ESTUDO DA PRODUÇÃO DE SOJA NO PARANÁ (2020-2024): ANÁLISE E PREVISÃO COM DADOS DO LPSA/IBGE

Projeto em Business Intelligence e Analytics

Fase 01

Sabrina Dantas de Castro

1. Introdução

A soja destaca-se como uma das commodities mais importantes para a economia agrícola do Brasil e do estado do Paraná. Segundo a Secretaria da Agricultura e do Abastecimento do Paraná, o Brasil consolidou-se nos últimos anos como líder mundial na produção de soja, respondendo por cerca de 40% da produção global [SEAB/PR - Secretaria da Agricultura e do Abastecimento do Paraná 2023]. Dentro desse contexto, o Paraná ocupa posição de destaque: é o segundo maior produtor de soja do país, tipicamente responsável por aproximadamente 14% da safra nacional. Por exemplo, na safra de 2022/23 o estado tinha previsão de colher 20,7 milhões de toneladas, evidenciando sua relevância no cenário agrícola brasileiro [SEAB/PR - Secretaria da Agricultura e do Abastecimento do Paraná 2023]. Essa elevada participação significa que decisões de produção no Paraná impactam não apenas o mercado interno, mas também a oferta global de soja. Vale ressaltar que o Brasil, além de maior produtor, também é o maior exportador de soja do mundo, o que reforça a importância econômica da soja em termos de geração de divisas e desenvolvimento regional [CONAB - Companhia Nacional de Abastecimento 2025].

Diante da magnitude dessa cultura, torna-se fundamental aprimorar as técnicas de previsão de safra. Prever com acurácia a produção de soja tem implicações diretas no planejamento logístico, na formação de preços e na definição de políticas de abastecimento. Estimativas confiáveis de produção são essenciais para o planejamento estratégico de diversos agentes, orientando desde o armazenamento e escoamento da safra até políticas públicas de apoio ao produtor. Órgãos oficiais, como a Companhia Nacional de Abastecimento (Conab) e o Instituto Brasileiro de Geografia e Estatística (IBGE), divulgam regularmente prognósticos e levantamentos de safra justamente para subsidiar essas decisões. Em particular, o IBGE realiza o Levantamento Sistemático da Produção Agrícola (LSPA), uma pesquisa de acompanhamento mensal que fornece estimativas atualizadas de área e produção das principais culturas. O LSPA reúne informações por meio de comissões municipais e estaduais, consolidando dados em nível de estado e avaliando-os em âmbito nacional. Por se tratar de um levantamento oficial e padronizado, os dados do LSPA/IBGE fornecem uma base confiável e coerente para análises temporais da produção agrícola.

Entretanto, qual é o melhor método para utilizar esses dados na previsão de safras futuras? A literatura em previsão de safras aponta diversas abordagens possíveis – desde métodos estatísticos clássicos até algoritmos modernos de aprendizado de máquina. Cada método possui pressupostos e características distintas, podendo apresentar melhor ou pior desempenho a depender das propriedades da série temporal e do horizonte de previsão. Neste projeto, serão aplicados três diferentes algoritmos de previsão – ARIMA, Holt-Winters e Random Forest – aos dados agregados da produção de soja no Paraná, disponibilizados pelo LSPA/IBGE, para o período de 2020 a 2024. O objetivo central é avaliar qual dos modelos gera as previsões mais acuradas para a produção estadual de soja, bem como identificar as vantagens e limitações de cada abordagem no contexto dos dados disponíveis.

Com isso, evidencia-se a importância do tema, a escolha do LSPA/IBGE como fonte de dados e a formulação do problema de previsão da produção de soja no Paraná. Assim, o estudo concentra-se na avaliação comparativa de diferentes métodos de previsão, destacando suas potencialidades e limitações no contexto da produção de soja no Paraná.

2. Referencial Teórico

Para embasar a pesquisa, esta seção aborda três eixos principais: (i) a produção de soja no Paraná e no Brasil, contextualizando sua relevância; (ii) conceitos de estatística e séries temporais aplicados à previsão agrícola; e (iii) fundamentos dos três algoritmos de previsão selecionados – ARIMA, Holt-Winters e Random Forest.

Produção de Soja no Paraná e no Brasil

A cultura da soja experimentou forte expansão nas últimas décadas, tornando-se carro-chefe do agronegócio brasileiro. No Paraná, a soja é cultivada em milhões de hectares, especialmente nas regiões oeste e norte do estado, beneficiando-se de solos férteis e clima favorável. O estado já atingiu recordes históricos de produção (como na safra 2019/20, com cerca de 20,8 milhões de toneladas, segundo dados estaduais) e normalmente disputa com o Rio Grande do Sul a posição de segundo maior produtor nacional [SEAB/PR - Secretaria da Agricultura e do Abastecimento do Paraná 2023]. Nacionalmente, o Brasil ultrapassou os Estados Unidos e assumiu a liderança global na produção de soja, com safras superiores a 120–150 milhões de toneladas nos anos recentes [CONAB - Companhia Nacional de Abastecimento 2025]. Essa predominância deve-se, entre outros fatores, ao desenvolvimento de cultivares adaptadas, políticas de crédito e expansão agrícola em novas fronteiras. No contexto paranaense, a soja é vital para a economia regional, movimentando a cadeia de insumos, gerando empregos e abastecendo indústrias de processamento e exportação [Amorim 2025].

Estatística e Séries Temporais na Previsão Agrícola

A estatística aplicada à agricultura fornece as bases para análise de safras e elaboração de previsões. Séries agrícolas costumam apresentar tendência de longo prazo (decorrente de avanços tecnológicos) e variações sazonais (relacionadas ao ciclo de plantio e colheita). Modelos de séries temporais permitem capturar esses padrões e projetar valores futuros, partindo do pressuposto de que o passado contém informações úteis sobre o futuro [Box et al. 2016]. Para que isso seja possível, frequentemente é necessário tornar a série estacionária, eliminando tendências ou sazonalidades que prejudiquem o ajuste do modelo [Ehlers 2009].

Órgãos oficiais como o IBGE e a Conab aplicam modelos estatísticos para gerar prognósticos de safra. O LSPA/IBGE é um levantamento mensal que fornece estimativas atualizadas de área, rendimento e produção agrícola em cada estado, consolidando-se como fonte confiável e padronizada de dados. Já a Conab publica relatórios mensais de acompanhamento da safra de grãos, que incluem soja, detalhando área plantada, produtividade e condições climáticas.

Algoritmos de Previsão Selecionados

1. **ARIMA (Autoregressive Integrated Moving Average):** É um dos modelos estatísticos mais utilizados em previsão de séries temporais. Combina autorregressão, diferenciação (integração) e médias móveis para capturar padrões lineares nos dados. A metodologia de Box-Jenkins orienta a identificação dos parâmetros (p, d, q) que melhor se ajustam à série. O ARIMA costuma apresentar bom desempenho em séries com tendência definida e poucos choques abruptos.

2. **Holt-Winters (Suavização Exponencial):** Baseia-se no alisamento de observações passadas, atribuindo pesos decrescentes no tempo. Atualiza três componentes a cada período: nível, tendência e sazonalidade. É amplamente empregado em séries curtas com comportamento sazonal, sendo flexível na adaptação a mudanças recentes. Sua

simplicidade e boa acurácia em horizontes curtos fazem dele um modelo comparável ao ARIMA em vários contextos agrícolas.

3. **Random Forest Regressor:** Representa uma abordagem de aprendizado de máquina baseada em árvores de decisão. O modelo gera múltiplas árvores em subconjuntos dos dados e combina os resultados para reduzir variabilidade e aumentar robustez. Diferentemente de ARIMA e Holt-Winters, não exige pressupostos de linearidade ou estacionariedade, podendo capturar padrões complexos e não lineares. Contudo, quando aplicado a séries anuais curtas, pode enfrentar limitações de generalização, sendo especialmente útil quando há maior volume de dados ou relações não lineares subjacentes.

Em síntese, cada algoritmo apresenta características distintas: ARIMA foca em relações lineares com pressupostos estatísticos claros; Holt-Winters enfatiza tendência e sazonalidade de forma simples e adaptativa; e Random Forest, como técnica de aprendizado de máquina, busca captar interações complexas sem assumir estrutura prévia dos dados. A comparação desses três métodos neste projeto permitirá avaliar qual deles gera previsões mais acuradas para a produção de soja no Paraná no período de 2020 a 2024.

3. Revisão da Literatura

A previsão de safras agrícolas é amplamente estudada devido à sua importância econômica e aos desafios impostos pela variabilidade climática e de mercado. Modelos estatísticos clássicos de séries temporais, como ARIMA e Holt-Winters, têm sido tradicionalmente empregados em projeções agrícolas por sua capacidade de capturar tendências e sazonalidades a partir de dados históricos [Box et al. 2016].

Em um estudo aplicado no Peru, Sernaque Herrera et al. (2022) compararam ARIMA e Holt-Winters para prever a produção de cereais, constatando que o Holt-Winters aditivo apresentou ajuste superior ao ARIMA em alguns casos, enquanto o multiplicativo mostrou-se mais adequado em séries com forte sazonalidade. Esse trabalho ilustra a relevância de comparar diferentes modelos para identificar aquele que melhor se adapta ao comportamento da série [Herrera et al. 2022].

No contexto brasileiro, Anjos et al. (2023) avaliaram ARIMA e Holt-Winters para previsão das exportações de soja pelos portos do Paraná. Os autores concluíram que ambos os modelos foram capazes de captar tendências gerais, mas ressaltaram que a acurácia das previsões foi limitada por fatores externos, como variações de mercado e logística, sugerindo a necessidade de metodologias complementares [Anjos et al. 2023].

Mais recentemente, técnicas de aprendizado de máquina têm ganhado destaque na previsão agrícola por sua capacidade de lidar com relações não lineares e múltiplas variáveis. Pereira, Beker e Schmoeller (2024) aplicaram modelos de machine learning à previsão da produtividade da soja no Paraná, comparando Regressão Linear, Random Forest e XGBoost. Os resultados apontaram o Random Forest como o modelo mais preciso, alcançando coeficiente de determinação $R^2 \approx 0,81$, enquanto a regressão linear teve desempenho muito inferior [Pereira, Beker e Schmoeller 2024].

De forma semelhante, Pessina (2024) utilizou Random Forest com dados meteorológicos para prever a produtividade mesorregional de soja no Brasil, obtendo erro percentual médio absoluto (MAPE) de aproximadamente 8%. Esse resultado reforça a vantagem de modelos baseados em aprendizado de máquina em contextos agrícolas, sobretudo quando há disponibilidade de dados climáticos e de manejo [Pessina 2024].

Órgãos oficiais, como o IBGE e a Conab, também produzem estimativas de safra. O IBGE, por meio do LSPA, divulga mensalmente previsões de área, rendimento e produção agrícola, enquanto a Conab apresenta relatórios de acompanhamento da safra. Por exemplo, a Conab estimou a safra de grãos 2024/25 em 345,2 milhões de toneladas, incluindo 169,7 milhões de toneladas de soja, o que representa crescimento de 14,8% em relação à safra anterior [CONAB - Companhia Nacional de Abastecimento 2025]. Esses prognósticos oficiais são fundamentais para orientar políticas e o mercado, mas a literatura acadêmica demonstra que modelos estatísticos e de aprendizado de máquina podem oferecer previsões mais refinadas ou complementares.

Em síntese, os estudos revisados convergem na relevância de investigar diferentes abordagens de previsão agrícola. Os modelos clássicos (ARIMA, Holt-Winters) oferecem boa capacidade de projeção em horizontes curtos e séries estáveis, enquanto algoritmos de aprendizado de máquina, como o Random Forest, têm se mostrado superiores em cenários mais complexos. Nesse sentido, o presente trabalho insere-se na literatura ao propor uma comparação direta entre três modelos – ARIMA, Holt-Winters e Random Forest – aplicados à produção

de soja no Paraná com dados do LSPA/IBGE (2020–2024), preenchendo uma lacuna prática e metodológica no campo da previsão agrícola.

4. Plano de Trabalho

4.1 Descrição do Problema

A soja é a principal commodity agrícola do Paraná, responsável por aproximadamente 14% da safra brasileira e consolidando o estado como o segundo maior produtor nacional. Nos últimos anos, entretanto, a produção apresentou fortes oscilações em função de fatores climáticos e de mercado. Em 2022, por exemplo, uma estiagem severa reduziu drasticamente a produção estadual, mas em 2023 e 2024 observou-se recuperação, alcançando 18,6 milhões de toneladas. Para 2025, projeções indicam nova alta, com estimativas próximas de 22,4 milhões de toneladas, cerca de 20% acima do ano anterior [SEAB/PR - Secretaria da Agricultura e do Abastecimento do Paraná 2023, CONAB - Companhia Nacional de Abastecimento 2025].

Essas variações revelam a importância de métodos de previsão acurados, capazes de antecipar resultados de safra e subsidiar o planejamento agrícola, logístico e econômico. Apesar de existirem estimativas oficiais divulgadas por órgãos como IBGE e Conab, ainda há necessidade de avaliar comparativamente diferentes metodologias preditivas aplicadas aos dados oficiais do estado.

Diante disso, este estudo busca responder à seguinte questão: **qual dos modelos preditivos – ARIMA, Holt-Winters ou Random Forest – apresenta melhor desempenho na previsão da produção de soja no Paraná utilizando os dados do LSPA/IBGE referentes ao período de 2020 a 2024?**

4.2 Hipóteses

Para responder a essa questão, formularam-se as seguintes hipóteses de pesquisa:

- **H1:** Modelos clássicos de séries temporais, como ARIMA e Holt-Winters, capturam de forma adequada tendências e sazonalidades da produção de soja no Paraná, fornecendo previsões satisfatórias em horizontes de curto prazo [Herrera et al. 2022, Anjos et al. 2023].
- **H2:** O algoritmo de aprendizado de máquina Random Forest apresenta desempenho superior, dada sua capacidade de modelar relações não lineares e lidar com maior variabilidade interanual [Pereira, Beker e Schmoeller 2024, Pessina 2024].
- **H3:** A comparação entre os três métodos permitirá identificar em quais condições cada abordagem é mais eficaz, oferecendo subsídios para recomendações práticas de previsão agrícola no estado.

4.3 Objetivos

Objetivo Geral: Comparar o desempenho dos algoritmos ARIMA, Holt-Winters e Random Forest na previsão da produção de soja no Paraná entre 2020 e 2024, utilizando dados oficiais do LSPA/IBGE.

Objetivos Específicos:

1. Coletar e organizar os dados do LSPA/IBGE referentes à produção de soja no Paraná.
2. Analisar a série temporal da produção, identificando padrões de tendência e sazonalidade.
3. Implementar e calibrar os três algoritmos selecionados.
4. Avaliar e comparar os modelos utilizando métricas como MAE, RMSE e MAPE.
5. Discutir os resultados e propor recomendações sobre a aplicabilidade dos modelos para previsão agrícola.

4.4 Atividades e Cronograma

O cronograma de execução está sintetizado na Tabela 4.1, distribuído em três meses (julho a setembro). Conforme indicado:

- **Atividade 1** (revisão da literatura; coleta e organização dos dados do LSPA/IBGE) ocorre em *julho e agosto*;
- **Atividade 2** (implementação e calibração dos modelos ARIMA, Holt-Winters e Random Forest) desenvolve-se em *agosto e setembro*;
- **Atividade 3** (validação, comparação de resultados, redação do relatório e conclusão) concentra-se em *setembro*.

Essa distribuição garante fluxo lógico entre preparação dos dados, modelagem e consolidação dos resultados.

Tabela 4.1: Cronograma de Atividades para 2025.3

Atividade	**Julho**	**Agosto**	**Setembro**
1. Revisão e coleta de dados	X	X	
2. Implementação e calibração		X	X
3. Validação e relatório			X

4.5 Metodologia

A metodologia é quantitativa, experimental e comparativa. Os dados do LSPA/IBGE constituem a série temporal base (IBGE, 2025). O pré-processamento inclui análise de consistência, identificação de tendências e sazonalidade.

- O **modelo ARIMA** seguirá a metodologia Box-Jenkins [Box et al. 2016].
- O **modelo Holt-Winters** será ajustado nas formas aditiva e multiplicativa [Herrera et al. 2022].
- O **Random Forest** será implementado com defasagens da série e variáveis explicativas complementares, conforme sugerido em estudos recentes de previsão agrícola [Pereira, Beker e Schmoeller 2024, Pessina 2024].

As previsões serão avaliadas com base em métricas de erro (RMSE, MAE e MAPE). Os experimentos serão conduzidos em Python (bibliotecas statsmodels e scikit-learn), garantindo reprodutibilidade.

4.6 Desenho da Solução

O processo metodológico definido para este estudo é representado no fluxograma da Figura 4.1, o qual organiza de forma sequencial as etapas necessárias para a previsão da produção de soja no Paraná, com base nos dados do LSPA/IBGE. As etapas são descritas a seguir.

1. **Coleta de Dados:** nesta etapa, são obtidos os arquivos em formato CSV disponibilizados pelo LSPA/IBGE, abrangendo os valores oficiais de produção de soja no Paraná entre 2020 e 2024. Essa base constitui a série histórica fundamental para os experimentos do estudo.

2. **Pré-processamento:** os dados coletados passam por tratamento inicial, incluindo a detecção e correção de inconsistências, análise exploratória para identificação de tendências de longo prazo e verificação de padrões de sazonalidade. Além disso, são avaliados possíveis outliers que possam distorcer a modelagem.

3. **Modelagem:** com a série temporal preparada, são implementados três algoritmos de previsão: ARIMA, Holt-Winters e Random Forest. Cada modelo é calibrado de acordo com suas premissas, permitindo comparar métodos estatísticos tradicionais com técnicas de aprendizado de máquina.

4. **Validação:** a etapa de validação é conduzida separando o período de 2020 a 2023 para treinamento e o ano de 2024 para teste. Dessa forma, é possível verificar se os modelos ajustados são capazes de reproduzir valores observados em um ano recente não utilizado no processo de treinamento.

5. **Avaliação:** após a validação, os modelos são avaliados comparativamente a partir de métricas estatísticas consagradas para previsão de séries temporais, como RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) e MAPE (Mean Absolute Percentage Error). Essa análise permite quantificar a acurácia de cada abordagem.

6. **Aplicação:** o modelo com melhor desempenho é então utilizado para gerar previsões da produção de soja para os anos de 2025 e 2026. Os resultados são discutidos em termos de plausibilidade, intervalos de confiança e utilidade prática para o setor agrícola.

7. **Tomada de Decisão:** por fim, os resultados das previsões são apresentados de forma a subsidiar o planejamento agrícola, contribuindo para decisões estratégicas em áreas como logística, armazenamento, comercialização e políticas públicas relacionadas à produção de soja.

Esse desenho metodológico assegura que o estudo avance além da análise descritiva, oferecendo uma comparação prática entre diferentes algoritmos de previsão e possibilitando a geração de cenários futuros para apoiar a tomada de decisão no agronegócio paranaense.

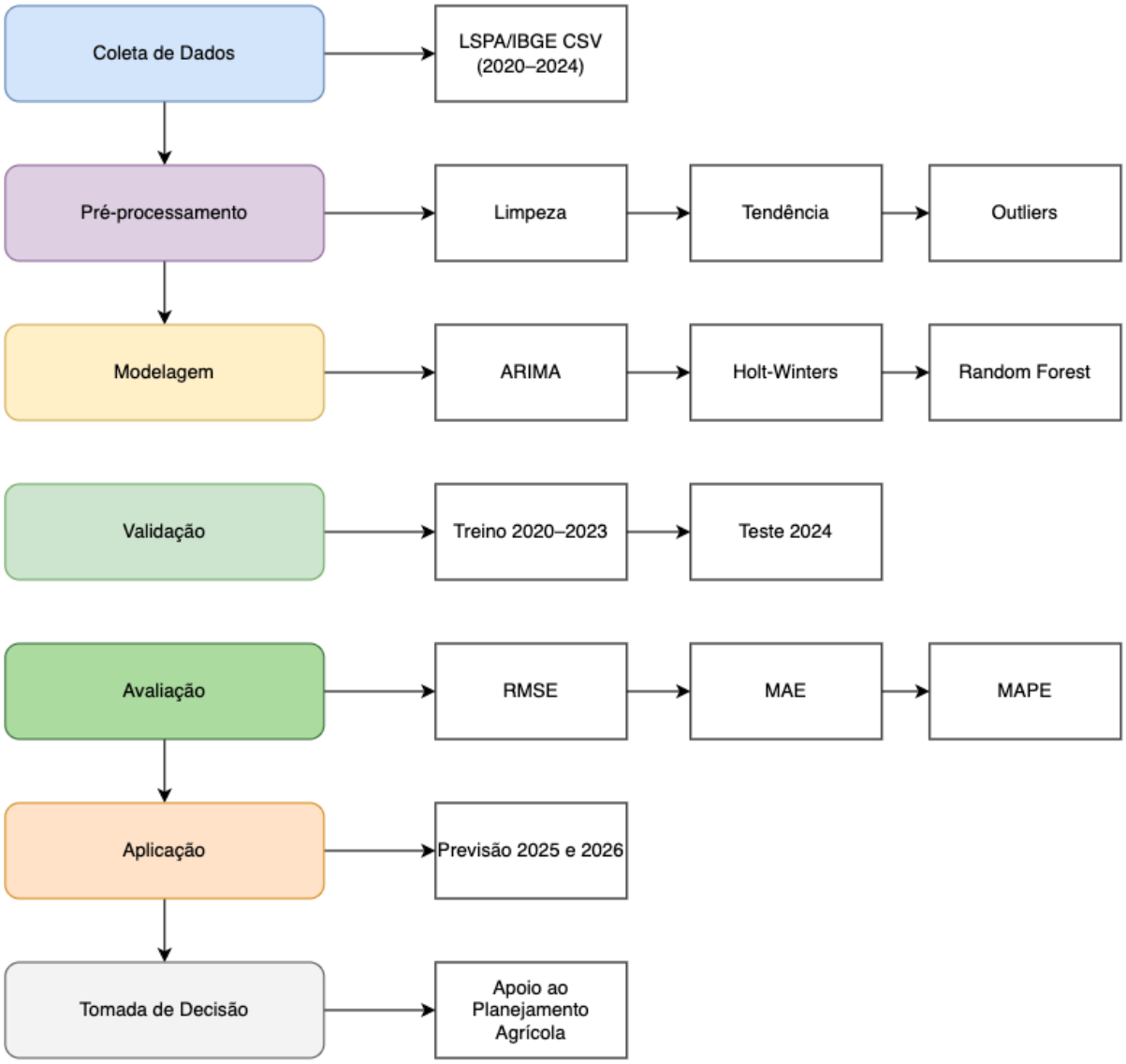

Figura 4.1: Desenho da Solução

5. Conclusão

A análise da produção de soja no Paraná entre 2020 e 2024, com base nos dados do LSPA/IBGE, evidencia tanto a relevância econômica da cultura quanto a vulnerabilidade às condições climáticas. A quebra de safra em 2022, causada por estiagem severa, ilustra os riscos inerentes à produção agrícola. Por outro lado, a recuperação observada em 2023 e 2024 confirma a resiliência do setor e a importância de métodos preditivos robustos para antecipar cenários. Este projeto propõe a comparação de três algoritmos de previsão – ARIMA, Holt-Winters e Random Forest – aplicados aos dados oficiais do Paraná. Espera-se que os modelos clássicos (ARIMA e Holt-Winters) apresentem bom desempenho em séries temporais estáveis e de curto prazo, enquanto o Random Forest tende a se destacar em contextos de maior variabilidade, capturando padrões não lineares e interações complexas.

Do ponto de vista prático, a aplicação desses modelos permitirá identificar a abordagem mais adequada para apoiar o planejamento agrícola no estado. Com previsões mais acuradas, agentes do agronegócio paranaense – produtores, cooperativas, indústria e gestores públicos – poderão planejar melhor o armazenamento, o transporte e as políticas de apoio, reduzindo riscos e aumentando a eficiência.

Em síntese, a contribuição esperada do trabalho é dupla:

1. **Acadêmica**, ao comparar metodologias distintas de previsão de safras no contexto paranaense, preenchendo lacunas metodológicas;

2. **Aplicada**, ao oferecer subsídios concretos para a tomada de decisão no agronegócio, reforçando a utilidade dos dados públicos do IBGE integrados a técnicas modernas de análise preditiva.

Apesar da relevância do estudo, algumas limitações abrem caminho para pesquisas futuras. Entre elas, destacam-se:

- Incorporação de variáveis climáticas (precipitação, temperatura, índices ENSO/El Niño e La Niña) para enriquecer os modelos de previsão.

- Análises em nível municipal ou mesorregional, permitindo captar diferenças intraestaduais que são mascaradas em dados agregados.

- Comparação com outros algoritmos de aprendizado de máquina, como XGBoost, LSTM e redes neurais recorrentes, que têm mostrado resultados promissores em séries temporais agrícolas.

- Integração de dados de sensoriamento remoto (imagens de satélite, NDVI) para aprimorar a detecção precoce de quebras de safra.

Essas possibilidades indicam que o presente trabalho pode ser expandido para além da comparação entre três algoritmos, contribuindo de forma contínua para o avanço das metodologias de previsão agrícola e para a resiliência do agronegócio frente a incertezas climáticas e de mercado.

6.Referências Bibliográficas

[Amorim 2025]AMORIM, D. E. C. *Safra de grãos 2024/25 é estimada em 345,2 milhões de toneladas com recorde na produção de milho e soja*. 2025. www.cnnbrasil.com.br/economia/mercado/safra-de-soja-deve-totalizar-recorde-de-1673-milhoes-de-t-em-2025-preve-ibge/. Acesso em: 24 ago. 2025.

[Anjos et al. 2023]ANJOS, B. C. et al. Análise de previsão da exportação de soja: Uma avaliação comparativa entre os modelos arima e holt-winters. In: *Anais do Congresso Brasileiro de Engenharia de Produção (CONBREPRO)*. [S.l.]: XIII Congresso Brasileiro de Engenharia de Produção, 2023. p. 12. Acesso em: 24 ago. 2025.

[Box et al. 2016]BOX, G. E. P. et al. *Time Series Analysis: Forecasting and Control*. 5. ed. Hoboken, New Jersey, USA: John Wiley & Sons, 2016. 720 p.

[CONAB - Companhia Nacional de Abastecimento 2025]CONAB - Companhia Nacional de Abastecimento. *Safra de grãos 2024/25 é estimada em 345,2 milhões de toneladas com recorde na produção de milho e soja*. 2025. www.gov.br/conab/pt-br/assuntos/noticias/safra-de-graos-2024-25-e-estimada-em-345-2-milhoes-de-toneladas-com-recorde-na-producao-de-milho-e-soja. Acesso em: 24 ago. 2025.

[Ehlers 2009]EHLERS, R. S. *Análise de séries temporais*. [S.l.: s.n.], 2009. 118 p.

[Herrera et al. 2022]HERRERA, H. et al. Comparison of arima and holt-winters forecasting models for time series of cereal production in peru. In: *Intelligent Human Systems Integration (IHSI 2022): Integrating People and Intelligent Systems*. [S.l.]: AHFE International, 2022.

[Pereira, Beker e Schmoeller 2024]PEREIRA, E. M.; BEKER, J. P. S.; SCHMOELLER, R. P. Aplicação de machine learning na previsão da produtividade da soja. *Revista Pleiade*, Centro Universitário Descomplica UniAmérica, v. 18, n. 45, p. 25–38, 2024. Disponível em: <https://pleiade.uniamerica.br/index.php/pleiade/article/view/1069/1271>.

[Pessina 2024]PESSINA, A. L. R. *Aprendizado de máquina para predição da produtividade mesorregional de soja utilizando dados públicos de estações meteorológicas no Brasil*. 49 p. Dissertação (Mestrado) — ICMC/USP, São Carlos/SP, Brasil, 2024. Monografia (MBA em Inteligência Artificial e Big Data).

[SEAB/PR - Secretaria da Agricultura e do Abastecimento do Paraná 2023]SEAB/PR - Secretaria da Agricultura e do Abastecimento do Paraná. *Segundo maior produtor de soja, Paraná responde por 14% da safra brasileira*. 2023. www.agricultura.pr.gov.br/Noticia/Segundo-maior-produtor-de-soja-Parana-responde-por-14-da-safra-brasileira. Acesso em: 24 ago. 2025.