Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Previsão de Produção de Açúcar no Estado de São Paulo Utilizando Aprendizado de Máquina: Análise de Interpretabilidade via *Shapley Additive Explanations*

Lucas A. B. Matos[1]
[1]Instituto de Química/UNESP, Araraquara, SP
Marilaine Colnago[2], Wallace Casaca[3]
[2]IBILCE/UNESP, São José do Rio Preto, SP

O Brasil é o maior produtor mundial de cana-de-açúcar, e o estado de São Paulo se destaca nesse cenário, concentrando 50% da área cultivada e sendo responsável por 60% da produção nacional e 30% da produção global [1]. Dada a relevância desse setor para a economia e o mercado global, a utilização de modelos matemáticos e métodos computacionais torna-se essencial para a análise de dados, otimização de processos e suporte a decisões estratégicas, contribuindo para a eficiência e sustentabilidade da produção.

Sendo assim, o presente trabalho tem como foco a análise de explicabilidade da previsão da produção de açúcar utilizando o algoritmo *Random Forest* (RF) e o método SHAP (*SHapley Additive exPlanations*) para interpretar o impacto de cada variável no modelo. Para a previsão, além das variáveis originais (produção de açúcar, cana-de-açúcar e etanol, períodos de safra, preço de revenda do açúcar e do etanol, preço do açúcar total recuperável (ATR) e valor do açúcar na bolsa de NY), aplicou-se engenharia de atributos para a criação de novas variáveis, bem como *Random Search* para otimização do modelo [3]. O conjunto de dados (mensais) foi dividido em 80% para treinamento e 20% para teste, abrangendo o período de 31/01/2012 a 31/03/2023.

A Tabela 1 apresenta as métricas MAE (Erro Médio Absoluto), MAPE (Erro Médio Percentual Absoluto) e RMSE (Raiz do Erro Quadrático Médio) [1], que indicaram um desempenho adequado do modelo, com MAPE inferior a 10%.

Tabela 1: Métricas de validação do modelo RF.

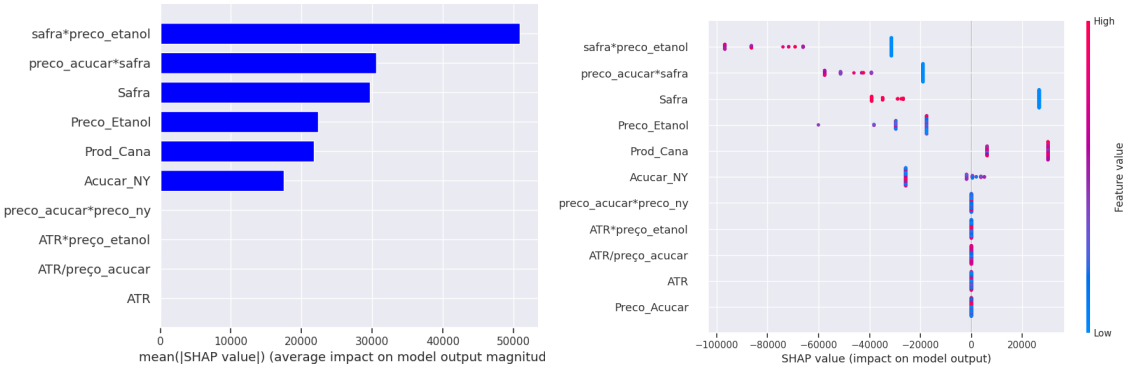
MAE (t)	MAPE (%)	RMSE (t)
123.515,40	9,60	185.188,68

A fim de tornar o modelo mais interpretável e fornecer transparência às previsões, utilizou-se o método SHAP, uma abordagem de Inteligência Artificial explicável que quantifica a contribuição de cada variável para a previsão do modelo, permitindo uma análise mais detalhada do impacto dos atributos nas estimativas [2]. Cabe mencionar que, para tal análise, as variáveis relacionadas à produção de etanol foram excluídas da análise, pois a transformação da produção de açúcar de anual para mensal dependia diretamente dessa variável, o que poderia distorcer a interpretação.

Conforme ilustrado na Figura 1 (a), as variáveis mais influentes no modelo foram as relacionadas ao preço do etanol e à safra. A Figura 1 (b) revela que o preço do etanol está associado a contribuições negativas de SHAP, indicando que o modelo tende a prever menores valores para a

[1]lucas.b.matos@unesp.br
[2]marilaine.colnago@unesp.br
[3]wallace.casaca@unesp.br

produção de açúcar. Isso sugere que, quando o preço do etanol está mais alto, uma maior parte da produção de cana é direcionada à fabricação de etanol, em detrimento da produção de açúcar. Por outro lado, a safra, uma variável binária em que 0 representa períodos de safra e 1 de entressafra, apresenta seus menores valores (0) associados a contribuições positivas de SHAP. Isso influencia o algoritmo a prever maiores valores para a produção de açúcar, o que é consistente com o fato de que, durante a safra, a produção de cana tende a ser mais alta, refletindo em maior produção de açúcar.



(a) Gráfico de barras. (b) Gráfico de resumo.

Figura 1: Método SHAP para a previsão da produção de açúcar. Fonte: dos autores.

Os resultados demonstram que a combinação entre Aprendizado de Máquina (AM) e métodos explicáveis, como o SHAP, aprimora a previsão e traz transparência ao modelo. Essa abordagem fornece subsídios valiosos para o setor sucroenergético, pois permite a análise dos fatores que impactam a produção e contribui para estratégias eficazes no planejamento agrícola e na dinâmica de mercado.

Agradecimentos

Agradecemos à FAPESP pelo fomento à pesquisa (Processos: 2023/05265-4, 2024/04718-8, 2023/14427-8).

Referências

[1] D. Chicco, M. J. Warrens e G. Jurman. “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”. Em: **Peerj computer science** 7 (2021), e623. DOI: 10.7717/peerj-cs.623.

[2] T. Le et al. “Classification and explanation for intrusion detection system based on ensemble trees and SHAP method”. Em: **Sensors** 22.3 (2022), p. 1154. DOI: 10.3390/s22031154.

[3] M. Paula et al. “Predicting Energy Generation in Large Wind Farms: A Data-Driven Study with Open Data and Machine Learning”. Em: **Inventions** 8.5 (2023), p. 126. DOI: 10.3390/inventions8050126.