UNIVERSIDADE ESTATUDAL PAULISTA JÚLIO DE MESQUITA FILHO

CÂMPUS DE JABOTICABAL

RANDOM FOREST NA PREVISÃO DA PRODUÇÃO DE CANA-DE-AÇÚCAR

Hugo Guiné Pinto Ferreira

Engenheiro agrônomo

2023

UNIVERSIDADE ESTATUDAL PAULISTA JÚLIO DE MESQUITA FILHO
CÂMPUS DE JABOTICABAL

RANDOM FOREST NA PREVISÃO DA PRODUÇÃO DE CANA-DE-AÇÚCAR

Discente: Hugo Guiné Pinto Ferreira
Orientador: Prof. Dr. Gener Tadeu Pereira
Coorientador: Prof. Dr. Alan Rodrigo Panosso

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Mestre em Agronomia (Ciência do Solo)

2023

F383r

Ferreira, Hugo Guiné Pinto
Random forest na previsão da produção de cana-de-açúcar / Hugo Guiné Pinto Ferreira. -- Jaboticabal, 2023

36 p. : tabs., mapas

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Gener Tadeu Pereira
Coorientador: Alan Rodrigo Panosso

1. Agricultura de precisão. 2. Árvore de decisão. 3. big data. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

unesp

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: RANDOM FOREST NA PREVISÃO DA PRODUÇÃO DE CANA-DE-AÇÚCAR

AUTOR: HUGO GUINÉ PINTO FERREIRA
ORIENTADOR: GENER TADEU PEREIRA
COORIENTADOR: ALAN RODRIGO PANOSSO

Aprovado como parte das exigências para obtenção do Título de Mestre em Agronomia (Ciência do Solo), pela Comissão Examinadora:

Prof. Dr. GENER TADEU PEREIRA (Participaçao Virtual)
Departamento de Engenharia e Ciencias Exatas DECEx / FCAV UNESP Jaboticabal

Prof. Dr. MARCÍLIO VIEIRA MARTINS FILHO (Participaçao Virtual)
Departamento de Ciências da Produção Agrícola / FCAV UNESP Jaboticabal

Dr. DIOGO MAZZA BARBIERI (Participaçao Virtual)
Athenas Consultoria Agrícola e Laboratório / Jaboticabal/SP

Jaboticabal, 23 de janeiro de 2023

DADOS CURRICULARES DO AUTOR

Hugo Guiné Pinto Ferreira, Jaboticabal, 20 de abril de 1994.

Graduado em Engenharia Agronômica pela Universidade de Araraquara no ano de 2018.

Trabalha na empresa Athenas Consultoria Agrícola e Laboratórios em Jaboticabal desde 2014.

Iniciou como analista de laboratório e atualmente trabalha na área de geoprocessamento.

AGRADECIMENTOS

Agradeço primeiramente a Deus,
Á minha família,
Á Athenas Consultoria Agrícola e Laboratórios,
Á usina COFCO International,
E á UNESP - FCAV

Sumário

1. INTRODUÇÃO 1
2. REVISÃO DE LITERATURA 2
2.1. CANA-DE-AÇÚCAR 2
2.2. AGRICULTURA 4.0 6
2.3. MINERAÇÃO DE DADOS 7
3 MATERIAL E MÉTODO 10
3.1 AMOSTRA 10
3.2 LOCAL DO ESTUDO 10
3.3 METODOLOGIA DE CAMPO 11
3.4 ANÁLISE DOS DADOS 13
4. RESULTADO E DISCUSSÃO 14
5. CONCLUSÃO 22
6. REFERÊNCIAS BIBLIOGRÁFICAS 23

RANDOM FOREST NA PREVISÃO DA PRODUÇÃO DE CANA-DE-AÇÚCAR

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 dados dos anos de 2016 a 2018, com um total de 70 mil hectares de área plantada. Os resultados indicam que a acurácia dos modelos encontrados estão entre 77 e 94%. As variáveis de manejo, tais como, número de cortes e época de colheita, foram as principais condicionantes da produtividade da cultura, em relação às variáveis químicas do solo. Os resultados indicam que o algoritmo RF apresentou precisão satisfatória, evidenciando sua aplicação na tomada de decisão dentro das unidades produtoras.

Palavras-chave: agricultura de precisão; árvore de decisão; big data.

RANDOM FOREST ALGORITHM IN THE FORECAST OF SUGARCANE PRODUCTION

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. Data from the years 2016 to 2018 were used, with a total of 70 thousand hectares of planted area. The results indicate that the accuracy of the models found were between 77 and 94%. Management variables such as number of harvests and harvest season were the main determinants of crop yield to soil chemical variables. Our results show that the RF algorithm was enough to be applied for decision making of sugar mill operation planning.

Keywords: precision agriculture; decision tree; big data.

1. 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, enquanto a produção de açúcar aumentará de 38,7 milhões de toneladas para 46,4 milhões de toneladas (SANCHES et al., 2019). Os mesmos autores afirmam que aumentar o rendimento agrícola da cultura da cana-de-açúcar deverá oferecer uma alternativa mais econômica e ambientalmente sustentável, pois, o aumento da produtividade acarretará a redução dos custos de produção e evitará a necessidade de novas expansões de cultivos. Projeções indicam que o Brasil continuará a ser o principal produtor de açúcar e etanol, produzindo cerca de 34% da cana do mundo até o ano de 2027, que será usada para gerar 20% da produção mundial de açúcar e 88% da produção mundial de etanol (FAO, 2018).

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. Tais variáveis estão intrinsecamente relacionadas, principalmente às práticas de manejo agrícola (variedade, idade, corte, maturação, ataque de pragas, estresses climáticos entre outras), deste modo 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 encontrados na literatura utilizaram técnicas de mineração de dados, como algoritmos de floresta aleatória ou “*random forest*” (RF) para estimar a produtividade da cana-de-açúcar (EVERINGHAM et al., 2009; BOCCA; RODRIGUES, 2016; EVERINGHAM et al., 2016; OLIVEIRA et al., 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, gerar a probabilidade de classe e são robustos sobre ajustes, termo usado para descrever quando um modelo se mostra eficaz no ajuste a um conjunto de dados previamente levantados, mas não na previsão de novos resultados, mesmo para

conjuntos de dados ligeiramente desbalanceados (KHOSHGOFTAAR et al., 2007; SANCHES et al., 2019).

Apesar de amplamente difundidos em vários campos da ciência, algoritmos de aprendizado de máquinas devem ser testados especificamente na área da agricultura, para ajudar o profissional na tomada de decisão quanto ao manejo específico de uma área, criando zonas específicas de manejo, visando a sustentabilidade do agrossistema. Assim, esse estudo se justifica pela necessidade de utilizar um algoritmo capaz de combinar várias árvores de decisão com as amostras relacionadas em seu banco de dados.

Deste modo, o objetivo deste trabalho é 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.

2. REVISÃO DE LITERATURA

2.1. CANA-DE-AÇÚCAR

A cana-de-açúcar possui alta relevância na economia nacional, sendo o açúcar um dos principais produtos comercializados desde o período colonial. Sua entrada no Brasil ocorreu por volta de 1532, na Capitania de São Vicente, seguida da Capitania de Pernambuco em 1535 e da Bahia em 1538; e já em 1580 levou o Brasil a se destacar no mercado mundial de açúcar (FIGUEIREDO, 2008; GAROFALO et al., 2020).

Segundo dados da Conab (2019), na safra 2019/20 estima-se que a área total de cana-de-açúcar colhida no Brasil foi de aproximadamente 8,4 milhões de ha, com uma produção estimada em aproximadamente 622 milhões de toneladas de cana-de-açúcar. A maior região produtora é a Centro-Sul que contribuiu, na safra em questão, com uma área de aproximadamente 7,5 milhões de ha e, com uma produção de aproximadamente 571 milhões de toneladas, enquanto a região Norte/Nordeste contribuiu com 879,2 mil ha de área plantada e, com uma produção de aproximadamente 50 milhões de toneladas. Nas

regiões produtoras, a maior produção foi encontrada na região Sudeste com aproximadamente 394 milhões de toneladas, seguida da região Centro-Oeste com aproximadamente 142 milhões de toneladas e, a terceira maior região produtora foi a região Nordeste com aproximadamente 47 milhões de toneladas (CONAB, 2019).

De acordo com a classificação taxonômica a cana-de-açúcar é uma gramínea e pertence à família *Poaceae* e ao gênero *Saccharum.* o provável centro de origem da cultura é a Polinésia e, suas espécies foram disseminadas por todo o Sudeste Asiático que culminou, em Papua Nova Guiné e Java (Indonésia), na criação de um moderno centro de diversidade (CHEAVEGATTI-GIANOTTO et al., 2011). A cana-de-açúcar possui hábito semi-perene e, de modo geral, o ciclo de cultivo é exercitado por aproximadamente cinco anos (LEITE et al., 2008). De acordo com Doorenbos; Kassan (1979) o clima típico da cultura é o tropical e subtropical, sendo cultivada sob condições de sequeiro em regiões úmidas, também possível o seu cultivo sob irrigação em regiões sub-úmidas com latitudes entre 35° N e 35° S. A planta da cana-de-açúcar é dividida em parte aérea (colmos, folhas e inflorescência) e radicular (raízes e rizomas) a qual possui perfilhamento desde a fase inicial do desenvolvimento (SCARPARI; BEAUCLAIR, 2008). A cana-de-açúcar possui o melhor rendimento na produção de etanol e, em relação aos combustíveis fósseis, é uma opção de combustível menos poluente, além de ser renovável (CARDOZO, 2012).

Além disso, apresenta metabolismo fotossintético C4, sendo considerada altamente eficiente na conversão de energia radiante em energia química (CASTRO, 2003). Além da alta taxa de ocorrência do processo fotossintético, apresenta eficiência na utilização e resgate de CO2 (gás carbônico) da atmosfera (SEGATO et al., 2006). Dentre as plantas de metabolismo C4, ela é uma das culturas de maior importância mundial no que tange à produção de alimentos contribuindo, referente a colheita mundial de açúcar destinado ao consumo humano, com aproximadamente 75% de participação (SOUZA et al., 2008). Em comparação com as plantas de metabolismo C3, ela apresenta taxa de crescimento e eficiência hídrica de duas a três vezes maiores (CASAGRANDE, 1996).

A cultura nas condições climáticas da região sudeste possui basicamente dois ciclos distintos, diferenciados pela época de plantio. A cana de ciclo de 12 meses, comumente chamada de cana de ano, é plantada entre setembro a outubro e atinge o maior desenvolvimento em meados de abril, tendo sua colheita iniciada a partir de setembro até outubro. Na cana de ciclo de 18 meses, também chamada de cana de ano e meio, o plantio ocorre entre os meses de janeiro até abril, sendo o crescimento intensificado durante a primavera e verão (outubro a abril), ocorrendo a colheita a partir de junho do ano seguinte ao plantio (RODRIGUES, 1995).

As fases de desenvolvimento da cana são fortemente influenciadas por vários fatores, tais como estado nutricional, tratos culturais e espécies utilizadas no plantio. (RODRIGUES; ROSS, 2020).

A etapa da colheita da cana-de-açúcar é crucial para usinas e fornecedores de cana, onde diversas medidas importantes são tomadas para o gerenciamento da área agrícola, entre elas destacam-se a produção de cana, a qualidade da cana (que é a medida do teor de sacarose da cana) e a produtividade da cana que é a produção dividida pela área. No centro oeste de São Paulo a medida de área mais comum é o hectare que equivale a 10.000 m² sendo a produtividade expressa por TCH (Tonelada de Cana por Hectare). O ATR (Açúcar Total Recuperável) é uma das principais medidas de qualidade da cana-de-açúcar e representa a quantidade de quilos de açúcar que é possível extrair em uma tonelada de cana (GARCIA; CAMOLESI Jr, 2015). As normas dessas análises são determinadas pelo CONSECANA-SP (Conselho dos Produtores de Cana-de-Açúcar, Açúcar e Álcool do Estado de São Paulo). Dessas análises são extraídas várias medidas como BRIX (porcentagem em massa de sólidos solúveis), POL (Porcentagem de sacarose), PC (Pol da cana), FIBRA, AR (Açúcar redutor), ATR (Açúcar total recuperável) (CONSECANA, 2013).

Diversos fatores podem influenciar na qualidade da cana. Alguns desses fatores estão relacionados com as características das variedades, época de colheita, idade da cana, entre outros. Além disso, cada variedade de cana tem sua curva de maturação padrão para o primeiro corte e para os demais e, de

acordo com essa curva, a variedade de cana pode ser precoce, normal ou tardia. Com base na curva de maturação e na produtividade é feito o planejamento de colheita para todos os talhões, considerando as restrições de capacidade diária de moagem da usina, capacidade das frentes de corte e transporte, área com aplicação de maturador, distância, período de colheita, e demais restrições consideradas no planejamento (GARCIA; CAMOLESI Jr, 2015).

O acompanhamento desses fatores nas áreas de cultivo e na produção, ao longo do tempo, contribui para o entendimento a respeito da dinâmica de produção da cana-de-açúcar e, o conhecimento dessa dinâmica pode auxiliar instituições governamentais, de pesquisa e privadas a se planejarem quanto às necessidades de formulações de políticas públicas, de novas tecnologias e mesmo de investimentos. Levantamentos oficiais internacionais e nacionais com séries históricas, dados estatísticos de associações de produtores e publicações técnicas são extremamente valiosos por servirem de base para o entendimento da dinâmica de produção. No caso específico da cana-de-açúcar, informações internacionais relevantes são disponibilizadas pela FAO e no Brasil consultas podem ser realizadas em fontes, como IBGE, Conab, Unica, Sindaçúcar, Sindálcool, entre outras instituições do setor (GAROFALO et al., 2020).

O surgimento do Programa Nacional do Álcool (Proálcool) teve grande contribuição na ascensão no setor de produção de etanol. Almejou-se com essa criação a diminuição da dependência das importações do petróleo (KOHLHEPP, 2010). Ele trouxe importantes inovações em áreas referentes à produção de bioetanol no país e na genética das plantas, como também, em tecnologias para o setor industrial beneficiando as usinas e destilarias e, na indústria automobilística (RODRIGUES; ORTIZ, 2006). Extensas áreas de cultivo como as regiões de Ribeirão Preto, Campinas e Bauru e fábricas de açúcar existentes influenciaram a instalação de redes de destilaria próximas a essas localidades, no Estado de São Paulo, num mesmo período (UDOP, 2016). A constante preocupação da sociedade com as emissões excessivas dos veículos automotores atreladas com a qualidade do ar dos grandes centros urbanos foram fatores importantes, os quais também contribuíram para o incentivo da utilização

e venda do etanol como um biocombustível desde a década de 1970 até atualmente (LEITE; LEAL, 2007).

2.2. **AGRICULTURA 4.0**

Com o advento da internet e da popularização do uso de computadores, a tecnologia chegou ao campo com o objetivo de otimizar a produção agrícola e consequentemente gerar produtividades cada vez maiores. O Brasil se destaca no cenário mundial como um dos maiores produtores de alimentos do mundo, e a tecnologia empregada no campo foi determinante para que a agricultura brasileira alcançasse o patamar atual. Com a evolução contínua consolidou-se uma nova era de tecnologia agrícola, onde não existe mais separação entre o mundo físico e virtual, conectados para facilitar a vida das pessoas. E por trás dessa fusão está o conceito da Agricultura 4.0 (Agro 4.0) (MASSRUHÁ; LEITE, 2017).

A Agricultura 4.0 emprega métodos computacionais de alto desempenho, rede de sensores, comunicação máquina para máquina (M2M), conectividade entre dispositivos móveis, computação em nuvem, métodos e soluções analíticas para processar grandes volumes de dados e construir sistemas de suporte a tomada de decisões de manejo (MASSRUHÁ; LEITE, 2017). Com isso, a agricultura nacional tende a ser mais tecnológica, surgindo novas oportunidades para a utilização de inovações na área de tecnologias da informação e comunicação (AgroTIC) em todas as etapas da cadeia produtiva (MASSRUHÁ; LEITE; MOURA, 2015). Com o uso destes vários recursos tecnológicos digitais da Agro 4.0 podemos elevar os índices de produtividade através da eficiência do uso de insumos, da redução de custos com mão de obra, oferecer melhor qualidade do trabalho e a segurança dos trabalhadores e diminuir os impactos ao meio ambiente (MASSRUHÁ; LEITE, 2017), englobando a agricultura de precisão, a automação e a robótica agrícola, além de técnicas de *bigdata*.

A tecnologia da informação e da comunicação (TIC) tem contribuído há vários anos para as diversas áreas de conhecimento. Ela permitiu o armazenamento e processamento de grandes volumes de dados, automatização de processos e o intercâmbio de informações e de conhecimento (MASSRUHÁ; LEITE, 2017). Aos poucos, produtores rurais vêm aderindo as TIC e sendo estas inseridas nas atividades rurais como fatores de competitividade (ASSAD; PACETTI, 2009).

2.3. **MINERAÇÃO DE DADOS**

O termo mineração de dados é utilizado para designar várias técnicas e procedimentos utilizados para examinar e processar elevadas quantidades de dados, criando um modelo para otimizar a forma de leitura dos dados existentes e permitir que sejam feitas projeções com base nos dados processados. Em resumo, a mineração de dados transforma grandes quantidades de dados em padrões, que são divididos em: direcionada e não direcionada. Na mineração de dados direcionada, busca-se prever um ponto de dados em particular, ao contrário da mineração não-direcionada, onde é necessária a criação de grupos de dados ou encontrar padrões em grupos de dados existentes. A mineração de dados moderna teve início na década de 1990, quando os computadores conseguiram chegar a um certo nível de processamento e armazenamento de dados e seus custos permitiram as empresas utilizar suas funções sem recorrer à utilização de recursos computacionais externos (ABERNETHY, 2010; RUFATTO, 2015).

A técnica de aprendizado de máquina é um subcampo da ciência da computação que trabalha com reconhecimento de padrões e teoria da aprendizagem computacional, em inteligência artificial na análise de dados. Na mineração de dados são utilizadas técnicas de aprendizado de máquina para descoberta de padrões e extração de conhecimento (HAN; KAMBER; PEI, 2012; BOLSONELLO et al, 2019). Como existem várias técnicas para a mineração de dados, um passo importante para a construção de um bom modelo, é a escolha

correta da técnica a ser utilizada. Após esse passo, o modelo construído deve ser refinado, para permitir a exploração mais eficiente das suas utilidades (RUFATTO, 2015).

O crescimento das aplicações de mineração de dados na agricultura vem se tornando mais eficientes e precisas, e com o volume de dados acumulado pelo setor sucroenergético, o potencial de aplicação dessa técnica é vasto e de grande importância para o setor. Em estudos com dados de menor escala, com intuito de estudar a variabilidade da produtividade de cana-de-açúcar, técnicas de mineração de dados têm se mostrado superiores a técnicas estatísticas convencionais (ANDERSON et al., 1999; FERRARO et al., 2009; BOCCA; 2014). Uma característica relevante da produção de cana-de-açúcar é a presença de diversos tomadores de decisão, uma vez que está inserida nas cadeias de valor do complexo da cana-de-açúcar (HIGGINS et al., 2007; BOCCA, 2014).

Duas técnicas de mineração bastante utilizadas na agricultura são as árvores de regressão (*Classification and Regression Trees*, CART) e máquinas de vetor de suporte (*Support Vector Machines*, SVM). Uma importante evolução dos modelos baseados em árvores são as *Random Forest (RF)*, sendo essa técnica bastante aplicada na agricultura em diversas culturas. Uma das principais vantagens da técnica de RF é a possibilidade de investigação não linear e relações hierárquicas entre os preditores e a resposta usando uma abordagem de aprendizagem em conjunto. O método envolve fazer várias tentativas de dados diferentes ou modelos para prever uma variável de resposta como a produtividade da cana-de-açúcar. Alguns autores reportaram aplicações das RF em modelos de análises de dados, como o caso de Everingham et al. (2016), que relataram o uso de RF para investigar quais atributos climáticos impactam na produção de cana-de-açúcar na Austrália.

A técnica do *Random Forest* pertence à classe de algoritmos supervisionados em que uma variável dependente é explicada em termos de N variáveis independentes medidas em qualquer escala (SANCHES, et al. 2019). RF é um método de aprendizado conjunto aplicado a problemas de classificação e regressão. A técnica do RF é capaz de manipular grandes volumes de dados, usar variáveis categóricas como preditores e medir o grau de importância das

variáveis preditivas no modelo final. 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 (CHAROEN-UM, et al. 2019).

Com a técnica do RF é possível classificar a importância relativa de cada variável preditora, sendo que a importância de cada variável é baseada no erro de previsão de regressão do *out-of-bag (OOB).* Aproximadamente 30% dos dados são OOB e não são usados na construção da árvore (ABDEL-RAHMAN et al., 2013; EVERINGHAM, et al., 2016). Com essa técnica a importância da variável do preditor é relatada como decréscimo da porcentagem média na taxa de classificação, para o modelo de classificação ou aumento médio no erro quadrático médio para o modelo de regressão se essa variável foi removida da análise (EVERINGHAM, et al., 2016).

Essas árvores de decisão apresentam alta variância e baixo viés, para cada nova amostra é construída uma árvore de decisão. Cada novo conjunto de seleção de treinamento utilizado é desenhado substituindo o original. Em seguida, uma nova árvore é cultivada no conjunto de treinamento, usando uma seleção aleatória de recursos. Quando utilizados esses recursos aleatórios, aparenta melhorar a precisão dessa seleção e são fornecidas estimativas do erro de generalização (PE *) deste conjunto combinado de árvores (BREIMAN, 2001).

O método de árvore de decisão é um método de simples interpretação que agrega mais vantagens do que desvantagens em sua aplicação, como por exemplo ser fácil de expor em gráficos e entender-se que suas predições se assemelham a decisões humanas podendo ser aplicados a um grande volume de dados, em contrapartida uma alteração nos dados que formarão as árvores aleatórias irá interferir significativamente na predição final dos dados.

Bagging é o nome dado ao método de melhorar as previsões para métodos de regressão em especial para árvores de decisão em que são calculados a média das previsões resultantes, as árvores não são podadas, apresentando alta variância e baixo viés. Este método apresenta melhorias na precisão pois reúne uma grande quantidade de dados em uma análise. A taxa de erro de teste é calculada com conjuntos dos dados utilizados no treinamento

do algoritmo e é utilizado um número suficientemente grande para que essa taxa se estabilize (JAMES, et al., 2017).

O algoritmo RF tem como vantagem, que como uma floresta aleatória descorrelaciona as árvores de modo que o algoritmo não tem permissão para considerar os preditores pela maioria, já que o número de preditores considerados em cada divisão é aproximadamente igual à raiz quadrada do número total de preditores. Fato que ocorre no *bagging*, pois a média de uma grande quantidade altamente correlacionada não reduz significativamente a variância de uma árvore quanto a média de grandes quantidades correlacionadas. No RF cada divisão considera apenas um subconjunto de preditores, sem influência de preditor que aparece mais vezes. Como as árvores são descorrelacionadas a média das árvores é menos variável e consequentemente mais confiável (JAMES, et al., 2017).

Para construção das árvores de decisão o algoritmo faz uma seleção aleatória de duas variáveis no banco de dados e determina o atributo que melhor separa os dados como nó principal, após essa determinação seleciona mais dois atributos aleatórios que não foram selecionados anteriormente para a formação de um novo nó e a partir dos subconjuntos das variáveis selecionadas são formadas sucessivas árvores de decisão. De maneira que podem ser formadas N árvores conforme o objetivo do treinamento (JAMES, et al., 2017).

3 MATERIAL E MÉTODO

3.1 AMOSTRA

Os dados utilizados no experimento foram coletados entre os anos de 2016 e 2018 a partir de uma parceria entre CSME, Athenas Agrícola e a usina COFCO International, com uma área total de aproximadamente 70 mil hectares. Totalizando mais de 50 mil pontos de observação (Figura 01).

3.2 LOCAL DO ESTUDO

Os dados são referentes às unidades produtoras de cana-de-açúcar do noroeste paulista da usina COFCO International nos municípios de Catanduva e Potirendaba.

3.3 METODOLOGIA DE CAMPO

Foram levadas em consideração as variáveis de manejo mais interligadas com a produtividade apresentadas na tabela a seguir:

Nº de cortes	Número de vezes que o canavial foi mecanicamente colhido;
Idade	Tempo que a cultura ficou no campo até a colheita do canavial;
Variedade	Foram consideradas 6 variedades plantadas nas áreas: RB966928, RB92579, RB867515, RB855536, CV7870, IACSP9550;
Época de colheita	Mês do ano em que fora realizada a operação de colheita: março a novembro.

Tabela 1: variáveis de manejo e produtividade da cana-de-açucar

Para análise do solo, a cada 3 hectares foi realizada uma amostragem desse material, com uma profundidade de 0,00 a 0,25 m e de 0,25 a 0,50 m, as amostras foram 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), a soma de bases (SB), capacidade de troca de cátions (CTC), saturação por bases (V%) e percentagem de saturação por alumínio (m%).

A figura 1 apresenta a região em que foram realizadas as demarcações dos pontos de amostragem encaminhados para análise e formação do banco de dados deste estudo. Os pontos escolhidos para compor o banco de dados foram os pontos coincidentes em toda a área de extensão nos três anos referentes à pesquisa (2016-2018), indicados de amarelo. Os pontos indicados de azul são os pontos demarcados não coincidentes nos três anos, portanto não utilizados na análise.

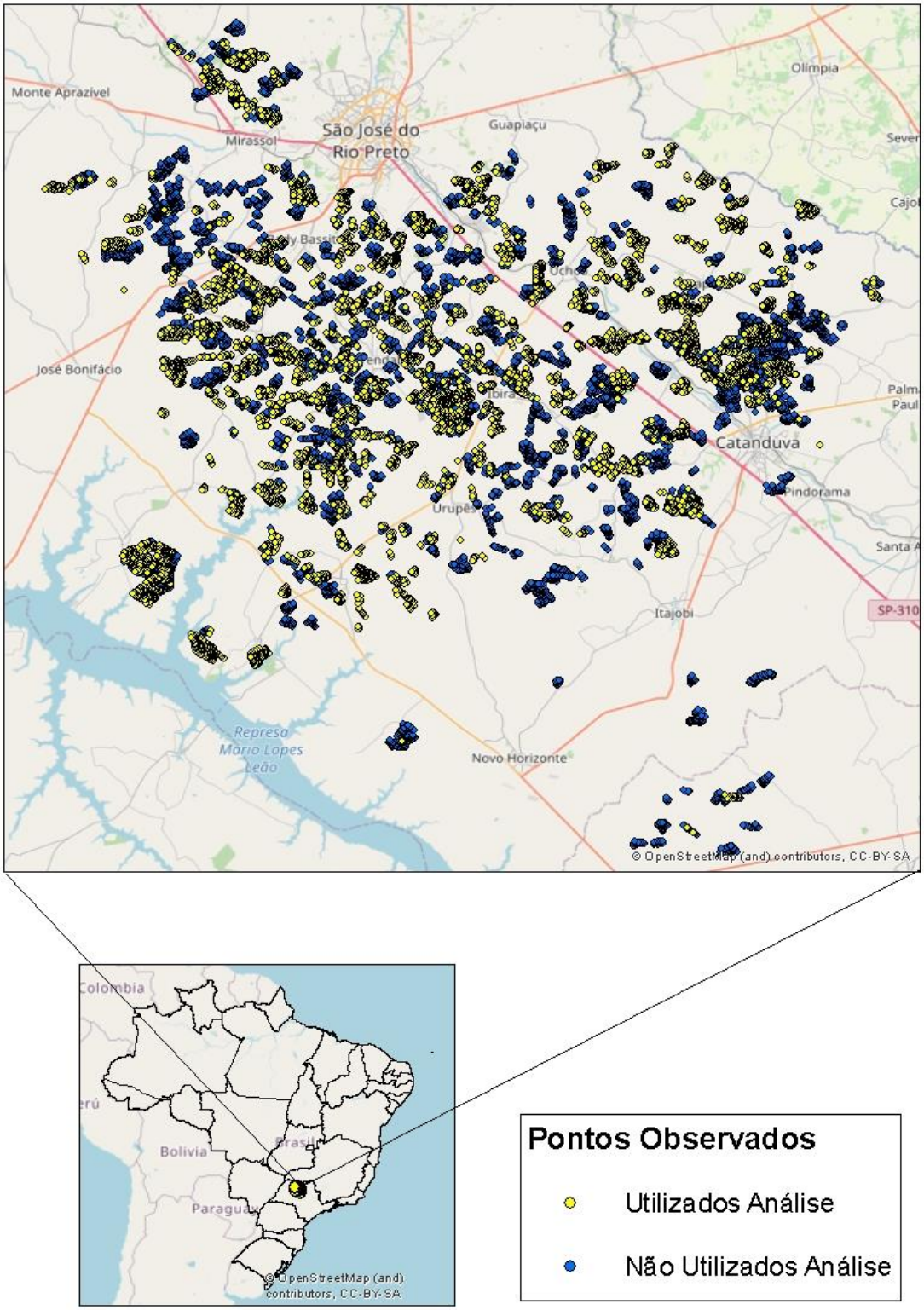


Figura 1: Pontos de amostragem demarcados no território

3.4 ANÁLISE DOS DADOS

Para análise foram utilizados os pontos de observação coincidentes nos três anos dos dados coletados, aproximadamente 7 mil amostras.

Para manipulação dos dados foi realizada a padronização dos valores das variáveis no programa R (http://www.r-project.org/) utilizando a seguinte fórmula:

$$Z = \frac{x_i - \bar{x}}{\sigma}$$

em que *Z* é o conjunto de dados com média igual a 0 e desvio padrão igual a 1 e o x_i corresponde ao valor da amostra. Dessa forma todas as variáveis seguem uma distribuição normal padrão.

Após a padronização das variáveis foi utilizado o programa R (http://www.r-project.org/) e o pacote "RandomForest" para identificar os atributos do solo e manejos específicos (variáveis independentes) que melhor explicam a variabilidade espaço temporal do

Uma parte do banco de dados é utilizada para treinamento do algoritmo e formação das primeiras árvores de decisão, após esse treino o restante dos dados é utilizado como teste percorrendo as árvores de decisão criadas. Dessa forma os dados utilizados para teste e treinamento são distintos, diminuindo os vieses envolvidos na progressão das análises e resultado da predição. A predição é apresentada a partir da média de resultados mostrados nas árvores de decisão. Neste estudo foi utilizado 75% dos dados disponíveis para o treinamento do algoritmo e 25% para teste.

O algoritmo foi escolhido para este estudo pois é bastante robusto na simulação de modelos lineares e não-lineares (WILEY, 2016), sendo o critério de parada do algoritmo de calibração o número total de ciclos igual a 1000 ou o erro médio quadrático menor que 1%. Para os resultados apresentados foram utilizados 100 ciclos no treinamento.

Para a avaliação da qualidade da previsão foi adotada a raiz do erro quadrático médio (RMSE) e o coeficiente de determinação múltiplo (R^2). Todas as análises foram realizadas por meio do software R utilizando os pacotes

"agricolae", "gclus", "cluster", "vegan", "lattice", "nortest", "rio", "ggplot2" e "randomForest".

4. RESULTADO E DISCUSSÃO

A partir da análise dos dados gerados, foi possível observar que as variáveis de manejo (nº de cortes, variedade, época de colheita e idade) são as mais correlacionadas com a produtividade da cana-de-açúcar na área total do estudo. Os atributos químicos do solo tiveram grau de importância baixo, tal que em nenhum ano ou unidade produtora este grau atingiu mais que 25% de importância, independente da profundidade analisada.

Em comparação geral entre as duas unidades produtoras, nos três anos analisados, a maior correlação com a produção da cana de açúcar por grau de importância é o número de cortes, representado nas figuras 2 e 3. Outros estudos apontam o mesmo resultado correlacionado, em que o aumento do número de cortes é diretamente proporcional ao caimento da produtividade do canavial. Hammer (2016) analisou dados de 18 usinas distribuídas em 05 cidades do estado de São Paulo entre os anos de 2012 e 2015 obteve resultado concordante de que a variável de maior impacto na produtividade de cana de açúcar é o número de cortes.

A segunda variável de manejo que mais apresentou grau de importância nos dados analisados foi a época de colheita. No ano de 2018 foi o item que liderou a importância nas duas unidades produtoras e nas duas profundidades analisadas Figuras 8,9,14 e 15. Entretanto somente na unidade de Potirendaba apresentou maior importância, enquanto na unidade de Catanduva esteve muito próxima do número de cortes. É um fator importante, entre outros associados, como apresentado em estudo que visa ordenar as variáveis que gera produtividade da cana de açúcar (HAMMER, 2016).

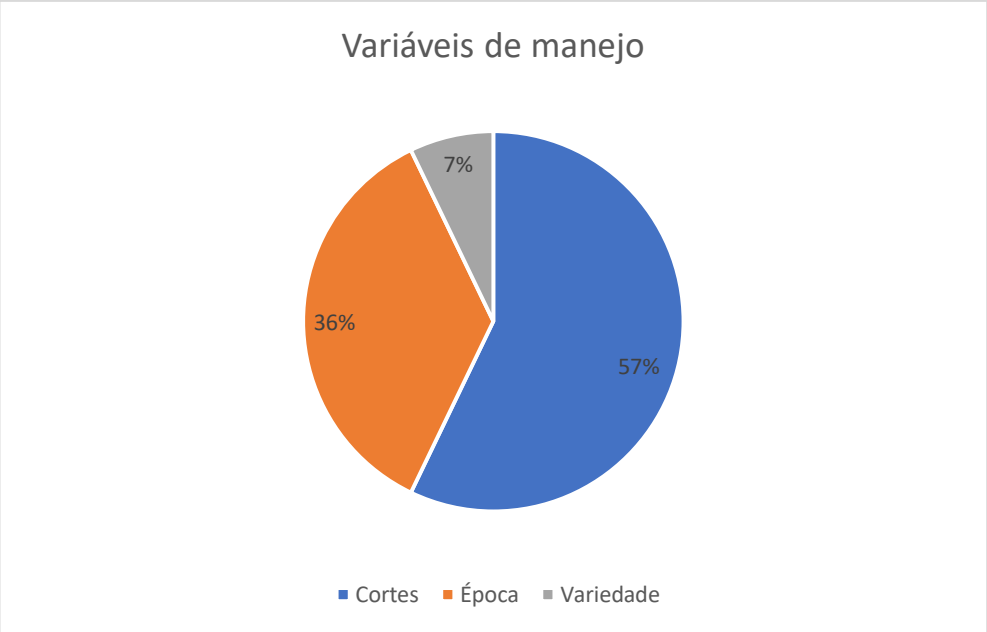


Figura 2: Grau de relação entre as variáveis de manejo e produtividade

Em relação a profundidade, de 0,00 a 0,25m (superficial) e 0,25 e 0,50m (sub superficial), não apresentaram diferença relevante para serem determinantes na produtividade do canavial. Nos três anos analisados as duas profundidades apresentaram fatores semelhantes nas unidades produtoras, conforme mostram as tabelas 2 e 3.

	Média		2016		2017		2018	
	0-25	25-50	0-25	25-50	0-25	25-50	0-25	25-50
Geral	m%	SB						
Catanduva			m%	m%	Ca	Ca	P	P
Potirendaba			P	K	pH	MO	Ca	Ca

Tabela 2: Variável de maior grau de importância entre os atributos químicos do solo

	Média		2016		2017		2018	
	0-25	25-50	0-25	25-50	0-25	25-50	0-25	25-50
Geral	Cortes	Cortes						
Catanduva			Cortes	Cortes	Época	Variedade	Época	Época
Potirendaba			Cortes	Cortes	Cortes	Cortes	Época	Época

Tabela 3: Variável de maior grau de importância entre as variáveis de manejo do solo

Nas figuras 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,13, 14, 15 e 16 são apresentadas as ordens de importância de cada item, em cada ano, por unidade produtora de cana de açúcar e o modelo gerado pelo algoritmo para previsão da produção com base nos dados obtidos neste estudo.

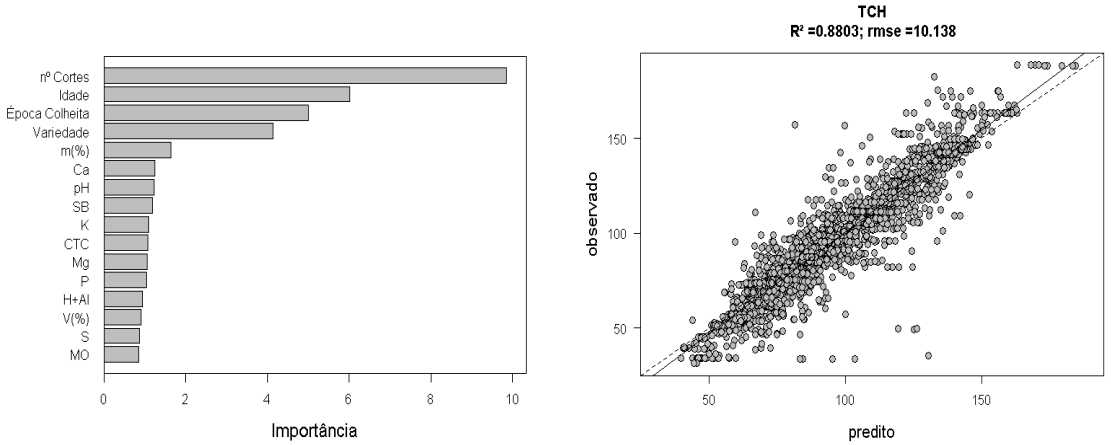


FIGURA 3. Ordem de importância dos atributos selecionados no ano 2016 a 2018 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de troca de cátions, 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.

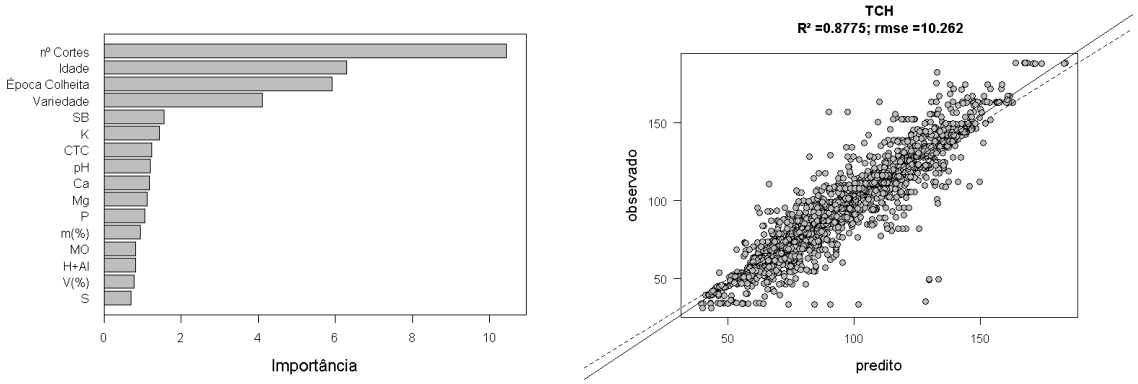


FIGURA 4. Ordem de importância dos atributos selecionados no ano 2016 a 2018 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de troca de cátions, 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.

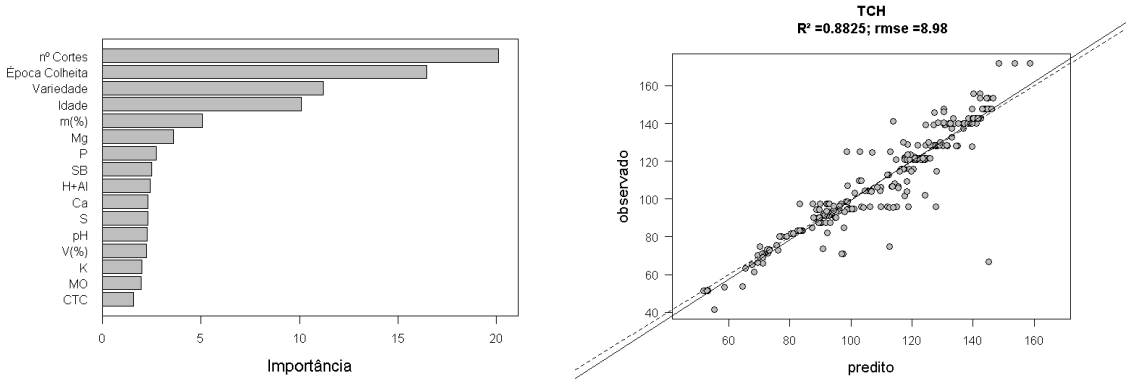


FIGURA 5. Ordem de importância dos atributos selecionados na unidade Catanduva no ano 2016 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de 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.

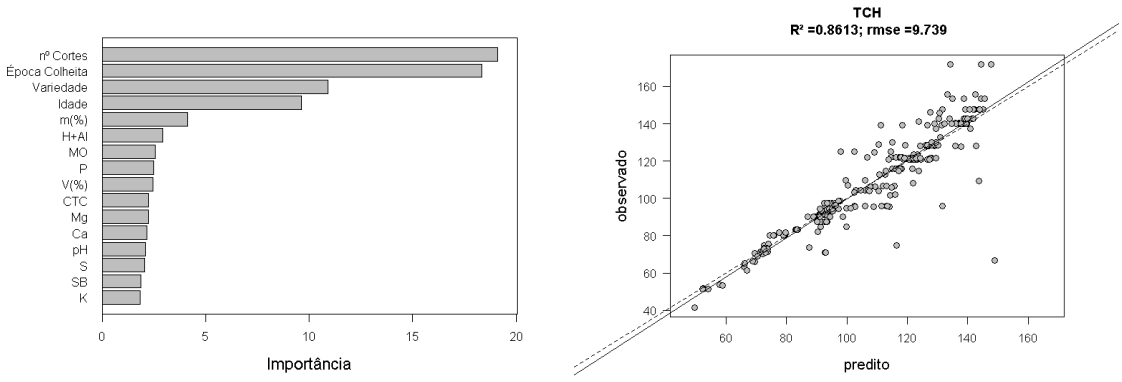


FIGURA 6. Ordem de importância dos atributos selecionados na unidade Catanduva no ano 2016 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de 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.

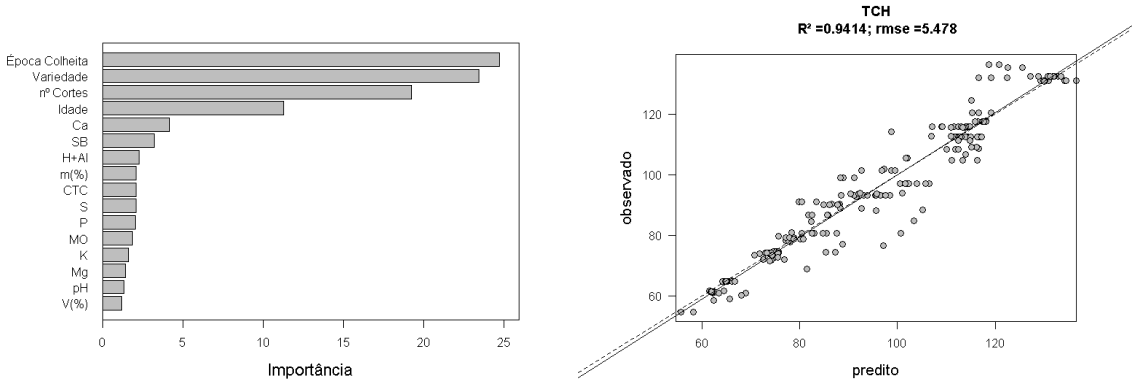


FIGURA 7. Ordem de importância dos atributos selecionados na unidade Catanduva no ano 2017 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de 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.

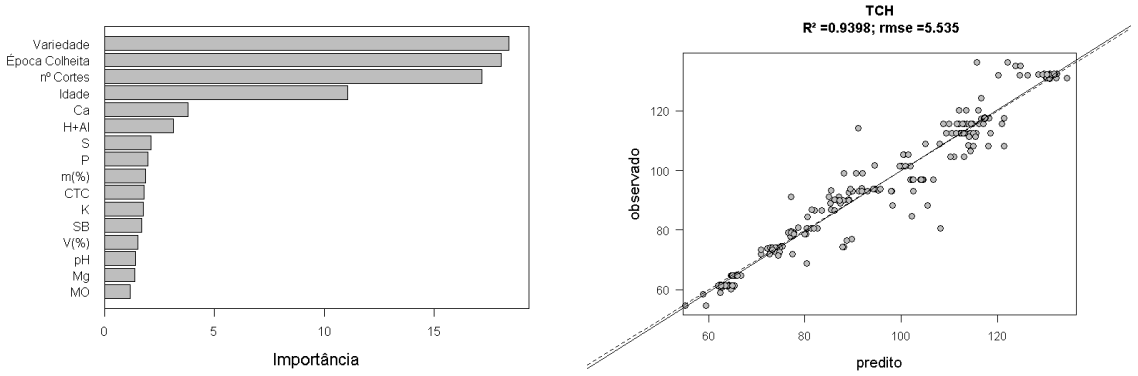


FIGURA 8. Ordem de importância dos atributos selecionados na unidade Catanduva no ano 2017 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de 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.

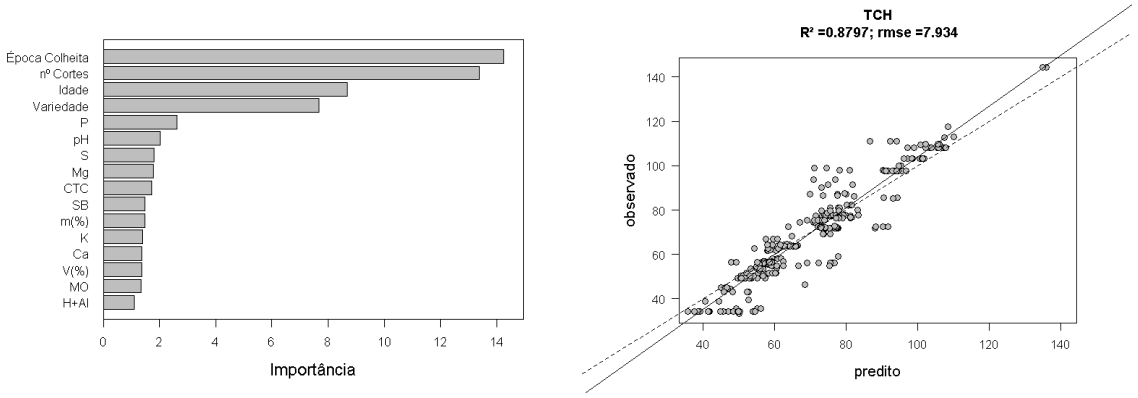


FIGURA 9. Ordem de importância dos atributos selecionados na unidade Catanduva no ano 2018 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de 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.

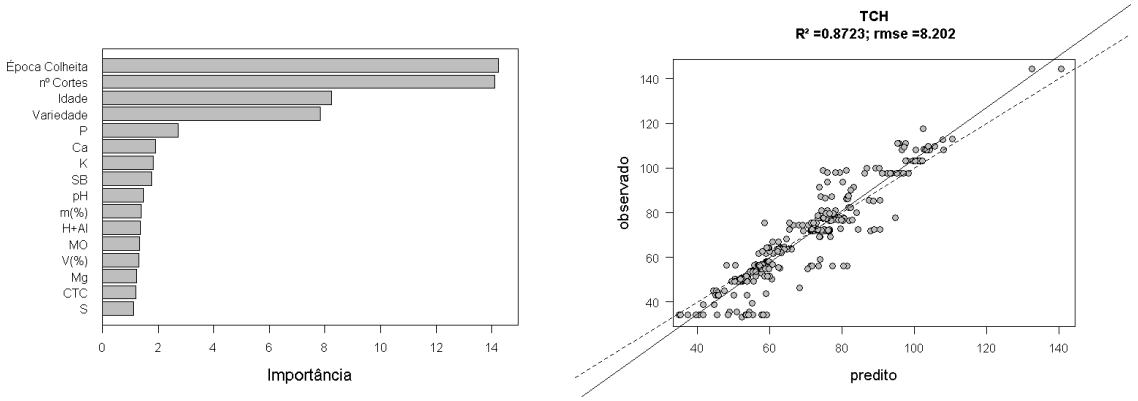


FIGURA 10. Ordem de importância dos atributos selecionados na unidade Catanduva no ano 2018 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade de 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.

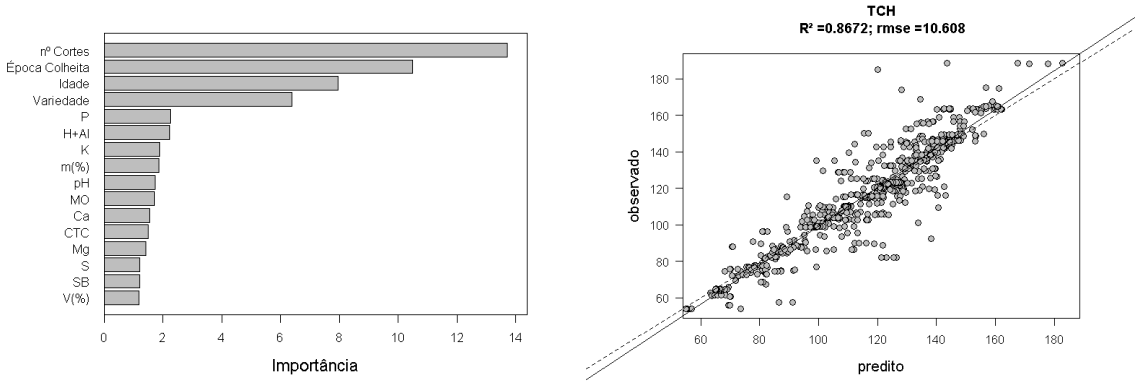


FIGURA 11. Ordem de importância dos atributos selecionados na unidade Potirendaba no ano 2016 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade 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.

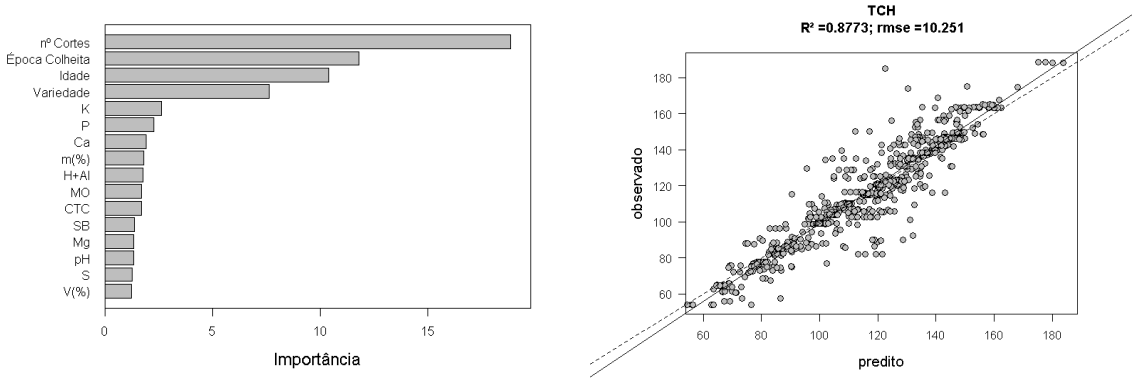


FIGURA 12. Ordem de importância dos atributos selecionados na unidade Potirendaba no ano 2016 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade 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.

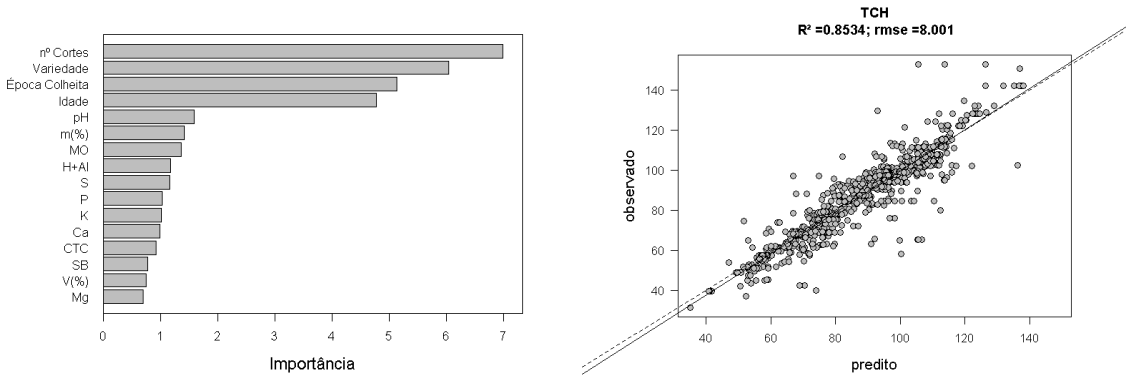


FIGURA 13. Ordem de importância dos atributos selecionados na unidade Potirendaba no ano 2017 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade 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.

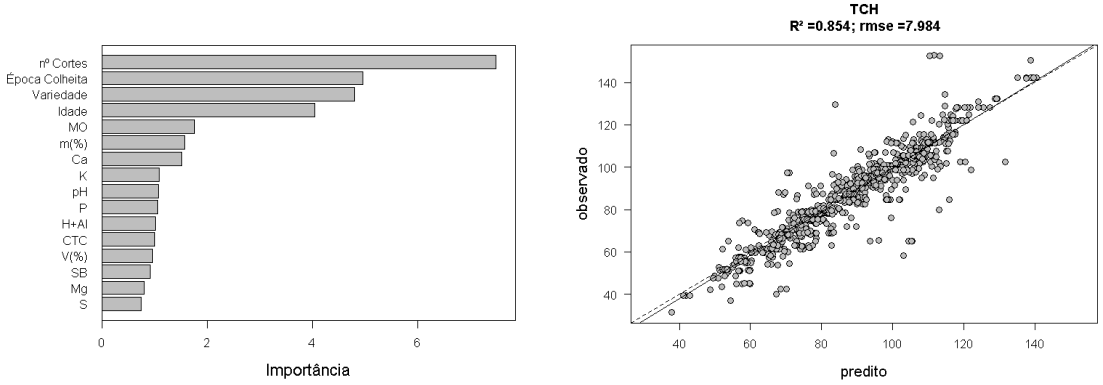


FIGURA 14. Ordem de importância dos atributos selecionados na unidade Potirendaba no ano 2017 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade 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.

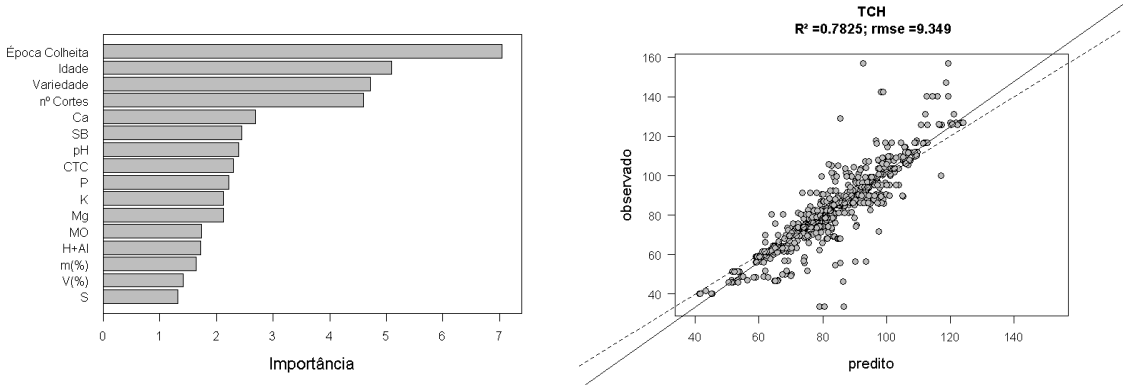


FIGURA 15. Ordem de importância dos atributos selecionados na unidade Potirendaba no ano 2018 na camada 0,00 a 0,25 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade 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.

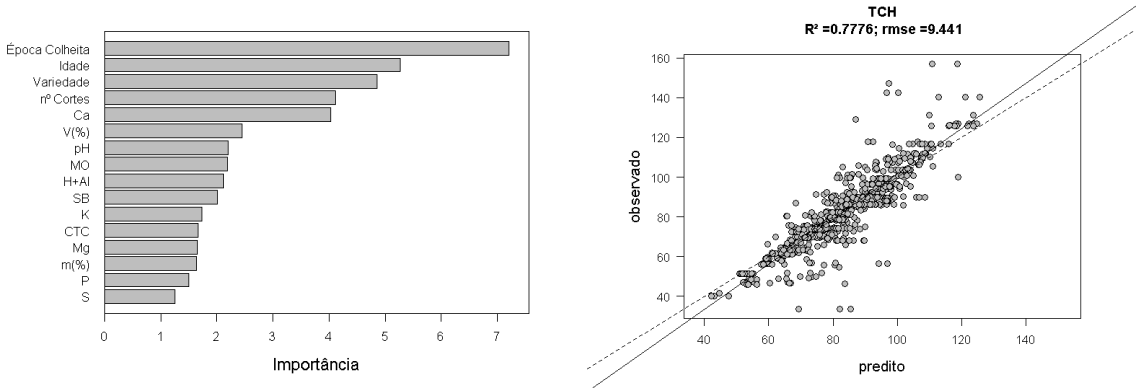


FIGURA 16. Ordem de importância dos atributos selecionados na unidade Potirendaba no ano 2018 na camada 0,25 a 0,50 m para utilização no modelo gerado pelo algoritmo *Random Forest* para previsão da produção (TCH) da cana-de-açúcar. Ca = teor de cálcio, pH = potencial hidrogeniônico, CTC = capacidade 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.

5. CONCLUSÃO

Os resultados indicam que os modelos preditivos a partir de árvores de decisão, gerados pelo algoritmo Random Forest apresentaram precisão

próximas ao observado, evidenciado pelos valores de R^2 maiores que 77% e RMSE máximo de 10.608. Assim, comparando os resultados com a prática profissional, a aplicação dessa metodologia mostrou-se promissora para suporte à tomada de decisão, visto que os resultados apresentados auxiliam no planejamento dentro das usinas e unidades produtoras.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ABERNETHY, M. Mineração de dados com Weka, parte 2: Classificação e armazenamento em cluster. Disponível em <http://www.ibm.com/developerworks/br/opensource/library/os-weka2/> Acesso em 13 mar. 2021.

ANDERSON, D. L.; PORTIER, K. M.; OBREZA, T. A.; COLLINS, M. E.; PITTS, D. J. Tree Regression Analysis To Determine Effects Of Soil Variability On Sugarcane Yields. **Soil Sci. Soc. Am. J.,** v. 63, n. 3, p. 592–600, 1999.

ASSAD, L.; PANCETTI, A. A silenciosa revolução das TICs na agricultura. **ComCiência (UNICAMP),** Campinas (SP), v. 110, 2009.

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.

BOCCA, F. F. Produtividade de cana-de-açúcar: caracterização dos contextos de decisão e utilização de técnicas de mineração de dados para modelagem. **Dissertação (mestrado)** - Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola, Campinas, SP. 2014. 86 f.

BOLSONELLO, J. E. K., FREDDO, A. R., GALLINA, A. L., ELIAS, F., STARIKOFF, K. R. Aplicação do aprendizado de máquina sobre a qualidade da

produção agroecológica de leite**, Anais da IX Jornada de Iniciação Científica e Tecnológica - IX JIC**, v. 1 n. 9, 2019.

BREIMAN, L. Random Forests. **Machine Learning,** v. 45, p. 5–32, 2001.

CARDOZO, N.P. **Modelagem da maturação da cana-de-açúcar em função de variáveis meteorológicas.** Dissertação (Mestrado em Ciências) – Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2012.

CASAGRANDE, A. A. **Crescimento da cana-de-açúcar**. Stab, Açúcar, Álcool e Subprodutos,Piracicaba, v.14, n.5, p.7-8, 1996.

CASTRO, P.R.C. Fotossíntese na cana-de-açúcar. STAB, v.20, n. 3, p. 26-27, 2003.

CECHIN, I. Uso de sistemas portáteis de fluorescência na avaliação do estresse. In: CONGRESSO DA SOCIEDADE BOTÂNICA DE SÃO PAULO, São Carlos. Anais. São Carlos, Universidade Federal de São Carlos, 1996, p.1-28.

CHAPMAN, P.; CLINTON, R.; KERBER, R.; et al. CRISP-DM 1.0 - Step-by-step data mining guide. SPSS Inc., 2000.

CHAROEN-UNG, P.; MITTRAPIYANURUK, P., Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques, 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1-6, 2018.

CHEAVEGATTI-GIANOTTO, A; et al. **Sugarcane (Saccharum X officinarum): a reference study for the regulation of genetically modified cultivars in Brazil.** Tropical Plant Biology, California, v. 4, n. 1, p. 62-89, mar. 2011.

Companhia Nacional de Abastecimento. **Acompanhamento da safra brasileira de cana-de-açúcar**, v. 1, Brasília : Conab, 2013.

CONSECANA, 2013. Procedimentos e normas para o acompanhamento de análise da qualidade da cana-de-açúcar. Disponível em < http://www.cana.com.br/biblioteca/manual_consecana_2013.pdf >. Acesso: 21 abr. 2021.

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.

EVERINGHAM, Y., SEXTON, J., SKOCAJ, D. INMAN-BAMBER, G. Accurate prediction of sugarcane yield using a random forest algorithm **Agron. Sustain. Dev.**, v. 36, n. 27, 2016.

FAO, 2018. OECD-FAO AGRICULTURAL OUTLOOK 2018-2027 © OECD/FAO 2018. Disponível em: < http://www.fao.org/3/i9166e/i9166e_Chapter5_Sugar.pdf > Acesso: 21 abr. 2021.

FIGUEIREDO, P. Breve história da cana-de-açúcar e do papel do Instituto Agronômico no seu estabelecimento no Brasil. In: DINARDO-MIRANDA, L. L.; VASCONCELOS, A. C. M. de; LANDELL, M. G. de A. (Ed.). Cana-de-açúcar. Campinas: Instituto Agronômico, 2008. p. 31-44.

GARCIA, E., CAMOLESI Jr, L. Aplicação do processo de descoberta de conhecimento em base de dados agrícola para reconhecimento de fatores que impactam na produtividade da cana-de-açúcar. In: **XI CONGRESSO NACIONAL DE EXCELÊNCIA EM GESTÃO**, anais..., 2015.

GAROFALO, D. F. T., PACKER, A. P., RAMOS, N. P., KONDO, V. Y., MATSUURA, M. I. da S. F., CABRAL, O. M. R. Dinâmica na cultura da cana-de-açúcar no Brasil: 1990 a 2018 **Jaguariúna: Embrapa Meio Ambiente,** 2020, 41p.

HAMMER, R, G. Modelagem da produtividade da cultura da cana de açúcar por meio do uso de técnicas de mineração de dados. **Dissertação (mestrado)** – Universidade de São Paulo, Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, SP. 2016. 99 f.

HAN, J.; KAMBER, M.; PEI,J. Data mining concepts and techniques. USA: **Elsevier,** 2012.HOSHGOFTAAR, 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.

ISAAKS, E. H.; SRIVASTAVA, R. M. Applied geostatistics. Nova York: **Oxford University Press**, 1989. 561 p.

JAMES, G.; et al. An Introduction to Statistical Learning with Applications in R. **Springer Texts in Statistics**. New York p.303-321, 2017.

KOHLHEPP, G. **Análise da situação da produção de etanol e biodiesel no Brasil.** Estudos Avançados, São Paulo, v. 24, n. 68, p. 223-253, 2010.

LEITE, G.M.V.; et al. **Efeitos de fontes e doses de silicato de cálcio no rendimento agrícola e na qualidade tecnológica da cana-de-açúcar, cultivar SP80-1816.** Ciência e Agrotecnologia, Lavras, v. 32, n. 4, p. 1120-1125, 2008.

LEITE, R.C.C.; LEAL, M.R.L.V. **O biocombustível no Brasil.** Novos Estudos - CEBRAP, São Paulo, v. 78, p. 15-21, jul. 2007.

MASSARUHÁ, S. M. F. S.; MOURA, M. F.; LEITE, M. A. A. Análise de tendências da produção técnico científica em tecnologias avançadas no agronegócio. **In: X Congresso Brasileiro de Agro informática**, 2015.

MASSRUHÁ, S. M. F. S; LEITE, M. A. de A. AGRO 4.0 – Rumo à agricultura digital. Disponível em: < https://ainfo.cnptia.embrapa.br/digital/bitstream/item/166203/1/PL-Agro4.0-JC-na-Escola.pdf >. Acesso: 21 abr. 2021.

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/.

RODRIGUES, D.; ORTIZ, L. **Em direção à sustentabilidade da produção de etanol de cana-de- -açúcar no Brasil**. Porto Alegre: Núcleo Amigos da Terra, 2006. 37p.

RUFATTO, L. Mineração de dados aplicada na agricultura - análise de insumos e safra. 2015. **Trabalho de Conclusão de Curso** (Tecnologia em Análise e desenvolvimento de Sistemas) - Universidade Tecnológica Federal do Paraná, Medianeira, 2015.

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).

SEGATO, S.V. et al. **Aspectos fenológicos da cana-de-açúcar.** In: SEGATO, S.V. Atualização em produção de cana-de-açúcar. Piracicaba, p.19-36, 2006.

SOUZA, A.P. et al. **Elevated CO2 increases photosynthesis, biomass and productivity, and modifies gene expression in sugarcane**. Plant, Cell & Environment, v.31, p.1116-1127, 2008.

UDOP – União dos Produtores de Bioenergia. **Energia da cana-de-açúcar no Brasil.** 2016. Disponível em: . Acesso em: 9 maio 2016.

WEBSTER, R.; OLIVER, M. A. Statistical methods in soil and land resource survey. **New York: Oxford University Press,** 1990. 328 p.

WILEY, J. F. R Deep Learning Essentials. Birmingham, UK: **Packt Publishing Ltda**, 2016. p.

WITTEN, I. H.; FRANK, E. Data mining : practical machine learning tools and techniques. 2nd ed. Amsterdam; Boston, MA: Morgan Kaufman, 2005.

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