UMA ANÁLISE EXPLORATÓRIA DA PRODUÇÃO DE CAFÉ DO ESTADO DE MINAS GERAIS ATRAVÉS DE TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

MARCELO SANTOS CARIELO[1], WESLEI ALVIM DE TARSO MARINHO[2], JOSÉ AUGUSTO DE LIMA PRESTES[3]

[1]Doutor em Matemática em Aplicada pela Universidade Estadual de Campinas (Unicamp), atua na Facti - Fundação de Apoio à Capacitação em Tecnologia da Informação, Av. João Scarparo Netto, 84, Salas 20D - 22D, Unique Village Offices, Campinas-SP, 13080-655, Brasil, (19) 3211-5033, marcelo.santos@facti.com.br
[2]Mestre em Ciência da Computação pela Universidade Federal de Pernambuco (UFPE), atua na Facti - Fundação de Apoio à Capacitação em Tecnologia da Informação, weslei.marinho@facti.com.br
[3]Graduado em Ciências Jurídicas e Sociais pela Pontifícia Universidade Católica de Campinas (PUC-Campinas), Mestrando em Política Científica e Tecnológica pela Unicamp, atua na Facti - Fundação de Apoio à Capacitação em Tecnologia da Informação, jose.prestes@facti.com.br

Apresentado no
XLIX Congresso Brasileiro de Engenharia Agrícola - CONBEA 2020
23 a 25 de novembro de 2020 - Congresso On-line

RESUMO: Pesquisas recentes mostram que o Brasil continua como o maior produtor de café do mundo. Em particular, o estado de Minas Gerais (MG) merece destaque devido ao protagonismo na qualidade e quantidade de sua produção. O objetivo deste trabalho é propor e avaliar uma análise exploratória, baseada em técnicas de aprendizagem de máquina, para a produção de café do estado de MG. O problema a ser resolvido foi delineado a partir de uma base de dados pública, formada por 24629 amostras. Foram empregados cinco algoritmos de aprendizagem de máquina à tarefa proposta. A principal contribuição é a estimativa com alta performance (R^2 = 0,63 e RMSE = 0,87 com o *XGBoost*) para a quantidade produzida de café. Os resultados obtidos corroboram estudos já publicados pela Companhia Nacional de Abastecimento (CONAB) e Instituto Brasileiro de Geografia e Estatística (IBGE).

PALAVRAS-CHAVE: *coffea arabica L.,* produtividade, modelagem preditiva.

AN EXPLORATORY ANALYSIS OF COFFEE PRODUCTION IN THE STATE OF MINAS GERAIS THROUGH MACHINE LEARNING TECHNIQUES

ABSTRACT: Recent research shows that the Brazil remains the largest coffee producer in the world. In particular, the state of Minas Gerais (MG) deserves to be highlighted due to its role in the quality and quantity of its production. The objective of this work is to propose and evaluate an exploratory analysis on coffee production for the state of MG, based on machine learning techniques. The problem to be solved was delineated from a public database, composed by 24629 samples. Five machine learning algorithms were used for the proposed task. The main contribution is the estimate with high reliability (R^2 = 0.63 and RMSE = 0.87 whith *XGBoost*) for the quantity of coffee produced. The results obtained confirm studies already published by the Companhia Nacional de Abastecimento (CONAB) and the Instituto Brasileiro de Geografia e Estatística (IBGE).

KEYWORDS: *coffea arabica L.*, productivity, predictive modeling.

INTRODUÇÃO: Em relatório publicado pela Organização Internacional do Café em agosto de 2020, o Brasil aparece como sendo o maior produtor mundial de café (ICO, 2020). Mesmo

com a queda de 10,9% na produção de 2019, influenciada pela bienalidade negativa, as projeções para a safra de 2020 eram, na ocasião, promissoras. Estimativas do Instituto Brasileiro de Geografia e Estatística (IBGE) para este ano indicavam um aumento de 12,9% - ou 3,4 milhões de toneladas - em relação aos dados de 2019 (IBGE, 2020). A produção total de café no Brasil é composta pelas espécies arábica e conilon. Responsável por mais de 70% da produção total de café no país, a produção de café arábica se concentra no estado de MG, que é o maior produtor de café arábica no país há mais de 10 anos (CONAP, 2014; EMBRAPA, 2020). O presente trabalho tem como objetivo propor e avaliar uma análise exploratória para a produção de café para o estado de MG, baseada em técnicas de aprendizagem de máquina (HASTIE, T., 2009). Após realizar o pré-processamento dos dados disponíveis relacionados à produção de café (EMBRAPA, 2020), os modelos foram treinados e validados. Finalmente, mediante o emprego de análises preditivas, os resultados obtidos foram comparados com aqueles disponíveis na literatura (CONAP, 2014; CONAP, 2016; LIAKOS, K. et. al., 2018; ROELLI, Y. E. et al., 2020). Este tipo de estudo serve para suportar a modelagem de políticas públicas dos mais diversos matizes, bem como para o planejamento empresarial.

MATERIAL E MÉTODOS: A fim de analisar a produção cafeeira no estado de Minas Gerais, este trabalho se valeu de uma base de dados da produção agrícola municipal disponibilizada pela Empresa Brasileira de Pesquisa Agropecuária – Embrapa (EMBRAPA, 2020). A base de dados ora selecionada contém informações sobre a produção total de café (arábica e conilon) do Brasil, no período de 1990 a 2018, para o estado de MG (EMBRAPA, 2020). Essa base é composta por 24629 amostras e 11 variáveis preditoras, quais sejam: a) Nome da Lavoura; b) Ano; c) Nome da UF; d) Sigla da Região Geográfica; e) Sigla da UF; f) Nome da Mesorregião; g) Nome da Microrregião; h) Nome do Município; i) Área Colhida (milhares de hectares); j) Quantidade Produzida (mil toneladas); e k) Valor da Produção (milhares de R$). Em vista disso, neste trabalho é proposta uma tarefa supervisionada (HASTIE, T., 2009) para estimar valores da variável *Quantidade Produzida* para os anos de 2009 até 2016, haja vista essa quantidade estar diretamente relacionada a indicadores de produtividade. A escolha do período a ser analisado e o delineamento do problema levaram em conta tanto os requisitos necessários para aplicar-se os algoritmos desejados quanto aspectos presentes em relatórios nesta área de produção agrícola (CONAP, 2014; CONAP, 2016; IOC, 2020). A análise exploratória foi iniciada com o pré-tratamento da base de dados - etapa que incluiu a imputação de dados faltantes, normalização, transformação e seleção de variáveis (HASTIE, T., 2009). Isso forneceu uma base de dados pré-tratada, formada por um subconjunto de 12695 amostras das 24629 inicias. Após essa etapa, as 12695 amostras pré-tratadas foram divididas em duas partes: uma para treinar os modelos (75%) e outra para a validação (25%) dos modelos preditivos. Os modelos utilizados foram os seguintes: Regressão Linear, *RANSAC Regressor*, *Theil-Sen Estimator*, *Huber Regressor* e o *XGBoost* (PEDREGOSA, F., 2011). Os quatro primeiros são modelos utilizados para identificar características lineares presentes nos dados. Já o *XGBoost*, por seu turno, é um algoritmo *gradient boosting* capaz de aprender padrões não-lineares subjacentes à estrutura dos dados (HASTIE, T., 2009). A etapa de validação desses modelos mostrou que as estimativas fornecidas pelos modelos ficaram dentro de uma margem de erros adequada.

RESULTADOS E DISCUSSÃO: Quanto à qualidade da base de dados disponibilizada (EMBRAPA, 2020), as variáveis *Área Colhida*, *Quantidade Produzida* e *Valor da Produção* possuíam, respectivamente, 6480, 6480 e 6479 valores faltantes, o que representa cerca de 26,31% do total disponível. Dessa maneira, como parte da análise exploratória, esses dados foram tratados através da imputação com o algoritmo *k-nearest neighbors* (HASTIE, T.,

2009), implementados no *Scikit-learn* (PEDREGOSA, F., 2011). Nos testes sem esta etapa, o treinamento e validação dos modelos de aprendizagem de máquina foram menos eficientes. Ainda na etapa de pré-tratamento dos dados, as assimetrias presentes foram minimizadas através do uso de uma transformação logarítmica. Isso assegurou que o coeficiente de variação ficasse com valores considerados adequados para o objetivo proposto no trabalho (PEDREGOSA, F., 2011; HASTIE, T., 2009). Após a transformação dos dados, devido à presença de valores discrepantes (*outliers*), foi retido seu 95º percentil. Isso equivale a dados correspondentes ao período de 1994 a 2018, representando, outrossim, 12695 amostras de interesse. Em posse das 12695 amostras pré-tratadas, contendo as variáveis *Área Colhida*, *Quantidade Produzida* e *Valor da Produção,* formulou-se a tarefa supervisionada de estimar valores da variável *Quantidade Produzida*. Para isso foram escolhidas 75% das amostras para treino, o que corresponde ao período de 1994 a 2009. O restante, relacionado aos demais anos, foi utilizado para validação. Na TABELA 1, temos a performance dos algoritmos utilizados.

TABELA 1. Síntese da performance das análises preditivas aplicadas as 12695 amostras.

Algoritmo	**RMSE (10^3 hectares)**	**R^2**
Regressão Linear	0.22	0.72
RANSAC Regressor	0.22	0.73
Theil-Sen Estimator	0.22	0.72
Huber Regressor	0.26	0.62
XGBoost	0.15	0.86

Destaca-se o *XGBoost*, tendo obtido um coeficiente de determinação (HASTIE, T., 2009) R^2 de 0,63 e um *root mean square error* RMSE de 0,87. Na FIGURA 1 percebe-se que, de 1994 a 2009, o valor médio fornecido pelos modelos preditivos (em vermelho, verde, azul, amarelo e laranja) aproxima-se dos valores em preto, já presentes na base de dados. O mesmo ocorre na faixa de valores correspondente ao período de validação, que vai de 2010 a 2018. Além disso, percebe-se que os modelos aprenderam o caráter de bienalidade da produção de café. Em termos de valores médios, as estimativas preditas pelos modelos, indo de 2009 em diante, ficam bastante próximas do que ocorreu na realidade, conforme os dados utilizados na base de dados (CONAB, 2014; EMBRAPA, 2020). Estimar a Quantidade Produzida anualmente contribui para análises de estimativas da produtividade da produção de café (CONAB, 2014; CONAB, 2016), favorecendo as diversas etapas do planejamento em políticas públicas.

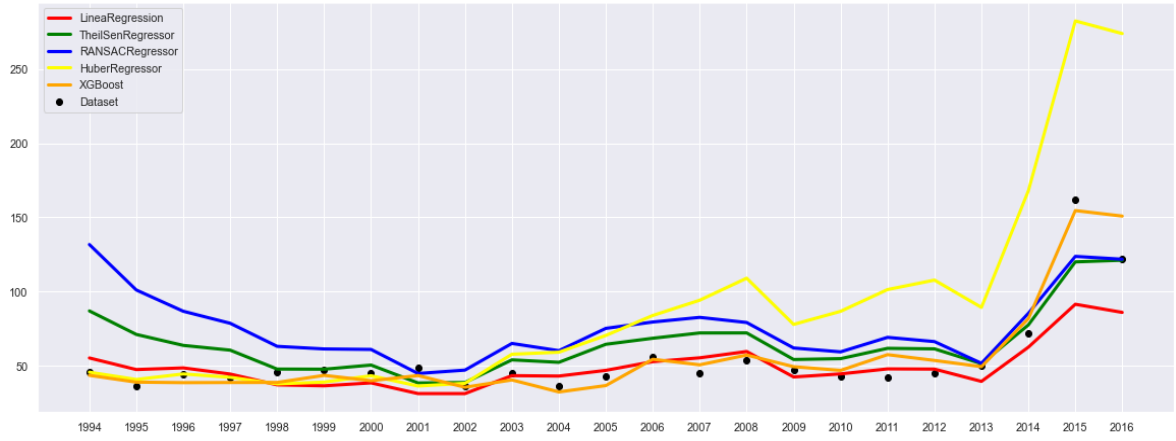

FIGURA 1. Estimativas das quantidades de café produzida por ano.

CONCLUSÕES: Os resultados obtidos neste trabalho através da modelagem preditiva sugerem um alto potencial de aplicações neste contexto, dando suporte à tomada de decisão em questões relacionadas à produção total de café para o estado de MG. A alta performance obtida na estimativa do valor médio da quantidade de produção anual, de 2010 a 2018, pode servir de base para novos estudos, incluindo períodos e variáveis adicionais.

AGRADECIMENTOS: Agradecimentos à Facti - Fundação de Apoio à Capacitação em Tecnologia da Informação, que apoiou a execução deste trabalho interno de Pesquisa, Desenvolvimento e Inovação (PD&I).

REFERÊNCIAS:

ICO. International Coffee Organization. **Relatório sobre o Mercado de Café (2019/20)**. Disponível em: < http://www.ico.org/pt/Market-Report-19-20-p.asp>. Acesso em 2020.

IBGE. Instituto Brasileiro de Geografia e Estatística. **IBGE prevê safra recorde de grãos em 2020**. 2020. Disponível em: <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-de-noticias/releases/26537-ibge-preve-safra-recorde-de-graos-em-2020>. Acesso em: 7 de setembro de 2020.

CONAB. Companhia Nacional de Abastecimento. **Acompamento da safra brasileira : café**. v. 1. 2014. Disponível em: <https://www.conab.gov.br/>. Acesso em: 7 de setembro de 2020.

CONAB. Companhia Nacional de Abastecimento. **Compêndio de Estudos.** v.1. 2016. Disponível em: <https://www.conab.gov.br/>. Acesso em: 7 de setembro de 2020.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Agropensa**. Disponível em: <https://www.embrapa.br/agropensa/bases-de-dados>. Acesso em: 7 de setembro de 2020.

LIAKOS, K. G., BUSATO, P., MOSHOU, D., PEARSON, S., BOCHTI, D. (2018). Machine learning in agriculture: A review. ***Sensors***, v. 18, n.8, pp. 2674-2703, 2018.

ROELLI, Y. E., BEUCHER, A., MOLLER, P. G., GREVE, M. B., GREVE, M. H. Comparing a random-forest-based prediction of winter wheat yield to historical yield potential. **Agronomy**, v. 10, n. 3, p. 395-412, 2020.

HASTIE, T., TISHIRANI, R., FRIEDMAN, J. **The Elements of Statistical Learning**. Springer Series in Statistics, 745 pp., 2009.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **The Journal of Machine Learning Research**, v. 12, p. 2825-2830, 2011.