da produtividade do milho safrinha na região de estudo.

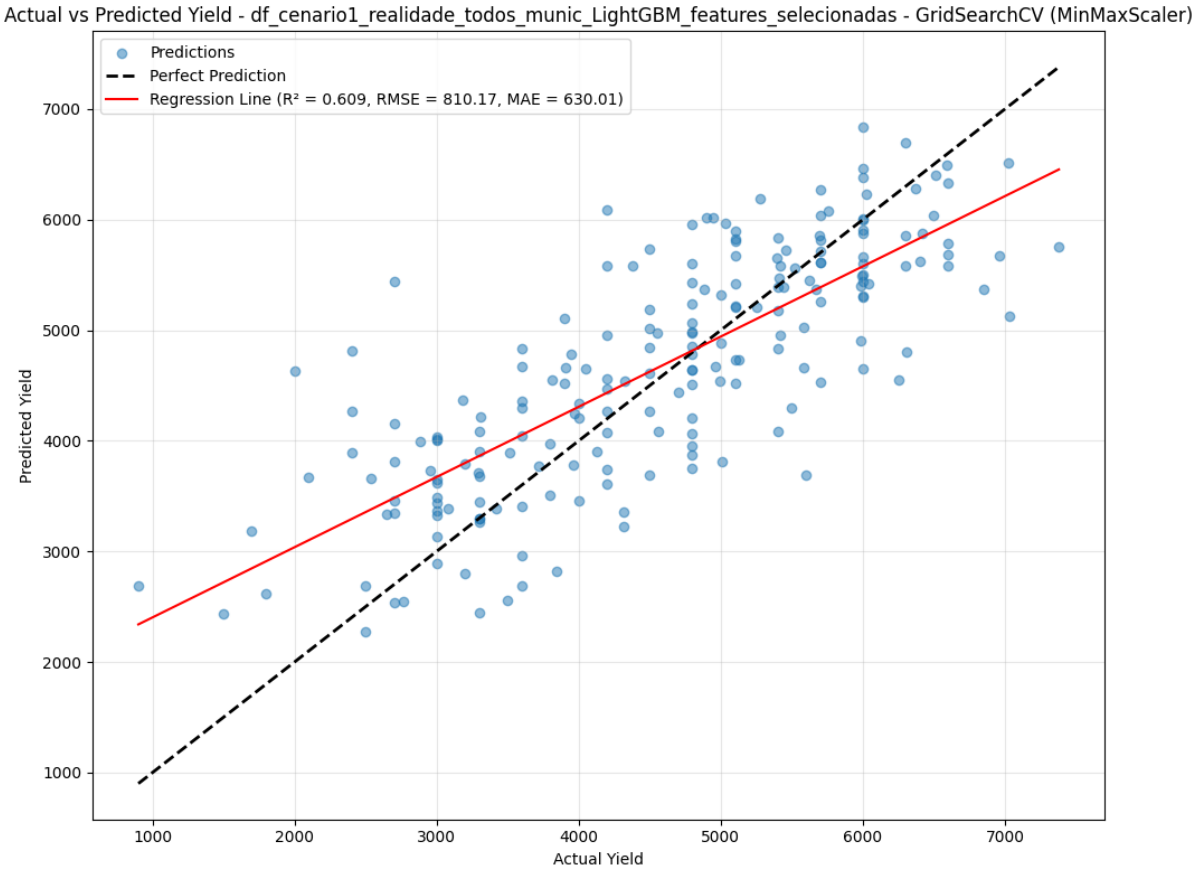


Figura 4.6. Gráfico de dispersão entre a produtividade real e a predita pelo melhor modelo (`LightGBM`). A linha tracejada representa a predição perfeita ($y = x$) e a linha vermelha, a regressão linear dos pontos.

Fonte: Elaborado pelo autor.

Os resultados apresentados consolidam o LightGBM como a abordagem mais robusta para a previsão da produtividade do milho safrinha na região de estudo. Para uma visão comparativa dos melhores desempenhos alcançados por cada modelo avaliado (Random Forest, XGBoost e LightGBM) em suas configurações otimizadas, a Tabela 4.4 sumariza os principais resultados. Detalhes completos de todas as combinações de cenários, pré-processamento e ajuste de hiperparâmetros para cada modelo podem ser consultados nas Tabelas A.1, A.2 e A.3 no Apêndice I.

Tabela 4.4. Melhores resultados de desempenho dos modelos avaliados na previsão da produtividade do milho.

Modelo	Cenário	Pré-processamento	Tunning	R^2	MAE (kg/ha)
Random Forest	1	RobustScaler	GridSearchCV	0.563	645.69
XGBoost	1	StandardScaler	GridSearchCV	0.598	631.02
LightGBM	1	MinMaxScaler	GridSearchCV	0.609	630.01

Fonte: Adaptado das Tabelas A.1, A.2 e A.3 do Apêndice I.

5 CONCLUSÃO

Este trabalho demonstrou com sucesso o desenvolvimento e a validação de um modelo de inteligência artificial, baseado no algoritmo `LightGBM`, para a previsão da produtividade do milho safrinha em Mato Grosso. O modelo final alcançou um desempenho robusto, explicando aproximadamente 61% da variabilidade da produtividade ($R^2 = 0,61$) e apresentando um erro médio absoluto de $630,01$ kg/ha, validando a metodologia proposta como uma ferramenta eficaz para o planejamento agrícola e a mitigação de riscos.

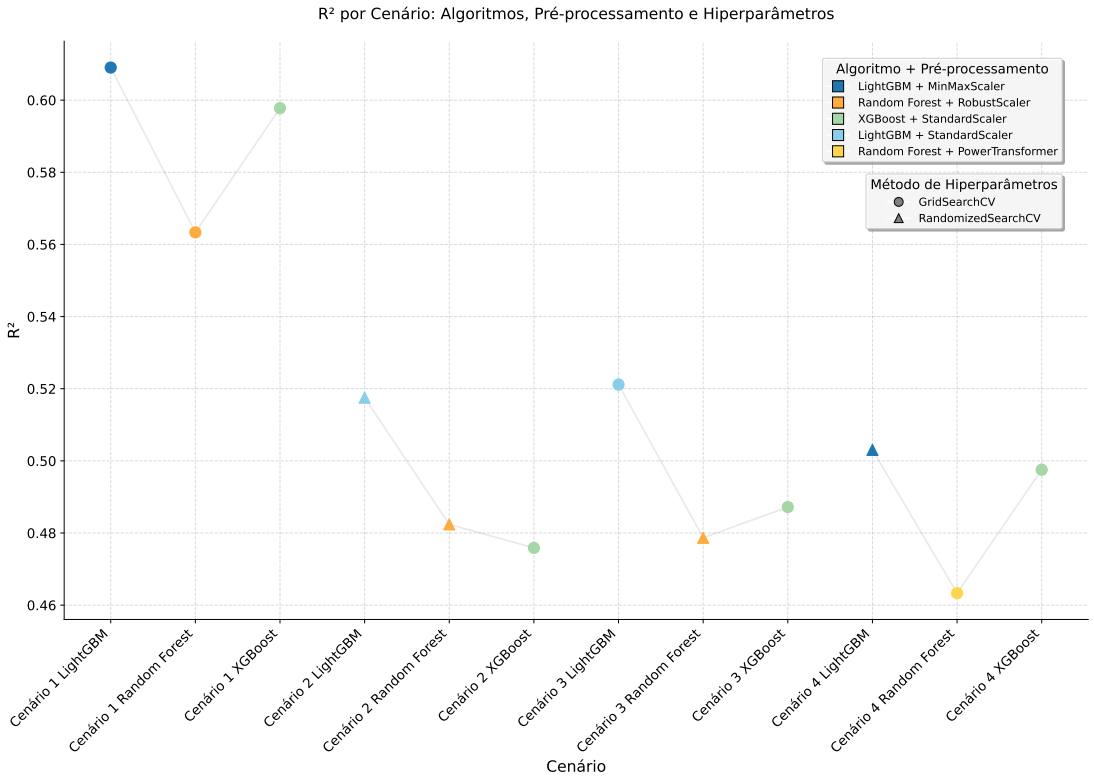


Figura 5.1. Métrica de avaliação R^2 por Cenário para todos os modelos.
Fonte: Elaborado pelo autor.

A análise aprofundada das variáveis, detalhada no ranking de features apresentado nas Tabelas 4.2 e 4.3, revelou padrões claros de influência na produtividade. Consistentemente em todos os cenários analisados, a produtividade_milho_2004 (produtividade da safra anterior) emergiu como a *feature* mais relevante, indicando uma forte dependência histórica na produção. Além disso, indicadores de demanda hídrica como a Evapotranspiração de Referência (ETo) nos meses do ano safra (Janeiro a Abril, com ETo_JAN e ETo_ABR frequentemente entre as posições de destaque) são preditores de alta influência. Outras variáveis climáticas, como a Temperatura Máxima (Tmax_FEV, Tmax_MAR, Tmax_ANN) e Mínima (Tmin_FEV, Tmin_JAN, Tmin_ANN), bem como a Precipitação (pr_ABR, pr_MAR, pr_ANN) e a area_2004 (área plantada na safra anterior), também demonstraram relevância, embora com variações no ranking entre os cenários.

Essa descoberta tem implicações práticas diretas, sugerindo que o monitoramento tanto dos resultados de safras anteriores quanto das condições hídricas e térmicas nos meses que antecedem o plantio pode servir como um valioso indicativo precoce para variações na produtividade. Isso oferece tempo hábil para a tomada de decisão por parte de produtores e gestores, permitindo a implementação de estratégias de manejo e mitigação de riscos.

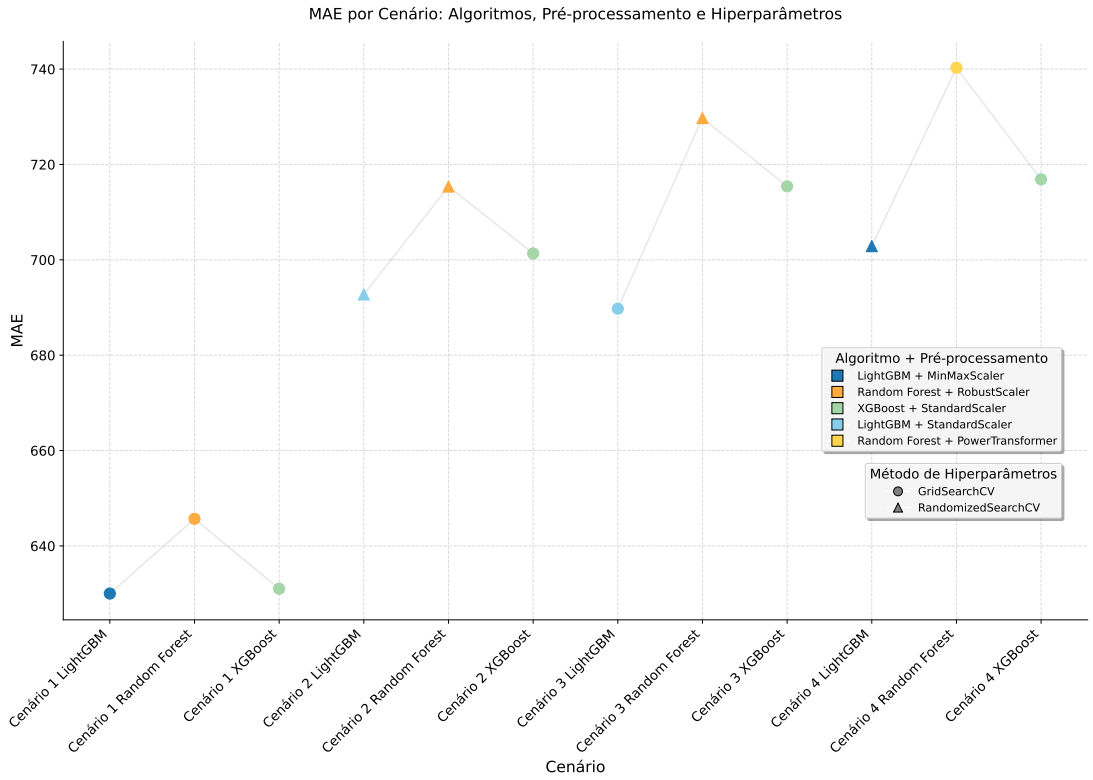


Figura 5.2. Métrica de avaliação MAE por Cenário para todos os modelos.
Fonte: Elaborado pelo autor.

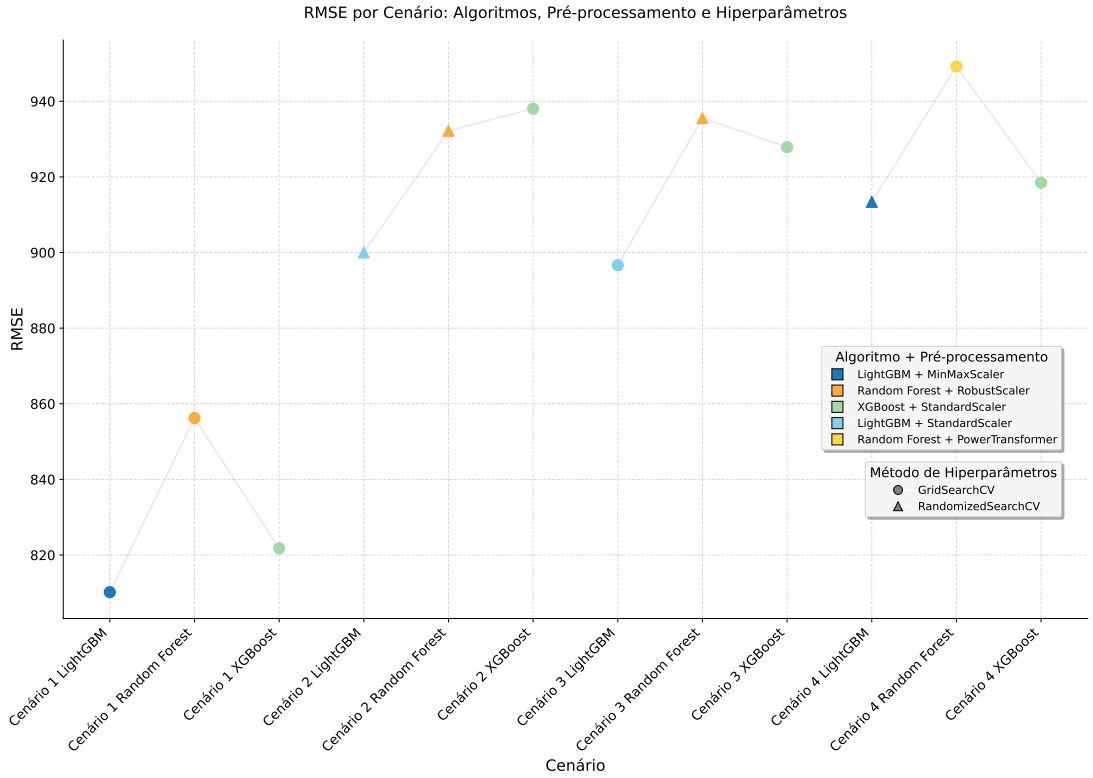


Figura 5.3. Métrica de avaliação RMSE por Cenário para todos os modelos.
Fonte: Elaborado pelo autor.

Para as 3 métricas analisadas o Cenário 3 (Jan-Mar) teve seu melhor modelo superando a eficiência do Cenário 2 (Jan-Abr), mesmo com um mês a menos de dados.

Metodologicamente, o estudo confirmou a importância da otimização de hiperparâmetros e da utilização de um conjunto de dados abrangente, uma vez que os modelos iniciais, com dados limitados a apenas três municípios, se mostraram insuficientes. Como trabalhos futuros, sugere-se a comparação do `LightGBM` com outros algoritmos de ensemble e a aplicação do modelo em outras regiões produtoras, a fim de validar e expandir sua aplicabilidade.

REFERÊNCIAS

Allen, R. G., L. S. Pereira, D. Raes, e M. Smith, 1998 *Crop Evapotranspiration: Guidelines for Computing Crop Water Requirements*. Number 56 in FAO Irrigation and Drainage Paper, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, ISBN 92-5-104219-5.

Anton, H. e C. Rorres, 2012 *Álgebra Linear com Aplicações*. Bookman, Porto Alegre, 10th edition.

Aparecido, L. E. d. O., G. B. Torsoni, D. Z. Mesquita, K. C. d. Meneses, e J. R. D. S. C. d. Moraes, 2020 Modelagem da produtividade do milho safrinha em função das condições climáticas do mato grosso do sul. Revista Brasileira de Climatologia **26**.

Bergamaschi, H. e R. Matzenauer, 2014 *O milho e o clima*. EMATER/RS-ASCAR, first edition.

Breiman, L., 2001 Random forests. Machine Learning **45**: 5–32.

Charles, H., J. Godfray, J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir, J. Pretty, S. Robinson, S. M. Thomas, e C. Toulmin, 2010 Food security: The challenge of feeding 9 billion people.

Chen, T. e C. Guestrin, 2016 Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, San Francisco, CA, USA, ACM.

de Souza, T. T., 2018 Simulação de cenários agrícolas futuros para a cultura do milho no brasil com bases em projeções de mudanças climáticas.

Géron, A., 2019 *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas para a Construção de Sistemas Inteligentes*. O'Reilly Media.

Guzzon, F., L. W. A. Rios, G. M. C. Cepeda, M. C. Polo, A. C. Cabrera, J. M. Figueroa, A. E. M. Hoyos, T. W. J. Calvo, T. L. Molnar, L. A. N. León, T. P. N. León, S. L. M. Kerguelén, J. G. O. Rojas, G. Vázquez, R. E. Preciado-Ortiz, J. L. Zambrano, N. P. Rojas, e K. V. Pixley, 2021 Conservation and use of latin american maize diversity: Pillar of nutrition security and cultural heritage of humanity. Agronomy **11**.

Haykin, S., 2001 *Redes Neurais: princípios e práticas*. Bookman, second edition.

IMEA - Instituto Mato-Grossense de Economia Agropecuária, 2023 Informativo IMEA Matogrossense. Boletim Informativo, Dados e projeções do agronegócio em Mato Grosso, com estimativas até maio de 2023.

Jones, P. G. e P. K. Thornton, 2003 The potential impacts of climate change on maize production in africa and latin america in 2055. Global Environmental Change **13**: 51–59.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, e T.-Y. Liu, 2017 Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 3146–3154, Curran Associates, Inc.

Khashei-Siuki, A., M. Kouchakzadeh, e B. Ghahraman, 2011 Predicting dryland wheat yield from meteorological data using expert system, khorasan province, iran. J. Agr. Sci. Tech. **13**: 627–640.

Kukreti, A., 2021 Power transformers in-depth understanding. Kaggle, Disponível em: `kaggle.com/code/abhikuks/power-transformers-in-depth-understanding#Box-Cox-Transformation`. Acesso em: 20 nov. 2024.

Marin, F. e D. S. P. Nassif, 2013 Mudanças climáticas e a cana-de-açúcar no brasil: Fisiologia, conjuntura e cenário futuro.

Monfreda, C., N. Ramankutty, e J. A. Foley, 2008 Farming the planet: 2. geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. Global Biogeochemical Cycles **22**.

Nyéki, A., C. Kerepesi, B. Daróczy, A. A. Benczúr, G. Milics, A. J. Kovács, e M. Neményi, 2019 Maize yield prediction based on artificial intelligence using spatio-temporal data. In *Precision Agriculture '19*, edited by J. V. Stafford, pp. 1011–1017, Wageningen Academic Publishers, Wageningen, The Netherlands.

Peres, T. C. e . L. B. Maier, 2019 Análise da variabilidade espaço-temporal da precipitação no brasil: Soi e pdo. In *Anais do XIX Simpósio Brasileiro de Sensoriamento Remoto*, pp. 1–11, Santos, SP, Instituto Nacional de Pesquisas Espaciais, Realizado de 14 a 17 de abril de 2019.

Pilario, K. E. S., Y. Cao, e M. Shafiee, 2021 A kernel design approach to improve kernel subspace identification. IEEE Transactions on Industrial Electronics **68**: 6171–6181.

Pinheiro, J. A. d., 2004 Modelo estocástico para estimação de produtividade potencial de milho em piracicaba-sp.

Santos, M. V. d., 2000 Relatório técnico consolidado de clima para o estado de mato grosso.

Scikit-learn, 2024 Power transform. scikit-learn.org, Disponível em: `scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.power_transform.html`. Acesso em: 20 nov. 2024.

Suresh, N., N. V. Ramesh, S. Inthiyaz, P. P. Priya, K. Nagasowmika, K. V. Kumar, M. Shaik, e B. N. Reddy, 2021 Crop yield prediction using random forest algorithm. In *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, pp. 279–282, Institute of Electrical and Electronics Engineers Inc.

Tanumihardjo, S. A., L. Mcculley, R. Roh, S. Lopez-Ridaura, N. Palacios-Rojas, e N. S. Gunaratna, 2019 Maize agro-food systems to ensure food and nutrition security in reference to the sustainable development goals. Agronomy .

Vieira, P. A., E. C. Gilmar, P. Henz, V. G. D. Caldas, e N. E. Técnicos, 2019 Geopolítica do alimento o brasil como fonte estratégica de alimentos para a humanidade.

Vieira Junior, P. A., 2006 *Previsão de atributos do clima e do rendimento de grãos de milho na região Centro-Sul do Brasil*. Ph.D. thesis, Escola Superior de Agricultura "Luiz de Queiroz"(ESALQ), Universidade de São Paulo, Piracicaba, Tese apresentada para a obtenção do título de Doutor em Agronomia. Área de concentração: Fitotecnia.

APÊNDICES

Apêndice I

Tabela A.1. Resultados Detalhados do Modelo Random Forest (RF) por Cenário, Pré-processamento e Ajuste de Tunning.

Cenário	Pré-processamento	Tunning	R^2	RMSE	MAE
1-RF	StandardScaler	None	0,553	865,820	656,593
1-RF	StandardScaler	RS-CV	0,562	857,087	646,366
1-RF	StandardScaler	GS-CV	0,563	856,322	646,311
1-RF	MinMaxScaler	None	0,553	866,393	657,686
1-RF	MinMaxScaler	RS-CV	0,562	857,361	646,817
1-RF	MinMaxScaler	GS-CV	0,563	856,752	646,663
1-RF	RobustScaler	None	0,553	866,107	656,277
1-RF	RobustScaler	RS-CV	0,563	856,963	646,054
1-RF	RobustScaler	GS-CV	0,563	856,210	645,686
1-RF	PowerTransformer	None	0,552	866,907	658,870
1-RF	PowerTransformer	RS-CV	0,562	857,269	647,124
1-RF	PowerTransformer	GS-CV	0,563	856,652	646,674
2-RF	StandardScaler	None	0,478	935,950	715,298
2-RF	StandardScaler	RS-CV	0,482	932,584	715,667
2-RF	StandardScaler	GS-CV	0,476	937,608	719,355
2-RF	MinMaxScaler	None	0,478	935,787	715,319
2-RF	MinMaxScaler	RS-CV	0,482	932,407	715,453
2-RF	MinMaxScaler	GS-CV	0,477	937,417	719,429
2-RF	RobustScaler	None	0,479	934,837	714,567
2-RF	RobustScaler	RS-CV	0,482	932,207	715,367
2-RF	RobustScaler	GS-CV	0,477	936,916	718,807
2-RF	PowerTransformer	None	0,478	935,943	716,343
2-RF	PowerTransformer	RS-CV	0,482	932,277	715,731
2-RF	PowerTransformer	GS-CV	0,477	937,157	719,345
3-RF	StandardScaler	None	0,467	946,123	736,775
3-RF	StandardScaler	RS-CV	0,478	935,853	729,936
3-RF	StandardScaler	GS-CV	0,472	941,216	740,906
3-RF	MinMaxScaler	None	0,466	947,172	738,074
3-RF	MinMaxScaler	RS-CV	0,478	935,983	730,400
3-RF	MinMaxScaler	GS-CV	0,472	941,659	741,652
3-RF	RobustScaler	None	0,468	945,230	736,152
3-RF	RobustScaler	RS-CV	0,479	935,545	729,740
3-RF	RobustScaler	GS-CV	0,473	940,714	740,502
3-RF	PowerTransformer	None	0,466	947,291	738,411
3-RF	PowerTransformer	RS-CV	0,478	935,935	730,448
3-RF	PowerTransformer	GS-CV	0,472	941,802	741,649
4-RF	StandardScaler	None	0,430	978,544	751,293
4-RF	StandardScaler	RS-CV	0,450	960,999	747,788
4-RF	StandardScaler	GS-CV	0,453	958,449	749,206
4-RF	MinMaxScaler	None	0,430	978,283	751,068
4-RF	MinMaxScaler	RS-CV	0,446	964,398	747,845
4-RF	MinMaxScaler	GS-CV	0,463	949,280	739,615
4-RF	RobustScaler	None	0,430	978,243	750,468
4-RF	RobustScaler	RS-CV	0,446	964,201	747,391
4-RF	RobustScaler	GS-CV	0,453	958,051	748,934
4-RF	PowerTransformer	None	0,431	977,626	749,719
4-RF	PowerTransformer	RS-CV	0,450	960,660	748,468
4-RF	PowerTransformer	GS-CV	0,463	949,231	740,283

Fonte: Elaborado pelo autor.

Tabela A.2. Resultados Detalhados do Modelo XGBoost (XGB) por Cenário, Pré-processamento e Ajuste de Tunning.

Cenário	Pré-processamento	Tunning	R^2	RMSE	MAE
1-XGB	StandardScaler	None	0,519	899,086	688,587
1-XGB	StandardScaler	RS-CV	0,583	836,530	640,860
1-XGB	StandardScaler	GS-CV	0,598	821,764	631,023
1-XGB	MinMaxScaler	None	0,519	899,086	688,587
1-XGB	MinMaxScaler	RS-CV	0,583	836,530	640,860
1-XGB	MinMaxScaler	GS-CV	0,598	821,764	631,023
1-XGB	RobustScaler	None	0,519	899,086	688,587
1-XGB	RobustScaler	RS-CV	0,583	836,530	640,860
1-XGB	RobustScaler	GS-CV	0,598	821,764	631,023
1-XGB	PowerTransformer	None	0,519	899,086	688,587
1-XGB	PowerTransformer	RS-CV	0,583	836,530	640,860
1-XGB	PowerTransformer	GS-CV	0,598	821,764	631,023
2-XGB	StandardScaler	None	0,409	996,178	765,983
2-XGB	StandardScaler	RS-CV	0,476	938,065	720,482
2-XGB	StandardScaler	GS-CV	0,476	938,052	701,336
2-XGB	MinMaxScaler	None	0,409	996,178	765,983
2-XGB	MinMaxScaler	RS-CV	0,476	938,065	720,482
2-XGB	MinMaxScaler	GS-CV	0,476	938,052	701,336
2-XGB	RobustScaler	None	0,409	996,178	765,983
2-XGB	RobustScaler	RS-CV	0,476	938,065	720,482
2-XGB	RobustScaler	GS-CV	0,476	938,052	701,336
2-XGB	PowerTransformer	None	0,409	996,178	765,983
2-XGB	PowerTransformer	RS-CV	0,476	938,065	720,482
2-XGB	PowerTransformer	GS-CV	0,476	938,052	701,336
3-XGB	StandardScaler	None	0,435	973,937	759,202
3-XGB	StandardScaler	RS-CV	0,474	939,297	725,372
3-XGB	StandardScaler	GS-CV	0,487	927,872	715,409
3-XGB	MinMaxScaler	None	0,435	973,937	759,202
3-XGB	MinMaxScaler	RS-CV	0,474	939,297	725,372
3-XGB	MinMaxScaler	GS-CV	0,487	927,872	715,409
3-XGB	RobustScaler	None	0,435	973,937	759,202
3-XGB	RobustScaler	RS-CV	0,474	939,297	725,372
3-XGB	RobustScaler	GS-CV	0,487	927,872	715,409
3-XGB	PowerTransformer	None	0,435	973,937	759,202
3-XGB	PowerTransformer	RS-CV	0,474	939,297	725,372
3-XGB	PowerTransformer	GS-CV	0,487	927,872	715,409
4-XGB	StandardScaler	None	0,385	1015,952	763,200
4-XGB	StandardScaler	RS-CV	0,480	934,706	731,194
4-XGB	StandardScaler	GS-CV	0,498	918,464	716,876
4-XGB	MinMaxScaler	None	0,385	1015,952	763,200
4-XGB	MinMaxScaler	RS-CV	0,480	934,706	731,194
4-XGB	MinMaxScaler	GS-CV	0,498	918,464	716,876
4-XGB	RobustScaler	None	0,385	1015,952	763,200
4-XGB	RobustScaler	RS-CV	0,480	934,706	731,194
4-XGB	RobustScaler	GS-CV	0,498	918,464	716,876
4-XGB	PowerTransformer	None	0,385	1015,952	763,200
4-XGB	PowerTransformer	RS-CV	0,480	934,706	731,194
4-XGB	PowerTransformer	GS-CV	0,498	918,464	716,876

Fonte: Elaborado pelo autor.

Tabela A.3. Resultados Detalhados do Modelo LightGBM (LGBM) por Cenário, Pré-processamento e Ajuste de Tunning.

Cenário	Pré-processamento	Tunning	R^2	RMSE	MAE
1-LGBM	StandardScaler	None	0,582	837,538	648,444
1-LGBM	StandardScaler	RS-CV	0,597	822,827	632,980
1-LGBM	StandardScaler	GS-CV	0,599	821,001	634,992
1-LGBM	MinMaxScaler	None	0,580	839,353	646,516
1-LGBM	MinMaxScaler	RS-CV	0,586	833,843	645,634
1-LGBM	MinMaxScaler	GS-CV	0,609	810,169	630,013
1-LGBM	RobustScaler	None	0,565	854,453	651,922
1-LGBM	RobustScaler	RS-CV	0,601	818,192	638,445
1-LGBM	RobustScaler	GS-CV	0,583	836,327	645,990
1-LGBM	PowerTransformer	None	0,578	842,214	648,536
1-LGBM	PowerTransformer	RS-CV	0,595	824,452	638,308
1-LGBM	PowerTransformer	GS-CV	0,596	823,932	641,584
2-LGBM	StandardScaler	None	0,481	933,849	716,210
2-LGBM	StandardScaler	RS-CV	0,518	900,019	692,733
2-LGBM	StandardScaler	GS-CV	0,480	934,126	713,986
2-LGBM	MinMaxScaler	None	0,479	935,142	708,634
2-LGBM	MinMaxScaler	RS-CV	0,516	901,588	690,185
2-LGBM	MinMaxScaler	GS-CV	0,510	907,028	695,298
2-LGBM	RobustScaler	None	0,474	940,014	720,999
2-LGBM	RobustScaler	RS-CV	0,515	902,580	691,107
2-LGBM	RobustScaler	GS-CV	0,513	903,774	696,429
2-LGBM	PowerTransformer	None	0,475	938,937	720,591
2-LGBM	PowerTransformer	RS-CV	0,507	909,406	697,885
2-LGBM	PowerTransformer	GS-CV	0,505	911,937	695,383
3-LGBM	StandardScaler	None	0,445	964,928	722,382
3-LGBM	StandardScaler	RS-CV	0,507	909,596	697,243
3-LGBM	StandardScaler	GS-CV	0,521	896,635	689,749
3-LGBM	MinMaxScaler	None	0,463	949,315	716,637
3-LGBM	MinMaxScaler	RS-CV	0,510	907,035	693,445
3-LGBM	MinMaxScaler	GS-CV	0,502	914,495	708,361
3-LGBM	RobustScaler	None	0,444	966,255	728,223
3-LGBM	RobustScaler	RS-CV	0,510	906,773	694,151
3-LGBM	RobustScaler	GS-CV	0,499	916,927	699,508
3-LGBM	PowerTransformer	None	0,459	953,112	711,929
3-LGBM	PowerTransformer	RS-CV	0,510	906,598	691,110
3-LGBM	PowerTransformer	GS-CV	0,483	932,063	708,550
4-LGBM	StandardScaler	None	0,461	951,645	714,538
4-LGBM	StandardScaler	RS-CV	0,488	927,575	719,515
4-LGBM	StandardScaler	GS-CV	0,494	921,751	704,210
4-LGBM	MinMaxScaler	None	0,450	960,789	726,954
4-LGBM	MinMaxScaler	RS-CV	0,503	913,417	702,874
4-LGBM	MinMaxScaler	GS-CV	0,495	920,637	705,017
4-LGBM	RobustScaler	None	0,479	935,454	706,812
4-LGBM	RobustScaler	RS-CV	0,496	920,189	714,740
4-LGBM	RobustScaler	GS-CV	0,488	927,108	716,806
4-LGBM	PowerTransformer	None	0,484	930,660	693,590
4-LGBM	PowerTransformer	RS-CV	0,483	931,852	725,935
4-LGBM	PowerTransformer	GS-CV	0,490	925,523	730,937

Fonte: Elaborado pelo autor.