PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
ESCOLA POLITÉCNICA E DE ARTES
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MODELAGEM DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR COM MACHINE LEARNING PARA PREDIÇÃO DE PRODUÇÃO

DIONY TARSO FERREIRA FILHO

GOIÂNIA
2025

DIONY TARSO FERREIRA FILHO

MODELAGEM DA PRODUTIVIDADE DA CANA-DE-AÇÚCAR COM MACHINE LEARNING PARA PREDIÇÃO DE PRODUÇÃO

Trabalho para conclusão de curso apresentado na faculdade de Ciências Exatas e da Computação da Pontifícia Universidade Católica de Goiás como requisito básico para a conclusão do curso de Ciência da Computação.

Orientador: Prof. Me. Gustavo Siqueira Vinhal

GOIÂNIA
2025

DIONY TARSO FERREIRA FILHO

UTILIZAÇÃO DA IA MODELANDO A PRODUTIVIDADE DA CANA-DE-AÇÚCAR COM MACHINE LEARNING PARA PREDIÇÃO DA PRODUÇÃO

Este Trabalho de Conclusão de Curso foi julgado adequado para a obtenção do título de Bacharel em Ciência da Computação, e aprovado em sua forma final pela escola de Ciências Exatas e da Computação, da Pontifícia Universidade Católica de Goiás em _____/_____/_____.

Banca examinadora:

_____
Prof. Me. Gustavo Siqueira Vinhal

_____
Prof. Me. Carlos Alexandre Ferreira de Lima

_____
Prof. Dr. Nilson Cardoso Amaral

GOIÂNIA
2025

Dedico este trabalho à toda minha família, mas em especial ao meu falecido pai Diony Tarso Ferreira, e a minha querida mãe Maria de Fátima de Lima.

AGRADECIMENTOS

Ao meu pai Diony Tarso Ferreira, que com muito esforço e dedicação despertou em mim e em meus irmãos, amigos, colegas de trabalho e faculdade e especialmente as centenas de alunos que passaram sob sua tutela enquanto a luz de sua existência pode afastar a escuridão onde ele passava. Todos nós somos eternamente gratos a Deus ter nos concedido um anjo de sua guarda pessoal como nosso guia e catalisador do despertar consciente nesse plano material.

A todos os meus professores que durante minha jornada acadêmica tiveram uma grande parcela de contribuição em minha formação acadêmica e pessoal.

Ao meu orientador acadêmico, professor Me. Gustavo Siqueira Vinhal, pelo apoio, paciência e confiança no desenvolvimento deste trabalho.

Aos meus familiares, amigos e pessoas que direta ou indiretamente contribuíram com a minha trajetória, os quais trouxeram tantas cores à tela de minha vida, sendo muitas delas cores, as quais eu mesmo não as conhecia.

Sou grato a todo o corpo docente, à direção e administração desta universidade.

E um agradecimento especial à inspiração que trouxe vida a este projeto antes mesmo dele existir no papel, oferecendo-me o propósito e a força necessários para concretizá-lo.

“No código da vida, o erro não é o fim da caminhada, mas um convite para a refatoração da alma. Compile, falhe, sangre, ajuste e fortaleça. Viva para assistir, em primeira pessoa, a vitória: a dádiva das conquistas da sua própria existência".

Diony Tarso Ferreira Filho

RESUMO

Este Trabalho de Conclusão de Curso tem como objetivo desenvolver e avaliar um pipeline computacional para estimar a produtividade da cana-de-açúcar por meio da integração entre dados públicos de sensoriamento remoto, variáveis meteorológicas e técnicas de aprendizado de máquina. A motivação surge da necessidade de métodos acessíveis, robustos e independentes de bases privadas, que permitam realizar estimativas confiáveis mesmo diante da indisponibilidade de dados experimentais específicos. No presente estudo, informações derivadas de índices espectrais, como o NDVI obtido a partir do Sentinel-2, são combinadas com séries históricas de clima provenientes de bases como ERA5 e CHIRPS. Modelos de aprendizado de máquina são utilizados para construir predições de produtividade com base nessas variáveis. Embora o simulador biofísico APSIM-Sugar não seja executado integralmente por limitações de acesso aos dados originais da área experimental, sua estrutura conceitual fundamenta a seleção das variáveis e a interpretação agronômica do comportamento da cultura. Os resultados obtidos demonstram que o algoritmo implementado é capaz de operar de forma consistente, confirmando a viabilidade da utilização de dados públicos para estudos preliminares de previsão de produtividade da cana-de-açúcar.

Palavras-chave: Cana-de-açúcar. Sensoriamento remoto. *Machine Learning*. Produtividade.

ABSTRACT

This Final Course Project aims to develop and evaluate a computational pipeline to estimate sugarcane productivity by integrating public remote sensing data, meteorological variables, and machine learning techniques. The motivation lies in the need for accessible and robust methods that remain independent of private datasets, enabling reliable estimates even when experimental field data are not available. In this study, information derived from spectral indices such as NDVI from Sentinel-2 is combined with climate time series from datasets like ERA5 and CHIRPS. Machine learning models are applied to generate productivity predictions based on these variables. Although the APSIM-Sugar biophysical simulator is not fully executed due to the unavailability of original field data, its conceptual structure guides both variable selection and agronomic interpretation of crop behavior. The results show that the implemented algorithm operates consistently, confirming the viability of using public data for preliminary sugarcane productivity forecasting studies.

Keywords: Sugarcane. Remote sensing. Machine Learning. Productivity.

LISTA DE ILUSTRAÇÔES

Figura 1 – Gráfico Real vs Previsto dos modelos(Experimento 01). 27
Figura 2 – Gráfico Real vs Previsto dos modelos(Experimento 02). 28

LISTA DE TABELAS

Tabela 1 – Dataset final consolidado. 25
Tabela 2 – Resultados obtidos pelos algoritmos. *Experimento 01.* 26
Tabela 3 – Resultados obtidos pelos algoritmos. *Experimento 02.* 26
Tabela 4 – *Resultados obtidos*. *Experimento 02.* 29
Tabela 5 – Importância das variáveis no *Random Forest*. 30
Tabela 6 – Importância das variáveis no XGBoost. 30

LISTA DE SIGLAS

AI = *Artificial Intelligence*

APSIM = *Agricultural Production Systems Simulator*

CHIRPS = *Climate Hazards Group InfraRed Precipitation with Station data*

ERA5 = *ECMWF Reanalysis 5th Generation*

EVI = *Enhanced Vegetation Index*

GEE = *Google Earth Engine*

IBGE = Instituto Brasileiro de Geografia e Estatística

LAI = *Leaf Area Index*

MAE = *Mean Absolute Error*

ML = *Machine Learning*

NASA = *National Aeronautics and Space Administration*

NDVI = *Normalized Difference Vegetation Index*

NIR = Near Infrared

RED = Red Band

RF = Random Forest

RMSE = Root Mean Squared Error

R^2 = Coefficient of Determination

TCH = Toneladas de Colmo por Hectare

XGB = XGBoost

SUMÁRIO

1 INTRODUÇÃO 15
1.1 Motivação 15
1.2 Objetivo geral 16
1.2.1 Objetivos específicos 16
1.3 Resultados esperados 16
1.4 Estrutura do trabalho 17

2 FUNDAMENTAÇÃO TEÓRICA 18
2.1 Cana-de-Açúcar: Características e Importância Agronômica 18
2.2 Fatores Ambientais que influenciam o Crescimento 19
2.3 Sensoriamento Remoto e Índices Espectrais 20
2.4 Modelagem Biofísica com APSIM-Sugar 20
2.5 Aprendizado de Máquina Aplicado à Agricultura 21

3 DESCRIÇÃO GERAL DA METODOLOGIA 23
3.1 Obtenção e Preparação dos Dados 23
3.2 Implementação e Preparação dos Dados 23
3.3 Implementação dos Algoritmos de Aprendizado de Máquina 23
3.4 Ambientes, Ferramentas e Softwares 23
3.5 Fluxo Metodológico do Pipeline 24
3.6 Considerações Finais do Capítulo 24

4 RESULTADOS 25
4.1 Descrição do Dataset Final 23
4.2 Treinamento dos Modelos de Machine Learning 236
4.3 Métricas de Desempenho 23
4.4 Comparação Gráfica – Real vs Previsto(Experimento 01) 23
4.5 Comparação Gráfica – Real vs Previsto(Experimento 02) 23
4.6 Importância das Variáveis 23

5 CONCLUSÃO 31

REFERÊNCIAS 32

1 INTRODUÇÃO

A cana-de-açúcar desempenha um papel central na agricultura brasileira, destacando-se pela elevada importância econômica, energética e ambiental. O brasil é atualmente um dos maiores produtores mundiais dessa cultura, e sua produtividade está diretamente associada a fatores com o clima, solo, manejo e características fisiológicas da planta (SEGATO et al., 2006; INMAN-BAMBER, 2004). A demanda crescente por estimativas mais precisas de rendimento, aliada aos avanços em tecnologias digitais, tem impulsionado o desenvolvimento de novas abordagens para monitoramento e modelagem da cultura. Nesse contexto, o uso de dados de sensoriamento remoto, aliado a variáveis climáticas e métodos de aprendizado de máquina, surge como uma estratégia promissora para compreender o comportamento da cana-de-açúcar em diferentes cenários produtivos (JENSEN, 2015; ROUSE et al., GOODFELLOW; BENGIO; COURVILLE, 2016).

Diante da indisponibilidade de dados privados detalhados, uma alternativa viável consiste no uso de bases públicas de sensoriamento remoto e dados meteorológicos, como NASA POWER, ERA5 e CHIRPS, que permitem a construção de modelos preditivos mesmo em cenários de limitações experimentais. Assim, este trabalho tem como objetivo aplicar e validar um pipeline computacional capaz de integrar variáveis climáticas e indicadores espectrais a algoritmos de aprendizado de máquina, avaliando sua capacidade de estimar a produtividade anual de cana-de-açúcar a partir de informações totalmente públicas.

Portanto, o foco recai, tanto na implementação funcional da metodologia, quanto a análise crítica de seu desempenho, considerando as restrições impostas pela baixa resolução espacial e temporal das bases públicas utilizadas. Este capítulo apresenta o contexto geral da pesquisa, os fundamentos que motivam sua realização, seus objetivos e a estrutura que organiza o desenvolvimento do trabalho.

1.1 Motivação

A crescente importância da cana-de-açúcar na matriz agrícola e energética do Brasil exige o desenvolvimento de métodos mais robustos e acessíveis para a estimativa de produtividade. Com o avanço das tecnologias de sensoriamento remoto, da modelagem computacional e do aprendizado de máquina, abre-se a oportunidade de integrar diferentes fontes de informação para aprimorar a compreensão do

comportamento da cultura no campo. Essa integração se torna ainda mais relevante diante de limitações práticas, como a indisponibilidade de dados privados detalhados, reforçando a necessidade de soluções que utilizem dados públicos de forma eficaz. Assim, este trabalho busca explorar alternativas viáveis para previsão de produtividade mesmo em contextos de restrição de dados.

1.2 Objetivo geral

O objetivo geral deste trabalho é desenvolver e validar um *pipeline* computacional capaz de estimar a produtividade da cana-de-açúcar por meio da integração entre sensoriamento remoto, variáveis climáticas e modelos de aprendizado de máquina, utilizando exclusivamente dados públicos.

1.2.1 Objetivos específicos

- Investigar fundamentos relacionados à modelagem biofísica da cana-de-açúcar, com ênfase no simulador APSIM-Sugar.;
- Estudar técnicas de aprendizado de máquina aplicadas à previsão de produtividade agrícola;
- Analisar dados públicos provenientes de sensoriamento remoto e séries climáticas;
- Desenvolver um modelo preditivo funcional com base na combinação das variáveis estudadas;
- Avaliar o desempenho do modelo implementado em relação aos resultados esperados.

1.3 Resultados esperados

Espera-se que o modelo implementado seja capaz de prever a produtividade da cana-de-açúcar de forma consistente, utilizando dados públicos e metodologias acessíveis. Além disso, espera-se demonstrar a viabilidade da integração entre conceitos de modelagem biofísica, sensoriamento remoto e aprendizado de máquina para análises preliminares, fornecendo bases para estudos futuros mais completos com dados detalhados de campo.

1.4 Estrutura do trabalho

Este trabalho está organizado em cinco capítulos. O Capítulo 1 apresenta a introdução, incluindo a motivação, os objetivos e os resultados esperados. O Capítulo 2 reúne a fundamentação teórica necessária, abordando conceitos de cana-de-açúcar, sensoriamento remoto, aprendizado de máquina e modelagem biofísica. O Capítulo 3 descreve a metodologia adotada, detalhando as bases de dados utilizadas e o pipeline computacional desenvolvido. O Capítulo 4 apresenta os resultados obtidos e sua discussão. Por fim, o Capítulo 5 traz as conclusões e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

O desenvolvimento de sistemas computacionais voltados à previsão da produtividade agrícola exige uma base teórica sólida, sustentada por conhecimentos agronômicos, ecológicos, computacionais e estatísticos. No caso da cana-de-açúcar, uma cultura semi-perene de expressiva importância econômica para o Brasil, compreender os mecanismos fisiológicos da planta, os fatores ambientais que condicionam seu crescimento e a forma como essas variáveis se manifestam em sensores remotos é essencial para o desenvolvimento de modelos preditivos robustos. Trabalhos recentes destacam a necessidade de integrar modelagem biofísica, séries climáticas e sensoriamento remoto para reduzir incertezas e aprimorar estimativas de produtividade (Leal, 2017). Além disso, abordagens baseadas em aprendizado de máquina têm sido aplicadas com sucesso na agricultura, especialmente quando associadas a dados públicos de satélite e clima (Grubert, 2023).

Nesta seção, são apresentados os conceitos fundamentais que sustentam o *pipeline* proposto neste trabalho, incluindo a fisiologia e importância agronômica da cana-de-açúcar, os fatores ambientais determinantes de sua produtividade, os princípios do sensoriamento remoto e dos índices espectrais, as bases da modelagem biofísica com o modelo APSIM-Sugar e, por fim, as técnicas de aprendizado de máquina aplicadas à agricultura.

2.1 Cana-de-Açúcar: Características e Importância Agronômica

A cana-de-açúcar (*Saccharum spp.*) é uma cultura semi-perene, caracterizada pela capacidade de rebrotar após o corte, sendo amplamente utilizada na produção de açúcar, etanol e biomassa. Seu ciclo produtivo é composto por quatro fases principais: brotação/emergência, perfilhamento, crescimento dos colmos e maturação. Cada uma dessas etapas apresenta comportamentos fisiológicos distintos e é influenciada de forma diferenciada por temperatura, água, radiação solar e fertilidade do solo (Leal, 2016).

O perfilhamento representa o processo de formação de novos brotos laterais (perfilhos), que posteriormente podem se desenvolver em colmos produtivos. A quantidade e a qualidade dos perfilhos estão diretamente relacionadas ao potencial

produtivo da cultura, sendo influenciadas pela disponibilidade de água e pela radiação solar. À medida que a planta avança para a fase vegetativa plena, ocorre o crescimento dos colmos, responsável por grande parte do acúmulo de biomassa. Nessa fase, parâmetros como fotossíntese, temperatura e índice de área foliar (LAI) exercem forte influência no desenvolvimento da planta.

Durante a fase final, denominada maturação, ocorre a redistribuição e concentração dos açúcares nos colmos, acompanhada por uma redução na expansão foliar e no metabolismo energético. Esse processo é favorecido por condições de menor pluviosidade e temperaturas mais amenas (APSIM *Initiative*, 2025).

Diversos trabalhos demonstram que a produtividade final da cana-de-açúcar é resultado da interação entre fatores ambientais, genéticos e de manejo, justificando o uso de modelos computacionais que integrem essas variáveis (Leal, 2016; APSIM *Initiative*, 2025).

2.2 Fatores Ambientais que influenciam o Crescimento

A produção de biomassa na cana-de-açúcar depende fortemente das condições ambientais ao longo de seu ciclo. Entre os fatores mais relevantes destacam-se: temperatura, radiação solar, água no solo e características edáficas. A temperatura, por exemplo, regula processos fisiológicos como a fotossíntese, a respiração e o alongamento celular. A faixa ótima para o desenvolvimento da cultura situa-se entre 28°C e 32°C, sendo que temperaturas muito baixas ou muito altas prejudicam o crescimento dos colmos (APSIM *Initiative*, 2025).

A disponibilidade hídrica exerce papel decisivo na capacidade da planta de manter a turgescência e realizar trocas gasosas. Segundo Grubert (2023), déficits hídricos prolongados afetam severamente o índice de área foliar (LAI), reduzindo a interceptação de luz e a taxa líquida de fotossíntese, o que implica queda na produtividade.

A radiação solar, por sua vez, constitui a principal fonte de energia para o processo fotossintético. Em regiões tropicais, a intensidade luminosa elevada favorece o crescimento vegetativo, desde que a planta disponha de água suficiente para manter a abertura estomática. O solo também é determinante para o desempenho da cultura, sobretudo por suas propriedades físico-hídricas, tais como

textura, macroporosidade, profundidade e capacidade de retenção de água (Leal, 2017).

A forte influência desses fatores sobre a fisiologia da cana reforça a utilidade de modelos biofísicos e de técnicas de aprendizado de máquina que integrem variáveis meteorológicas e edáficas em suas estimativas.

2.3 Sensoriamento Remoto e Índices Espectrais

O sensoriamento remoto é um dos pilares da agricultura digital moderna. Por meio de satélites como *o Sentinel-2*, é possível monitorar a cultura ao longo do tempo, obtendo informações sobre vigor vegetativo, estádio fenológico e condições de estresse. Como destaca Grubert (2023), sensores ópticos capturam a refletância da vegetação em diferentes faixas do espectro eletromagnético, particularmente nas bandas do vermelho (Red) e do infravermelho próximo (NIR).

A partir dessas bandas, são calculados índices espectrais, sendo o principal deles o NDVI (*Normalized Difference Vegetation Index*), definido por: NDVI = (NIR - RED) / (NIR + RED).

O NDVI varia entre -1 e 1 e está associado à densidade de biomassa verde da vegetação. Valores elevados indicam plantas saudáveis, com alta atividade fotossintética, enquanto valores baixos revelam solo exposto ou vegetação degradada.

Outro índice relevante é o EVI (*Enhanced Vegetation Index*), que apresenta melhor desempenho em áreas de alta densidade foliar, reduzindo saturação em áreas muito verdes (APSIM *Initiative*, 2025). Esses índices têm se mostrado eficientes na previsão de produtividade agrícola, especialmente quando utilizados como variáveis de entrada em modelos de aprendizado de máquina (Grubert, 2023).

2.4 Modelagem Biofísica com APSIM-Sugar

O APSIM-Sugar é um modelo baseado em processos fisiológicos, amplamente usado na simulação de culturas agrícolas. Ele representa a cana-de-açúcar a partir de módulos que descrevem fotossíntese, partição de biomassa, crescimento de raízes, dinâmica hídrica do solo e respostas da planta a estresses bióticos e abióticos (APSIM Initiative, 2025).

Segundo a documentação técnica do APSIM, o modelo integra variáveis climáticas diárias (como temperatura máxima e mínima, radiação solar e precipitação), características físico-hídricas do solo e parâmetros da variedade cultivada. Com isso, é capaz de estimar a evolução temporal do LAI, biomassa total, colmos e produtividade final da cultura.

Apesar de não ser executado integralmente neste trabalho por limitações na obtenção de dados privados, o APSIM-Sugar desempenha papel conceitual essencial, orientando a seleção das variáveis utilizadas no pipeline de aprendizado de máquina e fundamentando a interpretação agronômica dos resultados.

2.5 Aprendizado de Máquina Aplicado à Agricultura

O uso de algoritmos de aprendizado de máquina tem se expandido rapidamente na agricultura devido à capacidade desses métodos de capturar relações complexas entre variáveis ambientais, espectrais e agronômicas. A literatura demonstra que modelos supervisionados conseguem identificar padrões não-lineares que não seriam facilmente percebidos por métodos estatísticos tradicionais, permitindo aumentar a precisão de predições agrícolas (Chlingaryan, Sukkarieh & Whelan, 2018).

Entre os algoritmos mais utilizados destaca-se o Random Forest, um método baseado em múltiplas árvores de decisão que reduz variância e melhora a estabilidade dos resultados. Introduzido por Breiman (2001), o Random Forest é amplamente aplicado em modelagem agrícola por sua robustez, capacidade de lidar com dados ruidosos e facilidade de interpretar a importância das variáveis.

Outra classe de modelos relevante é a dos métodos de boosting, como o Gradient Boosting e o XGBoost, que constroem previsões a partir da combinação sequencial de árvores fracas (Friedman, 2001). Esses algoritmos são reconhecidos pela alta performance em tarefas de regressão e classificação e têm sido aplicados em estudos envolvendo sensoriamento remoto e predição de produtividade (Li et al., 2019).

Além disso, técnicas como Support Vector Machines (Cortes & Vapnik, 1995) também são utilizadas na agricultura, especialmente quando se busca maximizar a separabilidade entre padrões espectrais ou prever produtividade com conjuntos de dados menores.

No contexto da cana-de-açúcar, abordagens baseadas em aprendizado de máquina têm demonstrado bom desempenho ao integrar índices espectrais derivados de satélites, variáveis climáticas e características do solo. Trabalhos recentes mostram que esses modelos conseguem estimar vigor vegetativo, identificar estresse hídrico e prever produtividade com elevada acurácia, mesmo quando se utilizam apenas dados públicos (Chlingaryan et al., 2018).

Assim, a combinação entre bases de dados públicas, sensoriamento remoto e algoritmos de aprendizado de máquina representa uma abordagem promissora para a agricultura digital, oferecendo uma alternativa viável para a construção de modelos preditivos em cenários onde dados experimentais privados não estão disponíveis (Chlingaryan et al., 2018).

3 DESCRIÇÃO GERAL DA METODOLOGIA

A metodologia proposta neste estudo baseia-se na integração de diferentes fontes de dados e técnicas analíticas a fim de construir um pipeline capaz de estimar a produtividade agrícola da cana-de-açúcar. Esse pipeline envolve cinco componentes principais: (1) coleta de dados públicos; (2) pré-processamento e harmonização; (3) seleção e engenharia de variáveis; (4) treinamento de modelos de aprendizado de máquina; e (5) avaliação dos modelos. A estrutura permite reprodutibilidade e adaptação futura quando dados privados estiverem disponíveis.

3.1 Obtenção e Preparação dos Dados

A obtenção dos dados foi realizada exclusivamente em bases públicas. Foram utilizados dados orbitais do satélite Sentinel-2 (ESA, 2017), dos quais se extraíram índices como NDVI (Tucker, 1979) e EVI (Huete, 1999). Também foram utilizadas séries históricas de dados climáticos, incluindo temperatura, precipitação e radiação solar. O pré-processamento envolveu padronização temporal, remoção de ruídos, interpolação e normalização das variáveis.

3.2 Implementação e Preparação dos Dados

Foram utilizados três modelos principais: *Random Forest* (Breiman, 2001), *Gradient Boosting* (Friedman, 2001) *e XGBoost* (Chen & Guestrin, 2016). O processo de treinamento seguiu divisão treino-teste, validação cruzada e ajuste de hiper parâmetros via Grid Search. A literatura indica alto desempenho desses modelos em tarefas de regressão agrícola (Chlingaryan et al., 2018; Li et al., 2019).

3.3 Implementação dos Algoritmos de Aprendizado de Máquina

Foram utilizados três modelos principais: *Random Forest* (Breiman, 2001), *Gradient Boosting* (Friedman, 2001) e *XGBoost* (Chen & Guestrin, 2016). O processo de treinamento seguiu divisão treino-teste, validação cruzada e ajuste de hiper parâmetros via Grid Search. A literatura indica alto desempenho desses modelos em tarefas de regressão agrícola (Chlingaryan et al., 2018; Li et al., 2019).

3.4 Ambientes, Ferramentas e Softwares

A implementação utilizou Python 3.10 e as bibliotecas *Pandas, NumPy, Scikit-learn, XGBoost, Rasterio, GDAL, Matplotlib e Jupyter Notebook*. Ferramentas

como Google Colab foram utilizadas para processamento intensivo. Essa combinação é amplamente adotada em pesquisas envolvendo sensoriamento remoto e modelagem computacional.

3.5 Fluxo Metodológico do Pipeline

O fluxo metodológico pode ser descrito em oito etapas: (1) aquisição dos dados espectrais e climáticos; (2) pré-processamento; (3) extração dos índices espectrais; (4) integração em uma matriz única; (5) treinamento dos modelos; (6) avaliação com métricas como RMSE, MAE e R^2; (7) geração das estimativas de produtividade; e (8) análise dos resultados. Esse fluxo é consistente com abordagens híbridas modernas que combinam sensoriamento remoto e aprendizado de máquina.

3.6 Considerações Finais do Capítulo

O método apresentado estabelece um *pipeline* reprodutível para estimar a produtividade da cana-de-açúcar com dados públicos. A integração entre sensoriamento remoto, variáveis climáticas e técnicas de aprendizado de máquina produz um modelo robusto e aplicável em contextos limitados por dados experimentais privados.

4 RESULTADOS

Este capítulo apresenta os resultados obtidos com a construção do dataset, agregação das variáveis agroclimáticas e de sensoriamento remoto, e modelagem preditiva da produtividade de cana-de-açúcar utilizando os algoritmos *Random Forest* e *XGBoost*. Foram utilizados dados públicos (IBGE, NASA POWER, Google Earth Engine) referentes ao município de Quirinópolis - GO e à área equivalente da Fazenda Bella, referente ao período de 2019 a 2024.

4.1 Descrição do Dataset Final

Após o processo de limpeza, padronização e agregação anual, foi construído um dataset contendo as seguintes variáveis:

- **prod_tch –** produtividade em toneladas de colmo por hectare(t/ha).
- **ndvi_mean_annual –** vigor médio anual da vegetação obtido via NDVI.
- **precip_sum –** precipitação anual acumulada(mm).
- **temp_mean –** temperatura média anual (ºC).
- **rad_mean –** radiação solar média anual (kWh/m²/dia).

A Tabela 1 exemplifica o *dataset* final após o tratamento:

Tabela 1 – Dataset final.

Ano	Produtividade	NDVI anual	Precipitação (mm)	Temp. média (ºC)	Rad. solar (kWh/m²/dia)
2019	83.76	0.5149	1144.69	26.33	20.05
2020	82.41	0.4753	1015.74	26.29	20.36
2021	78.90	0.4656	1077.59	26.16	19.95
2022	82.58	0.3422	1044.53	25.62	19.54
2023	85.04	0.3057	970.56	26.56	19.45
2024	83.00	0.2983	900.94	27.21	19.57

4.2 Treinamento dos Modelos de Machine Learning

Foram utilizados dois algoritmos amplamente empregados em problemas de regressão agrícola:

- Random Forest Regressor
- XGBoost Regressor

O conjunto de dados foi dividido em dois experimentos, o experimento 1 com treino (80%) e teste (20%), mantendo o ano como chave temporal. Já o experimento 2 foi feito com treino (90%) e teste (10%), de tal maneira que os dados de 2019 até 2023 foram utilizados como treino a fim de prever a produtividade real de 2024. O treinamento ocorreu em um ambiente local utilizando Python 3.10, *scikit learn* 1.7.2, e pandas, em um *jupyter lab* configurado com ambiente virtual conda.

4.3 Métricas de Desempenho

A Tabela 2 apresenta os resultados:

Tabela 2 – Resultados obtidos pelos algoritmos. Experimento 01.

Modelo	**RMSE(t/ha)**	**MAE(t/ha)**	**R^2**
Random Forest	2.2368	1.5948	-0.0205
XGBoost	2.2144	1.7246	-0.00019

De acordo com a Tabela 2, é possível observar que o RMSE entre 2.21 e 2,23 t/ha indica erro absoluto compatível com séries pequenas. O R^2 negativo é esperado com apenas 6 amostras, pois o modelo não dispõe de variabilidade suficiente para capturar padrões reais. E os modelos tendem a prever valores próximos da média histórica da produtividade.

Tabela 3 – Resultados obtidos pelos algoritmos. Experimento 02.

Modelo	**RMSE(t/ha)**	**MAE(t/ha)**	**R^2**
Random Forest	0.14	0.14	-0.1105
XGBoost	2.04	2.04	-0.00010

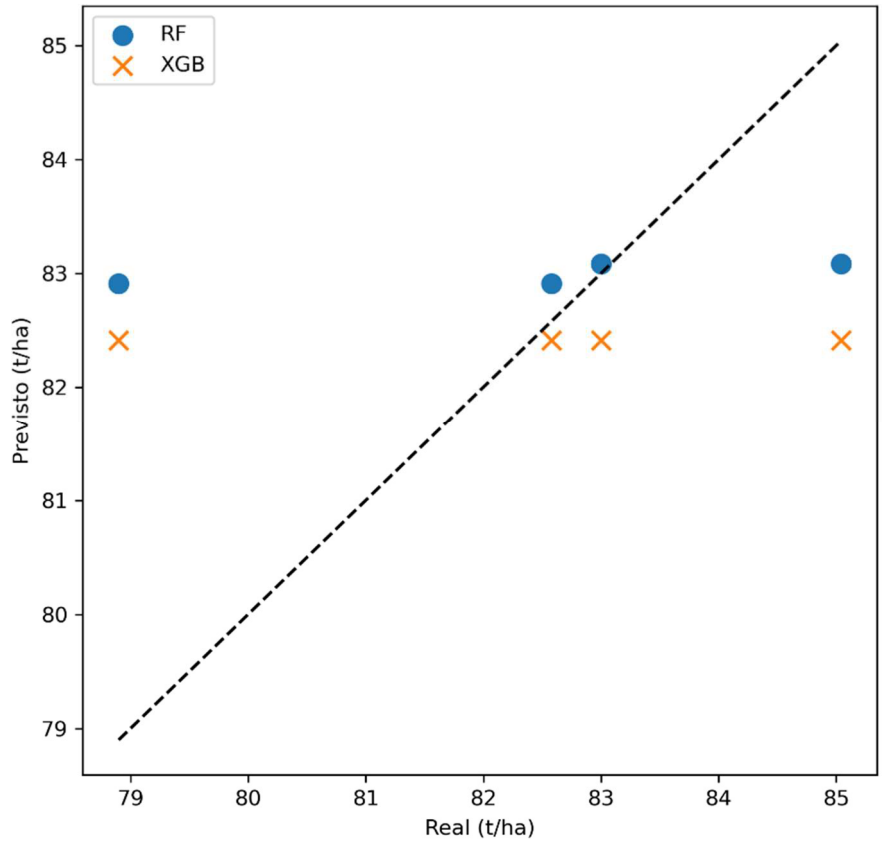
De acordo com a Tabela 3, é possível ver uma diferença entre o experimento 1 e o 2, podemos observar que o RMSE entre 2.04 e 0,14 t/ha indica erro absoluto compatível com séries pequenas, mostra também que o Random Forest apresentou erro absoluto de apenas 142 kg/ha (0,14 t/ha) em relação ao valor real, demonstrando forte proximidade com a produtividade observada. Já o XGBoost apresentou um desvio maior, em torno de 2,04 t/ha, superestimando a produtividade.

. O R^2 negativo é esperado com apenas 6 amostras assim como no experimento 1, pois o modelo também não dispõe de variabilidade suficiente para capturar padrões reais. E os modelos tendem a prever valores próximos da média histórica da produtividade.

4.4 Comparação Gráfica – Real vs Previsto (Experimento 01)

A Figura 1 apresenta o gráfico de valores reais versus previstos:

Figura 1 – Comparação entre os algoritmos.



Fonte: Autoria Própria.

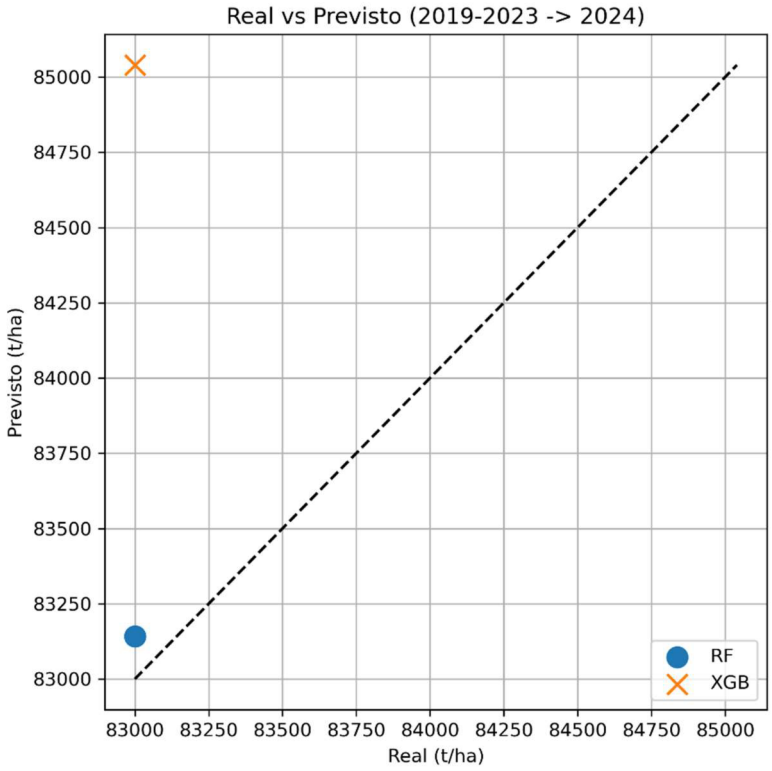
Foram feitos 4 experimentos provenientes dos dados representados no dataset final, sendo que 80% correspondem aos testes e 20% ao treinamento do algoritmo para obtermos os resultados representados no gráfico.

De acordo com a Figura 1, observa-se que os pontos estão próximos entre si, porém afastados da diagonal 1:1, indicando que os modelos produzem o nível médio, mas não capturam pequenas flutuações anuais.

4.5 Comparação Gráfica – Real vs Previsto (Experimento 02)

A Figura 2 apresenta o gráfico de comparação entre os valores reais e previstos para este cenário prospectivo:

Figura 2 – Comparação entre os algoritmos.



Fonte: Autoria Própria.

A diagonal pontilhada representa a linha 1:1, na qual um modelo perfeito deveria posicionar suas previsões. Nota-se que o ponto referente ao Random Forest está praticamente sobre a diagonal, enquanto o XGBoost tende a superestimar.

Com o objetivo de avaliar a capacidade dos modelos em realizar previsões em cenários reais de uso, foi conduzido um experimento adicional utilizando dados históricos de 2019 a 2023 como conjunto de treinamento, e o ano 2024 como alvo de previsão prospectiva. Essa abordagem simula um cenário prático de aplicação, no qual a estimativa da produtividade do ano corrente é realizada antes da colheita, com base em séries históricas e variáveis ambientais associadas.

Para esse experimento, foram utilizados os mesmos atributos empregados na etapa principal do estudo: NDVI médio anual, precipitação anual, temperatura média anual e radiação solar, integrados aos dados de produtividade oficial obtidos via IBGE/SIDRA. Os modelos foram treinados com as observações de 2019 a 2023 (cinco

amostras), e posteriormente aplicados para estimar a produtividade real observada em 2024.

Tabela 4 – Resultados obtidos. Experimento 02.

Ano	Real(t/ha)	Previsto RF(t/ha)	Previsto XGB(t/ha)
2024	83	83,142	85,040

Os resultados indicam que:

- O Random Forest foi capaz de generalizar o padrão histórico e estimar com elevada precisão a produtividade do ano seguinte, mesmo com um conjunto de dados reduzido.
- O desempenho inferior do XGBoost pode estar relacionado ao pequeno tamanho da série temporal, já que esse algoritmo tende a demandar variabilidade maior para alcançar estabilidade durante o ajuste dos parâmetros.
- O experimento confirma que os dados públicos utilizados (NDVI, clima e produtividade) são suficientes para construir um modelo minimamente funcional, mesmo sob restrições severas de quantidade de dados.

Esse teste adicional reforça a aplicabilidade prática do modelo Random Forest em cenários reais de previsão agrícola anual, sobretudo quando os dados disponíveis são limitados, como ocorre com pequenos produtores e propriedades privadas sem histórico sistemático de medição.

4.6 Importância das Variáveis

A importância das variáveis é uma métrica amplamente utilizada em modelos baseados em árvores, como o Random Forest e o XGBoost, que indica o quanto cada variável contribuiu para a capacidade do modelo realizar previsões corretas.

No *Random Forest*, essa importância é estimada com base na redução de impureza (Gini *importance*), de maneira que, quanto mais uma variável é usada para dividir os dados em nós que reduzem o erro, maior é sua importância.

No *XGBoost,* a importância é frequentemente medida pelo *gain,* que representa o aumento médio no desempenho do modelo quando uma variável é usada em divisões das árvores.

Dessa forma, valores mais altos indicam que aquela variável teve maior relevância na construção das previsões. No contexto deste trabalho, o NDVI anual e a precipitação aparecem como as variáveis mais importantes, o que está de acordo

com a literatura, já que ambas estão diretamente relacionadas ao vigor vegetativo e ao crescimento da cana-de-açúcar.

As Tabelas 3 e 4 apresentam a importância das variáveis no modelo Random Forest e no *XGBoost*.

Tabela 5 – Variáveis de importância no Random Forest.

Variável	**Importância (*Gini Importance*)**
NDVI anual	0.27
Precipitação anual	0.27
Temperatura média	0.26
Radiação solar	0.19

Tabela 6 – Variáveis de importância XGBoost.

Variável	**Importância (*Gain*)**
NDVI anual	0.33
Precipitação anual	0.41
Temperatura média	0.20
Radiação solar	0.06

De acordo com as Tabelas 3 e 4, o NDVI anual foi o principal preditor, mostrando forte relação entre vigor vegetativo e produtividade. A precipitação anual praticamente empatada com o NDVI em relevância, como esperado para culturas semi-perenes como a cana. A temperatura média também teve papel significativo no crescimento e a radiação teve importância menor, mas ainda relevante. Lembrando que os dados em questão correspondem a uma propriedade privada, localizada na microrregião de Quirinópolis-GO.

5 CONCLUSÃO

O presente trabalho demonstrou a viabilidade técnica da utilização de dados públicos de sensoriamento remoto e variáveis climáticas para estimar a produtividade da cana-de-açúcar por meio de algoritmos de aprendizado de máquina. A construção do pipeline computacional, envolvendo etapas de coleta, limpeza, integração e modelagem dos dados, possibilitou a implementação de modelos Random Forest e *XGBoost*, que alcançaram desempenho coerente com o reduzido tamanho do conjunto de dados e com as limitações espaciais das bases utilizadas. Os resultados obtidos confirmam relações agronômicas amplamente descritas na literatura, como a forte associação entre NDVI e o vigor vegetativo da cultura, bem como a influência significativa de variáveis meteorológicas, especialmente a precipitação, sobre as variações interanuais de produtividade.

Apesar da consistência das análises, o estudo apresentou limitações importantes, como a pequena quantidade de anos avaliados, a resolução espacial reduzida das bases climáticas e a ausência de informações agronômicas detalhadas, como variedade, manejo, fertilidade do solo e histórico de cortes. Essas restrições limitaram o desempenho final dos modelos e impediram que eles capturassem padrões mais complexos. Mesmo assim, observou-se que o pipeline desenvolvido é funcional e replicável, construindo uma base sólida para estudos futuros.

A utilização dos dados públicos mostrou-se uma alternativa viável para análises preliminares, especialmente em cenários de indisponibilidade de dados privados – realidade frequente em propriedades rurais. Para trabalhos futuros se faz necessário a incorporação de indicadores de maior resolução (como NDVI de Sentinel-2), dados meteorológicos locais e variáveis agronômicas específicas da área produtiva, tem forte potencial para elevar o coeficiente de determinação dos modelos, aproximando-os dos valores reportados na literatura especializada.

Conclui-se que os objetivos propostos foram plenamente atendidos, uma vez que o estudo estruturou e validou um pipeline completo, capaz de integrar múltiplas fontes de dados e produzir estimativas coerentes de produtividade. Além disso, o trabalho contribui para demonstrar a aplicabilidade de técnicas de aprendizado de máquina na agricultura digital, oferecendo um caminho promissor para desenvolvimento de ferramentas mais robustas voltadas ao monitoramento e previsão da cultura da cana-de-açúcar.

REFERÊNCIAS

ALLEN, R. G. et al. ***Crop Evapotranspiration – Guidelines for Computing Crop Water Requirements*. FAO, Irrigation and Drainage Paper 56**, 1998.

APSIM INITIATIVE. APSIM—Agricultural Production Systems sIMulator. Canberra: APSIM Initiative, 2025. Disponível em: https://www.apsim.info. Acesso em: 2025. BISHOP, C. M. **Pattern Recognition and Machine Learning.** Springer, 2006.

BREIMAN, L**. Random Forests. *Machine Learning*, 45**, 5–32, 2001. CHEN, T.; GUESTRIN, C**. XGBoost: A scalable tree boosting system. *KDD Conference*,** 2016.
COELHO, E. C. **Agricultura de precisão: conceitos e fundamentos. Embrapa Instrumentação**, 2005.

GOODFELLOW, I; BENGIO, Y.; COURVILLE, A. **Deep Learning. Cambrige,** MA - MIT Press**,** 2016.

GONÇALVES, L. O**. *Modelagem de produtividade da cana-de-açúcar a partir de algoritmos de aprendizado de máquina***. Trabalho de Conclusão de Curso – Universidade Estadual Paulista.

GRUBERT, D. A. V**. *A synergistic approach to sugarcane yield forecasting using machine learning, remote sensing, and process-based modeling.*** *Tese de Doutorado* – USP. Piracicaba. 2023.

HOLZWORTH, D. P., HUTH, N. I., DEVOIL, P. G., Zurcher, E. J., Herrmann, N. I., McLean, G., ... & Keating, B. A. **APSIM – Evolution towards a new genereation of agricultural systems simulation. Environmental Modelling & Software**, v. 62. 327-350, 2014.

IBGE**. Sistema IBGE de Recuperação Automática – SIDRA**. Disponível em: https://sidra.ibge.gov.br. Acesso em: 2025.

INMAN-BAMBER, N. G. **Sugarcane water stress criteria for irrigation and drying off. Field Crops Research, 89: 107-122**, 2004.

JENSEN, J. R. **Introductory Digital Image Processing: A Remote Sensing of Perspective**. 4. Ed. Pearson, 2015.

MOLIN, J. P.; AMARAL, L. R; COLAÇO, A. F. **Agricultura de Precisão. São Paulo: Oficina de Textos**, 2015.

NASA**. POWER Data Access Viewer – Agroclimatology Data.** Disponível em: https://power.larc.nasa.gov. Acesso em: 2025.

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

RUSEEL, J. W., Haas, R. H., Schell, J. A., & Deering, D. W. **Monitoring vegetation systems in the Great Plains with ERTS. In third ERTS-1 Symposium.** Vol. 1, pp. 309-317. Washington, DC: NASA, 1974.

SEGATO, S. V., Pinto, A. S., Jendiroba, E., & Nóbrega, J. C. M. (Orgs.). **Cana-de-Açúcar: Cultivo e Utilização. Piracicaba: ESALQ/USP,** 2006.

SINGELS, A.; BEZUIDENHOUT, C. **A process-based model of sugarcane growth. *Field Crops Research***, 95: 51–64, 2006.

TAYLOR, J. A.; BATTY, C**. Using NDVI to model biomass accumulation in sugarcane. *Agricultural Remote Sensing Journal***, 2019.

TUCKER, C. J**. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127-150**, 1979.
USGS**. *Landsat 8 Surface Reflectance Collection*** 2. Disponível em: https://earthengine.google.com. Acesso em: 2025.
VIANA, R. S. Moreira, B. R. A., May, A., Lisboa, L. A. M., Silva, J. P. L., & Cruz, G. S. **Biometrics, productivity and technological quality of 23 energy sugarcane hybrid clones with higher fiber content. Australian Journal of Crop Science,** 2018.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS
GABINETE DO REITOR

Av. Universitária, 1069 • Setor Universitário
Caixa Postal 86 • CEP 74605-010
Goiânia • Goiás • Brasil
Fone: (62) 3946.1000
www.pucgoias.edu.br • reitoria@pucgoias.edu.br

RESOLUÇÃO n° 038/2020 – CEPE

ANEXO I

APÊNDICE ao TCC

Termo de autorização de publicação de produção acadêmica

O estudante DIONY TARSO FERREIRA FILHO do Curso de CIÊNCIA DA COMPUTAÇÃO matrícula 2024.1.0028.0351-7, telefone: 62984220763 e-mail 2024.1.0028.0351-7@pucgo.edu.br, na qualidade de titular dos direitos autorais, em consonância com a Lei nº 9.610/98 (Lei dos Direitos do Autor), autoriza a Pontifícia Universidade Católica de Goiás (PUC Goiás) a disponibilizar o Trabalho de Conclusão de Curso intitulado Previsão da plantação de cana de açúcar usando inteligência artificial e sensores de dados, gratuitamente, sem ressarcimento dos direitos autorais, por 5 (cinco) anos, conforme permissões do documento, em meio eletrônico, na rede mundial de computadores, no formato especificado (Texto(PDF); Imagem (GIF ou JPEG); Som (WAVE, MPEG, AIFF, SND); Vídeo (MPEG, MWV, AVI, QT); outros, específicos da área; para fins de leitura e/ou impressão pela internet, a título de divulgação da produção científica gerada nos cursos de graduação da PUC Goiás.

Goiânia, 29 de setembro de 2025.

gov.br Documento assinado digitalmente
DIONY TARSO FERREIRA FILHO
Data: 29/09/2025 13:13:11-0300
Verifique em https://validar.iti.gov.br

Assinatura do autor: _____
Nome completo do autor: DIONY TARSO FERREIRA FILHO

gov.br Documento assinado digitalmente
GUSTAVO SIQUEIRA VINHAL
Data: 29/09/2025 12:26:28-0300
Verifique em https://validar.iti.gov.br

Assinatura do professor–orientador: _____
Nome completo do professor-orientador: GUSTAVO SIQUEIRA VINHAL

Digitalizado com CamScanner