Previsão da Produção Total de Etanol Derivado da Cana-de-Açúcar no Brasil

Marcelo Ladeira
Programa Pós-Graduação em Computação Aplicada (PPCA)
Universidade de Brasília (UnB)
Brasília, Brasil
mladeira@unb.br

Alexandre Quirino de Melo
Programa Pós-Graduação em Computação Aplicada (PPCA)
Universidade de Brasília (UnB)
Brasília, Brasil
alexandrequirino@hotmail.com

Awdrey Vieira Vilella
Programa Pós-Graduação em Computação Aplicada (PPCA)
Universidade de Brasília (UnB)
Brasília, Brasil
awdreyv@gmail.com

Lucas Rodrigues da Silva
Programa Pós-Graduação em Computação Aplicada (PPCA)
Universidade de Brasília (UnB)
Brasília, Brasil
lucas.mt7@hotmail.com

***Abstract*—Este artigo propõe uma abordagem inovadora para aprimorar a precisão e a agilidade das previsões da produção total de etanol derivado da cana-de-açúcar no Brasil. Através da implementação dos algoritmos de aprendizado de máquina XGBoost e LightGBM, em conjunto com os modelos estatísticos tradicionais ARIMA e SARIMA, busca-se superar os métodos manuais atualmente empregados pela Companhia Nacional de Abastecimento (Conab).**

Os dados históricos foram extraídos de boletins trimestrais da Conab, abrangendo o período de 2012 a 2024, e submetidos a um processo de ETL (Extração, Transformação e Carga) para assegurar consistência e qualidade. Foi aplicada a técnica de otimização de hiperparâmetros *Grid Search* para maximizar o desempenho dos modelos, resultando em previsões mais precisas e confiáveis. As métricas RMSE, MAPE, R^2, MAE e SMAPE foram utilizadas para avaliar a performance dos modelos e comparar seus resultados com as projeções da Conab.

Os resultados obtidos demonstram a superioridade dos modelos propostos em relação aos métodos tradicionais, proporcionando ganhos significativos em termos de acurácia e tempo de resposta. O artigo conclui que a adoção de modelos preditivos avançados pode transformar significativamente a gestão de safras no Brasil, contribuindo para a segurança alimentar e o fortalecimento do setor agrícola.

***Index Terms*—Modelos preditivos, Aprendizado de máquina, Processo ETL, Previsão de safra, ARIMA, SARIMA, XGBoost, LightGBM, Otimização de hiperparâmetros, Grid Search, Métricas de avaliação, Companhia Nacional de Abastecimento, Conab.**

I. Introdução

A Companhia Nacional de Abastecimento (Conab) é uma empresa pública brasileira vinculada ao Ministério do Desenvolvimento Agrário e Agricultura Familiar (MDA), criada em 1990 [6]. Sua principal função é gerir as políticas de abastecimento e estoques de produtos agrícolas, atuando no monitoramento e regulação do mercado de alimentos para assegurar o equilíbrio entre oferta e demanda [7].

A missão da Conab é regular e garantir o abastecimento de produtos agrícolas no Brasil, promovendo a segurança alimentar e contribuindo para o desenvolvimento sustentável do setor agrícola [8]. Suas atividades estão fundamentadas em cinco pilares principais:

- **Estoques Reguladores**: manutenção de estoques estratégicos de produtos agrícolas para assegurar a estabilidade de preços e o abastecimento em situações de escassez ou crise;
- **Acompanhamento de Safra**: coleta e análise de dados sobre as safras agrícolas, fornecendo informações cruciais para o planejamento e a tomada de decisões no setor;
- **Programas de Apoio**: iniciativas voltadas à agricultura familiar e pequenos produtores rurais, oferecendo recursos, tecnologias e assistência técnica para fomentar o desenvolvimento;
- **Comercialização e Leilões**: realização de operações de compra e venda de produtos agrícolas, visando equilibrar os preços e garantir a oferta adequada no mercado; e
- **Informações de Mercado**: produção e disseminação de análises sobre preços, oferta, demanda, estoques e tendências do mercado agrícola, promovendo maior transparência e previsibilidade.

Com essas ações, a Conab desempenha um papel estratégico na promoção da segurança alimentar e no fortalecimento da agricultura brasileira [8].

A Conab emprega uma ampla gama de métodos e fontes de informação para prever safras agrícolas, desempenhando um papel crucial no planejamento do abastecimento de alimentos no Brasil e na orientação das decisões dos agentes do setor **[7]**. Suas principais abordagens incluem:

- **Levantamentos de Campo**: realização de coletas diretas de dados em propriedades rurais, abrangendo informações sobre área plantada e desenvolvimento das culturas. Esses dados são utilizados para estimar a produtividade e a produção de forma precisa;
- **Parcerias e Colaborações**: estabelecimento de parce-

rias estratégicas com órgãos governamentais, instituições de pesquisa e entidades do setor agrícola, permitindo a obtenção de dados complementares e a troca de informações para enriquecer as análises; e

- **Modelos Estatísticos e Preditivos**: aplicação de técnicas avançadas de estatística e aprendizado de máquina, que utilizam dados históricos para prever resultados futuros. Esses modelos são especialmente importantes para a análise de séries temporais e para antecipar tendências no setor agrícola.

Por meio dessas abordagens, a Conab assegura informações confiáveis e atualizadas, contribuindo para um planejamento mais eficiente e decisões baseadas em evidências.

Além das abordagens estatísticas convencionais, a inclusão de variáveis econômicas e agronômicas, como área plantada e produtividade, é fundamental para a obtenção de projeções mais precisas e robustas. A previsão de safras, no entanto, é um desafio complexo, sujeito a incertezas decorrentes de fatores como condições climáticas e práticas agrícolas[5].

Atualmente, para culturas de grãos, a Conab adota modelos como ARIMA e SARIMA, que oferecem ferramentas poderosas para lidar com séries temporais que apresentam tendências e sazonalidades. O modelo ARIMA (*Auto-Regressive Integrated Moving Average*) combina componentes auto-regressivos (AR), integração (I) e média móvel (MA), sendo indicado para séries com padrões regulares [9]. Já o SARIMA (*Seasonal ARIMA*) expande as capacidades do ARIMA ao incorporar sazonalidade, permitindo análises mais detalhadas em séries com flutuações periódicas [1].

Para culturas como a cana-de-açúcar, no entanto, a Conab ainda não utiliza um modelo específico para projeções de acordo com os especialistas da área. Este artigo propõe avaliar a eficiência de modelos preditivos avançados, como ARIMA, SARIMA, XGBoost (*eXtreme Gradient Boosting*) e LightGBM (*Light Gradient-Boosting Machine*), na previsão da produção total de etanol, derivado da cana-de-açúcar. O XGBoost e o LightGBM são algoritmos de aprendizado de máquina baseados na técnica de *boosting*, que utiliza uma abordagem iterativa para combinar várias árvores de decisão simples (chamadas de "aprendizes fracos") em um modelo robusto e poderoso [10].

Esses algoritmos se destacam por sua eficiência em identificar padrões complexos em grandes volumes de dados, oferecendo alto desempenho mesmo em problemas desafiadores, como aqueles com dados desbalanceados ou alta dimensionalidade. Além disso, ambos integram otimizações específicas para lidar com grandes conjuntos de dados, tornando-os amplamente utilizados em competições e aplicações práticas de ciência de dados [10]. Enquanto o XGBoost se concentra em desempenho e eficácia ao construir árvores de decisão de forma sequencial, o LightGBM se destaca por sua otimização de memória e velocidade graças a uma abordagem *leaf-wise* (uma abordagem usada para crescer árvores de decisão em algoritmos) [11].

Nesse método, ao invés de dividir todos os nós no mesmo nível, como na abordagem *level-wise* (abordagem tradicional usada para crescer árvores de decisão) utilizada pelo XGBoost, o LightGBM escolhe expandir o nó folha com maior redução de erro. Isso permite criar árvores assimétricas, que podem capturar padrões mais complexos e gerar modelos mais precisos, especialmente em grandes conjuntos de dados [11].

O estudo deste artigo realizará uma análise comparativa entre os métodos automatizados propostos e o modelo manual atualmente empregado pela Conab. Serão utilizadas métricas como **RMSE** (Erro Quadrático Médio), **MAPE** (Erro Médio Absoluto Percentual), $\mathbf{R^2}$ (Coeficiente de Determinação), **MAE** (erro absoluto médio) e **SMAPE** (erro percentual médio simétrico) para avaliar a precisão e eficácia dos modelos. As previsões serão validadas contra dados reais do censo agrícola, garantindo maior confiabilidade nas estimativas.

Os dados de safra de cana-de-açúcar, com foco na produção de etanol, serão extraídos do portal da Conab [12]. A implementação desses novos modelos preditivos será realizada em *Python*, utilizando a plataforma *Google Colab*, que oferece um ambiente interativo e acessível para análise de dados e desenvolvimento de soluções. O objetivo é otimizar o processo de previsão e fortalecer a capacidade analítica da Conab, proporcionando suporte estratégico para decisões fundamentadas no setor agrícola.

Espera-se que esta iniciativa contribua significativamente para a melhoria da gestão de safras no Brasil, promovendo uma agricultura mais eficiente, sustentável e resiliente diante das incertezas do mercado e das condições climáticas.

II. Revisão da Literatura

O uso de algoritmos para previsões tem se expandido com o aumento da disponibilidade de dados e da capacidade computacional, sendo amplamente aplicados em diversos campos, incluindo meteorologia, finanças, saúde e marketing. Eles são capazes de modelar relações complexas entre variáveis e realizar previsões precisas, ajustando-se aos dados ao longo do tempo. A otimização de hiperparâmetros é uma etapa crucial para melhorar o desempenho dos modelos, aumentando sua acurácia em previsões futuras [13].

No artigo "Potencialidade da utilização de modelos de séries temporais na previsão do preço do trigo no estado do Paraná", os autores analisam a eficácia de diferentes modelos de séries temporais na previsão dos preços do trigo no Paraná, destacando os modelos ARIMA, SARIMA, ARCH, GARCH e TARCH. Os modelos ARCH, GARCH e TARCH mostraram melhor desempenho devido à capacidade de modelar a variância condicional, embora ARIMA e SARIMA também apresentassem boa performance [1]. Foi concluído que todos os modelos são úteis para auxiliar na tomada de decisões e na comercialização do trigo, diminuindo riscos e aumentando margens de retorno demonstrando proximidade entre os valores previstos e observados.

O artigo "Sistema de previsão da safra de soja para o Brasil", teve por objetivo avaliar um sistema de previsão de safra de soja para o Brasil, baseado em modelos empíricos regionalizados para estimativa da produtividade, a partir de um banco de dados de área cultivada em escala municipal, e de um

sistema de monitoramento agrometeorológico de abrangência nacional [2]. A análise estatística pelo teste t indica não haver diferença entre as estimativas e os dados oficiais.

Já o artigo *"Sugarcane Yield and Price Prediction Using Forecasting Models"*, destaca a importância da cana-de-açúcar na Índia e a aplicação da aprendizagem de máquina (ML) na agricultura, proporcionando a agricultores *insights* detalhados sobre a qualidade e produção de suas safras [3]. A utilização de algoritmos de aprendizado de máquina como: *Random Forest*, *Multi Linear Regression*, *Lasso Regression*, *Adaboost Regressor* e *Decision Tree Regressor*, visam prever o rendimento da cana-de-açúcar. Além disso, o modelo ARIMA é empregado para prever os preços da cultura.

No artigo *"Statistical Modeling for Prediction of Sugarcane Yield in India using ARIMA Model"*, os autores destacam a importância da cana-de-açúcar na Índia, onde buscou modelar e prever o rendimento da cana-de-açúcar, utilizando diferentes modelos ARIMA [4]. O resultado indicou que o modelo ARIMA foi o mais adequado para prever o rendimento da cultura, considerando métricas como: Erro Quadrático Médio, Erro Médio Absoluto, Erro Percentual Absoluto Médio e R2 Ajustado.

A dissertação de mestrado "Otimização de hiperparâmetros do XGBoost utilizando meta-aprendizagem"[10], explora a aplicação de meta-aprendizagem para otimizar hiperparâmetros do XGBoost, visando recomendar configurações eficientes para novos conjuntos de dados. A meta-aprendizagem, que aproveita metadados para acelerar processos como seleção de algoritmos, é destacada como solução promissora para reduzir custos e tempo no uso do XGBoost. Apesar de sua ampla aplicação, a literatura sobre recomendação de hiperparâmetros para esse algoritmo ainda é limitada. O estudo busca preencher essa lacuna, demonstrando o potencial da abordagem para melhorar a eficiência e os resultados em problemas reais.

O artigo "*LightGBM: A Highly Efficient Gradient Boosting Decision Tree*"[11], apresenta o LightGBM, uma implementação otimizada do *Gradient Boosting Decision Tree* (GBDT), focada em melhorar a eficiência e a escalabilidade em cenários de grandes volumes de dados e alta dimensionalidade de features. Para isso, propõe duas técnicas principais: *Gradient-based One-Side Sampling* (GOSS), que prioriza instâncias com gradientes maiores ao estimar ganhos de informação, e *Exclusive Feature Bundling* (EFB), que agrupa features mutuamente exclusivas para reduzir a dimensionalidade de maneira quase sem perdas. Os experimentos demonstraram que o LightGBM é até 20 vezes mais rápido que implementações convencionais, mantendo precisão comparável.

III. Metodologia

Para o presente artigo foi necessário uma abordagem sistemática e integrada para a previsão da produção total de etanol derivado da cana-de-açúcar, substituindo métodos manuais por processos automatizados e baseados em inteligência artificial. Os dados serão centralizados em uma plataforma única, o projeto busca aprimorar a eficiência, precisão e agilidade na geração de previsões agrícolas. Com a aplicação de modelos estatísticos e de *machine learning* (aprendizado de máquina), aliados a soluções tecnológicas avançadas, a metodologia também facilita o acesso, a análise e a visualização dos dados, fortalecendo a tomada de decisão estratégica no setor agrícola. A seguir, detalhamos as etapas do processo:

A. Processo ETL (Extração, Transformação e Carga)

A etapa inicial consistiu em um trabalho detalhado de coleta e organização de dados. Para isso, foram baixados boletins trimestrais de Safra de Cana-de-Açúcar publicados pela Conab, abrangendo o período de 2012 a 2024, totalizando dezenas de relatórios em formato PDF. Esses documentos passaram por um processo minucioso de extração e estruturação das informações relevantes, garantindo que os dados fossem organizados e consistentes para posterior análise.

O processo *ETL* foi executado com o objetivo de centralizar e preparar essas informações:

- **Extração**: Os boletins foram baixados e estruturados, extraindo dados relevantes manualmente.
- **Transformação**: Durante a transformação, os dados foram padronizados e validados para garantir consistência e qualidade, com tratamento de valores ausentes, remoção de duplicidades e normalização de formatos.
- **Carga**: Os dados coletados e transformados foram inseridos em uma planilha Excel, considerando o volume reduzido de informações. Para realizar as análises, a planilha foi importada utilizando a ferramenta *Google Colab (site)*, que permitiu a execução de *scripts* de análise e modelagem de forma eficiente e acessível, integrando os dados diretamente ao ambiente de trabalho.

B. Implementação de Modelos Preditivos

Com os dados consolidados, aplicamos diversos modelos preditivos para estimar a produção total de etanol derivado da cana-de-açúcar para uma Análise Global dos Levantamentos (2012–2024/25). Além disso, será realizada uma previsão dos levantamentos por períodos semestrais dos anos anteriores, abrangendo até o ciclo 2024-25, permitindo uma análise comparativa e uma projeção mais robusta da evolução da produção ao longo do tempo.

- **Modelos de Séries Temporais:** : ARIMA e SARIMA foram utilizados devido à capacidade de modelar padrões sazonais e tendências históricas. Essas técnicas permitiram capturar a variabilidade intrínseca do ciclo de produção agrícola.
- **Modelos de Machine Learning:** : As técnicas XGBoost e LightGBM foram empregadas para lidar com relações não lineares e identificar padrões mais complexos.
- **Análise Comparativa**: Os resultados obtidos pelos modelos foram avaliados com métricas como Erro Quadrático Médio da Raiz (RMSE), Erro Absoluto Percentual Médio (MAPE), Coeficiente de Determinação (R^2), Erro Absoluto Médio (MAE) e Erro Absoluto Percentual Médio Simétrico (SMAPE). Em seguida, foram comparados com

as estimativas atuais baseadas no censo manual realizado pela Conab, buscando identificar ganhos de eficiência e precisão.

C. Validação e Ajuste de Modelos

Para garantir a confiabilidade e robustez das previsões, foi realizada uma etapa de otimização de hiperparâmetros utilizando a técnica de *Grid Search*. Essa abordagem consiste em uma busca exaustiva por combinações de hiperparâmetros dentro de um espaço pré-definido. O *Grid Search* foi aplicado aos modelos XGBoost, LightGBM, ARIMA e SARIMA, avaliando sistematicamente todas as combinações possíveis de parâmetros. O objetivo foi identificar a configuração que maximiza o desempenho de cada modelo. Essa estratégia assegurou a escolha da melhor configuração para todos os modelos, levando em conta as características específicas dos dados e buscando sempre o ajuste ideal para as previsões.

A comparação com o método manual atualmente utilizado pela Conab permitiu uma avaliação clara dos ganhos em eficiência, precisão e escalabilidade. Essa análise contribui de forma significativa para a melhoria do processo de previsão e para a tomada de decisões estratégicas no setor agrícola brasileiro.

Essa metodologia vai além da automação e aprimoramento das previsões da produção total de etanol derivado da cana-de-açúcar, fortalecendo também a capacidade analítica da Conab ao integrar soluções avançadas de *data analytics* e *business intelligence*. *Data analytics* envolve a coleta, análise e interpretação de dados brutos para identificar padrões, gerar *insights* e tomar decisões informadas, enquanto *business intelligence* se refere ao uso de tecnologias e práticas para transformar dados em informações acionáveis, apoiando a tomada de decisões estratégicas e operacionais. O projeto agrega valor ao processo decisório no setor agrícola, especialmente nas operações relacionadas ao etanol, ao alinhar-se com as melhores práticas em gestão de dados e inteligência artificial, promovendo inovação e eficiência operacional.

D. Métricas utilizadas nos modelo

A avaliação da performance de modelos de previsão é fundamental para garantir a qualidade das previsões e a tomada de decisões informadas. A seguir, exploraremos as métricas mais utilizadas para avaliar os modelos ARIMA, SARIMA, XGBoost e LightGBM.

Métricas utilizadas:

- **Coeficiente de Determinação (R^2)**: é uma medida amplamente utilizada para avaliar a qualidade do ajuste de um modelo linear, indicando a proporção da variabilidade da variável dependente explicada pelo modelo. Embora seja útil para modelos ARIMA e SARIMA, sua interpretação pode ser mais complexa para modelos não-lineares como XGBoost e LightGBM.
- **Erro Quadrático Médio da Raiz (RMSE) / Erro Absoluto Percentual Médio (MAPE)**: são métricas mais versáteis e amplamente utilizadas para avaliar a performance de diferentes tipos de modelos, incluindo os mencionados. O RMSE mede a magnitude média do erro de previsão, enquanto o MAPE fornece uma medida do erro em termos percentuais. Ambas as métricas são particularmente úteis para comparar a performance de diferentes modelos e ajustar hiperparâmetros.
- **Erro Absoluto Médio (MAE)**: Mede a magnitude média dos erros em um conjunto de previsões, ignorando a direção do erro (positivo ou negativo).
- **Erro Absoluto Percentual Médio Simétrico (SMAPE)**: Mede o erro percentual entre os valores reais e previstos, com uma normalização que utiliza a soma do valor real e previsto no denominador, tornando-o simétrico e menos sensível a valores extremos.

Considerações adicionais:

- Ambas as métricas são particularmente úteis para comparar a performance de diferentes modelos e ajustar hiperparâmetros.
- A escolha da métrica ideal depende do contexto específico do problema e dos objetivos da análise.

Ao escolher as métricas adequadas, poderá ser avaliado de forma precisa a performance dos seus modelos de previsão e tomar decisões mais informadas sobre qual modelo utilizar em cada situação.

E. Descrição dos modelos

A) ARIMA (*AutoRegressive Integrated Moving Average*): desenvolvido por *George E. P. Box e Gwilym M. Jenkins* em 1970, é uma ferramenta estatística poderosa utilizada para analisar e prever séries temporais. Foi introduzido no contexto do trabalho de Box e Jenkins sobre análise de séries temporais e modelagem estatística. Eles popularizaram o uso de ARIMA em seu livro clássico *Time Series Analysis: Forecasting and Control* (1970).

O ARIMA combina três componentes principais:

- **AutoRegressivo (AR)**: Captura a dependência linear entre o valor atual da série e seus valores passados.
- **Integração (I)**: Remove tendências ou sazonalidades presentes na série, tornando-a estacionária. Isso é importante porque o ARIMA assume que a série é estacionária.
- **Média Móvel (MA)**: Modela a relação entre o valor atual da série e os erros de previsão passados. Isso ajuda a capturar o impacto de eventos aleatórios que podem afetar a série.

Notação: é representado por ARIMA(p,d,q), onde:

- **p**: Número de termos autoregressivos.
- **d**: Número de diferenciações necessárias para tornar a série estacionária.
- **q**: Número de termos de média móvel.

B) SARIMA (*Seasonal AutoRegressive Integrated Moving Average*): é uma extensão do modelo ARIMA por *George E. P. Box* e colaboradores, especialmente projetado para lidar com séries temporais que apresentam padrões repetitivos em intervalos regulares. Ele combina os componentes do ARIMA

(autoregressivo, integração e média móvel) com termos adicionais para capturar a sazonalidade:

Componentes sazonais:

- **Sazonal AutoRegressivo (SAR)**: Captura a dependência entre o valor atual e os valores correspondentes em períodos anteriores.
- **Sazonal Integração (SI)**: Remove a tendência sazonal da série, tornando-a estacionária em relação à sazonalidade.
- **Sazonal Média Móvel (SMA)**: Modela o impacto de choques sazonais aleatórios na série.

Notação: é representado por SARIMA(p,d,q)(P,D,Q,s), onde:

- **p, d, q**: Parâmetros não sazonais (ARIMA).
- **P, D. Q**: Parâmetros sazonais (SARIMA).
- **s**: Período da sazonalidade.

C) XGBoost (*Extreme Gradient Boosting*): desenvolvido por *Tianqi Chen* em 2014, é uma técnica de aprendizado de máquina que combina múltiplas árvores de decisão para criar modelos de alta precisão. Ele se destaca por sua eficiência, flexibilidade e capacidade de lidar com grandes conjuntos de dados.

Principais Características:

- ***Boosting***: Combina fracos modelos de árvores de decisão sequencialmente, com cada nova árvore corrigindo os erros do modelo anterior.
- **Regularização**: Emprega técnicas de regularização L1 e L2 para evitar *overfitting*, que ocorre quando o modelo aprende excessivamente os detalhes e o ruído do conjunto de dados de treinamento, a ponto de perder a capacidade de generalizar para novos dados. Essas técnicas ajudam a penalizar modelos excessivamente complexos, favorecendo aqueles que apresentam um equilíbrio entre precisão nos dados de treinamento e capacidade de generalização para dados não vistos, melhorando assim o desempenho do modelo em situações reais.
- **Otimização**: Utiliza algoritmos eficientes para otimizar a construção das árvores de decisão, resultando em modelos mais precisos e rápidos.
- **Flexibilidade**: Suporta diversos tipos de problemas, incluindo classificação, regressão e *ranking*. Permite a personalização através de diversos hiperparâmetros.
- **Tratamento de dados ausentes**: Possui mecanismos internos para lidar com valores faltantes nos dados.

D) LightGBM (*Light Gradient Boosting Machine*): desenvolvido pela *Microsoft* em 2017, é uma biblioteca de aprendizado de máquina de alta performance, especializada em árvores de decisão. Ele se destaca por sua velocidade, eficiência e capacidade de lidar com grandes volumes de dados.

Principais Características:

- **Crescimento de árvores *leaf-wise***: Ao focar no crescimento das folhas com maior potencial de redução de erro, o LightGBM consegue construir modelos mais precisos com menos árvores.
- ***Histogram-based***: Utiliza histogramas para agrupar os valores das variáveis (ou atributos), reduzindo a complexidade dos cálculos e acelerando o treinamento.
- **Paralelização**: Aproveita a capacidade de processamento de múltiplos núcleos, tornando o treinamento mais rápido.
- **Amostragem**: Emprega técnicas de amostragem de dados e atributos para evitar *overfitting*, que ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, capturando ruídos e padrões irrelevantes, o que prejudica sua capacidade de generalizar para novos dados.

IV. Análise e Resultados

A previsão da produção de etanol derivado da cana-de-açúcar é uma tarefa essencial para o planejamento estratégico do setor energético e para a formulação de políticas públicas relacionadas à sustentabilidade e ao abastecimento de combustíveis renováveis. A análise preditiva, baseada em dados históricos de produção agrícola e industrial, desempenha um papel crucial para compreender padrões e tendências que influenciam a eficiência produtiva e a oferta de etanol ao longo dos anos.

Este artigo explora diferentes abordagens de modelagem preditiva, utilizando modelos estatísticos e de aprendizado de máquina, para gerar estimativas de safras futuras com base em dados históricos. Para isso, são empregados os modelos ARIMA, SARIMA, XGBoost e LightGBM, abrangendo diversas granularidades temporais e análises específicas.

Dois estudos de caso foram definidos para avaliar a eficácia dos modelos em diferentes cenários:

A) Estudo de Caso 1: Considera todos os levantamentos realizados entre 2012 e a safra 2024/25, abrangendo uma visão global e completa dos dados disponíveis.

Os resultados obtidos foram comparados com os dados disponibilizados no site da CONAB, conforme detalhado na Tabela 1 a seguir:

TABLE I
Resultados dos Modelos - Estudo de Caso 1

Modelo	4º Lev. 23-24	1º Lev. 24-25	2º Lev. 24-25	3º Lev. 24-25
Arima	27325812,79	29397901,42	27796261,89	27796261,89
Sarima	28033573,92	29984455,37	26143696,19	26143696,19
XGBoost	30341614,00	26002018,00	25804398,00	25804398,00
LightGBM	28865592,97	28786520,79	28684339,28	28684339,28
CONAB	29689543,60	27316808,70	28468932,20	28856912,10

Os valores apresentados foram derivados das métricas calculadas, conforme detalhado na Tabela 2 a seguir:

B) Estudo de Caso 2: Analisa os levantamentos agrupados por períodos semestrais entre 2012 e a safra 2024/25, focando na segmentação temporal para identificar padrões sazonais mais precisos.

Os resultados obtidos foram comparados com os dados disponibilizados no site da CONAB, conforme detalhado na Tabela 3 a seguir:

TABLE II
MÉTRICAS DE DESEMPENHO - ESTUDO DE CASO 1

Modelo	Métrica	4º Lev. 23-24	1º Lev. 24-25	2º Lev. 24-25
ARIMA	RMSE	3764925,82	3740156,52	3709965,71
	MAPE	5,92%	5,98%	5,95%
	R^2	-0,85	-0,86	-0,86
	SMAPE	4,02%	4,03%	3,99%
SARIMA	RMSE	4736501,81	4882855,19	3877239,25
	MAPE	10,40%	10,64%	6,56%
	R^2	-2,21	-2,69	-1,03
	SMAPE	8,20%	6,90%	8,72%
XGBoost	RMSE	4965946,70	2961419,97	2913851,96
	MAPE	18,17%	9,34%	9,05%
	R^2	-19,22	-3,33	-4,53
	SMAPE	16,53%	8,76%	8,87%
LightGBM	RMSE	2584524,25	2354158,94	1961807,53
	MAPE	9,00%	7,87%	6,50%
	R^2	-4,48	-1,73	-1,51
	SMAPE	8,52%	7,48%	6,24%

TABLE III
RESULTADOS DOS MODELOS - ESTUDO DE CASO 2

Modelo	4º Lev.	1º Lev.	2º Lev.	3º Lev.
Arima	27471579,26	27531796,00	27968864,44	27986440,50
Sarima	29673028,22	27071831,66	27618532,77	26516845,45
XGBoost	34913688,00	29009144,00	29202596,00	32795626,00
LightGBM	29383883,75	28282021,11	28005328,67	29230403,33
CONAB	29689543,60	27316808,70	28468932,20	28856912,10

Os valores apresentados foram derivados das métricas calculadas, conforme detalhado na Tabela 4 a seguir:

TABLE IV
MÉTRICAS DOS MODELOS - ESTUDO DE CASO 2

Modelo	Métrica	4º Lev.	1º Lev.	2º Lev.	3º Lev.
ARIMA	RMSE	7662432,48	7380446,78	7092766,11	7523057,47
	MAPE	15,94%	16,14%	13,33%	16,32%
	R^2	-4,51	-8,52	-11,02	-5,34
	SMAPE	12,47%	12,17%	10,82%	12,26%
SARIMA	RMSE	7225017,50	7013399,81	6856564,72	6898904,87
	MAPE	12,04%	11,46%	11,11%	10,84%
	R^2	-3,90	-7,60	-10,23	-4,33
	SMAPE	21,10%	19,82%	19,44%	19,18%
XGBoost	RMSE	7155768,70	4924596,61	2947362,25	6437714,81
	MAPE	23,89%	16,53%	10,74%	24,13%
	R^2	-6,05	-3,45	-10,09	-23,38
	SMAPE	20,91%	14,88%	10,14%	21,40%
LightGBM	RMSE	2725269,14	2425517,00	1772079,45	3060653,55
	MAPE	9,14%	8,02%	5,92%	10,74%
	R^2	-0,02	-0,08	-3,01	-4,51
	SMAPE	9,09%	7,80%	5,69%	10,06%

V. ANÁLISE

Os dois estudos de caso proporcionam uma visão abrangente da avaliação de modelos preditivos para a previsão de safras de etanol derivados da cana-de-açúcar, utilizando diferentes abordagens temporais e métricas de desempenho. A análise permite identificar o modelo mais adequado para diferentes cenários e finalidades.

Com base nas previsões e nas métricas calculadas para os modelos ARIMA, SARIMA, XGBoost e LightGBM, avaliados em diferentes levantamentos, é possível analisar e comparar os resultados para identificar o modelo mais adequado.

Critérios de Comparação:

*__Erro Médio Absoluto Percentual (MAPE):__ Avalia a precisão percentual, sendo melhor quando menor.
* **Erro Médio Absoluto (MAE) e Erro Quadrático Médio (RMSE):** Representam os desvios absolutos e quadráticos das previsões, melhores quando menores.
* **Coeficiente de Determinação** (R^2)**:** Mede a qualidade do ajuste. Um valor próximo de 1 é desejável.
* **SMAPE:** Variante do MAPE que considera o erro simétrico, mais robusto em certos cenários.

Resumo da Análise nos estudos de casos:

Estudo de Caso 1: Análise Global dos Levantamentos (2012–2024/25). Neste cenário, que considera todos os levantamentos, os modelos foram avaliados em uma perspectiva ampla. A análise revelou que:

- **LightGBM** apresentou o melhor desempenho geral, com os menores valores de MAPE e SMAPE, especialmente em levantamentos recentes. Este modelo mostrou-se mais confiável para capturar padrões nos dados históricos.
- **ARIMA** demonstrou desempenho razoável em métricas percentuais (MAPE e SMAPE), mas com valores negativos de R², indicando dificuldade em modelar a variabilidade dos dados.
- **SARIMA** teve desempenho inferior ao ARIMA em geral, com maior imprecisão em levantamentos iniciais, refletindo a sensibilidade a padrões sazonais.
- **XGBoost** foi o modelo com desempenho mais fraco, exibindo os maiores erros e menor explicabilidade, sendo menos adequado para este conjunto de dados.

O **LightGBM**, ao considerar todos os levantamentos, é recomendado como a solução mais robusta devido à sua superioridade nas métricas e consistência.

Estudo de Caso 2: A análise dos levantamentos por períodos semestrais destacou o comportamento sazonal e permitiu uma avaliação mais granular dos modelos em relação aos valores da CONAB.

- **LightGBM** foi o modelo mais consistente em todos os períodos analisados. Ele apresentou:
  - Os menores valores de RMSE, MAPE e SMAPE, indicando alta precisão.
  - Previsões mais próximas dos valores fornecidos pela CONAB, reforçando sua confiabilidade.
- **SARIMA** teve um desempenho razoável, especialmente em períodos intermediários, mas mostrou-se menos confiável que o LightGBM devido a erros ligeiramente maiores.
- **ARIMA** apresentou métricas aceitáveis em algumas situações, mas foi limitado pela explicabilidade negativa (R² negativo) e pela menor precisão geral.
- **XGBoost** manteve seu desempenho inferior, com previsões menos alinhadas aos valores reais e maiores erros.

O **LightGBM** destacou-se como o modelo mais eficaz em ambos os estudos de caso, atendendo aos seguintes critérios:

- **Precisão:** Menores erros absolutos e percentuais, mesmo em cenários de alta variabilidade.
- **Alinhamento com a CONAB:** Previsões mais próximas dos valores reais, facilitando a tomada de decisão.
- **Consistência:** Excelente desempenho tanto na análise global quanto na segmentação temporal.

Dado seu desempenho superior e capacidade de modelar padrões complexos nos dados, recomenda-se a utilização do **LightGBM** como a principal abordagem preditiva para as safras de etanol derivado de cana-de-açucar na CONAB.

O gráfico com a comparação entre os modelos do Estudo de Caso 1 é apresentado na figura 1.

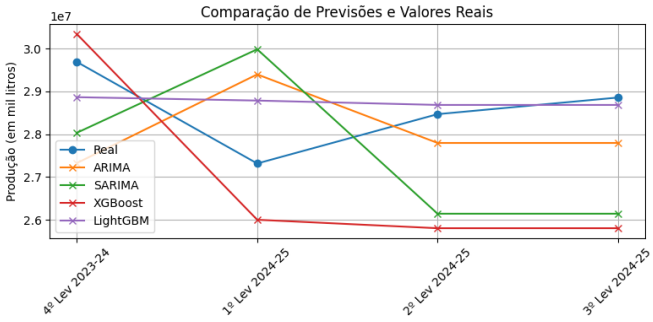

Fig. 1. Comparação de previsões dos modelos - Estudo de Caso 1.

O gráfico com a comparação entre os modelos do Estudo de Caso 2 é apresentado na figura 2.

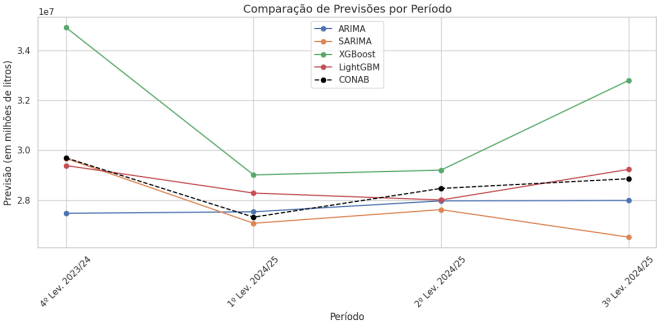

Fig. 2. Comparação de previsões dos modelos - Estudo de Caso 2.

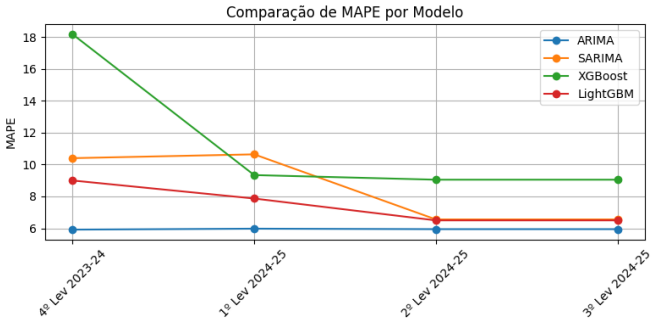

Fig. 3. Comparação de métricas - Estudo de Caso 1.

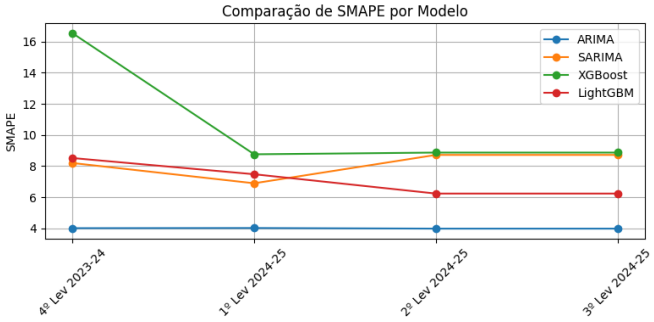

Fig. 4. Comparação de métricas - Estudo de Caso 1.

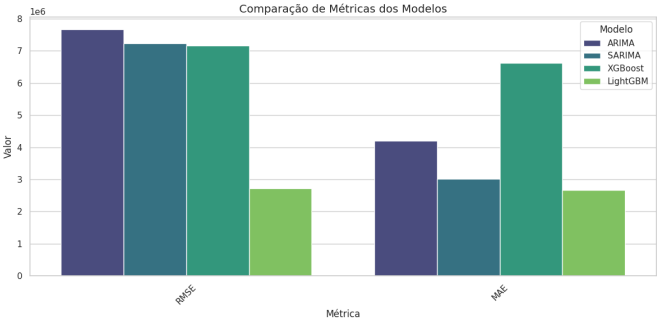

Fig. 5. Comparação de métricas - Estudo de Caso 2.

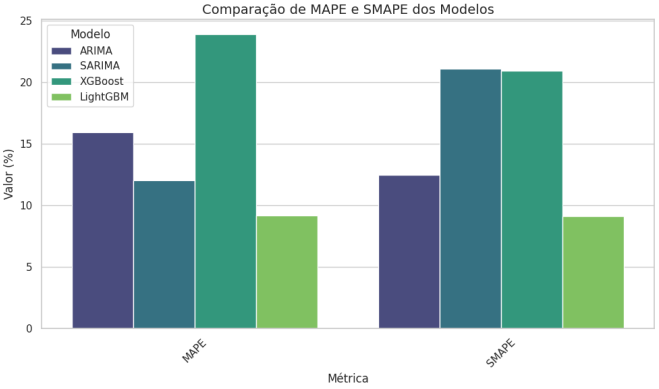

Fig. 6. Comparação de métricas - Estudo de Caso 2.

Com base nos resultados obtidos, o modelo LightGBM apresentou o melhor desempenho, embora não seja isento de limitações. Sua previsão de 28.684.339,28 mil litros para o 3º levantamento de 2024-25 está bastante próxima da estimativa manual da CONAB (28.856.912,1 mil litros). O modelo demonstrou um desempenho geral superior, especialmente nos levantamentos mais recentes, com valores de MAPE e SMAPE menores em comparação aos demais modelos analisados.

A previsão da CONAB foi utilizada como benchmark, fornecendo um valor de referência para avaliar as estimativas geradas pelos modelos. Nesse cenário, o LightGBM emerge como a melhor alternativa entre os modelos analisados, entregando previsões que, embora não perfeitas, demonstram maior alinhamento com as estimativas da CONAB.

VI. Conclusão

O estudo realizado demonstra a viabilidade e os benefícios da aplicação de modelos avançados de previsão para a estimativa da produção total de etanol derivado da cana-de-açúcar no Brasil. Ao substituir métodos manuais por abordagens automatizadas baseadas em inteligência artificial e aprendizado de máquina, como ARIMA, SARIMA, XGBoost e LightGBM, foi possível alcançar maior precisão, eficiência e agilidade no processo de previsão.

Os resultados obtidos evidenciam que os modelos de séries temporais são eficazes na captura de padrões sazonais e tendências históricas, enquanto os algoritmos de *machine learning* oferecem vantagens na identificação de relações complexas e não-lineares, ampliando a capacidade analítica da Conab. As métricas de avaliação utilizadas (RMSE, MAPE, R^2, MAE e SMAPE) confirmaram a superioridade dos modelos propostos em comparação com o método manual atualmente empregado.

Além de otimizar o processo de previsão, a integração de ferramentas avançadas de *data analytics* e *business intelligence* fortalece a tomada de decisões estratégicas no setor agrícola. A centralização dos dados e o uso de plataformas acessíveis, como o *Google Colab*, facilitam a análise e visualização, promovendo maior transparência e acessibilidade às informações.

Por fim, este trabalho contribui não apenas para a melhoria da gestão de safras, mas também para o avanço da inovação tecnológica na Conab, alinhando-se às melhores práticas de inteligência artificial e gestão de dados. Espera-se que os *insights* e resultados deste estudo sirvam de base para futuras aplicações em outras culturas agrícolas e que a metodologia proposta seja adotada como referência para o desenvolvimento de soluções sustentáveis e resilientes no setor agrícola brasileiro.

REFERÊNCIAS

[1] ARÊDES, A. F.; PEREIRA, M. W. G. Potencialidade da Utilização de Modelos de Séries Temporais na Previsão do Preço do Trigo no Estado do Paraná. Revista de Economia Agrícola, São Paulo, v. 55, n. 1, p. 63-76, jan./jun. 2008.

[2] ASSAD, E. D., MARIN, F. R., EVANGELISTA, S. R., PILAU, F. G., FARIAS, J. R. B., PINTO, H. S., / ZULLO JÚNIOR, J. Sistema de previsão da safra de soja para o Brasil. Pesquisa Agropecuária Brasileira, Brasília, v. 42, n. 5, p. 615-625, maio 2007.

[3] SNEHA, V.; BHAVANA, V. Sugarcane Yield and Price Prediction Using Forecasting Models. In: 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF), Chennai, Índia, 2023.

[4] KUMAR, Ajay; SAIN, Veer; JASLAM, P.K. Muhammed; DEEPANKAR; BHARDWAJ, Nitin; KUMAR, Vinay. Statistical Modeling for Prediction of Sugarcane Yield in India using ARIMA Model. Research Journal of Chemical and Environmental Sciences, Vol 9 [1], February 2021, p. 08-14. Academy for Environment and Life Sciences, INDIA.

[5] COMPANHIA NACIONAL DE ABASTECIMENTO. Boletim da Safra de Cana-de-Açúcar. Disponível em: https://www.conab.gov.br/info-agro/safras/cana/boletim-da-safra-de-cana-de-acucar. Acesso em: 2 dez. 2024.

[6] COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). Institucional. Disponível em: https://www.conab.gov.br/institucional. Acesso em: 1 dez. 2024.

[7] COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). Perguntas frequentes. Disponível em: https://www.conab.gov.br/perguntas-frequentes. Acesso em: 1 dez. 2024. Aqui está a referência para o documento solicitado:

[8] COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). Gestão Estratégica. Disponível em: https://www.conab.gov.br/institucional/gestao-estrategica. Acesso em: 1 dez. 2024.

[9] MELLO, Bruna Marques de; LUZ, Vivian Freire da; ROCHA, Ingrid Reis da; et al. Um estudo sobre o uso de ferramentas digitais na educação infantil: um olhar a partir da literatura. In: Anais do Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE). [S.l.]: SBC, 2023. Disponível em: https://sol.sbc.org.br/index.php/bresci/article/download/30589/30393. Acesso em: 1 dez. 2024.

[10] MARINHO, Tiago Lima. Otimização de hiperparâmetros do XGBoost utilizando meta-aprendizagem. 2021. 71 f. Dissertação (Mestrado em Informática) – Universidade Federal de Alagoas, Instituto de Computação, Maceió, 2021. Disponível em: ¡https://www.repositorio.ufal.br/bitstream/123456789/9851/1/Otimiza%C3%A7%C3%A3o%20de%20hiperpar%C3%A2metros%20do%20XGBoost%20utilizando%20meta-aprendizagem.pdf¿. Acesso em: 1 dez. 2024.

[11] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: a highly efficient gradient boosting decision tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

[12] BRASIL. Companhia Nacional de Abastecimento (CONAB). Boletim da Safra de Cana-de-Açúcar: 3º levantamento – safra 2024/2025. Brasília, 28 nov. 2024. Disponível em: https://www.conab.gov.br/info-agro/safras/cana/boletim-da-safra-de-cana-de-acucar. Acesso em: 1 dez. 2024.

[13] S. Gupta, B. Kishan and P. Gulia, "Comparative Analysis of Predictive Algorithms for Performance Measurement," in IEEE Access, vol. 12, pp. 33949-33958, 2024, doi: 10.1109/ACCESS.2024.3372082.