Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

MODELO PREDITIVO DO CICLO DE COLHEITA DE SEMENTES DE MILHO

Victor Augusto Borin Palma

Uberlândia-MG
2026

Victor Augusto Borin Palma

MODELO PREDITIVO DO CICLO DE COLHEITA DE SEMENTES DE MILHO

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Rogério de Melo Costa Pinto

Uberlândia-MG

2026

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, 05 de agosto de 2026

BANCA EXAMINADORA

Rogério de Melo Costa Pinto

Marcelo Tavares

Ednaldo Carvalho Guimarães

Uberlândia-MG

2026

Agradecimentos

Agradeço ao meu orientador, Prof. Rogério de Melo Costa Pinto, pela orientação, paciência e disponibilidade ao longo de todo o desenvolvimento deste trabalho. Agradeço também aos professores da banca examinadora, pela leitura cuidadosa e pelas contribuições que enriqueceram esta pesquisa.

Expresso minha sincera gratidão ao Lucas Queiroz, pelo compartilhamento de seu conhecimento técnico e pelo apoio prestado ao longo do desenvolvimento deste trabalho.

Por fim, agradeço à minha família, pelo apoio incondicional, incentivo e compreensão durante toda a minha graduação. Seu suporte foi fundamental para que eu pudesse superar os desafios dessa trajetória e alcançar esta conquista.

Resumo

A previsão do ciclo de colheita de sementes de milho é fundamental para a qualidade fisiológica do produto e para a eficiência operacional da produção e do beneficiamento. Este trabalho teve como objetivo desenvolver um modelo preditivo para estimar o ciclo de colheita de sementes de milho, utilizando técnicas de aprendizado de máquina automatizado por meio da biblioteca *AutoGluon*, com base em variáveis agronômicas, climáticas, fenológicas e características de híbridos. A base de dados histórica, proveniente de diferentes regionais produtoras, foi organizada em quatro grupos de variáveis (contexto produtivo, clima acumulado, fenologia/fisiologia e indicadores pré-colheita) e submetida a um processo de seleção de variáveis antes do treinamento. O *AutoGluon* avaliou múltiplos algoritmos – incluindo *Random Forest*, *Extra Trees*, *LightGBM*, *XGBoost*, *CatBoost* e redes neurais – e combinou suas previsões por meio de uma estratégia de *stacking* multinível, resultando no modelo final *WeightedEnsemble_L3_FULL*. Os resultados mostraram desempenho satisfatório, com coeficiente de determinação (R^2) de 0,918134: no consolidado Brasil, aproximadamente 72,6% dos contratos apresentaram erro de previsão igual ou inferior a 5 dias, com diferença média de 2,64 dias nessa faixa. A análise por regional revelou desempenho inferior na Regional Sul, associado à maior variabilidade climática local. Conclui-se que a abordagem de *AutoML* baseada em *ensemble learning* constitui uma alternativa promissora para apoiar o planejamento agrícola e a logística da indústria de beneficiamento de sementes de milho no Brasil.

Palavras-chave: Aprendizado de máquina, AutoML, AutoGluon, sementes de milho, ciclo de colheita.

Abstract

Predicting the harvest cycle of corn seeds is essential for the physiological quality of the product and for the operational efficiency of production and processing. This work aimed to develop a predictive model to estimate the corn seed harvest cycle, using automated machine learning techniques through the *AutoGluon* library, based on agronomic, climatic, phenological, and hybrid quality variables. The historical database, gathered from different producing regions, was organized into four groups of variables (production context, accumulated climate, phenology/physiology, and pre-harvest indicators) and underwent a variable selection process before training. *AutoGluon* evaluated multiple algorithms – including *Random Forest*, *Extra Trees*, *LightGBM*, *XGBoost*, *CatBoost*, and neural networks – and combined their predictions through a multi-level *stacking* strategy, resulting in the final model, *WeightedEnsemble_L3_FULL*. Results showed satisfactory performance, with a coefficient of determination (R^2) of 0.918134: for the Brazil-wide consolidated data, approximately 72.6% of contracts had a prediction error of 5 days or less, with an average difference of 2.64 days within that range. Regional analysis revealed lower performance in the South region, associated with greater local climate variability. It is concluded that the *AutoML* approach based on *ensemble learning* constitutes a promising alternative to support agricultural planning and processing industry logistics for corn seeds in Brazil.

Keywords: Machine learning, AutoML, AutoGluon, corn seeds, harvest cycle.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações e Símbolos

Lista de Abreviações

ANN Redes Neurais Artificiais (*Artificial Neural Networks*)

AutoML Aprendizado de Máquina Automatizado (*Automated Machine Learning*)

EVI Índice de Vegetação Aprimorado (*Enhanced Vegetation Index*)

GDU Graus-Dia Acumulados (*Growing Degree Units*)

GNDVI Índice de Vegetação por Diferença Normalizada Verde (*Green Normalized Difference Vegetation Index*)

LSTM Memória de Curto Prazo Longa (*Long Short-Term Memory*)

MAE Erro Médio Absoluto (*Mean Absolute Error*)

MAPE Erro Percentual Absoluto Médio (*Mean Absolute Percentage Error*)

MLP Perceptron Multicamadas (*Multi-Layer Perceptron*)

NDVI Índice de Vegetação por Diferença Normalizada (*Normalized Difference Vegetation Index*)

RMSE Raiz do Erro Quadrático Médio (*Root Mean Squared Error*)

SAVI Índice de Vegetação Ajustado ao Solo (*Soil Adjusted Vegetation Index*)

SVM Máquinas de Vetores de Suporte (*Support Vector Machines*)

VARI Índice Visível Resistente aos Efeitos Atmosféricos (*Visible Atmospherically Resistant Index*)

Lista de Símbolos

R^2 Coeficiente de determinação

1. Introdução

O milho consolidou-se como o cereal de maior volume de produção mundial, exercendo um papel central na economia e na segurança alimentar. No contexto da produção de sementes, a etapa da colheita é considerada o momento mais crítico, pois define o rendimento final e a qualidade fisiológica do produto. Estimativas apontam que a falta de um planejamento eficiente ou falhas na execução podem gerar perdas de produtividade durante a colheita mecanizada.

Além da relevância econômica da cultura, a crescente digitalização do agronegócio tem impulsionado o uso de ferramentas analíticas e modelos preditivos para apoiar a tomada de decisão. Nesse contexto, técnicas de aprendizado de máquina vêm sendo amplamente empregadas para modelar fenômenos agrícolas complexos, permitindo identificar padrões em grandes volumes de dados e gerar previsões mais precisas para atividades operacionais e estratégicas.

A problemática central reside na variabilidade do ciclo de maturação da cultura, influenciado por fatores genéticos e ambientais. Sem uma previsão assertiva da data de colheita, o setor de campo enfrenta gargalos logísticos severos: a organização de máquinas e equipes torna-se reativa, o que pode levar à ociosidade do maquinário ou à saturação da capacidade operacional.

Essa variabilidade é ainda mais evidente em um país de dimensões continentais como o Brasil, onde diferenças de temperatura, precipitação, radiação solar e condições edafoclimáticas influenciam diretamente o desenvolvimento fenológico das plantas. Consequentemente, híbridos semelhantes podem apresentar comportamentos distintos entre regiões, safras e ambientes produtivos, aumentando a complexidade do planejamento agrícola.

Paralelamente, a indústria de beneficiamento depende da previsibilidade para garantir a eficiência dos processos de recepção e secagem. A colheita de sementes exige que o grão entre na planta de beneficiamento com níveis específicos de umidade (frequentemente entre 18% e 25%) para garantir o máximo vigor e germinação. Um modelo preditivo robusto permite que a indústria prepare sua infraestrutura de secadores e silos no momento exato, evitando perdas pós-colheita. Assim, a aplicação de uma análise preditiva na produção de semente de milho surge como uma ferramenta essencial para a transição de uma gestão reativa para uma gestão proativa de riscos.

Embora existam estudos que utilizem técnicas de aprendizado de máquina para previsão de produtividade, monitoramento de culturas e suporte à agricultura de precisão, ainda são limitados os trabalhos voltados especificamente à previsão do ciclo de colheita de sementes de milho considerando simultaneamente fatores climáticos, fenológicos, fisiológicos e operacionais – lacuna discutida em maior profundidade no Capítulo 2 (Fundamentação Teórica).

Parte-se da hipótese de que a combinação de informações climáticas, características dos híbridos e indicadores obtidos durante o desenvolvimento da cultura permite prever o ciclo de colheita com precisão suficiente para apoiar o planejamento operacional no campo e a preparação da indústria de beneficiamento, reduzindo incertezas e aumentando a eficiência da cadeia produtiva.

1.1 Objetivos

1.1.1 Objetivo Geral

Desenvolver um modelo preditivo para estimar o ciclo de colheita de sementes de milho, utilizando técnicas de aprendizado de máquina por meio da biblioteca *AutoGluon*, com base em variáveis agronômicas, climáticas, fenológicas e características de híbridos, visando otimizar o planejamento operacional no campo e a logística da indústria de beneficiamento.

1.1.2 Objetivos Específicos

- Analisar a influência de variáveis meteorológicas, como precipitação, temperatura e radiação solar, sobre o ciclo de colheita das sementes de milho.
- Avaliar o impacto de variáveis fenológicas, como GDU e tempo entre plantio e despendoamento, na previsão do ciclo de colheita.
- Investigar a contribuição de indicadores de qualidade dos híbridos, como vigor, germinação e índices obtidos em pré-colheita, na acurácia do modelo preditivo.
- Aplicar técnicas de aprendizado de máquina automatizado utilizando o *AutoGluon* para treinar, validar e selecionar o melhor modelo preditivo.
- Analisar os resultados obtidos por regional.
- Avaliar o desempenho do modelo por meio de métricas estatísticas adequadas, verificando sua capacidade de generalização em diferentes ambientes e safras.

2. Fundamentação Teórica

2.1 Produção de Sementes de Milho e Qualidade Fisiológica

A produção de sementes de milho no Brasil representa um dos segmentos mais tecnificados e economicamente vitais do agronegócio nacional. Diferente da produção de grãos destinados ao consumo ou à indústria de rações, a multiplicação de sementes exige um rigoroso controle de qualidade que começa muito antes da semeadura e se estende até a entrega do produto final ao agricultor [10]. O milho híbrido, em particular, é apontado como responsável por aproximadamente 50% do rendimento final de uma lavoura, o que coloca sobre a indústria de sementes a responsabilidade de garantir que o potencial genético e a qualidade fisiológica (traduzida em germinação e vigor) sejam preservados integralmente durante todo o ciclo produtivo [3].

A previsão do ciclo de colheita de sementes de milho é relevante porque o momento da colheita interfere diretamente na qualidade fisiológica da semente, na eficiência operacional da produção e no desempenho da etapa de beneficiamento. O Ministério da Agricultura e Pecuária destaca que, no período de colheita, a semente deve estar fisiologicamente madura e suficientemente seca para permitir uma colheita segura, ou então ser colhida úmida e submetida à secagem artificial. Estudos científicos também mostram que as condições de secagem e armazenamento influenciam a manutenção da qualidade fisiológica das sementes e dos grãos de milho, reforçando que a colheita não é apenas uma etapa agrícola, mas um ponto decisivo para preservar vigor, germinação e valor comercial [12].

A justificativa se fortalece ainda mais quando se considera a heterogeneidade ambiental do Brasil. O país apresenta grande diversidade climática e biológica, com diferenças marcantes entre regiões quanto a temperatura, regime de chuvas, umidade e condições de solo. O IBGE destaca a existência de climas equatorial, tropical e temperado no território nacional, enquanto a Embrapa ressalta que o zoneamento agroecológico se baseia justamente nas potencialidades e vulnerabilidades ambientais de cada região, considerando clima, solo, vegetação e geomorfologia. Na prática, isso significa que o ciclo de colheita de sementes de milho não ocorre da mesma forma em todas as regiões do país, pois o ambiente altera a taxa de secagem, o avanço fenológico e a janela ideal de colheita [5].

Nesse contexto, o presente trabalho se justifica pelo uso de aprendizado de máquina com *AutoGluon* em Python para prever o ciclo de colheita de sementes de milho a partir de variáveis

agronômicas, climáticas, fenológicas e de híbridos. As variáveis relacionadas ao ambiente, como precipitação acumulada, radiação solar, temperatura e indicadores de vegetação, são coerentes com a literatura, porque o período entre maturação fisiológica e colheita depende fortemente das condições meteorológicas e da perda de umidade do grão. Já as variáveis pré-colheita, baseadas em testes amostrais dos híbridos, são especialmente importantes para incorporar diferenças de vigor, germinação e desempenho fisiológico entre materiais genéticos. Essa combinação amplia a capacidade do modelo de representar tanto a realidade do campo quanto as exigências da indústria de beneficiamento, que precisa se preparar para receber as sementes no momento correto, com menor risco de sobrecarga operacional e melhor preservação da qualidade [4].

2.2 Aprendizado de Máquina Aplicado à Agricultura

A literatura recente demonstra o crescente interesse pela aplicação de técnicas de aprendizado de máquina em problemas agrícolas. Entretanto, a maior parte dos estudos concentra-se na previsão de produtividade, monitoramento de culturas, classificação de doenças ou estimativas de rendimento.

Trabalhos como os de Liu et al. [7] exploram redes neurais recorrentes para previsão de tempo de colheita utilizando variáveis climáticas e produtivas, enquanto Almeida et al. [2] investigam diferentes algoritmos de aprendizado de máquina aplicados ao planejamento operacional agrícola. De forma semelhante, Kavitha et al. [6] avaliam abordagens baseadas em modelos *ensemble* para melhorar a capacidade preditiva em sistemas agrícolas complexos.

Nos últimos anos, a agricultura de precisão tem incorporado técnicas de aprendizado de máquina para apoiar decisões relacionadas ao manejo, produtividade e planejamento operacional das lavouras. Em problemas de previsão agrícola, diferentes algoritmos vêm sendo empregados, incluindo Regressão Linear, *Random Forest*, *Support Vector Machines* (SVM), Redes Neurais Artificiais (ANN), *Gradient Boosting* e métodos *ensemble*. Esses modelos são capazes de identificar padrões complexos entre variáveis ambientais, agronômicas e fisiológicas, produzindo estimativas mais precisas do que métodos estatísticos tradicionais em diversos cenários.

A avaliação do desempenho de modelos preditivos pode ser realizada por diferentes métricas, sendo o coeficiente de determinação (R^2) uma das mais utilizadas para mensurar a qualidade do ajuste. Essa métrica representa a proporção da variabilidade da variável resposta que é explicada pelo modelo em relação à variabilidade total dos dados. Seus valores variam entre 0 e 1, em que valores próximos de 1 indicam que o modelo consegue explicar grande parte da variação observada, enquanto valores próximos de 0 indicam baixa capacidade explicativa. Dessa forma, quanto maior o valor do R^2, melhor é o ajuste do modelo aos dados, embora um elevado coeficiente de determinação não garanta, por si só, um bom desempenho preditivo em novos conjuntos de dados, sendo recomendável sua interpretação em conjunto com outras métricas de avaliação. Essas métricas permitem comparar diferentes algoritmos e identificar aqueles mais adequados para cada contexto agrícola.

Entre os estudos relacionados à previsão de colheita, Liu et al. [7] propuseram um modelo denominado LSTMFS, que combina seleção de variáveis com redes neurais do tipo *Long Short-Term Memory* (LSTM) para previsão do tempo de colheita. Os autores utilizaram dados reais de produção e variáveis climáticas, incluindo temperatura acumulada, radiação solar, precipitação, umidade relativa do ar e velocidade do vento. Os resultados demonstraram desempenho superior aos modelos tradicionais de redes neurais recorrentes, alcançando RMSE de 0,199 e MAPE de 4,84%, evidenciando o potencial de arquiteturas profundas para capturar relações temporais associadas ao desenvolvimento das culturas.

Em uma abordagem voltada diretamente à previsão da data de colheita de soja e milho, Accorsi [1] comparou modelos de Regressão Linear, *Random Forest*, Redes Neurais MLP e *XGBoost* utilizando variáveis relacionadas ao solo, fertilidade, matéria orgânica, temperatura, umidade, precipitação e radiação solar. O estudo demonstrou que, mesmo diante da crescente adoção de algoritmos complexos, a Regressão Linear apresentou o melhor desempenho, com erro médio absoluto de aproximadamente 3,84 dias. Os resultados reforçam que a escolha do modelo depende não apenas da complexidade do algoritmo, mas também das características da base de dados utilizada.

Outro trabalho relevante foi desenvolvido por Marra [9], que investigou a predição da produtividade da soja utilizando imagens multiespectrais de satélite e índices de vegetação. Foram avaliados modelos de Regressão Linear, *Random Forest* e Redes Neurais Artificiais a partir dos índices NDVI, GNDVI (*Green Normalized Difference Vegetation Index* – Índice de Vegetação por Diferença Normalizada Verde), VARI (*Visible Atmospherically Resistant Index* – Índice Visível Resistente aos Efeitos Atmosféricos), SAVI (*Soil Adjusted Vegetation Index* – Índice de Vegetação Ajustado ao Solo) e EVI (*Enhanced Vegetation Index* – Índice de Vegetação Aprimorado). O desempenho dos modelos foi avaliado por meio das métricas MAE, MAPE e RMSE. Os resultados indicaram que as Redes Neurais Artificiais apresentaram a maior capacidade preditiva, especialmente quando associadas aos índices NDVI e GNDVI, demonstrando a importância da integração entre sensoriamento remoto e aprendizado de máquina para aplicações agrícolas.

De forma semelhante, Almeida, Silva e Simões [2] desenvolveram um modelo de aprendizado de máquina para prever a produtividade de colhedoras florestais em plantações de eucalipto. O estudo avaliou 24 algoritmos pertencentes a diferentes grupos de aprendizado supervisionado, incluindo métodos lineares, árvores de decisão, redes neurais artificiais e técnicas *ensemble*. Foram utilizadas variáveis operacionais e meteorológicas, como idade do povoamento, volume individual das árvores, densidade, temperatura, umidade, precipitação, radiação solar e velocidade do vento. O modelo *CatBoost* apresentou o melhor desempenho, alcançando coeficiente de determinação de 0,70, demonstrando que informações climáticas podem aumentar significativamente a capacidade preditiva de modelos voltados ao planejamento operacional da colheita.

O crescente interesse por técnicas *ensemble* também é observado em Kavitha et al. [6], que propuseram um sistema de previsão e monitoramento agrícola baseado na integração de múltiplos algoritmos de aprendizado de máquina. Os autores destacam que modelos *ensemble*

apresentam maior robustez e capacidade de generalização ao combinar previsões de diferentes algoritmos, reduzindo erros individuais e melhorando a estabilidade das estimativas. Essa característica é particularmente relevante em sistemas agrícolas, nos quais as relações entre clima, solo, genética e manejo frequentemente apresentam elevada complexidade e não linearidade.

Embora esses estudos apresentem resultados promissores, observa-se que a maior parte da literatura se concentra na previsão de produtividade, monitoramento de culturas ou estimativas gerais de rendimento. Trabalhos especificamente voltados à previsão do ciclo de colheita de sementes de milho ainda são escassos, especialmente aqueles que integram simultaneamente informações climáticas, fenológicas, fisiológicas e operacionais obtidas em ambientes produtivos reais. Além disso, poucos estudos incorporam variáveis relacionadas à qualidade fisiológica dos híbridos, como vigor e germinação, provenientes de avaliações pré-colheita.

Dessa forma, a principal contribuição deste estudo consiste na aplicação de uma abordagem de *AutoML* baseada em *ensemble learning*, utilizando a plataforma *AutoGluon*, para prever o ciclo de colheita de sementes de milho em escala operacional. Diferentemente dos trabalhos encontrados na literatura, a proposta integra dados históricos de múltiplas regiões produtoras, variáveis meteorológicas, características dos híbridos e indicadores pré-colheita, buscando gerar previsões mais precisas e aplicáveis tanto à gestão agrícola quanto ao planejamento industrial do beneficiamento de sementes.

3. Metodologia

O presente estudo caracteriza-se como uma pesquisa aplicada de abordagem quantitativa, cujo objetivo é desenvolver um modelo preditivo para estimar o ciclo de colheita de sementes de milho. Do ponto de vista estatístico, trata-se de um problema de regressão supervisionada, no qual busca-se prever uma variável contínua (ciclo de colheita) a partir de um conjunto de variáveis explicativas relacionadas às condições ambientais, características dos híbridos e indicadores agronômicos observados ao longo do ciclo produtivo.

3.1 Base de Dados e Variáveis

A base de dados utilizada neste estudo, proveniente de uma empresa produtora de sementes, é composta por informações históricas de produção de sementes de milho oriundas de diferentes regionais produtoras. Os dados contemplam variáveis agronômicas, climáticas, fenológicas, fisiológicas e operacionais, permitindo representar os diversos fatores que influenciam o desenvolvimento da cultura e a definição do momento ideal de colheita.

As variáveis explicativas testadas foram organizadas em quatro grupos principais. O primeiro grupo corresponde às variáveis de contexto produtivo, incluindo ciclo de despendoamento, safra, estação do ano, regional e tipo de colheita. O segundo grupo compreende variáveis climáticas acumuladas entre os períodos de plantio, despendoamento e colheita, incluindo precipitação acumulada, temperatura média, radiação solar descendente e radiação solar líquida. O terceiro grupo reúne variáveis fenológicas e fisiológicas, com destaque para o GDU (Graus-Dia Acumulados) e o NDVI (Índice de Vegetação por Diferença Normalizada). O GDU é uma métrica agrometeorológica que quantifica o acúmulo de energia térmica diária absorvida pela planta, sendo fundamental para mapear a taxa de desenvolvimento biológico do milho desde a emergência até a maturação fisiológica e o ponto ideal de colheita. Já o NDVI atua como um indicador remoto de biomassa e atividade fotossintética; na prática, ele mensura o "nível de verde" e o vigor da vegetação avaliando a forma como a planta reflete a luz solar. Esse índice é crucial para monitorar a saúde da lavoura e a progressão do processo de senescência (secagem) da cultura na fase de pré-colheita. Além desses índices, o grupo inclui medições de vigor e germinação. Por fim, o quarto grupo contempla variáveis pré-colheita obtidas por meio de avaliações amostrais dos híbridos, incluindo indicadores de doenças, estimativas produtivas e métricas agronômicas relacionadas ao desempenho dos materiais genéticos.

A utilização conjunta desses grupos de variáveis foi fundamentada na literatura científica

[8], que demonstra que o desenvolvimento fenológico do milho é influenciado simultaneamente por fatores climáticos, genéticos e fisiológicos. Dessa forma, buscou-se construir um conjunto de dados capaz de representar tanto a dinâmica ambiental quanto as características intrínsecas dos híbridos avaliados.

3.2 Seleção de Variáveis e Divisão da Amostra

Antes do treinamento dos modelos, os dados passaram por um processo de seleção de variáveis. Essa etapa foi conduzida utilizando dois critérios complementares: análise de correlação entre as variáveis explicativas e a variável resposta, e análise de importância das variáveis (*feature importance*) calculada pelos modelos preditivos. O objetivo foi reduzir redundâncias, eliminar atributos pouco informativos e aumentar a capacidade de generalização do modelo final.

A divisão dos dados em conjuntos de treinamento e teste foi realizada de forma estratificada, preservando a representatividade das diferentes regionais presentes na base histórica. O conjunto de treinamento foi composto pela maior parte dos registros e utilizado para construção dos modelos, enquanto o conjunto de teste correspondeu a 15% dos dados não utilizados no treinamento, sendo reservado exclusivamente para avaliação do desempenho preditivo.

3.3 Modelagem com AutoGluon

Para a modelagem foi utilizada a biblioteca *AutoGluon*, desenvolvida para aplicações de *Automated Machine Learning* (*AutoML*). Essa abordagem automatiza etapas tradicionalmente realizadas manualmente pelos cientistas de dados, incluindo pré-processamento, seleção de algoritmos, ajuste de hiperparâmetros e combinação de modelos. O objetivo do *AutoML* é identificar automaticamente a combinação de técnicas que produz o melhor desempenho para o problema analisado.

Durante o processo de treinamento, o *AutoGluon* avaliou diferentes algoritmos de regressão, incluindo modelos lineares, árvores de decisão, *Random Forest*, *Extra Trees*, *Gradient Boosting*, *LightGBM*, *XGBoost* e redes neurais artificiais. Cada algoritmo apresenta vantagens específicas na identificação de padrões presentes nos dados, sendo capaz de capturar diferentes tipos de relações entre as variáveis explicativas e a variável resposta.

Os principais algoritmos avaliados pelo *AutoGluon* podem ser agrupados em diferentes categorias. Os modelos baseados em árvores de decisão, como *Random Forest* e *Extra Trees*, são capazes de identificar relações não lineares entre as variáveis e são menos sensíveis à presença de valores extremos. O *Random Forest* constrói diversas árvores utilizando subconjuntos aleatórios dos dados e combina suas previsões por meio de média, reduzindo a variância do modelo. O *Extra Trees* utiliza uma estratégia semelhante, porém introduz maior aleatoriedade na construção das árvores, o que pode aumentar a capacidade de generalização.

Os algoritmos *LightGBM*, *XGBoost* e *CatBoost* pertencem à família dos métodos de *Gradi-*

ent Boosting. Nessa abordagem, árvores de decisão são construídas sequencialmente, de forma que cada nova árvore busca corrigir os erros cometidos pelas anteriores. O *XGBoost* destaca-se pela eficiência computacional e pelo uso de técnicas de regularização para reduzir sobreajuste. O *LightGBM* foi desenvolvido para lidar com grandes volumes de dados e possui treinamento rápido, mantendo elevada capacidade preditiva. Já o *CatBoost* apresenta vantagens no tratamento de variáveis categóricas, reduzindo a necessidade de pré-processamento e frequentemente apresentando bom desempenho em bases com atributos qualitativos.

O pacote *AutoGluon* baseia-se na técnica conhecida como *ensemble learning*, amplamente utilizada em problemas de aprendizado de máquina. O princípio dos métodos *ensemble* consiste em combinar diversos modelos individuais para produzir uma previsão final mais robusta do que aquela obtida por qualquer modelo isoladamente. A justificativa é que diferentes algoritmos tendem a cometer erros distintos; ao combinar suas previsões, parte desses erros pode ser compensada, resultando em maior precisão e estabilidade.

A combinação de modelos é construída por meio de uma estratégia denominada *stacking* multinível. Na primeira camada (L1), diversos algoritmos independentes são treinados utilizando os dados originais. Em seguida, suas previsões são utilizadas como entradas para modelos de níveis superiores, que aprendem a combinar as respostas dos modelos anteriores. Esse processo é repetido até a formação da camada final.

A escolha desse pacote é justificada por sua capacidade de capturar relações lineares e não lineares simultaneamente, reduzir a variância associada aos modelos individuais e aumentar a capacidade de generalização para novos dados. Essas características são particularmente importantes no contexto agrícola, em que fenômenos biológicos e ambientais frequentemente apresentam comportamento complexo e elevada variabilidade espacial e temporal.

3.4 Critérios de Avaliação

A avaliação dos modelos foi realizada a partir do coeficiente de determinação (R^2), que mede a proporção da variabilidade observada na variável resposta explicada pelo modelo. Adicionalmente, buscou-se minimizar a diferença média entre os valores preditos e os valores observados reais (real – predição do modelo), visando aumentar a precisão prática das estimativas.

Como critério complementar de decisão, foi considerada a proporção de contratos com erro absoluto inferior a cinco dias entre o valor previsto e o valor real. Esse indicador foi escolhido por representar diretamente a aplicabilidade operacional do modelo, uma vez que pequenas diferenças na previsão tendem a gerar impactos significativamente menores sobre o planejamento agrícola e industrial.

Por fim, os resultados obtidos foram analisados não apenas sob a perspectiva estatística, mas também considerando sua aplicabilidade prática nas diferentes regionais produtoras, avaliando o potencial da ferramenta como suporte ao planejamento logístico, à gestão de recursos operacionais e à programação da recepção de sementes na indústria de beneficiamento.

4. RESULTADOS

O modelo final selecionado foi o *WeightedEnsemble_L3_FULL*, identificado como aquele que apresentou o melhor desempenho durante o processo de validação, como mostrado na Tabela 4.1.

Tabela 4.1: Coeficiente de determinação (R^2) por modelo testado

Modelo	R^2
WeightedEnsemble_L3	0,918134
WeightedEnsemble_L2	0,916549
LightGBMXT_BAG_L1	0,914706
LightGBM_BAG_L2	0,912931
LightGBM_r131_BAG_L2	0,911378
LightGBM_BAG_L1	0,910588
LightGBMLarge_BAG_L2	0,906644
LightGBMLarge_BAG_L1	0,905372
LightGBMXT_BAG_L2	0,904009
LightGBM_r131_BAG_L1	0,676219
KNeighborsDist_BAG_L1	0,625930

Fonte: elaborado pelo autor.

O *WeightedEnsemble_L3_FULL* é o modelo final gerado pelo processo de treinamento do *AutoGluon*, combinando as previsões de diversos modelos individuais para obter um desempenho superior. Esse método utiliza a técnica de *Weighted Ensemble*, na qual as previsões de diferentes algoritmos são agregadas por meio de uma média ponderada, atribuindo pesos maiores aos modelos que apresentaram melhor desempenho durante a etapa de validação e menor influência aos modelos menos precisos. Além disso, a denominação L3 (*Level 3*) indica que o *ensemble* foi construído utilizando a estratégia de *stacking* em três níveis, em que os modelos das camadas superiores utilizam como entrada as previsões produzidas pelas camadas anteriores, aprendendo a corrigir seus erros e a combinar seus pontos fortes. Por fim, o termo *FULL* indica que, após a definição da estrutura do *ensemble*, o modelo foi retreinado utilizando 100% dos dados disponíveis para treinamento, incluindo os conjuntos originalmente destinados ao treinamento e à validação, buscando maximizar sua capacidade de generalização e desempenho em novos dados.

No presente estudo, esse modelo apresentou um coeficiente de determinação (R^2) de 0,918134, indicando que aproximadamente 91,81% da variabilidade da produtividade foi explicada pelo modelo, evidenciando um elevado nível de ajuste aos dados e uma alta capacidade

preditiva.

Dentre todas as variáveis avaliadas, foram selecionadas aquelas que apresentaram maior importância (*feature importance*) para o modelo. A importância de uma variável representa o quanto ela contribui para a capacidade preditiva do algoritmo, indicando sua influência na construção das previsões. Em outras palavras, quanto maior o valor de importância atribuído a uma variável, maior é sua contribuição para explicar a variável resposta e, consequentemente, para o desempenho do modelo. As variáveis selecionadas, bem como seus respectivos valores de importância, são apresentadas na Tabela 4.2.

Tabela 4.2: Importância das variáveis selecionadas (*feature importance*)

Feature	**Importância**
Temperatura Média	0,697223
Dias até o despendoamento	0,682088
NDVI	0,579142
Radiação Solar Máxima	0,504436
GDU Médio por dia	0,332416
Tipo de colheita	0,325772
Umidade	0,308302
Regional	0,295727
Safra	0,290754
Híbrido	0,271992
Vigor	0,266125
Germinação	0,264828
Doenças	0,252589
Precipitação Média	0,224032

Fonte: elaborado pelo autor.

Os resultados do modelo preditivo foram analisados com base na distribuição dos erros entre os valores preditos e os valores reais do ciclo de colheita, segmentados por regionais, conforme apresentado na Figura 4.1. Para facilitar a interpretação, os erros foram classificados em três faixas: diferenças menores ou iguais a 5 dias, entre 5 e 7 dias e maiores ou iguais a 7 dias. De maneira geral, observa-se que o modelo apresentou desempenho satisfatório, com predominância de previsões com erro inferior ou igual a 5 dias em todas as regionais analisadas.

No consolidado Brasil, aproximadamente 72,6% dos contratos apresentaram erro dentro dessa faixa, evidenciando boa capacidade preditiva para aplicação prática no planejamento agrícola e industrial. Ao analisar especificamente a classe de maior interesse operacional (erro menor ou igual a 5 dias), verificou-se uma diferença média de 2,64 dias entre os valores preditos e observados. Esse resultado indica que, para aproximadamente três quartos dos contratos avaliados, o modelo foi capaz de prever o ciclo de colheita com elevada proximidade em relação à data efetivamente observada. Considerando a complexidade dos fatores que influenciam o desenvolvimento da cultura do milho, esse desempenho demonstra potencial para utilização como ferramenta de apoio à tomada de decisão.

A análise regional revelou diferenças no desempenho do modelo. A Regional Sul apresentou menor taxa de acerto quando comparada às regionais Cerrados Norte e Cerrados Centro. Uma

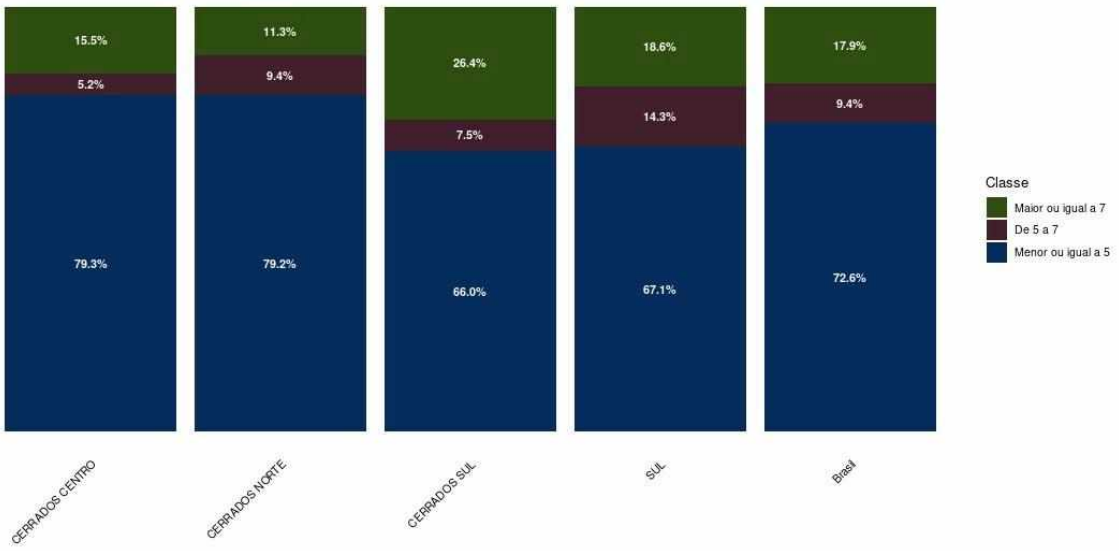

Figura 4.1: Distribuição percentual dos erros de previsão do ciclo de colheita, por regional
Fonte: elaborado pelo autor.

possível explicação para esse comportamento está relacionada à maior variabilidade climática observada nos contratos dessa regional, especialmente em relação à temperatura, precipitação e umidade. Como essas variáveis influenciam diretamente o desenvolvimento fenológico da cultura, ambientes mais heterogêneos tendem a aumentar a dificuldade de modelagem e reduzir a capacidade de generalização dos algoritmos. Esses resultados reforçam a importância das diferenças ambientais entre as regiões analisadas. As variações climáticas observadas entre estados como Mato Grosso, Goiás, Minas Gerais, Distrito Federal e São Paulo influenciam diretamente o desenvolvimento da cultura e, consequentemente, o ciclo de colheita. Essa heterogeneidade ambiental impacta o desempenho dos modelos preditivos e evidencia a necessidade de incorporar variáveis capazes de representar adequadamente as condições locais de produção.

Os resultados obtidos apresentam convergência com a literatura recente sobre aplicações de aprendizado de máquina na agricultura. Liu et al. [7], ao desenvolverem um modelo de previsão do tempo de colheita baseado em rede neural LSTM e seleção de variáveis, demonstraram que a combinação de informações climáticas e produtivas aumenta significativamente a capacidade preditiva dos modelos. Os autores utilizaram 27 variáveis relacionadas ao ambiente e à produção agrícola, alcançando baixos níveis de erro preditivo. De forma semelhante, o presente estudo utilizou informações climáticas, fenológicas, fisiológicas e operacionais, evidenciando que a integração de múltiplas fontes de dados contribui para representar adequadamente a dinâmica do ciclo de colheita.

A importância das variáveis climáticas observada neste trabalho também está alinhada aos resultados apresentados por Almeida et al. [2], que compararam 24 algoritmos de aprendizado de máquina utilizando atributos relacionados à colheita e às condições meteorológicas. Segundo os autores, fatores ambientais possuem elevada relevância para o desempenho dos modelos preditivos, resultado que também foi identificado neste estudo por meio da seleção de variáveis e da análise de importância dos atributos.

Outro aspecto relevante refere-se à utilização de modelos *ensemble*. Conforme destacado por Kavitha et al. [6], a combinação de múltiplos algoritmos tende a produzir modelos mais robustos e menos sensíveis às variações presentes nos dados. O modelo selecionado neste trabalho, *WeightedEnsemble_L3_FULL*, segue essa mesma lógica ao combinar previsões de diferentes algoritmos por meio de uma estratégia de *stacking* multinível. Os resultados observados sugerem que essa abordagem foi capaz de capturar relações complexas entre fatores climáticos, características dos híbridos e indicadores agronômicos, contribuindo para a obtenção de previsões mais estáveis.

Além do desempenho estatístico, os resultados possuem relevância operacional significativa. No campo, previsões mais precisas permitem melhor planejamento de equipes, máquinas e recursos logísticos. Na indústria de beneficiamento, a antecipação da chegada dos lotes possibilita programar secadores, silos e linhas de processamento de forma mais eficiente, reduzindo gargalos operacionais e contribuindo para a preservação da qualidade fisiológica das sementes.

5. CONCLUSÕES

O presente estudo desenvolveu um modelo preditivo para estimar o ciclo de colheita de sementes de milho, utilizando a biblioteca *AutoGluon* e uma abordagem de *AutoML* baseada em *ensemble learning*. O modelo final, *WeightedEnsemble_L3_FULL*, apresentou desempenho satisfatório, com coeficiente de determinação (R^2) de 0,918134 e aproximadamente 72,6% dos contratos avaliados apresentando erro de previsão igual ou inferior a 5 dias no consolidado Brasil, com diferença média de 2,64 dias na classe de maior interesse operacional.

Os objetivos propostos foram atendidos: analisou-se a influência de variáveis meteorológicas, fenológicas e de qualidade dos híbridos sobre o ciclo de colheita – destacando-se temperatura média, dias até o despendoamento, NDVI e radiação solar máxima entre as variáveis de maior importância; aplicaram-se técnicas de aprendizado de máquina automatizado para treinar, validar e selecionar o melhor modelo preditivo; os resultados foram analisados por regional, evidenciando diferenças de desempenho associadas à heterogeneidade climática entre regiões produtoras; e o desempenho do modelo foi avaliado por meio de métricas estatísticas adequadas (R^2, erro absoluto e proporção de contratos dentro de faixas de erro operacionalmente relevantes).

Embora os resultados demonstrem o potencial da abordagem proposta, as diferenças observadas entre regionais – com destaque para o menor desempenho na Regional Sul, associado à maior variabilidade climática local – indicam oportunidades para aprimoramentos futuros. A inclusão de novas variáveis ambientais, o aumento da base histórica e a realização de ajustes específicos por regional podem contribuir para reduzir ainda mais os erros preditivos.

Ainda assim, os resultados obtidos demonstram que a utilização de aprendizado de máquina automatizado representa uma alternativa promissora para apoiar o planejamento da produção de sementes de milho em diferentes ambientes produtivos brasileiros, com relevância tanto para o planejamento operacional no campo quanto para a logística da indústria de beneficiamento.

Referências Bibliográficas

[1] Accorsi, L.: *Estudo comparativo de modelos de machine learning para estimar a data de colheita no contexto da agricultura de precisão*, 2025.

[2] Almeida, R. O., Silva, R. B. G. da e Simões, D.: *Cut-to-Length Harvesting Prediction Tool: Machine Learning Model Based on Harvest and Weather Features.* Forests, 15(8):1398, 2024. DOI: 10.3390/f15081398. Acesso em: 30 jun. 2026.

[3] Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA): *Cultivares.* Em *Cultivo do Milho*, Sistemas de Produção, 2. Embrapa Milho e Sorgo, Sete Lagoas, 6ª ed., 2010. https://ainfo.cnptia.embrapa.br/digital/bitstream/item/27051/1/Cultivares.pdf, Acesso em: 30 jun. 2026.

[4] Erickson, N. *et al.*: *AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data*, 2020. https://arxiv.org/abs/2003.06505, arXiv preprint arXiv:2003.06505. Acesso em: 30 jun. 2026.

[5] Instituto Brasileiro de Geografia e Estatística (IBGE): *Clima no Brasil*, s.d. https://educa.ibge.gov.br/jovens/conheca-o-brasil/territorio/20644-clima.html, Acesso em: 30 jun. 2026.

[6] Kavitha, H. S. *et al.*: *Optimized Crop Prediction and Monitoring Using Ensemble Machine Learning Algorithms.* Em *2024 Second International Conference on Networks, Multimedia and Information Technology (NMITCON)*. IEEE, 2024. DOI: 10.1109/NMITCON62075.2024.10698915. Acesso em: 30 jun. 2026.

[7] Liu, S. C., Jian, Q. Y., Wen, H. Y. e Chung, C. H.: *A Crop Harvest Time Prediction Model for Better Sustainability, Integrating Feature Selection and Artificial Intelligence Methods.* Sustainability, 14(21):14101, 2022. DOI: 10.3390/su142114101. Acesso em: 30 jun. 2026.

[8] Magalhães, P. C. *et al.*: *Desenvolvimento do milho segunda safra: fatores genético-fisiológicos, plataforma de conhecimento e práticas de manejo de cultivo e uso, visando sustentabilidade de produção e produtividade no binômio soja/milho.* Rel. Téc. 258, Embrapa Milho e Sorgo, Sete Lagoas, 2020. https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/1128757/1/Documentos-258.pdf, 44 p. Acesso em: 30 jun. 2026.

[9] Marra, T. M.: *Abordagem para predição de soja por índices de vegetação e modelos de machine learning*, 2023.

[10] Ministério da Agricultura e Pecuária: *Guia de inspeção de campos para produção de sementes*, s.d. https://www.gov.br/agricultura/pt-br/assuntos/insumos-agropecuarios/insumos-agricolas/sementes-e-mudas/publicacoes-sementes-e-mudas/guia-de-inspecao-de-campos-para-producao-de-sementes.pdf, Acesso em: 30 jun. 2026.

[11] Organização das Nações Unidas para Alimentação e Agricultura (FAO): *World Food Situation – Crop Prospects and Food Situation*, s.d. https://www.fao.org/worldfoodsituation/, Acesso em: 30 jun. 2026.

[12] Silva, R. O. *et al.*: *Planejamento otimizado da produção e logística de empresas produtoras de sementes de milho: um estudo de caso*. Gestão & Produção, 2008. https://www.scielo.br/j/gp/a/KKQhN38GGwjrLX8Y3rGtqRy/, Acesso em: 30 jun. 2026.

[13] Verde AgriTech: *Ciclo do milho: entenda as fases de desenvolvimento da cultura*, s.d. https://blog.verde.ag/pt/nutricao-de-plantas/ciclo-do-milho/, Acesso em: 30 jun. 2026.