PREVISÃO DA PRODUÇÃO DA CANA-DE-AÇÚCAR A PARTIR DA TÉCNICA DE RANDOM FOREST

Hugo Guiné Pinto Ferreira[1], Diogo Mazza Barbieri[2], Alan Garcia Santos[3], Alex Antonio Affonso[4], Gener Tadeu Pereira[5], Alan Rodrigo Panosso[6]

[1] Engenheiro Agrônomo, Universidade de Araraquara, (16)-99703-0482, hugo_hgpf@hotmail.com
[2] Dr. em Agronomia, UNESP/Jaboticabal, (16)-3209-1100, diogo@athenasagricola.com.br
[3] Mestre em Engenharia Mecânica, Fundação Hermínio Ometto, (16)-991926867 alan.garcia@fho.edu.br
[4] Doutor em Engenharia Elétrica, USP São Carlos, (16)-3373-9908, alex.affonso@usp.br
[5] Professor Adjunto, UNESP/Jaboticabal, (16)-3209-7215, gener.t.pereira@unesp.br
[6] Professor Assistente Doutor, UNESP/Jaboticabal, (16)-3209-7210, alan.panosso@unesp.br

Apresentado no
XLVIII Congresso Brasileiro de Engenharia Agrícola - CONBEA 2019
17 a 19 de setembro de 2019 - Campinas - SP, Brasil

RESUMO: A cana-de-açúcar é a principal fonte de energia renovável no Brasil com um futuro promissor em todo o mundo, tanto na questão econômica quanto na ambiental. Assim, buscar métodos que melhorem a capacidade de previsão do rendimento da cultura é estratégico para a sustentabilidade da produção. O objetivo do trabalho foi testar o algoritmo de aprendizado de máquina *Random Forest* (RF) para previsão da produtividade da cana-de-açúcar a partir dos atributos do solo e práticas de manejos da cultura em áreas comerciais no noroeste paulista. Foram utilizados dois cultivos comerciais (C1 e C2) na mesma região, durante os anos de 2016 a 2018, com um total de 55 mil hectares de área plantada. Os resultados indicam que a acurácia dos modelos encontrados foram entre 77 e 87%. As variáveis de manejo como número de cortes, variedade e época de colheita foram as principais condicionantes da produtividade da cultura, quando comparadas às variáveis químicas do solo. Nossos resultados indicam que o algoritmo RF apresentou acurácia e precisão satisfatórias, evidenciando sua aplicação na tomada de decisão dentro das unidades produtoras.

PALAVRAS-CHAVE: aprendizado de máquina, big data, agricultura de precisão.

FORECASTING OF SUGARCANE PRODUCTION USING RANDOM FOREST TECHNIQUE

ABSTRACT: Sugarcane is the main source of renewable energy in Brazil with a promising future throughout the world, both economically and environmentally. Thus, the forecast of crop yield is strategic for the sustainability of production. The objective of this work was to test the Random Forest (RF) machine learning algorithm to predict sugarcane productivity based on soil attributes and crop management practices in commercial areas in northwest São Paulo, Brazil. Two commercial crops were used in the same region during the years 2016 to 2018, with a total of 55 thousand hectares of planted area. The results indicate that the accuracy of the models found were between 77 and 87%. Management variables such as number of harvests, variety and harvest season were the main determinants of crop yield when compared to soil chemical variables. Our results show that the RF algorithm was accurate enough to be applied for decision making of sugar mill operation planning.

KEYWORDS: machine learning, big data, precision agriculture.

INTRODUÇÃO: De acordo com as metas estabelecidas pelo governo brasileiro na COP21, a produção de etanol em 2030 deverá ser de 54 bilhões de litros, aproximadamente o dobro dos níveis atuais de produção, já, por sua vez, a produção de açúcar aumentará de 38,7 milhões de toneladas para 46,4 milhões de toneladas (SANCHES; GRAZIANO MAGALHÃES; JUNQUEIRA FRANCO, 2019). A produção agrícola é uma complexa interação entre as variáveis ambientais, os atributos do solo e a dinâmica de nutrientes no sistema solo-planta-atmosfera, assim novos estudos sobre o potencial produtivo da cultura em solos brasileiros são necessários. Avanços na ciência de dados e *big data* podem ser capazes de atuar nesse tema (WOLFERT et al., 2017). Alguns estudos são encontrados na literatura utilizando técnicas de mineração de dados, como algoritmos de *Random Forest* (floresta aleatória - RF) para estimar a produtividade da cana-de-açúcar (EVERINGHAM; SMYTH; INMAN-BAMBER, 2009; BOCCA; RODRIGUES, 2016; OLIVEIRA; BOCCA; RODRIGUES, 2017), mostrando o potencial dessas ferramentas. Os algoritmos de RF podem manipular grandes volumes de dados, usar variáveis categóricas como preditores, medir o grau de importância das variáveis preditivas e gerar a probabilidade de classe e são robustos contra o sobre ajustes, termo usado para descrever quando um modelo mostra-se eficaz no ajuste a um conjunto de dados anteriormente observado, mas ineficaz na previsão de novos resultados, mesmo para conjuntos de dados ligeiramente desbalanceados (KHOSHGOFTAAR; GOLAWALA; HULSE, 2007; SANCHES; GRAZIANO MAGALHÃES; JUNQUEIRA FRANCO, 2019). Em trabalhos recentes, Charoen-Ung; Mittrapiyanuruk (2018) testaram os algoritmos de RF e árvore de decisão (Gradient Boosting Tree Classification) para previsão da produção de cana-de-açúcar em talhões individuais na Tailândia a partir de variáveis como tipo de solo, manejo, método e práticas de irrigação, utilização de fertilizantes e volume de chuva, encontrando acurácia de aproximadamente 77% para os dois algoritmos utilizados. Nesse contexto, o objetivo do trabalho foi investigar a relação entre atributos físicos e químicos do solo e práticas de manejo na produtividade de cana-de-açúcar, a partir da aplicação da técnica de Random Forest a fim de identificar os fatores determinantes da variabilidade espaço-temporal dos rendimentos da cultura em toneladas de colmos por hectare (TCH).

MATERIAIS E MÉTODOS: Para o estudo foram utilizados dados de duas unidades produtoras de cana-de-açúcar entre os anos de 2016 a 2018, localizadas no interior do estado de São Paulo, com um total de aproximadamente 55 mil ha, em dois cultivos comerciais, C1 e C2. Foram levadas em consideração as variáveis de manejo que são as mais interligadas com a produtividade: **nº cortes**) é o número de vezes que o canavial foi mecanicamente colhido; **idade)** tempo que a cultura ficou no campo na colheita do canavial; **variedade**) foram consideradas as 6 principais variedades plantadas nas áreas; **época de colheita)** mês do ano em que fora realizada a operação de colheita. Foram tomadas uma observação a cada 3 hectares, para a determinação da produção da cultura (TCH) e amostragem de solo nas profundidades de 0,00 a 0,25 m e de 0,25 a 0,50 m, submetidas a análises laboratoriais para a determinação dos atributos químicos do solo: matéria orgânica do solo (**MO**), **pH**, fósforo (**P**), potássio (**K**), cálcio (**Ca**), magnésio (**Mg**), hidrogênio + alumínio (**H + Al**), alumínio (**Al**), enxofre (**S**), soma de bases (**SB**), troca catiônica capacidade (**CTC**), saturação por bases (**V%**) e percentagem de saturação por alumínio (**m%**). Foi utilizado o algoritmo de *Random Forest* (RF) para identificar os atributos do solo e manejos específicos (variáveis independentes) que melhor explicaram a variabilidade da produção da cana-de-açúcar (TCH, variável dependente). O RF pertence à classe de algoritmos supervisionados em que uma variável dependente é explicada em termos de ***n*** variáveis independentes medidas em qualquer escala (SANCHES; GRAZIANO MAGALHÃES; JUNQUEIRA FRANCO, 2019). Os algoritmos de RF operam com várias árvores de decisão no momento do treinamento e

permitem a identificação e classificação dos atributos mais significativos na descrição da variável dependente (SANCHES; GRAZIANO MAGALHÃES; JUNQUEIRA FRANCO, 2019). Para a validação do modelo, dividimos os dados no treinamento e conjuntos de testes, onde ¾ de dados foi usado para treinamento e ¼ para teste, resultando em 1624 e 542 observações, respectivamente. Durante a etapa de treinamento, foi utilizada o processo de validação cruzada para o ajuste dos hiperparâmetros do modelo. Em sequência, as métricas: coeficiente de determinação (R^2) e raiz do erro médio quadrático (RMSE) foram estimadas para o conjunto de testes, consideradas medidas padrões de avaliação do algoritmo. Todas as análises foram realizadas por meio do software R (R DEVELOPMENT CORE TEAM, 2019).

RESULTADOS E DISCUSSÃO: De maneira geral, a variável **nº cortes** foi a característica mais importante no conjunto de dados para determinação da produção da cultura, tendo mais que o triplo de importância comparado aos atributos químicos do solo (FIGURA 1). Em segundo lugar, a variável **idade** foi a segunda variável que mais se destacou na modelagem computacional do TCH. Já, a **variedade** e **época de colheita** apresentaram o mesmo grau de importância, comparando com os atributos químicos do solo, destacaram-se mais que o dobro. Os melhores resultados do algoritmo foram observados quanto os atributos químicos do solo foram da segunda camada, a uma profundidade de 0,25 a 0,50 m, mostrando assim uma maior importância relacionado a produtividade. As variáveis **S**, **pH** e **MO** foram excluídas dos modelos finais. Os modelos avaliados, para os diferentes anos e para cada cultivo comercial, individual, foram satisfatórios, apresentando valores de precisão superiores a 77% e acurária inferior a 0,45 toneladas de colmos por hectares (TABELA 1).

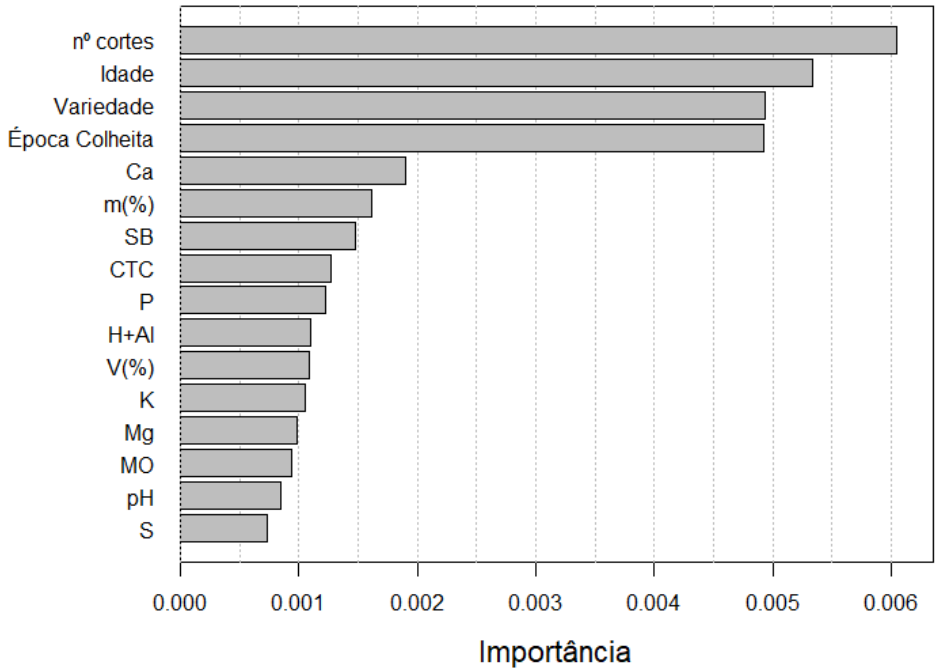

FIGURA 1. Ordem de importância dos atributos selecionados para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, CTC = capacidade de troca de cátions, H+Al = teor de hidrogênio mais alumínio, K, = teor de potássio, m(%) = saturação por alumínio, Mg = teor de magnésio, MO = teor de matéria orgânica do solo, P = teor de fósforo disponível, S = teor de enxofre, SB = soma de bases, V(%) = saturação por bases.

TABELA 1. Precisão e acurácia para avaliação do desempenho preditivo do modelo gerado a partir do algoritmo *Random Forest* para estimativa da produção da cana-de-açúcar (TCH), para dos plantios comerciais no noroeste do Estado de São Paulo.

Cultivo	2016		2017		2018	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
C1	83,21	0,3982	86,73	0,3607	81,52	0,1859
C2	85,76	0,3772	84,76	0,3869	77,11	0,4471

CONCLUSÕES: Os resultados indicam que os modelos preditivos a partir de árvores de decisão, gerados pelo algoritmo *Random Forest* apresentaram acurácia e precisão satisfatórias, evidenciando pelos valores de R^2 maiores que 80% e RMSE inferiores a 0,45. Assim, a aplicação dessa metodologia mostrou-se promissora para suporte à tomada de decisão dentro das usinas e unidade produtoras.

REFERÊNCIAS:

BOCCA, F. F. E RODRIGUES, L. H. A. (2016). The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. **Computers and Electronics in Agriculture**, doi:10.1016/j.compag.2016.08.015.

EVERINGHAM, Y. L.; SMYTH, C. W.; INMAN-BAMBER, N. G. Ensemble data mining approaches to forecast regional sugarcane crop production. **Agricultural and Forest Meteorology**, v. 149, n. 3-4, p. 689–696. doi: 10.1016/j.agrformet.2008.10.018, 2009.

KAISER, HENRY F. The varimax criterion for analytic rotation in factor analysis. **Psychometrika**, 1958, 23.3: 187-200.

KHOSHGOFTAAR, GOLAWALA AND HULSE, 2007. An empirical study of learning from imbalanced data using random forests. **In 19th IEEE international conference on tools with artificial intelligence**, 2, pp. 310-317.

OLIVEIRA, M. P. G. de; BOCCA, F. F.; RODRIGUES, L. H. A. From spreadsheets to sugar content modeling: A data mining approach. **Computers and Electronics in Agriculture**, v. 132, p. 14–20, 2017.

R DEVELOPMENT CORE TEAM. 2019: R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, https://www.rproject.org/.

SANCHES, GUILHERME M.; GRAZIANO MAGALHAES, PAULO S.; JUNQUEIRA FRANCO, HENRIQUE C. Site-specific assessment of spatial and temporal variability of sugarcane yield related to soil attributes. **Geoderma**, v. 334, p. 90-98, JAN 15 2019. (13/50942-2, 08/09265-9, 14/14965-0, 15/01587-0).

WOLFERT S., GE L., VERDOUW C. & BOGAARDT M.J. (2017) “Big data in smart farming, a review”, **Agricultural Systems**, 153, pp. 69-80.