

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP/FCAV
CÂMPUS DE JABOTICABAL**

**PREDIÇÃO DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR
UTILIZANDO DADOS DE ALTA RESOLUÇÃO E
MINERALOGIA DO SOLO**

Luis Alberto Rocha Rodrigues Junior

Engenheiro Agrônomo

Jaboticabal – SP

2025

**UNIVERSIDADE ESTADUAL PAULISTA – UNESP/FCAV
CÂMPUS DE JABOTICABAL**

**PREDIÇÃO DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR
UTILIZANDO DADOS DE ALTA RESOLUÇÃO E
MINERALOGIA DO SOLO**

Luis Alberto Rocha Rodrigues Junior

**Orientadora: Profa. Dra. Tatiana Fernanda Canata
Coorientadores: Dr.^a. Samira Luns Hatum de Almeida
Prof.^o. Dr. Glauco de Sousa Rolim**

Dissertação apresentada à Faculdade de Ciências Agrárias e Veterinárias – UNESP, Câmpus de Jaboticabal, como parte das exigências para obtenção do título de Mestre em Agronomia.

Jaboticabal – SP

2025

R696p	Rodrigues Junior, Luis Alberto Rocha Predição da produtividade da cana-de-açúcar utilizando dados de alta resolução e mineralogia do solo / Luis Alberto Rocha Rodrigues Junior. -- Jaboticabal, 2026 54 p. : il., tabs., mapas Dissertação (mestrado) - Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal Orientadora: Tatiana Fernanda Canata Coorientadora: Samira Luns Hatum de Almeida 1. GLCM. 2. Aprendizado de máquina. 3. LiDAR. 4. Índices de Vegetação. 5. Bandas Espectrais. I. Título.
-------	---

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

CERTIFICADO DE APROVAÇÃO

TÍTULO DA DISSERTAÇÃO: PREDIÇÃO DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR UTILIZANDO DADOS DE ALTA RESOLUÇÃO E MINERALOGIA DO SOLO

AUTOR: LUIS ALBERTO ROCHA RODRIGUES JUNIOR

ORIENTADORA: TATIANA FERNANDA CANATA COORIENTADOR: GLAUCO DE SOUZA ROLIM

COORIENTADORA: SAMIRA LUNS HATUM DE ALMEIDA

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

Documento assinado digitalmente
 **TATIANA FERNANDA CANATA**
Data: 28/02/2026 13:34:48-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. TATIANA FERNANDA CANATA (Participação Virtual)
Departamento de Engenharia / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **MARCELO RODRIGUES BARBOSA JUNIOR**
Data: 27/02/2026 10:37:13-0300
Verifique em <https://validar.iti.gov.br>

Dr. MARCELO RODRIGUES BARBOSA JÚNIOR (Participação Virtual)
Departamento de Ciência e Tecnologia Vegetal / University of Missouri (MU) - Portageville/MO

Documento assinado digitalmente
 **TERESA CRISTINA TARLE PISSARRA**
Data: 27/02/2026 20:26:51-0300
Verifique em <https://validar.iti.gov.br>

Profa. Dra. TERESA CRISTINA TARLE PISSARRA (Participação Virtual)
Departamento de Engenharia / FCAV UNESP Jaboticabal

Jaboticabal, 27 de fevereiro de 2026.

DADOS CURRICULARES DO AUTOR

LUIS ALBERTO ROCHA RODRIGUES JUNIOR – Nasceu em 26 de julho de 1999, no município de Viana, estado do Maranhão. Em agosto de 2018 ingressou no curso de Bacharel em Agronomia na Universidade Estadual do Maranhão, São Luís, recebendo o título de engenheiro agrônomo em dezembro de 2023. Em Março de 2024 ingressou na Faculdade de Ciências Agrárias e Veterinárias – FCAV/UNESP, Câmpus de Jaboticabal/SP, no Programa de Pós-graduação em Agronomia – Produção Vegetal em nível de Mestrado Acadêmico.

Ao meu querido avô José Diniz Rodrigues (*in memoriam*). Sua memória e ensinamentos ainda vivem em cada objetivo alcançado.

Dedico.

PREDIÇÃO DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR UTILIZANDO DADOS DE ALTA RESOLUÇÃO E MINERALOGIA DO SOLO

RESUMO – A estimativa antecipada da produtividade é fundamental para o planejamento agrícola, o manejo localizado e a tomada de decisões em áreas comerciais, especialmente diante da elevada variabilidade espacial presente nos canaviais. No entanto, os métodos tradicionais de estimativa de produtividade são, em geral, destrutivos, onerosos e pouco eficientes na representação dessa variabilidade. Nesse contexto, a integração de dados de sensoriamento remoto de alta resolução, atributos do solo e técnicas de aprendizado de máquina surge como uma alternativa promissora para a predição da produtividade da cana-de-açúcar. Objetivou-se nesta pesquisa aplicar modelos de predição da produtividade da cana-de-açúcar em escala de talhão, a partir da integração de dados orbitais de alta resolução, informações estruturais derivadas do LiDAR (Light Detection and Ranging) e atributos mineralógicos da fração argila. O estudo foi conduzido em uma área comercial, no estado de São Paulo. Os dados de produtividade real foram obtidos por sensores embarcados em colhedora de cana-de-açúcar, georreferenciados por GNSS com correção RTK. Foram utilizadas imagens multiespectrais da constelação PlanetScope, adquiridas em diferentes estádios fenológicos da cultura, das quais foram extraídas as bandas espectrais, os índices de vegetação e os atributos de textura baseados na Matriz de Co-ocorrência em Níveis de Cinza (GLCM). A altura das plantas foi obtida a partir de levantamentos LiDAR aerotransportados, enquanto a mineralogia do solo foi caracterizada por meio de medições de suscetibilidade magnética, convertidas em um índice representativo da fração argila. Foram avaliados os modelos Random Forest, XGBoost, KNN e Regressão Linear Múltipla. Os resultados indicaram que o modelo XGBoost apresentou o melhor desempenho comparado aos outros modelos testados, com os atributos de textura derivados da GLCM se destacando como as variáveis mais importantes para a predição da produtividade, superando os dados de altura de plantas e mineralogia do solo.

Palavras-chave: GLCM, Aprendizado de Máquina, LiDAR, Índices de Vegetação, Bandas espectrais.

PREDICTING SUGARCANE YIELD USING HIGH RESOLUTION DATA AND SOIL MINERALOGY

Abstract – Early yield estimation is essential for agricultural planning, site-specific management, and decision-making in commercial fields, especially considering the high spatial variability observed in sugarcane crops. However, traditional yield estimation methods are generally destructive, costly, and inefficient in capturing this variability. In this context, the integration of high-resolution remote sensing data, soil attributes, and machine learning techniques emerges as a promising alternative for sugarcane yield prediction. This study aimed to apply yield prediction models at the field scale by integrating high-resolution orbital data, structural information derived from LiDAR (Light Detection and Ranging), and mineralogical attributes of the clay fraction. The study was conducted in a commercial area in the state of São Paulo, Brazil. Actual yield data were obtained using sensors mounted on a sugarcane harvester, georeferenced by GNSS with RTK correction. Multispectral images from the PlanetScope constellation were used, acquired at different phenological stages of the crop, from which spectral bands, vegetation indices, and texture attributes based on the Gray-Level Co-occurrence Matrix (GLCM) were extracted. Plant height was obtained from airborne LiDAR surveys, while soil mineralogy was characterized through magnetic susceptibility measurements, converted into an index representative of the clay fraction. The models evaluated included Random Forest, XGBoost, KNN, and Multiple Linear Regression. The results indicated that the XGBoost model achieved the best performance among the tested models, with texture attributes derived from GLCM standing out as the most important variables for yield prediction, outperforming plant height and soil mineralogy data.

Keywords: GLCM; Machine Learning; LiDAR; Vegetation Indices, Spectral bands

SUMÁRIO

1	INTRODUÇÃO	11
2	REVISÃO DE LITERATURA	14
2.1	A cultura da Cana-de-açúcar	14
2.2	Sensoriamento Remoto na Agricultura	15
2.2.1	Índices de Vegetação	17
2.2.2	Textura GLCM	18
2.3	Uso do LiDAR na agricultura	20
2.4	Mineralogia do solo	21
2.5	Aprendizado de máquina	22
2.5.1	Floresta Aleatória	23
2.5.2	XGBoost	24
2.5.3	KNN	24
2.5.4	Regressão Linear Múltipla	25
3	MATERIAL E MÉTODOS	25
3.1	Área experimental	25
3.2	Dados de produtividade	26
3.3	Dados de Imagem de satélite	27
3.3.1	Aquisição de Refletância e Índices de Vegetação	28
3.3.2	Aquisição de Textura GLCM	29
3.4	Dados tridimensionais	30
3.5	Dados de mineralogia do solo	31
3.6	Análise estatística e remoção de outliers	32
3.7	Modelagem dos modelos preditivos	32
3.7.1	Geração do Banco de dados	32
3.7.2	Modelos de aprendizado de máquina	33
4	RESULTADOS	35
4.1	Análise estatística da produtividade	35
4.2	Predição da Produtividade	35
4.2.1	Desempenho dos modelos com todas as variáveis de entrada	35
4.2.2	Variáveis selecionadas a partir do Selection Features	36
4.2.3	Desempenho dos modelos com as variáveis selecionadas	37
4.3	Mapas preditivos dos modelos	39
5	DISCUSSÃO	40
6	CONCLUSÕES	46
	REFERÊNCIAS	47

LISTA DE ABREVIATURAS

NDVI	Índice de Vegetação por Diferença Normalizada
GNDVI	Índice de Vegetação por Diferença Normalizada no Verde
WDRVI	Índice de Vegetação com Ampla Faixa Dinâmica
EVI	Índice de Vegetação Aprimorado
ML	Machine Learning
RLM	Regressão Linear Multipla
KNN	<i>K nearest neighbors</i>
RF	<i>Random Forest</i>
GLCM	<i>Gray Level Co-occurrence Matriz</i>
LiDAR	<i>Light Detection and Ranging</i>
MDT	Modelo Digital de Terreno
MDS	Modelo Digital de Superfície
CHM	Modelo de altura de dossel
IDW	Inverso da distância
SR	Sensoriamento Remoto

LISTA DE FIGURAS

Figura 1	Ciclo fenológico da cana-de-açúcar, desde o plantio até o crescimento completo
Figura 2	Resposta espectral da vegetação sob diferentes níveis de estresse
Figura 3	Esquema conceitual do sensoriamento LiDAR aplicado ao dossel vegetal
Figura 4	Arquitetura de funcionamento do algoritmo Random Forest
Figura 5	Localização da área experimental e descrição da temperatura média e precipitação acumulada ao decorrer do ciclo da cana-de-açúcar estudado
Figura 6	Dados obtidos da imagem de satélite: Bandas espectrais, índices de vegetação e textura (A); Dados obtidos a partir da nuvem de pontos do LiDAR: MDS, MDT e Altura de plantas (CHM) (B); Dados de mineralogia do solo com a grade amostral e variabilidade da área (C)
Figura 7	Boxplot da produtividade real por talhões
Figura 8	Impacto das 20 variáveis selecionadas na predição da produtividade para os modelos KNN (A), RLM (B), XGBoost (C) e RF (D)
Figura 9	Desempenho dos modelos KNN (A), Random Forest (B), Regressão Linear Múltipla (C) e XGBoost (D) após a remoção da multicolinearidade e a seleção de variáveis, avaliados com base nos dados de teste
Figura 10	Mapa de produtividade real (A) e produtividade predita com as 20 variáveis selecionadas para os modelos RLM (B); RF (C); XGBoost (D) e KNN (E)
Figura 11	Mapa de erro percentual nos dos valores preditos com as 20 variáveis selecionadas para o modelo RLM (ARF (B); Mapa de XGBoost (C) e KNN (D)

LISTA DE TABELAS

Tabela 1	$P(i, j)$: Elemento da matriz GLCM, que representa a probabilidade de ocorrência do par de níveis de cinza i e j ; N_g : Número total de níveis de cinza; μ_i, μ_j : médias marginais de i e j ; $\sigma_i \sigma_j$: desvios padrão marginais de i e j .
Tabela 2	Descrição das datas de aquisição das cenas do satélite Planet e indicação do estágio vegetativo da cana-de-açúcar.
Tabela 3	Especificações espectrais do sensor Doble Classic PS2 – Planet.
Tabela 4	Índices de vegetação utilizados para composição do banco de dados.
Tabela 5	Medidas de textura de imagem baseadas na matriz GLCM e suas equações de cálculo.
Tabela 6	Desempenho dos modelos com todas as variáveis utilizando os dados de teste

1 INTRODUÇÃO

A cana-de-açúcar (*Saccharum officinarum*) é uma das culturas agrícolas mais relevantes do mundo, destacando-se como a principal fonte de açúcar e como matéria-prima para a produção de etanol e bioenergia (Figueroa-Rodríguez et al., 2019). Seu elevado potencial produtivo e importância socioeconômica tornam essencial o desenvolvimento de estratégias que possibilitem o acompanhamento da produtividade com maior eficiência (Waters et al., 2025). No entanto, as estratégias convencionais para mensurar a produtividade são, em geral, trabalhosos e incapazes de capturar a variabilidade espacial existente nas lavouras (Wei et al., 2020). Nesse contexto, o uso de tecnologias associadas à agricultura digital, tais como o sensoriamento remoto (SR) e o aprendizado de máquina (*machine learning*, ML) tem se mostrado uma abordagem promissora para a predição da produtividade em ambientes agrícolas complexos (Wang et al., 2021).

O sensoriamento remoto é uma das tecnologias que hoje estão abrangendo as formas de monitorar as lavouras. As Informações extraídas por meio do SR, como os dados espectrais extraídos de imagens orbitais por exemplo, permitem descrever, de forma indireta, as doenças nas culturas (Shen et al., 2024), o nível de nitrogênio (You et al., 2023), estrutura e o vigor do dossel de plantas (Dhakal et al., 2023). Além disso, os índices de vegetação derivados das bandas espectrais, como o NDVI (Índice de vegetação por diferença normalizada), também são amplamente utilizados para representar a biomassa e o vigor das plantas (Tian et al., 2025). Em uma outra abordagem, a incorporação de atributos texturais, obtidos a partir de algoritmos de Matriz de Co-ocorrência de Níveis de Cinza (GLCM), permite captar a rugosidade e a heterogeneidade do dossel (Mryka Hall-Beyer, 2017) por meio de estatísticas de textura, conhecidas como atributos de Haralick, que sintetizam padrões estruturais relacionados aos níveis de cinza da imagem (Lee et al., 2024).

Além das informações espectrais e texturais, as variáveis biométricas, como a altura de plantas, contribuem para a modelagem da produtividade. Uma das formas de mensurar a altura de plantas em campo é com o uso do LiDAR (Light Detection and Ranging), que, por meio de varredura a laser da superfície, possibilita a aquisição de dados tridimensionais de alta resolução, gerando uma nuvem de pontos altamente densa (Sofonia et al., 2019). Na cana-de-açúcar, o sistema LiDAR se mostra promissor

na aquisição de dados, apresentando correlações significativas de atributos derivados do LiDAR e biomassa (Shendryk et al., 2020), fornecendo uma representação fiel da estrutura da cana-de-açúcar (Canata et al., 2022). A aplicabilidade desses dados quando integradas a outros tipos de variáveis de entradas, têm fortalecido os modelos de predição capazes de capturar a variabilidade espacial e retornar informações relevantes para a cultura (Xu et al., 2020).

Outro fator determinante para a produtividade da cana-de-açúcar é o solo (Virani et al., 2025), especialmente a fração argila, que exerce papel central nos processos pedogenéticos e na disponibilidade de nutrientes (Barrios et al., 2017). Os índices derivados do estudo de argilas, como suscetibilidade magnética, têm sido utilizados como indicadores indiretos de atributos químicos e físicos do solo, apresentando correlação com capacidade de troca catiônica e disponibilidade de nutrientes (Peluco et al., 2013). O estudo desse índice reflete propriedades mineralógicas do solo, que influenciam na retenção de água e estabilidade dos agregados, afetando diretamente o desenvolvimento radicular. Nesse sentido, a relação entre a mineralogia da fração argila e a produtividade da cana-de-açúcar já foi demonstrada por Catelan et al. (2022), que verificaram a possibilidade de definir zonas de manejo para a cultura a partir desse atributo. Dessa forma, a incorporação de informações mineralógicas no treinamento de modelos de aprendizado de máquina possui potencial para ampliar a compreensão das interações entre o solo e a produtividade no sistema de produção.

Nesse contexto, os modelos de aprendizado de máquina têm sido amplamente aplicados na predição da produtividade da cana-de-açúcar (Akbarian et al., 2022; Amaro et al., 2025; De Oliveira et al., 2022; Picoli et al., 2009), uma vez que possuem capacidade de modelar relações não lineares e integrar múltiplas fontes de dados ambientais e agronômicos, como os descritos acima. Os algoritmos como Random Forest, Regressão Linear Multipla e redes neurais artificiais têm desempenhado positivamente em diferentes predições na cana-de-açúcar. Barbosa Junior et al. (2024), por exemplo, conseguiram predizer matérias-primas de sacarina e bioenergia na cana-de-açúcar integrando os dados de sensoriamento remoto algoritmos de Regressão Linear Multipla, *Random Forest*, *Árvore de decisão* e *Support Vector Machine*. Uma outra abordagem, integrando dados de imagens orbitais e ambientais foi vista por Shendryk et al., (2021), que conseguiram predizer a

produtividade da cana-de-açúcar, concentração de açúcar e produtividade de açúcar em nível de campo com $R^2 = 0,51$, $0,63$ e $0,62$ respectivamente, utilizando algoritmos de aprendizado de máquina como o XGBoost. Esses estudos reforçam que o uso de diferentes modelos pode ser uma ferramenta promissora para predições que busquem integrar novas variáveis no intuito de prever a produtividade da cana-de-açúcar.

No entanto, grande parte dos estudos concentra-se na utilização isolada de índices espectrais combinados com conjuntos restritos de variáveis, como bandas espectrais. Além disso, ainda são escassos os trabalhos que integram simultaneamente informações espectrais, atributos texturais derivados de imagens orbitais, variáveis estruturais obtidas por LiDAR e indicadores mineralógicos do solo para a predição da produtividade da cana-de-açúcar em escala de talhão. Nesse sentido, a exploração entre essas diferentes fontes de dados pode contribuir para o desenvolvimento de modelos preditivos mais robustos e para uma melhor compreensão dos fatores que determinam a variabilidade nas lavouras, possibilitando a captura da complexidade temporal e espacial dos ambientes produtivos.

Dessa forma, sugere-se que a integração dos dados supracitados, podem gerar resultados que indiquem caminhos no manejo da cultura no que tange tomadas de decisões antecipadas e inferências mais localizadas nos talhões. Portanto, o presente estudo tem como objetivo aplicar modelos de predição da produtividade da cana-de-açúcar em escala de talhão, a partir da integração de dados orbitais de alta resolução, informações estruturais derivadas do LiDAR e atributos mineralógicos da fração argila.

2 REVISÃO DE LITERATURA

2.1 A cultura da Cana-de-açúcar

A cana-de-açúcar é uma gramínea perene alta do gênero *Saccharum*, utilizada para a produção de açúcar. As plantas geralmente atingem de 2 a 6 m de altura, com colmos robustos, nodulares e fibrosos, ricos em sacarose e considerada uma das culturas mais importantes para a produção de bioenergia e açúcar no mundo (Diniz et al., 2019; Lima et al., 2001). Desde a sua colonização, o Brasil se caracterizou como um potencial e grande produtor da cultura da cana-de-açúcar (Rodrigues et al., 2020), com safra estimada para 2024/2025 em 676,98 milhões de toneladas (Conab, 2025). A produção é predominante principalmente nas regiões Centro-Sul e Norte-Nordeste, o que permite dois períodos distintos de safra por ano, sendo que de abril a novembro a produção ocorre no Centro-Sul e de setembro a março no Norte-Nordeste (Shikida, 2013).

Morfologicamente, a cana-de-açúcar apresenta um hábito ereto e inflorescência em espiga, com crescimento do caule em colmos (Marafon, 2012). Suas folhas possuem caráter estreito e bainha aberta. Além disso por ser uma planta com metabolismo fotossintético C4, é altamente eficiente na conversão de energia radiante em energia química, o que resulta em uma taxa fotossintética elevada. Essa característica torna a cana-de-açúcar bem adaptada a ambientes com alta intensidade luminosa, além de ser eficiente no uso da água (Diniz et al., 2019).

A escolha da época de plantio é crucial para o bom desenvolvimento da cana-de-açúcar, que pode ser plantada em três diferentes períodos: sistema de ano-e-meio, sistema de ano e plantio de inverno. No sistema de ano-e-meio (18 meses), o plantio ocorre entre janeiro e março, período que oferece condições ideais de temperatura e umidade para o crescimento da planta, favorecendo a brotação rápida e a redução da incidência de doenças. No sistema de ano (12 meses), a cana-de-açúcar é plantada entre outubro e novembro, com o desenvolvimento sendo interrompido entre março e abril, retomando nos meses seguintes até a maturação. Já o plantio de inverno é indicado para propriedades com irrigação, sendo realizado na estação seca. Nesse sistema, o canavial alcança altas produtividades no primeiro ano, uma vez que é possível controlar a disponibilidade de água no solo (Rossetto; Santiago, 2022). Dependendo da variedade e do ambiente no qual se desenvolve, o ciclo fenológico (Figura 1) da cultura da cana-de-açúcar, pode ter duração inicial de 16 meses para

cana-planta e 12 meses para cana-soca, com os mesmos estádios de desenvolvimento, entretanto com duração do ciclo diferente (Rossetto; Santiago, 2022; Silva; Cavichioli, 2025).

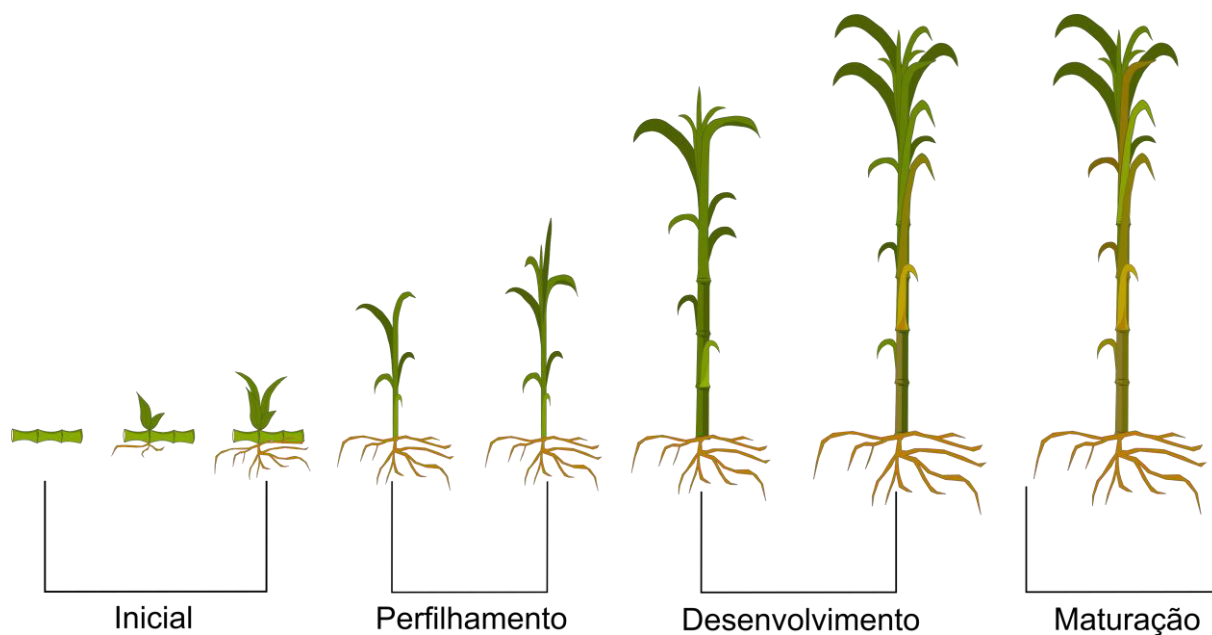


Figura 1: Ciclo fenológico da cana-de-açúcar, desde o plantio até o crescimento completo.

Fonte: Adaptado de Silva, (2019)

Porém, a produção e a qualidade da cana-de-açúcar são favorecidas com a adequada disponibilidade de água e nutrientes no solo, principalmente nos estádios de germinação, perfilamento e crescimento intenso e de restrição hídrica no período de maturação (Brunini; Turco, 2016). A prática da irrigação associada aos outros métodos de manejo e monitoramento da cultura de cana-de-açúcar, evita os danos causados pelo estresse hídrico ou deficiências nutricionais e permite ampliar o tempo de exploração da planta, o número de colheitas e busca o melhor rendimento do ponto de vista econômico, principalmente nas fases de germinação, perfilamento e alongamento dos colmos período no qual há o enriquecimento em sacarose na planta (Almeida et al., 2008; Barbosa Júnior et al., 2024).

2.2 Sensoriamento Remoto na Agricultura

O sensoriamento remoto consolidou-se como uma ferramenta estratégica para estudos ambientais e agrícolas a partir da segunda metade do século XX (Kuhn et al., 2015), impulsionado pelo avanço da exploração espacial e pelo desenvolvimento de sensores orbitais. Ao longo das últimas décadas, a evolução

tecnológica no que tange às resoluções dos sensores, ampliou significativamente a capacidade de monitoramento espacial e temporal das áreas agrícolas, permitindo análises em diferentes escalas e contribuindo para o planejamento, a gestão e a avaliação da dinâmica dos sistemas produtivos (Kuchler, 2021).

Do ponto de vista conceitual, o sensoriamento remoto é definido como a ciência que obtém informações sobre alvos sem contato físico direto, por meio da detecção da radiação eletromagnética refletida ou emitida pelos materiais da superfície terrestre (Formaggio e Saches, 2017). A interpretação das imagens depende fundamentalmente do entendimento das interações entre a radiação eletromagnética e os alvos, como solo, a vegetação (Figura 2) e a água, os quais apresentam respostas espectrais distintas em função de suas propriedades físicas, químicas e estruturais.

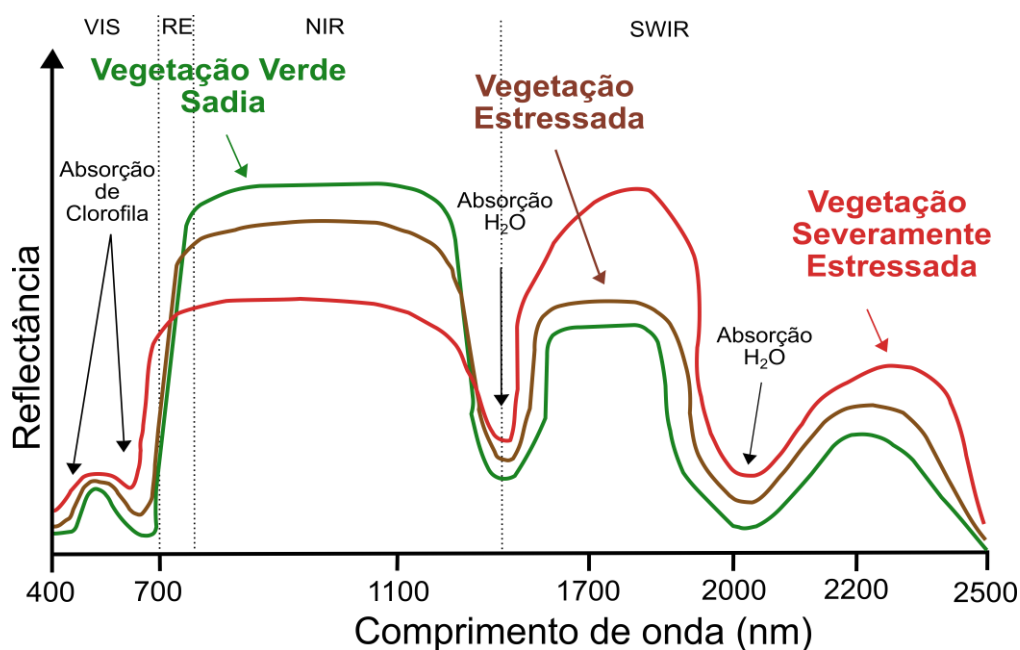


Figura 2: Resposta espectral da vegetação sob diferentes níveis de estresse.

Fonte: Adaptado de Marion et al, (2021)

Nesse contexto, a compreensão do espectro eletromagnético e do comportamento espectral dos alvos é essencial para a extração de informações confiáveis, sendo a base para aplicações agrícolas voltadas ao monitoramento do uso e cobertura da terra, da variabilidade espacial e das mudanças ao longo do tempo (Ponzoni et al., 2012).

Os sensores aplicados ao sensoriamento remoto agrícola podem ser embarcados em plataformas terrestres, aéreas ou orbitais e apresentam características técnicas que determinam a sua adequação os diferentes objetivos de estudo. Entre essas características, destacam-se as resoluções espacial, espectral, radiométrica e temporal, que influenciam diretamente o nível de detalhamento, a sensibilidade às variações dos alvos e a frequência de aquisição das imagens. A escolha adequada do sensor e da plataforma é, portanto, um fator decisivo para a qualidade dos dados e para a precisão das análises em estudos agrícolas, especialmente aqueles voltados ao monitoramento espaço-temporal, à caracterização de áreas produtivas e ao suporte à tomada de decisão em sistemas de produção agropecuários (Barros et al., 2021; Meneses et al., 2025).

2.2.1 Índices de Vegetação

Os índices de vegetação são métricas derivadas da combinação matemática de bandas espectrais obtidas por sensores embarcados em plataformas orbitais, aéreas ou proximais (Ponzoni et al., 2012), desenvolvidas com o objetivo de realçar características da vegetação e reduzir as interferências provenientes do solo, da atmosfera e da geometria de aquisição das imagens. Esses índices fundamentam-se no comportamento espectral típico da vegetação, que apresenta forte absorção da radiação nas regiões do azul e do vermelho do espectro eletromagnético, associada à presença de pigmentos fotossintéticos (Moriwaki et al., 2023), e elevada refletância no infravermelho próximo, relacionada à estrutura celular das folhas (Almalki et al., 2022). Essa resposta espectral diferenciada permite que os índices de vegetação sejam utilizados como indicadores indiretos de atributos biofísicos da cobertura vegetal (Vera-Esmeraldas et al., 2025).

No contexto agrícola, os índices de vegetação têm ampla aplicação no monitoramento espaço-temporal da vegetação, possibilitando a avaliação do vigor, da densidade e da dinâmica dos agroecossistemas ao longo do ciclo produtivo (Lofton et al., 2012). A utilização desses índices em séries temporais de imagens orbitais tem contribuído significativamente para a identificação de padrões de variabilidade espacial, acompanhamento fenológico e a análise de mudanças no uso e cobertura do solo (Pande et al., 2023; Tian et al., 2025).

Dentre os índices comumente utilizados pode-se destacar alguns. O índice de Vegetação por Diferença normalizada (NDVI) por exemplo, é uma medida amplamente aplicada em avaliações da saúde e densidade da vegetação em diversas áreas. Ele é calculado a partir da diferença normalizada entre a refletância na banda do infravermelho próximo (NIR) e na banda do vermelho do espectro eletromagnético (Rouse, 1974). Os valores obtidos a partir da aplicação dos dados de refletância na equação do NDVI variam entre -1 até +1, com valores mais altos indicando maior densidade e saúde da vegetação (Rouse, 1974).

Além disso, outro Índice de Vegetação comum é o Índice de Vegetação Melhorado (EVI), no qual sua aplicação possui vantagens sobre o NDVI ao corrigir condições atmosféricas e ruído provocados por dosséis muito densos. Os índices derivados da banda do verde como o índice de Vegetação Normalizado pela cor do Verde (GNVDI) também contribuem para a análise de áreas agrícolas de forma direta ao mensurar a atividade fotossintética da vegetação. Uma característica marcante deste índice se concentra na sua eficácia na detecção de mudanças na vegetação causadas por fatores ambientais, como o estresse hídrico, as doenças, ou as deficiências de nutrientes (Gitelson; Kaufman; Merzlyak, 1996; Huete et al., 2002).

2.2.2 Textura GLCM

A matriz de Co-ocorrência de níveis de cinza (*Gray Level Co-occurrence Matriz* - GLCM) é um método estatístico para descrição de texturas proposto por Haralick et al. (1973). A GLCM armazena a probabilidade de que dois valores de intensidade de cinza estejam envolvidos por uma determinada relação espacial (Bravo, 2019). A partir desta matriz de probabilidades, diferentes medidas estatísticas (Tabela 1) (Haralick et al, 1973) são extraídas a fim de caracterizar a textura presente na imagem (Martins et al., 2011). Além disso, os parâmetros como a distância d entre os pixels e o ângulo θ caracterizado pela orientação da reta que passa pelos pixels são importantes na caracterização da relação espacial.

Tabela 1: Equações dos atributos texturais extraídos pela Matriz de Co-ocorrência em níveis de Cinza.

Medidas de textura	Equação
Variância	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i - u)^2 P(i, j)$

Segundo momento angular	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} P(i,j)^2$
Média	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} i(P(i,j))$
Homogeneidade	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} \frac{P(i,j)}{1 + i - j }$
Entropia	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} P(i,j) \log_2 P(i,j)$
Dissimilaridade	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} i - j P(i,j)$
Contraste	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i - j)^2 P(i,j)$
Correlação	$\frac{\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i - \mu_i)(j - \mu_j) P(i,j)}{\sigma_i \sigma_j}$

Legenda: $P(i, j)$: Elemento da matriz GLCM, que representa a probabilidade de ocorrência do par de níveis de cinza i e j ; N_g : Número total de níveis de cinza; μ_i, μ_j : médias marginais de i e j ; $\sigma_i \sigma_j$: desvios padrão marginais de i e j .

A textura baseada em GLCM tem sido amplamente explorada na agricultura como um recurso capaz de capturar a organização espacial de alvos agrícolas, complementando a informação espectral tradicional (De Freitas et al., 2024; Huang; Liu; Zhang, 2014). Desse modo, diferentemente das análises baseadas apenas na refletância, a matriz de Co-ocorrência dos níveis de cinza, destaca-se como um dos métodos mais consolidados para a análise textural de imagens. A efetividade dessa abordagem depende, entretanto, de escolhas metodológicas adequadas, como o tamanho da janela de cálculo, o nível de quantização dos tons de cinza e a escala espacial da imagem (De Freitas et al., 2024; Gonçalves; Sette, 2009; Mryka Hall-Beyer, 2017).

Em áreas agrícolas, essas características estão diretamente relacionadas ao arranjo das plantas, ao espaçamento, à arquitetura foliar e às condições do solo exposto, o que torna as métricas de textura particularmente relevantes para diferenciar os estádios fenológicos, identificar as áreas com estresse ou falhas de plantio e discriminar culturas com comportamento espectral semelhante (Gonçalves, 2021).

Alguns estudos têm demonstrado que a integração das métricas de textura GLCM a modelos estatísticos e de aprendizado de máquina potencializa o desempenho em tarefas de classificação e predição agrícola. Variáveis como

contraste, entropia, homogeneidade e correlação são frequentemente utilizadas como atributos de entrada em algoritmos como Random Forest, Support Vector Machines e métodos de regressão, contribuindo para maior robustez dos modelos em contraponto ao uso exclusivo de bandas espectrais ou índices de vegetação (Khojastehnazhand et al., 2025; Lee et al., 2024).

2.3 Uso do LiDAR na agricultura

O LiDAR (*Light Detection and Ranging*) tem se consolidado como uma tecnologia importante no sensoriamento remoto agrícola por permitir a caracterização tridimensional detalhada da superfície e da vegetação (Farhan et al., 2024; Rivera et al., 2023). Fundamentada na emissão ativa de pulsos de laser, a tecnologia LiDAR opera com base nos princípios da telemetria a laser e na medição do tempo de voo (time of flight), no qual o intervalo entre a emissão do pulso e o seu retorno ao sensor é utilizado para o cálculo preciso da distância até os alvos interceptados (Novero et al., 2019).

Ao interagir com diferentes componentes do dossel vegetal e com o solo, os pulsos podem gerar múltiplos retornos, possibilitando a discriminação vertical da estrutura da vegetação (Figura 3). Esse processo viabiliza a obtenção direta de coordenadas tridimensionais (x, y, z), resultando na geração de nuvens de pontos de alta densidade, que representam com elevado nível de detalhamento a geometria da superfície e dos elementos vegetados (Canata, 2021; Eitel et al., 2016).

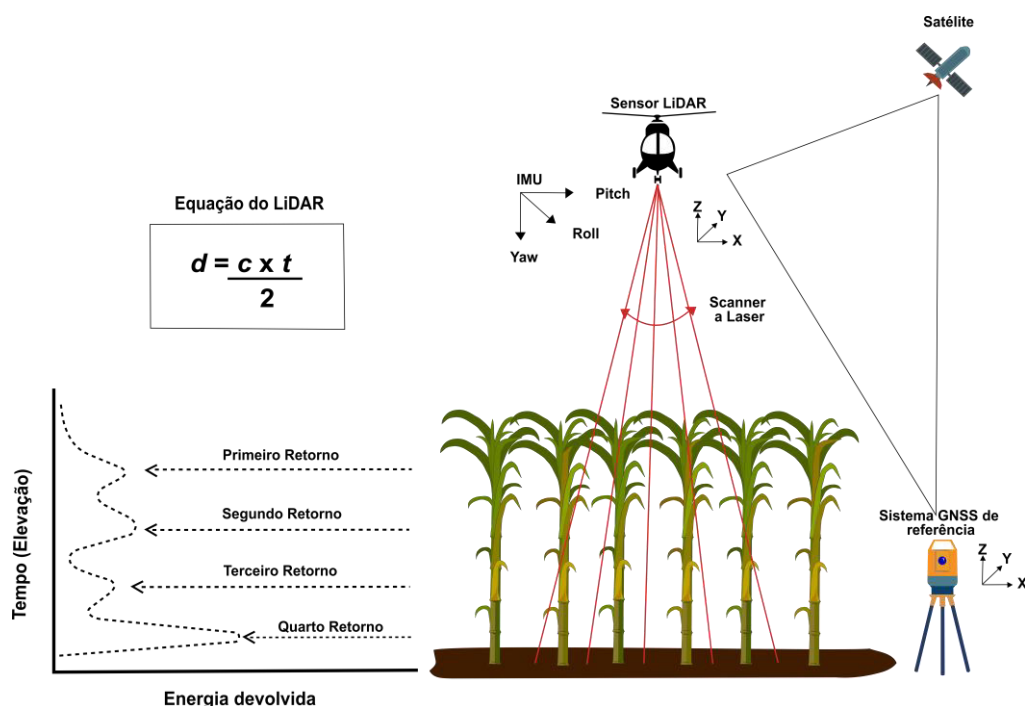


Figura 3: Esquema conceitual do sensoriamento LiDAR aplicado ao dossel vegetal.

Fonte: Adaptado de Silva, (2025).

Legenda: D = Distância entre o sensor e o alvo; C = Velocidade de propagação da luz em m.s^{-1} T tempo de emissão e retorno ao sensor.

Os produtos oriundos da nuvem de pontos como modelo digital de elevação e de superfície, permitem a extração de métricas estruturais como a altura das plantas, a densidade do dossel, volume da copa e rugosidade da superfície, as quais podem apresentar forte relação com variáveis biofísicas e agronômicas, sendo particularmente relevantes para o monitoramento do crescimento das culturas e para a avaliação da variabilidade espacial aos talhões (Rivera et al., 2023; Sofonia et al., 2019). A literatura evidencia que as métricas derivadas do LiDAR podem ser trabalhadas de forma eficiente a modelos de inteligência artificial, ampliando a acurácia na estimativa de produtividade, biomassa e vigor vegetativo (Shendryk et al., 2020; Xu et al., 2020).

2.4 Mineralogia do solo

A susceptibilidade magnética do solo se mostra como uma propriedade física indicadora da composição mineralógica e dos processos pedogenéticos que controlam a dinâmica dos nutrientes e a qualidade do solo (Fontes et al., 1985). Essa propriedade expressa a capacidade do solo de se magnetizar na presença de um

campo magnético externo, sendo fortemente influenciada pela presença e concentração de minerais magnéticos, como magnetita, maghemita e hematita, os quais estão associados aos óxidos de ferro formados durante o intemperismo (Siqueira, 2010). Em um contexto agrônomo, a susceptibilidade magnética apresenta correlação com atributos químicos e físicos do solo, como teor de argila, capacidade de troca catiônica e disponibilidade de nutrientes (Peluco et al., 2013), o que a torna uma ferramenta promissora para o mapeamento indireto da fertilidade e da variabilidade espacial dos solos agrícolas (Siqueira et al., 2016).

Além disso, a susceptibilidade magnética tem sido empregada como uma variável eficiente para a identificação de zonas de manejo e para a compreensão da heterogeneidade espacial dos atributos do solo em diferentes sistemas de produção (Catelan et al., 2022). A literatura aponta que áreas com maior susceptibilidade magnética tendem a apresentar maior teor de óxidos de ferro pedogenéticos, frequentemente associados a solos mais intemperizados, estruturados e com maior potencial produtivo (Brito et al., 2022; Souza Junior et al., 2010), evidenciando sua utilidade como indicador indireto do desempenho agrônomo das culturas.

Na cultura da cana-de-açúcar, a susceptibilidade magnética tem se mostrado particularmente relevante devido à forte relação entre mineralogia do solo, dinâmica do carbono e produtividade agrícola (Catelan et al., 2022). Além disso, há a possibilidade do seu uso discriminar os ambientes de produção, auxiliar na delimitação de zonas de manejo e contribuir para a modelagem da produtividade da cana-de-açúcar, sobretudo quando combinada aos dados de sensoriamento remoto e modelos estatísticos ou de aprendizado de máquina (Siqueira et al., 2016). A incorporação da de fatores mineralógicos do solo em estudos agrônomo pode ampliar a compreensão da interação solo planta e fortalecer abordagens integradas com as ferramentas de inteligência artificial.

2.5 Aprendizado de máquina

O Aprendizado de máquina, também conhecido como machine learning, pode ser definido como um subconjunto da inteligência artificial (IA) concentrado em algoritmos que podem "aprender" os padrões dos dados de treinamento e, posteriormente, fazer inferências precisas sobre novos dados (Chen, 2025). Essa ferramenta consegue identificar padrões complexos e relações não lineares em

conjuntos de dados multivariados, característica especialmente relevante em aplicações agrícolas que integram informações de sensoriamento remoto, atributos do solo e variáveis agrônomicas (Mukhamediev et al., 2022).

Diferentemente de abordagens estatísticas tradicionais, os algoritmos de ML permitem explorar grandes volumes de dados heterogêneos, aprendendo diretamente a partir dos dados observados, sem a necessidade de pressupostos rígidos sobre a distribuição das variáveis (Botero-Valencia et al., 2025). Nesse sentido, diversos modelos podem ser adotados para análises multivariadas e não lineares como por exemplo: Floresta aleatória, XGBoost, K-Vizinhos mais próximos e Regressão Linear Multipla. Os modelos exemplificados, de maneira geral seguem uma linha de funcionamento no qual envolvem a divisão do banco de dados em duas fases: treinamento e teste. Quando o desempenho atinge um nível satisfatório, medido por métricas matemáticas e estatísticas, o processo é concluído. O modelo resultante pode então ser utilizado para tarefas como classificação, agrupamento ou predição (Oliveira, 2023).

2.5.1 Floresta Aleatória

O algoritmo de Floresta Aleatória desenvolvido por Breiman (2001) combina a saída de múltiplas árvores de decisão para chegar a um único resultado (Sruthi, 2025). Se caracteriza como uma ferramenta de aprendizado de máquina, com arquitetura baseado em árvores de decisão (Figura 4) e com a capacidade preditiva tanto para regressões quanto para classificações (Vieira; Oliveira, 2015).

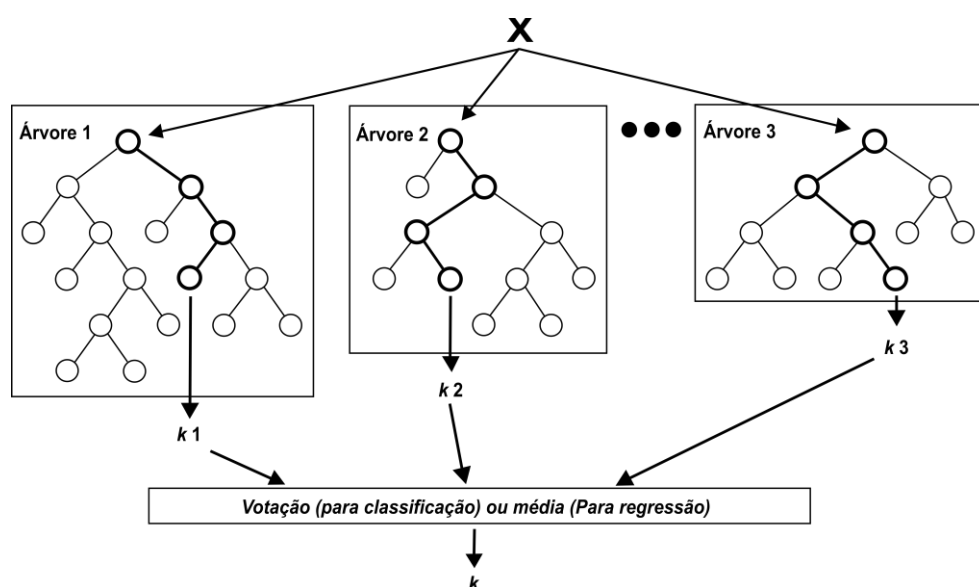


Figura 4: Arquitetura de funcionamento do algoritmo Random Forest.

Fonte: Adaptado de Verikas et al., (2016).

Sua ampla popularidade decorre de sua adaptabilidade, permitindo que lide com problemas de regressão ou classificação de forma eficaz. A robustez do algoritmo Random Forest está associada à sua aptidão para modelar conjuntos de dados complexos, reduzindo os efeitos do sobreajuste por meio da combinação de múltiplas árvores de decisão (Breiman, 2001). Essa característica torna o método particularmente eficiente em diferentes tarefas preditivas no contexto do aprendizado de máquina.

Além disso, o Random Forest apresenta elevada flexibilidade ao lidar tanto com variáveis contínuas, comumente empregadas em problemas de regressão, quanto com variáveis categóricas, vistas em problemas de classificação, ampliando sua aplicabilidade em análises agrícolas e ambientais com dados não lineares (Charoen-Ung; Mittrapiyanuruk, 2018; Everingham et al., 2016).

2.5.2 XGBoost

O XGBoost, ou *eXtreme Gradient Boosting*, é uma biblioteca open-source que implementa algoritmos de aprendizagem de máquina que funcionam sobre um framework de gradient boosting (Chen e Guestrin, 2016). Seu funcionamento consiste na construção sequencial de árvores de regressão, em que cada nova árvore é ajustada para corrigir os erros residuais das anteriores, utilizando gradientes de primeira e segunda ordem da função de perda, o que melhora a convergência e a estabilidade do modelo (Gonçalves et al., 2025; Sahin, 2020).

A importância do XGBoost reside principalmente no controle da complexidade do modelo e redução do sobreajuste, tratamento eficiente de dados esparsos, paralelização e otimizações de cache e memória, permitindo seu uso em grandes volumes de dados, sendo uma ferramenta que potencializa tarefas de regressão, classificação e ranqueamento (Mahdizadeh Gharakhanlou e Perez, 2024; Xie e Tummala, 2025).

2.5.3 KNN

O algoritmo K-Nearest Neighbors (KNN) é um método de classificação e regressão amplamente utilizado, baseado na proximidade entre os dados (Cover;

Hart, 1952). Para classificação, ele atribui a um ponto não classificado a categoria mais comum entre seus k vizinhos mais próximos. Em tarefas de regressão, ele prediz o valor de uma variável contínua com base na média ou média ponderada dos valores de seus k vizinhos mais próximos (Curbani, 2023).

2.5.4 Regressão Linear Múltipla

A Regressão Linear Múltipla (RLM) é uma técnica estatística utilizada para investigar e modelar relações entre uma variável resposta e diversas variáveis explicáveis de resposta, descrito pela Equação 1. O modelo parte do pressuposto da não colinearidade entre elas (Grego, 2014). A RLM ajusta os coeficientes de regressão para minimizar o erro de previsão, sendo menos propensa ao overfitting, especialmente quando há poucos dados em relação ao número de características preditoras (Silva, 2025).

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots \beta_k X_k + u_i \quad \text{Equação 1}$$

Onde:

Y_i é a variável dependente,

β_0 é o intercepto,

$\beta_1, \beta_2 \dots \beta_k$ são os coeficientes de regressão que medem o impacto de cada variável independente,

$X_1, X_2 \dots X_k$ são as variáveis independentes que afetam a variável dependente Y_i

u_i é o termo de erro ou resíduo, que captura a variação de Y_i que não pode ser explicada pelas variáveis independentes.

No entanto, devido à falta de operadores avançados e à sensibilidade aos outliers, a RLM pode ter limitações na previsão de fenômenos complexos que contam com grande quantidade de variáveis, afetando o desempenho do modelo (De; Aparecido, 2019).

3 MATERIAL E MÉTODOS

3.1 Área experimental

A área experimental (Figura 5) está localizada em um campo comercial, com coordenada geográfica 22°41'28.66"S; 48°16'51.82"O, dividida em quatro talhões

localizados no município de Botucatu, São Paulo, Brasil, com altitude média de 480 m.

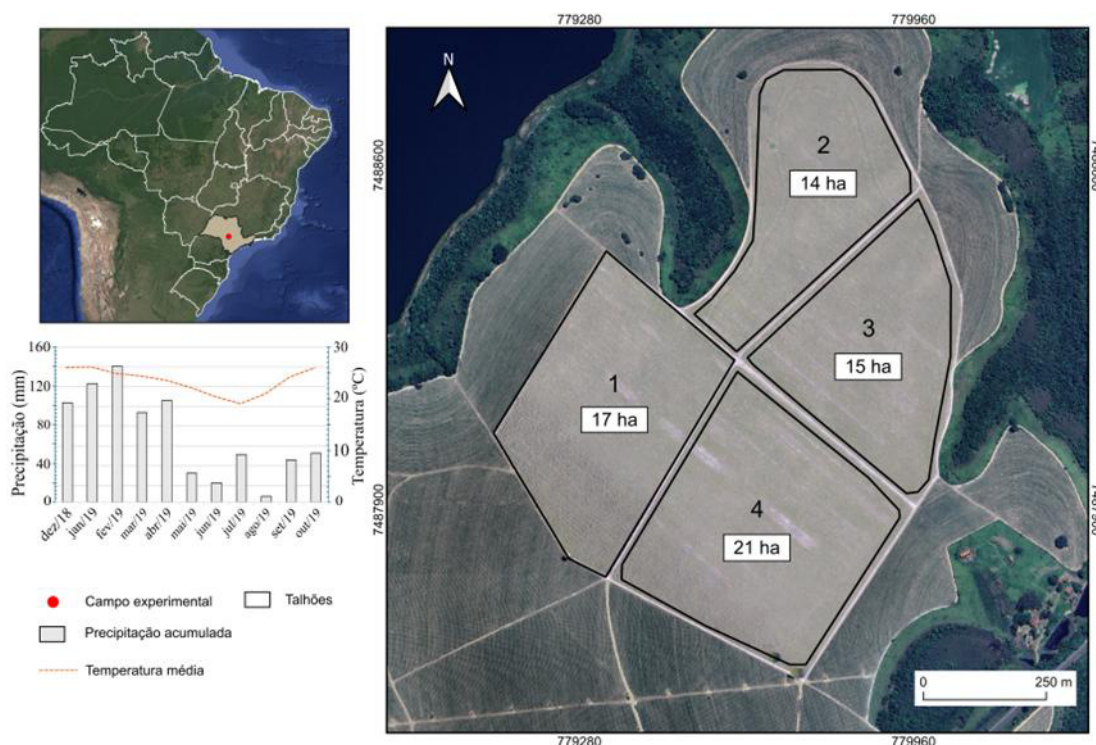


Figura 5: Localização da área experimental e descrição da temperatura média e precipitação acumulada ao decorrer do ciclo da cana-de-açúcar estudado.

O clima da região é classificado, segundo Koppen, como mesotérmico, Cwa (Alvares et al., 2013), caracterizando-se como subtropical úmido, com seca no inverno (de junho a setembro) e chuvas durante os meses de novembro a abril. A precipitação média anual é de 1.433 mm. Além disso, a umidade relativa da região é de 70%, com temperatura média anual de 21° C. A cultivar de cana-de-açúcar avaliada neste estudo foi a cultivar SP83-2847 no quarto ano de colheita, plantada em solo do tipo argissolo com espaçamento de 1,5 m entre linhas.

3.2 Dados de produtividade

Os dados de produtividade da cana-de-açúcar utilizado como referência para modelagem foram gerados na safra 2018/2019 a partir de sensores comerciais embarcados na colhedora que mede a diferença de pressão hidráulica do picador da colhedora calibrado para a produtividade total real, combinados com um receptor de Sistema Global de Navegação por Satélite (GNSS) com sinal de correção diferencial

Real-Time Kinematic (RTK) (GPS L1/L2 + Glonass) para georreferenciar os dados a cada cinco segundos (0,2 Hz). Os dados originais validados por (Maldaner et al., 2022) foram convertidos para valores de produtividade com base na distribuição de peso do transporte e filtrados conforme descrito por Canata et al., (2021).

Os dados de produtividade real extraídos da colhedora, foram interpolados para os quatro talhões por meio da técnica de krigagem ordinária, com tamanho de pixel de 10 x 10 m, utilizando o plugin do QGIS Smart-Map (versão 1.5) (Pereira et al., 2022). O semivariograma gerado para a produtividade predita apresentou ajuste pelo modelo linear, com alcance de aproximadamente 395,7 m, indicando a distância máxima em que há dependência espacial significativa entre as amostras. O efeito pepita de 8,62 foi com um patamar 53,38, com $R^2 = 0,995$ e erro de ajuste reduzido (RMSE = 27,54).

3.3 Dados de Imagem de satélite

A data da colheita anterior (11 de novembro de 2018) foi utilizada como referência para selecionar as cenas das imagens orbitais e calcular o intervalo de dias após o corte (DAC). A partir dessa data, foram selecionadas 11 imagens que correspondem aos principais estádios fenológicos da cana-de-açúcar (Tabela 2) até a fase de maturação. As cenas foram adquiridas na plataforma PlanetScope (<https://www.planet.com/explorer/>), cuja constelação de aproximadamente 130 nanossatélites fornece dados diários da superfície terrestre já corrigidos para reflectância no topo da atmosfera. Todas as imagens foram reprojctadas para o sistema de coordenadas UTM zona 22s, DATUM WGS 84 e recortadas utilizando o contorno dos talhões de interesse.

As cenas foram selecionadas baseado nos seguintes critérios: Cobertura de nuvens < 2%; inicialmente, assumiu-se um intervalo de aproximadamente 30 (com margem de ± 5 dias) dias após o corte (DAC), porém, esse intervalo sofreu alterações devido a inconsistência das imagens no que diz respeito à cobertura de nuvens e captura das cenas pelos nanossatélites, conforme descrito na Tabela 2.

Tabela 2: Descrição das datas de aquisição das cenas do satélite Planet e indicação do estágio vegetativo da cana-de-açúcar.

ID	DAC	Data da imagem de satélite – safra 2018/2019	Estádio fenológico
I1	30	14-12-2018	Inicial
I2	60	19-01-2019	Inicial
P1	90	09-02-2019	Perfilhamento
P2	120	23-03-2019	Perfilhamento
P3	150	20-04-2019	Perfilhamento
D1	180	22-05-2019	Desenvolvimento
D2	210	26-06-2019	Desenvolvimento
D3	240	24-07-2019	Desenvolvimento
M1	270	27-08-2019	Maturação
M2	300	17-09-2019	Maturação
M3	330	17-10-2019	Maturação

3.3.1 Aquisição de Refletância e Índices de Vegetação

Foram utilizadas quatro bandas espectrais fornecidas pelo sensor Dove classic PS2 do satélite Planet (Tabela 3). Ao total, foram calculados quatro IVs (Tabela 4), baseados no que a literatura (Jabed et al., 2024; Tian et al., 2025) menciona para o monitoramento de culturas agrícolas relacionados ao vigor e biomassa.

Para o processamento das imagens orbitais, a extração de refletância e a geração dos índices de vegetação, foi utilizado o software de acesso aberto QGIS (versão 3.42.1).

Tabela 3: Especificações espectrais do sensor Doble Classic PS2 - Planet.

Banda espectral	Comprimento de onda (nm)	Resolução		Radiométrica (bits)
		Espacial (m)	Temporal (dias)	
B1 - Blue	455 – 515	3	Diária	16
B2 - Green	500 - 590			
B3 - Red	590 - 670			
B4 - NIR	780 - 860			

Tabela 4: Índices de vegetação utilizados para composição do banco de dados.

Índice de vegetação	Equação	Referência
---------------------	---------	------------

Índice de Vegetação por Diferença Normalizada	$NDVI = \frac{B4 - B3}{B4 + B3}$	Rouse (1974)
Índice de Vegetação por Diferença Normalizada no Verde	$GNDVI = \frac{B4 - B2}{B4 + B2}$	Gitelson; Kaufman; Merzlyak, (1996)
Índice de Vegetação com Ampla Faixa Dinâmica	$WDRVI = \frac{0,1 \times B4 - B3}{0,1 \times B4 + B3}$	Gitelson (2004)
Índice de Vegetação Aprimorado	$EVI = 2,5 \times \frac{(B4 - B3)}{(B4 + 6) + (B3 \times 7,5) - (B1 \times 1)}$	Huete et al., (2002)

Legenda: B1: Banda do Azul; B2: Banda do Verde; B3: Banda do Vermelho; B4: Banda do infravermelho.

3.3.2 Aquisição de Textura GLCM

As características texturais das imagens orbitais foram mensuradas utilizando o algoritmo da Matriz de Co-ocorrência em níveis de Cinza (GLCM), proposto por Haralick et al. (1973). A GLCM refere-se à variação espacial dos tons de cinza em uma imagem e descrevem padrões locais relacionados à rugosidade, uniformidade e estrutura dos alvos na superfície terrestre (Dhakal et al., 2023). Esses padrões foram processados em cinco atributos de textura: Variância (var), Homogeneidade (hom), Entropia (ent), Contraste (con) e Correlação (cor). A opção por utilizar apenas cinco atributos se deu pela falta de recurso computacional para processar todas as imagens para todos os atributos. A Tabela 5 descreve as equações GLCM utilizadas neste estudo para cada atributo.

Tabela 5: Medidas de textura de imagem baseadas na matriz GLCM e suas equações de cálculo.

Medidas de textura	Equação
Variância	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i - u)^2 P(i, j)$
Homogeneidade	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} \frac{P(i, j)}{1 + i - j }$
Entropia	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} P(i, j) \log_2 P(i, j)$
Contraste	$\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i - j)^2 P(i, j)$
Correlação	$\frac{\sum_{i=0}^{Ng-1} \sum_{j=0}^{Ng-1} (i - \mu_2)(j - \mu_2)P(i, j)}{\sigma_i \sigma_j}$

Legenda: Equações dos oito atributos texturais extraídos pela Matriz de Co-ocorrência em níveis de Cinza. $P(i, j)$: Elemento da matriz GLCM, que representa a probabilidade de ocorrência do par de níveis de cinza i e j ; N_g : Número total de níveis de cinza; μ_i, μ_j : médias marginais de i e j ; σ_i, σ_j : desvios padrão marginais de i e j .

O processamento foi feito utilizando as quatro bandas (Blue = B1, Green = B2, Red = B3 e NIR = B4) de todas as datas coletadas, seguido do nome do atributo extraído e a data de aquisição da imagem. O tamanho da janela utilizada foi de 11x11 pixels com deslocamento angular de 45°. O número de cinza da função glcm (n_gray) foi fixado em 65.536 devido às imagens orbitais possuírem resolução radiométrica de 16 bits.

3.4 Dados tridimensionais

A nuvem de pontos foi obtida por meio de um sensor LiDAR acoplado a um helicóptero, operando a uma altitude média de aproximadamente 500 m em relação ao solo. O sistema, configurado em ângulo nadir para realizar a varredura das áreas de cana-de-açúcar, era composto por um sensor a laser (ALTM Gemini 56,100 OPTECH), receptores GNSS e uma unidade inercial (IMU), responsáveis pelo georreferenciamento dos dados do dossel das plantas com base em pontos de controle de validação. Uma IMU foi integrada para corrigir o erro de posicionamento no conjunto de dados de acordo com a oscilação da plataforma nos ângulos transversal, lateral e longitudinal.

A coleta dos dados foi conduzida em duas operações distintas no fim safra de 2018/2019 e medições manuais das plantas em campo foram executadas imediatamente após cada varredura a laser para validação. O primeiro levantamento ocorreu cerca de 10 dias antes da colheita, abrangendo os talhões previamente definidos pelos pontos de validação. O segundo levantamento foi realizado 10 dias após a colheita, utilizando o mesmo sistema, com o objetivo de gerar o Modelo Digital de Terreno (MDT) de referência.

O processamento da nuvem de pontos consistiu na geração de um Modelo digital de Terreno e de Superfície (MDT e MDS) com o objetivo de extrair, por meio da diferença entre o MDS e o MDT, o Modelo de altura de dossel (CHM) da cana-de-açúcar com resolução espacial de 0,5 x 0,5 m. Para esta etapa, foi utilizado o software estatístico RStudio (versão 2025.05.1+513), especificamente a função `grid_terrain` do pacote `lidR` (Roussel et al., 2020).

3.5 Dados de mineralogia do solo

A composição mineralógica do solo foi avaliada por meio da técnica de diagnóstico da suscetibilidade magnética, utilizando um sensor da marca Bartington. Toda a metodologia de análise da mineralogia, incluindo método de interpolação e geração de zonas de manejo, baseou-se na investigação dos processos pedogenéticos do solo (QUANTICUM, 2023). Os dados brutos obtidos a partir das leituras de suscetibilidade magnética foram processados e convertidos em um índice denominado “Índice de Tipologia da Argila” (ITA). Esse índice representa a proporção e a presença relativa de minerais com diâmetro inferior a 0,002 mm naturalmente presentes no solo. Esses minerais, classificados como fração argila, refletem indiretamente os fatores e os processos responsáveis pela formação do solo, sendo, portanto, indicativos de sua evolução pedogenética.

O indicador empregado neste estudo fundamenta-se na concepção proposta por Fontes et al. (1985), originalmente descrito como atributo diagnóstico para o mapeamento de séries de solos. Posteriormente, o índice foi incorporado a abordagens de zoneamento agrícola, sendo aplicado na delimitação de zonas de manejo e na identificação de áreas com distintos potenciais produtivos para a cultura da cana-de-açúcar (Catelan et al., 2022). Para a análise espacial, foi estabelecida uma malha regular de 50 × 50 m, utilizada para a extração dos valores do Índice de Terras Agrícolas (ITA). A interpolação foi conduzida pelo método do inverso da distância ponderada (IDW), resultando na definição de duas zonas de manejo, classificadas com base nos intervalos de valores do ITA.

- Zona 1: $ITA < 5$, associada a ambientes pedogenéticos com maior influência de processos redutores, em outras palavras, a estrutura dos minerais presentes neste tipo pode resultar em menor disponibilidade de nutrientes para as plantas;
- Zona 2: $ITA > 5$, associada a ambientes de formação mais oxidativos, indicando solos mais evoluídos quimicamente. Se bem manejado, a facilidade de disponibilizar nutrientes para as plantas é maior.

3.6 Análise estatística e remoção de outliers

As estatísticas descritivas foram realizadas no ambiente de programação Python (versão 3.12.7). Os outliers foram identificados e removidos com base na distribuição dos dados dentro de cada grupo experimental, considerando observações que excederam os limites estabelecidos pelo intervalo interquartilico.

3.7 Modelagem dos modelos preditivos

3.7.1 Geração do Banco de dados

O banco de dados utilizado para o treinamento dos modelos foi construído mediante o alinhamento de todos os raster pertencentes aos diferentes grupos de variáveis (Texturas, Índices de Vegetação, Reflectância, Mineralogia do Solo e Altura de Plantas). Em seguida, aplicou-se o plugin Point Sampling Tool (versão 0.5.5) do QGIS (Jurgiel, 2012), tomando como referência a grade amostral correspondente aos pontos de produtividade real (10 x 10 m), a partir da qual foram extraídos os valores de cada raster. Esse procedimento permitiu consolidar todas as informações em um único banco de dados estruturado. Ao final, o conjunto de dados resultante foi composto por 302 colunas de variáveis de entrada, que compreendem os dados de textura GLCM das 11 datas e das quatro bandas espectrais, os dados espectrais referentes às 11 datas, os índices de vegetação para cada período e os dados de suscetibilidade magnética e altura de plantas.

Para garantir a uniformidade dos resultados, a nomenclatura das variáveis foi padronizada da seguinte maneira: as métricas de textura foram identificadas pelas iniciais v (variância), cor (correlação), con (contraste), e (entropia) e h (homogeneidade) e os estádios fenológicos da cana-de-açúcar foram representados pelas letras I (Inicial), P (Perfilhamento), D (Desenvolvimento) e M (Maturação). As bandas espectrais foram codificadas como R, G, B e NIR, referentes ao vermelho, verde, azul e infravermelho próximo além dos IVs, que seguiram a nomenclatura original, indicada na Tabela 4. Essa padronização tornou o banco de dados mais organizado e facilitou a identificação das combinações entre atributos, estádios fenológicos e bandas empregadas na modelagem.

A organização das colunas seguiu uma estrutura padronizada conforme o grupo de variáveis. Para o grupo de textura, cada coluna foi nomeada combinando: (i) o atributo de textura (GLCM), (ii) o estágio fenológico e (iii) a banda espectral utilizada

no cálculo do atributo de textural (ex. v_D1_R). De forma semelhante, as variáveis de reflectância e IVs foram nomeadas iniciando pela banda espectral (R, G, B, NIR), seguida do estágio fenológico (ex. R_D1), assim como os IVs seguido do estágio de desenvolvimento (ex. $NDVI_D1$). As variáveis referentes à altura das plantas e mineralogia do solo foram nomeadas, respectivamente, como *chm* e *solo* e não foram seguidas de nenhum outro codificador pois não foram extraídas acompanhando os estádios de desenvolvimento da cultura.

3.7.2 Modelos de aprendizado de máquina

O banco de dados possui um total 5.435 pontos que compreende os quatro talhões avaliados. Estes dados foram divididos em 80% para o treinamento dos modelos e 20% para a etapa de teste. Os algoritmos de aprendizado de máquina (ML) Floresta Aleatória (RF), Regressão Linear Múltipla (RLM), XGBoost e K-Vizinhos mais próximos (KNN) foram utilizados para desenvolver os modelos preditivos de produtividade da cana-de-açúcar. Os métodos de ML foram implementados usando a biblioteca Scikit-learn em Python (versão 3.12.7). Para reduzir a multicolinearidade entre as variáveis independentes, aplicou-se o Índice de Inflação da Variância (VIF – *Variance Inflation Factor*), esse procedimento permitiu identificar e remover preditores altamente correlacionados.

Os modelos foram inicialmente testados com um banco de dados completo (302 variáveis de entrada) e, posteriormente, após a aplicação do VIF, com um subconjunto de 20 variáveis de entrada selecionadas pelo método interno de seleção de variáveis do *RandomForest* da biblioteca Scikit-learn, visando também à redução da dimensionalidade do banco de dados. A otimização dos hiperparâmetros foi conduzida pelo método de busca em grade (GridSearchCV) da biblioteca Scikit-learn (Pedregosa et al., 2011), empregando validação cruzada com cinco folds para garantir maior robustez na seleção e reduzir o risco de overfitting. O desempenho dos modelos foi avaliado e comparado por meio das métricas coeficientes de determinação (R^2), raiz do erro quadrático médio (RMSE) e erro percentual absoluto médio (MAPE).

Mapas temáticos com a produtividade predita dos quatro modelos foi gerado a partir do conjunto de teste e comparados ao mapa de produtividade real. Além disso, o erro relativo absoluto entre a produtividade real e predita também foi gerado para os quatro modelos utilizados aplicando a Equação 2:

$$Erro\ relativo = \frac{|pr - pp|}{pp} \times 100$$

Equação 2

Onde:

pr é a produtividade real

pp é a produtividade predita

O desenho metodológico desta pesquisa baseia-se em uma abordagem multivariada integrada que articula sistematicamente três componentes principais: (A) dados multiespectrais texturais de imagens de satélite, (B) dados de altura de plantas extraídos da nuvem de pontos (C) Dados relacionados a estrutura de argilas presentes na área experimental. Essa estrutura metodológica é ilustrada na Figura 6, que descreve desde a aquisição até a geração do produto no qual foi extraído os dados.

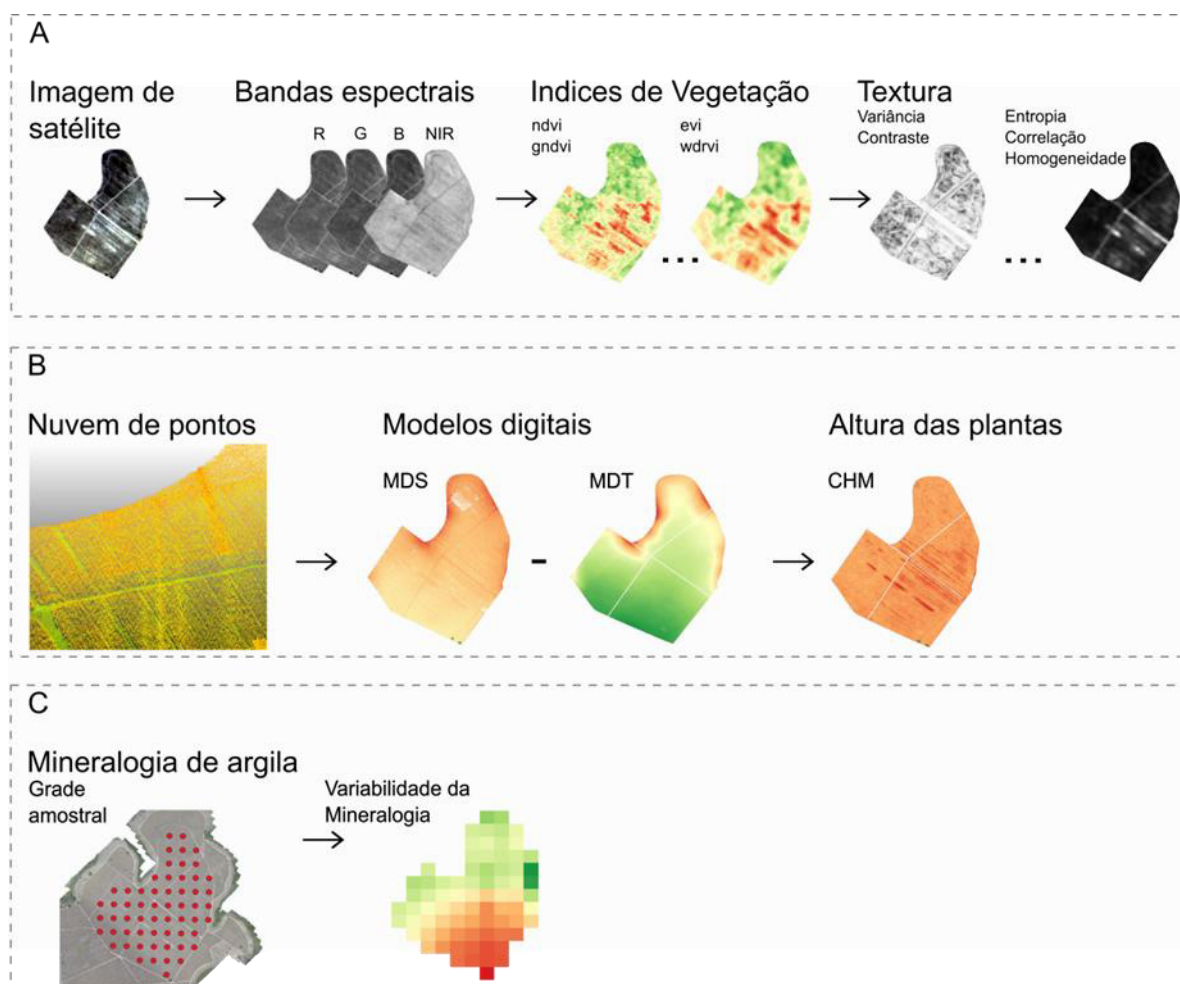


Figura 6: Dados obtidos da imagem de satélite: Bandas espectrais, índices de vegetação e textura (A); Dados obtidos a partir da nuvem de pontos do LiDAR: MDS, MDT e Altura de plantas (CHM) (B); Dados de mineralogia do solo com a grade amostral e variabilidade da área (C).

4 RESULTADOS

4.1 Análise estatística da produtividade

Antes da etapa de modelagem, foi realizada a análise descritiva da variável resposta a fim de caracterizar a distribuição dos dados utilizados neste estudo (Figura 7).

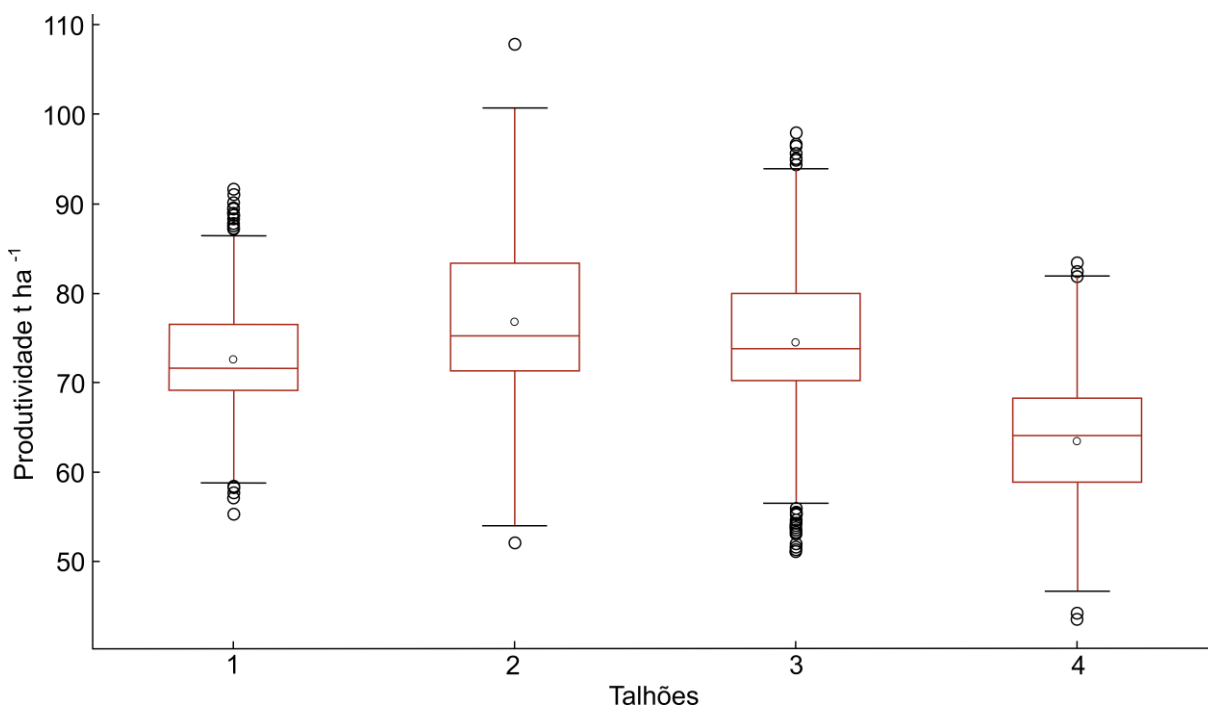


Figura 7: Boxplot da produtividade real por talhões.

A produtividade média entre os talhões variou de 63,33 t ha⁻¹ a 77,07 t ha⁻¹, apresentando diferenças de produtividade entre eles. Além disso, o talhão dois apresentou a maior variação (52,13 a 107,90 t ha⁻¹), por outro lado, o talhão quatro apresentou menor produtividade média e menor dispersão dos dados, refletida pelo menor intervalo de valores apresentado pelo boxplot do talhão.

4.2 Predição da Produtividade

4.2.1 Desempenho dos modelos com todas as variáveis de entrada

As métricas de desempenho dos modelos avaliados com o conjunto de variáveis sem redução da dimensionalidade e multicolinearidade do banco de dados podem ser visualizadas na Tabela 6. O XGBoost apresentou o melhor desempenho geral, com o maior R² e valores de erro menores (RMSE: 5,28 e MAPE 5,41%). O

modelo Random Forest teve desempenho semelhante ao XGBoost, com leve redução no valor do R^2 e um leve acréscimo na acurácia. Por outro lado, o modelo de Regressão Linear Múltipla apresentou os piores resultados.

Tabela 6: Desempenho dos modelos com todas as variáveis utilizando os dados de teste.

	Modelos	R^2	RMSE (t ha ⁻¹)	MAPE (%)
Todas as variáveis	KNN	0,58	5,82	5,93
	RF	0,61	5,56	5,83
	RLM	0,48	6,44	6,80
	XGboost	0,65	5,28	5,41

O modelo KNN obteve desempenho razoável ao utilizar todas as entradas, com erro absoluto médio pouco superior ao RF, mas com redução no RMSE em relação ao RLM. Nota-se que as diferenças entre os modelos não foram drásticas no RMSE, com variação de 1,32 t ha⁻¹ entre o maior e o menor valor. No entanto, os valores de R^2 indicam variações da variabilidade explicada na predição da produtividade.

4.2.2 Variáveis selecionadas a partir do Selection Features

Verificou-se a importância das variáveis utilizadas para prever os valores de produtividade nos quatro modelos utilizados (Figura 8). O maior impacto para predição foi ocasionado pela variância do nível de cinza (GLCM) da banda do vermelho no estágio inicial do perfilhamento para todos os modelos. Seguindo, o índice WDRI obtido no estágio inicial de desenvolvimento da cana-de-açúcar, no qual obteve a segunda maior relevância na predição. As demais, mostram um leve impacto em todos os modelos, concentrando em variáveis de textura extraídas da banda do infravermelho próximo e vermelho, variando do início do ciclo da cultura até o estágio de maturação.

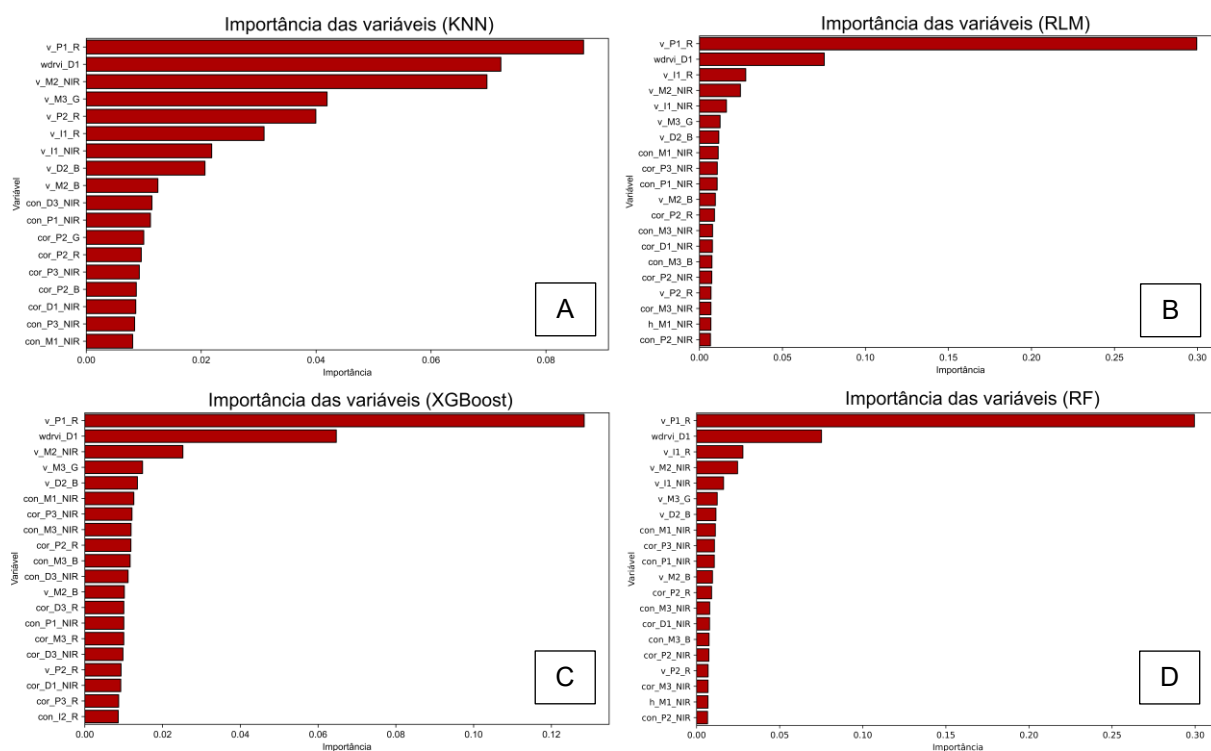


Figura 8: Impacto das 20 variáveis selecionadas na predição da produtividade para os modelos KNN (A), RLM (B), XGBoost (C) e RF (D).

4.2.3 Desempenho dos modelos com as variáveis selecionadas

Ao observar valores preditos após redução do banco de dados, os modelos baseados em árvores como o Random Forest e XGBoost, obtiveram os menores RMSE (5,61 e 5,53 t ha⁻¹) e os mesmos valores de MAPE (5,79%), além de apresentarem precisões melhores em relação aos outros modelos treinados, com de R² 0,61 e 0,62 respectivamente (Figura 9).

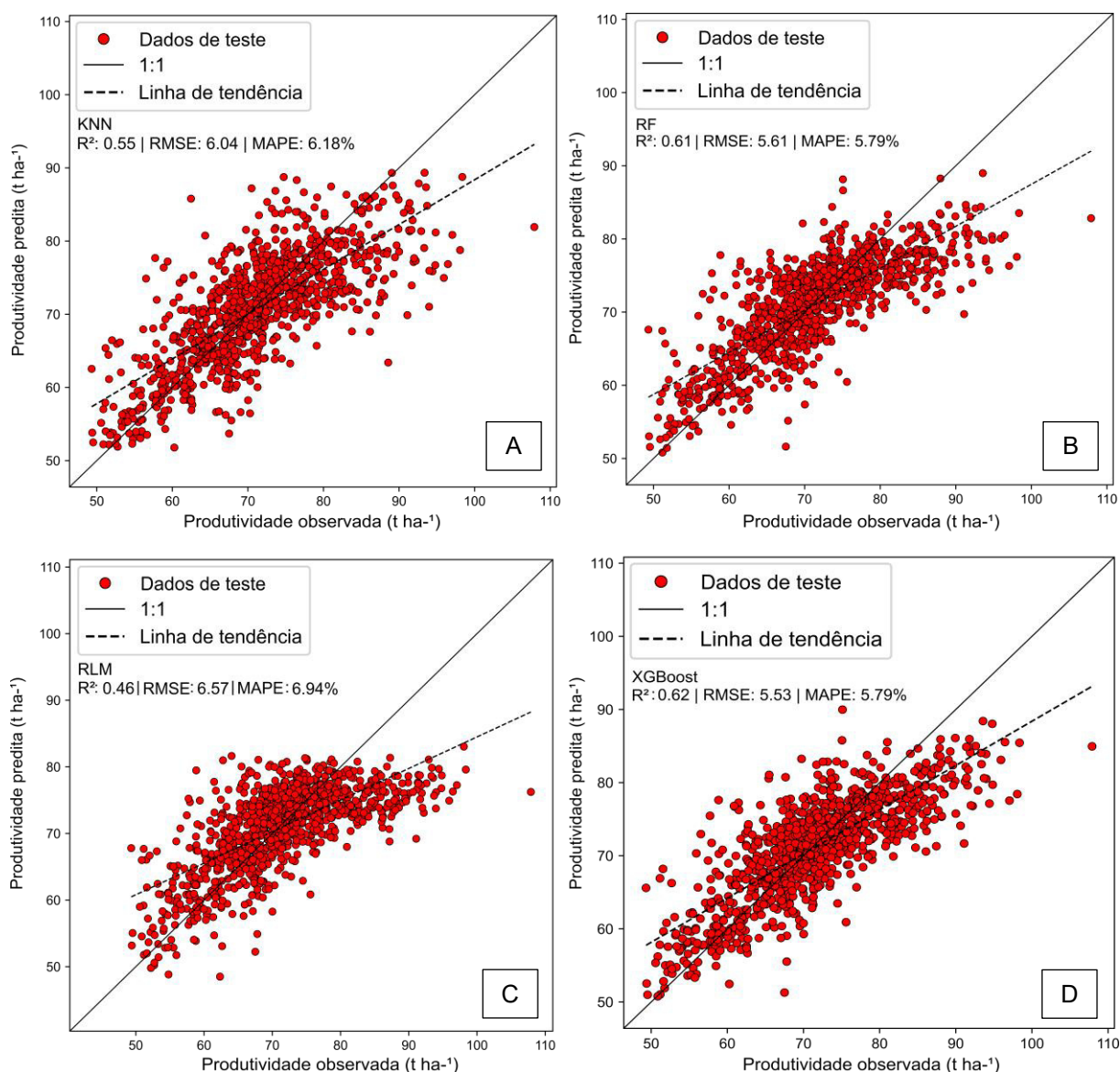


Figura 9: Desempenho dos modelos KNN (A), Random Forest (B), Regressão Linear Múltipla (C) e XGBoost (D) após a remoção da multicolinearidade e a seleção de variáveis, avaliados com base nos dados de teste.

O modelo de KNN após a redução do banco de dados pela seleção das variáveis, apresentou métricas inferiores ao comparar com os resultados apresentados na Tabela 6. O coeficiente de determinação passou por uma leve queda, com um poder explicativo que antes era de 0,58 para 0,55. Já as métricas de acurácia obtiveram um aumento para 6,04 t ha⁻¹ no RMSE e 6,18% no MAPE. Os modelos baseados em árvores de decisão mantiveram destaque superior quanto ao desempenho.

4.3 Mapas preditivos dos modelos

Considerando os resultados obtidos a partir dados preditos, os valores de produtividade predita pelos diferentes modelos foram interpolados para a geração dos mapas e comparados ao mapa de produtividade observada em campo (Figura 10). Os parâmetros dos semivariogramas ajustados aos mapas de produtividade predita são apresentados no APÊNDICE A.

De modo geral, os mapas derivados dos modelos RLM, RF, XGBoost e KNN (Figuras 10B–E) foram capazes de reproduzir os principais padrões de variabilidade espacial observados nos dados reais, especialmente no que se refere à distinção entre áreas de maior e menor produtividade dentro do talhão. Observa-se que todos os modelos capturaram a tendência espacial predominante, com maior concentração de valores elevados na porção norte da área, que corresponde ao talhão dois, e menores produtividades na região sul, que corresponde ao talhão quatro.

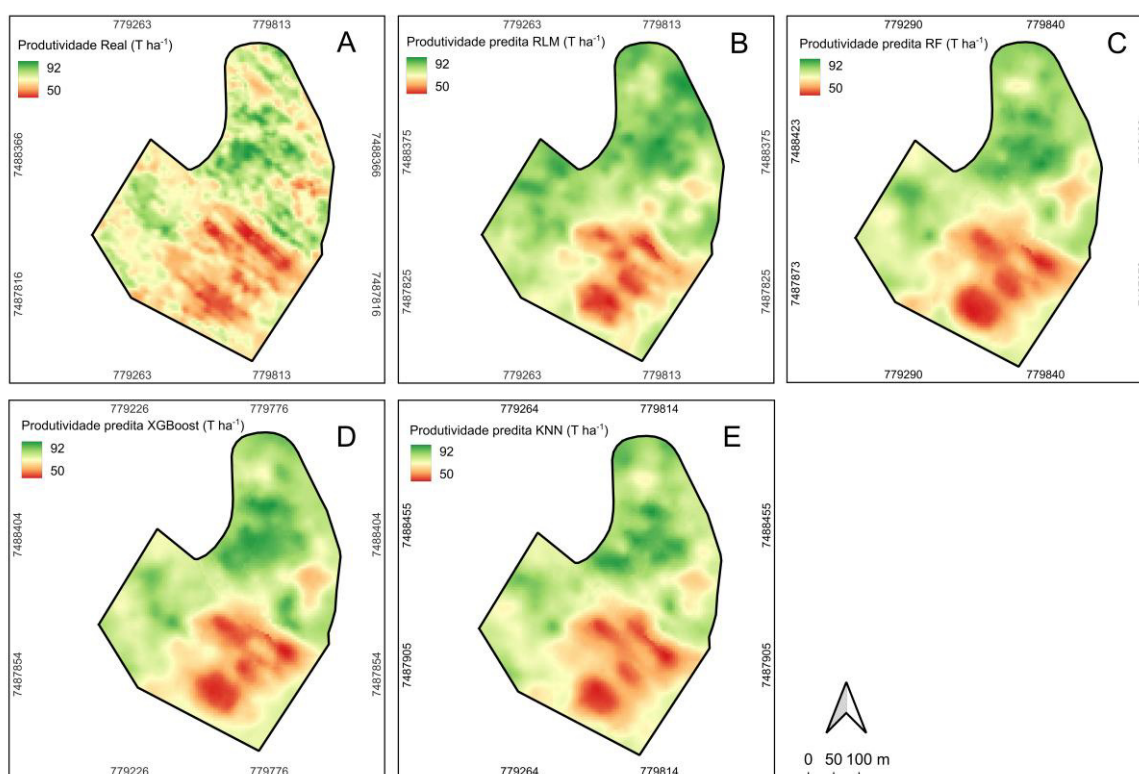


Figura 10: Mapa de produtividade real (A) e produtividade predita com as 20 variáveis selecionadas para os modelos RLM (B); RF (C); XGBoost (D) e KNN (E).

A Figura 11 apresenta os mapas de erro relativo associados aos modelos RLM, RF, XGBoost e KNN, permitindo avaliar a distribuição espacial das discrepâncias

entre os valores de produtividade preditos e observados. O modelo RLM (Figura 11 A) apresentou erros relativos que atingiram valores de até 22%. Já o modelo RF (Figura 11 B) exibiu comportamento semelhante, porém com valores máximos de erro ligeiramente superiores, alcançando cerca de 25%. Em contraste, o modelo XGBoost (Figura 11 C) apresentou os menores erros relativos máximos, limitados a aproximadamente 15%, além de uma distribuição espacial mais uniforme. Já o modelo KNN (Figura 11 D) apresentou erros relativos de até cerca de 23%.

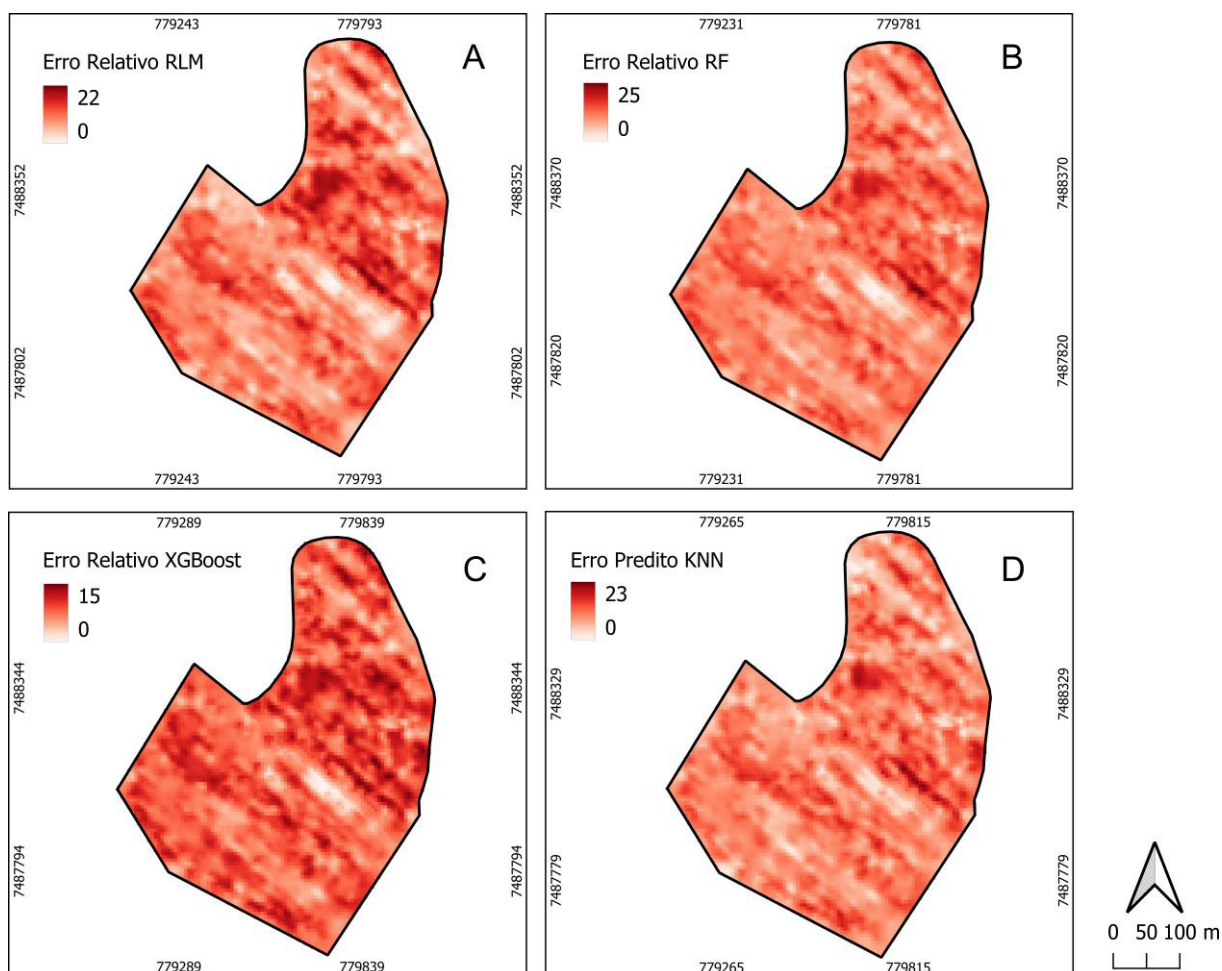


Figura 11: Mapa de erro percentual nos dos valores preditos com as 20 variáveis selecionadas para o modelo RLM (ARF (B); Mapa de XGBoost (C) e KNN (D).

5 DISCUSSÃO

Neste estudo, avaliou-se se produtos derivados de imagens de satélite, como a reflectância, índices de vegetação (IVs) e texturas, além do parâmetro biofísico, altura de plantas e atributos relacionados ao solo (Figura 6), são capazes de prever a

produtividade da cana-de-açúcar com antecedência à colheita. Comumente, verifica-se a aplicação de aprendizado de máquina para prever a produtividade da cana-de-açúcar levando como referência o seu estágio vegetativo (You et al., 2023; Zhang et al., 2025), integrando índices de vegetação, bandas espectrais e variáveis biométricas (Argolo dos Santos et al., 2025; Xu et al., 2020). Entretanto, os resultados obtidos neste estudo indicam que os atributos de textura GLCM apresentam expressiva capacidade preditiva da produtividade.

Os valores descritivos da produtividade indicados na Figura 7 apresentam observações importantes a respeito da variável dependente. Os valores de produtividade média revelam que os talhões um, dois e três possuem maior produtividade quando comparado com o talhão quatro. No entanto, ao observar os pontos extremos do boxplot, os maiores valores estão nos talhões que possuem maiores produtividades, sugerindo uma variabilidade maior dos dados em torno da média. Por fim, nota-se, principalmente pelos quartis, que há variabilidade espacial entre os talhões que serviram de base para os modelos preditivos utilizado neste estudo.

Nesse sentido, a Tabela 6 apresenta a primeira abordagem testada, em que os modelos treinados com todas as variáveis utilizadas como entrada, demonstram como os algoritmos respondem à alta dimensionalidade do banco de dados utilizado. A priori, o XGBoost apresentou o melhor desempenho geral, explicando 65% da variabilidade observada na produtividade, indicando um bom desempenho. Além disso, o Random Forest, embora um pouco inferior apresentou resultados similares ao XGBoost, confirmando a robustez de modelos baseados em árvores para conjuntos de dados heterogêneos. No entanto, é importante destacar que o uso de todas as variáveis no modelo pode introduzir redundância decorrente da multicolinearidade. Essa limitação também se aplica aos demais modelos testados, que, apesar de apresentarem arquiteturas distintas e métricas de desempenho satisfatórias, tornam-se pouco aplicáveis em função do elevado número de variáveis de entrada envolvidas.

Após a redução de dimensionalidade, apenas 20 variáveis foram selecionadas como entradas para os modelos treinados (Figura 8). As métricas de precisão e acurácia do modelo KNN foram pouco afetadas no seu desempenho. Da mesma forma, os modelos baseados em árvore de decisão, que não sofreram redução

considerável em seus coeficientes de determinação (Figura 9). A não redução tão acentuada do R^2 , pode indicar que somente com as variáveis selecionadas, indicadas na Figura 8, o modelo não sofreu uma penalização no que tange a explicar a variabilidade da produtividade, a partir eliminação das outras variáveis (Goyal, 2021) que por sua vez, podem não ter sido tão influentes na predição.

Entre as variáveis descartadas, destacam-se os dados de altura de planta e de mineralogia do solo, que não apresentaram relevância na predição. Embora os dados provenientes do LiDAR apresentem reconhecido potencial para estimativas de biomassa (Sofonia et al., 2019; Xu et al., 2020), sua limitação neste estudo pode estar associada ao momento da aquisição, que não contemplou todos os estádios fenológicos da cultura. De modo semelhante, os dados de mineralogia, ainda que relacionados a fatores de causa e efeito sobre o rendimento da cana-de-açúcar, possivelmente não contribuíram para a predição devido à diferença de resolução espacial em relação às demais variáveis utilizadas no modelo.

A capacidade dos modelos XGboost e RF lidarem com variáveis de entrada semelhantes ao deste estudo, já foi descrita por outros autores ao utilizarem principalmente dados de sensoriamento remoto (Militino et al., 2024). Além disso, Trentin et al., (2025), ao utilizarem os modelos de mesma arquitetura, chegaram a valores de R^2 semelhantes aos apresentados por este estudo em seu trabalho feito com a cultura do tomate. Em contrapartida, a Regressão Linear múltipla, não conseguiu desempenhar bem na predição. Uma hipótese para essa resposta pode estar relacionada arquitetura do modelo, que é considerado menos flexível no que tange à interpretação de dados espaciais com características não lineares, comportamento observado por Barbosa Júnior et al., (2023) ao utilizar o RLM para estimativa de Brix em cana-de-açúcar.

Ao observar as variáveis de maior importância para os modelos testados (Figura 8), verificamos que as texturas da imagem de satélite, obtiveram maior poder preditivo. Conceitualmente, as texturas não representam medidas físicas diretas, descrevem a disposição espacial dos níveis de cinza em imagens geradas por operadores matemáticos (Huang et al., 2014). Assim, o valor numérico calculado por uma janela móvel varia conforme o deslocamento e indica diferenças de superfície em imagens de satélite (Hall-Beyer, 2017). Em coberturas vegetais densas e relativamente homogêneas, variáveis derivadas da GLCM (*Gray-Level Co-occurrence*

Matrix) mostraram-se úteis como preditores em modelos de aprendizado de máquina (Rao et al., 1979).

Sendo assim, os atributos de textura acrescentam informação complementar aos índices de vegetação e têm melhorado o desempenho de modelos preditivos (Nogueira Martins et al., 2023; Shen et al., 2024). Nesse sentido, é plausível inferir que as características de textura desempenharam um papel importante na previsão da produtividade da cana-de-açúcar vistos neste estudo (Figura 8). Grande parte das variáveis utilizadas são derivadas de texturas, com destaque para a variância ($v_{_}$), que mede a variabilidade dos níveis de cinza e das bordas visuais (Freitas et al., 2024), a correlação ($cor_{_}$), que permite distinguir padrões de classes (Hall-Beyer, 2017), e o contraste ($con_{_}$), que avalia a variação local, ou seja, a mudança nas vizinhanças dos pixels (Freitas et al., 2024). No estudo de You et al. (2023), os autores utilizaram modelos como o Random Forest e obtiveram estimativas relevantes ao integrar variáveis de textura na previsão de nitrogênio em cana-de-açúcar nos estádios de perfilhamento e desenvolvimento. Esses resultados são parcialmente consistentes com os nossos, já que, além dos estádios mencionados, também destacamos o período de maturação para a nossa predição. O atributo correlação GLCM também foi destacado por Freitas et al. (2022) em seu estudo com pastagens, no qual os autores concluem que essa variável foi um bom preditor de biomassa, especialmente quando combinada com o atributo de média GLCM. Esse achado pode ser explicado pela semelhança na estrutura do dossel vegetativo da cana-de-açúcar, que, assim como as pastagens, apresenta uma superfície heterofílica em estádios mais avançados de desenvolvimento e maturação.

Em contrapartida às variáveis de textura, uma possível explicação para a relevância do único índice de vegetação (WDRVI) na predição da produtividade pode estar relacionada à sua maior sensibilidade em relação ao NDVI, sobretudo em dosséis mais densos (Tian et al., 2025), como o da cana-de-açúcar. Proposto por Gitelson (2004), o WDRVI reduz o efeito de saturação observado em índices tradicionais, permitindo capturar sutis variações fisiológicas e fenológicas ao longo do ciclo da cultura. Essa característica pode ter contribuído para que o índice se destacasse entre as variáveis dos modelos, refletindo melhor as diferenças estruturais do dossel. Resultados semelhantes foram relatados por Zhang et al. (2025), que observaram um bom desempenho ($R^2 = 0,77$) ao empregar o WDRVI combinado a

outros índices em modelos de com arquitetura boosting, sugerindo que esse índice pode representar um bom preditor de produtividade em estádios de desenvolvimento da cana-de-açúcar.

Observa-se que os estádios com maior influência na predição da produtividade se concentram entre o perfilhamento e a maturação (Figura 8). Além disso, de forma consistente, as texturas derivadas das bandas do vermelho e do infravermelho próximo apresentaram maior relevância, o que é fisiologicamente plausível, uma vez que a banda do vermelho está diretamente associada à absorção de luz pela clorofila e à atividade fotossintética, enquanto a banda do infravermelho próximo apresenta baixa absorção da energia incidente e elevada refletância, característica de dosséis saudáveis (Ayyangar; Rao; Rao, 1980; Vicardi, 2015).

Ao analisarmos as cinco primeiras variáveis nos modelos KNN e XGBoost (Figura 8), nota-se também o impacto da banda do verde (v_M3_G) e da banda do azul (v_D2_B). Barbosa Junior et al., (2024), em seu trabalho com predição de parâmetros de qualidade, detectaram a importância da banda do azul e associaram isso a fatores fisiológicos e atividade fotossintética ligados ao fitocromo à medida que a planta se desenvolve. Neste estudo, a banda do azul no estágio de desenvolvimento foi onde mais impactou na predição do XGBoost. Outra perspectiva tratada por Canata et al., (2024), associa a influência dessa banda com a síntese do açúcar na planta. Barbosa Junior et al., (2024), também discute sobre a associação da banda do verde com fatores ligados a celulose da planta e acúmulo de açúcar no caule, sugerindo uma provavelmente explicação sobre o porquê da importância da variável v_M3_G na predição da produtividade, exatamente no estágio de maturação, período que há maior acúmulo de açúcar nos colmos (Aude, 1993).

De maneira geral, observa-se um bom comparativo ao analisar a produtividade real comparada a predita pelo modelo de aprendizado de máquina (Figura 10), com os erros relativos (Figura 11) não se distribuem de forma homogênea ao longo do talhão. Os mapas evidenciam que, embora todos os modelos sejam capazes de reproduzir os padrões espaciais gerais da produtividade comparado ao mapa de produtividade real (Figura 10 - A), há diferenças importantes na magnitude e na organização espacial dos erros, refletindo distintas capacidades de representação da variabilidade entre os talhões. O mapa de produtividade predita evidencia que o modelo XGBoost (Figura 10 – D) conseguiu reproduzir a variabilidade espacial da

produtividade entre os talhões com menor erro relativo, na casa dos 15%. Os valores de RMSE e MAPE indicam boa acurácia na precisão dos valores preditos, indicando um potencial para representar a variabilidade espacial da produtividade de forma antecipada a colheita, sendo uma alternativa promissora no que tange o uso de sensores embarcados e métodos destrutivos para extrair produtividade em talhões.

6 CONCLUSÕES

Foi possível prever a produtividade da cana-de-açúcar com o modelo XGBoost apresentando o melhor desempenho entre os modelos avaliados, explicando 65% da variabilidade da produtividade e apresentando erro médio de 5,41%. Dentre as variáveis utilizadas, os resultados indicam que os atributos de textura derivados da Matriz de Co-ocorrência em Níveis de Cinza (GLCM) foram os principais responsáveis pela capacidade preditiva do modelo, superando variáveis relacionadas à altura das plantas e à mineralogia da fração argila. Os mapas preditos foram capazes de representar a produtividade real com erro relativo em torno de 15%, indicando boa concordância espacial entre valores observados e preditos.

REFERÊNCIAS

- AKBARIAN, Sharareh *et al.* Sugarcane yields prediction at the row level using a novel cross-validation approach to multi-year multispectral images. **Computers and Electronics in Agriculture**, v. 198, 2022. <https://doi.org/10.1016/j.compag.2022.107024>
- ALMALKI, Raid *et al.* Monitoring and mapping vegetation cover changes in arid and semi-arid areas using remote sensing technology: a review. **Remote Sensing**, v. 14, n. 20, p. 5143, 2022. <https://doi.org/10.3390/rs14205143>
- ALMEIDA, Alexsandro Claudio dos Santos *et al.* Desenvolvimento vegetativo e produção de variedades de cana-de-açúcar em relação à disponibilidade hídrica e unidades térmicas. **Ciência e Agrotecnologia**, v. 32, n. 5, p. 1441-1448, 2008. <https://doi.org/10.1590/S1413-70542008000500013>
- AMARO, Rafaella Pironato *et al.* Regional Model to Predict Sugarcane Yield Using Sentinel-2 Imagery in São Paulo State, Brazil. **Sugar Tech**, v. 27, n. 1, p. 108–118, 1 fev. 2025. <https://doi.org/10.1007/s12355-024-01468-z>
- ARGOLO DOS SANTOS, R.; MANTOVANI, E. C.; BUFON, V. B.; FERNANDES-FILHO, E. I. Sugarcane evapotranspiration based on machine learning with remote sensing approaches using SAR sensor and NASA POWER data. **International Journal of Remote Sensing**, 2025. <https://doi.org/10.1080/01431161.2025.2513563>
- AUDE, M. I. da S. Estádios de desenvolvimento da cana-de-açúcar e suas relações com a produtividade. **Ciência Rural**, v. 23, n. 2, p. 241-248, 1993. <https://doi.org/10.1590/S0103-84781993000200022>
- AYYANGAR, R. S.; RAO, P. P. N.; RAO, K. R. Crop cover and crop phenological information from red and infrared spectral responses. **Indian Society of Photo-Interpretation & Remote Sensing**, v. 8, n. 1, 1980.
- BARBOSA JÚNIOR, M. R. *et al.* Integrated sensing and machine learning: predicting saccharine and bioenergy feedstocks in sugarcane. **Industrial Crops and Products**, v. 215, p. 118627, 2024. <https://doi.org/10.1016/j.indcrop.2024.118627>
- BARBOSA JÚNIOR, M. R. *et al.* UAV imagery data and machine learning: a driving merger for predictive analysis of qualitative yield in sugarcane. **Frontiers in Plant Science**, v. 14, 2023. <https://doi.org/10.3389/fpls.2023.1114852>
- BARRIOS, M. D. O. S. R. *et al.* Magnetic susceptibility as indicator of soil quality in sugarcane fields. **Revista Caatinga**, v. 30, n. 2, p. 287-295, 2017. <https://doi.org/10.1590/1983-21252017v30n203rc>

BARROS, Pedro Paulo da Silva et al. Monitoramento fitossanitário utilizando sensoriamento remoto: avanços e desafios. **Revista Brasileira de Cartografia**, v. 73, n. 2, p. 489-515, 2021.

BOTERO-VALENCIA, Juan et al. Machine learning in sustainable agriculture: systematic review and research perspectives. **Agriculture**, v. 15, n. 4, p. 377, 2025. <https://doi.org/10.3390/agriculture15040377>

BREIMAN, Leo. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

BRITO, Wildson Benedito Mendes et al. Usos e potencialidades das medidas de suscetibilidade magnética nos solos da Amazônia: uma revisão. **Revista Valore**, v. 7, 2022.

BRUNINI, Rodrigo G.; TURCO, José E. P. Water stress indices for the sugarcane crop on different irrigated surfaces. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 20, n. 10, p. 925-929, 2016. <https://doi.org/10.1590/1807-1929/agriambi.v20n10p925-929>

CANATA, T. F. et al. 3D data processing to characterize the spatial variability of sugarcane fields. **Sugar Tech**, v. 24, p. 419-429, 2022. <https://doi.org/10.1007/s12355-021-01048-5>

CANATA, T. F. et al. AI-driven prediction of sugarcane quality attributes using satellite imagery. **Sugar Tech**, v. 26, n. 3, p. 741-751, 2024. <https://doi.org/10.1007/s12355-024-01399-9>

CANATA, T. F.; WEI, M. C. F.; MALDANER, L. F.; MOLIN, J. P. Sugarcane yield mapping using high-resolution imagery data and machine learning technique. **Remote Sensing**, v. 13, n. 2, p. 232, 2021. <https://doi.org/10.3390/rs13020232>

CATELAN, M. G. et al. Sugarcane yield and quality using soil magnetic susceptibility. **Scientia Agrícola**, v. 79, n. 4, 2022. <https://doi.org/10.1590/1678-992X-2020-0329>

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Acompanhamento da safra brasileira de cana-de-açúcar**. Brasília: CONAB, 2019.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Acompanhamento da safra brasileira de cana-de-açúcar**. Brasília: CONAB, 2025.

DE OLIVEIRA, Romario Porto *et al.* Predicting Sugarcane Biometric Parameters by UAV Multispectral Images and Machine Learning. **AGRONOMY-BASEL**, v. 12, n. 9, set. 2022. <https://doi.org/10.3390/agronomy12091992>

DHAKAL, Rakshya *et al.* Utilizing Spectral, Structural and Textural Features for Estimating Oat Above-Ground Biomass Using UAV-Based Multispectral Data and Machine Learning. **Sensors**, v. 23, n. 24, 1 dez. 2023. <https://doi.org/10.1016/j.micpro.2021.103822>

DINIZ, Augusto L. *et al.* Genomic resources for energy cane breeding in the post-genomics era. **Computational and Structural Biotechnology Journal**, v. 17, p. 1404-1414, 2019. <https://doi.org/10.1016/j.csbj.2019.10.006>

EITEL, Jan U. H. *et al.* Beyond 3-D: the new spectrum of LiDAR applications for earth and ecological sciences. **Remote Sensing of Environment**, v. 186, p. 372-392, 2016. <https://doi.org/10.1016/j.rse.2016.08.018>

EVERINGHAM, Y. *et al.* Accurate prediction of sugarcane yield using a random forest algorithm. **Agronomy for Sustainable Development**, v. 36, p. 27, 2016. <https://doi.org/10.1007/s13593-016-0364-z>

FIGUEROA-RODRÍGUEZ, Katia A. *et al.* What Has Been the Focus of Sugarcane Research? A Bibliometric Overview. **International Journal of Environmental Research and Public Health**, v. 16, n. 18, p. 3326, 10 set. 2019. <https://doi.org/10.3390/ijerph16183326>

FORMAGGIO, Antonio Roberto; SANCHES, Ieda Del'Arco. **Sensoriamento remoto em agricultura**. São Paulo: Oficina de Textos, 2017.

GITELSON, Anatoly A. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. **Journal of Plant Physiology**, v. 161, 2004.

GITELSON, Anatoly A.; KAUFMAN, Yoram J.; MERZLYAK, Mark N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. **Remote Sensing of Environment**, v. 58, n. 3, p. 289-298, 1996.

HARALICK, R. M.; SHANMUGAM, K.; DINSTEN, I. Textural features for image classification. **IEEE Transactions on Systems, Man, and Cybernetics**, v. SMC-3, n. 6, p. 610-621, 1973. <https://doi.org/10.1109/TSMC.1973.4309314>

HUETE, A. *et al.* Overview of the radiometric and biophysical performance of the MODIS vegetation indices. **Remote Sensing of Environment**, v. 83, n. 1-2, p. 195-213, 2002.

KUCHLER, Patrick Calvano. Utilização do sensoriamento remoto para o mapeamento dos sistemas integrados de produção agrícola: contribuição ao monitoramento da agricultura de baixa emissão de carbono no estado do Mato Grosso, Brasil. 2021.

KHOJASTEHNZHAND, Mostafa; NASIRI, Sajjad; KAZEMI, Amir. Detection of Tomato Disease Using Texture Features and Genetic Bee Colony Algorithm. **SN Computer Science**, v. 6, n. 5, 1 jun. 2025.

LEE, Chun Han; CHEN, Kuang Yu; LIU, Li yu Daisy. Effect of Texture Feature Distribution on Agriculture Field Type Classification with Multitemporal UAV RGB Images. **Remote Sensing**, v. 16, n. 7, 1 abr. 2024. <https://doi.org/10.3390/rs16071221>

LIMA, MARCELO AGUIAR COSTA *et al.* Morfogênese in vitro e susceptibilidade de calos de variedades nacionais de cana-de-açúcar (*Saccharum officinarum* L.) a agentes seletivos utilizados em sistemas de transformação genética. **Revista Brasileira de Botânica**, v. 24, n. 1, p. 73–77, mar. 2001. <https://doi.org/10.1590/S0100-84042001000100008>

LOFTON, J. *et al.* Estimating sugarcane yield potential using an in-season determination of normalized difference vegetative index. **Sensors**, v. 12, n. 6, p. 7529–7547, 2012. <https://doi.org/10.3390/s120607529>

MALDANER, L. F.; MOLIN, J. P. Data processing within rows for sugarcane yield mapping. **Scientia Agricola**, v. 77, n. 5, e20180391, 2020. <https://doi.org/10.1590/1678-992X-2018-0391>

MENESES, PAULO ROBERTO; TATI DE ALMEIDA; EDSON EYJI SANO. **Processamento de imagens de sensoriamento remoto**. 1. ed. Brasília: [S.n.].

MORIWAKI, Thaise *et al.* Chloroplast and outside-chloroplast interference of light inside leaves. **Environmental and Experimental Botany**, v. 208, p. 105258, abr. 2023. <https://doi.org/10.1016/j.envexpbot.2023.105258>

MUKHAMEDIEV, Ravil I. *et al.* Review of artificial intelligence and machine learning technologies: classification, restrictions, opportunities and challenges. **Mathematics**, v. 10, n. 15, p. 2552, 2022. <https://doi.org/10.3390/math10152552>

ODAIR JOSÉ KUHN *et al.* **CIÊNCIAS AGRÁRIAS Tecnologias e Perspectivas**. 1. ed. Paraná: [S.n.].

PELUCO, Rafael Gonçalves *et al.* Suscetibilidade magnética do solo e estimação da capacidade de suporte à aplicação de vinhaça. **Pesquisa Agropecuária Brasileira**, v. 48, n. 6, p. 661–672, jun. 2013.

PICOLI, Michelle Cristina Araujo *et al.* Vegetation index from MODIS sensor to estimate sugarcane yield. **Bragantia**, v. 68, n. 3, p. 789–795, 2009. <https://doi.org/10.1590/S0006-87052009000300028>

PONZONI, Flávio Jorge; SHIMABUKURO, Yosio Edemir; KUPLICH, Tatiana Mora. **Sensoriamento remoto da vegetação**. 2. ed. São Paulo: Oficina de Textos, 2012.

RIVERA, Gilberto et al. LiDAR applications in precision agriculture for cultivating crops: a review of recent advances. **Computers and Electronics in Agriculture**, v. 207, p. 107737, 2023. <https://doi.org/10.1016/j.compag.2023.107737>

SHENDRYK, Yuri et al. Fine-scale prediction of biomass and leaf nitrogen content in sugarcane using UAV LiDAR and multispectral imaging. **International Journal of Applied Earth Observation and Geoinformation**, v. 92, p. 102177, 2020. <https://doi.org/10.1016/j.jag.2020.102177>

SHENDRYK, Yuri; DAVY, Robert; THORBURN, Peter. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. **Field Crops Research**, v. 260, 2021. <https://doi.org/10.1016/j.fcr.2020.107984>

SHIKIDA, Pery. Expansão canavieira no Centro-Oeste: limites e potencialidades. v. ano XXII, p. 122–137, jan. 2013.

SOFONIA, Jeremy et al. Monitoring sugarcane growth response to varying nitrogen application rates: a comparison of UAV SLAM LiDAR and photogrammetry. **International Journal of Applied Earth Observation and Geoinformation**, v. 82, 2019. <https://doi.org/10.1016/j.jag.2019.05.011>

TIAN, Xianchao *et al.* Evaluating the sensitivity of vegetation indices to leaf area index variability at individual tree level using multispectral drone acquisitions. **Agricultural and Forest Meteorology**, v. 364, p. 110441, 1 abr. 2025. <https://doi.org/10.1016/j.agrformet.2025.110441>

VERA-ESMERALDAS, Adrián *et al.* UAV-Based Spectral and Thermal Indices in Precision Viticulture: A Review of NDVI, NDRE, SAVI, GNDVI, and CWSI. **Agronomy**, v. 15, n. 11, p. 2569, 7 nov. 2025. <https://doi.org/10.3390/agronomy15112569>

VIRANI, V. B.; KUMAR, Neeraj; MOTE, B. M. Integration of Remote Sensing and Meteorological Data for Rapid Sugarcane Yield Estimation Using Machine Learning. **Journal of the Indian Society of Remote Sensing**, v. 53, n. 4, p. 1109–1124, 2025. <https://doi.org/10.1007/s12524-024-02066-y>

WANG, M. et al. Mapping sugarcane in complex landscapes by integrating multi-temporal Sentinel-2 images and machine learning algorithms. **Land Use Policy**, v. 88, 2019. <https://doi.org/10.1016/j.landusepol.2019.104190>

WANG, Pengwen; HAFSHEJANI, Behzad Aalipur; WANG, Daluyo. An improved multilayer perceptron approach for detecting sugarcane yield production in IoT based smart agriculture. **Microprocessors and Microsystems**, v. 82, 2021. 10.1016/j.micpro.2021.103822

WATERS, Ethan Kane; CHEN, Carla Chia Ming; RAHIMI AZGHADI, Mostafa. Sugarcane health monitoring with satellite spectroscopy and machine learning: A review. **Computers and Electronics in Agriculture**, v. 229, p. 109686, 1 fev. 2025. <https://doi.org/10.1016/j.compag.2024.109686>

WEI, M. C. F. et al. **Carrot yield mapping: a precision agriculture approach based on machine learning**. *AI*, v. 1, n. 2, p. 229-241, 2020. <https://doi.org/10.3390/ai1020015>

XU, J. X. et al. Estimation of sugarcane yield using a machine learning approach based on UAV-LiDAR data. **Remote Sensing**, v. 12, n. 17, 2020. <https://doi.org/10.3390/rs12172823>

YOU, H. et al. Sugarcane nitrogen nutrition estimation with digital images and machine learning methods. **Scientific Reports**, v. 13, 2023. <https://doi.org/10.1038/s41598-023-42190-2>

ZHANG, S. et al. Sensitive multispectral variable screening method and yield prediction models for sugarcane based on gray relational analysis and correlation analysis. **Remote Sensing**, v. 17, n. 12, 2025. <https://doi.org/10.3390/rs17122055>

APÊNDICE

APÊNDICE A

Observou-se dependência espacial moderada para todos os modelos. O alcance espacial foi semelhante entre os modelos (395–432 m), com o ajuste do modelo teórico Linear com patamar que apresentou R^2 elevado ($>0,995$) para todos os casos, embora o RMSE tenha sido menor para RLM (12,06).

Caracterização dos modelos teóricos de semivariograma aplicados à interpolação dos mapas de produtividade predita

Modelo	Modelo Teórico	Pepita (C_0)	Patamar ($C_0 + C$)	($C_0 / C_0 + C$)	Alcance (m)	RMSE	R^2
RLM	Linear com patamar	10,648	66,881	66.9	424,855	26.909	0.997
RF	Linear com patamar	7,36	49,151	49.2	432,451	18.663	0.996
XGBoost	Linear com patamar	2,489	38,587	38.6	411,086	12.058	0.996
KNN	Linear com patamar	8,22	53,375	53.4	395,669	27.539	0.995

Legenda: C_0 = efeito pepita; $C_0 + C$ = patamar; ($C_0 / C_0 + C$) = proporção pepita/patamar; Alcance = distância máxima de dependência espacial.