International Journal of Technology

http://ijtech.eng.ui.ac.id

Research Article

Design of a Sugarcane Yield and Productivity Prediction Model

Napthaleni [1], Muhammad Asrol [1,*]

[1]*Industrial Engineering Department, BINUS Graduate Program – Master of Industrial Engineering, Bina Nusantara University, Jakarta, 11480, Indonesia*
**Corresponding author: muhammad.asrol@binus.edu, Tel.: +62 21 5345830*

Abstract: Indonesia's sugar industry has yet to become self-sufficient in sugar production. This is due to the unpredictable and fluctuating relationship between yield (sugar content [%]) and sugarcane productivity (Ton/Ha) in all state-owned and private Indonesian sugar mills. As a result, Indonesia's domestic sugar consumption is still balanced by sugar imports. This study aimed to identify the main criteria and predict sugarcane yield and productivity using vegetative growth indicators in sugarcane cultivation using data science and a machine learning technique based on SVR and RF. The study found that the essential features for predicting sugarcane yield are clear juice, Pol, purity, Brix, and maturity factor, whereas the number of stems, stem height, stem weight, rainfall, and juring factor are important for predicting sugarcane productivity. The best model to predict sugarcane yield (%) was generated using RF with an average absolute error rate of 0.074% and accuracy in predicting yield with an average absolute percentage error of 0.010% and a sugarcane yield prediction error rate of 0.129%. The best sugarcane productivity prediction model was generated using SVR with an average absolute error rate of 0.051 tons/ha and accuracy in forecasting productivity with an average absolute percentage error of 0.001% and a sugarcane productivity prediction error rate of 0.058 tons/ha. This model may be used to optimize sugar cane cultivation and harvesting times, resulting in increased productivity and yields, which benefits corporate performance and increases national sugar output.

Keywords: Prediction; Productivity; Random forest; Sugarcane yield; Support vector regression

1. Introduction

The state of Indonesia's sugar agro-industry has barred it from reaching food self-sufficiency. As a result, the government has designated sugar as a strategic priority. According to Indonesia BPS (2023), national sugar production in 2022 will be 2.41 million tons, while the need for sugar in Indonesia will be 7.3 million tons. With a consumption sugar of 3.2 million tons and the sugar needs of industry of 4.1 million tons, there is a gap in sugar needs, and the government imports sugar to cover the deficit and stabilize sugar prices due to high demand and low supply (BPS, 2023). The sugar agroindustry's problem is directly related to the upstream side, notably sugarcane yield and productivity. (Rafiqi et al., 2023) asserted that low productivity and yield (produce) are internal issues that impede the development of alternative policies for sugar self-sufficiency. (Sulaiman et al., 2023) added that the lack of sugar output is due to diminishing land area, Low sugar cane

This work was supported by the research grant from the Ministry of Education, Culture, Research and Technology, Republic of Indonesia, under the Penelitian Tesis Magister 2024 research scheme (contract numbers: 105/E5/PG.02.00.PL/2024; 784/LL3/AL.04/2024; 092/VR.RTT/VI/2024), which was awarded on June 21, 2024.

https://doi.org/10.14716/ijtech.v16i5.7233
Received August 2024; Revised September 2024; Accepted January 2025

productivity and yield. (BPS, 2023) demonstrates that the area of sugar cane plantations in Indonesia increased by 74,350 ha, or 17.9%, over the last five years (2018–2022). This indicates that the link between sugar output and sugarcane crop area is nonlinear.

According to data collected by the Directorate General of Plantations, Ministry of Agriculture, all Indonesian sugar mills' productivity and sugar cane production fluctuated in 2023. An unstable relationship pattern between sugarcane yield and productivity is revealed. This phenomenon creates uncertainty related to the correlation between increasing yield and increasing sugarcane productivity. The sugar mill has been conducting sugarcane maturity analysis to predict the appropriate harvesting time so that the processed sugarcane is in optimum condition, which has been done every two weeks since the plants were 8 months old, with yield samples utilized. (Indrawanto et al., 2017). Sugarcane production is estimated twice a year, in December and March, to predict the number of milling days needed. Owing to the high plant variability and the significant influence of environmental factors, sugarcane maturity analysis and production estimation are susceptible to errors. Additionally, a simple prediction model with linear regression is unable to capture the complexity of the interaction between multiple factors that affect sugarcane maturity and production. Thus, an accurate prediction model that can manage complex data, adjust to change, and find complex patterns can increase the efficiency of sugar production and make more accurate decisions while managing sugarcane plantations. Furthermore, accurate agricultural production forecasts may boost industry sustainability by improving both environmental and economic consequences (Everingham et al., 2016), and sugar shortages can emerge if sugar production changes are not adequately predicted (Jaelani et al., 2022).

Previous research has modeled predictions of sugarcane productivity and yield, such as: (Jaelani et al., 2022) long short-term memory (LSTM) machine learning methods and linear regression using annual agency and journal data from 1968 to 2020 with year variables, sugarcane production, sugar production, sugar consumption, and population (Respati, 2022) Sugar production for 2023-2026 was forecasted using the ARIMA method, VAR, and a transfer function based on sugar production data from a training (1972-2016) and testing series (2017-2022). The three techniques produce the same growth estimate but based on the mean absolute percentage error (MAPE) value, the VAR method has the lowest value for testing series data, indicating that it is appropriate for estimating Indonesian sugar production. (Paidipati et al., 2022) used adaptive regression methods, such as multivariate splines (MARS), support vector regression (SVR), partial least squares regression (PLSR), elastic net regression, and multiple linear regression (MLR), to estimate sugarcane productivity in India, with SVR outperforming other regression non-parametric methods. Asrol et al. (2020) used the relief methodology and the support vector machine (SVM) method, where soil pH, humidity, and sugarcane age are the primary elements influencing sugar content.

The SVM method can be utilized to estimate sugar content and harvest time for sugarcane mills. Hammer et al. (2020) developed a model to predict sugarcane productivity using the random forest (RF), gradient boosting machine (GBM), and SVM methods and identified the main variables that influence sugarcane yields according to their relative importance using an operational data set from 18 sugar factories during three growing seasons, including variety, soil type, age of sugarcane, average air temperature, rainfall, wind speed, and solar. It was found that the SVM-generated one was marginally superior. Gaffar and Sitanggang (2019) used the SVR approach to develop a sugarcane productivity prediction model based on climate parameters that the model utilized performs rather well in estimating sugarcane productivity. Shah et al. (2018) forecast agricultural yields per hectare from crop yield and meteorological data using three regression-based methods: multivariate polynomial regression (MPR), SVM, and random forest. Support vector machine regression is the best method for predicting crop yield. Compared with multiple linear regression and decision tree regression, RF regression is a very successful technique for predicting sugarcane crop yields (Erick et al., 2023) and has shown effectiveness in tasks such as regression (Lárraga-Altamirano et al., 2024). In their research, Maldaner et al. (2021) used artificial neural networks (ANNs), RFs, and MLRs to estimate sugarcane yield, and RFs proved to be the most effective model.

There is a knowledge gap in understanding the critical elements for enhancing sugarcane yield and productivity, and earlier studies only projected sugarcane yield or productivity independently, making a correlation between productivity and yield unknown. In addition, the yield and productivity of sugar cane in Indonesian sugar factories are still determined manually, and cane maturity is measured through assessments in March and December.

Therefore, this study aimed to identify important features in predicting sugarcane yield and productivity using vegetative growth indicators in sugarcane cultivation using data science and machine learning techniques. Data for this study were collected from sugar plants in Malang and Madiun, East Java, Indonesia. These secondary data was gathered from observations of vegetative growth and analytical findings in 2023. The techniques used in this study are based on SVR and RF regression. These approaches were chosen based on earlier research, which found that SVR and RF were the top performing regression-based ML algorithms. The data utilized in this study are continuous and numeric with a numerical output. A continuous dependent variable can be predicted from a series of independent inputs using regression analysis (Panigrahi et al., 2022). The selection of the regression method should evaluate the various variables considered, as well as the type and distribution of the data (Tatachar, 2021).

This paper is organized to help solve the problems faced by the sugarcane industry, provide innovation, and contribute to the development of accurate predictive models to recommend crop decisions to maximize production in uncertain conditions. According to Bocca and Rodrigues (2016), crop yield models may help decision-makers in any agro-industrial supply chain, even when they relate to issues unrelated to crop production.

2. Related Works

Machine learning is the process of automatically identifying designs in data without making any assumptions about the data structure (Noorsaman et al., 2023). According to Van Klompenburg et al. (2020), machine learning is a useful technique for forecasting crop yields, determining what crops should be grown, and what must be done with these crops during their growth season. A subfield of artificial intelligence known as "machine learning" looks for patterns and links in past data to forecast or make choices. Harvest yield prediction may be developed using various machine learning techniques, including regression, classification, grouping, and prediction. Several algorithms, including ANNs, decision trees, support vector machines, naïve Bayes, and linear and logistic regression, can also be used (Palanivel and Surianarayanan, 2019). (Singla et al., 2020) used remote sensing data to forecast sugarcane yield using ensemble machine learning. (Mahesh, 2020) reported that machine learning techniques are frequently used. The type of model that is most suitable, the number of variables, and the type of issue to be addressed all influence the type of method employed.

Machine learning has been applied in many sectors, including developing prediction models for many important features in the production process of the industry. In machine learning, the prediction model is divided into 3 parts, numerical prediction, classification, and time series prediction. Since the research predicted the yield and productivity, which are stated numerically, we will explore the numerical prediction. Previous studies in predicting sugarcane yield and productivity have used algorithms such as multiple linear regression, decision tree regression, random forest regression, multivariate adaptive regression splines, support vector regression, partial least squares regression, and K-nearest neighbors, as shown in Table 1. The random forest and support vector regression have the best performance in determining productivity and yield. Regression-based machine learning builds models that forecast numerical (continuous) values based on input data by applying statistical approaches. Regression approaches, such as random forest (RF) and support vector regression (SVR), are commonly employed. Regression and classification can be accomplished using the supervised machine learning algorithm RF. RF represents decision tree ensemble learning (Charoen-Ung and Mittrapiyanuruk, 2019). SVR is a regression approach that maps input to output via a support vector machine. SVR searches for a

hyperplane that maximizes the margin between the data and the hyperplane, resulting in a more stable and resistant to overfitting (Smola and Schölkopf, 2004).

Table 1 Previous studies on productivity and sugarcane yield prediction

No	Author(s)	Year	Method	Features
1	Erick et al.	(2023)	Multiple linear regression, decision tree regression, and random forest regression	Sugarcane yield, area, sugarcane age, sugarcane crop cycle, temperature, rainfall, soil, pH
2	Jaelani et al.	(2022)	Long- and short-term memory	Year, sugarcane production, sugar consumption, population
3	Paidipati et al.	(2022)	Multivariate adaptive regression splines, support vector regression, partial least squares regression, elastic-net regression, multiple linear regression, t-test regression, t-test regression	Sugarcane yield, production, and area
4	Maldaner et al.	(2021)	Keywords: Multiple linear regression, random forest, artificial neural network	CAN data from the sugarcane harvester
5	dos Santos Luciano et al.	(2021)	Random forest	Remote sensing data, meteorological data (solar radiation, wind speed, relative humidity, precipitation, minimum and maximum temperature), and agronomic data (soil, variety, yield, harvest date, number of harvests, and production environment)
6	Asrol et al.	(2020)	Support vector machine (SVM)	Soil pH, temperature, rainfall, humidity, sugarcane age, area height, early sugar content, Pol, Brix,
7	Hammer et al.	(2020)	Keywords: Random forest, gradient boosting machine, support vector machine	Varieties, soil type, number of sugarcanes cut, sugarcane age, planting spacing, rainfall, average air temperature, wind speed, and solar radiation
8	Singla et al.	(2020)	Keywords: Support vector regression, random forest, K-nearest neighbors, classification trees, regression trees	Satellite data
9	Charoen-Ung and Mittrapiyanuruk	(2019)	Random forest	Cane class, water type, soil type, area, fertilizer, rainfall, distance, contract area,
10	Gaffar and Sitanggang	(2019)	Support vector regression	Year, area, province, minimum temperature, maximum temperature, average humidity, rainfall, duration of sunshine, wide area, production, productivity

Machine learning applications for forecasting must be developed through several important stages, including data preprocessing, modeling, and evaluation. Feature selection techniques may be applied to data pre-processing to accomplish effective data reduction, according to (Jović et al., 2015). This helps locate precise data models. By eliminating redundant and unnecessary data,

feature selection offers a practical solution to the problem. This can speed up computation, increase learning accuracy, and improve the understanding of the learning model or data (Cai et al., 2018). Many techniques have been applied for feature selection, including statistical tests, correlations, clustering, and feature-important analysis. The optimal feature is chosen using the univariate feature selection approach, which is based on statistical tests such as the best scoring feature, best percentile feature, false positive rate, false discovery rate, family-wise error, and hyper-parameter search estimator (Medar et al., 2019). Correlation features may also be applied to feature selection (Chen et al., 2021; Mohamad et al., 2021). Table 1 presents some previous studies on sugarcane productivity and yield.

Based on the numerous investigations shown in previous research, it is often restricted to forecasting sugar cane output or productivity. In addition, the primary factors influencing sugar cane productivity are not identified, and the range of variables examined is constrained and typically depends on climate-related data. (Asrol et al., 2020) defined the major criteria. However, there was no evaluation of the performance of machine learning. This research will identify the primary factors influencing productivity and yield using assessment data and a preliminary analysis of sugar factories during the 2023 milling season. The features considered in this study are based on vegetative growth data from sugarcane cultivation. Two machine learning techniques, SVM and random forest, will be developed to predict sugarcane productivity and yield. Furthermore, the effectiveness of the ML model will be assessed for further implementation in decision-making at the mill.

3. Methods

3.1. Research Flow

The research workflow is shown in Figure 1. Data were collected from the sugar agroindustry in Malang and Madiun, East Java, Indonesia. These are secondary data generated from observations of vegetative growth of sugarcane and the results of sugarcane analysis in 2023, which will be useful in predicting sugarcane productivity and yield. Second, after data collection and acquisition, the raw data must be pre-processed by cleaning, simplifying, and transforming. Data must be transformed into an accurate, relevant, and consistent dataset for further analysis and modeling. This stage includes verifying the raw data to ensure that they are useful and useful for use.

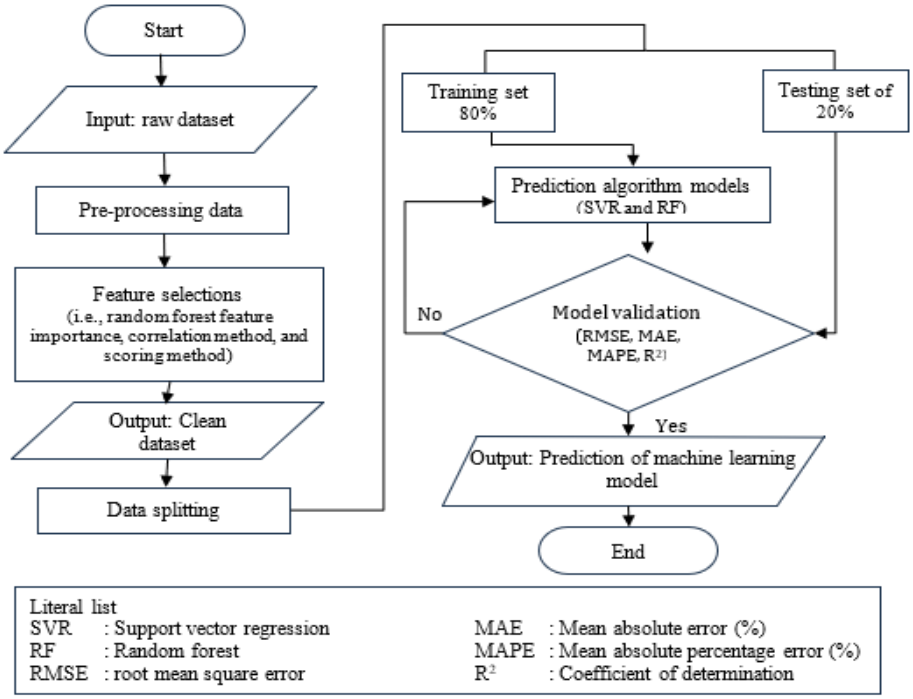


Figure 1 The research workflow

Next, feature selection is performed using the correlation filter, scoring, and random forest feature importance methods, which are used to determine the important features for each prediction target. In data analysis, feature selection is used to improve model performance by eliminating unnecessary information. After the feature selection step, the entire dataset is separated into a training set and a testing set, where the training data are used to build a regression-based machine learning model. When the model is trained using training data, the testing data are considered unseen data (ELhadad et al., 2022).

The prediction model is designed using a regression-based machine learning model. Previous studies have used a variety of productivity and yield prediction methods for sugarcane, such as linear regression (Jaelani et al., 2022), support vector regression (SVR) (Paidipati et al., 2022; Gaffar and Sitanggang, 2019; Shah et al., 2018), support vector machine (SVM) (Asrol et al., 2020; Hammer et al., 2020), random forest (RF) (Lárraga-Altamirano et al., 2024; Erick et al., 2023; dos Santos Luciano et al., 2021; Maldaner et al., 2021), multiple linear regression (Paidipati et al., 2022; Maldaner et al., 2021), partial least squares regression (PLSR) (Paidipati et al., 2022), and artificial neural network (ANN) (Maldaner et al., 2021). Based on these investigations, SVR and RF are the regression-based ML approaches with the best performance. Therefore, both models are used in this investigation.

The next stage is model evaluation. The SVR and RF output values were analyzed using the RMSE, MAE, and coefficient of determination (R^2) statistical models to determine the accuracy of the model used to forecast the output. The ability of a model to forecast sugarcane productivity and yield can be assessed by considering its RMSE, MAE (Shetty et al., 2020), and (R2) (Nikhil et al., 2024; Canata et al., 2021). The details of each stage are delivered in the following subsections.

3.2. Data collection and analysis

This study employs secondary data collected from manual observations of sugarcane vegetation growth and yield analysis conducted in two sugar factories in different regions during 2023. These two sugar factories have different production capacities. The sugar factory in Malang has a production capacity of 12,000 tons of cane per day with a sugarcane area of 21,838 ha, whereas the sugar factory in Madiun has a production capacity of 6,000 tons of cane per day with a sugarcane area of 9,987 ha. A total of 10 features are used for yield prediction with 2,225 rows of data whose complete data can be seen in supplementary material 1, while 13 features are used for sugarcane production forecasting with 2,656 rows of data to develop a prediction model with complete data which can be seen in supplementary material 2. The features used in forecasting sugarcane production and yield consist of two types of data, namely numeric and categorical. The features included in sugarcane production and yield forecasting consist of two data types, namely, numeric and categorical.

More than 4,500 instances have been collected consisting of related features for yield and productivity. These data represent the sugar agroindustry conditions for the current year's production. The agroindustry performs a single cycle of production for each year, and the number of instances represents more than 50% of the total dataset.

3.3. Pre-processing Data

At this stage, a data cleaning procedure is performed, during which the data are examined to identify any mistakes, inconsistencies, or anomalies that may exist in the dataset. These problems may affect the correctness and caliber of the dataset, which might include missing values, duplicate entries, outliers, improper data formats, and other data quality concerns. This stage is crucial for determining how good the data are and how much cleanup is needed. According to Pandey et al. (2020), data pre-processing is one of the most crucial stages in the creation of any ML model because it directly affects the model's quality and efficiency. If we neglect this step and create a model using datasets with missing values, the resulting model will be inconsistent and less effective. According to Sari et al. (2023), poor data quality can hamper accurate predictions when underfitting occurs due to dispersed data quality or overfitting, which restricts predictions to a small range of data.

According to Alexandropoulos et al. (2019), the data preparation process can be discretization or normalization, noise reduction, outlier detection, feature selection, instance selection, and missing value imputation based on the raw data conditions. This study outlines the steps involved in data preparation, including preprocessing, as shown in Figure 2. The first step is the data transformation technique by continuing the discrete to change the categorical data type to numeric, which provides a numeric data type for all features used. Then, the data is cleaned by removing features that have more than 5% missing data values, and features that have missing values below 5% are imputed to fill in the missing values on the features with the average value or the most frequent value in the feature. Finally, feature selection is performed. The purpose of feature selection in preprocessing is to eliminate the variables or features that are most relevant to the study's target.

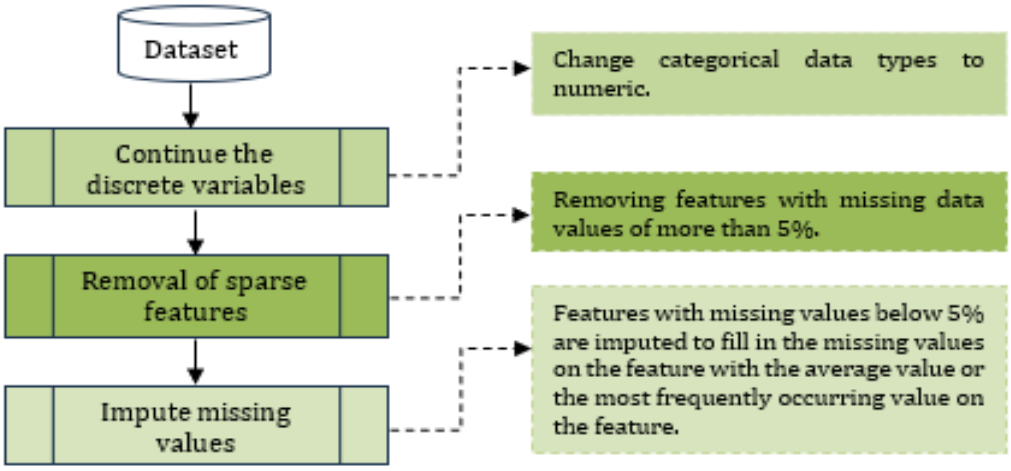


Figure 2 Preprocessing steps in data preparation

3.4. Feature Selection

There are two approaches to dimensionality reduction. Feature selection and extraction. In addition to lowering the data burden, the feature selection approach helps prevent overfitting of the model (Venkatesh and Anuradha, 2019). Feature selection by filtering is the strategy used to reduce dimensionality. The selection of characteristics most relevant to the target variable is done by applying the correlation and scoring approach known as the filter method. In machine learning models, feature selection based on importance is an essential stage as it guides the usage of variables to what works best and most efficiently for a particular machine learning model (AlSagri and Ykhlef, 2020).

To deepen the analysis, this study employs multiple methods for feature selections. This study ensures that the selected features have an important position in predicting yield and productivity. Therefore, three methods for feature selection and analysis are employed, namely, correlation methods, scoring, and feature importance by random forest. This multimethod is applied to confirm the importance of features in the prediction. The detailed technique for feature selections applied in this study is described in this section.

3.4.1. Importance of Random Forest Feature

The relative relevance of each feature and the degree to which its removal reduces accuracy or its inclusion increases accuracy may be determined using feature importance metrics (AlSagri and Ykhlef, 2020). According to Gregorutti et al. (2017), due to various important measurements, the random forest method enables us to simultaneously assess a predictor's relevance. The original random forest algorithm calculates three metrics: the z-score, Gini importance, and permutation importance. The permutation significance measure has demonstrated strong performance for top variable selection methods, among other criteria. Equation 1 is used to calculate the feature importance for each feature in the dataset:

$$\text{Feature Importance } (F_j) = y_{j\_true} - y_{j\_premuted} \quad (1)$$

Where F_j is importance of feature j, y_{j_true} is model's performance before the eth feature permutation, and $y_{j\_premuted}$ is model's performance after the eth permutation. The obtained feature importance values obtained are then normalized to be in the range of 0–1.

3.4.2. Corellation Method

The Pearson correlation coefficient is used for normally distributed numerical data by following Equation 2.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (2)$$

where a significant positive or negative correlation is indicated by a coefficient value that falls between -1 and 1. The correlation is weaker if the value is near zero, or if it is exactly zero, it implies no association at all.

3.4.3. Scoring Method

Univariate regression examines which factors among a set of available variables have the strongest individual connection with the target score. Single-feature evaluation and ranking are common outcomes of univariate feature filters (Jović et al., 2015). This model expresses the dependent variable (Y, the filter score) as a linear function of the independent variable (X, the only factor under consideration). Mathematical Equation 3:

$$y = \beta_0 + \beta_1 X + \varepsilon \quad (3)$$

Here, Y is the dependent variable (filter score), X is the independent variable (single factor), β_0 is the intercept (constant term representing the average score when X is zero), β_1 is the slope (coefficient representing the change in score for a unit increase in X), and ε is the error term (accounts for random noise and unexplained variance).

3.5. Prediction Algorithm Models

3.5.1. Support vector regression (SVR)

Supervised machine learning models called SVM are used to analyze data for regression and classification. Regression analysis was conducted using the SVM. This non-parametric regression model plays a major role in the presence of outliers and is highly helpful for predicting when nonlinearities impact the data (Paidipati et al., 2022).

The SVR model is expressed as the following functional Equation 4:

$$f(x) = \{w, \emptyset(x)\} + b, \text{w} \in \text{X}, \text{b} \in \text{R} \quad (4)$$

where w is the weight vector of the inputs, b is the bias, $\emptyset(x)$ is a kernel function. When a nonlinear input is converted into a linear input using a nonlinear function. The objective is to identify, for each training set of data, the function f(x) with the largest ε-deviations from the achieved objective yi. As long as the mistakes are inside the ε-insensitive band, they are ignored. Vapin introduced the concept of an insensitive loss function to SVR ε, which can be represented as Equation 5:

$$L_\varepsilon = (f(x) - (y) = \begin{cases} |f(x) - y| - \varepsilon & if |f(x) - y| \geq \varepsilon, \\ 0 & Otherwise \end{cases} \quad (5)$$

where the ε-insensitive area is marked by ε. There is no loss if the predicted values are inside the band region; however, if the expected values fall outside the band, the loss is equal to the difference between the anticipated value and the margin. It is possible to describe the restrictions and the goal function as Equations 6, 7, and 8:

$$\min \frac{1}{2}(w,w) + C \sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{6}$$

$$Subject\ to = ((w, \emptyset(x_i) + b) - y_i \leq \varepsilon + \xi_i, \tag{7}$$

$$y_{i-}\big((w, \emptyset(x_i) + b\big) \leq \varepsilon + \xi_i, \tag{8}$$

$$\xi_i;\ \xi_i^* \geq 0, i = 1,2,\ldots,n$$

where $(\xi_i + \xi_i^*)$ is the empirical risk, n is the quantity of training data, ξ_i and ξ_i^* are the slack variables, and C is the modifying coefficient, which provides the trade-off between training error and model complexity. The optimal value of each parameter is determined using the Lagrange function after choosing a band width (ε), kernel function (φ), and altering coefficient (C).

3.5.2. Random Forest

The second technique employed is regression using random forests. (Criminisi et al., 2012) The average of all outputs of the tree. The t-th tree at input point v follows Equation 9:

$$p\,(y|v) = \frac{1}{T}\sum_{t}^{T} p_t(y|v) \tag{9}$$

T represents the total number of trees in the random forest.

3.6. Model Evaluation

The prediction model's performance is evaluated by comparing the predicted values to the actual observed values using the RMSE and MAE metrics. Model evaluation is performed to determine the accuracy with which the model predicts output and the extent of the error in the outcomes. The RMSE is used to determine the accuracy of the model, and the MAE is used to determine the magnitude of the error in the results (Hammer et al., 2020).

The RMSE is calculated by the formula in Equation 10:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \check{y}_i)^2} \tag{10}$$

Where, y_i is actual value for the with data point, $\check{y}_i$ is Predicted value for the with data point. The average of all absolute differences is calculated and expressed in Equation 11:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \check{y}_i|^2 \tag{11}$$

The MAPE is used to assess forecast accuracy across many series with varying scales (Hyndman, 2014). Measuring the average absolute percentage error is shown in Equation 12.

$$MAPE = \frac{\sum_{i=1}^{n}|y_i - \check{y}_i|/y_i}{n} \times 100\% \tag{12}$$

The determination coefficient (R-squared) is a standard metric for assessing regression analysis in any scientific subject (Chicco et al., 2021). R-squared follows the mathematical Equation 12:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(\check{y}_i - y_i)^2}{\sum_{i=1}^{m}(\bar{y} - y_i)^2} \tag{13}$$

The projected with value is represented by $\check{y}_i$ in the following formulas, and the actual with value is represented by the y_i element. For each y_i element in the ground truth dataset, the regression approach predicts the $\check{y}_i$ element.

4. Results and Discussion

4.1. Pre-processing

The dataset was collected from the estimated sugarcane production and sugarcane yield analysis data collected in 2023 at two sugar factories located in East Java, Indonesia. The dataset shows variability because the observations of sugarcane fields are in several areas, not only in Malang and Madiun but also around them. Sugar factories estimate sugarcane production data twice a year in March and December. In this study, the estimated sugarcane production data was taken from observation data in March for the 2022-2023 sugarcane planting season because the sugarcane milling production process in Indonesia starts in May to early November, when the sugarcane production estimation observation data in December shows that the sugarcane plants have not grown optimally. Thus, the observation data has a high error rate that can impact model performance.

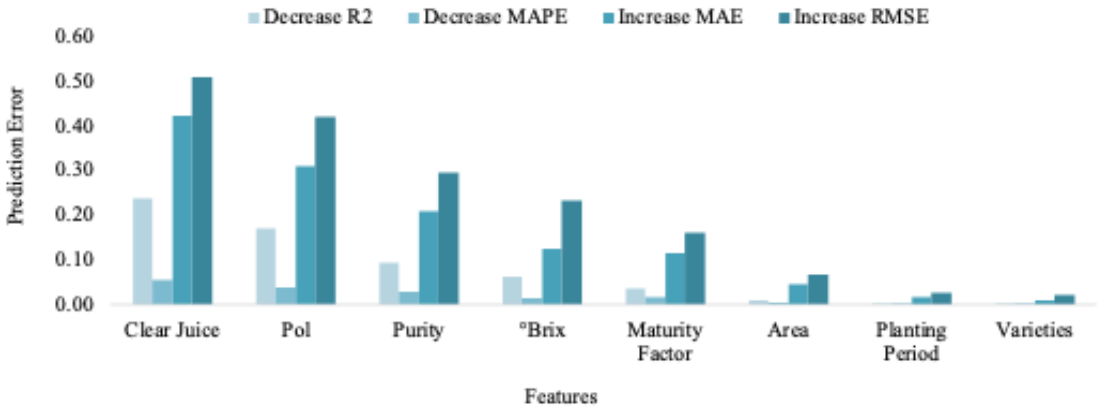
After collection, the raw dataset must be handled to ensure data validity with the preprocessing stage. Among the dataset preprocessing steps are imputing missing values by deleting rows that contain missing values or using the most common value (for discrete attributes) or the average value (for continuous attributes). Subsequently, continuous discrete variables are applied to categorical-type features to convert the data to numeric by treating the data as ordinal using standard procedures, and feature gaps are eventually eliminated by deleting features that have missing values greater than 5% of the time.

In the preprocessing stage, in terms of productivity, the preprocessing results showed that 13 features became 12 features with a total of 2656 rows of data, whereas in the preprocessing stage there was 1 feature that had a missing value of 59%, so this feature was removed. Meanwhile, for sugarcane yield, the findings showed a decrease from 10 features to 8 features with 2225 rows data with 2 missing due to having a missing value of 40%. Descriptive statistics before preprocessing can be fully seen in supplementary material 3 for sugarcane yield and supplementary material 4 for productivity, and after preprocessing, complete data can be seen in supplementary material 5 for sugarcane yield and supplementary material 6 for productivity.

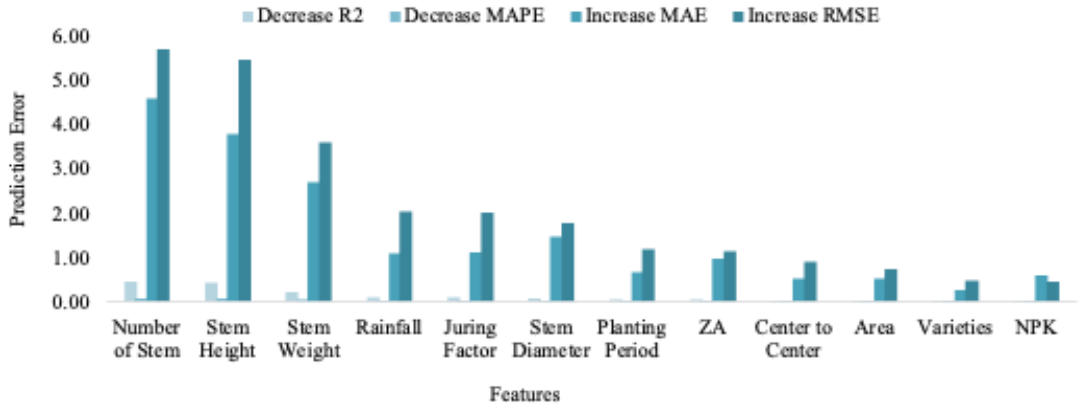
4.2. Feature engineering and selection

The contribution of each feature to the prediction of sugarcane yield and productivity was determined through feature selection using the random forest technique. This was done by measuring the increase in model prediction error after randomizing the feature values, which destroys the relationship between the features and targets. The importance of variables is determined by fitting a model that includes all predictors and updating the model after permuting each variable. The link between each predictor and the result is then examined (Maldaner et al., 2021). Figure 3 shows the results of the feature importance random forest processing in this study.

The plot features are sorted based on their relevance. The results of the feature importance show that the number of stems is the most important feature in the prediction of sugarcane productivity, and clear juice is the most important feature in the prediction of yield. In their study, AlSagri and Ykhlef (2020) revealed that the five most and least significant features were removed separately in the analysis, and the RF was recalculated. In this study, the main features, namely, clear juice, Pol, purity, Brix, and maturity factor, were selected to be used as training and testing data in the yield prediction modeling. In addition, features such as the number of stems, stem height, stem weight, rainfall, and juring factor will be used to predict sugarcane productivity. Features with high weights were considered important. The larger the weight of the feature, the greater the probability that this results in feature importance sampling-based adaptive random forest (ARF) selecting the feature (Cao et al., 2011)



(a)



(b)

Figure 3 Variable importance of the model's performance (a) sugarcane yield and (b) productivity

Based on the relationship between two dependent and independent variables using the correlation and scoring methods, there are five features that have a strong correlation relationship and most influence the dependent variable if there is a change in the independent variable, namely, clear juice, purity, Pol, maturity factor, and °Brix in yield prediction modeling, while the features in sugarcane productivity modeling are the number of stems, stem weight, rainfall, juring factor, and center-to-center, which can be seen in Tables 2 and 3. Both methods provide weight-based characteristics where both techniques produce the same features in producing scores from highest to lowest, where the higher the weight indicates that the features have a strong relationship and influence each other in determining the prediction target. Statistical measurements are used in the filter approach to assign a score value to each feature. The features are ranked and arranged in descending order according to their scores (Venkatesh and Anuradha, 2019).

Table 2 Selected features of the sugarcane yield

Features	Pearson Correlation	Univariate Regression
Clear Juice	0.982	58,589.42
Purity	0.865	6,615.97
Pol	0.802	4,011.35
Maturity Factor	0.693	2,056.48
°Brix	0.559	1,011.39
Planting Period	0.314	243.97
Varieties	0.188	81.24
Area	0.134	40.42

4.3. Modeling

The prediction model is designed using 80% and 20% of the training and testing data, respectively. SVR with polynomial kernel functions was used in this study. The findings of the experiment (Cheng et al., 2007) that combined spatial and temporal dimensions nonlinearly showed that using SVMs for nonlinear regression increased prediction accuracy compared to using linear regression and other conventional methods. The kernel method provides a highly effective way to add nonlinearity to the SVR (Joshua et al., 2022). The goal of SVR is to identify the optimal line for the provided data. In this case, the hyperplane is the optimum line. The data are converted into the desired format using a mathematical function known as a kernel, and borders are created at a distance ε that indicates the margin between data points (Tatachar, 2021). C is also an essential parameter that governs the trade-off between increasing the margin and reducing the training error (Cheng et al., 2007). This study uses the SVR parameters with values using default settings. Hanka and Santosa (2021) revealed that polynomial SVM is the best prediction method compared to RBF SVM and KNN, using a kernel degree of 3.0, gamma 44, and a cost (C) value of 1.00. In evaluating sugarcane yields in Thailand, Som-ard et al. (2024) used SVR with the model's optimal hyperparameters of cost (C) 1.00 and gamma 0.1 using the RBF kernel.

Table 3 Selected productivity features

Features	Pearson Correlation	Univariate Regression
Number of stem cells	0.529	1,032.84
Stem Weight	0.497	869.62
Rainfall	0.469	749.97
Juring Factor	0.404	519.08
Center to Center	0404	519.08
Planting Period	0.285	234.45
Stem Diameter	0.279	224.65
Area	0.230	148.62
Stem Height	0.169	78.35
ZA	0.159	65.87
Varieties	0.100	26.66
NPK	0.068	12.26

The complete default settings for RF regression can be seen in supplementary material 7. The model consists of 10 decision trees, and each node requires at least 5 data samples before further separation to prevent overfitting on nodes with insufficient data. The parameter settings of the random forest model are configured with replicable training parameters to ensure that the results received each time the model is trained are the same. This is beneficial to ensure learning repeatability. Next, sugar cane yield and productivity forecasts were obtained by validating the model using test data.

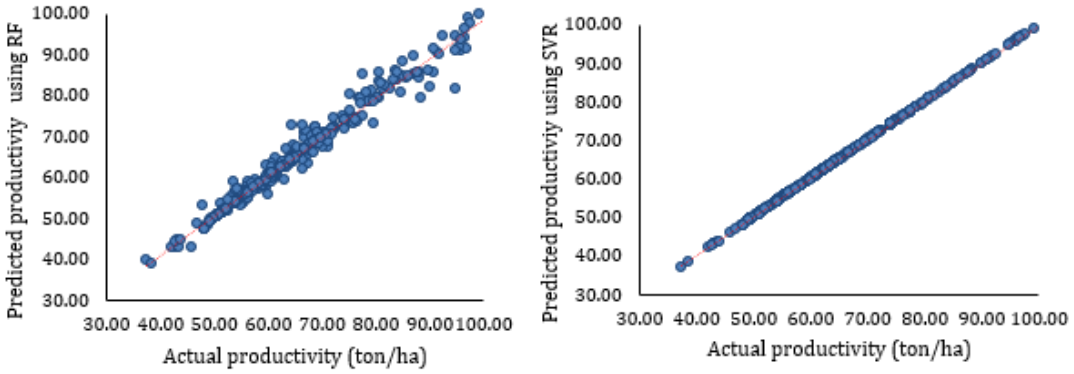
4.4. Models Performance

Sugar cane yield and productivity forecasts were obtained by validating the model using test data. Subsequently, we contrasted the productiveness and sugarcane yield of datasets with predicted sugarcane yield and production values calculated using RMSE, MAE, MAPE, and R-squared. Singla et al. (2020) studied the performance and behavior of a predictive model used to estimate sugarcane yield using performance assessment criteria such as MAE, RMSE, and R2. The best results based on the 5 features, which are the main criteria in predicting sugarcane yield and productivity, can be seen in Table 4, where the SVR model for predicting sugarcane productivity has an average absolute error rate of 0.051 tons/ha and an accuracy value in predicting productivity with an average absolute percentage error of 0.001% and a prediction error rate of 0.058 tons/ha with data variability of 100%.

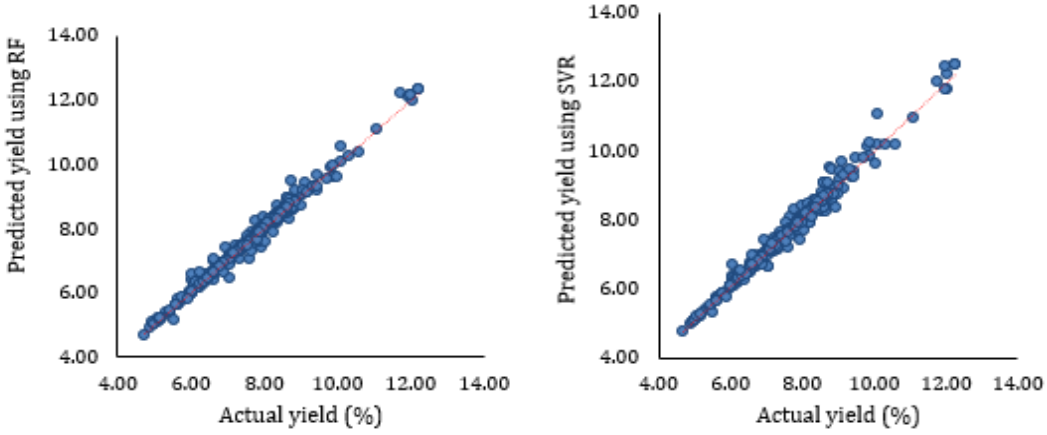
The RF model was used to determine the yields with a prediction error rate of 0.129%, an average absolute error difference of 0.074%, and an accuracy value in forecasting yields with an average percentage absolute error of 0.010% and data variability of 98.8%. Figure 4 compares the actual and predicted productivity and yield values using SVR and RF, respectively. In this test, the proposed algorithm has slightly different values from the predicted and actual values. This indicates that the model can predict productivity and yield soon according to model input and parameters.

Table 4 Performance evaluation statistics of machine learning models

		Productivity				Sugarcane Yield	
		RF	SVR			RF	SVR
RMSE	Ton/ha	1.621	0.058	RMSE	%	0.129	0.151
MAE	Ton/ha	0.740	0.051	MAE	%	0.074	0.090
MAPE	%	0.011	0.001	MAPE	%	0.010	0.011
R^2		0.977	1.000	R^2		0.988	0.983



(a)



(b)

Figure 4 Actual vs. predicted values for (a) productivity and (b) sugarcane yield

This study indicates that the optimal model for predicting sugarcane production is SVR, whereas the most effective model for forecasting sugarcane yield is RF techniques. Furthermore, the findings of this study facilitate efficiency and expediency in estimating productivity and sugarcane production within the sugar companies, thereby aiding in decision-making regarding harvest timing, workforce readiness, and factory operations. The digital shift in sugarcane productivity and yield forecasting can be seen in the figure in supplementary material 8. The existing manual procedure takes a long time to obtain results. The data obtained from the field is then manually summarized into a database that takes a long time and then calculated manually using existing formulas, so the results obtained have not been calculated for the level of accuracy of the results.

However, by using machine learning, the results can be known more quickly and accurately. The obtained models can be developed in an integrated manner with the smart farming system, allowing the data received in the field to be directly input into a model whose results can be immediately known. A digital transformation approach with the implementation of a machine learning model to predict sugarcane yield and productivity may improve business process efficiency and lead time.

The model is retrained at least once a year because the sugar mill analyzes the potential yield and productivity of sugarcane every year before the sugarcane milling season. The resulting prediction model can be developed by adding environmental factors due to potential changes in environmental conditions.

4.5. Managerial implications and contributions

This study demonstrates that regression-based techniques can be used to predict sugarcane productivity and yield. SVR is the best model for predicting sugarcane production, while the RF methodology is the most successful model for predicting sugarcane yield. The findings of this study indicate that sugarcane yield and productivity potential may be more accurately predicted by improving key parameters that influence their target to achieve better outcomes. Furthermore, this study presents yield prediction findings using the RF model that are superior to those of Maldaner et al. (2021), with an average absolute error rate of 0.074% as opposed to 5.6%. Furthermore, Medar et al. (2019) showed an accuracy level of 83.49% in forecasting sugarcane harvest outcomes, whereas the SVR model's prediction results for sugarcane productivity in this study have an average absolute prediction error rate of 0.001%.

This study has practical implications for production planning, including adjusting production targets, machine readiness, labor, and raw material supplies. Knowing the potential yield and productivity of sugarcane makes it possible to determine the best time to harvest the crop and start the milling period. Moreover, ML supports quick and accurate decision-making.

This research helps the government make decisions about the amount of sugar import quota to meet domestic needs and keep consumer sugar prices stable by estimating the potential amount of sugar production that will be produced. This information will help the government to determine the balance of sugar commodities. Furthermore, the academic contributions of this research demonstrate that the regression model, which differs from earlier models, may be utilized to estimate sugarcane yield and productivity. The performance of the model generated in this study also displays positive outcomes. This study's exploration has a lot of potential to increase the productivity and efficiency of the sugar industry by combining multiple ML models to decrease bias, enhance model generalization, and increase prediction accuracy by combining heterogeneous data, including soil, weather, satellite imagery, and field sensor data.

5. Conclusions

In conclusion, this study successfully designed a prediction model for sugarcane yield and productivity based on ML using yield and vegetation growth analysis data. The important feature random forest method was used to determine features that can improve the performance of the model. There are 5 important features in developing an accurate yield prediction model, including clear juice, Pol, purity, Brix, and maturity factor, while the number of stems, stem height, stem weight, rainfall, and juring factor are used as input variables to train the RF and SVR models. The prediction model designed from the important attributes shows that SVR is the best model for predicting sugarcane productivity. The yield prediction model is derived from the RF model. This study presents a feasible approach to determining sugarcane yield and productivity through data-based methods, moving from old manual techniques to digital transformation, which is very important in industrial applications. For further research, it is important to apply the model to support the decision-making process in the sugar mill, especially in the harvesting time schedule, and to simulate sugar production efficiency.

Acknowledgements

We would like to express our gratitude to the research grant from the Ministry of Education, Culture, Research and Technology, Republic of Indonesia, under the Penelitian Tesis Magister 2024 research scheme (contract numbers: 105/E5/PG.02.00.PL/2024; 784/LL3/AL.04/2024; 092/VR.RTT/VI/2024), which was awarded on June 21, 2024.

Author Contributions

Conceptualization, M.A; methodology, M.A. and N; software, N; validation, M.A. and N; formal analysis, M.A. and S; investigation, N; resources, M.A. and N; data curation, N; writing—original draft preparation, N; writing—review and editing, M.A.; visualization, N; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Conflict of Interest

The authors have declared no conflict of interest

Data availability statement

The data that support the findings of this study are available in Figshare at DOI: 10.6084/m9.figshare.29955986

References

Alexandropoulos, SAN, Kotsiantis, SB & Vrahatis, MN 2019, 'Data preprocessing in predictive data mining', *The Knowledge Engineering Review*, vol. 34, article e1, https://doi.org/10.1017/S026988891800036X

AlSagri, H & Ykhlef, M 2020, 'Quantifying feature importance for detecting depression using random forest', *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 628-635, https://doi.org/10.14569/IJACSA.2020.0110577

Asrol, M, Marimin, M & Yani, M 2020, 'Business intelligence model construction to improve sugarcane yield for the sustainable sugar industry', *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, special issue 6, pp. 109–118, http://dx.doi.org/10.5373/JARDCS/V12SP6/SP20201013

Badan Pusat Statistik (BPS) 2023, *Statistik tebu Indonesia 2022*, Badan Pusat Statistik, Jakarta

Bocca, FF & Rodrigues, LHA 2016, 'The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling', *Computers and Electronics in Agriculture*, vol. 128, pp. 67-76, https://doi.org/10.1016/j.compag.2016.08.015

Cai, J, Luo, J, Wang, S & Yang, S 2018, 'Feature selection in machine learning: A new perspective', *Neurocomputing*, vol. 300, pp. 70-79, https://doi.org/10.1016/j.neucom.2017.11.077

Canata, TF, Wei, MCF, Maldaner, LF & Molin, JP 2021, 'Sugarcane yield mapping using high-resolution imagery data and machine learning technique', *Remote Sensing*, vol. 13, no. 2, article 232, https://doi.org/10.3390/rs13020232

Cao, D-S, Liang, Y-Z, Xu, Q-S, Zhang, L-X, Hu, Q-N & Li, H-D 2011, 'Feature importance sampling-based adaptive random forest as a useful tool to screen underlying lead compounds', *Journal of Chemometrics*, vol. 25, no. 4, pp. 201–207, https://doi.org/10.1002/cem.1375

Charoen-Ung, P & Mittrapiyanuruk, P 2019, 'Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning', *In:* Recent Advances in Information and Communication Technology 2018: Proceedings of the 14th International Conference on Computing and Information Technology (IC2IT 2018), Springer International Publishing, pp. 33–42, https://doi.org/10.1007/978-3-319-93692-5_4

Chen, P, Li, F & Wu, C 2021, 'Research on intrusion detection method based on Pearson correlation coefficient feature selection algorithm', *Journal of Physics: Conference Series*, IOP Publishing, vol. 1757, no. 1, article 012054, https://doi.org/10.1088/1742-6596/1757/1/012054

Cheng, T, Wang, J & Li, X 2007, 'The support vector machine for nonlinear spatio-temporal regression', *In:* Proceedings of Geocomputation

Chicco, D, Warrens, MJ & Jurman, G 2021, 'The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation', *PeerJ Computer Science*, vol. 7, article e623, https://doi.org/10.7717/peerj-cs.623

Criminisi, A, Shotton, J & Konukoglu, E 2012, 'Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning', *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2–3, pp. 81–227, http://dx.doi.org/10.1561/0600000035

Dos Santos Luciano, AC, Picoli, MCA, Duft, DG, Rocha, JV, Leal, MRLV & Le Maire, G 2021, 'Empirical model for forecasting sugarcane yield on a local scale in Brazil using Landsat imagery and random forest algorithm', *Computers and Electronics in Agriculture*, vol. 184, article 106063, https://doi.org/10.1016/j.compag.2021.106063

ELhadad, R, Tan, YF & Tan, WN 2022, 'Anomaly prediction in electricity consumption using a combination of machine learning techniques', *International Journal of Technology*, vol. 13, no. 6, pp. 1317–1325, https://doi.org/10.14716/ijtech.v13i6.5931

Erick, Y, Umezuruike, C, Jossy, N & Gusite, B 2023, 'Development of a machine learning regression model for accurate sugarcane crop yield prediction, Jinja–Uganda', *Journal of Applied Sciences, Information and Computing*, vol. 4, no. 1, pp. 25–33, https://doi.org/10.59568/JASIC-2023-4-1-03

Everingham, Y, Sexton, J, Skocaj, D & Inman-Bamber, G 2016, 'Accurate prediction of sugarcane yield using a random forest algorithm', *Agronomy for Sustainable Development*, vol. 36, pp. 1–9, https://dx.doi.org/10.1007/s13593-016-0364-z

Gaffar, AWM & Sitanggang, IS 2019, 'Spatial model for predicting sugarcane crop productivity using support vector regression', *In*: IOP Conference Series: Earth and Environmental Science, IOP Publishing, vol. 335, no. 1, article 012009, https://doi.org/10.1088/1755-1315/335/1/012009

Gregorutti, B, Michel, B & Saint-Pierre, P 2017, 'Correlation and variable importance in random forests', *Statistics and Computing*, vol. 27, pp. 659–678, https://doi.org/10.1007/s11222-016-9646-1

Hammer, RG, Sentelhas, PC & Mariano, JCQ 2020, 'Sugarcane yield prediction through data mining and crop simulation models', *Sugar Tech*, vol. 22, no. 2, pp. 216–225, https://doi.org/10.1007/s12355-019-00776-z

Hanka, MKF & Santosa, B 2021, 'Analisis kualitas bahan baku tebu melalui teknik pengklasteran dan klasifikasi kadar gula sebelum giling (studi kasus pabrik gula PT. XYZ)' ('Analysis of the quality of sugar cane raw materials through clustering techniques and classification of sugar content before milling (case study of PT. XYZ sugar factory)), *Jurnal Teknik ITS*, vol. 10, no. 2, pp. F100–F107

Hyndman, RJ 2014, 'Measuring forecast accuracy', *Business Forecasting: Practical Problems and Solutions*, pp. 177–183

Indrawanto, C, Purwono, Syakir, M, Siswanto, Soetopo, D, Munarso, SJ, Pitono, J & Rumini, W 2017, *Budidaya dan pascapanen tebu*, IAARD Press, Jakarta

Jaelani, T, Yamin, M & Mahandari, CP 2022, 'Machine learning untuk prediksi produksi gula nasional' (Machine learning for national sugar production prediction), *JMPM (Jurnal Material dan Proses Manufaktur)*, vol. 6, no. 1, pp. 31–36, https://doi.org/10.18196/jmpm.v6i1.14897

Joshua, SV, Priyadharson, ASM, Kannadasan, R, Khan, AA, Lawanont, W, Khan, FA, Rehman, AU & Ali, MJ 2022, 'Crop yield prediction using machine learning approaches on a wide spectrum', *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5663–5679, http://dx.doi.org/10.32604/cmc.2022.027178

Jović, A, Brkić, K & Bogunović, N 2015, 'A review of feature selection methods with applications', *In:* 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), IEEE, pp. 1200–1205, https://doi.org/10.1109/MIPRO.2015.7160458

Lárraga-Altamirano, HR, Hernández-López, DR, Piedad-Rubio, AM & Blanco-Martínez, JR 2024, 'Machine-learning model for estimating sugarcane production at crop level', *Journal of Technology and Innovation*, pp. 11–28

Mahesh, B 2020, 'Machine learning algorithms – a review', *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386

Maldaner, LF, De Paula Corrêdo, L, Canata, TF & Molin, JP 2021, 'Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches', *Computers and Electronics in Agriculture*, vol. 181, article 105945, https://doi.org/10.1016/j.compag.2020.105945

Medar, RA, Rajpurohit, VS & Ambekar, AM 2019, 'Sugarcane crop yield forecasting model using supervised machine learning', *International Journal of Intelligent Systems and Applications*, vol. 11, no. 8, p. 11-20, https://doi.org/10.5815/ijisa.2019.08.02

Mohamad, M, Selamat, A, Krejcar, O, Crespo, RG, Herrera-Viedma, E & Fujita, H 2021, 'Enhancing big data feature selection using a hybrid correlation-based feature selection', *Electronics*, vol. 10, no. 23, article 2984, https://doi.org/10.3390/electronics10232984

Nikhil, UV, Pandiyan, AM, Raja, SP & Stamenkovic, Z 2024, 'Machine learning-based crop yield prediction in South India: Performance analysis of various models', *Computers*, vol. 13, no. 6, article 137, https://doi.org/10.3390/computers13060137

Noorsaman, A, Amrializzia, D, Zulfikri, H, Revitasari, R & Isambert, A 2023, 'Machine learning algorithms for failure prediction model and operational reliability of onshore gas transmission pipelines', *International Journal of Technology*, vol. 14, no. 3, pp. 680-689, https://doi.org/10.14716/ijtech.v14i3.6287

Paidipati, KK, Banik, A, Shah, B & Sangwa, NR 2022, 'Forecasting of sugarcane productivity estimation in India – a comparative study with advanced non-parametric regression models', *Journal of Algebraic Statistics*, vol. 13, no. 2, pp. 760–778

Palanivel, K & Surianarayanan, C 2019, 'An approach for prediction of crop yield using machine learning and big data techniques', *International Journal of Computer Engineering and Technology*, vol. 10, no. 3, pp. 110–118

Pandey, N, Patnaik, PK & Gupta, S 2020, 'Data pre-processing for machine learning models using python libraries', *International Journal of Engineering and Advanced Technology*, vol. 9, no. 4, pp. 1995–1999, http://dx.doi.org/10.35940/ijeat.D9057.049420

Panigrahi, B, Kathala, KCR & Sujatha, M 2023, 'A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models', *Procedia Computer Science*, vol. 218, pp. 2684–2693, https://doi.org/10.1016/j.procs.2023.01.241

Rafiqi, Saidin, OK, Lubis, MY & Ikhsan, E 2021, 'Optimization of sustainable sugar industry towards food security', *In:* IOP Conference Series: Earth and Environmental Science, IOP Publishing, vol. 782, no. 2, article 032039, https://doi.org/10.1088/1755-1315/782/3/032039

Respati, E 2022, *Outlook komoditas perkebunan tebu 2022 (Sugarcane plantation commodity outlook 2022)*, Pusat Data dan Sistem Informasi Pertanian, Kementerian Pertanian Republik Indonesia, Indonesia

Sari, M, Berawi, MA, Larasati, SP, Susilowati, SI, Susantono, B & Woodhead, R 2023, 'Developing machine learning model to predict HVAC system of healthy building: A case study in Indonesia', *International Journal of Technology*, vol. 14, no. 7, pp. 1438–1448, https://doi.org/10.14716/ijtech.v14i7.6682

Shah, A, Dubey, A, Hemnani, V, Gala, D & Kalbande, DR 2018, 'Smart farming system: Crop yield prediction using regression techniques', *In:* Proceedings of International Conference on Wireless Communication: ICWiCom 2017, Springer Singapore, pp. 49–56, https://doi.org/10.1007/978-981-10-8339-6_6

Shetty, SA, Padmashree, T, Sagar, BM & Cauvery, NK 2021, 'Performance analysis on machine learning algorithms with deep learning model for crop yield prediction', *In:* Data Intelligence and Cognitive Informatics: Proceedings of ICDICI 2020, Springer Singapore, pp. 739–750, https://doi.org/10.1007/978-981-15-8530-2_58

Singla, SK, Garg, RD & Dubey, OP 2020, 'Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information', *Revue d'Intelligence Artificielle*, vol. 34, no. 6, pp. 731-743, https://doi.org/10.18280/ria.340607

Smola, AJ & Schölkopf, B 2004, 'A tutorial on support vector regression', *Statistics and Computing*, vol. 14, pp. 199-222, https://doi.org/10.1023/B:STCO.0000035301.49549.88

Som-ard, J, Suwanlee, SR, Pinasu, D, Keawsomsee, S, Kasa, K, Seesanhao, N, Ninsawat, S, Borgogno-Mondino, E & Sarvia, F 2024, 'Evaluating sugarcane yield estimation in Thailand using multi-temporal Sentinel-2 and Landsat data together with machine-learning algorithms', *Land*, vol. 13, no. 9, article 1481, https://doi.org/10.3390/land13091481

Sulaiman, AA, Arsyad, M, Amiruddin, A, Teshome, TT & Nishanta, B 2023, 'New trends of sugarcane cultivation systems toward sugar production on the free market: A review', *AGRIVITA Journal of Agricultural Science*, vol. 45, no. 2, pp. 395–406, http://dx.doi.org/10.17503/agrivita.v45i2.4066

Tatachar, AV 2021, 'Comparative assessment of regression models based on model evaluation metrics', *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9, pp. 853–860

Van Klompenburg, T, Kassahun, A & Catal, C 2020, 'Crop yield prediction using machine learning: A systematic literature review', *Computers and Electronics in Agriculture*, vol. 177, article 105709, https://doi.org/10.1016/j.compag.2020.105709

Venkatesh, B & Anuradha, J 2019, 'A review of feature selection and its methods', *Cybernetics and Information Technologies*, vol. 19, no. 1, pp. 3–26