UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Gabriel Senosien Viotto

Técnicas de Aprendizado de Máquina Aplicadas na Predição da Produtividade de Soja no Estado do Tocantins

Goiânia
2025

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): Gabriel Senosien Viotto.

Título do trabalho: Técnicas de Aprendizado de Máquina Aplicadas na Predição da Produtividade de Soja no Estado do Tocantins.

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [x] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;

- Submissão de artigo em revista científica;

- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **David Henriques Da Matta**, **Professor do Magistério Superior**, em 08/12/2025, às 09:32, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Gabriel Senosien Viotto**, **Discente**, em 12/12/2025, às 09:24, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Gabriel Senosien Viotto

Técnicas de Aprendizado de Máquina Aplicadas na Predição da Produtividade de Soja no Estado do Tocantins

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Goiás para aprovação no componente curricular TCC, como parte das exigências para a obtenção do título de bacharel em Estatística. **Orientador**: David Henriques da Matta

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Viotto, Gabriel Senosien
Técnicas de Aprendizado de Máquina Aplicadas na Predição da Produtividade de Soja no Estado do Tocantins [manuscrito] / Gabriel Senosien Viotto. - 2025.
35 f.: il.

Orientador: Prof. David Henriques da Matta.
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Goiás, Instituto de Matemática e Estatística (IME), Estatística, Goiânia, 2025.
Bibliografia. Anexos. Apêndice.
Inclui siglas, mapas, abreviaturas, símbolos, gráfico, tabelas, lista de figuras, lista de tabelas.

1. Estatísticas. 2. Redução de risco. 3. Matopiba. 4. Dados climáticos. I. Matta, David Henriques da, orient. II. Título.

CDU 519.22

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

ATA DE DEFESA DE TRABALHO DE CONCLUSÃO DE CURSO

Aos vinte e seis dias do mês de novembro do ano de 2025 iniciou-se a sessão pública de defesa do Trabalho de Conclusão de Curso (TCC) intitulado “Técnicas de Aprendizado de Máquina Aplicadas na Predição da Produtividade de Soja no Estado do Tocantins”, de autoria de Gabriel Senosien Viotto, do curso de Estatística, do Instituto de Matemática e Estatística da UFG. Os trabalhos foram instalados pelo Prof. Dr. David Henriques da Matta com a participação dos demais membros da Banca Examinadora: Valdivino Vargas Junior (IME/UFG), Joelmir Divino Carlos Feliciano (IME/UFG) e Leonardo José Motta Campos (Embrapa/GO). Após a apresentação, a banca examinadora realizou a arguição do estudante. Posteriormente, de forma reservada, a Banca Examinadora atribuiu a nota final de 9,3 , tendo sido o TCC considerado aprovado.

Proclamados os resultados, os trabalhos foram encerrados e, para constar, lavrou-se a presente ata que segue assinada pelos Membros da Banca Examinadora.

Documento assinado eletronicamente por **Joelmir Divino Carlos Feliciano**, **Professor do Magistério Superior**, em 26/11/2025, às 16:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **David Henriques Da Matta**, **Professor do Magistério Superior**, em 26/11/2025, às 19:05, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Leonardo José Motta Campos**, **Usuário Externo**, em 27/11/2025, às 09:02, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Valdivino Vargas Junior**, **Professor do Magistério Superior**, em 27/11/2025, às 10:09, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5793586** e o código CRC **38635926**.

Referência: Processo nº 23070.056228/2025-92 SEI nº 5793586

Agradecimentos

Agradeço primeiramente a Deus pela minha vida e sabedoria, meus pais Fabio e Fabiana pelo apoio incondicional e carinho mesmo de longe foram fundamentais, minha namorada Geovana pela compreensão e carinho, meus avós Alzira, Luiz, Olga e Valentim pelo apoio, meus tios-avós Natalino e Maria José por me receber em Goiânia, ao IME e a UFG pelo curso e ao meu orientador David pela oportunidade e aprendizados.

Resumo

Este estudo aplica e compara técnicas de aprendizado de máquina na predição da produtividade de soja no estado do Tocantins, utilizando variáveis categorizadas via k-means oriundas de dados climáticos da plataforma NASA POWER e dados de produtividade de soja, obtidos em experimentos realizados em diversos municípios do Tocantins, de 2013 à 2023. Foram implementados cinco algoritmos: Random Forest, Árvore de Decisão, XGBoost, Support Vector Machine e Bagging. Os modelos foram treinados com dados de municípios de produtividade de Soja e avaliados mediante validação cruzada utilizando as métricas RMSE, MAE e R^2. Os resultados demonstraram a superioridade dos métodos de *ensemble*, com o Bagging apresentando o melhor desempenho (RMSE = 498,64 kg/ha, MAE = 380,07 kg/ha, R^2 = 0,724). O XGBoost e Random Forest também obtiveram resultados muito próximos. A análise de importância de variáveis revelou que fatores climáticos, especialmente precipitação no período reprodutivo e radiação solar no período vegetativo, são determinantes cruciais para a produtividade. O estudo conclui que as técnicas de aprendizado de máquina, particularmente os métodos de *ensemble*, são ferramentas promissoras para a predição de produtividade da soja, podendo auxiliar no planejamento agrícola e na tomada de decisão no setor, para a redução do alto risco inerente a esta atividade.

Palavras-chave: Estatísticas. Redução de Risco. Matopiba. Dados Climáticos.

Abstract

This study applies and compares machine learning techniques for predicting soybean yield in the state of Tocantins, using categorized variables via k-means derived from climate data from the NASA POWER platform and soybean yield data obtained from experiments conducted in several municipalities of Tocantins from 2013 to 2023. Five algorithms were implemented: Random Forest, Decision Tree, XGBoost, Support Vector Machine, and Bagging. The models were trained using yield data from municipalities and evaluated through cross-validation using the metrics RMSE, MAE, and R^2. The results demonstrated the superiority of ensemble methods, with Bagging showing the best performance (RMSE = 498.64 kg/ha, MAE = 380.07 kg/ha, R^2 = 0.724). XGBoost and Random Forest also achieved very similar results. The variable importance analysis revealed that climatic factors—especially precipitation during the reproductive period and solar radiation during the vegetative period—are crucial determinants of productivity. The study concludes that machine learning techniques, particularly ensemble methods, are promising tools for predicting soybean yield and can support agricultural planning and decision-making in the sector, contributing to reducing the high inherent risk of this activity.

Keywords: Statistics. Risk Reduction. Matopiba. Climate Data.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

RMSE: *Root Mean Squared Error*

MAE: *Mean Absolute Error*

Lista de símbolos

Σ	Letra grega Sigma

Sumário

Introdução

A soja ocupa uma posição de destaque no agronegócio brasileiro, sendo uma das principais commodities exportadas pelo país, líder mundial na produção e exportação de soja (Companhia Nacional de Abastecimento (CONAB), 2024; Food and Agriculture Organization of the United Nations (FAO), 2023). O cultivo da soja não apenas movimenta a economia nacional, mas também exerce forte impacto nas dinâmicas regionais, sobretudo em estados que vêm ampliando sua fronteira agrícola, como o Tocantins.

Prever a produtividade da soja no Tocantins é vital para a gestão econômica e o planejamento estratégico em múltiplas escalas (Centro de Estudos Avançados em Economia Aplicada (CEPEA), 2023). Para o produtor rural, essas estimativas antecipadas são fundamentais para o gerenciamento de risco (ASSAD; PINTO, 2021), orientando decisões críticas sobre épocas de plantio, comercialização e planejamento financeiro para a safra seguinte (BACCHI; CALDAS, 2021). Em nível estadual e nacional, a soja é um pilar da economia e um dos principais produtos da pauta de exportações (MDIC, 2024). Portanto, prever sua produtividade permite ao produtor, ao governo e aos agentes de mercado planejarem a safra (época de plantio e colheita principalmente), estabilizarem os preços de venda (BARROS, 2022), gerenciarem estoques e assegurarem a logística de escoamento (CASTRO *et al.*, 2023) transformando incertezas climáticas em informações administráveis para o desenvolvimento sustentável do agronegócio, podendo portanto reduzir os efeitos negativos das mudanças climáticas sobre a produção agrícola..

Nos últimos anos, o estado do Tocantins tem consolidado sua relevância na produção de grãos, especialmente pela expansão de áreas cultivadas (ofertas de terras baratas) e pela busca de práticas agrícolas mais eficientes e adaptadas às suas condições climáticas e do solo (Instituto Brasileiro de Geografia e Estatística (IBGE), 2022; SILVA; LIMA; SOUZA, 2021).

A produtividade da soja é resultado da interação de múltiplos fatores, entre os quais destacam-se as características do solo e as condições climáticas durante seu ciclo fenológico, fatores como a temperatura, que rege a velocidade dos processos bioquímicos, a disponibilidade hídrica, determinada pela precipitação e pela evapotranspiração e a radiação solar, fonte primária de energia para a fotossíntese, são fatores cruciais para o desenvolvimento da planta e, consequentemente, para o resultado final do rendimento de grãos. A variabilidade espaço-temporal desses elementos pode ser robustamente capturada pela base de dados NASA POWER (NASA Power Project, 2024), a qual fornece séries históricas de parâmetros agroclimatológicos de forma global e de acesso livre, permitindo a modelagem da relação entre clima e produtividade com significativa confiabilidade.

Diante da complexidade desses fatores, torna-se evidente a necessidade de metodologias computacionais avançadas para análise e previsão de produtividade. Nesse sentido, as técnicas de

Aprendizado de Máquina (*Machine Learning*) surgem como ferramentas de destaque, capazes de identificar padrões complexos em grandes volumes de dados e oferecer previsões mais acuradas do que métodos estatísticos tradicionais (JAMES *et al.*, 2021; HASTIE; TIBSHIRANI; FRIEDMAN, 2017).

Modelos como Árvores de Decisão, Random Forest, XGBoost, Support Vector Machines e Bagging têm apresentado resultados consistentes em diferentes aplicações, incluindo estudos voltados para gestão de risco na agricultura (ZHANG; LI; WANG, 2019). A capacidade desses algoritmos em lidar com relações não-lineares e interações complexas entre variáveis os torna muito adequados para problemas agronômicos.

Neste contexto, o presente trabalho tem como objetivo principal comparar diferentes técnicas de aprendizado de máquina na previsão da produtividade de soja em áreas de cultivo no Tocantins, considerando dados obtidos em experimentos e climáticos provenientes da plataforma NASA POWER. Especificamente, busca-se:

- Implementar e otimizar cinco algoritmos de aprendizado de máquina para predição de produtividade;
- Comparar o desempenho dos modelos utilizando métricas estatísticas robustas;
- Identificar o modelo mais adequado para auxiliar na tomada de decisão no setor agrícola;
- Disponibilizar ferramentas preditivas que possam apoiar o planejamento de safras e a adoção de estratégias mais seguras e sustentáveis.

A relevância deste estudo reside no potencial de fornecer ferramentas que auxiliam na tomada de decisão para produtores rurais, técnicos agrícolas e gestores públicos, possibilitando uma agricultura mais segura, eficiente e adaptada às condições específicas do estado do Tocantins. A abordagem metodológica adotada pode ser replicada em outras regiões e culturas, ampliando o impacto dos resultados obtidos. Além disso, o impacto cada vez maior das mudanças climáticas sobre a produção agrícola, torna a importância deste estudo ainda maior.

A estrutura deste trabalho está organizada em quatro capítulos. No Capítulo 1, é apresentada a fundamentação teórica sobre os algoritmos de aprendizado de máquina utilizados. O Capítulo 2 descreve a metodologia empregada, incluindo a coleta e preparação dos dados. No Capítulo 3, são expostos os resultados obtidos com a aplicação dos modelos e discute os resultados baseados na literatura existente. Por fim, as considerações finais e recomendações são apresentadas na Conclusão.

1 Revisão Bibliográfica

1.1 Aprendizado de Máquina na Agricultura

O Aprendizado de Máquina (*Machine Learning*) tem se consolidado como uma das principais abordagens da Estatística e da Inteligência Artificial (IA) aplicadas ao setor agrícola, permitindo a modelagem de relações complexas entre variáveis ambientais, fisiológicas e agronômicas (SHARMA; KUMAR; SINGH, 2025). Essa área se destaca pela capacidade de identificar padrões ocultos em grandes volumes de dados e de realizar predições precisas, mesmo em contextos caracterizados por alta variabilidade espacial e temporal, condições típicas dos sistemas agrícolas modernos.

Segundo Mitchell (1997), "um programa de computador é dito aprender a partir de uma experiência E com relação a algumas classes de tarefas T e medida de desempenho P, se o seu desempenho em T, medido por P, melhora com a experiência E". No contexto agrícola, a "experiência E" pode ser representada pelos dados históricos de produtividade, parâmetros meteorológicos, características físico-químicas do solo, práticas de manejo e imagens de sensoriamento remoto. A "tarefa T" frequentemente se refere à predição de produtividade, à estimativa de parâmetros de crescimento ou à detecção de estresses bióticos e abióticos, enquanto a "medida de desempenho P" envolve métricas quantitativas de avaliação, como o Erro Quadrático Médio (RMSE), o Erro Absoluto Médio (MAE) e o Coeficiente de Determinação (R^2).

O uso de técnicas de Aprendizado de Máquina na agricultura moderna é um dos pilares centrais do paradigma da Agricultura 4.0, que representa a quarta revolução agrícola (WOLFERT *et al.*, 2017). Nesta fase, a agricultura torna-se um sistema ciber-físico, integrando massivamente dados heterogêneos, provenientes de sensores de solo, imagens de satélites e drones, e estações meteorológicas, em plataformas analíticas. O Aprendizado de Máquina atua como o cérebro desses sistemas (LIAKOS *et al.*, 2018) transformando esse grande volume de dados em ideias para o apoio à decisão, seja para a aplicação de insumos em taxa variada, a detecção precoce de pragas ou a previsão de produtividade. Dessa forma, o Aprendizado de Máquina permite não apenas aumentar a eficiência produtiva e reduzir custos, mas também promover a sustentabilidade ambiental por meio do uso racional e preciso de recursos como água, fertilizantes e defensivos.(RODRIGUES; SANTANA; GUIMARãES, 2021).

Diversos estudos têm demonstrado a eficácia dos algoritmos de aprendizado de máquina na predição de produtividade agrícola. (ZHANG; LI; WANG, 2019) aplicaram Random Forest e XGBoost na predição de produtividade de soja na China, obtendo resultados superiores aos modelos tradicionais. (KLOMPENBURG; KASSAHUN; CATAL, 2020) compararam múltiplos algoritmos e identificaram que métodos de ensemble tendem a apresentar melhor desempenho em

dados agrícolas. No contexto brasileiro, (RODRIGUES; SANTANA; GUIMARãES, 2021) aplicaram técnicas de aprendizado de máquina na predição de produtividade de soja no país, destacando a importância da integração de dados climáticos e de solo para melhorar a acurácia dos modelos.

1.2 Algoritmos de Aprendizado de Máquina Aplicados

A predição precisa da produtividade agrícola é um desafio complexo, dada a natureza multifatorial e as relações não-lineares que governam a interação entre o genótipo da cultura e as variáveis edafoclimáticas. Nesse contexto, o Aprendizado de Máquina, tendo em vista sua capacidade de modelar padrões complexos e interações de alta dimensão, é apropriada (KLOMPENBURG; KASSAHUN; CATAL, 2020). Entre as abordagens mais utilizadas, serão avaliadas a Árvore de Decisão, o Random Forest, o XGBoost, o Support Vector Machines e os métodos de Bagging e Boosting, no presente estudo.

1.2.1 Árvores de Decisão

As Árvores de Decisão são modelos supervisionados que segmentam o espaço de atributos por meio de divisões hierárquicas sucessivas, baseadas em critérios de impureza como a entropia ou o índice de Gini (BREIMAN *et al.*, 1984). Cada divisão é realizada de forma a maximizar a homogeneidade dos subconjuntos resultantes, produzindo uma estrutura lógica e facilmente interpretável, o que as torna especialmente atrativas em aplicações agrícolas, onde a transparência das decisões é essencial (BREIMAN *et al.*, 2017).

Esses modelos são amplamente empregados na agricultura de precisão para tarefas como a classificação de solos (DEMATTE *et al.*, 2018), o diagnóstico de deficiências nutricionais (MOHANTY; HUGHES; SALATHé, 2016) e a previsão de produtividade com base em variáveis climáticas e de manejo (KLOMPENBURG; KASSAHUN; CATAL, 2020).

1.2.2 Random Forest

O Random Forest, desenvolvido por Breiman (2001), é um método de *ensemble* que combina múltiplas árvores de decisão através da técnica de *bagging* criando múltiplos subconjuntos de dados a partir do conjunto original usando amostragem com substituição. Essa abordagem reduz a variância do modelo e melhora sua capacidade de generalização, sendo particularmente robusta diante de dados ruidosos ou colineares.

Em aplicações agrícolas, o Random Forest tem sido amplamente utilizado para estimar rendimento de culturas (KLOMPENBURG; KASSAHUN; CATAL, 2020). Mapear propriedades do solo (HENGL *et al.*, 2018) e classificar imagens de sensoriamento remoto (BELGIU; DRăGUţ, 2016) devido à sua alta acurácia e capacidade de lidar com conjuntos de dados heterogêneos (LIAKOS *et al.*, 2018).

1.2.3 XGBoost (Extreme Gradient Boosting)

O XGBoost, proposto por Chen e Guestrin (2016), representa uma evolução dos métodos de *boosting* construindo árvores de decisão de forma sequencial, onde cada nova árvore busca corrigir os erros residuais das anteriores. Sua eficiência computacional, capacidade de paralelização e mecanismos de regularização que previnem um sobreajuste, o tornando um dos algoritmos mais poderosos e competitivos em tarefas de previsão.

Na agricultura, o XGBoost tem sido aplicado com sucesso em problemas de predição de produtividade de culturas (KLOMPENBURG; KASSAHUN; CATAL, 2020), classificação de doenças em folhas (ATHANASIOS *et al.*, 2022) e otimização de manejo agrícola. Esta última aplicação se concretiza na recomendação de aplicação de insumos em taxa variada (FENG *et al.*, 2021) e no manejo preciso da irrigação (WANG *et al.*, 2021), onde o modelo analisa dados multivariados para prescrever as ações mais eficientes e sustentáveis.

1.2.4 Support Vector Machines

Os Support Vector Machines, introduzidos por Cortes e Vapnik (1995), são algoritmos de aprendizado supervisionado baseados na Teoria Estatística de Aprendizado e no princípio da Minimização do Risco Estrutural. Diferente de métodos que minimizam o erro empírico, as Support Vector Machines buscam encontrar o hiperplano ótimo no espaço de características que maximiza a margem de separação entre classes, o que, teoricamente, leva a uma melhor generalização para dados não vistos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Na versão para regressão (Support Vector Regression), o objetivo é ajustar uma função que capture a tendência central dos dados, mantendo os erros dentro de uma margem de tolerância ϵ e penalizando de forma linear apenas os desvios absolutos superiores a esse limite.

O poder de modelagem das Support Vector Machines é ampliado pelo uso de funções de *kernel*, como as funções linear, polinomial e de base radial, que mapeiam implicitamente os dados para um espaço de maior dimensionalidade onde um hiperplano linear se torna uma fronteira de decisão complexa no espaço original. Essa característica, conhecida como *Kernel Trick*, as torna especialmente úteis em contextos agrícolas com número limitado de observações e alta dimensionalidade, típicos de experimentos de pequena escala ou de análises de dados espectrais de solo e planta (THENKABAIL, 2018). No entanto, seu desempenho é sensível à calibração dos hiperparâmetros de regularização (C) e do *kernel*, o que demanda uma validação cruzada cuidadosa.

1.2.5 Bagging

O *Bagging* (Bootstrap Aggregating), introduzido por (BREIMAN, 1996), é uma técnica de *ensemble* projetada primariamente para reduzir a variância de modelos com alta instabilidade, como árvores de decisão profundas. O método opera gerando múltiplos modelos base (por

exemplo, árvores de decisão) independentes, cada um treinado em uma amostra bootstrap diferente, um subconjunto aleatório do conjunto de dados original, obtido com reposição. A previsão final é obtida pela agregação dos resultados de todos os modelos, seja por votação majoritária (para classificação) ou pela média (para regressão). Esse processo de "democracia de modelos"suaviza o ruído e mitiga o sobreajuste (*overfitting*), resultando em um preditor mais robusto e generalizável (GOODFELLOW; BENGIO; COURVILLE, 2016).

Na agricultura de precisão, onde os dados são inerentemente complexos e com alta variabilidade espacial e temporal, o *Bagging* é amplamente aplicado. Ele é a base fundamental para algoritmos como o *Random Forest*, que agrega o princípio do *Bagging* com a aleatorização de características. Sua aplicação é comum em tarefas como a previsão de rendimento a partir de múltiplas variáveis ambientais e de manejo, e na classificação de padrões de saúde vegetal ou estresse hídrico em imagens multiespectrais obtidas por drones e satélites, onde a estabilidade do modelo é crucial para a confiabilidade do diagnóstico.

1.3 Otimização de Hiperparâmetros dos Modelos

1.4 Avaliação dos Modelos

A avaliação do desempenho de modelos de Aprendizado de Máquina é uma etapa essencial para garantir a validade das previsões e a confiabilidade das conclusões inferidas. As métricas de desempenho permitem quantificar a diferença entre os valores observados e os valores estimados pelo modelo, fornecendo subsídios para comparação entre diferentes algoritmos e configurações de parâmetros. No contexto agrícola, onde os dados frequentemente apresentam alta variabilidade e ruído, a escolha adequada da métrica é determinante para avaliar corretamente a capacidade preditiva do modelo (HYNDMAN; KOEHLER, 2006) (CHAI; DRAXLER, 2014).

As principais métricas utilizadas em modelos de regressão, como os aplicados à previsão de produtividade, incluem o Coeficiente de Determinação (R^2), o Erro Quadrático Médio (RMSE) e o Erro Absoluto Médio (MAE). Cada uma delas oferece uma perspectiva distinta sobre o desempenho do modelo e, quando utilizadas em conjunto, permitem uma avaliação mais abrangente (VENKATESH; PARTHIBAN, 2022).

Para além da seleção adequada de métricas, é importante garantir que a avaliação do modelo não seja influenciada por uma divisão específica dos dados entre treinamento e teste, assegurando que seu desempenho seja generalizável para novas observações. Para este fim, a validação cruzada destaca-se como uma técnica indispensável. Trata-se de um procedimento de reamostragem que fornece uma estimativa mais robusta e confiável do erro preditivo do modelo, mitigando o risco de superajuste a particularidades de um único subconjunto de dados. A estratégia mais consolidada é a validação cruzada k-partes, na qual o conjunto de dados original é dividido aleatoriamente em k grupos de tamanho similar. O processo é então repetido

k vezes: em cada iteração, um dos grupos é utilizado como conjunto de teste, enquanto os k-1 grupos restantes formam o conjunto de treinamento. O modelo é treinado e, posteriormente, avaliado no grupo de teste, gerando uma estimativa das métricas de desempenho. Ao final do ciclo, as k estimativas obtidas para cada métrica são consolidadas por meio de sua média e desvio padrão, produzindo uma medida final de acurácia que é significativamente menos variável e mais representativa da performance real do algoritmo. No contexto da previsão de produtividade agrícola, onde os dados são inerentemente complexos e sujeitos a alta variabilidade, a validação cruzada é crucial não apenas para uma avaliação final realista, mas também para comparar diferentes algoritmos e para o afinamento dos seus hiperparâmetros, garantindo que o modelo selecionado seja o mais apto a realizar previsões precisas sob condições variáveis (ARLOT; CELISSE, 2010).

1.4.1 Coeficiente de Determinação (R²)

O R² mede a proporção da variabilidade da variável dependente que é explicada pelo modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.1}$$

Onde y_i representa o valor observado, $\hat{y}_i$ o valor predito pelo modelo, $\bar{y}$ a média dos valores observados e n o tamanho amostral.

O R² varia entre 0 e 1, sendo que valores mais próximos de 1 indicam um melhor ajuste do modelo aos dados, representando a proporção da variabilidade explicada. No entanto, é crucial interpretar esse valor com cautela, pois um R² elevado pode ser um indício de sobreajuste.

O sobreajuste ocorre quando um modelo é excessivamente complexo e acaba capturando não apenas os padrões gerais dos dados, mas também o ruído e as flutuações aleatórias específicas do conjunto de treinamento. Embora isso resulte em um ajuste perfeito aos dados conhecidos, o modelo perde a capacidade de generalização, apresentando desempenho significativamente pior quando aplicado a novos dados. É análogo a memorizar respostas em vez de compreender o conceito, portanto, um R² muito alto, especialmente em modelos com muitas variáveis, deve ser avaliado com criticidade. Para uma análise mais confiável, é essencial utilizar o R² ajustado, que penaliza a adição de variáveis irrelevantes, e validar o modelo em um conjunto de teste independente, onde um desempenho consistente confirma sua robustez (MONTGOMERY; PECK; VINING, 2012).

1.4.2 Root Mean Square Error (RMSE)

O *Root Mean Square Error* (RMSE) representa a raiz quadrada da média dos erros quadráticos e é amplamente utilizado para quantificar a magnitude média dos erros entre valores

observados e preditos (CHAI; DRAXLER, 2014):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1.2}$$

Onde y_i representa o valor observado, $\hat{y}_i$ o valor predito pelo modelo e n o tamanho amostral.

O RMSE fornece uma medida direta da precisão do modelo, sendo expresso nas mesmas unidades da variável resposta. Valores menores indicam maior precisão, refletindo menor dispersão dos erros em torno da linha de regressão perfeita.

Essa métrica é particularmente sensível a grandes discrepâncias individuais *(outliers)*, pois o erro é elevado ao quadrado antes da média. Assim, em contextos agrícolas onde podem ocorrer valores extremos como uma época de seca por exemplo, o RMSE pode aumentar significativamente, mesmo que o restante das previsões seja satisfatório.

1.4.3 Mean Absolute Error (MAE)

O Mean Absolute Error (MAE) mede o erro médio absoluto entre os valores observados e preditos, refletindo o desvio médio sem considerar a direção do erro (HYNDMAN; KOEHLER, 2006):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1.3}$$

Onde y_i representa o valor observado, $\hat{y}_i$ o valor predito pelo modelo e n o tamanho amostral.

O MAE é uma métrica simples e intuitiva, expressa nas mesmas unidades da variável analisada, o que facilita a interpretação prática dos resultados. Diferentemente do RMSE, o MAE atribui peso igual a todos os erros, sendo menos sensível a valores extremos.

Em termos práticos, ele representa o erro médio esperado entre as observações e as previsões do modelo, o que o torna uma métrica bastante útil em aplicações agrícolas voltadas à previsão de rendimento e avaliação de modelos empíricos de produtividade, onde a robustez frente a valores discrepantes é desejável.

Todavia, por não penalizar erros grandes com a mesma severidade que o RMSE, o MAE pode subestimar o impacto de observações discrepantes, portanto é recomendável utilizá-lo conjuntamente com outras métricas.

2 Metodologia

2.1 Área de Estudo

O estudo foi conduzido em diferentes cidades do estado do Tocantins, uma das regiões em grande expansão da fronteira agrícola brasileira. Todos os dados foram obtidos por meio da experimentação agrícola, mantendo os princípios básicos da experimentação (repetição, aleatoriedade e controle local). O estado apresenta elevada variabilidade edafoclimática (condições de solo e clima), com diferentes tipos de solos e classificação predominantemente de Latossolos e Plintossolos. O regime climático caracterizado por duas estações bem definidas, uma chuvosa e outra seca, com altas temperaturas médias, fatores que influenciam fortemente o desenvolvimento da cultura da soja.

Foram considerados dados provenientes de municípios representativos da produção agrícola do estado, incluindo Aparecida do Rio Negro, Guaraí, Lagoa da Confusão, Palmas, Paraíso do Tocantins, Pedro Afonso, Pium e Porto Nacional. A seleção dessas localidades buscou abranger diferentes condições de solo, altitude e regime pluviométrico, de modo a capturar a variabilidade natural dos ambientes produtivos do Tocantins.

2.1.1 Aquisição e Estruturação do Conjunto de Dados

Os dados utilizados neste estudo foram obtidos a partir de duas principais fontes, integradas para compor uma base abrangente de informações. Os dados de produtividade e de solo foram cedidos pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa), abrangendo o período de 2013 a 2023 e organizados por município. Este conjunto de dados engloba informações agronômicas fundamentais, como a produtividade média (kg/ha), textura do solo, a cultivar plantada, o ciclo da cultivar. Após a aquisição, esses dados foram organizados em planilhas e padronizados para garantir sua integração com as demais fontes.

Complementarmente, os dados climáticos foram coletados por meio da plataforma NASA POWER *Prediction of Worldwide Energy Resources*, que fornece séries temporais de dados meteorológicos em resolução diária e espacial ajustada. As variáveis climáticas consideradas incluíram temperatura média, máxima e mínima (*T2M, T2M MAX, T2M MIN*) precipitação total (*PRECTOTCORR*), radiação solar incidente (*ALLSKY SFC SW DWN*) (NASA Power Project, 2024). A combinação sistemática dessas fontes permitiu a criação de um conjunto de dados robusto, estruturado por município e ao longo de uma série histórica de uma década, essencial para a modelagem precisa da produtividade da soja com técnicas de aprendizado de máquina.

A etapa de categorização de variáveis teve como objetivo extrair informações derivadas das variáveis climáticas originais, representando de forma mais significativa os padrões observados ao

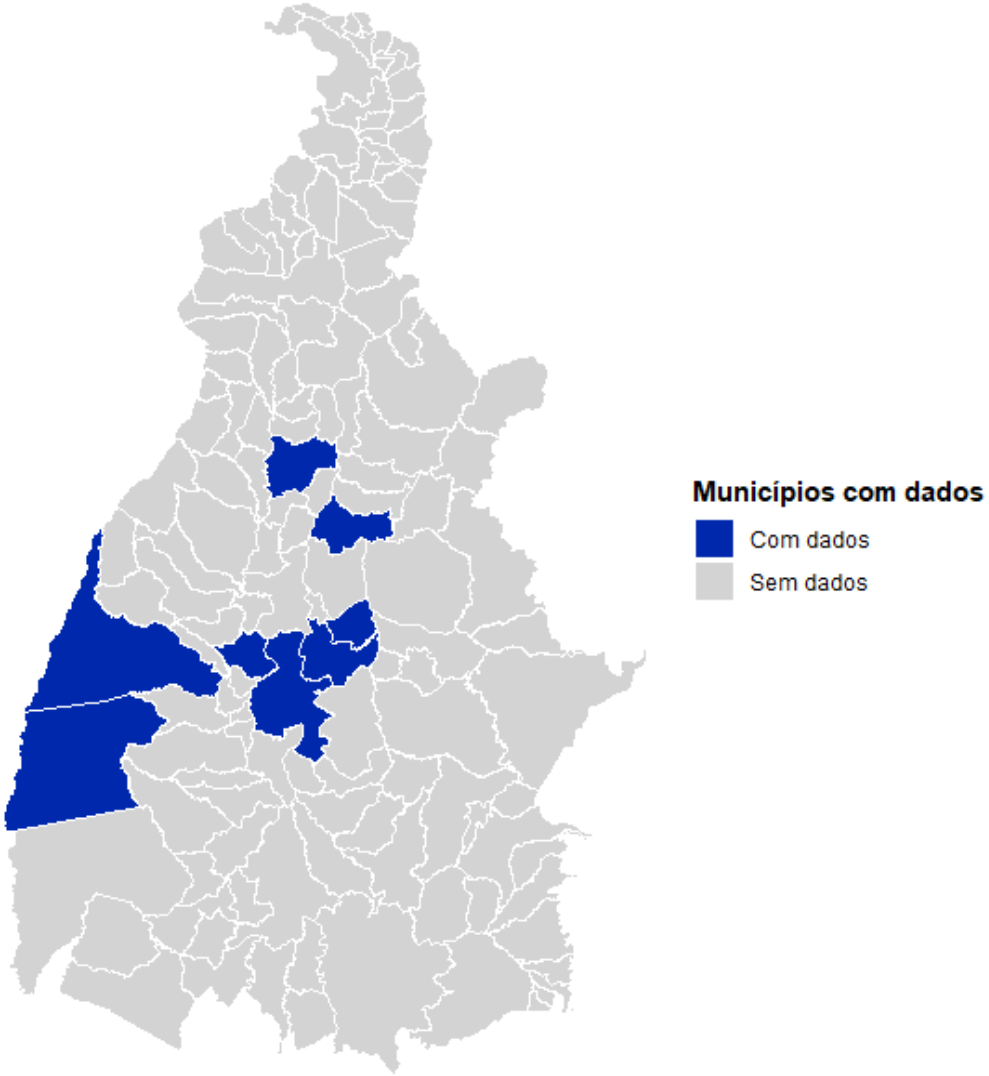

Figura 1 – Mapa da distribuição dos municípios do estado do Tocantins

Fonte: Elaborado pelo autor

longo do ciclo fenológico da soja. As variáveis contínuas foram agrupadas por meio do algoritmo K-means (LLOYD, 1982), gerando três categorias representativas dos níveis baixo, médio e alto de cada variável. Essa discretização auxilia na interpretação agronômica dos resultados e na aplicação de algoritmos mais sensíveis a variáveis categóricas.

Foram criadas 48 variáveis climáticas derivadas, considerando:

- Indicadores de precipitação, temperatura e radiação em três estágios da cultura (vegetativo, reprodutivo e de maturação);
- Medidas de tendência central (média e mediana);
- Medidas de dispersão (intervalo interquartil — IQR);

- Proporções de valores acima do percentil 90 e abaixo do percentil 10, para identificação de extremos climáticos;
- Acumulados de precipitação durante cada fase fenológica.

Essas variáveis sintetizam as condições ambientais predominantes ao longo do ciclo produtivo, possibilitando que os modelos capturem interações complexas entre clima, solo e produtividade.

O pré-processamento foi realizado integralmente no ambiente R (R Development Core Team, 2024), conforme as diretrizes apresentadas por Kuhn e Johnson (2013), assegurando a qualidade, consistência e integridade dos dados antes da aplicação dos algoritmos. Essa etapa é fundamental, pois dados agrícolas frequentemente apresentam ruídos, lacunas e escalas heterogêneas, o que pode comprometer o desempenho dos modelos preditivos se não forem devidamente tratados.

2.1.2 Divisão Treino-Teste

O conjunto de dados final foi dividido em subconjuntos de treino (70%) e teste (30%), utilizando amostragem estratificada para preservar a distribuição da variável resposta. Essa divisão visa avaliar a capacidade de generalização dos modelos, evitando sobreajuste e assegurando que o desempenho medido reflita o comportamento esperado em dados não obtidos.

Para a validação e otimização dos hiperparâmetros, utilizou-se o método de validação cruzada 5-fold (ARLOT; CELISSE, 2010). Nesse procedimento, o conjunto de treino é dividido em cinco partes de tamanhos aproximadamente iguais; a cada iteração, uma parte é usada para validação e as demais para treinamento. Essa técnica garante uma avaliação mais robusta do desempenho médio do modelo, reduzindo o viés decorrente de particionamentos específicos dos dados.

2.2 Algoritmos de Aprendizado de Máquina

Foram implementados e comparados cinco algoritmos de aprendizado de máquina supervisionado, amplamente reconhecidos na literatura por seu desempenho em modelagem agrícola: Árvore de Decisão, Random Forest, XGBoost, Support Vector Machine e Bagging. A seguir, descrevem-se suas principais características e os parâmetros ajustados e todo o processamento e análise de dados foram realizados na linguagem R (versão 4.3.0) (R Core Team, 2023), no ambiente integrado RStudio.

2.2.1 Random Forest (RF)

O algoritmo Random Forest (BREIMAN, 2001) foi implementado por meio do pacote `ranger`, com otimização de hiperparâmetros via *grid search*. Trata-se de um método de *ensemble* baseado em múltiplas árvores de decisão, treinadas a partir de amostras aleatórias dos dados através da técnica de *bootstrap*. Foram otimizados os parâmetros `mtry` (número de variáveis consideradas em cada divisão), `splitrule` (critério de divisão, como *variance* ou *extratrees*) e `min.node.size` (tamanho mínimo dos nós terminais). O Random Forest é amplamente utilizado em aplicações agrícolas devido à sua robustez contra ruído e capacidade de lidar com interações não lineares entre variáveis climáticas e agronômicas.

2.2.2 Árvore de Decisão (AD)

As Árvores de Decisão foram implementadas com o pacote `rpart` (THERNEAU; ATKINSON, 2019), seguindo os princípios estabelecidos por Breiman *et al.* (1984). Utilizou-se o parâmetro de complexidade (`cp`) como fator de regularização para evitar sobreajuste. Esta técnica é especialmente útil por fornecer uma estrutura interpretável, permitindo a identificação direta das variáveis mais relevantes para a produtividade e a criação de cenários baseados nas condições que levam a determinados níveis de produtividade estimada.

2.2.3 XGBoost (XGB)

O XGBoost (CHEN; GUESTRIN, 2016) foi implementado com o pacote *xgboost*, utilizando o princípio do *boosting* onde modelos sequenciais são treinados de forma que cada nova árvore corrija os erros residuais da anterior. Foram otimizados os parâmetros `nrounds` (número de iterações de boosting), `max depth` (profundidade máxima das árvores), `eta` (taxa de aprendizado) e `gamma` (redução mínima de perda para permitir nova partição).

2.2.4 Support Vector Machine (SVM)

O modelo SVM (CORTES; VAPNIK, 1995) foi implementado com o pacote `e1071`, utilizando o *kernel* radial. Os parâmetros de custo (`C`) e largura do *kernel* (`sigma`) foram otimizados por validação cruzada. Este algoritmo é eficaz para modelar relações não lineares e funciona bem em conjuntos de dados com alta dimensionalidade, características frequentes em dados agrícolas que combinam variáveis climáticas, de solo e de manejo.

2.2.5 Bagging

O método Bagging (*Bootstrap Aggregating*) foi implementado como um caso especial do Random Forest usando o pacote `ranger`, configurando `mtry` igual ao número total de variáveis preditoras conforme proposto originalmente por Breiman (1996). Esta abordagem busca reduzir

a variância dos modelos através da agregação de múltiplas árvores de decisão treinadas em diferentes subconjuntos dos dados, gerando previsões mais estáveis e robustas.

2.3 Métricas de Desempenho

O desempenho dos modelos foi quantificado por meio de três métricas complementares — RMSE, MAE e R^2 —, amplamente empregadas em estudos de regressão agrícola.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.3}$$

Onde y_i representa o valor observado, $\hat{y}_i$ o valor predito pelo modelo, $\bar{y}$ a média dos valores observados e n o tamanho amostral.

Essas métricas permitem mensurar tanto a precisão absoluta das predições (RMSE e MAE) quanto a proporção da variância explicada pelo modelo (R^2), possibilitando uma análise abrangente da qualidade do ajuste e da capacidade preditiva.

3 Resultados

3.1 Desempenho Comparativo dos Modelos

Os métodos de *ensemble* (Bagging e XGBoost) e o Random Forest apresentaram os melhores desempenhos, com valores de RMSE próximos de 500 kg/ha e R^2 acima de 0,72, além de um MAE em torno de 378kg/ha. O modelo Bagging obteve o melhor desempenho geral, seguido muito próximo pelo XGBoost e Random Forest. A Tabela 1 apresenta o desempenho dos cinco modelos testados no conjunto de teste. A média da produtividade nos dados de treino foi de 3971.31 kg/ha.

Tabela 1 – Desempenho comparativo dos modelos de aprendizado de máquina na predição de produtividade de soja

Modelo	**RMSE (kg/ha)**	**MAE (kg/ha)**	**R^2**	**Média Modelo (kg/ha)**
Bagging	498.64	380.07	0.724	3933.39
XGBoost	500.80	377.03	0.723	3926.31
Random Forest	503.97	380.22	0.721	3922.39
SVM	515.09	390.10	0.707	3934.34
Árvore de Decisão	525.66	394.39	0.696	3904.33

Fonte: Elaborado pelo autor

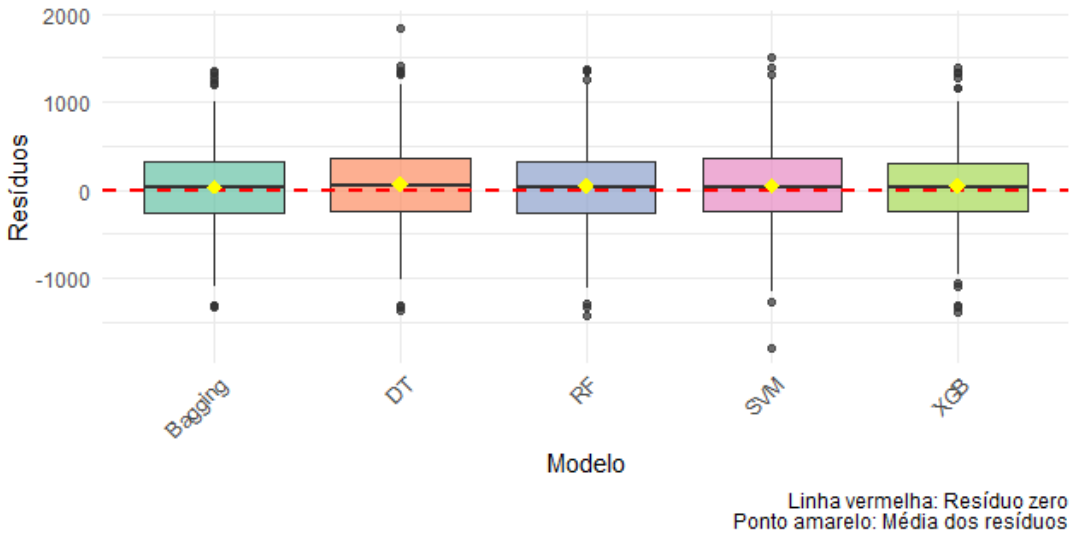

Figura 2 – Distribuição dos resíduos dos modelos de aprendizado de máquina

Fonte: Elaborado pelo autor

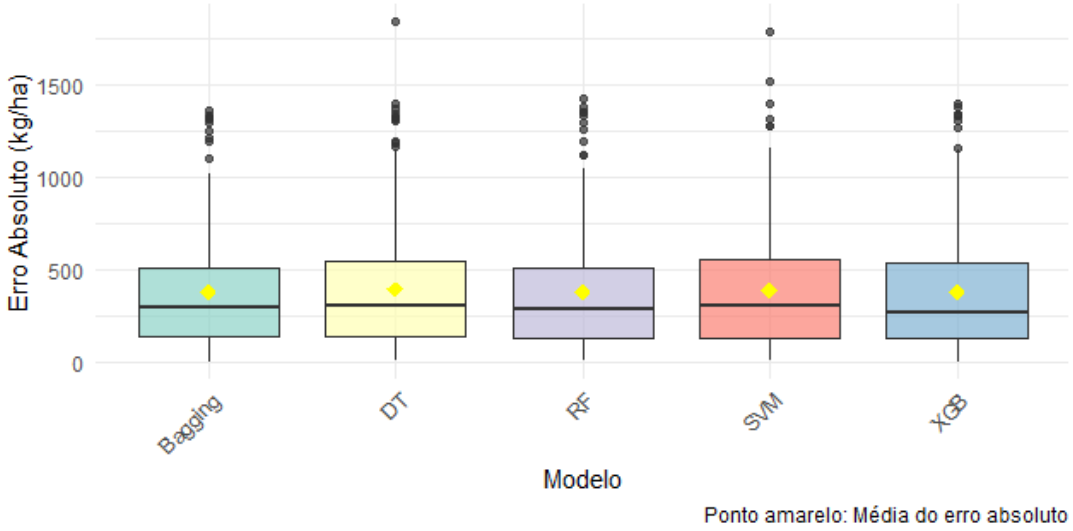

Figura 3 – Distribuição do erro absoluto dos modelos de aprendizado de máquina

Fonte: Elaborado pelo autor

A Figura 2 revela que todos os modelos apresentam distribuição aproximadamente simétrica em torno de zero, indicando ausência de viés sistemático. Os métodos de *ensemble* mostraram menor dispersão nos resíduos, como mostra a menor amplitude da caixa corroborando seu melhor desempenho nas métricas numéricas.

O erro absoluto médio variou entre 377 kg/ha (XGBoost) e 394 kg/ha (Árvore de Decisão), representando aproximadamente 10-11% da produtividade média observada na região de estudo. A Figura 3 compara a distribuição do erro absoluto entre os modelos, proporcionando uma visão complementar sobre a precisão das predições.

3.2 Importância das Variáveis

Para o modelo Random Forest, que apresentou bom desempenho e permite análise de importância de variáveis, foi possível identificar as variáveis climáticas mais relevantes para a predição de produtividade. As variáveis climáticas categorizadas relacionadas à precipitação no período reprodutivo e radiação solar no período vegetativo emergiram como as mais importantes.

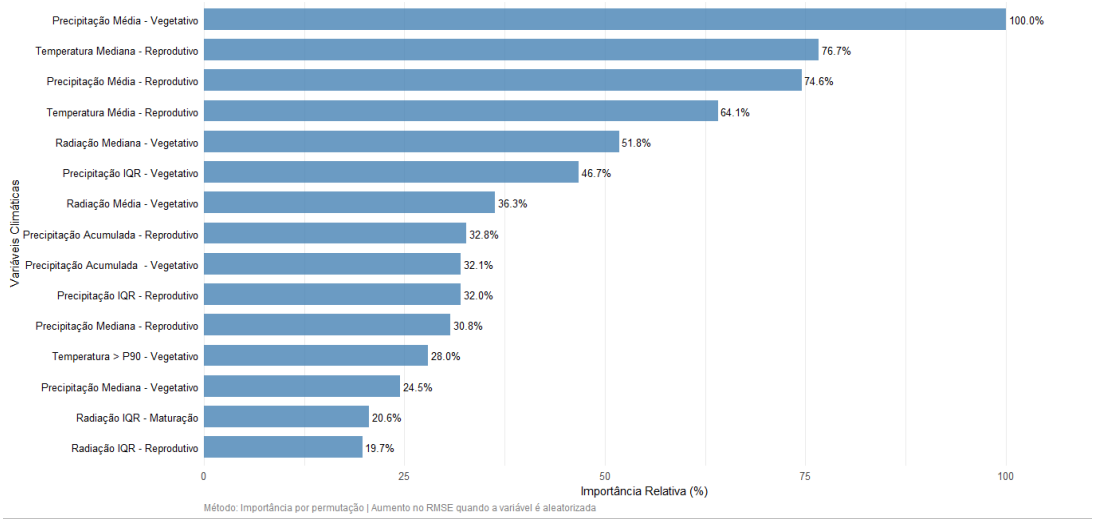

Figura 4 – Importância das variáveis segundo o Modelo Random Forest

Fonte: Elaborado pelo autor

3.3 Desempenho dos Algoritmos de Aprendizado de Máquina

Os resultados demonstram a superioridade dos métodos de *ensemble* (Bagging, XGBoost e Random Forest) sobre abordagens individuais (Árvore de Decisão) e outros paradigmas (SVM, Redes Neurais) comprovados por Zhang, Li e Wang (2019) que frequentemente reportam o melhor desempenho de técnicas de *ensemble* em problemas de predição agrícola.

A pequena diferença de desempenho entre Bagging, XGBoost e Random Forest (RMSE entre 498-504 kg/ha) sugere que, na prática, qualquer um desses três métodos poderia ser utilizado com resultados similares.

Uma importante consideração prática diz respeito ao trade-off entre desempenho preditivo e interpretabilidade. Enquanto o XGBoost e Redes Neurais são considerados "caixas-pretas"ou seja, possuem maiores dificuldades de interpretação, o Random Forest e especialmente a Árvore de Decisão oferecem maior transparência na interpretação dos resultados. Para aplicações onde a compreensão dos fatores determinantes é crucial, o Random Forest pode representar um bom equilíbrio entre acurácia e interpretabilidade.

3.4 Implicações para o Setor Agrícola

Os valores de RMSE obtidos (aproximadamente 500 kg/ha) representam um erro de predição da ordem de 9-12% considerando a produtividade média na região. Este nível de acurácia pode ser considerado satisfatório para aplicações de planejamento estratégico, como estimativas de safra e definição de políticas agrícolas.

Para aplicações em nível de talhão, que é a divisão de uma área agrícola (como um campo de cultivo ou floresta) e servem como unidades de manejo para facilitar o controle e a otimização de atividades como plantio, irrigação, adubação e colheita, onde decisões de manejo específicas são tomadas, pode ser necessário refinar ainda mais os modelos, possivelmente incorporando variáveis adicionais como adubação, práticas de manejo e dados de solo mais detalhados.

3.5 Relevância das Variáveis Climáticas

O fato de variáveis climáticas categorizadas terem emergido como as mais importantes nos modelos reforça a conhecida influência das condições meteorológicas na produtividade da soja. Especificamente, a importância da precipitação e radiação no período reprodutivo está alinhada com o conhecimento agronômico estabelecido, que identifica esta fase como crítica para a definição do rendimento final (FARIAS; NEPOMUCENO; NEUMAIER, 2021).

3.6 Limitações do Estudo

Algumas limitações devem ser consideradas na interpretação dos resultados obtidos neste estudo, em relação às variáveis não consideradas, é importante destacar que fatores agronômicos relevantes como a incidência de pragas e doenças, práticas de manejo específicas adotadas pelos produtores e a água disponível no solo não puderam ser incluídos nos modelos, o que pode representar fontes de variabilidade não capturadas pelas análises. Quanto à generalização dos resultados, os modelos foram calibrados e validados especificamente para as condições edafoclimáticas e de manejo predominantes no estado do Tocantins, sendo necessária validação adicional para garantir sua aplicabilidade em outras regiões com características distintas.

3.7 Direções para Pesquisas Futuras

Com base nas limitações identificadas e nos resultados obtidos, uma direção promissora para pesquisas futuras é a integração de dados de sensoriamento remoto e imagens de satélite, permitindo capturar variáveis espaciais complementares e ampliar a capacidade preditiva dos modelos. Além disso, o desenvolvimento de modelos específicos calibrados para diferentes regiões e tipos de solo surge como uma estratégia relevante, sobretudo para considerar de forma mais precisa a heterogeneidade pedoclimática que caracteriza os ambientes agrícolas.

Outra frente importante envolve a implementação de sistemas de predição em tempo real, capazes de oferecer suporte ágil ao processo de tomada de decisão no manejo agrícola. Tais sistemas podem contribuir para intervenções mais rápidas e eficientes em campo, aumentando a precisão das recomendações e favorecendo a adoção de práticas agrícolas mais inteligentes e sustentáveis.

Mais uma ampliação deste trabalho, seria a recomendação do ciclo de cultivares para o plantio em determinada região com base nas previsões climáticas e na água disponível do solo, algo que pode ser encontrado em outras bases de dados.

3.8 Contribuições do Trabalho

Este estudo contribui significativamente ao fornecer uma comparação abrangente de múltiplos algoritmos de aprendizado de máquina aplicados em condições reais de cultivo no Tocantins, demonstrando a viabilidade prática da predição de produtividade utilizando exclusivamente variáveis categóricas derivadas de dados climáticos. Adicionalmente, a pesquisa oferece informações valiosas sobre a relativa importância de diferentes fatores climáticos em distintas fases do ciclo fenológico da soja, enquanto disponibiliza uma metodologia replicável que pode ser adaptada para o estudo de outras culturas e regiões, ampliando assim o potencial de aplicação dos resultados obtidos.

Os resultados obtidos representam um passo importante no sentido de desenvolver ferramentas preditivas robustas que possam apoiar a tomada de decisão no setor agrícola, contribuindo para uma agricultura mais resiliente, eficiente e rentável.

Conclusão

Este trabalho apresentou como objetivo principal aplicar e comparar diferentes técnicas de aprendizado de máquina na predição da produtividade de soja no estado do Tocantins, utilizando variáveis categóricas oriundas de dados climáticos. Os resultados obtidos permitem extrair conclusões relevantes tanto do ponto de vista metodológico quanto prático.

Em relação ao desempenho dos algoritmos, constatou-se que os métodos de *ensemble* - particularmente Bagging, XGBoost e Random Forest, apresentaram os melhores resultados. Esta superioridade dos métodos de *ensemble* está alinhada com a literatura da área e reforça sua adequação para problemas de predição em agricultura.

A análise de importância de variáveis revelou que fatores climáticos, especialmente aqueles relacionados à precipitação no período reprodutivo e radiação solar no período vegetativo, são determinantes cruciais para a produtividade da soja na região estudada.

Do ponto de vista prático, o nível de acurácia alcançado é promissor para aplicações de planejamento agrícola em nível regional. Os modelos desenvolvidos podem auxiliar produtores e gestores públicos em estimativas de safra, planejamento logístico e definição de políticas agrícolas.

Contudo, é importante reconhecer as limitações do estudo. A ausência de informações detalhadas sobre práticas de manejo e a especificidade regional dos modelos desenvolvidos impõem cautela na generalização dos resultados. Pesquisas futuras deveriam incorporar variáveis adicionais e validar os modelos em outras regiões.

Como perspectiva final, recomenda-se a implementação dos modelos em plataformas acessíveis aos produtores e técnicos agrícolas, potencialmente integrados com sistemas de monitoramento em tempo real. Esta direção poderia maximizar o impacto prático dos resultados obtidos, transformando conhecimento acadêmico em ferramentas concretas de apoio ao desenvolvimento agrícola regional.

Referências

ARLOT, S.; CELISSE, A. A survey of cross-validation procedures for model selection. **Statistics Surveys**, v. 4, p. 40–79, 2010. Citado 2 vezes nas páginas 20 e 24.

ASSAD, E. D.; PINTO, H. S. **Mudança Climática e Desafios para a Agricultura Brasileira**. [S.l.]: Editora Embrapa, 2021. Citado na página 14.

ATHANASIOS, P. *et al.* A comparison of machine learning methods for crop disease detection and severity estimation. **Agronomy**, MDPI, v. 12, n. 3, p. 1–25, 2022. Citado na página 18.

BACCHI, M. R. P.; CALDAS, M. M. Percepção de riscos e estratégias de gestão em propriedades rurais. **Revista de Economia e Sociologia Rural**, v. 59, n. 2, 2021. Citado na página 14.

BARROS, G. S. d. C. Análise de políticas agrícolas e formação de preços. In: **Agricultura, Transformação Produtiva e Sustentabilidade**. [S.l.]: Editora FGV, 2022. Citado na página 14.

BELGIU, M.; DRăGUţ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, Elsevier, v. 114, p. 24–31, 2016. Citado na página 17.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, p. 123–140, 1996. Citado 2 vezes nas páginas 18 e 25.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, p. 5–32, 2001. Citado 2 vezes nas páginas 17 e 25.

BREIMAN, L. *et al.* **Classification and Regression Trees**. [S.l.]: Routledge, 1984. Citado 2 vezes nas páginas 17 e 25.

BREIMAN, L. *et al.* **Classification and Regression Trees**. 1st. ed. [S.l.]: Routledge, 2017. Citado na página 17.

CASTRO, S. G. *et al.* Logística de escoamento da soja no brasil: desafios e oportunidades. **Revista Transporte e Território**, v. 25, 2023. Citado na página 14.

Centro de Estudos Avançados em Economia Aplicada (CEPEA). **PIB do Agronegócio Brasileiro**. Piracicaba: [s.n.], 2023. Citado na página 14.

CHAI, T.; DRAXLER, R. Root mean square error (rmse) or mean absolute error (mae)?-arguments against avoiding rmse in the literature. **Geoscientific Model Development**, v. 7, n. 3, p. 1247–1250, 2014. Citado 2 vezes nas páginas 19 e 21.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.: s.n.], 2016. p. 785–794. Citado 2 vezes nas páginas 18 e 25.

Companhia Nacional de Abastecimento (CONAB). **Acompanhamento da safra brasileira de grãos: Safra 2023/24**. Brasília: [s.n.], 2024. Citado na página 14.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995. Citado 2 vezes nas páginas 18 e 25.

DEMATTE, J. A. *et al.* Geospatial soil sensing system (geos3): A powerful data mining procedure to retrieve soil information from satellite data. **Remote Sensing of Environment**, Elsevier, v. 212, p. 161–175, 2018. Citado na página 17.

FARIAS, J.; NEPOMUCENO, A.; NEUMAIER, N. Ecofisiologia da soja. **Circular Técnica, EMBRAPA Soja**, n. 48, p. 1–9, 2021. Citado na página 30.

FENG, L. *et al.* A recommendation model for fertilization based on xgboost and iot data. **Computers and Electronics in Agriculture**, Elsevier, v. 185, p. 106125, 2021. Citado na página 18.

Food and Agriculture Organization of the United Nations (FAO). **FAO Statistical Yearbook 2023**. Roma: [s.n.], 2023. Citado na página 14.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. Citado na página 19.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer, 2009. Citado na página 18.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. [S.l.]: Springer, 2017. Citado na página 15.

HENGL, T. *et al.* Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, PeerJ Inc., v. 6, p. e5518, 2018. Citado na página 17.

HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. **International Journal of Forecasting**, v. 22, p. 679–688, 2006. Citado 2 vezes nas páginas 19 e 21.

Instituto Brasileiro de Geografia e Estatística (IBGE). **Produção agrícola municipal 2022**. Rio de Janeiro: [s.n.], 2022. Citado na página 14.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. 2nd. ed. [S.l.]: Springer, 2021. Citado na página 15.

KLOMPENBURG, T. V.; KASSAHUN, A.; CATAL, C. Crop yield prediction using machine learning: A systematic literature review. **Computers and Electronics in Agriculture**, Elsevier, v. 177, p. 105709, 2020. Citado 3 vezes nas páginas 16, 17 e 18.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. [S.l.]: Springer, 2013. Citado na página 24.

LIAKOS, K. G. *et al.* Machine learning in agriculture: A review. **Sensors**, MDPI, v. 18, n. 8, p. 2674, 2018. Citado 2 vezes nas páginas 16 e 17.

LLOYD, S. Least squares quantization in pcm. **IEEE Transactions on Information Theory**, v. 28, p. 129–137, 1982. Citado na página 23.

MITCHELL, T. M. **Machine Learning**. [S.l.]: McGraw-Hill, 1997. Citado na página 16.

MOHANTY, S. P.; HUGHES, D. P.; SALATHé, M. Using deep learning for image-based plant disease detection. **Frontiers in Plant Science**, Frontiers Media SA, v. 7, p. 1419, 2016. Citado na página 17.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to Linear Regression Analysis**. 5th. ed. [S.l.]: Wiley, 2012. Citado na página 20.

NASA Power Project. **Prediction of Worldwide Energy Resources (POWER) Data Access Viewer**. 2024. Disponível em: <https://power.larc.nasa.gov>. Citado 2 vezes nas páginas 14 e 22.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>. Citado na página 24.

RODRIGUES, M. L.; SANTANA, C. A.; GUIMARãES, F. G. Machine learning for agricultural yield prediction in brazil. **Computers and Electronics in Agriculture**, Elsevier, v. 184, p. 106106, 2021. Citado 2 vezes nas páginas 16 e 17.

SHARMA, P.; KUMAR, R.; SINGH, S. **Machine Learning for Agriculture: Methods and Applications**. [S.l.]: Elsevier, 2025. Citado na página 16.

SILVA, A.; LIMA, B.; SOUZA, C. **Dinâmicas da expansão agrícola no Tocantins**. Palmas: Editora Universitária do Tocantins, 2021. Citado na página 14.

THENKABAIL, P. S. (Ed.). **Remote Sensing Handbook (Three-Volume Set)**. Boca Raton, FL: CRC Press, 2018. Citado na página 18.

THERNEAU, T. M.; ATKINSON, E. J. **An Introduction to Recursive Partitioning Using the RPART Routines**. [S.l.]: Mayo Foundation, 2019. Citado na página 25.

VENKATESH, A.; PARTHIBAN, R. Comparative evaluation and comprehensive analysis of machine learning models for regression analysis. **Data Intelligence**, MIT Press, v. 4, n. 3, p. 620–641, 2022. Citado na página 19.

WANG, Y. *et al.* Predicting soil moisture for smart irrigation using xgboost and iot sensor data. **Agricultural Water Management**, Elsevier, v. 255, p. 107050, 2021. Citado na página 18.

WOLFERT, S. *et al.* Big data in smart farming–a review. **Agricultural Systems**, Elsevier, v. 153, p. 69–80, 2017. Citado na página 16.

ZHANG, X.; LI, Y.; WANG, J. Application of machine learning in crop yield prediction. **Agricultural Systems**, v. 175, p. 50–60, 2019. Citado 3 vezes nas páginas 15, 16 e 29.