UNIVERSIDADE ESTADUAL DE CAMPINAS

FACULDADE DE ENGENHARIA AGRÍCOLA

JHONNATAN ALEXANDER YEPES GUARNIZO

MÉTODOS SUPERVISIONADOS DE MACHINE LEARNING APLICADOS À PRODUTIVIDADE AGRÍCOLA DE CANA-DE-AÇÚCAR

CAMPINAS

2021

JHONNATAN ALEXANDER YEPES GUARNIZO

MÉTODOS SUPERVISIONADOS DE MACHINE LEARNING APLICADOS À PRODUTIVIDADE AGRÍCOLA DE CANA-DE-AÇÚCAR

Dissertação apresentada à Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Agrícola, na área de Máquinas Agrícolas.

Orientador: Prof. Dra. Barbara Janet Teruel Mederos

ESTE TRABALHO CORRESPONDE À VERSÃO FINAL DA DISSERTAÇÃO DEFENDIDA PELO ALUNO JHONNATAN ALEXANDER YEPES GUARNIZO, E ORIENTADO PELA PROF. DRA. BARBARA JANET TERUEL MEDEROS .

CAMPINAS

2021

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Rose Meire da Silva - CRB 8/5974

Y43m

Yepes Guarnizo, Jhonnatan Alexander, 1991-
Métodos supervisionados de machine learning aplicados à produtividade agrícola de cana-de-açúcar / Jhonnatan Alexander Yepes Guarnizo. – Campinas, SP : [s.n.], 2021.

Orientador: Barbara Janet Teruel Mederos.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Engenharia Agrícola.

1. Aprendizado de máquina. 2. Agricultura e tecnologia. 3. Cana-de-açúcar - Rendimento. 4. Floresta aleatória. I. Teruel Medeiros, Barbara Janet, 1966-. II. Universidade Estadual de Campinas. Faculdade de Engenharia Agrícola. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Supervised machine learning methods applied to sugarcane yield
Palavras-chave em inglês:
Machine learning
Agriculture and technology
Sugarcane - Yield
Random forest
Área de concentração: Máquinas Agrícolas
Titulação: Mestre em Engenharia Agrícola
Banca examinadora:
Barbara Janet Teruel Mederos [Orientador]
Antonio Pires de Camargo
Inácio Henrique Yano
Data de defesa: 26-05-2021
Programa de Pós-Graduação: Engenharia Agrícola

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-9313-9834
- Currículo Lattes do autor: http://lattes.cnpq.br/2977125307776889

Este exemplar corresponde à redação final da **Dissertação de Mestrado** defendida por **Jhonnatan Alexander Yepes Guarnizo**, aprovada pela Comissão Julgadora em 26 de maio de 2021, na Faculdade de Engenharia Agrícola da Universidade Estadual de Campinas.

Profa. Dra. Barbara Janet Teruel Mederos – Presidente e Orientador

Prof. Dr. Antonio Pires de Camargo – **Membro Titular**

Dr. Inácio Henrique Yano – **Membro Titular**

A Ata da defesa com as respectivas assinaturas dos membros encontra-se no SIGA/Sistema de Fluxo de Dissertação e na Secretaria do Programa da Unidade.

DEDICATÓRIA

Aos meus pais, em reconhecimento aos seus incontáveis ensinamentos, sacrifícios feitos e valores infundidos.

Aos meus irmãos por sua voz de apoio e conselhos em momentos difíceis.

A minha filha pela motivação, por ser o combustível da minha e razão das minhas alegrias.

AGRADECIMENTOS

A Deus por suas bênçãos, força e sabedoria ao longo da minha vida

A minha família, pelo suporte e apoio em cada projeto que proponho.

A meus amigos que se tornaram em minha família Felipe, Iván, Jeison e Nicolás por seu incondicional apoio nos dias mais difíceis, seus conselhos e pelo convite para realizar este sonho.

A minha orientadora professora doutora Bárbara Teruel Medeiros, por ser minha bússola, a partir de sua experiência e conhecimento que foram transmitidos para minha formação.

A Faculdade de Engenharia Agrícola (FEAGRI) pela disponibilização de sua infraestrutura para o desenvolvimento de meu projeto.

A Centro de Ensino e Pesquisa em Agricultura (CEPAGRI), pelo fornecimento dos dados meteorológicos usados para minha pesquisa.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

RESUMO

A indústria canavieira brasileira é responsável por 39% da produção mundial, considerada a maior produtora e processadora dessa matéria-prima. A previsão da produtividade da cana-de-açúcar é valiosa para a tomada de decisões precisas, sendo os métodos de aprendizado de máquinas uma das ferramentas que auxiliam na predição. O estudo teve como objetivo avaliar o desempenho de três algoritmos de aprendizado de máquina supervisionado (*Random Forest*, *Naive Bayes* e Árvores de Decisão) para obter modelos de classificação de produtividade, baseado nos atributos de fertilidade do solo em plantio de cana-de-açúcar. Os experimentos ocorreram no campo experimental da Faculdade de Engenharia Agrícola, e foi plantada a variedade de cana-de açúcar IACSP97-4039, que tem grande participação nas usinas do estado de São Paulo, é de ciclo curto e com alto conteúdo energético. A produtividade obtida em campo foi classificada em três faixas de valores (baixo, médio e alto), a comparação de desempenho dos métodos explorados apontou a *Random Forest* como o maior índice de instâncias corretamente classificadas, superando o desempenho das Árvores de Decisão. O *Naive Bayes*, que é pouco usado para estes fins, apresentou melhor desempenho na distinção entre as classes "média" e "baixa", demostrando potencial para a descrição de modelos de classificação da produtividade agrícola.

Palavras-chave: Previsão de safra, classificação multiclasse, agricultura computacional, *Radom Forest*, *Naive Bayes*, Árvores de Decisão

ABSTRACT

The Brazilian sugarcane industry is responsible for 39% of world production, considered the largest producer and processor of this raw material. The forecast of sugarcane yield is valuable for making accurate decisions and the supervised machine learning algorithms tools help in estimating the biomass obtained at the end of the harvest. The study aimed to evaluate the performance of three supervised machine-learning algorithms (Random Forest, Naive Bayes and Decision Trees) to obtain productivity classification models based on soil fertility attributes in sugarcane plantations. The experiments were carried out in the experimental field of the School of Agricultural Engineering, the variety of sugarcane IACSP97-4039 was planted, which has a large participation in the sugar mills of the state of São Paulo, is a short-cycle species and a high-energy content. The productivity obtained in the field was classified into three ranges of values (low, medium and high), the performance comparison of the explored methods pointed to Random Forest as the highest rate of correctly classified instances, surpassing the performance of the decision trees. Naive Bayes, which is little used for these purposes, performed better in distinguishing between the “middle” and “low” classes, showing potential for the description of models for classification of agricultural productivity.

Keywords: Yield forecast, multi-class classification, computational agriculture, Random Forest, Naive Bayes, Decision Trees

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

B: Boro.

Ca: Cálcio.

CEPAGRI: Centro de Ensino e Pesquisa em Agricultura.

CONAB: Companhia Nacional de Abastecimento.

EDA (*Exploratory Data Análise*): Análise Exploratório de Dados.

FAO: Organização das Nações Unidas para a Alimentação e a Agricultura.

Fe: Ferro

FEAGRI: Faculdade de Engenharia Agrícola.

IAC: Instituto Agronômico de Campinas.

IDW (*Inverse Distance Weighted*): Ponderação pelo Inverso da Distância.

K: Potássio.

ML: Machine Learning.

MAE (*Mean Absolute Error*): Erro médio absoluto.

Mn: Manganês.

M.O.: Matéria orgânica.

P: Fósforo.

PIB: Produto Interno Bruto.

TPH: Toneladas por hectare.

RF (*Random Forest*): Floresta aleatória.

RMSE (*Root Mean Square Error*): Raiz do erro médio quadrático.

ROC (*Receiver Operating Characteristics*): Características Operacionais do Receptor.

SiBCS: Sistema Brasileiro de Classificação de Solos.

UNICAMP: Universidade Estadual de Campinas.

Zn: Zinco.

SUMÁRIO

1. INTRODUÇÃO

O agronegócio brasileiro cada ano aumenta sua participação do Produto Interno Bruto (PIB). Para o ano de 2020 atingiu uma participação de 16,18%, representando um crescimento de 2,78% em relação ao ano imediatamente anterior, sendo que isso é produto de políticas públicas, investimentos de infraestrutura, inclusão de subsídios e tecnologia para agricultura a fim de garantir um bem-estar económico e social a pesar dos impactos negativos da pandemia sobre diferentes atividades do setor (CNA, 2021). De acordo com o Ministério da Economia para o ano 2019, o Valor Bruto da Produção Agropecuária (VBP) alcançou uma cifra de R$651,5 bilhões. O ranking dos maiores faturamentos da agroindústria brasileira está composto pela soja (R$175,63 bilhões), carne bovina (R$139,71 bilhões), milho (R$90,70 bilhões) e leite (R$50,86 bilhões), sendo que a quinta posição é ocupada pela indústria da cana-de-açúcar com uma cifra de R$47,43 bilhões. A cana-de-açúcar é foco de múltiplos estudos no território nacional pela relevância na economia brasileira, consequência da flexibilidade como matéria prima a qual brinda um grande portfólio de produtos obtidos a partir do seu processamento, que além de açúcar e etanol, gera energia elétrica a través da queima do bagaço, vinhaça para a recuperação de solos de vocação agrícola, óleo fóssil usado na indústria química como componente de cosméticos e bebidas alcoólicas entre outros.

Segundo o censo realizado pela Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) em 2017, o Brasil é o maior produtor mundial de cana-de-açúcar, com uma produção de 736 milhões de toneladas que representa 39% da produção mundial, seguido pela Índia (18%) e China (7%). Em dados mais recentes, fornecidos pela Companhia Nacional de Abastecimento (CONAB, 2020), para novembro de 2020, o Brasil tinha 8.409,8 mil hectares da cobertura total do país destinadas à produção de cana-de-açúcar. No entanto, em termos de produtividade, ocupa a vigésima oitava posição do mundo, com uma produtividade de 76,3 toneladas por hectare (TPH), longe do Peru, Guatemala e Senegal, países com capacidade de produzir 121,2, 121 e 108,1 TPH, respetivamente (FAO, 2017).

Devido à importância desta cultura na economia do país, diferentes centros de pesquisa promovidos pelo governo brasileiro estão gerando múltiplos avanços de ordem científica e tecnológica para melhorar a produtividade desta cultura. A estimativa da safra está se tornando um elemento fundamental para o desenvolvimento rural, como um indicador na segurança alimentar e produtiva de um país. Conhecer a massa fresca projetada em uma colheita é importante para o comércio na agroindústria, garantindo a demanda nacional e ordenando a

logística correspondente ao transporte da cana e seu posterior processamento (BASTIAANSSEN; ALI, 2003).

Por este motivo, um dos ramos de interesse para os agricultores e o Governo é a previsão de safra a partir de modelos matemáticos semi-empíricos que relacionam fatores climatológicos, edáficos e culturais, ou através de diferentes índices de vegetação, sendo o Índice de Vegetação da Diferença Normalizada (NDVI) o método mais utilizado para esta tarefa (MIHAI; FLORIN, 2016).

Nesse cenário, o *Machine Learning* (ML) ou aprendizado de máquinas, nas últimas décadas vem sendo responsável por encontrar correlações, conexão de informações ou padrões entre atributos que compõem uma rede de dados através da adoção de diferentes algoritmos para a geração de modelos com a finalidade de agrupar, classificar ou predizer eventos (TERDAL, 2019). O desempenho de um modelo criado a partir de um algoritmo de aprendizado de máquina é medido por uma métrica de desempenho aprimorada com a experiência ao longo do tempo com a alimentação contínua da base de dados. Usando a experiência adquirida durante o processo de treinamento é gerado conhecimento desde os padrões criados para prever tendências futuras e gerar um valor agregado ao conjunto de dados explorado (MEDAR; RAJPUROHIT, 2014; LIAKOS *et al*., 2018).

Algoritmos de aprendizado de máquinas são aplicáveis em projetos interdisciplinares. Essas técnicas são divididas em três grupos: aprendizagem não supervisionada, aprendizagem por reforço e aprendizagem supervisionada. Neste último grupo, a literatura está focada no desempenho de métodos preditivos para determinação de eventos futuros. A presente pesquisa busca relacionar as propriedades químicas do solo usando técnicas de aprendizado de máquinas supervisionados para a avaliação de algoritmos de classificação da produtividade em cultura de cana-de-açúcar de ciclo curto.

2. HIPÓTESE

É possível gerar um modelo estatisticamente confiável para a previsão e classificação da produtividade de cana-de-açúcar de ciclo curto avaliando diferentes algoritmos de aprendizado de máquinas supervisionados, em busca de incorporar o estudo simultâneo da interação existente entre as propriedades químicas do solo e a resposta produtiva da planta como alternativa dos modelos agrometereológicos e modelos baseados em índices de vegetação existentes no setor canavieiro para a projeção de planos de ordenamento territorial, segurança alimentar, logística, armazenamento e processamento de produto fresco.

3. OBJETIVOS

3.1. Objetivo geral

Realizar a avaliação de desempenho de diferentes algoritmos de aprendizado de máquinas supervisionados, partindo das propriedades químicas do solo em busca de um modelo estatisticamente satisfatório de regressão ou classificação da produtividade agrícola de um lote de cana de açúcar da variedade IACSP97-4039.

3.2. Objetivos específicos

- Determinar as propriedades químicas do solo que mais influenciam a produção de biomassa na cultura de cana-de-açúcar em base à categorização de variáveis dos métodos de aprendizado de máquinas utilizados.
- Avaliar o desempenho do *Random Forest* como algoritmo preditor da produtividade de cana-de-açúcar.
- Comparar as métricas de análise estatística dos modelos de classificação dos algoritmos *Naive Bayes*, Árvores de Decisão e *Random Forest.*
- Estabelecer recomendações para direcionar futuras pesquisas que envolvam métodos supervisionados de aprendizado de máquinas com a produtividade agrícola.

4. REVISÃO BIBLIOGRÁFICA

4.1. Importância socioeconômica

A cana-de-açúcar (*Saccharum spp.*) é uma planta semiperene. Atualmente, a origem da cana é um tema de discussão. Rodolfo Junior em 2015, insiste que é originária da Oceania, no começo foi utilizada como planta ornamental pelos habitantes da Nova Guiné; enquanto Scarpari em 2002 apoiado por Doorenbos & Kassam (1979), datam que a origem desta espécie vegetal está no sudeste da Ásia para iniciar a sua expansão no ano 8000 a.C.

Martin Alfonso de Souza foi quem introduziu a cana-de-açúcar em território brasileiro em 1502 com variedades provenientes da Ilha da Madera e a partir de 1530 iniciou-se um período de ascensão econômica na área agrícola. O primeiro engenho canavieiro chamado “São Jorge dos Erasmos” data do ano de 1532 na Capitania de São Vicente, atualmente é considerado Monumento Nacional e faz parte do patrimônio da Universidade de São Paulo (USP) usado como centro de pesquisa cultural e um espaço turístico e arqueológico do sudeste brasileiro. Não obstante, esta cultura teve um crescimento mais acelerado e favorável no Nordeste, principalmente nos estados do Pernambuco e Bahia que se tornaram em portos de exportação deste produto para território europeu, trazendo como consequência um papel relevante na economia nacional entre os séculos XVI e XVIII (CARVALHO *et al*., 2013).

Em meados do século XVIII, a instalação de usinas para o processamento em grande escala de cana-de-açúcar começou a ser promovida como consequência do declínio da produção de café no estado de São Paulo. Produtores e novos investidores tiveram acesso a créditos para garantir infraestrutura e tecnologia. Isto gerou um ponto de inflexão na produção de açúcar no estado, e inumeráveis engenhos foram instalados nas regiões de Campinas, Itu, Moji Guaçu, Piracicaba e ao Norte do Estado, nas vizinhanças de Ribeirão Preto, denominado "quadrilátero do Açúcar", onde o centro era o atual Município de Piracicaba. (RODOLFO JUNIOR, 2015).

A produção de biocombustíveis começou em 1920, no campo das energias renováveis, a cana-de-açúcar se tornou um negócio confiável e seguro. No entanto, a produção não satisfazia os requisitos do mercado nacional devido a problemas fitossanitários constantes e baixos teores energéticos em termos de quantidade de açúcar produzida em relação à biomassa colhida (CARVALHO *et al*., 2013).

Kezzy de Morais *et al*., em 2015, afirmam que estas situações abriram a porta à necessidade do melhoramento genético focado tanto para a produção industrial como para a alimentação animal. A criação de centros de pesquisa desta índole busca gerar variedades altamente produtivas com capacidade de adaptação, tolerância e resistência a diferentes fenômenos climáticos, solos, estresse hídrico, pragas e doenças presentes no ambiente brasileiro.

Atualmente se mantêm vigentes quatro programas manejados por: o Instituto Agronómico de Campinas (IAC), a Rede Interuniversitária para o Desenvolvimento do Setor Sucroalcooleiro (RB), Canavialis (CV) e o Centro Tecnológica Canavieira (CTC) (BEZERRA *et al*., 2018).

A instabilidade dos preços e constantes aumentos nas tarifas de importação de combustíveis fósseis no final da década de 1960 foram fatores condicionais para os países em desenvolvimento que importaram petróleo geraram problemas inflacionários (THEODORO, 2011). Como resposta à situação foi promovida uma política de governo brasileiro em 1975 na qual surgiu o Programa Nacional do Álcool (Proálcool), esta proposta nasceu com a ideia de estimular a produção de etanol e fortalecer o proeminente negócio como alternativa à substituição aos derivados do petróleo, fortalecimento de instituições envolvidas no melhoramento genético e usinas para o processamento de matéria-prima (BRINKMAN *et al*., 2018).

Os diferentes subprodutos gerados a partir da cana-de-açúcar fizeram dela uma das culturas mais importantes do mundo, sendo cultivada em mais de 100 países (HELENA *et al*., 2016). É uma importante fonte de emprego nas áreas rurais das regiões tropicais e subtropicais do mundo, devido à sua vocação na produção de etanol. Atualmente, 85% das áreas mecanizadas estão concentradas no estado de São Paulo (FRANCO *et al*., 2013).

O impacto da atividade Sucroenergética no Brasil no âmbito social trouxe consigo, além da geração de emprego, a necessidade de mão de obra qualificada. As usinas promoveram o acesso à educação e criaram estruturas de capacitação contínua para a evolução do setor.

Estima-se que a indústria canavieira é responsável por cerca de 3,8 milhões de empregos diretos e 700 mil indiretos, representando 6% dos empregos totais do país. Isto se traduz em que este agronegócio tem uma participação de 2,6% no PIB brasileiro, ressaltando sua importância na estabilidade econômica do país (BRINKMAN et al., 2018).

No entanto, o impacto social da atividade sucroenergética no Brasil deve ser avaliado não apenas pelos empregos gerados. Igualmente importante é saber que o setor, espalhado por grande parte do território nacional, contribui para descentralizar a renda (SOUSA; MACEDO, 2009).

Atualmente, o Brasil é o maior produtor e exportador de açúcar do mundo com uma participação no plano internacional de 20%, sendo o continente europeu seu maior cliente. Embora a produção do biocombustível seja para o comércio interno, os recordes de desempenho que se marcam a cada ano permitiram que o país seja o segundo maior produtor de etanol no mundo, só atrás dos Estados Unidos (GIRARDI, 2019;GILIO; DIAS DE MORAES, 2016).

Neste contexto, o sucesso do posicionamento do setor sucroenergético brasileiro em mercados internacionais se deve a uma forte política de estado que incentiva o desenvolvimento e implementação de tecnologias no agronegócio, condições edafológicas, climáticas e pessoal qualificado adequado para a expansão e fortalecimento da cultura de cana-de-açúcar como substituto de combustíveis fósseis.

4.2. Produtividade agrícola brasileira

O rendimento ou produtividade agrícola é definida como a relação do produto fresco por unidade de área. A planta atua como um integrador dos estímulos. A produção final é o resultado das consequências e as interações dos efeitos: (I) ambientais e climáticos (temperatura, umidade, precipitação e radiação), (II) biológicos (fauna, flora e microrganismos), (III) a intervenção do homem como o único fator externo com a possibilidade de intervir na tomada de decisões, (IV) da geografia e topografia onde se estabelecem as culturas e (V) a distribuição dos atributos químicos, nutricionais e estrutura física da composição do solo (SINGH; SARWAR; SHARMA, 2017; TERDAL, 2019).

A inserção de motores que usam bicombustível (flex) no mercado brasileiro e a geração de energia elétrica a partir da cana-de-açúcar, em 2002, impactaram a competitividade do setor canavieiro, na constante busca por produzir maiores quantidades de biomassa. Isto trouxe aumento da oferta e redução de custos, contribuindo para ampliar a sustentabilidade do setor (CONAB, 2019). Apenas nas safras 2001/02 e 2019/20, a produtividade passando de 293 a 643 milhões de toneladas. A produção de etanol naquele período passou de 11 para 32 bilhões de reais (CONAB, 2020).

O Brasil, segundo o último censo realizado pela Organização das Nações Unidas para a Alimentação e a Agricultura (FAO) em 2017, é o principal produtor de cana-de-açúcar no mundo. Em termos de produção de etanol, é o terceiro maior produtor do mundo, atrás apenas dos Estados Unidos e da China. Porém, ao contrário de outros países, o Brasil pode aumentar sua produção de biocombustíveis, gerando risco no abastecimento de alimentos (SILVA; CASTAÑEDA-AYARZA, 2021).

O relatório de acompanhamento da safra realizado pela Companhia Nacional de Abastecimento (CONAB) em dezembro de 2020, cita que 10.039 milhões de hectares da área total do país está ocupada com cana-de-açúcar. De esta, 265,3 mil hectares estão ao cultivo de mudas; 1.272,6 mil hectares em áreas de plantio e 8.481,2 mil de hectares em área colhida. Somente no Estado de São Paulo está concentrada 50,86% da produção total do país.

No entanto no mesmo censo de 2017 publicado pela FAO, em termos de produtividade, o Brasil ocupa a vigésima oitava posição do mundo, com uma produtividade média de 74,5 toneladas por hectare (TPH), longe do Peru, Guatemala e Senegal, países com capacidade de produzir 121,2, 121 e 108,1 TPH. Para a safra de 2019, a média nacional bateu um novo recorde de 75,78 TPH, sendo o Nordeste e o Sul brasileiros os que apresentam as maiores limitações na produtividade com rendimento de 58,82 e 63,95 TPH, respectivamente. Os investimentos realizados por parte do governo, bem como as melhores condições climáticas ocorridas, indicam uma produtividade 9,8% maior que na última safra. Entretanto, em São Paulo, o rendimento estimado é de 79,23 TPH, representando aumento de 5,4% em relação à safra anterior. Para esse aumento de produtividade, as chuvas ocorridas em abril e maio contribuíram, além da evolução nos manejos que envolvem a cultura, desde a preparação de solo, manejo de lavoura até a colheita e processamento, otimizando a produção como um todo (CONAB, 2019).

4.3. Modelos de produtividade

Os primeiros trabalhos publicados para a previsão da produtividade em cana-de-açúcar em território brasileiro foram realizados em base à estatística clássica mediante regressões lineares como foi o caso de Friedrich (1990). Posteriormente, mediante o desenvolvimento de ferramentas computacionais realizou-se uma análise estatística com maior

base de dados, utilizando regressão linear múltipla e modelos polinomiais (Oliveira *et al*., 2012; Marcari; Rolim; Aparecido, 2015).

Estas pesquisas tinham dois fatores em comum: dados climáticos como parâmetros de entrada para integrá-lo com a resposta biológica da cana-de-açúcar, e, que não são dados consistentes, já que as estações meteorológicas brasileiras apresentaram deficiências relacionadas a manutenção dos equipamentos e complexidade de armazenamento e processamento das medições.

Embora o desempenho dos modelos seja alto, os métodos de regressão da estatística clássica como são os modelos de regressão linear e polinomial, se caracterizam pelo baixo custo de operações computacionais e rápido tempo de processamento. Porém, uma das desvantagens consiste na filtragem dos modelos torna inviável a possibilidade de ter operações matemáticas de resposta rápida com a capacidade de se adaptar a diferentes conjuntos de dados. Como resposta a estas limitações, a gestão integrada de uma grande rede de fenômenos que são mensuráveis, armazenáveis e processáveis deu origem à "Revolução dos Dados Verdes", que nasce do otimismo de que dados massivos podem e irão fornecer benefícios para as indústrias agrícolas e para a sociedade global, semelhantes ao aumento da produtividade agrícola provocado pela Revolução Verde entre as décadas de 1960 e 1980 (EVERINGHAM *et al*., 2016).

4.4. Aprendizado de máquina

O aprendizado de máquina surgiu com tecnologias de manipulação e análise de uma base de dados. A área da computação gerou uma ferramenta de alto desempenho para criar oportunidades de interpretação de valores e achar correlações dificilmente perceptíveis com estatística descritiva com aplicação multidisciplinar como é o caso das agro-tecnologias. O *Machine Learning* (ML) ou aprendizado de máquinas, consiste na identificação de padrões válidos, úteis e compreensíveis na busca de regularidades estatísticas presentes em um conjunto de dados para geração e transmissão de conhecimento, melhorando permanentemente o desempenho da máquina com a alimentação constante da base de dados que permitem a detecção, caracterização e avaliação da consistência dos atributos e dados da unidade (TERDAL, 2019 ; JIANG; GRADUS; ROSELLINI, 2020).

Os algoritmos do aprendizado de máquina estão divididos particularmente em três grupos de acordo com o resultado desejado:

- **Aprendizagem não supervisionada:** O principal objetivo deste tipo de aprendizagem é descobrir similaridades entre os objetos que compõem uma rede de dados sem rótulos. As aplicações da aprendizagem não supervisionada estão divididas em seis categorias: aprendizagem hierárquica (intimamente relacionado com o *"deep learning"* e redes neurais multicamadas), *"clustering"* ou agrupamento de dados, modelos de variáveis latentes, redução de dimensionalidade e detecção de *"outliers"* ou dados discrepantes (USAMA *et al*., 2019).
- **Aprendizagem por reforço:** É descrito como um problema de controle teórico em que o algoritmo aprende como agir com um ambiente desconhecido maximizando uma recompensa de aprendizado. Cada ação tem algum impacto no ambiente e o ambiente fornece *feedback* que orienta o algoritmo (JIN *et al*., 2019).
- **Aprendizagem supervisionada:** Busca descrever fenômenos e relações ocultas no conjunto de dados composto por atributos de entrada (denominadas variáveis independentes) e um atributo meta ou variável dependente**.** A relação existente entre os atributos gera uma estrutura chamada modelo para descrever as tarefas de regressão porque o objetivo é prever eventos futuros ou classificar um intervalo de valores ou características específicas de interesse (NASTESKI, 2017). Além das tarefas preditivas, a aprendizagem supervisionada também se encarrega de classificar atributos, estes modelos são usados para discriminar diferentes objetos ou observações a partir de características específicas presentes dentro dos atributos. Este mecanismo de aprendizagem é normalmente dividido em três etapas: treinamento, teste e avaliação. No processo de treinamento, as amostras são adotadas pelo algoritmo para gerar um modelo de aprendizagem. No processo de teste, o modelo de aprendizado usa o mecanismo de execução para fazer a previsão ou classificação no conjunto de dados armazenados para o teste. Finalmente, a avaliação de desempenho do modelo gerado é realizada de acordo com as métricas que definem o sucesso.

Liakos *et al*. em 2018 propõem que os trabalhos de aprendizado de máquinas supervisionados aplicados na agricultura estão divididos em (a) manejo de culturas: incluindo

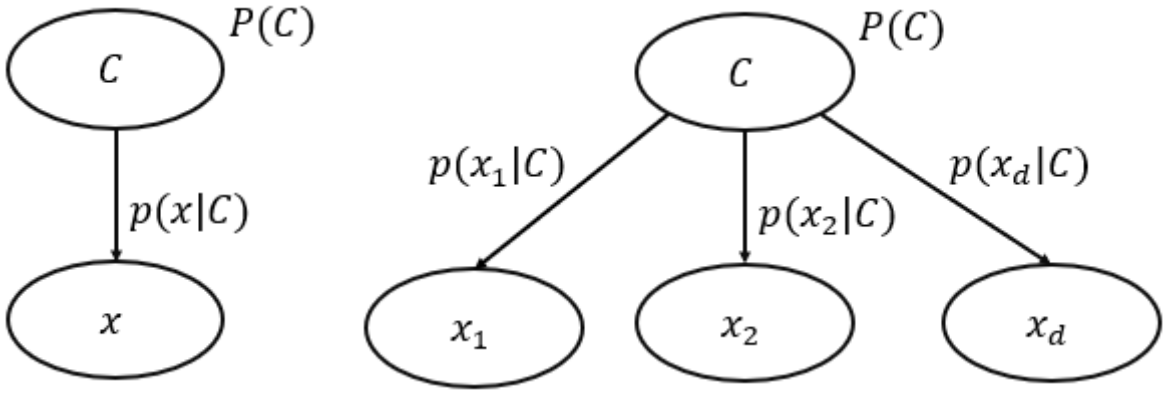
aplicações na previsão e classificação da produtividade agrícola, detecção de doenças, qualidade de culturas para detecção de plantas daninhas e reconhecimento de espécies; (b) gestão de animais, incluindo aplicações em bem-estar animal e produção animal; (c) gestão da água; e (d) manejo do solo.

4.4.1. *Naive Bayes*

Naive Bayes é uma técnica de classificação baseada no Teorema de *Baye*s, que consiste em assumir que o valor de uma característica particular é independente do valor de qualquer outra característica, dada a variável de classe (ZHANG, 2004). Em termos mais simples, este classificador supõe que a presença de uma característica particular em uma classe não está relacionada com a presença de nenhuma outra característica. Isto é chamado de independência condicional.

A suposição condicional nem sempre é a mais apropriada em aplicações da vida cotidiana, isso é gerado porque assume um modelo probabilístico subjacente e permite que a incerteza sobre o modelo seja capturada de forma baseada em princípios ao determinar as probabilidades dos resultados. Um procedimento simples para superar a limitação deste algoritmo é estender sua estrutura para representar explicitamente o maior número de dependências e características entre os atributos que compõem a rede de dados (GRANIK; MESYURA, 2017).

Isto permite que a classificação bayesiana proporcione uma perspectiva útil para compreender e avaliar os algoritmos de aprendizagem já que calcula probabilidades explícitas para hipótese e atenua o ruído nos dados de entrada. Em um conjunto de dados, dois atributos podem depender entre si, mas a dependência pode ser distribuída uniformemente em cada classe. Claramente, neste caso, é violado o pressuposto de independência condicional, mas a *Naive Bayes* continua a ser um ótimo classificador. Além disso, o que finalmente afeta a classificação é a combinação de dependências entre todos os atributos (BAŞTANLAR; OZUYSAL, 2014).



(a) (b)

Figura 1 (a) Modelo gráfico de classificação. (b) Classificador *Naive Bayes* assume entradas independentes

O modelo de classificação *Naive Bayes* é um modelo simples e efetivo para conjuntos de dados grandes. A Figura 1 representa o estudo individual dos preditores descritos pelos atributos (x) para determinar a classe (C). Usando o teorema de *Bayes*, é possível determinar a probabilidade que aconteça C, dado que x ocorreu $P(C|x)$. Em consequência, $P(x|C)$ responde à probabilidade gerada pelo atributo x dada a classe C, $P(C)$ e $P(x)$ correspondem à probabilidade de ocorrência do classificador e dos atributos respectivamente. Matematicamente, esta relação é determinada pela seguinte equação.

$$P(C|x) = \frac{P(x|C)P(C)}{P(x)} \quad (1)$$

Na agricultura, este algoritmo vem sendo aplicado a diferentes frentes. Como é a detecção e classificação de doenças bacterianas e lesões nas folhas de plantas de arroz com base em imagens RGB (AHMED *et al*., 2019), recomendação e plantio de culturas de acordo com o comportamento das mudanças climáticas (SETIADI *et al*., 2020) sugestão das quantidades de fertilizante com base no tipo de solo e irrigação das culturas através de uso de sensores de temperatura, umidade do sole e sensores de pH (PADALALU *et al*., 2017) e geração de modelos de classificação de produtos alimentícios de acordo a suas características morfológicas, de cor e textura (VEERNAGOUDA GANGANAGOWDER; KAMATH, 2017).

Em cada estudo feito com classificador *Naive Bayes* apresenta resultados promissórios no setor agrícola, porém é um algoritmo com uma menor quantidade de publicações em comparação a outros métodos de classificação como o *Support Vector Machine* (SVM), Árvores de Decisão, *Random Forest* por apresentar índices de desempenho inferiores

aos de seus concorrentes diretos devido à simplicidade probabilística com que categoriza as variáveis.

4.4.2. Árvores de Decisão

O modelo de Árvore de Decisão é um modelo supervisionado capaz de cumprir com tarefas de regressão e de classificação. Este modelo hierárquico realiza sequências de divisões em menos etapas. Uma Árvore de Decisão é composta por nós internos de decisão e folhas terminais. Os nós implementam a função de teste com resultados discretos que rotulam os ramos.

O nó de decisão ou também chamado nó raiz, está sempre no topo da estrutura, e os nós internos contêm um teste de valor sobre um dos atributos mais relevantes. Os resultados desses testes formam os ramos. Os nós-folha, nas extremidades, referem-se às classes da variável resposta e representam o resultado da predição obtida pelo modelo (WITTEN; FRANK; HALL, 2011).

Um nó folha define uma região localizada no espaço de entrada onde as instâncias que caem nesta região têm os mesmos rótulos (na classificação) ou saídas numéricas muito semelhantes (na regressão). Os limites das regiões são definidos pelos discriminantes que são codificados nos nós internos no caminho da raiz ao nó folha. A estrutura física de uma Árvore de Decisão é mostrada na Figura 2

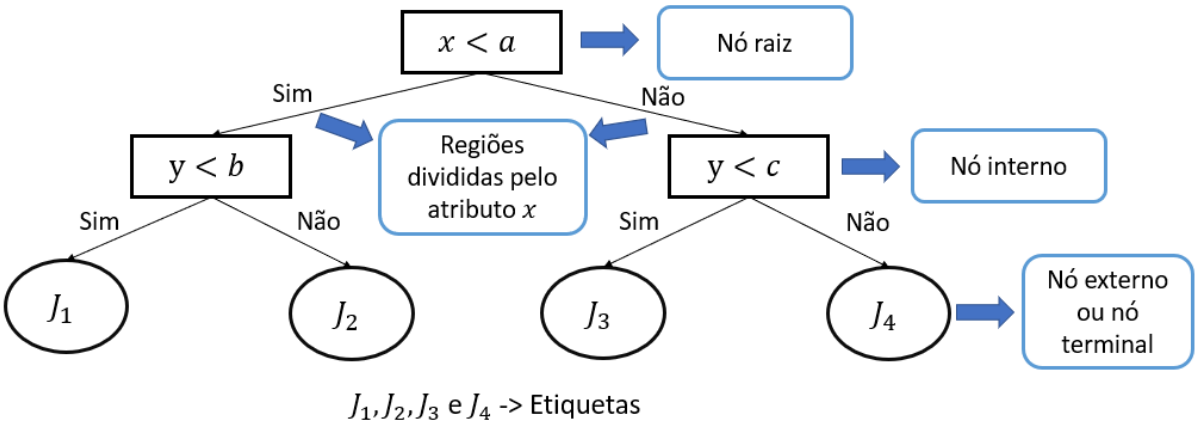


Figura 2 Estrutura clássica de uma Árvore de Decisão, adaptado de González Perea *et al*. (2019)

As equações que especificam o valor de saída previsto de um determinado vetor de entrada são os rótulos do nó principal no caso das árvores de regressão. Por outro lado, nas árvores de classificação de atributos os nó folha contêm um rótulo que indica o grupo ou classe ao qual faz parte determinado vetor de recurso (GONZÁLEZ PEREA *et al*., 2019).

Ao mesmo tempo, este é um modelo não paramétrico no sentido que não é assumido nenhuma forma paramétrica para as densidades de classe e a estrutura da árvore não é fixa a priori, a árvore que descreve o problema cresce porque são adicionados ramos e folhas durante a aprendizagem dependendo da complexidade do problema, a alimentação dos dados, e treinamento contínuo (BAŞTANLAR; OZUYSAL, 2014).

A Árvore de Decisão é uma das estruturas de dados mais comuns e poderosas em toda a ciência da computação porque o custo computacional para a construção de um árvore é mais baixo que outros algoritmos (MARSLAND, 2014). Outro dos grandes benefícios é o fato de fácil entendimento e sua interpretação pode ser transformada em um conjunto de regras de "se-então" de acordo com a estrutura formada pelos dados. Seguir este conjunto de regras permite obter uma resposta visível e transparente, o que faz que os cientistas de dados tenham uma maior confiança do que obter uma resposta de um algoritmo "caixa preta".

A árvore é construída com o atributo que fornece maior informação, ou seja, aquela com a entropia mais alta. Depois de usar esse recurso, e reavaliada a entropia de cada atributo e novamente é escolhida aquele com a entropia mais alta. Em consequência, o desempenho deste método de aprendizado de máquinas é altamente sensível pela divisão do conjunto de dados de treinamento, gerando uma instabilidade pela criação de árvores que seguem a tendência de classes dominantes.

Para o sector agropecuário, a aplicação de Árvores de Decisão tem apresentado um grande número de trabalhos transversais que incluem a classificação da qualidade da água como foi o estudo conduzido por Chandanapalli, Reddy e Lakshmi (2018), que consistiu em fazer uma rede de sensores sem fios capazes de realizar uma exploração contínua da água que deriva de um estudo meticuloso das propriedades físico-químicas para fins de irrigação e como fonte de dessedentação para o gado.

Habib *et al*. (2020) usaram visão computacional para diagnóstico, reconhecimento e classificação de doenças comuns como são a mancha preta, a mancha marrom e antracnose em mamão obtendo uma com 90% de eficiência média usando imagens capturadas por câmeras de dispositivos móveis. Neste caso, as fotos foram o meio para extrair os atributos de classificação fazendo uso de um recurso típico como é a extração e transformação de informação de imagens através de pixels em vetores para o aprendizado de máquinas.

Continuando com a linha de classificação de atributos mediante a interpretação de imagens aplicadas à agricultura de precisão mediante o sensoriamento remoto, Conţiu e Groza (2016), desenvolveram um sistema inteligente híbrido capaz de explorar o processamento de imagens do Lansat 8 conjugando a determinação do Índice de Vegetação por Diferença Normalizada (NDVI) e as imagens RGB do satélite para classificar de culturas de milho, algodão, arroz e soja encapsulando um modelo robusto de Árvores de Decisão e Redes Neurais para estes fins.

4.4.3. *Random Forest*

O algoritmo supervisionado *Random Forest* é um método baseado nos princípios das Árvores de Decisão, por este motivo tem a possibilidade de cumprir com a classificação de atributos e previsão de eventos futuros. O *Random Forest* foi definido por Leo Breiman do Departamento de Estatística da Universidade da Califórnia em 2001.

Este método divide aleatoriamente o conjunto de dados em vários subconjuntos homogêneos em tamanho para gerar diferentes Árvores de Decisão em paralelo. Esta fase é chamada *bootstrapping*, sua finalidade é garantir que cada árvore seja única, reduzindo a variância do classificador. Posteriormente, são combinados aleatoriamente para que seja criada uma única árvore capaz de conter o máximo de observações possíveis o que gera uma resistência a valores discrepantes e ruídos das bases de dados como é evidenciado na Figura 3 (ABDEL-RAHMAN; AHMED; ISMAIL, 2013; MISRA; LI, 2020).

Segundo Elavarasan e Vincent (2021), determinar o número ideal de árvores que comporão o algoritmo é uma tarefa tediosa. Portanto, uma abordagem alternativa para melhorar o desempenho da floresta aleatória é via reforço. O desempenho do algoritmo aumenta com a adição de mais árvores, mas por sua vez, leva a alta complexidade e sobre ajuste de dados. Isso, por sua vez, degrada a precisão do algoritmo. Portanto, uma quantidade considerável de árvores é necessária para garantir o aprimoramento da precisão preditiva ou classificatória.

A maior limitação da aplicação do *Random Forest* consiste em que seu algoritmo, ao gerar uma grande quantidade de Árvores de Decisão que, não tem a capacidade de capturar a influência de variáveis de baixa importância como o são considerados nos modelos empíricos baseados em estatística clássica. Por esse motivo, o conjunto de dados de treinamento do modelo deve ser selecionado para cobrir o maior intervalo possível para as variáveis e situações preditivas mais críticas.

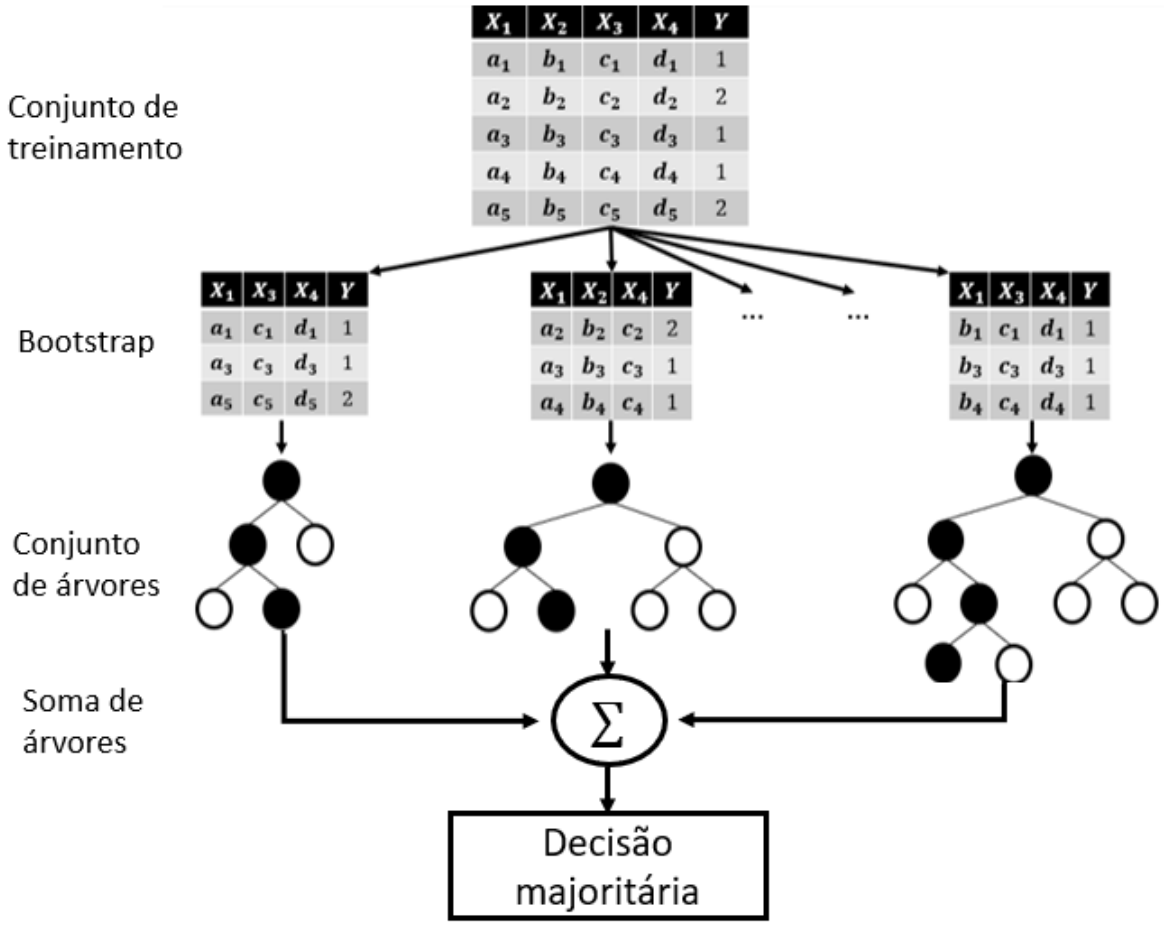


Figura 3 Operação do algoritmo *Random Forest*, adaptado de Misra e Li (2020)

Diferentes publicações foram geradas nos últimos anos, reafirmando a importância deste tipo de projetos e pesquisas para a previsão de produtividade agrícola e modelamento dos elementos químicos que compõem o solo. Charoen-Ung e Mittrapiyanuruk (2018) desenvolveram um modelo na Tailândia usando o *Random Forest* para predizer a produtividade de três variedades diferentes de cana-de-açúcar com um desempenho na previsão de 72%. Os dados que foram modelados estão relacionados com base nas características da parcela (tipo de solo, área de produção, e rendimento produtivo), características da cana (classe e tipo de cana), cuidados e estabelecimento da parcela (tipo de fonte da água, método de irrigação, método de controle de epidemias e aplicação de fertilizantes) e volume de chuvas como único parâmetro meteorológico envolvido no desenvolvimento da cultura.

Kouadio *et al*. (2018) geraram um modelo para analisar propriedades de fertilidade do solo e para gerar uma estimativa precisa da produtividade do café Robusta. Neste estudo, consideraram 17 elementos e nutrientes do solo divididos em dois grupos: minerais e não minerais. Este estudo confirmou a utilidade potencial do acoplamento de algoritmos de

inteligência artificial com modelos de cultivo biofísico em sistemas de apoio à decisão que implementam agricultura de precisão, para melhorar o rendimento em pequenas propriedades de conjuntos de dados de fertilidade do solo cuidadosamente selecionados.

Em termos gerais, os modelos de previsão e classificação da safra podem ser divididos em dinâmicos, empíricos ou estatísticos, dependendo da metodologia adotada para sua formulação. Os modelos dinâmicos gerados com ajuda de técnicas de mineração de dados, têm a capacidade de se adaptar a mudanças que ocorrem nos atributos descritores com flexibilidade, sem afetar o desempenho dos modelos base.

Por este motivo, a pesquisa em curso tem como contribuição a inclusão dos elementos químicos que interagem no solo para estimar a produtividade agrícola em cana-de-açúcar mediante uma divisão de classes, oferecendo uma solução viável para a usinas canavieiras para o planejamento de atividades relacionadas com o uso de maquinaria agrícola, logística interna e armazenamento. Este tipo de modelo é susceptível com as variações da produção agrícola devido a eventos climáticos extremos, como secas e altas temperaturas, que são os principais riscos para os agricultores, governos, seguros agrícolas e mercados, reforçando a necessidade de previsões precisas e oportunas da produção agrícola em um clima incerto.

Especialistas em produção de cana-de-açúcar não têm apenas como único propósito estimar o rendimento com base na produção histórica da região, as características do solo, a modelagem de fatores climáticos e a ocorrência de pragas e doenças. Os pesquisadores também fortalecem continuamente seus bancos de dados para melhorar o desempenho de seus modelos com base nos parâmetros de entrada e procuram novos descritores envolvidos para melhorar seus desempenhos. Everingham *et al.* (2002) definiram as decisões da indústria e usaram a previsão climática para melhorar a tomada de decisões em quatro aspectos desses processos identificados, incluindo: (a) previsão de rendimento e seu efeito nas vendas de açúcar em mercados futuros; (b) uso de previsão climática para tomar decisões sobre gestão da irrigação; (c) determinações de rendimento no início e no final da colheita; e (d) práticas de colheita. Veenadhari (2014) descreve que prever o rendimento do cultivo antes da colheita é uma ferramenta para os responsáveis políticos e agricultores para tomar medidas adequadas e planejar da melhor forma, a logística envolvida no processo para a sua comercialização e armazenamento do produto fresco.

Para superar as limitações associadas à capacidade preditiva associada ao uso de *Random Forest* para aprender classes minoritárias, é possível combinar suas respostas com

modelos de regressão linear múltipla capazes de modelar a produtividade das culturas em condições extremas, onde os dados de observação são escassos (EVERINGHAM *et al*., 2016).

A revisão bibliográfica contribuiu para entender a hipótese e metodologias aplicadas na obtenção de modelos de previsão de safra, constatando-se que, ainda há necessidade de explorar metodologias que incorporem características e composição do solo, e a relação destas com as variáveis climatológicas na produtividade agrícola. O uso da metodologia de Árvores de Decisão em termos de classificação e organização das variáveis envolvidas de acordo com o impacto gerado ao longo de um processo, se apresenta como uma alternativa promissora devido a sua capacidade de abranger eventos de pouca ocorrência, eficiência em termos da rapidez de processamento computacional na formulação do modelo, remoção de ruídos na base de dados, o alto nível de acurácia e seu natureza aleatória na construção de um só árvore minimiza o sobre ajuste do modelo gerado.

5. MATERIAL E MÉTODOS

5.1. Descrição da área de estudo

A experiência de campo da pesquisa foi realizada entre os dias 11 de julho de 2019 e 23 de junho de 2020, na parcela experimental Faculdade de Engenharia Agrícola (FEAGRI) da Universidade Estadual de Campinas (UNICAMP), no município de Campinas-SP. A parcela contava com uma área total de trezentos e vinte metros quadrados (320m^2), compreendida entre as coordenadas geográficas 22°49’12.85”S e 47°03’41.7”W (Figura 4) e uma altitude média de 625 metros acima do nível do mar. A Tabela 1 apresenta as coordenadas UTM correspondentes aos vértices da área de estudo.

De acordo com o Sistema Brasileiro de Classificação de Solos (SiBCS), o solo da parcela experimental corresponde a um Latossolo Vermelho Distroférrico e uma distribuição granulométrica de 57% de argila, 25% de areia e 17,8% de silte. A fim de eliminar plantas daninhas que habitavam a parcela, descompactar o solo e gerar uma camada de solo estável e aerada para fixação das raízes o solo foi submetido a uma passagem de roçadeira e posteriormente realizou-se duas operações com arado de discos de 26”.

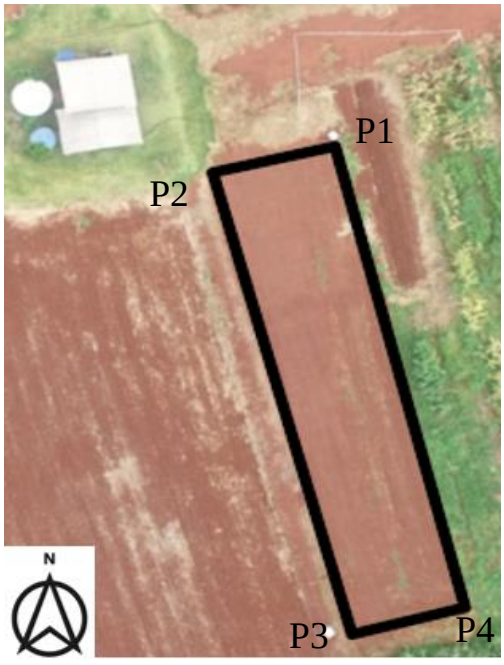


Figura 4 Pontos de contorno da parcela experimental (P1, P2, P3 e P4) usada para o desenvolvimento da pesquisa.

Tabela 1 Coordenadas UTM do contorno da parcela experimental utilizada para o desenvolvimento da pesquisa

Ponto	Norte	Leste	Cota
P1	7474944,62	288400,00	624
P2	7474940,78	288390,76	623
P3	7474903,44	288406,26	625
P4	7474907,27	288415,50	624

O clima é de transição entre Cwa e Cfa qualificado como subtropical de altitude, com inverno seco e um verão úmido e quente, de acordo com a classificação Köppen. Durante o estabelecimento da cultura, desde a semeadura até a colheita, a temperatura média foi de 22,7°C, umidade relativa média de 64,24% e uma precipitação acumulada de 1213,1 mm, sendo que o 25% da precipitação se concentrou entre os meses de janeiro e maio de 2020. Os dados meteorológicos foram fornecidos pela estação meteorológica do Centro de Ensino e Pesquisa em Agricultura (CEPAGRI) localizada na Universidade Estadual de Campinas (UNICAMP) georreferenciada nas coordenadas (22°48´56"S, 47°03´28"W), a qual tem uma frequência de medição e armazenamento das variáveis climáticas com um intervalo de tempo de dez minutos. A Figura 5 apresenta a série histórica destes fatores climáticos medidos no decorrer do período experimental.

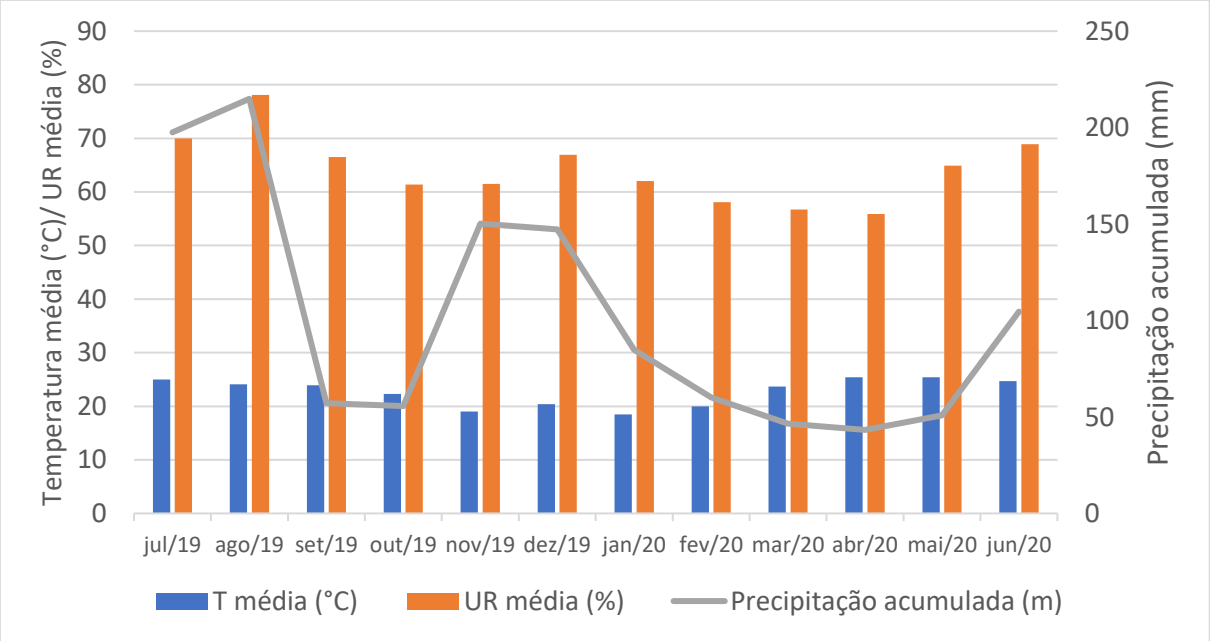


Figura 5 Valores mensais médios de temperatura (°C), umidade relativa (%) e precipitação mensal acumulada (mm) entre julho de 2019 e junho de 2020.

A pesquisa foi executada em parceria com estudos desenvolvidos pelo Prof. Dr. Hugo Enrique Hernandez Figueroa da Faculdade de Engenharia Elétrica e de Computação (FEEC) da UNICAMP. Quem de forma paralela executava o projeto “Radar Transportado por Drone para Agricultura de Precisão na Cana de Açúcar” (Processo No 17194163, 11/2018 – 12/2020 FAPESP).

Em consequência, a fim de garantir o crescimento e cumprir com os requerimentos hídricos das plantas e das pesquisas que estavam em andamento de forma paralela, foram instalados dois sistemas de irrigação expostos na Figura 6.

O primeiro foi um sistema por micro aspersão instalado no início de junho de 2019 antes de semear as mudas pre-brotadas em campo. Estava constituído por duas linhas laterais com 16 emissores cada, micro aspersores não autocompensantes (marca RSB ®) com vazão de 50 L h^{-1} operando na pressão de 200 kPa.

Em outubro de 2019 foi realizada uma troca para um sistema de irrigação por gotejamento superficial. Cada linha de plantio contava com duas linhas de gotejadores, utilizando o tubo gotejador não autocompensante *Drip-tech* (marca *Drip-Plan* ®), os emissores estão espaçados a cada 0,20 m e vazão nominal de 2,25 L h^{-1} a uma pressão de 200 kPa.

(a) Sistema de irrigação por micro aspersão

(b) Sistema de irrigação por gotejamento superficial

Figura 6 Detalhes dos sistemas de irrigação implantados no decorrer da pesquisa.

5.2. Unidade experimental

Na condução do experimento, foram adquiridas 400 mudas de cana-de-açúcar da variedade IACSP97-4039 previamente brotadas pelo centro de cana em Ribeirão Preto (SP) do Instituto Agronômico de Campinas (IAC).

A variedade IACSP97-4039 foi lançada ao setor canavieiro em novembro de 2013. É caracterizada por ser uma cultivar de ciclo curto (até 13 meses), alto teor de sacarose sustentado nas quatro safras que se conseguem obter com esta variedade. Por este motivo a Usina São Matinho, sede Iracemápolis, empresa parceira do projeto, vem destinando diferentes talhões em seus terrenos usando este cultivar.

Outros dos grandes destaques deste material vegetal são a resistência ao déficit hídrico e as mudanças bruscas de temperatura. Não obstante, tem um perfil raro que se caracteriza por ter no início da safra alto potencial de acúmulo de sacarose e por apresentar estabilidade na produtividade agrícola ao longo da safra de acordo com diferentes publicações do IAC.

O estabelecimento das mudas de cana-de-açúcar em campo foi realizado manualmente, seguindo a configuração de plantio da Usina São Martinho onde foram conduzidos trabalhos de validação das pesquisas associadas a este projeto, foram semeadas seis linhas equidistantes a 1,50 m entre elas, com um espaçamento de 0,75 m entre plantas como é representado na Figura 7.

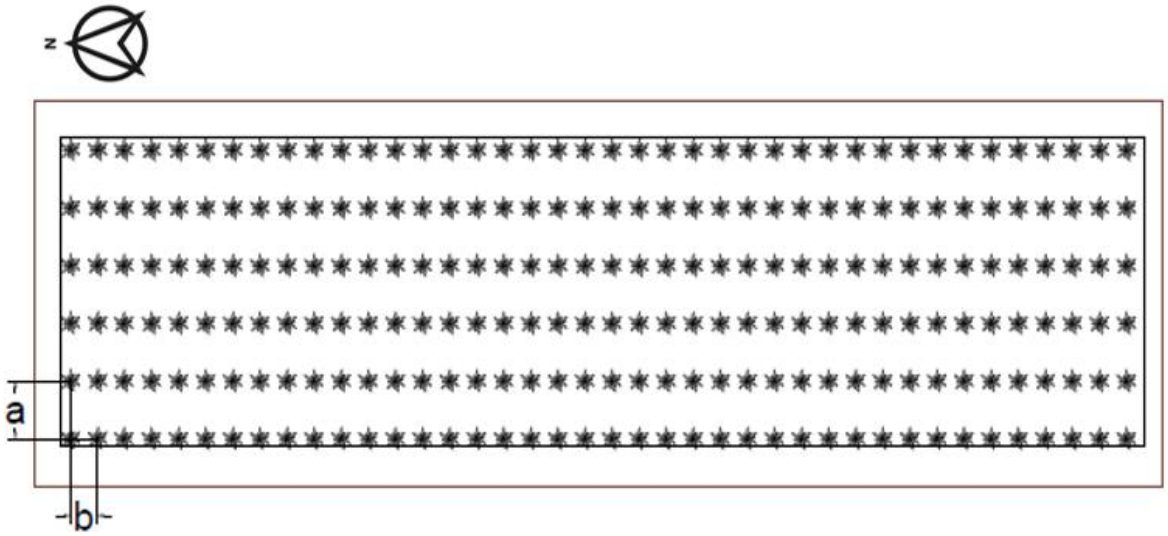


Figura 7 Configuração de plantio: (a) espaçamento entre linhas (1,50 m) e (b) espaçamento entre plantas (0,75 m).

5.3. Amostragem e análise química do solo

A amostragem de solo foi realizada a partir do amadurecimento do material vegetal, em dezembro de 2019 até a colheita em junho de 2020 com uma periodicidade mensal com o propósito de mensurar a variabilidade espacial e temporal das propriedades químicas do solo.

Para estes fins, foi realizada uma amostragem por ponto, em oito posições devidamente georreferenciados (Tabela 2), e em cada um destes foram extraídas as amostras de solo, embalando-as e identificadas. As amostras continham trezentos gramas de solo, retiradas a uma profundidade de 0,20m. Referência típica na amostragem da usina São Martinho.

Tabela 2 Coordenadas UTM dos pontos amostrais para a determinação das propriedades de fertilidade do solo na área de estudo

Ponto	Norte	Leste	Cota
AM1	7474939,84	288399,31	623
AM2	7474930,42	288403,22	623
AM3	7474921,00	288407,13	624
AM4	7474911,57	288411,04	624
AM5	7474937,92	288394,68	624
AM6	7474928,50	288398,59	624
AM7	7474919,08	288402,50	624
AM8	7474909,65	288406,41	624

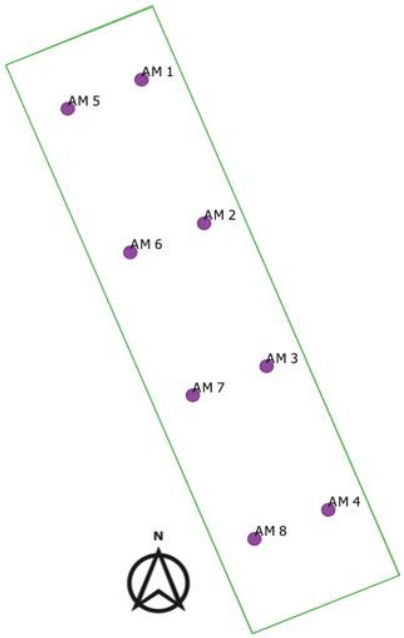


Figura 8 Distribuição dos pontos amostrais para a determinação das propriedades químicas do solo na parcela de experimentação

A determinação do teor de cada micro e macro elemento nas amostras foram analisadas pelo Laboratório de Fertilidade do Solo do Instituto Agronômico de Campinas (IAC) da cidade de Campinas-SP. A Tabela 3 apresentada as propriedades químicas do solo contempladas para a constituição dos modelos.

Tabela 3 Parâmetros de fertilidade do solo usados como preditores e metodologia adotada pelo laboratório para sua determinação

Propriedades químicas		**Método de análise**
Macronutrientes	pH	Cloreto de cálcio ($CaCl_2$)
	Fósforo (P)	Resina
	Potássio (K)	Resina
	Cálcio (Ca)	Resina
	Magnésio (Mg)	Resina
	Matéria orgânica (MO)	Fotométrico
Micronutrientes	Boro (B)	Água quente
	Cobre (Cu)	Ácido dietilenotriaminopentaacético (DTPA)
	Ferro (Fe)	Ácido dietilenotriaminopentaacético (DTPA)
	Manganês (Mn)	Ácido dietilenotriaminopentaacético (DTPA)
	Zinco (Zn)	Ácido dietilenotriaminopentaacético (DTPA)

5.4. Colheita

Em termos do aprendizado de máquinas, o atributo meta desejado foi a produtividade, definida como a quantidade de massa produzida a partir de um material vegetal por unidade de área. Desde o nono mês após plantio, foram coletadas mensalmente amostras dentro do canavial a fim de avaliar o Índice de Maturação (IM).

Este ensaio destrutivo que permitiu determinar o momento adequado para efetuar a colheita com ajuda de um refratômetro analógico foram medidos os graus brix internódio pertencente a última folha cuja bainha destaca-se facilmente e para a base foram adotados os graus brix no quarto internódio acima do nível do solo e posteriormente realizada a divisão entre estes resultados.

$$\mathrm{IM} = \frac{\text{Brix do ponteiro}}{\text{Brix da base do colmo}} \qquad (2)$$

Segundo Cesnik e Mioque (2004), ponto de corte da cana-de-açúcar deve ser realizado quando 0,9<IM<1,0. Já quando a relação entre os graus brix do ponteiro e base do colmo são maiores que 1,0 a cana começa um processo de declínio de maturação e inversão da sacarose.

Em meados de junho de 2020, o ponto ótimo de maturidade foi atingido e procedeu-se manualmente com o corte do canavial. A pesagem da produtividade foi realizada com uma balança Toledo® composta com célula de carga cuja carga máxima suportada é de 500 kg e uma precisão de 0,01 kg.

No momento da colheita da cana madura, a massa correspondente a duas plantas contíguas correspondeu ao valor adotado como a produção de biomassa por metro quadrado como é representado na Figura 9.

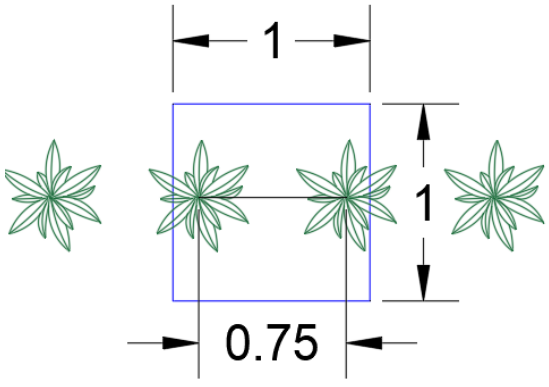


Figura 9 Método adotado para medição da produtividade (cotas em metros)

5.5. Processamento de dados e aplicação dos modelos

O banco de dados composto pelo macro e microelementos do solo (onze elementos) e a produtividade como atributo alvo, passaram por diferentes filtros até chegar à aplicação dos algoritmos de classificação de *Machine Learning.* Uma vez coletado o banco de dados, procedeu-se à instância de operação com a máquina dividida pelas etapas de: a) preparação dos dados, b) análise exploratória, c) modelagem e aprendizagem da máquina e d) avaliação do modelo. O fluxograma exposto na Figura 10 resume algumas das atividades realizadas em cada etapa do processamento dos dados.

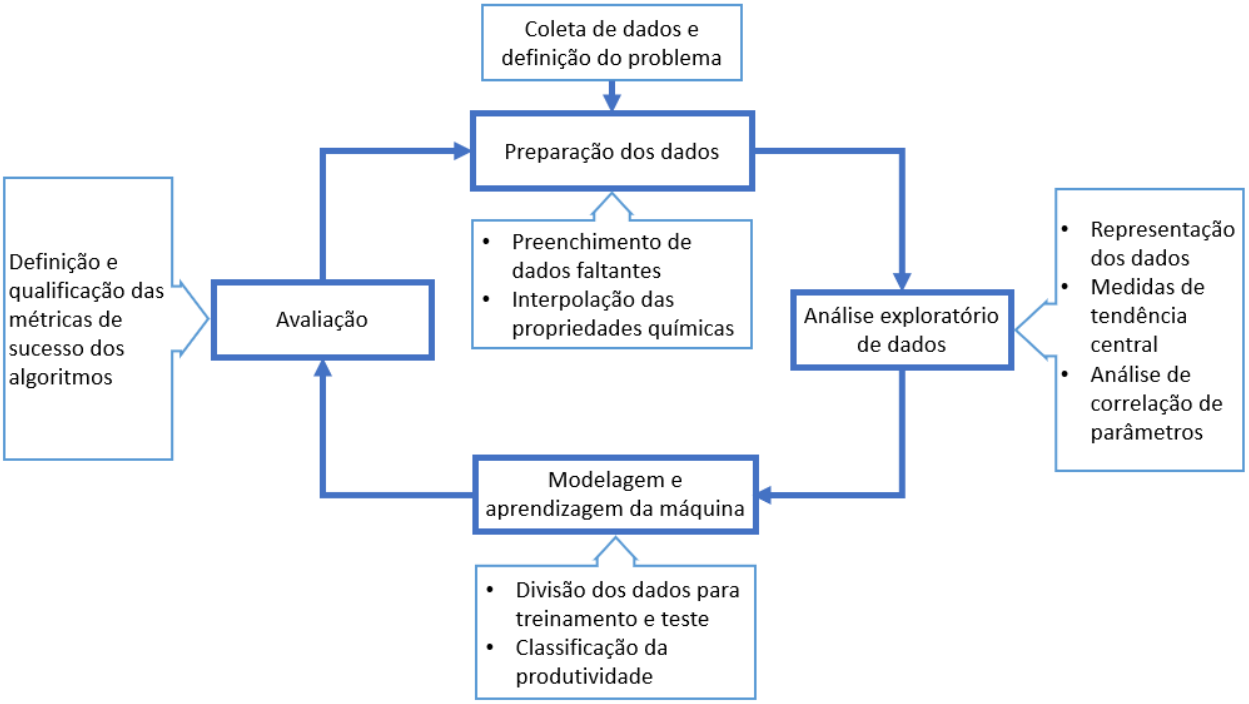


Figura 10 Fluxograma explicativo do procedimento para o processamento dos dados.

5.5.1 Preparação dos dados

A preparação ou pré-processamento dos dados é um passo vital no aprendizado de máquinas porque permite explorar a consistência dos dados e detecção de ruídos existentes no conjunto de dados com a capacidade de gerar problemas na aplicação dos algoritmos. Para o desenvolvimento desta etapa que consistiu em preencher valores faltantes, remoção de dados atípicos também denominados *"outliers"* e modelamento dos atributos químicos do solo e produtividade dentro da área de estudo foi escolhido o software de código aberto QGIS em sua versão 3.10.12.

Durante o seguimento feito mensalmente dentro do canavial, as plantas que não se desenvolveram foram detectadas e georreferenciadas. Os dados provenientes destes locais foram catalogados como dados faltantes, por outra parte, a partir da análise do histograma da produtividade outliers foram detectados.

Uma vez removidos estes valores do conjunto de dados, procedeu-se com a interpolação mediante o método do vizinho mais próximo. Ao realizar este procedimento, obteve-se uma imagem ráster a qual foi vetorizada com o propósito de obter o correspondente valor da coordenada de interesse. O mesmo princípio foi usado para interpolar e consequentemente obter as propriedades químicas do solo dentro da área experimental já que parte da hipótese que norteia o projeto consiste na influência da variabilidade espacial destes fatores com a produtividade agrícola.

5.5.2 Análise exploratória de dados

A linguagem de programação Python executada no editor de código Visual Studio Code, foi escolhida para o Análise Exploratório de Dados ou EDA por suas siglas em inglês (*Exploratory Data Analysis*) principalmente porque conta com uma interface gráfica, bibliotecas que continuamente são atualizadas e fortalecidas, uma linguagem eficiente com ambiente flexível para desenvolvimento de projetos de computação técnica e interpretação de dados.

Na EDA medidas de tendência central foram determinadas, a elaboração de uma análise de correlação de variáveis, realização de provas de hipótese e execução de testes de hipóteses que permitiram remover parâmetros que dentro do aprendizado de máquinas não estavam correlacionados com o atributo alvo, reduzindo a dimensionalidade do problema.

5.5.3 Modelagem e aprendizagem da máquina

Depois de examinar a consistência dos dados e realizar uma remoção de informação desnecessária, os dados foram divididos em dois conjuntos. O primeiro, com 70% do total de dados permitiu o treinamento para cada algoritmo de classificação, enquanto os 30% restante foram reservados para validar o desempenho dos modelos treinados.

A divisão dos dados, treinamento dos modelos, visualização de resultados e posteriormente avaliação das métricas dos modelamentos foi executado com as bibliotecas *Pandas*, *Scikit-learn* e *Matplotlib*.

5.5.4 Avaliação dos modelos

A análise das características de operação do receptor (ROC) por suas siglas em inglês *Receiver Operating Features*, foi o método empregado para avaliar, comparar e selecionar o melhor classificador com base em seu desempenho. Geralmente é confundido termo "análise ROC" o é a mais geral, porque faz referência a um estudo global de classificação, enquanto a "curva ROC" na verdade denota uma curva no "gráfico ROC" (MAJNIK; BOSNIĆ, 2013).

A primeira fase desta análise consistiu na geração da matriz contingência ou matriz de confusão (pois representa a confusão entre as classes) e mostrada na Tabela 4. Existem quatro possíveis saídas da classificação para cada instância:

- **Verdadeiro Positivo (TP):** Corresponde aos dados positivos classificados corretamente pelo algoritmo.
- **Falso Negativo (FN):** O número de dados positivos classificados incorretamente.
- **Verdadeiro Negativo (TN):** quando a instância negativa é classificada como tal.
- **Falso Positivo (FP):** No caso de dados positivos classificação incorretamente.

Tabela 4 Estrutura da matriz de confusão para problemas de classificação binária

		Valor verdadeiro (Confirmado pela análise)	
Valor previsto (predito pelo teste)		Positivos	Negativos
	Positivos	Verdadeiro positivo (TP)	Falso Negativo (FN)
	Negativos	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Portanto, a diagonal da tabela de contingência contém as decisões corretas do classificador, enquanto outros elementos da tabela representam o número de classificações incorretas. A tabela de contingência é a fonte para o cálculo adicional das métricas de sucesso dos algoritmos, incluindo a taxa de verdadeiros positivos a qual também é chamada de sensibilidade ou *recall* (TPR, Eq. (3)), o termo especificidade denota a taxa de verdadeiros negativos (TNR, Eq. (4)), a taxa de falsos positivos (FPR, Eq. (5)) e a taxa de falsos negativos (FNR, Eq. (6)).

$$\text{Sensibilidade} = \text{Taxa de verdadeiros positivos (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{Especificidade} = \text{Taxa de verdadeiros negativos (TNR)} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (4)$$

$$\text{Taxa de falsos positivos (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (5)$$

$$\text{Taxa de falsos negativos (FNR)} = \frac{\text{FN}}{\text{TP} + \text{FN}} \quad (6)$$

Por outra parte, a precisão refere-se à dispersão do conjunto de valores obtidos a partir de medidas repetidas de uma magnitude. Quanto menor for a dispersão, maior é a precisão. A métrica é representada pela razão entre o número de previsões corretas (positivas e negativas) e o número total de previsões (ODED; ROKACH, 2005).

$$\text{Precisão} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (7)$$

Posteriormente, foi construído o gráfico bidimensional ROC considerando a Taxa de Falso Positivo (FPR) no eixo da abscissa contra a sensibilidade (TPR) nas ordenadas como é representado pela Figura 11. Com base nessa projeção, foi determinada a área sob a curva ROC (AUC) por sua sigla em inglês *Area Under the Curve*.

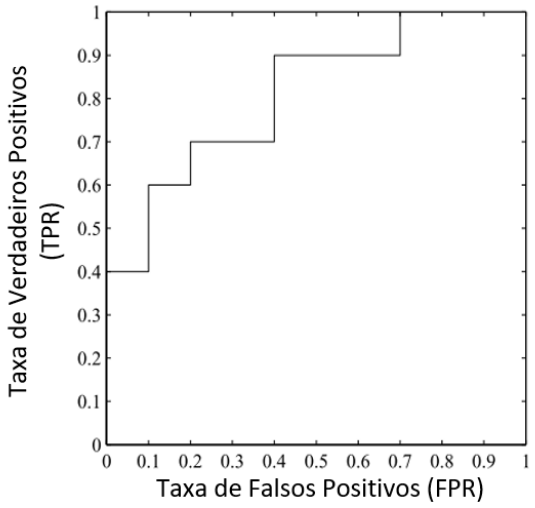


Figura 11 Exemplo de curva ROC para uma amostra

Majnik e Bosnić em 2013, descrevem que a AUC de um classificador é equivalente à probabilidade de que o classificador considere uma instância positiva escolhida aleatoriamente mais alta do que uma instância negativa escolhida aleatoriamente.

A Pontuação F1 é uma medida que resume a precisão e a sensibilidade de um teste em uma única métrica. Normalmente, é útil quando a distribuição das classes é desigual. Pode

ter uma pontuação máxima de 1 (precisão e sensibilidade perfeitas) e mínima de 0. Em geral, é uma medida associada à precisão e robustez do modelo.

$$\text{Pontuação F1} = \frac{2 * (\text{Precisão} * \text{Sensibilidade})}{\text{Precisão} + \text{Sensibilidade}} \tag{8}$$

Para efeitos de referência, avaliação e comparação, adicionalmente foram determinados o *Mean Absolute Error* (MAE, Eq. (9)), *Root Mean Square Error* (RMSE, Eq. (10)) e *Relative Absolute Error* (ERA, Eq. (11)).

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{9}$$

$$\text{Root Mean Square Error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|} \tag{10}$$

$$\text{Relative Absolute Error (RAE)} = \sum_{i=1}^{n}\frac{|y_i - x_i|}{|\overline{y_i} - x_i|} \tag{11}$$

Onde n corresponde ao número da amostra, y_i é a resposta da variável prevista e x_i é a resposta variável observada.

Por fim, foi calculado o Índice de Concordância Kappa (k, Eq. (12)) e representa a proporção de acordos observados além do acaso. Varia geralmente entre 0 e 1 (embora os números negativos sejam possíveis), quanto mais próximo de 1 for seu valor, denota maior confiabilidade, valores próximos ou menores que zero sugerem que a concordância é puramente aleatória.

$$\mathrm{k} = \frac{\mathrm{P_0} - \mathrm{P_e}}{1 - \mathrm{P_e}} \tag{12}$$

Sendo P_0 (Eq. 13) a proporção de acordos observados e P_e corresponde à proporção de acordos esperados na hipótese de independência entre os classificadores, ou seja, acordos por acaso.

$$\mathrm{P_0} = \frac{\mathrm{VP} + \mathrm{TN}}{\mathrm{n}} \tag{13}$$

$$\mathrm{P_e} = \frac{[(\mathrm{VP} + \mathrm{FP}) * (\mathrm{VP} + \mathrm{TN})] * [(\mathrm{FP} + \mathrm{TN}) * (\mathrm{FN} + \mathrm{TN})]}{\mathrm{n}^2} \tag{14}$$

Se tem o intuito que a partir da metodologia proposta, é possível gerar um modelo de classificação de produtividade em cana-de-açúcar aplicando algoritmos de aprendizado de máquinas comparecendo às propriedades químicas presentes no solo para estes fins.

Com o anterior, foram estabelecidas bases para um monitoramento do macro e microelementos do solo e a influência com a produção de material vegetal e desta forma monitorar a cultura espacial e temporalmente, aportando conhecimento na tomada de decisões da lavoura.

6. RESULTADOS E DISCUSSÃO

6.1. Análise de dados experimentais

A análise química é a principal ferramenta para avaliar a fertilidade do solo e consequentemente, a necessidade da intervenção do homem mediante adubação dentro do processo de estabelecimento e crescimento da cultura para uma pronta e efetiva prevenção de doenças pela deficiência de nutrientes que gerem uma queda na produtividade.

Em termos de análise da variabilidade temporal, a primeira camada de solo da parcela experimental onde foi realizado o experimento apresentou comportamento uniforme. Esta informação é suportada principalmente pelo comportamento do desvio padrão nos resultados das análises químicas dentro do ciclo de maturação e corte correspondentes aos oito pontos previamente georreferenciados dentro do contorno delimitado pela cultura resumidos na Tabela 5.

Tabela 5 Resumo da análise estatística descritiva dos fatores de fertilidade do solo avaliados em mg dm-3, MO expresso em g dm-3 e pH em escala própria.

Estatística descritiva	M.O.	pH	P	K	Ca	Mg	B	Cu	Fe	Mn	Zn
Média	25,7	4,53	11,3	12,9	90,8	6,1	0,30	15,1	18,6	22,6	4,1
Desvio padrão	1,7	0,18	2,7	2,5	30,4	1,4	0,006	3,5	3,3	0,4	2,5
Mínimo	22,3	4,2	8,2	9,0	44,0	3,6	0,29	9,0	13,2	21,3	0,8
25%	24,4	4,3	8,7	10,8	62,0	4,8	0,296	11,7	16,1	22,5	1,7
50%	25,8	4,5	10,7	12,9	86,4	6,0	0,3	15,0	18,9	22,7	3,2
75%	27,5	4,7	13,9	15,2	119,0	7,3	0,30	18,1	20,6	22,8	6,3
Máximo	28,2	4,8	16,5	16,8	144,0	8,6	0,31	21,0	26,5	23,3	9,8

Esta premissa não é satisfeita apenas no caso do cálcio, cujo desvio padrão foi de 30 mg dm^{-3}. Este efeito é atribuído ao fato de que este elemento é absorvido principalmente no momento do enraizamento da cultura de cana-de-açúcar, a substituição de cátions de base trocável cálcio na forma Ca^{2+} melhora a estrutura, permeabilidade e infiltração de água no solo e ajuda a planta a suportar o estresse por salinidade (RAHMAN *et al*., 2018).

Segundo as pesquisas de publicadas por Aldon *et al*. (2018) e Demidchik *et al*. (2018), a presença e consumo de cálcio como macronutriente nas plantas é essencial para o aumento da mineralização da matéria orgânica, aumento de atividade de bactérias capazes de decompor nitrogênio em formas assimiláveis pelas plantas e diminuição da toxidez pela

interação dos cátions H^+, Al^{3+} e Mn^{2+}. Como parte de sua função de mensageiro envolvido em vários processos celulares, induzindo ao desenvolvimento da planta e a respostas bióticas e abióticas ao estresse. A variação de Ca^{2+} livre está entre os primeiros eventos após a percepção da planta das mudanças ambientais.

Esta situação ocorreu entre dezembro de 2019 e março de 2020 (Tabela 6), quando a precipitação mensal acumulada sofreu uma redução de um terço, passando de 147 mm para 47 mm respetivamente, a umidade relativa média mensal também teve um declínio de 10,2%; enquanto a temperatura aumentou em 14% passando de 20,4 para 23,7°C. Nesse intervalo temporal houve aumento do consumo de Ca^{2+} disponível no solo.

Tabela 6 Cálcio médio disponível na primeira camada de solo

Ca (g dm^{-3})					
Dez./19	Jan./20	Fev./20	Mar./20	Abr./20	Maio/20
98,5	90	85	86,5	88,5	81,5

Parte do rigoroso monitoramento mensal realizado dentro do canavial com as investigações realizadas em paralelo, consistiu em realizar medições biométricas das plantas para medir sua taxa de crescimento e gerar um modelo de previsão de produtividade a partir de imagens capturadas por três radares embarcados em drone que atuam em distintas frequências (ORÉ *et al*., 2020; LUEBECK *et al*., 2020).

Essa série de testes permitiram detectar que 20 plantas não conseguiram se desenvolver na parcela, gerando um conjunto de dados faltantes dentro da rede de dados. Adicionalmente, no momento da análise estatística detectou-se um grupo de 8 dados anômalos e foram considerados outliers. Havendo verificado a presença de *outliers* detectou-se um grupo de 28 observações, correspondente ao 20% total dos dados foi removido do conjunto de banco de dados e posteriormente modelado pela Ponderação pelo Inverso da Distância (IDW) usando o software Qgis 3.10.12.

As mudanças realizadas sob a base de dados principal permitiram: a) diminuir em 26% o desvio padrão, b) reduzir o intervalo de valores entre a produtividade máxima e mínima da cultura em 36,5%, c) gerar um incremento do rendimento por metro quadrado de 29,14 para 32 kg m^{-2} para a produtividade estimada, em consequência o valor global de massa fresca passou de 4196,2 para 4616,5 kg m^{-2} tendo uma interpretação gráfica na Figura 12.

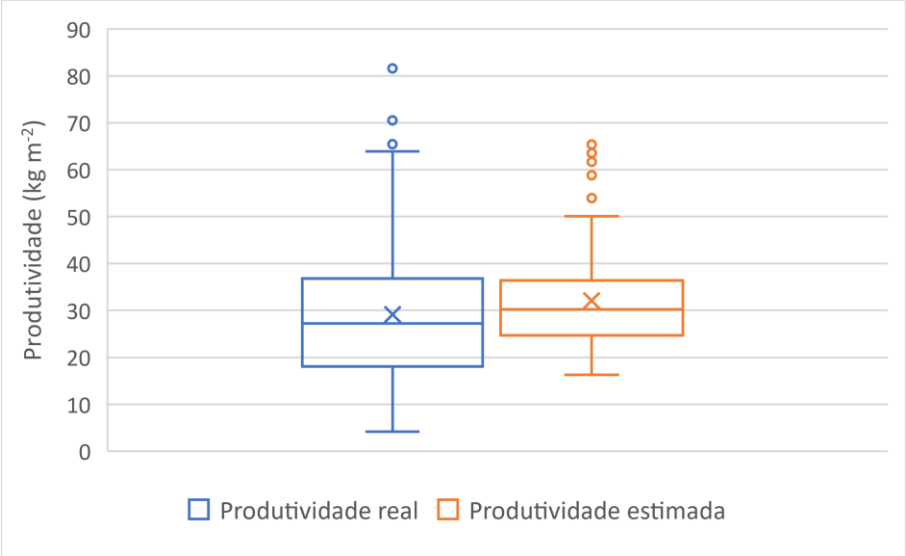
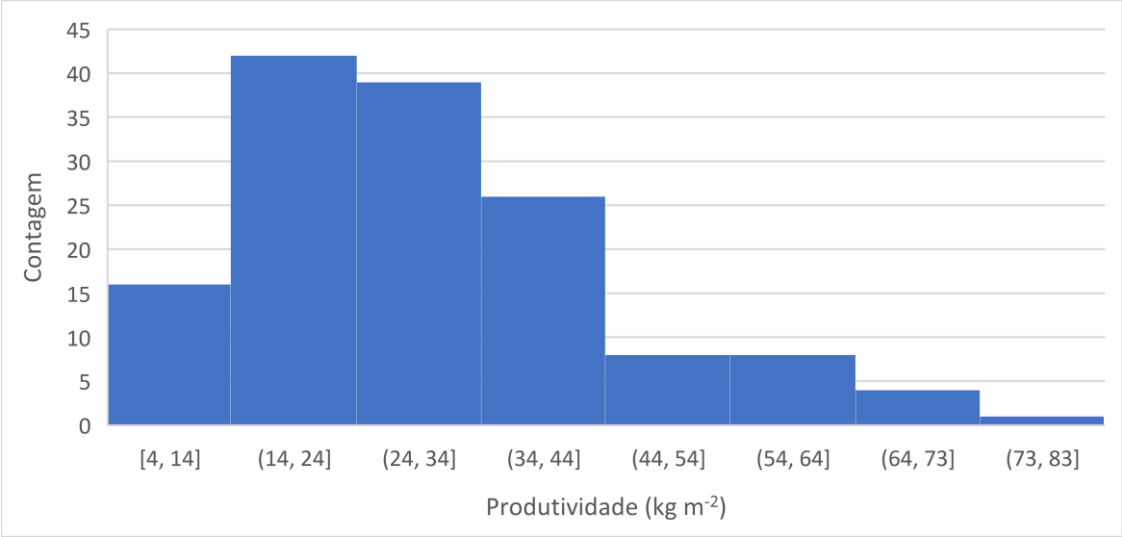
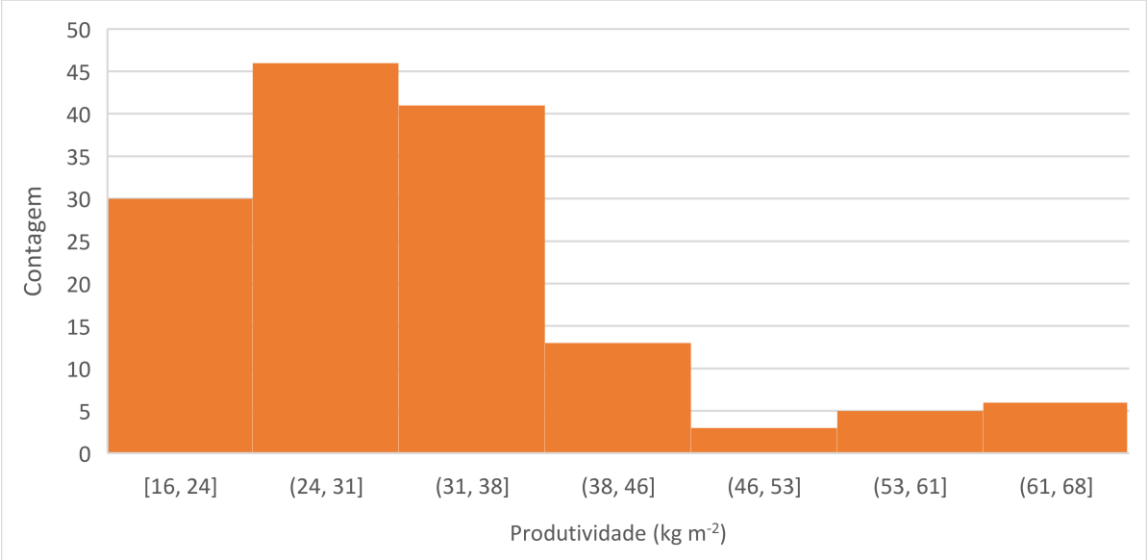


Figura 12 Análise *Boxplot* da produtividade agrícola, diferenciando a real obtida em campo e a estimada através do modelamento IDW

O histograma de frequência (Figura 13) para os casos da produtividade real e estimada exibiu uma distribuição assimétrica positiva, isso indica que a maioria dos dados está acima da média aritmética dos dados. Adicionalmente, os diagramas de frequências permitem discernir a heterogeneidade dos dados da variável dependente.



(a)



(b)

Figura 13 Histograma de frequência para a produtividade real e estimada. a)- Histograma de frequência da produtividade real; b)- Histograma de frequência da produtividade estimada pela ponderação do inverso da distância.

O comportamento assimétrico positivo da produção de matéria fresca é apoiado pelos resultados publicados por (Hammer; Sentelhas e Mariano 2020), através da combinação de 18 usinas de produção e transformação de cana-de-açúcar, em diferentes cidades do estado de São Paulo, com três safras e vários ciclos entre os anos 2011 e 2015.

A análise e interpretação do histograma de frequência em paralelo com as medidas de tendência central, permitiram categorizar a produtividade em três classes: baixa (Eq. (15)), média (Eq. (16)) e alta (Eq. (17)). Em consequência, a faixa correspondente a cada conjunto de dados foi dividida partir da média e desvio padrão da variável dependente.

$$\text{Baixa} \leq \bar{x}_{prod} - \frac{\sigma_{prod}}{2} \tag{15}$$

$$\bar{x}_{prod} - \frac{\sigma_{prod}}{2} \leq \text{Média} < \bar{x} + \frac{\sigma_{prod}}{3} \tag{16}$$

$$\text{Alta} \geq \bar{x}_{prod} + \frac{\sigma_{prod}}{3} \tag{17}$$

O resultado do fracionamento exercido sobre o conjunto de dados é representado na Figura 14. A classe média foi a que mais absorveu registros, seguida pela classe alta e por último a categoria denominada baixa. Porém, o percentual de participação dos três grupos de produtividade é semelhante, com uma diferença máxima de 4% correspondendo a 6 observações do volume total de dados.

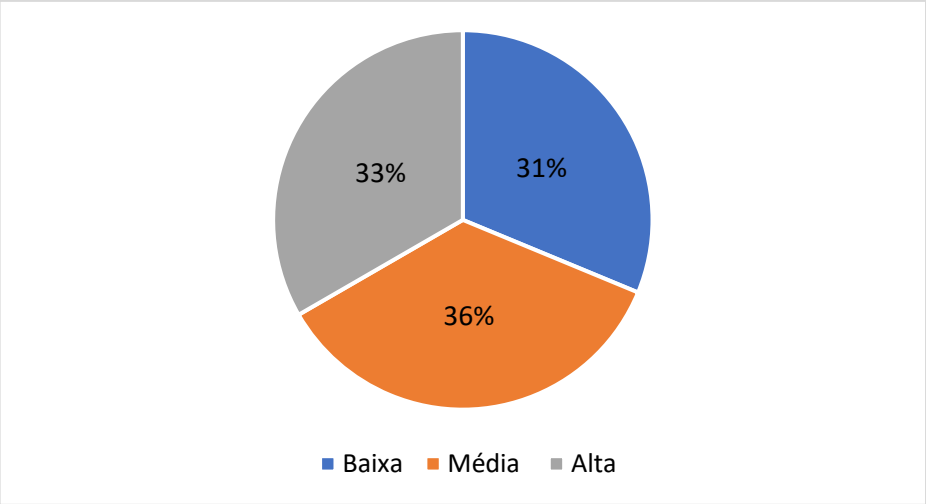


Figura 14 Representação da porcentagem de dados pertencente a cada classe

Uma vez concluída a inspeção quantitativa, pela qual se avaliou a saúde dos dados, a fim de proceder a limpá-los e ser capaz de processá-los em algoritmos de aprendizado de máquinas a partir do preenchimento de dados faltantes, interpolação das propriedades químicas, modelagem da produtividade e o estudo das medidas de tendência central para as variáveis envolvidas para o desenvolvimento da pesquisa foram realizados testes de hipóteses. Nesta etapa, com o intuito de reduzir a dimensionalidade das variáveis de entrada dos modelos, realizou-se a matriz de correlação de dados. Os micronutrientes Boro (B) e Manganês (Mn), que participavam como variáveis de entrada não estão correlacionadas de acordo com o observado na Figura 15, por este motivo foram descartados como partícipes do grupo de elementos químicos do conjunto de treinamento dos modelos.

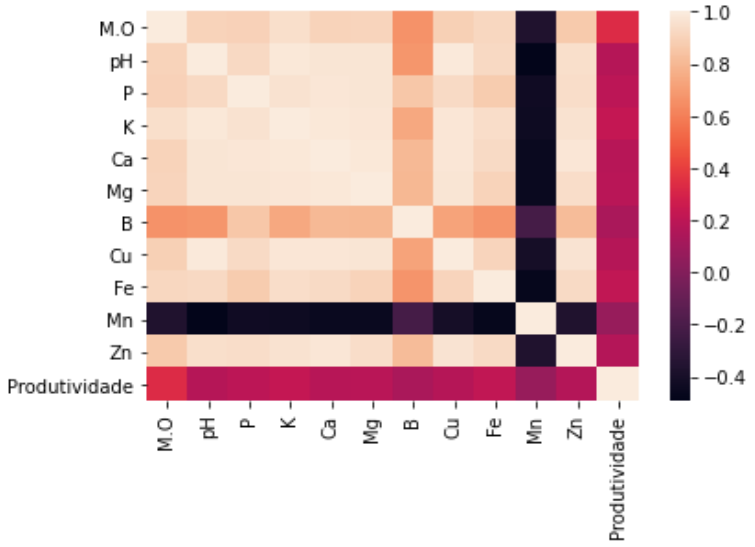


Figura 15 Matriz de correlação das propriedades químicas e a produtividade da cana-de-açúcar

O Boro (B) é um micronutriente aniônico e como todo o ânion sofre os efeitos de uma maior lixiviação que está presente no solo em forma de BO_4^{-2}. A principal função deste microelemento consiste na estabilização das paredes celulares de forma a estar envolvido na estrutura e funcionamento da parede celular e da membrana (YOSHINARI; TAKANO, 2017a); portanto, participa de inúmeras reações de transporte de íons, metabólitos e hormônios (YOSHINARI; TAKANO, 2017b). Este ânion, é absorvido pelas raízes principalmente na forma de ácido bórico, atua como um transportador e facilitador dentro da membrana plasmática. Por esse motivo, a deficiência de Boro afeta a capacidade fotossintética e o transporte de produtos fotossintéticos como a produção de açúcar (BRDAR-JOKANOVIĆ, 2020).

Por sua parte, o Manganês que tem como particularidade ser um cátion que está disponível no solo em três estados de oxidação, apenas a forma bivalente (Mn^{2+}) está disponível para absorção nas plantas mediante a troca com outros cátions bivalentes como o Cálcio, cobalto (Co), Cobre (Cu), Magnésio (Mg) ou Zinco (Zn) (ALEJANDRO *et al*., 2020); pois o Mn^{3+} é instável e o Mn^{4+} forma óxidos e precipitados altamente insolúveis (SCHMIDT, 2019). O Mn desempenha um papel em diversos processos do ciclo de vida de uma planta, como fotossíntese, respiração, eliminação de espécies reativas de oxigênio, formação de clorofila, defesa de patógenos, hidrólise e sinalização hormonal (EISENHUT *et al*., 2018; LI *et al*., 2019).

6.2. Desempenho dos modelos

Após de realizar a detecção e remoção de *outliers*, e pré-processamento dos dados, o conjunto de dados no treinamento e teste foi dividido aleatoriamente com uma proporção de 70:30; a fim de aliviar possibilidades de sobre ajuste ou *overfitting* no desempenho dos modelos de previsão e classificação. O número de registros no treinamento do algoritmo e o conjunto de dados de teste são 100 e 44 observações, respectivamente.

6.2.1. Previsão da produtividade empregando o algoritmo *Random Forest*

A seguir, foi treinado o algoritmo *Random Forest* como preditor sendo a produtividade da cana-de-açúcar a resposta nominal desta avaliação. De acordo com Charoen-Ung e Mittrapiyanuruk em 2018, a previsão de um atributo é considerada como a classificação correta apenas se o grau de rendimento previsto for o mesmo que o grau de rendimento real no conjunto de dados.

O modelo foi criado a partir da iteração de 50 árvores diferentes para evitar a representação de classes minoritárias dentro do modelo. Além disso, não foi definida uma profundidade máxima de exploração, garantindo que os nós fossem expandidos até que todas as folhas estivessem puras ou até que todas as folhas tivessem um mínimo de exemplares necessários para o treinamento interno do nó. Posteriormente, foi representada graficamente a comparação dos valores previstos e reais da produtividade (Figura 16). O valor máximo que o modelo gerado consegue predizer é 47,20 kg m^{-2} enquanto o valor máximo da produtividade com que o algoritmo foi trenado é de 65,42 kg m^{-2} visitando a incapacidade de estimar 15% do intervalo total de dados.

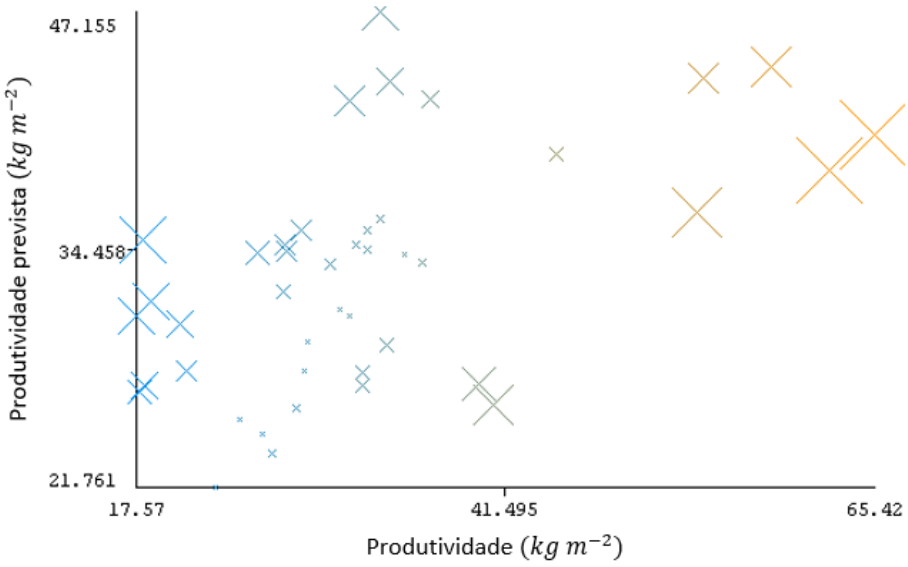


Figura 16 Dispersão da produtividade prevista pelo algoritmo *Random Forest*.

De acordo com o coeficiente de correlação e o erro relativo absoluto, o desempenho deste algoritmo foi insatisfatório e de pouca confiança, a Tabela 7 resume o desempenho estatístico obtido a partir deste algoritmo preditor. Por outra parte a RMSE demonstra o intervalo de valores que pode adotar um ponto dentro dos valores previstos.

O desempenho está longe do obtido na pesquisa de Jeong *et al*. em 2016, que analisou um histórico de produtividade de 30 anos em trigo, milho e batata no Nordeste dos Estados Unidos. De acordo com seus resultados, o modelo *Random Forest* explicou 96% da variância no desempenho com boa concordância entre os valores previstos e observados nos dados de teste. Alcançando um RMSE médio de 1,9 ton. ha^{-1}, isto como resposta de um reforço no algoritmo com um modelo de regressão múltipla para gerar maior robustez na previsão dos valores.

Tabela 7 Desempenho estatístico do preditor *Random Forest*

Coeficiente de correlação	Erro médio absoluto (kg m^{-2})	Raiz do erro quadrático (RMSE)(kg m^{-2})	Erro relativo absoluto
0,51	7,33	9,86	92,43%

Segundo a análise de importância dos elementos de fertilidade do solo dentro do teste preditivo, a matéria orgânica é o fator que está maiormente associado com a produção vegetal (Figura 17). Hipótese amparada pela matriz de correlação realizada na análise exploratória de dados, que a sua vez descreveu a baixa associação do Manganês e Boro com a variável objetivo.

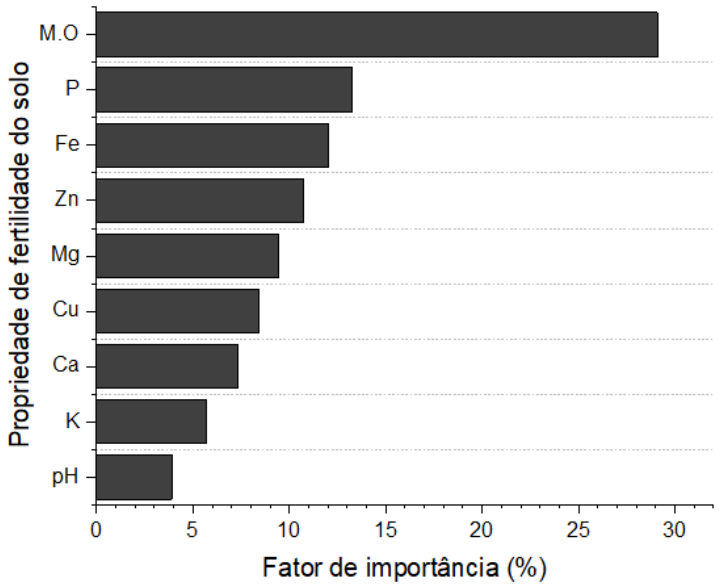


Figura 17 Análise de importância de atributos do modelo preditor *Random Forest*

Segundo Moreno-Barriga *et al*. (2017), a estabilização do teor de matéria orgânica (MO) do solo é necessária para manter o depósito de carbono orgânico no solo, com efeitos positivos na estrutura do solo, retenção de água, aeração, fertilidade, enraizamento, desenvolvimento de fauna e biomassa e diversidade microbiana. Embora Juriga *et al*. em 2018 esclarece que o Ca^{2+} gera uma ponte entre a matéria orgânica e as partículas minerais para a formação de agregados no solo que por sua vez controlam a dinâmica da MO e influenciam diretamente a capacidade do solo de capturar e estabilizar o carbono orgânico.

No âmbito de um estudo agro toxicológico, Zamulina *et al*. (2021) realizou um estudo do efeito da contaminação do solo pela presença de Zinco e Cobre sobre a atividade

bioquímica dos microrganismos que habitam o solo e o efeito tampão da matéria orgânica para mitigar a presença destes metais.

Do ponto de vista ecológico e biológico, o solo é um habitat dinâmico para uma grande variedade de seres vivos. Onde ocorrem interações bióticas para o funcionamento dos ecossistemas, nesta ordem de ideias, a macrofauna e a microfauna são cruciais para o esmagamento e transformação da matéria orgânica. A distribuição espaço-temporal da fauna contribui para a criação de nichos ecológicos heterogêneos e o impacto disso se reflete na fertilidade do solo (SOFO; MININNI; RICCIUTI, 2020; FROUZ, 2018).

Na relação água, solo, planta e meio ambiente a MO também demonstrou ter uma relação estreita na retenção de água no solo (ANKENBAUER; LOHEIDE, 2017). Minasny e Mcbratney (2018) mostraram que um aumento de 1% na matéria orgânica aumenta o conteúdo volumétrico de água no solo em 1,16%.

Finalmente, em resposta à preocupação da comunidade científica com os gases de efeito estufa, a gestão da água doce e a produtividade Zornoza *et al*. (2018), realizou um experimento sobre o envolvimento de matéria orgânica com déficit de irrigação, como resultado obteve-se que sujeitar as plantas ao estresse hídrico reduziu a liberação de CO_2, aumento da MO, deixando claro que isso ocorreu sem afetar o volume de produção.

6.2.2. Desempenho dos algoritmos de classificação

Em seguida, tomando o mesmo conjunto de dados para treinamento e com base na produtividade como variável categórica. Pelo desempenho individual em pesquisas anteriores de acordo com a revisão bibliográfica, os três algoritmos de classificação foram treinados e avaliados: *Naive Bayes*, Árvore de Decisão e *Random Forest*.

O método *Naive Bayes* é capaz de ajustar parâmetros mais rapidamente e usar uma faixa contínua de valores; entretanto, é computacionalmente mais caro do que as Árvores de Decisão porque essa pesquisa é feita individualmente para cada atributo do descritor. A busca bayesiana assume uma distribuição subjacente desconhecida e tenta aproximar a função desconhecida com modelos substitutos, como o processo gaussiano. A otimização bayesiana incorpora crenças anteriores sobre a função subjacente e a atualiza com novas observações (LANGFORD, 2017). Agilizando o ajuste dos parâmetros e garante a busca por uma solução aceitável, desde que seja observado um número suficiente de observações. Em cada iteração, a otimização Bayesiana coleta observações com a maior quantidade de informações e tem como

objetivo equilibrar a exploração (a partir de parâmetros incertos) coletando observações de parâmetros próximos do ótimo.

Por outro lado, as Árvores de Decisão e *Random Forest* são baseados no conceito de *bagging*. O objetivo do *bagging* é reduzir a variância da predição, calculando a média das predições feitas com o conjunto de dados disponível para o treinamento do algoritmo. O *Random Forest* adiciona um novo recurso ao *bagging*, que é escolher aleatoriamente um número aleatório de recursos e construir uma árvore com eles e repetir este procedimento várias vezes e eventualmente, calcular a média de todas as previsões feitas por todas as árvores. Portanto, *Random Forest* aborda os componentes de polarização e variância do erro, tornando-o uma ferramenta mais robusta (SHAHHOSSEINI *et al*., 2021).

As Tabela 8, Tabela 9Tabela 10 expõem os desempenhos de cada algoritmo correspondente aos operadores do receptor (ROC) obtidos com os modelos de aprendizado de máquina utilizados. Supletivamente, a Tabela 11 mostra os resultados da análise estatística de cada modelo. De acordo com o sumário dos resultados, o algoritmo *Random Forest* conseguiu atender melhor desempenho na classificação das etiquetas propostas para o desenvolvimento da pesquisa. Resultado semelhante ao obtido por Tripathi e Maktedar em 2020, na classificação de frutas e hortaliças

No entanto, fazendo um estudo mais aprofundado dentro da taxa de classificação de verdadeiros positivos e verdadeiros negativos abre um debate sobre a escolha do melhor classificador. Levando em consideração que o algoritmo *Naive Bayes* cumpriu um melhor desempenho na determinação de dados positivos corretamente classificados (TP) e na classificação no caso de dados positivos classificados incorretamente (TN) nas classes denominadas “média” e “baixa” acima dos seus concorrentes. A dificuldade em categorizar o conjunto de dados referentes à produtividade alta, reduziu o desempenho global desse algoritmo.

Outro fato notável que suporta a premissa de *Naive Bayes* como o algoritmo que melhor se adaptou aos dados é extraído da área sob a curva ROC (AUC), a qual é interpretável como à probabilidade de que o classificador considere uma instância positiva escolhida aleatoriamente mais alta do que uma instância negativa escolhida aleatoriamente. Nos três rótulos usados para dividir a produtividade teve um melhor despenho manifestando-se como o método mais robusto na categorização da produtividade.

Tabela 8 Resumo das características ROC do algoritmo *Naive Bayes*

Naive Bayes						
Classe	Taxa de verdadeiros positivos	Taxa de verdadeiros negativos	Precisão	Sensibilidade	Pontuação F1	Área ROC
Alta	0,36	0,18	0,40	0,36	0,38	0,70
Média	0,40	0,22	0,62	0,40	0,49	0,70
Baixa	0,83	0,32	0,50	0,83	0,63	0,84
Média ponderada	0,51	0,24	0,53	0,51	0,50	0,74

Tabela 9 Resumo das características ROC da Árvore de Decisão

Árvore de Decisão						
Classe	Taxa de verdadeiros positivos	Taxa de verdadeiros negativos	Precisão	Sensibilidade	Pontuação F1	Área ROC
Alta	1,00	0,78	0,30	1,00	0,47	0,61
Média	0,00	0,00	--	--	--	0,60
Baixa	0,50	0,03	0,86	0,50	0,63	0,73
Média ponderada	0,40	0,21	--	0,40	--	0,64

Tabela 10 Resumo das características ROC do algoritmo *Random Forest*

Random Forest						
Classe	Taxa de verdadeiros positivos	Taxa de verdadeiros negativos	Precisão	Sensibilidade	Pontuação F1	Área ROC
Alta	0,64	0,28	0,44	0,64	0,52	0,70
Média	0,35	0,26	0,54	0,35	0,42	0,52
Baixa	0,75	0,16	0,64	0,75	0,69	0,79
Média ponderada	0,58	0,24	0,54	0,54	0,52	0,64

Tabela 11 Métricas de desempenho dos métodos de classificação

Métrica	*Naive Bayes*	Árvore de Decisão	*Random Forest*
Instâncias corretamente classificadas (%)	51,16	39,53	53,48
Instâncias classificadas incorretamente (%)	48,84	60,46	46,51
Kappa estatística	0,27	0,18	0,30
Erro absoluto médio	0,34	0,4	0,38
Raiz do erro quadrático médio	0,47	0,44	0,47
Erro relativo absoluto	0,77	0,89	0,85

O valor de AUC está na faixa de 0 a 1. Uma vez que qualquer modelo de classificação útil deve estar acima da diagonal de um gráfico ROC, a AUC de tal modelo excede o valor 0,5. É mostrado que a AUC tem algumas características convenientes: um erro padrão que diminui quando a AUC e o número de amostras de teste aumentam; é independente de um limite de decisão; é invariante para as probabilidades da classe anterior; e indica em que grau as classes negativas e positivas são separadas (MAJNIK; BOSNIĆ, 2013).

Da mesma forma, a sensibilidade responsável por avaliar quão bem o modelo prevê a classe positiva quando o resultado real é positivo. Mostrou que nos três modelos utilizados tiveram uma melhor interpretação da faixa de valores correspondentes à produtividade denominada “alta”, enquanto as demais duas classes exibiram comportamento descontínuo, sendo a classe média a mais suscetível ao erro de classificação apesar de ser o rótulo com a quantidade de dados no momento da categorização da produtividade. Essa situação se reflete nos resultados obtidos na pontuação F1, a partir dos quais se pode argumentar que os modelos dificilmente são confiáveis na classificação de baixa produtividade.

Por sua vez, a Árvore de Decisão foi o método que menos sucesso teve na classificação de rótulos, segundo a Tabela 9 o algoritmo só tem a capacidade de distinguir entre classes baixa e alta, com a particularidade de acertar a totalidade dos dados rotulados como “alto” demonstrando um problema de sobre ajuste do modelo. Não obstante, de acordo com o resultado gráfico da Figura 18, o nó raiz corresponde à matéria orgânica e uma regra chega aos seus nós terminais para classificar a produtividade como baixa ou média quando realmente não tem possibilidade alguma de acertar observações correspondentes a esse último grupo de valores.

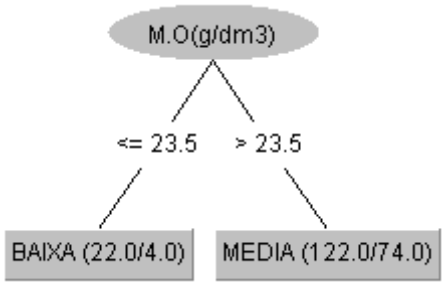


Figura 18 Classificação da produtividade pelo método de Árvores de Decisão

No entanto, os resultados obtidos no índice de concordância Kappa são insatisfatórios para os três métodos utilizados para resolver o problema. O máximo atingido foi de apenas 0,30 para o uso de *Random Forest* seguido de *Naive Bayes* com índice Kappa de 0,27.

Estas situações têm fundamento na estrutura interna de cada algoritmo. No caso das Árvores de Decisão e *Random Forest* são métodos que automatizam a de interações e efeitos principalmente não lineares, adicionalmente os preditores ou classificadores se dividem para acoplar uma floresta mais sólida. Nesse processo se geram os problemas de *overfitting* dos modelos, sendo o *Random Forest* o algoritmo menos propenso que a Árvore de Decisão simples (JIANG; GRADUS; ROSELLINI, 2020).

A pesquisa de Jeong *et al*. de 2016, demonstrou limitações com o *Random Forest* para o modelamento da produtividade em grande escala de trigo e milho. De acordo com seus resultados, o algoritmo *Random Forest* é menos intuitivo que os métodos de regressão tradicionais porque cada uma das árvores que compõem a floresta não pode ser descrita mecanicamente. Além disso, com *Random Forest*, apenas os valores incluídos no conjunto de dados de treinamento são usados para dividir as árvores de regressão para posteriormente agrupar as previsões fora das faixas de treinamento, o que gera um baixo desempenho na previsão de cenários futuros com os quais não houve a oportunidade de ser treinado.

Apesar dessas limitações, o *Random Forest* tem mostrado que é capaz de melhorar o desempenho da previsão de variáveis categóricas desequilibradas devido à sua capacidade de dividir árvores de mesmo tamanho em paralelo no processo de aprendizagem (ZHOU *et al*., 2020). Conforme mostrado anteriormente, o desequilíbrio nos dados é refletido nas métricas de precisão e sensibilidade. Por esse motivo, o *Random Forest* tem um desempenho melhor do que a Árvore de Decisão em termos de previsão de eventos positivos corretos.

De acordo com Deepa e Ganesan (2018), a maior desvantagem da Árvore de Decisão passa porque é um algoritmo instável, um pequeno desvio nos dados leva à criação de diferentes árvores de decisão e conjuntos de dados com poucas observações não funcionam bem. Além disso, é sensível a atributos ruidosos e irrelevantes, por vezes gerando uma ausência da função global das variáveis e, portanto, perda de representatividade nas variáveis classificatórias.

Por outro lado, o algoritmo *Naive Bayes* é adequado para resolver problemas com mais de dois classes quando as variáveis de entrada são categóricas; seu desempenho é geralmente superior ao de outros classificadores se sua suposição de independência das características for verdadeira com um volume menor de dados (FENG *et al*., 2018). No entanto, a principal limitação do uso desse algoritmo é que ele assume implicitamente que todos os atributos são independentes entre si, limitando sua aplicabilidade em casos da vida real, pois é difícil obter um conjunto de preditores completamente independentes entre si (POROIKOV *et al*., 2019). Ainda, se uma variável categórica tiver um rótulo no conjunto de dados de teste que não foi observado durante o treinamento do modelo, ela atribuirá um valor zero e não executará a classificação. Para resolver este problema, os dados de treinamento devem ser limpos e verificados ou suavizados com as funções Laplace (DRURY *et al*., 2017).

O gráfico de pares ou *pairplot* da Figura 19 é útil para descrever a incerteza gerada nos modelos, tendo em vista que a produtividade apresenta uma eminente heterogeneidade e dispersão nos dados.

Por sua vez exibe o potássio (K^+) como um atributo capaz de descrever o comportamento do resto de macro e microelementos presentes no solo. A importância deste macro elemento toma valor na biomassa já que constitui em média 10% da matéria seca das plantas (SUSTR; SOUKUP; TYLOVA, 2019). Em termos globais, o K^+ tem protagonismo na fotossíntese pelo papel que desempenha na regulação do movimento dos estomas, transporte de açúcares e produtividade (BLATT, 2016), o desenvolvimento e crescimento radicular dependem da interação do K^+ com o pH no processo de síntese de proteínas e ativação enzimática (FU *et al*., 2020).

Dentro do citoplasma, é o cátion monovalente mais importante para a ativação de várias enzimas do metabolismo celular, a interação com Ca^{2+} ocorre na célula pela ação do citoesqueleto e para manter a planta turgente (KLINSAWANG *et al*., 2018), adicionalmente está relacionado à resistência aos solos salinos, principalmente com a presença de NaCl, o

acúmulo de Na^+ gera perda de K^+ (SUSTR; SOUKUP; TYLOVA, 2019) e por fim, a deficiência de K^+ limita a absorção de N, diminuindo a eficiência na produtividade (HOU *et al*., 2019).

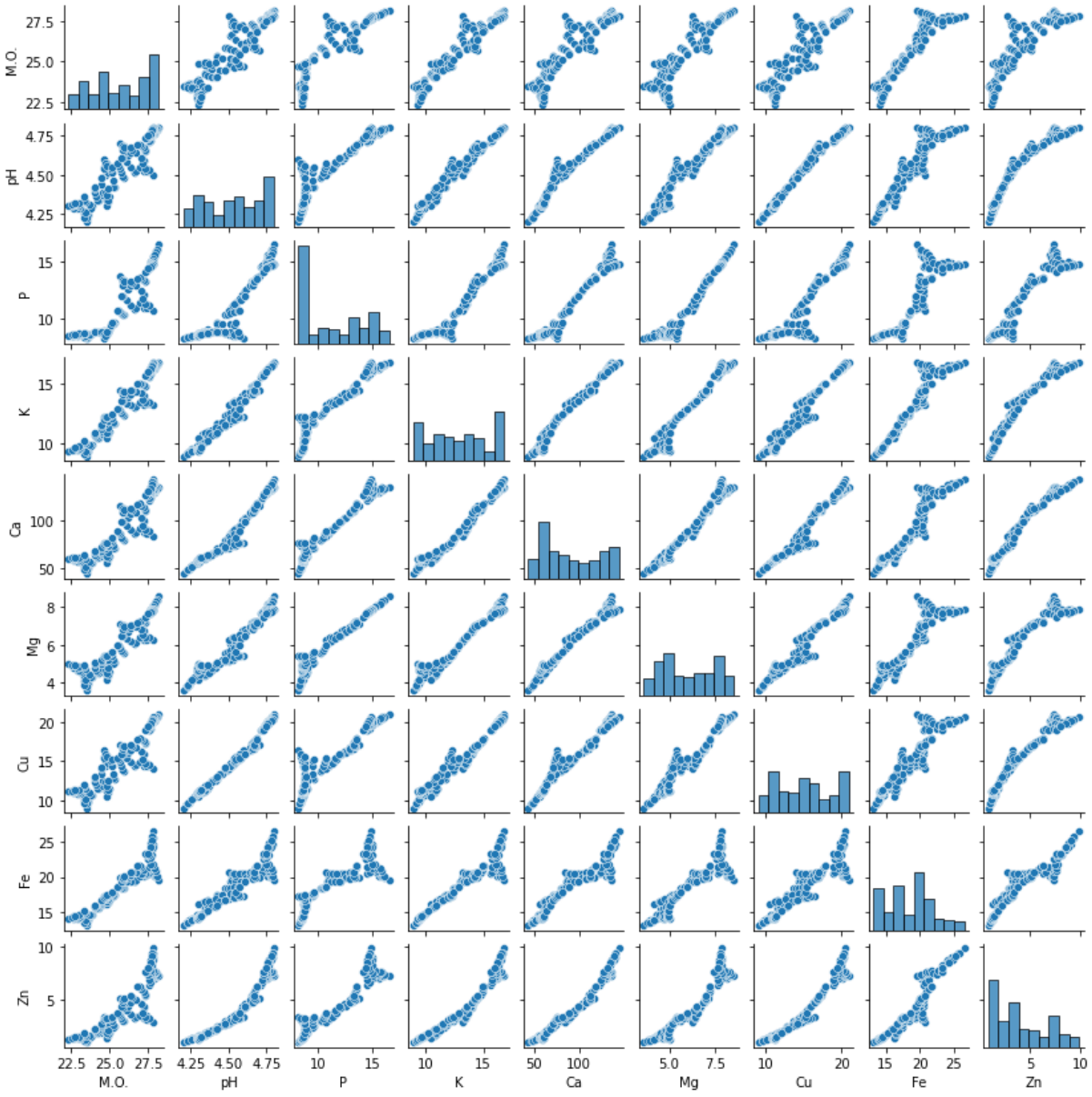


Figura 19 Interação do potássio (K^+) com as propriedades de fertilidade do solo analisados.

Para estudos futuros recomenda-se replicar o experimento em diferentes locais para além aumentar o volume de dados, incluir os fatores climáticos dentro das séries temporais e desta forma antecipar eventos futuros de produção permitindo criar modelos robustos e confiáveis.

Existe uma estreita relação entre o comportamento climático e a produtividade agrícola, principalmente na cana-de-açúcar, essa interação tem sido explorada em diferentes escalas. Hammer; Sentelhas e Mariano em 2020 realizaram um modelo agrometeorológico para

predizer a produtividade da cana-de-açúcar em 18 usinas canavieiras distribuídas no estado de São Paulo e corroborou a participação da variabilidade climática em distintas variedades de cana-de-açúcar.

Por sua vez, Linnenluecke *et al*. em 2020 realizou um estudo sobre o impacto das mudanças climáticas na indústria canadense na Austrália a partir de um conjunto de dados meteorológicos dos anos 1964-2012. Sua pesquisa corroborou que a produção de cana-de-açúcar foi afetada pela concentração de carbono atmosférico a partir de 1995 e pelo aumento contínuo das temperaturas máximas registradas anualmente. Abrindo a porta para a inclusão de práticas de desenvolvimento adicionais como a inclusão da irrigação, o gerenciamento da análise de dados climáticos na colheita, a melhoria dos nutrientes aplicados no solo e a importância do melhoramento genético das plantas para mitigar os efeitos do aquecimento global.

Por outro lado, há pesquisas que se concentram em estudar a mitigação do impacto das mudanças climáticas na produtividade agrícola e na geração de ecossistemas sustentáveis (PAREEK; DHANKHER; FOYER, 2020; ZHAO; LI, 2015), as implicações e importância do papel que a precipitação desempenha nos sistemas de produtivos (KNAPP; CIAIS; SMITH, 2017) e na China em particular estão usando modelos de previsão do clima para a formulação de projetos para a tomada de decisões que gerem impactos positivos na produtividade da cana-de-açúcar (RUAN *et al*., 2018).

Outro fator a ser abordado em pesquisas futuras é a inclusão de sistemas híbridos de aprendizado de máquina. Este é um campo que tem sido explorado na agricultura de precisão justamente pelo alto desempenho e confiabilidade exigidos na área (CHLINGARYAN; SUKKARIEH; WHELAN, 2018; UPRETI *et al*., 2019; YAHATA *et al*., 2017).

Em particular, *Random Forest* é o método que melhor se fusiona com outros algoritmos. Como é o caso de fortalecer seu desempenho com regressão linear múltipla para resolver os problemas típicos de escalabilidade neste sistema estatístico tradicional (RASSOUL *et al*., 2021). Ele também atua como um auxiliar em outro algoritmo de classificação como os *Support Vector Machine* (SVM) (DEMIDOVA; KLYUEVA; PYLKIN, 2019) e inclusive fortalece sistemas de redes neurais profundas (KONG; YU, 2018).

Especificamente na classificação multiclasse como o problema que foi desenvolvido nesta pesquisa, Yan *et al*. (2020) propõem um sistema híbrido para Árvores de

Decisão e *Random Forest* denominado "um contra todos". Esta estratégia é caracterizada por produzir m problemas binários com m número de problemas cada um para cada classe. Cada problema é resolvido por meio de um classificador binário, que se encarrega de identificar uma classe de todas as outras. Consequentemente, todos os dados de treinamento são usados na fase de aprendizagem, considerando os exemplos em uma classe como positivos e os exemplos nas demais classes como negativas. Para posteriormente na fase de validação, todos os exemplos desconhecidos são enviados para todos os classificadores.

6.3. Considerações suplementares

A previsão de produtividade agrícola é um dos problemas mais desafiadores e motivo de pesquisa dentro da agricultura, até agora vem sendo propostos, validados e exploradas diferentes metodologias para atender este caso comum dentro diferentes atores implicados como produtores, processadores de matéria fresca, transportadores e planejadores de logística.

Este problema requer o uso de vários conjuntos de dados, dado que o rendimento agrícola é consequência de diferentes fatores, como clima, fertilidade do solo, uso de defensivos agrícolas, interação biológica e a intervenção do homem. Este pensamento indica que a produtividade e, por conseguinte a previsão da safra não é uma tarefa trivial, realmente consiste na análise de várias etapas complexas que atuam com diferente nível de protagonismo ao longo do ciclo de crescimento das plantas.

Neste sentido, o aprendizado de máquinas vem tomando um papel importante para geração de modelos confiáveis para estimar a produção agrícola e as respostas das plantas a diferentes estímulos durante seu ciclo vegetativo, numa época em que a tecnologia interveio com a presença de sensores, radares e instrumentos que armazenam e processam diferentes grandezas para robustecer bases de dados históricas que treinam e melhoram continuamente os algoritmos preditores.

Este desenvolvimento traz consigo o problema de valores anômalos e ausentes em uma rede de dados transversalmente, uma vez que um dos pontos mais sensíveis dentro da mineração de dados é a maneira como eles são preenchidos e substituídos corretamente para evitar maiores danos, para depois selecionar o algoritmo que melhor se adapta a um banco de dados.

Transversalmente, o histórico do tratamento da parcela experimental teve um impacto negativo no desenvolvimento do projeto, uma vez que a heterogeneidade com que os nutrientes disponíveis no solo foram distribuídos, por sua vez, impactou o crescimento das plantas. Como resultado desta situação, os algoritmos avaliados na pesquisa não atenderam às expectativas que nortearam a hipótese preliminar do trabalho.

7. CONCLUSÕES

Para a classificação das três faixas de produtividade (baixa, média e alta), o *Random Forest* teve um maior índice de instâncias corretamente classificadas se comparado com os algoritmos *Naive Bayes* e Árvores de Decisão.

O *Naive Bayes* apresentou maior índice de sucesso na distinção entre as classes "média" e "baixa" que os algoritmos *Random Forest* e Árvores de Decisão. Esses resultados indicam grande potencial de uso na agricultura para a descrição de modelos de classificação da produtividade agrícola.

Devido à heterogeneidade da produtividade de cana-de-açúcar obtida em campo, o desempenho do algoritmo *Random Forest* teve maior êxito como classificador que como preditor da variável dependente.

A matéria orgânica mostrou maior relação com a produção de biomassa da cana-de-açúcar que outros atributos, segundo os resultados com *Random Forest* e árvores de decisão.

O índice médio de correlação do potássio (K^+) com os demais atributos químicos de solo determinado foi de 0,94; nas amostras da área de estudo, indicando o potencial na geração de modelos empíricos que estimam as propriedades de fertilidade de solos agrícolas.

Para estudo futuro, sugere-se aumentar a quantidade de amostragens de solo, incluir fatores climáticos dentro das séries temporais e incluir sistemas híbridos de aprendizado de máquina.

9. BIBLIOGRAFIA

(CONAB), C. N. de A. **Acompanhamento da Safra Brasileira (Cana-de-açúcar)**. Brasília.

(CONAB), C. N. de A. Análise mensal cana-de-açúcar: Outubro-Novembro de 2020. [s. l.], n. Mmc, 2020. a.

(CONAB), C. N. de A. **Acompanhamento da safra brasileira de cana-de-açúcar para dezembro de 2020**. Brasília.

ABDEL-RAHMAN, E. M.; AHMED, F. B.; ISMAIL, R. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. **International Journal of Remote Sensing**, [s. l.], v. 34, n. 2, p. 712–728, 2013. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/01431161.2012.713142>

AHMED, K.; SHAHIDI, T. R.; ALAM, S. I.; MOMEN, S. Rice Leaf Disease Detection Using Machine Learning Techniques. **2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)**, [s. l.], n. December, p. 1–5, 2019.

ALDON, D.; MBENGUE, M.; MAZARS, C.; GALAUD, J. P. Calcium signalling in plant biotic interactions. **International Journal of Molecular Sciences**, [s. l.], v. 19, n. 3, p. 1–19, 2018.

ALEJANDRO, S.; HÖLLER, S.; MEIER, B.; PEITER, E. Manganese in Plants : From Acquisition to Subcellular Allocation. **Frontiers in Plant Science**, [s. l.], v. 11, n. March, p. 1–23, 2020.

ANKENBAUER, K. J.; LOHEIDE, S. P. The effects of soil organic matter on soil water retention and plant water use in a meadow of the Sierra Nevada, CA. **Hydrological Processes**, [s. l.], v. 31, n. 4, p. 891–901, 2017.

BAŞTANLAR, Y.; OZUYSAL, M. **Introduction to Machine Learning Second Edition**. [s.l: s.n.]. v. 1107 Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24272434>

BASTIAANSSEN, W. G. M.; ALI, S. A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan. **Agriculture, Ecosystems**

and Environment, [s. l.], v. 94, n. 3, p. 321–340, 2003.

BEZERRA, J. D. C.; FERREIRA, G. D. G.; OLIVEIRA, M. W. De; CAMPOS, J. M. de S.; ANDRADE, A. P. De; NASCIMENTO JÚNIOR, J. R. S. Do. Sugar Cane: Genetic Improvement and Forage Purposes. **Nucleus Animalium**, [s. l.], v. 10, n. 2, p. 131–147, 2018.

BLATT, M. R. Plant Physiology : Redefining the Enigma of Metabolism in Stomatal Movement. **Current Biology**, [s. l.], v. 26, n. 3, p. R107–R109, 2016. Disponível em: <http://dx.doi.org/10.1016/j.cub.2015.12.025>

BRDAR-JOKANOVIĆ, M. Boron toxicity and deficiency in agricultural plants. **International Journal of Molecular Sciences**, [s. l.], v. 21, n. 4, 2020.

BREIMAN, L. Random Forests. **Machine Learning**, [s. l.], v. 45, n. 1, p. 5–32, 2001.

BRINKMAN, M. L. J.; DA CUNHA, M. P.; HEIJNEN, S.; WICKE, B.; GUILHOTO, J. J. M.; WALTER, A.; FAAIJ, A. P. C.; VAN DER HILST, F. Interregional assessment of socio-economic effects of sugarcane ethanol production in Brazil. **Renewable and Sustainable Energy Reviews**, [s. l.], v. 88, n. September 2017, p. 347–362, 2018. Disponível em: <https://doi.org/10.1016/j.rser.2018.02.014>

CARVALHO, L.; BUENO, R.; CARVALHO, M.; FAVORETO, A.; GODOY, A. Cana-de-açúcar e álcool combustível: histórico, sustentabilidade e segurança energética. **Biosfera**, [s. l.], v. 9, p. 530–543, 2013.

CESNIK, R.; MIOQUE, J. **Melhoramento na cana-de-açúcar**. Embrapa In ed. Brasíli: Embrapa Meio Ambiente, 2004.

CHANDANAPALLI, S. B.; REDDY, E. S.; LAKSHMI, D. R. DFTDT: distributed functional tangent decision tree for aqua status prediction in wireless sensor networks. **International Journal of Machine Learning and Cybernetics**, [s. l.], v. 9, n. 9, p. 1419–1434, 2018.

CHAROEN-UNG, P.; MITTRAPIYANURUK, P. Sugarcane Yield Grade Prediction using Random Forest and Gradient Boosting Tree Techniques. **Proceeding of 2018 15th International Joint Conference on Computer Science and Software Engineering, JCSSE 2018**, [s. l.], p. 1–6, 2018.

CHLINGARYAN, A.; SUKKARIEH, S.; WHELAN, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. **Computers and Electronics in Agriculture**, [s. l.], v. 151, n. November 2017, p. 61–69, 2018. Disponível em: <https://doi.org/10.1016/j.compag.2018.05.012>

CNA. **PIB do agronegócio avança novamente em outubro**, 2021. Disponível em: <https://www.cnabrasil.org.br/boletins/pib-do-agronegocio-cresce-em-outubro>

CONAB. Acompanhamento de safra brasileira cana-de-açúcar - safra 2019/2020. **Companhia Nacional de Abastecimento**, [s. l.], v. 6, p. 58, 2019.

CONŢIU, Ş.; GROZA, A. Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning. **Expert Systems with Applications**, [s. l.], v. 64, p. 269–286, 2016.

DEEPA, N.; GANESAN, K. Multi-class classification using hybrid soft decision model for agriculture crop selection. **Neural Computing and Applications**, [s. l.], v. 30, n. 4, p. 1025–1038, 2018.

DEMIDCHIK, V.; SHABALA, S.; ISAYENKOV, S.; CUIN, T. A.; POTTOSIN, I. Calcium transport across plant membranes: mechanisms and functions. **New Phytologist**, [s. l.], v. 220, n. 1, p. 49–69, 2018.

DEMIDOVA, L. A.; KLYUEVA, I. A.; PYLKIN, A. N. Hybrid approach to improving the results of the SVM classification using the random forest algorithm. **Procedia Computer Science**, [s. l.], v. 150, p. 455–461, 2019. Disponível em: <https://doi.org/10.1016/j.procs.2019.02.077>

DOORENBOS, J.; KASSAM, A. H. **Efectos del agua sobre el rendimiento de los cultivos**. Roma: Food and Agriculture Organization of the United Nations, 1979. v. 33

DRURY, B.; VALVERDE-REBAZA, J.; MOURA, M. F.; DE ANDRADE LOPES, A. A survey of the applications of Bayesian networks in agriculture. **Engineering Applications of Artificial Intelligence**, [s. l.], v. 65, n. January, p. 29–42, 2017. Disponível em: <http://dx.doi.org/10.1016/j.engappai.2017.07.003>

EISENHUT, M.; HOECKER, N.; SCHMIDT, S. B.; BASGARAN, R. M.; FLACHBART, S.; JAHNS, P.; ESER, T.; GEIMER, S.; HUSTED, S.; WEBER, A. P. M.; LEISTER, D.; SCHNEIDER, A. The Plastid Envelope CHLOROPLAST MANGANESE

TRANSPORTER1 Is Essential for Manganese Homeostasis in Arabidopsis. **Molecular Plant**, [s. l.], v. 11, n. 7, p. 955–969, 2018. Disponível em: <10.1016/j.molp.2018.04.008>

ELAVARASAN, D.; VINCENT, P. M. D. R. A reinforced random forest model for enhanced crop yield prediction by integrating agrarian parameters. **Journal of Ambient Intelligence and Humanized Computing**, [s. l.], n. 0123456789, 2021. Disponível em: <https://doi.org/10.1007/s12652-020-02752-y>

EVERINGHAM, Y. L.; MUCHOW, R. C.; STONE, R. C.; INMAN-BAMBER, N. G.; SINGELS, A.; BEZUIDENHOUT, C. N. Enhanced risk management and decision-making capability across the sugarcane industry value chain based on seasonal climate forecasts. **Agricultural Systems**, [s. l.], v. 74, n. 3, p. 459–477, 2002.

EVERINGHAM, Y.; SEXTON, J.; SKOCAJ, D.; INMAN-BAMBER, G. Accurate prediction of sugarcane yield using a random forest algorithm. **Agronomy for Sustainable Development**, [s. l.], v. 36, n. 2, 2016. Disponível em: <http://dx.doi.org/10.1007/s13593-016-0364-z>

FAO. **FAOSTAT**. 2017. Disponível em: <http://www.fao.org/faostat/en/#data/QC>. Acesso em: 10 jun. 2019.

FENG, X.; LI, S.; YUAN, C.; ZENG, P.; SUN, Y. Prediction of Slope Stability using Naive Bayes Classifier. **KSCE Journal of Civil Engineering**, [s. l.], v. 22, n. 3, p. 941–950, 2018.

FRANCO, H. C. J.; PIMENTA, M. T. B.; CARVALHO, J. L. N.; GRAZIANO MAGALHÃES, P. S.; ROSSELL, C. E. V.; BRAUNBECK, O. A.; VITTI, A. C.; ẄLLN, O. T.; NETO, J. R. Assessment of sugarcane trash for agronomic and energy purposes in Brazil. **Scientia Agricola**, [s. l.], v. 70, n. 5, p. 305–312, 2013.

FRIEDRICH, B.; RUDORFF, T.; TEIXEIRA BATISTA, G. Yield estimation of sugarcane Based on Agrometeorological-Spectral Models. **Elsevier Science Publishing**, [s. l.], v. 4257, n. 90, p. 183–192, 1990.

FROUZ, J. Geoderma E ff ects of soil macro- and mesofauna on litter decomposition and soil organic matter stabilization. **Geoderma**, [s. l.], v. 332, n. October 2016, p. 161–172, 2018. Disponível em: <https://doi.org/10.1016/j.geoderma.2017.08.039>

FU, X.; ZHAO, C.; MA, S.; TIAN, H.; DONG, D.; LI, G. L. Determining available

potassium in soil by laser-induced breakdown spectroscopy combined with cation exchange membrane adsorption. **Journal of Analytical Atomic Spectrometry**, [s. l.], v. 35, n. 11, p. 2697–2703, 2020.

GILIO, L.; DIAS DE MORAES, M. A. Sugarcane industry's socioeconomic impact in São Paulo, Brazil: A spatial dynamic panel approach. **Energy Economics**, [s. l.], v. 58, p. 27–37, 2016.

GIRARDI, E. P. Agronegócio sucroenergético e desenvolvimento no Brasil. **Revista franco-brasilera de geografia**, [s. l.], n. 40, p. 20, 2019. Disponível em: <http://journals.openedition.org/confins/19517>

GONZÁLEZ PEREA, R.; CAMACHO POYATO, E.; MONTESINOS, P.; RODRÍGUEZ DÍAZ, J. A. Prediction of irrigation event occurrence at farm level using optimal decision trees. **Computers and Electronics in Agriculture**, [s. l.], v. 157, n. December 2018, p. 173–180, 2019. Disponível em: <https://doi.org/10.1016/j.compag.2018.12.043>

GRANIK, M.; MESYURA, V. Fake news detection using naive Bayes classifier. **2017 IEEE 1st Ukraine Conference on Electrical and Computer Engineering, UKRCON 2017 - Proceedings**, [s. l.], p. 900–903, 2017.

HABIB, M. T.; MAJUMDER, A.; JAKARIA, A. Z. M.; AKTER, M.; UDDIN, M. S.; AHMED, F. Machine vision based papaya disease recognition. **Journal of King Saud University - Computer and Information Sciences**, [s. l.], v. 32, n. 3, p. 300–309, 2020. Disponível em: <https://doi.org/10.1016/j.jksuci.2018.06.006>

HAMMER, R. G.; SENTELHAS, P. C.; MARIANO, J. C. Q. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. **Sugar Tech**, [s. l.], v. 22, n. 5, p. 216–225, 2020. Disponível em: <https://doi.org/10.1007/s12355-019-00776-z>

HELENA, M. C. de M. N.; MARCIA, X. P.; ADROALDO, D. R.; RICHARD, A. R. iacute guez P.; CICERO, U. N.; MARCOS, V. L. Irrigation depths in sugarcane crop with drip irrigation system. **African Journal of Agricultural Research**, [s. l.], v. 11, n. 27, p. 2423–2432, 2016.

HOU, W.; XUE, X.; LI, X.; RIZWAN, M.; YAN, J.; REN, T.; CONG, R.; LU, J. Field Crops Research Interactive effects of nitrogen and potassium on : Grain yield , nitrogen uptake and nitrogen use efficiency of rice in low potassium fertility soil in China. **Field Crops**

Research, [s. l.], v. 236, n. October 2018, p. 14–23, 2019. Disponível em: <https://doi.org/10.1016/j.fcr.2019.03.006>

JEONG, J. H.; RESOP, J. P.; MUELLER, N. D.; FLEISHER, D. H.; YUN, K.; BUTLER, E. E.; TIMLIN, D. J.; SHIM, K. M.; GERBER, J. S.; REDDY, V. R.; KIM, S. H. Random forests for global and regional crop yield predictions. **PLoS ONE**, [s. l.], v. 11, n. 6, p. 1–15, 2016.

JIANG, T.; GRADUS, J. L.; ROSELLINI, A. J. Supervised Machine Learning: A Brief Primer. **Behavior Therapy**, [s. l.], v. 51, n. 5, p. 675–687, 2020. Disponível em: <https://doi.org/10.1016/j.beth.2020.05.002>

JIN, C.; WANG, Z.; YANG, Z.; JORDAN, M. I. Provably efficient reinforcement learning with linear function approximation. **arXiv**, [s. l.], v. 125, p. 1–7, 2019.

JURIGA, M.; ŠIMANSKÝ, V.; HORÁK, J.; KONDRLOVÁ, E.; IGAZ, D.; POLLÁKOVÁ, N.; BUCHKINA, N.; BALASHOV, E. The effect of different rates of biochar and biochar in combination with N fertilizer on the parameters of soil organic matter and soil structure. **Journal of Ecological Engineering**, [s. l.], v. 19, n. 6, p. 153–161, 2018.

KEZZY DE MORAIS, L.; CURSI, D. E.; MESSIAS, J.; SANTOS, D.; SAMPAIO, M.; MAXWELL, T.; CÂMARA, M.; DE ALBUQUERQUE, P.; GERALDO, S.; HERMANN, V. B.; HOFFMANN, P.; GIACOMINI, R.; ANTÔNIO, C.; FERNANDES, R.; GAZAFFI, J. R. **Melhoramento Genético da Cana-de-Açúcar Tabuleiros Costeiros Ministério da Agricultura, Pecuária e Abastecimento Melhoramento Genético da Cana-de-Açúcar Documentos 200**. Primeira e ed. Aracaju. Disponível em: <www.embrapa.com.br>

KLINSAWANG, S.; SUMRANWANICH, T.; WANNARO, A.; SAENGWILAI, P. Effects of root hair length on potassium acquisition in rice (Oryza sativa L.). **Applied Ecology and Environmental Research**, [s. l.], v. 16, n. 2, p. 1609–1620, 2018.

KNAPP, A. K.; CIAIS, P.; SMITH, M. D. Reconciling inconsistencies in precipitation–productivity relationships: implications for climate change. **New Phytologist**, [s. l.], v. 214, n. 1, p. 41–47, 2017.

KONG, Y.; YU, T. A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification. **Scientific Reports**, [s. l.], v. 8, n. 1, p. 1–9, 2018.

KOUADIO, L.; DEO, R. C.; BYRAREDDY, V.; ADAMOWSKI, J. F.; MUSHTAQ, S.; PHUONG NGUYEN, V. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. **Computers and Electronics in Agriculture**, [s. l.], v. 155, n. April, p. 324–338, 2018. Disponível em: <https://doi.org/10.1016/j.compag.2018.10.014>

LANGFORD, J. **Encyclopedia of Machine Learning and Data Mining**. [s.l: s.n.].

LI, J.; JIA, Y.; DONG, R.; HUANG, R.; LIU, P.; LI, X. Advances in the Mechanisms of Plant Tolerance to Manganese Toxicity. **International Journal of Molecular Sciences**, [s. l.], v. 20, n. 20, p. 15, 2019.

LIAKOS, K. G.; BUSATO, P.; MOSHOU, D.; PEARSON, S.; BOCHTIS, D. Machine learning in agriculture: A review. **Sensors (Switzerland)**, [s. l.], v. 18, n. 8, p. 1–29, 2018.

LINNENLUECKE, M. K.; ZHOU, C.; SMITH, T.; THOMPSON, N.; NUCIFORA, N. The impact of climate change on the Australian sugarcane industry. **Journal of Cleaner Production**, [s. l.], v. 246, p. 118974, 2020. Disponível em: <https://doi.org/10.1016/j.jclepro.2019.118974>

LUEBECK, D.; WIMMER, C.; MOREIRA, L. F.; ALCÂNTARA, M.; ORÉ, G.; GÓES, J. A.; OLIVEIRA, L. P.; TERUEL, B.; BINS, L. S.; GABRIELLI, L. H.; HERNANDEZ-FIGUEROA, H. E. **Drone-borne differential SAR interferometry**, 2020. Disponível em: <https://www.mdpi.com/2072-4292/12/5/778>

MAJNIK, M.; BOSNIĆ, Z. ROC analysis of classifiers in machine learning: A survey. **Intelligent Data Analysis**, [s. l.], v. 17, n. 3, p. 531–558, 2013. Disponível em: <https://www.medra.org/servlet/aliasResolver?alias=iospress&doi=10.3233/IDA-130592>

MARCARI, M. A.; ROLIM, G. de S.; APARECIDO, L. E. de O. Agrometeorological models for forecasting yield and quality of sugarcane. **Australian Journal of Crop Science**, [s. l.], v. 9, n. 11, p. 1049–1056, 2015.

MARSLAND, S. **Machine Learning An Algorithmic Perspective**. Second Edi ed. [s.l.] : Chapman & Hall Book, 2014. Disponível em: <https://b-ok.cc/book/2543746/ef80cb>

MEDAR, R. A.; RAJPUROHIT, V. S. A survey on Data Mining Techniques for

Crop Yield Prediction. **International Journal of Advance Research in Computer Science and Management Studies**, [s. l.], v. 2, n. 9, p. 59–64, 2014.

MIHAI, H.; FLORIN, S. Biomass prediction model in maize based on satellite images. **AIP Conference Proceedings**, [s. l.], v. 1738, 2016.

MINASNY, B.; MCBRATNEY, A. B. Limited effect of organic matter on soil available water capacity. **European Journal of Soil Science**, [s. l.], v. 69, n. 1, p. 39–47, 2018.

MINISTÉRIO DA ECONOMIA. **Valor Bruto da Produção Agropecuária**. Brasília.

MISRA, S.; LI, H. **Noninvasive fracture characterization based on the classification of sonic wave travel times**. First edit ed. [s.l.] : Elsevier Inc., 2020. a. Disponível em: <http://dx.doi.org/10.1016/B978-0-12-817736-5.00009-0>

MISRA, S.; LI, H. Random Forest Noninvasive fracture characterization based on the classification of sonic wave travel times Bone cancer detection using machine learning techniques. **Machine Learning for Subsurface Characterization**, [s. l.], 2020. b.

MORENO-BARRIGA, F.; DÍAZ, V.; ACOSTA, J. A.; MUÑOZ, M. Á.; FAZ, Á.; ZORNOZA, R. Organic matter dynamics, soil aggregation and microbial biomass and activity in Technosols created with metalliferous mine residues, biochar and marble waste. **Geoderma**, [s. l.], v. 301, n. October 2016, p. 19–29, 2017.

NASTESKI, V. An overview of the supervised machine learning methods. **Horizons.B**, [s. l.], v. 4, n. December 2017, p. 51–62, 2017.

ODED, M.; ROKACH, L. **The Data Mining and Knowledge Discovery Handbook**. First Edit ed. New York, USA: Springer USA, 2005.

OLIVEIRA, R. A.; SANTOS, R. S.; RIBEIRO, A.; ZOLNIER, S.; BARBOSA, M. H. P. Estimativa da produtividade da cana-de-açúcar para as principais regiões produtoras de Minas Gerais usando-se o método ZAE 1 Yield estimate of sugarcane in main producing regions of Minas Gerais using the AEZ method. **Revista Brasileira de Engenharia Agrícola e Ambiental**, [s. l.], v. 16, n. 31, p. 549–557, 2012. Disponível em: <http://www.agriambi.com.br/revista/v16n05/v16n05a11.pdf>

ORÉ, G.; ALCÂNTARA, M. S.; GÓES, J. A.; OLIVEIRA, L. P.; YEPES, J.;

TERUEL, B.; CASTRO, V.; BINS, L. S.; CASTRO, F.; LUEBECK, D.; MOREIRA, L. F.; GABRIELLI, L. H.; HERNANDEZ-FIGUEROA, H. E. Crop growth monitoring with drone-borne DInSAR. **Remote Sensing**, [s. l.], v. 12, n. 4, p. 615, 2020. Disponível em: <www.mdpi.com/journal/remotesensing>

PADALALU, P.; MAHAJAN, S.; DABIR, K.; MITKAR, S.; JAVALE, D. Smart Water Dripping System for Agriculture / Farming. [s. l.], p. 659–662, 2017.

PAREEK, A.; DHANKHER, O. P.; FOYER, C. H. Mitigating the impact of climate change on plant productivity and ecosystem sustainability. **Journal of Experimental Botany**, [s. l.], v. 71, n. 2, p. 451–456, 2020.

POROIKOV, V. V.; FILIMONOV, D. A.; GLORIOZOVA, T. A.; LAGUNIN, A. A.; DRUZHILOVSKIY, D. S.; RUDIK, A. V.; STOLBOV, L. A.; DMITRIEV, A. V.; TARASOVA, O. A.; IVANOV, S. M.; POGODIN, P. V. Computer-aided prediction of biological activity spectra for organic compounds: the possibilities and limitations. **Russian Chemical Bulletin**, [s. l.], v. 68, n. 12, p. 2143–2154, 2019.

RAHMAN, M. A.; LEE, S. H.; JI, H. C.; KABIR, A. H.; JONES, C. S.; LEE, K. W. Importance of mineral nutrition for mitigating aluminum toxicity in plants on acidic soils: Current status and opportunities. **International Journal of Molecular Sciences**, [s. l.], v. 19, n. 10, 2018.

RASSOUL, A.; MOHAMMAD, Z.; MAHMOUDI, R.; SHABANI, A. Investigating of the climatic parameters effectiveness rate on barley water requirement using the random forest algorithm , Bayesian multiple linear regression and cross - correlation function. **Paddy and Water Environment**, [s. l.], v. 19, n. 1, p. 137–148, 2021. Disponível em: <https://doi.org/10.1007/s10333-020-00825-4>

RODOLFO JUNIOR, F. **CARACTERIZAÇÃO DE VARIEDADES DE CANA-SOCA SOB DIFERENTES REGIMES HÍDRICOS NO CERRADO**. 2015. UNIVERSIDADE DE BRASÍLIA, [s. l.], 2015.

RUAN, H.; FENG, P.; WANG, B.; XING, H.; O'LEARY, G. J.; HUANG, Z.; GUO, H.; LIU, D. L. Future climate change projects positive impacts on sugarcane productivity in southern China. **European Journal of Agronomy**, [s. l.], v. 96, n. April, p. 108–119, 2018. Disponível em: <https://doi.org/10.1016/j.eja.2018.03.007>

SCARPARI, M. S. **Modelos para a previsão da produtividade da cana-de-açúcar (Saccharum spp.) através de parâmetros climáticos.** 2002. Universidade de São Paulo, Piracicaba, 2002. Disponível em: <http://www.teses.usp.br/teses/disponiveis/11/11136/tde-17122002-165859/>

SCHMIDT, S. B. The Biochemical Properties of Manganese in Plants. **Plants**, [s. l.], v. 8, n. 381, p. 15, 2019.

SETIADI, T.; NOVIYANTO, F.; HARDIANTO, H.; TARMUJI, A.; FADLIL, A.; WIBOWO, M. Implementation Of Naïve Bayes Method In Food Crops Planting Recommendation. **INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH**, [s. l.], v. 9, n. 02, 2020.

SHAHHOSSEINI, M.; HU, G.; HUBER, I.; ARCHONTOULIS, S. V. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. **Scientific Reports**, [s. l.], v. 11, n. 1, p. 1606, 2021. Disponível em: <https://doi.org/10.1038/s41598-020-80820-1>

SILVA, A. L. Da; CASTAÑEDA-AYARZA, J. A. Macro-environment analysis of the corn ethanol fuel development in Brazil. **Renewable and Sustainable Energy Reviews**, [s. l.], v. 135, n. September 2019, 2021.

SINGH, V.; SARWAR, A.; SHARMA, V. Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. **International Journal of Advanced Research in Computer Science**, [s. l.], v. 8, n. 5, 2017. Disponível em: <www.ijarcs.info>

SOFO, A.; MININNI, A. N.; RICCIUTI, P. Soil macrofauna: A key factor for increasing soil fertility and promoting sustainable soil use in fruit orchard agrosystems. **Agronomy**, [s. l.], v. 10, n. 4, 2020.

SOUSA, E.; MACEDO, I. **Etanol e bioeletricidade "A cana-de-açúcar no futuro da matriz energética"**. Saõ Paulo: União da Indústria de cana-de-açúcar, 2009. v. 1

SUSTR, M.; SOUKUP, A.; TYLOVA, E. Potassium in Root Growth and Development. **Plants**, [s. l.], v. 8, n. 10, p. 435, 2019. Disponível em: <https://www.mdpi.com/2223-7747/8/10/435>

TERDAL, S. Evaluation of Machine Learning Algorithms for Crop Yield Prediction. **International Journal of Engineering and Advanced Technology**, [s. l.], v. 8, n.

6, p. 4082–4086, 2019.

THEODORO, A. D. **Expansão da cana-de-açúcar no Brasil: ocupação da cobertura vegetal do cerrado**. 2011. Faculdade de Tecnologia de Araçatuba, [s. l.], 2011.

TRIPATHI, M. K.; MAKTEDAR, D. D. A role of computer vision in fruits and vegetables among various horticulture products of agriculture fields: A survey. **Information Processing in Agriculture**, [s. l.], v. 7, n. 2, p. 183–203, 2020. Disponível em: <https://doi.org/10.1016/j.inpa.2019.07.003>

UPRETI, D.; HUANG, W.; KONG, W.; PASCUCCI, S.; PIGNATTI, S.; ZHOU, X.; YE, H.; CASA, R. A comparison of hybrid machine learning algorithms for the retrieval of wheat biophysical variables from sentinel-2. **Remote Sensing**, [s. l.], v. 11, n. 5, 2019.

USAMA, M.; QADIR, J.; RAZA, A.; ARIF, H.; YAU, K. L. A.; ELKHATIB, Y.; HUSSAIN, A.; AL-FUQAHA, A. Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges. **IEEE Access**, [s. l.], v. 7, p. 65579–65615, 2019.

VEENADHARI, S.; MISRA, B.; SINGH, C. D. Machine learning approach for forecasting crop yield based on climatic parameters. **2014 International Conference on Computer Communication and Informatics: Ushering in Technologies of Tomorrow, Today, ICCCI 2014**, [s. l.], p. 1–5, 2014.

VEERNAGOUDA GANGANAGOWDER, N.; KAMATH, P. Intelligent classification models for food products basis on morphological, colour and texture features. **Agroindustry and Food Science**, [s. l.], v. 66, n. 4, p. 486–494, 2017. Disponível em: <https://revistas.unal.edu.co/index.php/acta_agronomica/article/view/60049>

WITTEN, I.; FRANK, E.; HALL, M. **Data mining**. Morgan Kau ed. Waikato: Elsevier, 2011. v. 54 Disponível em: <http://www.cs.waikato.ac.nz/~ml/weka/book.html%5Cnhttp://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569>

YAHATA, S.; ONISHI, T.; YAMAGUCHI, K.; OZAWA, S.; KITAZONO, J.; OHKAWA, T.; YOSHIDA, T.; MURAKAMI, N.; TSUJI, H. A hybrid machine learning approach to automatic plant phenotyping for smart agriculture. **Proceedings of the International Joint Conference on Neural Networks**, [s. l.], v. 2017- May, p. 1787–1793,

2017.

YAN, J.; ZHANG, Z.; LIN, K.; YANG, F.; LUO, X. A hybrid scheme-based one-vs-all decision trees for multi-class classification tasks. **Knowledge-Based Systems**, [s. l.], v. 198, p. 105922, 2020. Disponível em: <https://doi.org/10.1016/j.knosys.2020.105922>

YOSHINARI, A.; TAKANO, J. Insights into the mechanisms underlying boron homeostasis in plants. **Frontiers in Plant Science**, [s. l.], v. 8, n. November, p. 1–8, 2017. a.

YOSHINARI, A.; TAKANO, J. Insights into the Mechanisms Underlying Boron Homeostasis in Plants. **Frontiers in Plant Science**, [s. l.], v. 8, n. 3, p. 201–227, 2017. b. Disponível em: <http://journal.frontiersin.org/article/10.3389/fpls.2017.01951/full>

ZAMULINA, I. V.; GOROVTSOV, A. V.; MINKINA, T. M.; MANDZHIEVA, S. S.; BAUER, T. V.; BURACHEVSKAYA, M. V. The influence of long-term Zn and Cu contamination in Spolic Technosols on water-soluble organic matter and soil biological activity. **Ecotoxicology and Environmental Safety**, [s. l.], v. 208, n. June 2020, p. 111471, 2021. Disponível em: <https://doi.org/10.1016/j.ecoenv.2020.111471>

ZHANG, H. The optimality of Naive Bayes. **Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004**, [s. l.], v. 2, p. 562–567, 2004.

ZHAO, D.; LI, Y.-R. Climate Change and Sugarcane Production: Potential Impact and Mitigation Strategies. **International Journal of Agronomy**, [s. l.], v. 2015, p. 1–10, 2015.

ZHOU, X.; LU, P.; ZHENG, Z.; TOLLIVER, D.; KERAMATI, A. Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree. **Reliability Engineering and System Safety**, [s. l.], v. 200, n. January, p. 106931, 2020. Disponível em: <https://doi.org/10.1016/j.ress.2020.106931>

ZORNOZA, R.; ACOSTA, J. A.; GABARRÓN, M.; GÓMEZ-GARRIDO, M.; SÁNCHEZ-NAVARRO, V.; TERRERO, A.; MARTÍNEZ-MARTÍNEZ, S.; FAZ, Á.; PÉREZ-PASTOR, A. Greenhouse gas emissions and soil organic matter dynamics in woody crop orchards with different irrigation regimes. **Science of the Total Environment**, [s. l.], v. 644, p. 1429–1438, 2018. Disponível em: <https://doi.org/10.1016/j.scitotenv.2018.06.398>

APÊNDICE

Matrizes de confusão obtidas a partir da classificação da produtividade nas classes alta, média e baixa.

Naive Bayes		
Alta	Média	Baixa
4	3	4
6	8	6
0	2	10

Árvore de Decisão		
Alta	Média	Baixa
11	0	0
19	0	1
6	0	6

Random Forest		
Alta	Média	Baixa
7	3	1
9	7	4
0	3	9