ESTUDO COMPARATIVO ENTRE DIFERENTES REGRESSORES PARA ESTIMAR PRODUTIVIDADE DE CANA-DE-AÇÚCAR

Luiz Antonio Falaguasta Barbosa[1], *Daniel Carlos Guimarães Pedronette*[2] *e Ivan Rizzo Guilherme*[3]

Universidade Estadual Paulista "Júlio de Mesquita Filho", Av. 24 A, 1515 - Bela Vista, Rio Claro - SP, 13506-700 -
[1]luiz.falaguasta@unesp.br, [2]daniel.pedronette@unesp.br, [3]ivan.guilherme@unesp.br

RESUMO

Modelos de aprendizado de máquina têm obtido resultados notáveis em diversas áreas do conhecimento, inclusive em agricultura de precisão. Contudo, a diversidade de modelos distintos disponíveis e a carência de análises comparativas torna a seleção de um modelo uma tarefa desafiadora. Este artigo faz uma avaliação de diferentes regressores em tarefas de predição de produtividade de cana-de-açúcar, modeladas a partir de características compostas por índices vegetativos. O estudo realizado considerou os regressores Multi Layer Perceptron (MLP), Support Vector Regressor (SVR), Random Forest, AdaBoost e Gradient Boosting, ajustados a partir de hiperparâmetros obtidos com o método *grid search*. Foram utilizados conjuntos de dados públicos, referente a áreas cultivadas em 2 campos experimentais na Austrália. Foram considerados 5 modelos obtidos a partir de 10 índices vegetativos multiespectrais e espectro visível *red-green-blue* (RGB), escolhidos com base no redutor de dimensionalidade Análise de Componentes Principais (PCA). Além de imagens multiespectrais, bandas *Light Detection And Ranging* (LiDAR) também foram utilizadas no estudo comparativo.

Palavras-chave – Grid search, cana-de-açúcar, estimativa de produtividade, aprendizado de máquina.

ABSTRACT

Machine learning models have obtained remarkable results in several areas of knowledge, including precision agriculture. However, the diversity of different models available and the lack of comparative analyzes make the selection of a model a challenging task. This article makes an evaluation of different regressors in sugarcane productivity prediction tasks, modeled from features composed by vegetative indices. The study carried out considered the Multi Layer Perceptron(MLP), Support Vector Regression (SVR), Random Forest, AdaBoost and Gradient Boosting regressors, adjusted from hyperparameters obtained with the grid search method. Public datasets were used, referring to cultivated areas in 2 experimental fields in Australia. We considered 5 models obtained from 10 multispectral and visible spectrum red-green-blue (RGB) vegetative indices, chosen based on the dimensionality reducer Principal Component Analysis (PCA). In addition to multispectral imaging, Light Detection And Ranging (LiDAR) bands were also used in the comparative study.

Key words – *Grid search, sugarcane, yield crop, machine learning.*

1. INTRODUÇÃO

A visibilidade da estimativa de produção com antecedência de meses da colheita representa uma vantagem ao produtor, uma vez que permite executar o planejamento da comercialização, antecipar o investimento na aquisição de insumos para a safra seguinte e subsidiar a tomada de decisão sobre qual proporção e qual o melhor momento de iniciar a reforma de canaviais. A modelagem desses dados tem potencial para auxiliar os canavicultores a fazerem uma estimativa mais precisa de safras futuras. Enfim, há uma profusão de decisões que podem ser tomadas com base nessa estimativa, sendo possível realizar diferentes cálculos para melhorar a gestão tanto dos canaviais como das usinas que processam a matéria-prima. Ademais, tais tecnologias podem servir de base científica para assessorar iniciativas de políticas governamentais para o setor.

A previsão da safra é definida pela Organização para Agricultura e Alimentação (FAO) das Nações Unidas como a arte de prever a produção da safra antes que a colheita realmente ocorra, normalmente com alguns meses de antecedência [1]. Assim, desenvolver um modelo de aprendizado de máquina baseado em diferentes tipos de dados para estimar a produtividade esperada em talhões comerciais de cana-de-açúcar assume grande importância para a gestão e otimização das tomadas de decisões no setor sucroalcooleiro.

Existem hoje formas de estimar a produtividade de cana com base em dados meteorológicos e modelo de crescimento de plantas [2], baseado em dados de sensoriamento remoto [3] e utilizando-se os tipos de dados de sensoriamento remoto e meteorológicos somados a dados agronômicos, fazendo uso de aprendizado de máquina [4].

Tais possibilidades são viabilizadas por meio do uso de dados georreferenciados de imagem, meteorológicos e agronômicos. Dados de satélite, embora já amplamente utilizados como rotina há décadas pelo setor de pesquisa agrícola e de estudos governamentais [5], apenas recentemente tornaram-se mais acessíveis para o uso de produtores, com a maior oferta de imagens públicas de maior resolução espectral, espacial e temporal. Além disso, estimativas mais fidedignas podem ser obtidas quando as abordagens são combinadas com dados coletados no campo (*on farm*), tais como informações agronômicas da cultura, dados ambientais coletados por estações meteorológicas, avaliações biométricas coletadas no campo, e logicamente, dados de produtividade e qualidade de matéria-prima.

Dado a evolução dos modelos de aprendizado de máquina e os resultados promissores obtidos, há muitas opções de regressores disponíveis. Contudo, muitos trabalhos consideram um único modelo avaliado na área de estudo. Além disso, regressores distintos podem atingir diferentes resultados para um mesmo conjunto de dados, modelados com as mesmas variáveis independentes. Dessa forma, a seleção de um modelo para uma tarefa de predição de produtividade torna-se uma tarefa desafiadora.

Este trabalho tem como objetivo contribuir nessa tarefa, apresentando um estudo comparativo entre diferentes regressores baseados em aprendizado de máquina supervisionado. Em [6], uma revisão sistemática sobre predição de produtividade usando aprendizado de máquina, os mais utilizados foram Redes Neurais (27), Regressão Linear (14), Random Forest (12), Vetores de Suporte (10) e Gradient Boosting (4). Considerando direções apontadas em [6], alguns regressores foram avaliados em um conjunto

de dados público referente a áreas cultivadas em 2 campos experimentais na Austrália.

Este trabalho considerou diferentes regressores para avaliar, por meio do coeficiente de determinação ajustado ($\overline{R}^2$), a capacidade de acerto das predições realizadas em diferentes amostras ao longo do tempo e sob diferentes modelos construídos a partir de imagens multiespectrais e dados de LiDAR.

2. MATERIAL E MÉTODOS

2.1. Visão geral

Este trabalho trata da comparação entre diferentes regressores para obtenção da estimativa de produtividade de cana-de-açúcar em campos experimentais por meio de índices vegetativos e bandas espectrais obtidas por bandas multiespectrais como verde (*green* - G), vermelho (*red* - R), azul (*blue* - B), infravermelho próximo (*near infrared* - NIR) e borda vermelha (*red edge* - RE), incluindo o *Light Detection And Ranging* - LiDAR. Os índices vegetativos multiespectrais são um total de 10 (Normalized Difference Vegetation Index - NDVI, Normalized Difference Red Edge Index - NDRE, Green Normalized Difference Vegetation Index - GNDVI, Enhanced Vegetation Index - EVI, Modified Anthocyanin Content Index - MACI, Optimized Soil Adjusted Vegetation Index - OSAVI, Simplified Canopy Chlorophyll Content Index - SCCCI, Transformed Chlorophyll Absorption and Reflectance Index - TCARI, Triangular Greenness Index - TGI, Visible Atmospherically Resistant Index - VARI) e as características geradas a partir deles, de 7 variáveis estatísticas: máximo, mínimo, média e 4 percentis; resultando em 70 características. Já as bandas LiDAR, um total de 48. Foram utilizados 56 pontos de coleta de amostras de biomassa em parcelas de 2m x 2m distribuídas aleatoriamente ao longo das linhas 1 e 6 (Figura 3), distribuídas em 6 datas diferentes com intervalo de 7 dias.

A Figura 1 ilustra o workflow considerado, que consiste em: (1) coleta de imagens multiespectrais por meio de Veículo Aéreo Não Tripulado (VANT); (2) definição da área de estudo e extração de índices vegetativos (IVs) obtidos a partir das bandas multiespectrais; (3) redução de dimensionalidade por meio de *Principal Component Analysis (PCA)*; (4) obtenção de diferentes modelos de estimativa de produtividade a partir dos IVs, com os devidos ajustes dos melhores valores em diferentes hiperparâmetros para modelos definidos em [7] por meio de Grid Search; e (5) comparação entre esses modelos a fim de identificar aquele que apresenta melhores resultados.

2.2. Área de estudo

A área de estudo consiste de dois campos experimentais, de 1 hectare (ha) cada, localizados no nordeste australiano, ilustrado na Figura 2. O local apresenta precipitação que ultrapassa os 4000 mm ao ano, com chuvas concentradas entre dezembro e março. Cada campo experimental apresenta diferentes tipos de solo, sendo o campo experimental 1 com solo aluvial bem drenado e o 2, solo aluvial mal drenado.

2.3. Regressores

A fim de comparar os resultados apresentados em [7], foram avaliados cinco diferentes regressores:

- *Support Vector Regression*: uma generalização de SVM (*Support Vector Machine*) que é obtida pela introdução de uma região insensível ε ao redor da função, denominada de tubo ε. Este tubo reformula o problema de otimização para encontrar o tubo que melhor se aproxima da função de valor contínuo, enquanto equilibra a complexidade do modelo e o erro de predição;
- *Random Forest regressor*: um meta-estimador que ajusta um número de árvores de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo;
- *Gradient Boosting regressor*: estimador que constrói um modelo aditivo de forma progressiva, permitindo a otimização de funções de perda diferenciáveis arbitrárias. Em cada estágio, uma árvore de regressão é ajustada no gradiente negativo da função de perda dada;
- *AdaBoost regressor*: é um meta-estimador que começa ajustando um regressor no conjunto de dados original e depois ajusta cópias adicionais do regressor no mesmo conjunto de dados, mas onde os pesos das instâncias são ajustados de acordo com o erro da predição atual.
- *MLP regressor*: é um estimador que otimiza o erro quadrado usando LBFGS (algoritmo *Limited-memory* Broyden–Fletcher–Goldfarb–Shanno) ou gradiente descendente estocástico.

2.4. Aspectos de implementação

O estudo foi realizado a partir do código-fonte distribuído pelos autores de [7], que foi alterado para incorporar outros regressores. O protocolo experimental estabelecido em [7] foi seguido e os resultados obtidos foram comparados aos reportados no texto original. Para os demais regressores foram utilizadas implementações disponíveis na biblioteca Scikit-learn [8].

2.5. Delineamento experimental

O desenho experimental contém 5 tratamentos de Nitrogênio (N), à taxa de 0, 70, 110, 150, 190 kgN/ha, com 4 repetições em delineamento de blocos completos ao acaso (Figura 3). Cada bloco com 10 m de largura, 30 m de comprimento e composto de 6 linhas de plantio de cana-de-açúcar. Os tratamentos foram aplicados 52 e 55 dias após a colheita da safra anterior no campo 1 e campo 2, respectivamente.

2.6. Grid search

Grid search é um processo que pesquisa exaustivamente por meio de um subconjunto especificado manualmente do espaço de hiperparâmetros do algoritmo de um dado regressor. Tal processo foi utilizado neste trabalho, fornecendo um conjunto de hiperparâmetros a cada um dos regressores enumerados em 2.3, a fim de obter o regressor que melhor estimasse a produtividade para cada uma das 6 amostras nos respectivos modelos avaliados em [7]. A Figura 4 apresenta os resultados do regressor *Ordinary Least Squares* (OLS), originalmente utilizado em [7].

3. RESULTADOS

Considerando os resultados originais reportados em [7] (Figura 4), foi possível melhorar o coeficiente de determinação ajustado ($\overline{R}^2$) para todas as amostras e para todos os modelos utilizados, com exceção do regressor Random Forest e Gradient Boosting, que não tiveram bons

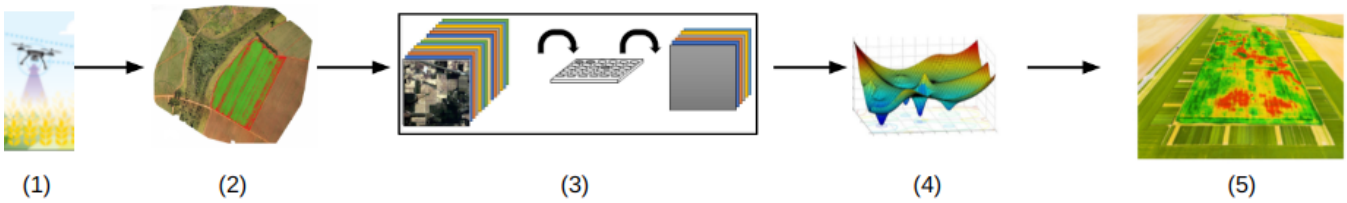

Figura 1: Workflow para obtenção de estimativa de produtividade a partir de IVs extraídos de imagens capturadas com VANT.

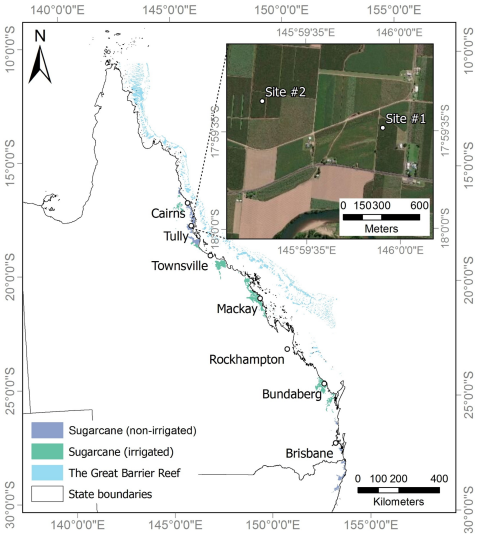

Figura 2: Localização dos campos experimentais [7].

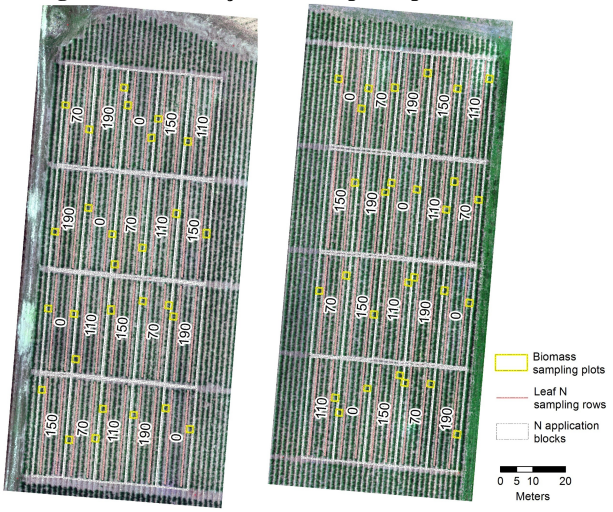

Figura 3: Desenho experimental contendo pontos de coleta de biomassa e Nitrogênio [7].

resultados para o modelo Multiespectral + LiDAR. Tal coeficiente de determinação ajustado é obtido por meio do cálculo:

$$\overline{R}^2 = 1 - (1 - R^2) * \left(\frac{n-1}{n-p}\right) \quad (1)$$

onde:

- R^2: valor de R ao quadrado para uma regressão linear;
- n: tamanho da amostra usada na regressão;
- p: número de preditores usados na regressão, incluindo a constante.

Em se tratando do SVR e do Random Forest regressor, estes apresentaram curvas semelhantes para 4 modelos, com

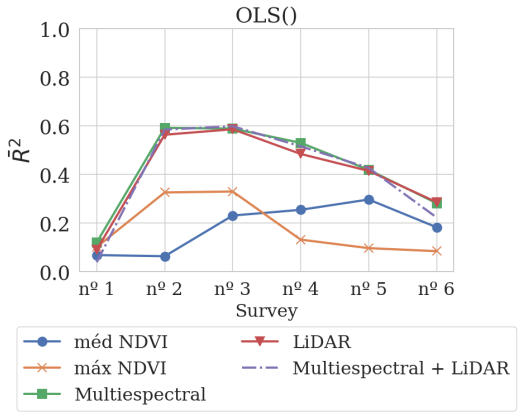

Figura 4: Regressor OLS.

diferença para o modelo Multiespectral + LiDAR, onde, no Random Forest, assumiu valores inferiores e mantendo praticamente a mesma silhueta. O regressor Gradient Boosting diferiu dos anteriores por apresentar valores iguais ou inferiores a 0,6 em todas as amostras, não apresentando resposta ao modelo Multiespectral + LiDAR. Já o regressor MLP foi o pior deles, não apresentando $\overline{R}^2$ às modelagens com NDVI médio e máximo. A Figura 5 exibe os resultados obtidos com os diferentes regressores.

4. DISCUSSÃO

Em [7], os resultados obtidos para a predição das 6 amostras nos 5 modelos elaborados são apresentados na Figura 4. Cada curva representa a modelagem baseada na seleção de variáveis independentes, que são: NDVI médio (curva azul), NDVI máximo (curva laranja), componentes principais multiespectrais (curva verde), componentes principais LiDAR (curva vermelha) e componentes principais multiespectrais com LiDAR (curva cinza). A Figura 5 mostra que os resultados obtidos com os demais regressores explorados neste trabalho foram melhores do que aquele apresentado na Figura 4, com exceção do regressor MLP.

A Tabela 1 sumariza as médias dos valores encontrados, por meio dos regressores estudados neste trabalho, para os diferentes modelos apresentados em [7], além de compará-los com o regressor originalmente apresentado naquele trabalho. Nela, é possível notar os melhores valores obtidos nas células em destaque, onde, para o modelo de NDVI médio, SVR e Random Forest apresentaram melhores resultados e para todos os demais modelos, o regressor AdaBoost obteve as melhores predições.

O problema central abordado neste trabalho trata da seleção de regressores em um cenário onde há diferentes deles disponíveis. O estudo apresenta dados experimentais que permitem a comparação entre os resultados obtidos com cada um deles. Foi feita a sumarização dos resultados encontrados na Tabela 1 e, neste conjunto de dados e com estas modelagens de características, o regressor AdaBoost

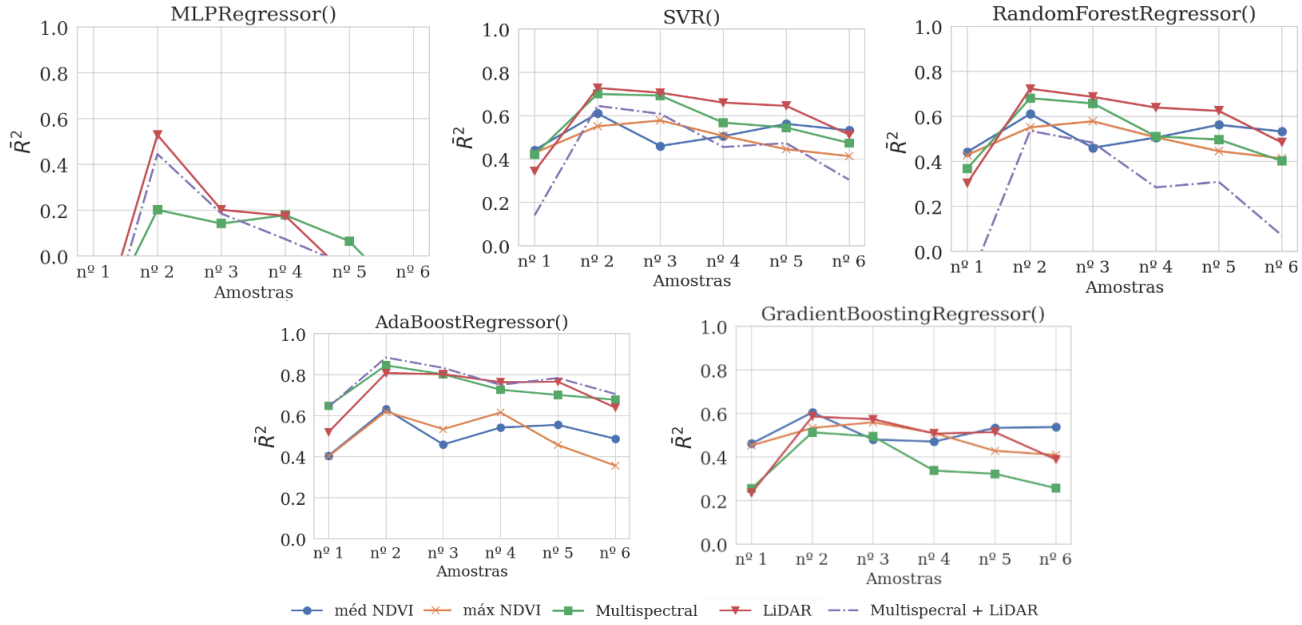

Figura 5: Regressores MLP, SVR, Random Forest, AdaBoost e Gradient Boosting.

	méd NDVI	máx NDVI	LiDAR	Multiespec	Multiespec+LiDAR	RMSE
OLS	0.17	0.18	0.40	0.42	0.39	5.87
MLP	-0.46	-0.42	0.00	0.04	-0.03	7.78
SVR	**0.52**	0.48	0.63	0.64	0.39	5.81
Random Forest	**0.52**	0.48	0.63	0.62	0.30	4.43
AdaBoost	0.51	**0.50**	**0.64**	**0.65**	**0.50**	**4.04**
Gradient Boosting	0.51	0.49	0.49	0.54	0.00	4.62

Tabela 1: Sumarização de médias de valores de $\overline{R}^2$ para os diferentes modelos, obtidos a partir dos regressores avaliados neste trabalho; bem como o RMSE calculado para cada regressor.

atingiu os melhores resultados, tanto de $\overline{R}^2$ quanto de RMSE (*Root Mean Square error*). Dessa forma, há indicativos de que explorar comparativamente diferentes regressores pode melhorar os resultados obtidos na predição de produtividade de cana-de-açúcar.

5. CONCLUSÕES

Em relação ao único regressor utilizado em [7], neste trabalho foram obtidos melhores resultados, utilizando o mesmo protocolo e os mesmos dados (Tabela 1). Este trabalho é o primeiro estudo comparativo entre regressores para estimar produtividade de cana-de-açúcar tendo como variáveis independentes os índices vegetativos extraídos de imagens multiespectrais capturadas por meio de VANT.

Em [9], também é realizada tal comparação, porém sem dados de LiDAR e avaliando outros regressores. O trabalho apresentado em [10] trata da comparação entre modelos de regressão para estimativa de produtividade nacional de cana da Índia, porém, por meio de imagens de satélite, utilizando como variáveis independentes a produtividade nos maiores estados produtores.

6. REFERÊNCIAS

[1] R Gommes, M Bernardi, and F Petrassi. Agrometeorological crop forecasting. *Sustainable Development Dimensions, FAO website, www. fao. org/waicent/faoinfo/sustdev/Eidirect/AGROMET/FORECAST. HTM*, 1996.

[2] Allard de Wit, Hendrik Boogaard, Davide Fumagalli, Sander Janssen, Rob Knapen, Daniel van Kraalingen, Iwan Supit, Raymond van der Wijngaart, and Kees van Diepen. 25 years of the wofost cropping systems model. *Agricultural Systems*, 168:154–167, 2019.

[3] Jeferson Lobato Fernandes, Nelson Francisco Favilla Ebecken, and Júlio César Dalla Mora Esquerdo. Sugarcane yield prediction in brazil using ndvi time series and neural networks ensemble. *International Journal of Remote Sensing*, 38:4631–4644, 2017.

[4] Ana Cláudia dos Santos Luciano, Michelle Cristina Araújo Picoli, Daniel Garbellini Duft, Jansle Vieira Rocha, Manoel Regis Lima Verde Leal, and Guerric le Maire. Empirical model for forecasting sugarcane yield on a local scale in brazil using landsat imagery and random forest algorithm. *Computers and Electronics in Agriculture*, 184:106063, 2021.

[5] W.A. Dorigo, R. Zurita-Milla, A.J.W. de Wit, J. Brazile, R. Singh, and M.E. Schaepman. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International Journal of Applied Earth Observation and Geoinformation*, 9:165–193, 2007.

[6] Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177:105709, 2020.

[7] Yuri Shendryk, Jeremy Sofonia, Robert Garrard, Yannik Rist, Danielle Skocaj, and Peter Thorburn. 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*, 92:102177, 2020.

[8] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[9] Sharareh Akbarian, Chengyuan Xu, Weijin Wang, Stephen Ginns, and Samsung Lim. An investigation on the best-fit models for sugarcane biomass estimation by linear mixed-effect modelling on unmanned aerial vehicle-based multispectral images: A case study of australia. *Information Processing in Agriculture*, 2022.

[10] Kiran Kumar Paidipati, Arjun Banik, Bhavin Shah, and Narpat Ram Sangwa. Forecasting of sugarcane productivity estimation in india - a comparative study with advanced non-parametric regression models. *JOURNAL OF ALGEBRAIC STATISTICS*, 13:760–778, 2022.