International Journal of Technology v(i) pp-pp (YYYY)
Received Month Year / Revised Month Year / Accepted Month Year

International Journal of Technology

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

Design of a Prediction Model for Sugarcane Yield and Productivity

Abstract. Indonesia's sugar industry has yet to become self-sufficient in sugar. This is due to the unpredictable and fluctuating relationship between yield (sugar content (%)) and sugarcane productivity (Ton/Ha) in all Indonesian sugar mills, whether state-owned or private. As a result, Indonesia's domestic sugar consumption is still balanced by imports of sugar. The objective of this study is 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 support vector regression (SVR) and random forest (RF). The study found that essential features for predicting sugarcane yield are clear juice, Pol, purity, Brix, and maturity factor, whereas important features for predicting sugarcane productivity are number of stem, stem height, stem weight, rainfall, and juring factor. 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 percentage absolut 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 up until now. 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 (Indonesia BPS, 2023). The sugar agroindustry's problem is directly related to the upstream side, notably sugarcane yield and productivity. (Saidin et al., 2023) asserts that low productivity and yield (produce) are internal issues impeding the development of alternative policies for sugar self-sufficiency. (Sulaiman et al., 2023) adds that the lack of sugar output is due to diminishing land area,

*Corresponding author's email: , Tel.: +00-00-000000; Fax: +00-00-000000
doi: 10.14716/ijtech.v0i0.0000

low sugar cane productivity, and low sugar yield. (Indonesia BPS, 2023) demonstrates that, while sugar output declined over the same period, 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, productivity and sugar cane production are fluctuating in all Indonesian sugar mills in 2023. An unstable pattern of relationship between sugarcane yield and productivity is revealed. This phenomenon creates uncertainty related to the correlation between increasing yield along with increasing sugarcane productivity. All this time, 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). In order to predict the number of milling days needed, sugarcane production is estimated twice a year, in December and March. Because of the high plant variability and the significant influence of environmental factors, sugarcane maturity analysis and production estimation are susceptible to error. 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. such that an accurate prediction model that can manage complex data, adjust to change, and find complex patterns. so that it can increase the efficiency of sugar production and make more accurate decisions while managing sugarcane plantations. Furthermore, an 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, Vector Auto Regression (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 estimate of growth 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 proper 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) using the relief methodology and the Support Vector Machine (SVM) method, where soil pH, humidity, and sugar cane age are the primary elements influencing sugar content, the SVM method can be utilized to estimate sugar content and harvest time for sugar cane mills, (Hammer et al., 2020) developed a model to predict sugarcane productivity using the Random Forest (RF), Gradient Boosting Machine (GBM), and Support Vector Machine (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 sugar cane, 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 Support Vector Regression (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), support vector machine (SVM), and random forest. Support vector machine regression is the best result in predicting the crop yield. Random forest (RF) regression is a very successful technique for predicting sugarcane crop yields compared to multiple linear regression and decision tree regression (Erick et al., 2023) and shown effectiveness in tasks such as regression (Lárraga-Altamirano wt al., 2024). Maldaner et al. (2021) used artificial neural network (ANN), RF, and MLR in his research to estimate sugarcane yield, RF proved to be the most effective model

There is a knowledge gap in understanding about 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 measurements are through assessments in March and December.

Therefore, the objective of this study is to identify important features in predicting sugarcane yield and productivity using vegetative growth indicators in sugarcane cultivation using data science and machine learning techniques. The data for this study was collected at sugar plants in Malang and Madiun, East Java, Indonesia. This secondary data was gathered from observations of sugarcane vegetative growth and sugarcane analytical findings in 2023. The techniques used in this study are based on support vector regression (SVR) and random forest regression (RF). These approaches were chosen based on earlier research, which found that SVR and RF were the top performing regression-based machine learning algorithms. Also, the data utilized in this study is continuous and numeric with a numerical output. A continuous dependent variable may be predicted from a series of independent inputs using regression analysis (Panigrahi et al., 2022). Regression method selection should evaluate various variables that are 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 conditions of uncertainty. Crop yield models, according to (Bocca and Rodrigues, 2016), may help decision-makers in any agro-industrial supply chain, even when they relate to issues unrelated to crop production.

2. Related Works

The process of automatically identifying designs in data without making any assumptions about the data's structure is known as machine learning (Noorsaman et al., 2023). According to (Van Klompenburg et al., 2020), machine learning is a useful technique for making decisions in forecasts crop yields, what crops should be grown, and what has to 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. Predicting harvest yields may develop using a variety of machine learning techniques, including regression, classification, grouping, and prediction. Several algorithms, including artificial neural networks, decision trees, support vector machines, naïve bayes, and linear and logistic regression, can also be used (Palanivel and Surianarayanan, 2019). (Singla et al., 2020) using remote sensing data to forecast sugarcane yield using ensemble machine learning. (Mahesh, 2020) reports that a number of frequently used machine learning techniques. The kind of model that is most suitable, the number of

variables, and the kind of issue to be addressed all influence the type of method that is employed.

Machine learning has been applied in many sectors, including it is possible to developing prediction models for many important features in production process of the industry. In machine learning, prediction model is divided into 3 parts, numerical prediction, classification and time series prediction. Since the goal of the research is to predict the yield and productivity which are stated in numerical manner, therefore 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 neighbours, which can be seen in Table 1. From the algorithms that have been used, the random forest and support vector regression have the best performance in determining productivity and yield. The Regression-based machine learning builds models that forecast numerical (continuous) values based on input data by applying statistical approaches. Regression approaches are commonly employed by algorithms such as random forest (RF) and support vector regression (SVR). Regression and classification may be accomplished with the help of 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 model that is more stable and resistant to overfitting (Smola and Schölkopf, 2004).

Machine learning application for forecasting must develop through several important stage, including data preprocessing, modelling and evaluation. Firstly, feature selection techniques may be applied to data pre-processing in order to accomplish effective data reduction, according to (Jović et al., 2015). This is helpful in locating 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 help the learning model or data be understood better (Cai et al., 2018). Many techniques has been applied for feature selection, including statistical test, correlations, clustering or feature important. 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 (Mohamad et al., 2021; Chen et al., 2021). Table 1 shows some previous studies on sugarcane productivity and yield.

Table 1 Previous studies on prediction of productivity and sugarcane yield

No.	Author(s)	Year	Method	Features
1	Erick et al.	2023	Multiple linier regression, decision tree regression, random forest regression	Sugarcane yield, area, age of sugarcane, sugarcane crop cycle, temperature, rainfall, Soil, pH
2	Jaelani et al.	2022	Long-short term memory	Year, sugarcane production, sugar production, sugar consumption, population
3	Paidipati et al.	2022	Multivariate adaptive regression splines, support vector regression, partial least square regression, elastic-net regression, multiple linear regression	Sugarcane yield, production, area

4	Maldaner et al.	2021	Multiple linier regression, random forest, artificial neural network	CAN data from sugarcane harvester
5	dos Santos et al.	2021	Random forest	Remote sensing data, meteorological data (solar radiation, wind speed, relative humidity, precipitation, min. temperature, max. temperature), agronomic data (Soil, variety, yield, harvest date, number of harvests, production environment)
6	Asrol et al.	2021	Support vector machine	Soil pH, temperature, rainfall, humidity, sugarcane ages, area height, early sugar content, Pol, Brix,
7	Hammer et al.	2020	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, solar radiation
8	Singla et al.	2020	Support vector regression, random forest, K-nearest neighbors, classification and regression trees	Satellite data
9	Charoen-Ung & Mittrapiyanuruk	2019	Random forest	Cane class, type, water type, soil type, area, fertilizer, rainfall, distance, contract area,
10	Gaffar & Sitanggang	2019	Support vector regression	Year, area, province, min. temperature, max. temperature, average humidity, rainfall, duration of sunshine, wide area, production, productivity

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 data related to climate. (Asrol et al., 2020) defined the major criteria. However, there was no evaluation of the machine learning performance. The primary factors influencing productivity and yield will be identified by this research 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 machine learning model will be assessed for further implementation in decision making at the mill.

3. Methods

3.1. Research Flow

The research flow begins with data collection (see Figure 1). The data was collected from sugar agroindustry in Malang and Madiun, East Java, Indonesia. These are secondary data generated from observations of sugarcane vegetative growth and the results of sugarcane analysis in 2023, which will be features in predicting sugarcane productivity and yield. Secondly, after data collection and acquisition, the raw data required to be pre-processed by cleaning, simplifying, and transforming. It needs to transform data into an accurate, relevant, and consistent dataset for further analysis and modelling. This stage includes verifying the raw data to ensure that the data is useful and efficacious for use.

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

The prediction model is design with the regression-based machine learning model. Previous studies have used a variety of productivity and yield prediction for sugarcane methods, 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) (Maldaner et al., 2021; Erick et al., 2023; Lárraga-Altamirano et al., 2024; dos Santos et al., 2021), multiple linear regression (Paidipati et al., 2022; Maldaner et al., 2021), partial least squares regression (PLSR) (Paidipati et al., 2022), artificial neural network (ANN) (Maldaner et al., 2021). Based on these investigations, SVR and RF are the regression-based machine learning approaches with the best performance. Therefore, both models will be used in this investigation.

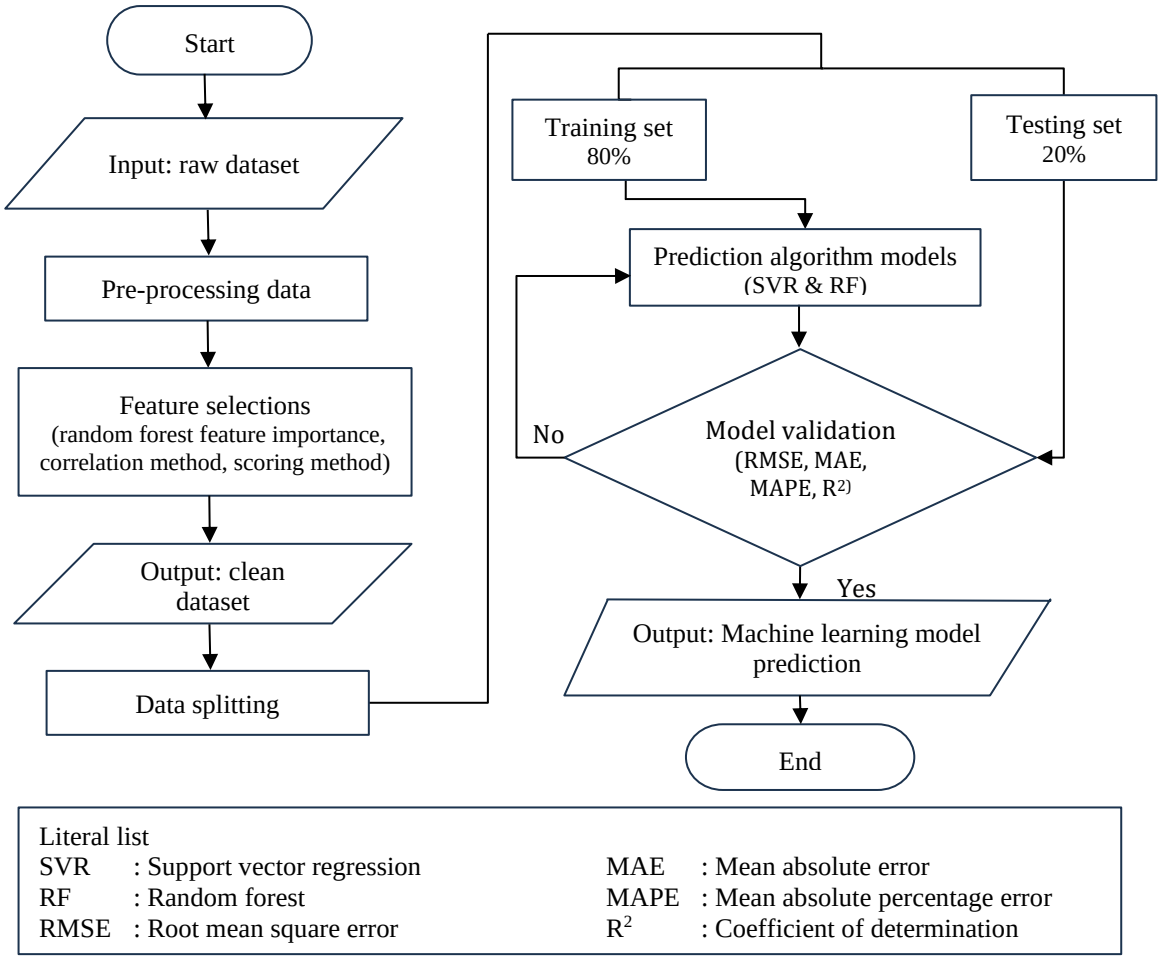


Figure 1 Research flow

In the next stage is model evaluation. The SVR and RF output values were analysed using the root mean square error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) statistical models to determine the accuracy of the model used to forecast the output. A model's ability to forecast sugarcane productivity and yield may be assessed by looking at its RMSE, MAE (Shetty et al., 2020), and (R2) (Canata et al., 2021; Nikhil et al., 2024). Further, the detail of each stage will be delivered in the following subsections.

3.2. Data Collection and Acquisition

This study employs secondary data collected from manual observations of sugarcane vegetation growth and yield analysis conducted in two sugar factories with different regions during 2023. These two sugar factories have different production capacities, where the sugar factory located in Malang has a production capacity of 12,000 tons of cane per day with a sugarcane area of 21,838 ha and 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. Table 2 delineates the ten features employed for yield prediction with 2,225 row of data, whereas Table 3 encompasses thirteen features utilized for forecasting sugarcane production with 2,656 row data as features for prediction model development. The features included in forecasting sugarcane production and yield consist of two data types, numeric and categorical.

A total more than 4,500 instances has been collected consist of related features for yield and productivity. These data have been represented the sugar agroindustry condition for the current year production. The agroindustry does a single cycle of production for each year and the number of instances has been represents more than 50% of the total dataset.

Table 2 Features for determining sugarcane yield

No.	Feature	Unit	Type	Description
1	Area	Ha	Numeric	A scalar quantity that shows the area of the plant.
2	Planting Period	A / B	Categorical	Sugarcane planting age where A shows the planting period in the 1st & 2nd week while B shows the 3rd & 4th week of the current month.
3	Varieties	-	Categorical	Groups of sugarcane plants that have characteristics and properties that are different from each other.
4	Total soluble solid (°Brix)	%	Numeric	The dissolved solids in a liquid, used to estimate the sugar content of an aqueous solution.
5	Polarization value (Pol)	%	Numeric	Sucrose content
6	Purity		Numeric	The figure indicates the sap's purity level in the sugarcane processing sector.
7	Clear Juice	%	Numeric	The first or initial milling of sugarcane juice.
8	Maturity Factor	%	Numeric	Indicator of sugarcane maturity level that determines optimal harvest time.
9	Coefficient of Improvement	%	Numeric	An indicator that shows the potential increase in sugar content in sugarcane stem after a certain periode of time.
10	Coefficient of Resistance	%	Numeric	An indicator shows the ability of sugar cane to maintain its sugar content after a certain period.

Table 3 Features for determining productivity

No.	Feature	Unit	Type	Description
1	Area	Ha	Numeric	A scalar quantity that shows the area of the plant.
2	Planting Period	A / B	Categorical	Sugarcane planting age where A shows the planting period in the 1st & 2nd week while B shows the 3rd & 4th week of the current month.
3	Varieties	-	Categorical	Groups of sugarcane plants that have characteristics and properties that are different from each other.
4	NPK	kgs/Ha	Numeric	Stands for Nitrogen (N), Phosphorus (P), and Potassium (K), the three primary macronutrients essential for plant growth.
5	ZA	kgs/Ha	Numeric	Provides nitrogen and sulfur, which are essential for vegetative growth.
6	KCL	kgs/Ha	Numeric	Provides potassium, which is crucial for overall plant health, flowering, and fruit production.
7	Center to Center	m	Numeric	The distance between the midpoints of two adjacent sugar cane plants.
8	Juring Factor	m	Numeric	The percentage of land area planted with crops compared to the total land area,
9	Number of Stems	pc	Numeric	Count the number of sugarcane stem within one meter.
10	Stem Height	m	Numeric	Stem height is measured from the soil surface to the top ring or node before the shoot.
11	Stem Diameter	Cm	Numeric	Calculate the width of a sugarcane stem.
12	Stem Weight	Kgs	Numeric	Measuring the weight of sugarcane stem per meter.
13	Rainfall	mm	Numeric	the volume of rain that falls and gathers in a place during a predetermined period

3.3. Pre-processing Data

At this stage, a data cleaning procedure is performed, during which the data is examined to find any mistakes, inconsistencies, or anomalies that may exist in the dataset. The correctness and caliber of the dataset may be impacted by these problems, which might include missing values, duplicate entries, outliers, improper data formats, and other data quality concerns. This stage is crucial for figuring out how good the data is and how much cleanup is needed. Data pre-processing, according to Pandey et al. (2020), is one of the most crucial stages in the creation of any machine learning model because it directly affects the model's quality and efficiency. If we neglect this step and create a model using data sets that have missing values, the resulting model will be inconsistent and less effective. According to Sari et al. (2023), accurate predictions can be hampered by poor data quality when it results in underfitting owing to dispersed data quality or overfitting, which restricts predictions to a small range of data.

The data preparation process, according to Alexandropoulos et al. (2019), 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 preparing the data, including preprocessing as seen 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 cleaning the data 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 carried out. The purpose of feature selection in preprocessing is to eliminate variables or features that are most relevant to the target in this study.

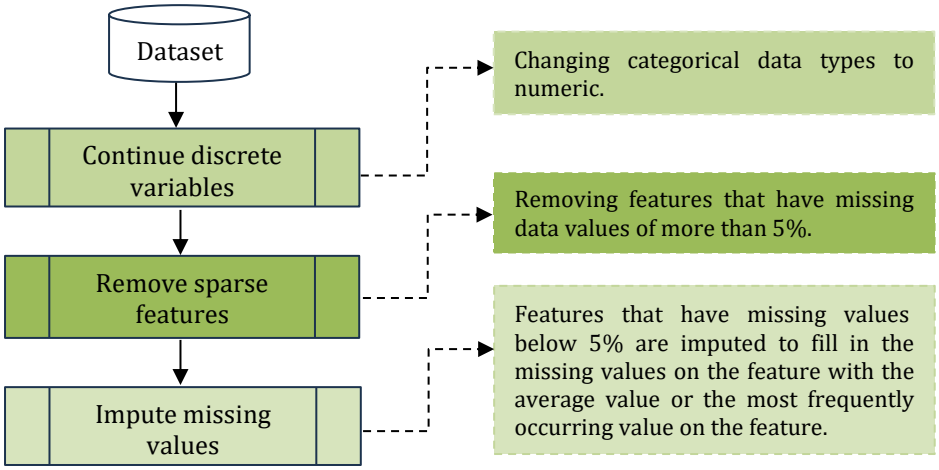


Figure 2 Preprocessing steps

3.4. Feature Selection

There are two approaches to dimensionality reduction. Feature Selection and Feature Extraction. In addition to lowering the data burden, the feature selection approach helps prevent overfitting the model (Venkatesh and Anuradha, 2019). Feature selection by filtering is the strategy used at this point to reduce dimensionality. Choosing a selection of characteristics that are most relevant to the target variable is done by applying the correlation and scoring approach known as the filter method. Also, 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 multi method for feature selections. This study ensures the selected features have importance position to predict yield and productivity. Therefore, three methods for feature selection and analysis are employed, including correlation methods, scoring and feature importance by random forest. These multimethod is applied to confirm the importance of features in predictions. In this following part, the detail technique for feature selections applied in this study are described.

3.4.1. Random Forest Feature Importance

Using feature importance metrics may determine the relative relevance of each feature and the degree to which its removal reduces accuracy or its inclusion increases accuracy (AlSagri and Ykhlef, 2020). According to (Gregorutti et al., 2017), due to various important measurements, the random forests method enables us to assess a predictor's relevance in simultaneously. Three metrics are calculated by the original random forest algorithm: the z-score, the Gini importance, and the permutation importance. The permutation significance measure has demonstrated strong performance for top variable selection methods, among other criteria. Equation 1 is used to calculate 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 jth feature permutation, and $y_{j\_premuted}$ is model's performance after the jth permutation. The feature importance values obtained are then normalized to be in the range 0 to 1.

3.4.2. Corellation Method

Pearson correlation coefficient is used for normally distributed numerical data by following the following mathematical 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 it implies no association at all if it is exactly zero.

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). The mathematical Equation 3:

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

Where 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. Predictions Algorithm Models

3.5.1. Support Vector Regression (SVR)

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

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

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

w is the weight vector of inputs, b is the bias, $\emptyset(x)$ is a kernel function. When a non-linear input is converted into a linear input by use of a non-linear function. The objective is to identify, for each training set of data, the function f(x) with the largest ε-deviations from the achieved objectives 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 difference between the anticipated value and the margin equals the loss. It is possible to describe the restrictions and the goal function as Equation 6, 7, and 8:

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

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

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

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

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. Utilizing the Lagrange function, the optimal value of each parameter is determined after choosing a band width (ε), kernel function (φ), and altering coefficient (C).

3.5.2. Random Forest

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

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

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

3.6. Model Evaluation

The performance of the prediction model is evaluated by comparing the predicted values to the actual observed values using the root mean square error (RMSE) and mean absolute error (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 and RMSE is used to determine the accuracy of the model and MAE to determine the magnitude of the error in the results (Hammer et al., 2020).

RMSE is calculate by formula in Equation 10:

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

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

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

To assess forecast accuracy across many series with varying scales, use the MAPE (Hyndman, 2014). Measuring the average absolute percentage error can be seen in Equation 12.

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

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

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

The projected ith value is represented by $\check{y}_i$ in the following formulas, and the actual ith 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 analysis data and sugarcane yield analysis data carried out in 2023 at two sugar factory located in East Java, Indonesia. The dataset shows variability because observations of sugarcane fields are in several areas not only in Malang and Madiun but also around them. Estimated sugarcane production data is carried out by sugar factories twice a year in March and December. In this study, 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, where the sugarcane production estimation observation data in December shows that the sugarcane plants have not grown optimally, so the observation data has a high error rate that can impact model performance.

The raw dataset after collection need to be handled to ensure the data validity with preprocessing stage. Among the pre-processing steps of the dataset perfomed are imputing missing values by deleting rows that contain missing values or by using the most common value (for discrete attributes) or the average value (for continuous attributes). Afterwards, 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.

Table 4 Descriptive statistics of sugarcane yield before pre-processing

No.	Features	Sat.	N	Min.	Max.	Mean	Std. Deviation
1	Area	Ha	2,225	0.1	20	3.69	4.09
2	Planting Period	-	2,225	-	-	-	-
3	Varieties	-	2,225	-	-	-	-
4	°Brix	%	2,225	8.25	20.05	14.77	1.64
5	Pol	%	2,225	8.85	22.41	13.81	2.35
6	Purity	%	2,225	57.64	93.86	80.88	4.63
7	Clear Juice	%	2,225	6.26	19.11	11.64	1.81
8	Maturity Factor	%	2,225	19.43	91.19	46.24	9.06
9	Coefficient of Improvement	%	1,335	77.72	162.09	114.29	10.86
10	Coefficient of Resistance	%	1,335	0.00	127.28	106.84	6.26

In the preprocessing stage, in terms of productivity, the preprocessing results showed that 13 features become 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 a total of 2225 rows data with 2 missing due to having a missing value of 40%. Features For a comparison, Tables 4 and 5 present descriptive statistics of the dataset before pre-processing and Tables 6 and 7 after pre-processing.

Table 5 Descriptive statistics of productivity before pre-processing

No.	Features	Sat.	N	Min.	Max.	Mean	Std. Deviation
1	Area	Ha	2,656	0.06	40.00	3.21	3.72
2	Planting Period	A / B	2,656	-	-		-
3	Varieties	-	2,656	-	-		-
4	NPK	kgs/Ha	2,656	200	700	463.29	150.87
5	ZA	kgs/Ha	2,656	600	1,000	780.20	174.45
6	KCl	kgs/Ha	1,099	100	100	100	-
7	Center to Center	m	2,656	1.10	1.15	1.11	0.02
8	Juring Factor	m	2,656	8,695.65	9,090.91	9,008.46	160.62
9	Number of Stem	m	2,656	4.00	9.85	6.61	0.89
10	Stem Height	m	2,656	0.89	3.42	2.41	0.34
11	Stem Diameter	cm	2,652	1.06	3.65	2.41	0.20
12	Stem Weight	Kgs	2,656	0.30	0.78	0.45	0.07
13	Rainfall	mm	2,656	1,682	3,050	2,172,02	374.99

Table 6 Descriptive statistics of sugarcane yield after pre-processing

No.	Features	Sat.	N	Min.	Max.	Mean	Std. Deviation
1	Area	Ha	2,225	0.1	20	3.69	4.09
2	Planting Period	-	2,225	0.00	7	4.07	1.73
3	Varieties	-	2,225	0.00	12	2.46	3.78
4	°Brix	%	2,225	8.25	20.05	14.77	1.64
5	Pol	%	2,225	8.85	22.41	13.81	2.35
6	Purity	%	2,225	57.64	93.86	80.88	4.63
7	Clear Juice	%	2,225	6.26	19.11	11.64	1.81
8	Maturity Factor	%	2,225	19.43	91.19	46.24	9.06

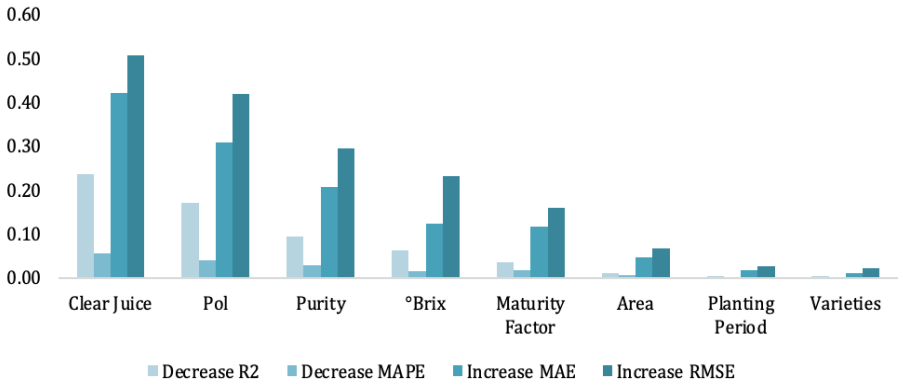
Table 7 Descriptive statistics of productivity after pre-processing

No.	Features	Sat.	N	Min.	Max.	Mean	Std. Deviation
1	Area	Ha	2,656	0.06	40.00	3.21	3.72
2	Planting Period	A / B	2,656	0.00	15	7.29	3.38
3	Varieties	-	2,656	0.00	27	3.18	7.10
4	NPK	kgs/Ha	2,656	200	700	463.29	150.87
5	ZA	kgs/Ha	2,656	600	1,000	780.20	174.45
6	Center to Center	m	2,656	1.10	1.15	1.11	0.02
7	Juring Factor	m	2,656	8,695.65	9,090.91	9,008.46	160.62
8	Number of Stem	m	2,656	4.00	9,85	6.61	0.89
9	Stem Height	m	2,656	0.89	3.42	2.41	0.34
10	Stem Diameter	cm	2,656	1.06	3.65	2.41	0.20
11	Stem Weight	Kgs	2,656	0.30	0.78	0.45	0.07
12	Rainfall	mm	2,656	1,682	3,050	2,172.02	374.99

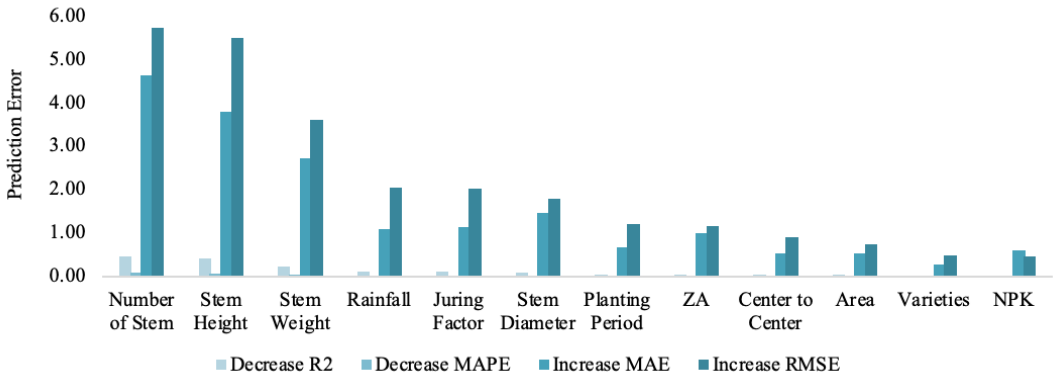
4.2. Feature Engineering and Selection

To the contribution of each feature to the prediction of sugarcane yield and productivity was determined through feature selection using the feature importance random forest technique. This was done by measuring the increase in model prediction error after

randomizing feature values, which destroys the relationship between features and targets. Variable importance is determined by fitting a model that includes all predictors and updating the model after permuting each predictor variable. Then, the link between each predictor and the result is examined (Maldaner et al., 2021). The results of the feature importance random forest processing in this study can be seen in Figure 3.



(a)



(b)

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

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

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 8 and 9. 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 8 Selected features of 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

Table 9 Selected features of productivity

Features	Pearson Correlation	Univariate Regression
Number of Stem	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

4.3. Modelling

The prediction model is designed using 80% of the training data and 20% of the testing data. Support Vector Regression (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 support vector machines 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. The hyperplane is the optimum line in this case. The data is 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) and C is also an essential parameter that governs the trade-off between increasing margin and reducing training error (Cheng et al., 2007). This study uses the SVR parameters shown in Table 10 with values using default settings. Hanka & Santosa (2021) in their research 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. Som-ard et al.

(2024) in evaluating sugarcane yields in Thailand used SVR with the model's optimal hyperparameters of cost (C) 1.00 and gamma 0.1 using the RBF kernel.

Table 10 shows the default settings for random forest regression. 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 random forest model's parameter settings 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, forecasts of sugar cane yields and productivity were obtained by validating the model using test data.

Table 10 The parameters and descriptions/values used

Technique	Parameters	Descriptions/Values
SVR	SVM kernel function	Polynomial
	Cost (C)	1.00
	Epsilon (ε)	0.10
	Gamma (g)	Auto
	Degree of kernel (d)	3.0
RF	Number of tree (N_{tree})	10
	Do not split subsets size	5

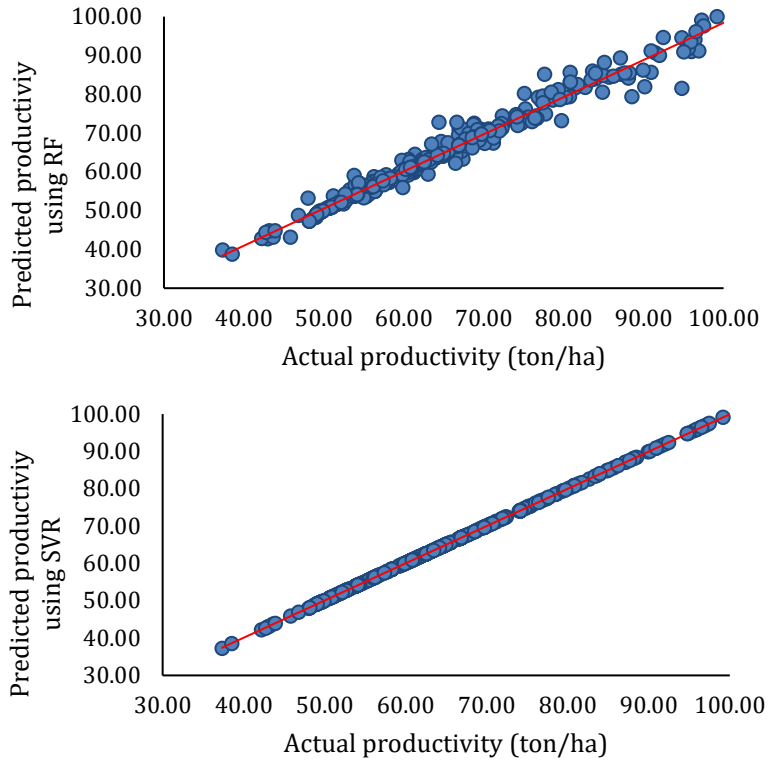
4.4. Models Performance

Forecasts of sugar cane yields and productivity were obtained by validating the model using test data. Following this, 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 the predictive model used to estimate sugarcane yield, utilizing 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 11, 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%.

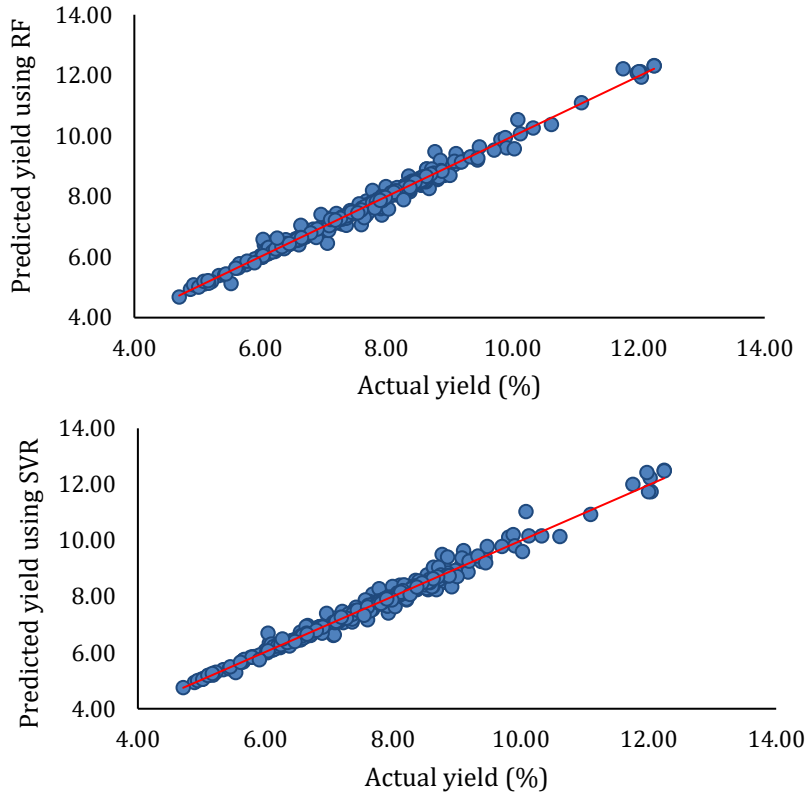
The determination of the yields was obtained from the RF model with a prediction error rate of 0.129%, an average absolute error difference of 0.074%, and an accuracy value in forecasting yields with an average percentage absolute error of 0.010% and data variability of 98.8%. The graph in Figures 4 compares the actual and predicted values using SVR and RF for productivity and yield, respectively. In this test, the proposed algorithm has slightly different value from the predicted values and actual values. It indicated that the model is possible to predict the productivity and yield in the near future according to model input and parameters.

Table 11 Machine learning models performance evaluation statistics

		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 prediction values for (a) productivity and (b) sugarcane yield

This study indicates that the optimal model for predicting sugarcane production is Support Vector Regression (SVR), whereas the most effective model for forecasting sugarcane yield is Random Forest techniques. Furthermore, the findings of this study facilitate efficiency and expediency in estimating productivity and sugarcane production within the sugar companies, transforming from manual methods to digitalization, thereby aiding in decision-making regarding harvest timing, workforce readiness, and factory operations. Figure 5 illustrates the digital shift in forecasting sugarcane productivity and yields. The existing manual procedure takes a long time to get 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 utilizing machine learning, the results can be known more quickly and accurately. The models obtained 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 known immediately. A digital transformation approach with the implementation of a machine learning model to predict sugarcane yield and productivity may improve the business process efficiency and lead time.

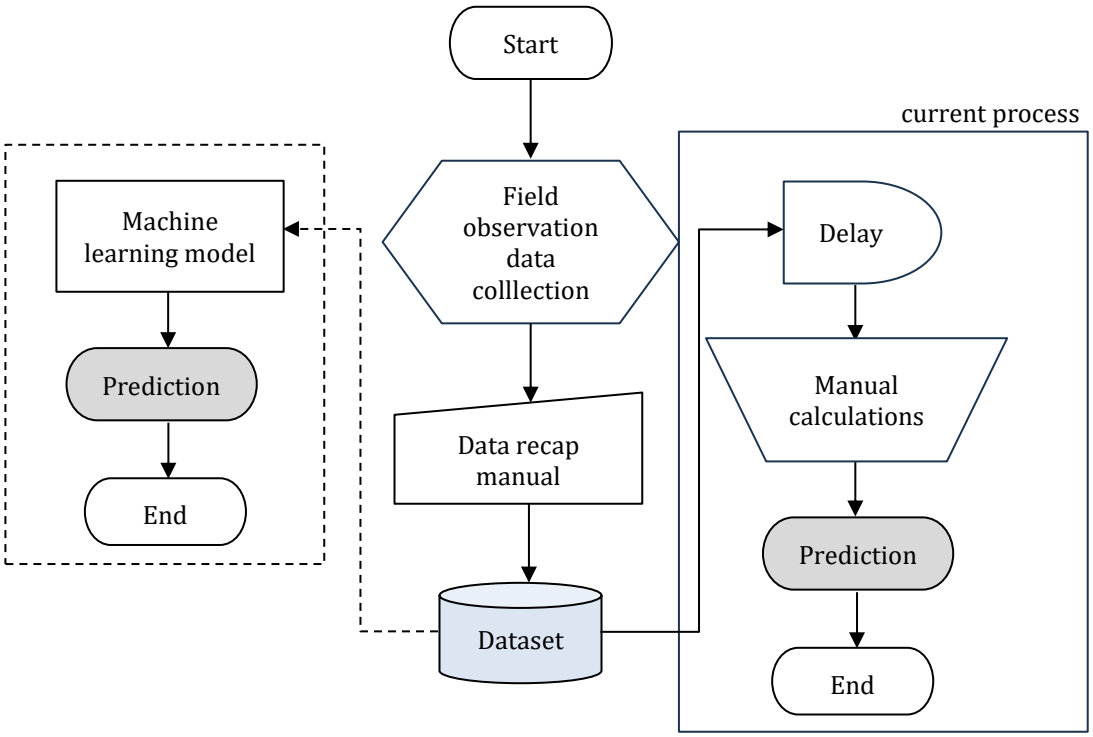


Figure 5 The flow of the current prediction method and with machine learning

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

4.5. Managerial Implications and Contributions

This study demonstrates that regression-based techniques may be used to anticipate sugarcane productivity and yield. Support Vector Regression (SVR) is the best model for predicting sugarcane production, while the Random Forest (RF) methodology is the most successful model for predicting sugarcane yield. The study's findings theoretically indicate

that sugarcane yield and productivity potential may be more accurately predicted by improving key parameters that influence their target in order 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. By knowing the potential yield and productivity of sugarcane, it is possible to determine the best time to harvest the crop and when starting the milling period. Moreover, machine learning 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 determine the balance of sugar commodities. Further from that, the academic contributions of this research demonstrate that the regression model may be utilized to estimate sugarcane yield and productivity, which differs from earlier models. The performance of the model generated in this research 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 machine learning 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. Conclusion

In conclusion, this study successfully designed a prediction model for sugarcane yield and productivity based on machine learning using yield and vegetation growth analysis data. The important feature random forest method was used to determine features that can improve model performance. There are 5 important features in developing an accurate yield prediction model, including clear juice, Pol, purity, Brix, and maturity factor, while important features in predicting sugarcane productivity are the number of stems, stem height, stem weight, rainfall, and juring factor, which 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 to predict sugarcane productivity. While for the yield prediction model, this model is derived from the RF model.

This study presents a feasible approach to determining the yield and productivity of sugarcane 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 applying the model to support in sugar mill decision making process especially in harvesting time schedule and simulate sugar production efficiency.

Acknowledgments

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

References

Alexandropoulos, S.A.N., Kotsiantis, S.B. and Vrahatis, M.N., 2019. Data preprocessing in predictive data mining. *The Knowledge Engineering Review*, *34*, p.e1

AlSagri, H., Ykhlef, M., 2020. Quantifying feature importance for detecting depression using random forest. *International Journal of Advanced Computer Science and Applications*, *11*(5)

Asrol, M., Marimin, M., Yani, M., 2020. Business intelligence model construction to improve sugarcane yield for the sustainable sugar industry. *J. Adv. Res. Dyn. Control Syst.*, Volume 12(Special Issue 6), pp.109-118

Bocca, F.F., Rodrigues, L.H.A., 2016. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. *Computers and electronics in agriculture*, Volume 128, pp.67-76

Cai, J., Luo, J., Wang, S., Yang, S., 2018. Feature selection in machine learning: A new perspective. *Neurocomputing*, Volume 300, pp.70-79

Canata, T.F., Wei, M.C.F., Maldaner, L.F. and Molin, J.P., 2021. Sugarcane yield mapping using high-resolution imagery data and machine learning technique. *Remote Sensing*, Volume 13(2), p.232

Cao, D.S., Liang, Y.Z., Xu, Q.S., Zhang, L.X., Hu, Q.N. and Li, H.D., 2011. Feature importance sampling-based adaptive random forest as a useful tool to screen underlying lead compounds. *Journal of Chemometrics*, *25*(4), pp.201-207

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

Chen, P., Li, F., Wu, C., 2021. Research on intrusion detection method based on Pearson correlation coefficient feature selection algorithm. In *Journal of Physics: Conference Series*. IOP Publishing, Volume 1757(1), p. 012054

Cheng, T., Wang, J. and Li, X., 2007. The support vector machine for nonlinear spatio-temporal regression. In *Proc Geocomputation*

Chicco, D., Warrens, M.J., 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*, Volume 7, p.e623

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*, Volume 7(2–3), pp.81-227

dos Santos Luciano, A.C., Picoli, M.C.A., Duft, D.G., Rocha, J.V., Leal, M.R.L.V. and 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*, *184*, p.106063

ELhadad, R., Tan, Y.F. and Tan, W.N., 2022. Anomaly prediction in electricity consumption using a combination of machine learning techniques. *International Journal of Technology*, *13*(6), pp.1317-1325.

Erick, Y., Umezuruike, C., Jossy, N. and 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*, Volume 4(1), pp.25-33

Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agronomy for sustainable development*, Volume 36, pp.1-9

Gaffar, A.W.M., Sitanggang, I.S., 2019. Spatial model for predicting sugarcane crop productivity using support vector regression. In *IOP Conference Series: Earth and Environmental Science*. IOP Publishing, Volume 335(1), p. 012009

Gregorutti, B., Michel, B., Saint-Pierre, P., 2017. Correlation and variable importance in random forests. *Statistics and Computing*, *27*, pp.659-678

Hammer, R.G., Sentelhas, P.C., Mariano, J.C., 2020. Sugarcane yield prediction through data mining and crop simulation models. *Sugar Tech*, Volume 22(2), pp.216-225

Hyndman, R.J., 2014. Measuring forecast accuracy. *Business forecasting: Practical problems and solutions*, pp.177-183

Indonesia, B. P. S., 2023. Statistik Tebu Indonesia 2022. *Jakarta: Badan Pusat Statistik*

Indrawanto, C., Purwono, Syakir, M., Siswanto, Soetopo, D., Munarso, S. J., Pitono, J., Rumini, W., 2017. Budidaya dan Pascapanen Tebu. *IAARD Press*. *Jakarta*

Jaelani, T., Yamin, M., Mahandari, C. P., 2022. Machine Learning untuk Prediksi Produksi Gula Nasional. *JMPM* (*Jurnal Material dan Proses Manufaktur*), Volume 6(1), pp.31-36

Joshua, S.V., Priyadharson, A.S.M., Kannadasan, R., Khan, A.A., Lawanont, W., Khan, F.A., Rehman, A.U., Ali, M.J., 2022. Crop yield prediction using machine learning approaches on a wide spectrum. *Computers, Materials & Continua*, *72*(3), pp.5663-5679

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

Lárraga-Altamirano, H.R., Hernández-López, D.R., Piedad-Rubio, A.M. and Blanco-Martínez, J.R., 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)*, Volume 9(1), pp.381-386

Maldaner, L.F., de Paula Corrêdo, L., Canata, T.F. and Molin, J.P., 2021. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. *Computers and Electronics in Agriculture*, *181*, p.105945

Medar, R. A., Rajpurohit, V. S., Ambekar, A. M., 2019. Sugarcane crop yield forecasting model using supervised machine learning. *International Journal of Intelligent Systems and Applications*, Volume 11(8), p.11

Mohamad, M., Selamat, A., Krejcar, O., Crespo, R. G., Herrera-Viedma, E., Fujita, H., 2021. Enhancing big data feature selection using a hybrid correlation-based feature selection. *Electronics*, Volume 10(23), p.2984

Hanka, M.K.F. and Santosa, B., 2021. Analisis kualitas bahan baku tebu melalui teknik pengklasteran dan klasifikasi kadar gula sebelum giling (studi kasus pabrik gula PT. XYZ). *Jurnal Teknik ITS*, *10*(2), pp. F100-F107

Nikhil, U.V., Pandiyan, A.M., Raja, S.P. and Stamenkovic, Z., 2024. Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models. *Computers*, Volume 13(6), p.137

Noorsaman, A., Amrializzia, D., Zulfikri, H., Revitasari, R. and Isambert, A., 2023. Machine Learning Algorithms for Failure Prediction Model and Operational Reliability of Onshore Gas Transmission Pipelines. *International Journal of Technology*, *14*(3)

Paidipati, K. K., Banik, A., Shah, B., Sangwa, N. R., 2022. Forcasting of Sugarcane Productivity Estimation in India-A Comparative Study with Advanced Non-Parametric Regression Models. *Journal of Algebric Statistics*, Volume 13(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*, Volume 10(3), pp.110-118

Pandey, N., Patnaik, P. K., Gupta, S., 2020. Data pre-processing for machine learning models using python libraries. *Int. J. Eng. Adv. Technol*, Volume 9(4), pp.1995-1999

Panigrahi, B., Kathala, K.C.R. and Sujatha, M., 2023. A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. *Procedia Computer Science*, Volume 218, pp.2684-2693

Respati E., 2022. Outlook Komoditas Perkebunan Tebu 2022. Pusat Data dan Sistem Informasi Pertanian, Kementerian Pertanian Republik Indonesia

Saidin, O.K., Lubis, M.Y., Ikhsan, E., 2021. Optimization of sustainable sugar industry towards food security. In *IOP Conference Series: Earth and Environmental Science*. IOP Publishing, Volume 782(2), p. 032039

Sari, M., Berawi, M.A., Larasati, S.P., Susilowati, S.I., Susantono, B. and Woodhead, R., 2023. Developing Machine Learning Model to Predict HVAC System of Healthy Building: A Case Study in Indonesia. *International Journal of Technology*, *14*(7), pp.1438-1448

Shah, A., Dubey, A., Hemnani, V., Gala, D., Kalbande, D. R., 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

Shetty, S.A., Padmashree, T., Sagar, B.M. and Cauvery, N.K., 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

Singla, S.K., Garg, R.D. and Dubey, O.P., 2020. Ensemble Machine Learning Methods to Estimate the Sugarcane Yield Based on Remote Sensing Information. *Revue d'Intelligence Artificielle*, Volume 34(6)

Smola, A. J., Schölkopf, B., 2004. A tutorial on support vector regression. *Statistics and computing*, Volume 14, pp.199-222

Som-ard, J., Suwanlee, S.R., Pinasu, D., Keawsomsee, S., Kasa, K., Seesanhao, N., Ninsawat, S., Borgogno-Mondino, E. and Sarvia, F., 2024. Evaluating Sugarcane Yield Estimation in Thailand Using Multi-Temporal Sentinel-2 and Landsat Data Together with Machine-Learning Algorithms. *Land*, *13*(9), p.1481

Sulaiman, A. A., Arsyad, M., Amiruddin, A., Teshome, T. T., Nishanta, B., 2023. New Trends of Sugarcane Cultivation Systems Toward Sugar Production on the Free Market: A Review. *AGRIVITA Journal of Agricultural Science*, Volume 45(2), pp.395-406

Tatachar, A.V., 2021. Comparative assessment of regression models based on model evaluation metrics. *International Journal of Innovative Technology and Exploring Engineering*, Volume 8(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*, Volume 177, p.105709

Vasconcelos, J.C.S., Speranza, E.A., Antunes, J.F.G., Barbosa, L.A.F., Christofoletti, D., Severino, F.J. and de Almeida Cançado, G.M., 2023. Development and validation of a model based on vegetation indices for the prediction of sugarcane yield. *AgriEngineering*, Volume 5(2), pp.698-719

Venkatesh, B., Anuradha, J., 2019. A review of feature selection and its methods. *Cybernetics and information technologies*, Volume 19(1), pp.3-26