



Research article

Machine learning based sugar recovery prediction model in sugarcane agroindustry

Fatata A'izza Rosyada^{a,*}, Marimin^b, Muhammad Asrol^c

^a Agro-Maritime Logistics Study Program, Graduate School, Bogor Agricultural University, Bogor 16680, Indonesia

^b Department of Agroindustrial Technology, Faculty of Agricultural Technology, Bogor Agricultural University, Bogor 11480, Indonesia

^c Industrial Engineering Department, BINUS Graduate Program - Master of Industrial Engineering, Bina Nusantara University, Jakarta 11480, Indonesia

Article Info

Article history:

Received 2 August 2025

Revised 24 October 2025

Accepted 7 April 2026

Available online 12 June 2026

Keywords:

Agroindustry,
Machine learning,
Prediction,
Sugar mill,
Sugar recovery

Abstract

Importance of the work: Sugar recovery is crucial for mill efficiency. In Indonesia, recovery rates declined by 1.91% from 2019 to 2023, hitting a decade low of 6.6% in 2022. These fluctuations indicate inefficiencies. This work introduced a machine learning-based model for early prediction and process optimization to achieve better production planning.

Objectives: To develop a model for predicting final sugar recovery with machine learning, using a multistage process and related variables in processing terms.

Materials and Methods: Day-to-day data from Sugar Mill XYZ (2020–2024) in West Java, Indonesia were used, including Brix, purity and pol values from multiple processing stages. Random forest, extreme gradient boosting (XGBoost) and artificial neural network (ANN) methods were used to develop a model and its evaluation using mean squared error (MSE) and mean absolute error (MAE). mean absolute percentage error (MAPE), coefficient of determination (R^2) and feature importance analysis.

Results: The model was developed using 699 daily milling records from 2020–2024, comprising 18 initial features. XGBoost outperformed random forest and ANN, achieving an MAE of 0.116, MSE of 0.03852, MAPE of 1.81% and a coefficient of determination (R^2) of 86.84% on the testing set. Feature significance analysis, which combines machine learning insights with empirical plant data, identified the key variables that had the greatest impact on sugar recovery, such as boiling house recovery, winter recovery, Pol in cane, milling potential efficiency and Pol in bagasse. The model correctly predicted the daily sugar recovery for production in 2024.

Main finding: This work provides a decision-support tool for sugar mill optimization. It illustrates how well XGBoost and random search optimization work together to predict sugar recovery depending on process variables.

* Corresponding author.

E-mail address: fatata.aizza@apps.ipb.ac.id (F.A. Rosyada)

online 2452-316X print 2468-1458/Copyright © 2026. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), production and hosting by Kasetsart University Research and Development Institute on behalf of Kasetsart University.

<https://doi.org/10.34044/j.anres.2026.60.3.06>

Introduction

The sugar industry is one of the major sectors that substantially contributes to the national economy and food security in Indonesia, with the primary source of sugar being sugar factories that are crucial to converting sugar cane into consumable sugar for the populace (Ramadhan et al. 2024). However, the industry faces various challenges, including fluctuations in production yields each season, process inefficiencies and competitive pressures in domestic and international markets (Asrol et al., 2020; 2021; Silalahi, 2024).

Sugar recovery (the ratio of sugar produced to the weight of sugarcane processed) has an important function in production recovery. According to data from the Ministry of Agriculture's Directorate General of Plantations (2024), Indonesia's sugar recovery varied over the previous 5 yr, declining by an average of 1.91% between 2019 and 2023. In addition, 2022 saw the lowest sugar recovery value in the previous 10 yr, reaching only 6.6% (Prasetyo, 2023). A high sugar recovery value indicates good efficiency in the production process, whereas a low value reflects potential economic losses. Several factors must be considered, including the quality of raw sugarcane and the efficiency of factory processing, which influence this variation in sugar recovery (Bantacut et al., 2012; Jaelani, et al., 2022).

Improving production process efficiency begins with the accurate prediction of sugar recovery, as this allows sugar mills to adopt a more proactive approach early on to optimize the sugar extraction process (Akbarian et al., 2022). In practice, many sugar mills in Indonesia still rely on conventional methods to predict yield values. The commonly used method involves calculating sugarcane yield based on the Pol and Brix analysis of the cane juice before the production process is completed (Subiyanto, 2017), while the final sugar yield is calculated based on the amount of crystal sugar produced divided by the amount of cane milled, expressed as a percentage (Pusdatin Kementan, 2021). In conventional practice, Pol and Brix values are measured manually using laboratory instruments such as a polarimeter and refractometer, respectively. The process requires multiple sampling points and is highly dependent on operator precision. However, this approach has substantial limitations. The Pol-Brix analysis is time-consuming and only reflects the sucrose concentration in the cane juice at the sampling point. This condition fails to consider the influence of process parameters such as maturity

factors, juice purity or boiling house recovery. Consequently, such methods fail to represent the nonlinear and cross-stage relationships between raw material quality and process losses occurring throughout milling, purification and crystallization stages.

In addition to conventional methods, the linear regression model developed by Cahyani, Marimin and Sukardi (2017) showed limitations in representing the complex and nonlinear relationships between variables affecting sugar recovery. The model's dependence on a few linear predictors, such as the MBS (Mature, Clean, Fresh) criteria and the Brix and Pol values of the sugarcane, indicates that conventional statistical methods are insufficient to capture process variability. Satpathi et al. (2025) studied sugarcane yield prediction based on a statistical approach using multiple linear regression (MLR) to assess the correlation between independent variables on weather and the dependent variable, sugarcane yield. The performance of MLR in that study was the lowest compared to penalized regression and machine learning approaches for detecting sugar recovery across several sub-districts, with the training R^2 value below 80% and the testing R^2 value below 10%, indicating the limited generalization capability of the model. Another study by Istifanus et al. (2018) applied a process-based simulation model using the MATLAB software to predict sugar yield at a Nigerian sugar factory, achieving an efficiency of about 75% with no significant difference from field data. This simulation approach effectively replicated process outputs; however, there was a lack of adaptive learning capability. Several yield prediction studies focused on predicting the sugarcane yield as raw material before the production process (Mishra et al., 2022; Nihar et al., 2022), rather than the final sugar recovery. Therefore, the current study focused on filling the research gap by applying machine learning based on process modeling.

In the current study, the prediction model was designed using regression-based machine learning. Other published research has explored various models for predicting sugarcane yield and sugar recovery, including support vector regression (Paidipati et al. 2022; Satpathi et al. 2025), support vector machine (Asrol et al., 2020), Random Forest (RF; Xu et al., 2020; Todd et al., 2021; Kumar et al., 2024) and artificial neural networks (ANN; Maldaner et al., 2021). Among these, tree-based ensemble models such as RF, XGBoost and ANN have shown superior performance due to their ability to handle nonlinear interactions, multicollinearity and process heterogeneity in agro-industrial data (Satpathi et al., 2025; Napthaleni and Asrol, 2025). Therefore, these three machine

learning models were selected in the current study to develop a predictive model of sugar recovery that integrates process efficiency indicators and quality control data across multiple factory stations.

The current study focused on several key issues: 1) identifying the factors that contribute significantly to variations in the final sugar yield from sugarcane; and 2) developing a predictive model for final sugar recovery using supervised machine learning algorithms (RF, XGBoost, ANN). The scope of this research covered: 1) the research object, which focused on the sugar agroindustry, using sugarcane as the raw material at Sugar Mill XYZ; and 2) the extent of sugar yield prediction, which was traced from sugarcane reception at the factory gate to the packaging process of white crystal sugar.

Materials and Methods

Data description and research flow

This study was conducted at Sugar Factory XYZ, a sugarcane-based agroindustry in West Java, Indonesia, using daily quality control data collected from various production stations ranging from cane reception to the final spinning stage during the 2020–2024 milling seasons. The name of the mill has been made generic to protect confidential commercially valuable data. Each observation represented an independent milling batch with unique raw material composition and process

conditions. Thus, the final sugar recovery (target variable) was modeled through a static regression framework rather than a time-series approach due to the absence of temporal dependency between consecutive days. The selection of model attributes was combined with a literature review, expert adjustment with plant managers and sugarcane researchers and data-driven feature importance analysis to identify key variables influencing sugar recovery, forming the structured research workflow presented in Fig. 1.

Data processing

Sugar recovery prediction was performed using quality control-related data collected during production at the sugar mill located in West Java Province, Indonesia. The initial dataset used for model development consisted of 18 features (Table 1), with 699 unprocessed records representing daily milling data from 2020 to 2024 in sugarcane milling seasons.

The process began with the identification of missing values, where some attribute values in the dataset were unfilled or empty (Alfarisi et al., 2013). Empty entries were removed or imputed to avoid affecting model building. In this case, missing data were addressed using a median-based imputation method for numerical variables in the predictor variables (X) to preserve the central tendency of the predictor variables while reducing the influence of outliers. For missing target data, the data row was deleted to preserve the central tendency of numerical distributions.

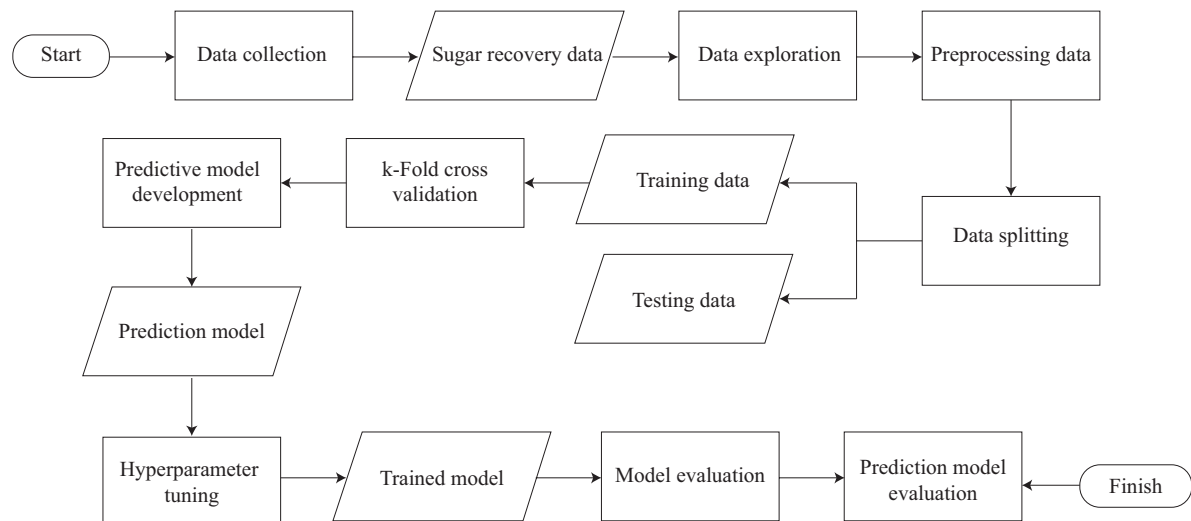


Fig. 1 Research modeling workflow

Table 1 Base parameters used in modelling.

Technique	Parameter	Value
Random forest	Number of trees (n_estimator)	100
	Maximum depth	10
	Minimum samples leaf	1
	Minimum samples split	2
Extreme gradient boosting	Number of trees (n_estimator)	100
	Maximum depth	6
	Learning rate	0.1
	Subsample	0.8
	Scale pos weight	1
	Gamma	0
Artificial neural network (mixed linear programming -regression)	Number of hidden layers	3
	Neurons per layer	[64, 32, 16]
	Activation function	ReLU (hidden), Linear (output)
	Optimizer	Adam
	Batch size	8
	Epochs	300

Next, the distributional characteristics of the features, correlations between the variables and the existence of outliers were assessed using descriptive statistical analysis. Boxplot visualization was used to detect outliers. Min-max scaling was used to normalize all numerical variables to standardize the feature space and avoid scale dominance.

Following normalization, feature importance scores from early model runs were used to select features. To cut down on noise and dimensionality, features that consistently contributed little to the target variable were removed. After that, the dataset was split 80% and 20% at random into the training and testing subsets, respectively (Karaogul et al., 2025). Random sampling ensured that both subsets maintained representative distributions of important statistical parameters to reduce sampling bias and possible overfitting.

The model's generalizability was evaluated by additional validation using a k-fold cross-validation procedure with five folds ($k = 5$) applied to the training set. Each iteration used one fold for validation and four for training, with the data divided into five equal folds. To ensure that every fold was used as the validation set once, this procedure was carried out five times. This technique enhances the reliability of performance estimation by averaging results across multiple training-validation steps to avoid variance and estimation bias (Hafid, 2023).

Feature importance analysis

The feature importance analysis was conducted in two stages. Experts in sugarcane processing participated in structured interviews to provide qualitative input during

the first phase. The aim of these expert consultations was to determine the main agronomic and operational factors affecting final sugar recovery.

The second phase applied quantitative analysis, using machine learning techniques on historical production data, to determine objectively each feature's relative contribution. Two algorithms were used: RF, implemented with the scikit-learn library (Nihar et al., 2022) and XGBoost, using the plot_importance() function from the xgboost library (Kurniawanda and Tobing 2022).

Feature importance in RF was evaluated based on each decision tree's average feature importance calculations. The feature importance calculation was based on the reduction in impurity produced by each feature when dividing nodes in the decision tree (Anshul, 2024). This is formally expressed in Eq. (1):

$$n_i = \frac{N_i}{N} \left[\text{impurity} - \left(\frac{N_{i(\text{right})}}{N_i} \times \text{right impurity} \right) - \left(\frac{N_{i(\text{left})}}{N_i} \times \text{left impurity} \right) \right] \quad (1)$$

For the XGBoost algorithm, feature importance was derived from the reduction in impurity after a feature-based split, measured using either the Gini index or information gain. The contribution of a feature was calculated based on the weighted average impurity of a parent node compared to the sum of the weighted impurities of the left and right child nodes after the split (Elith, Leathwick and Hastie 2008), as shown in Eq. (2):

$$\begin{aligned} \text{Gini index} &= 1 - \sum_{i=1}^n (P_i)^2 \\ &= 1 - [(P_+)^2 + (P_-)^2] \end{aligned} \quad (2)$$

where P_i = probability of observations belonging to class i , P_+ = probability of the positive class, P_- = probability of the negative class and n = total number of classes in the target variable (binary or multiclass classification).

Machine learning algorithms

Three machine learning algorithms were used to carry out the prediction modelling (RF, XGBoost, ANN). For the RF regression model, prediction involved constructing an ensemble of k decision trees and averaging their outputs to reduce variance. The prediction function is defined in Eq. (3):

$$\hat{f}(x) = \frac{1}{k} \sum_{i=1}^k \hat{h}_i(x) \quad (3)$$

where k is the total number of trees in the forest and $\hat{h}_i(x)$ is the prediction from the i^{th} tree.

Default settings for random forest regression parameters are shown in Table 1. XGBoost used a stage-wise additive model to minimize the residual errors of previous models (Kurniawanda and Tobing 2022). The process began by initializing predictions with the mean of the target variable, as shown in (Eq. (4) and Eq. (5):

$$h_0(x) = \text{mean}(Y) \quad (4)$$

$$\hat{Y} = Y - h_0(x) \quad (5)$$

where $h_0(x)$ = initial prediction (base learner) equal to the mean of the target variable in the training dataset, Y = observed (actual) target value and $\hat{Y}$ = residual, representing the difference between the observed value and the initial prediction.

Each subsequent model $f_i(x)$ was trained on the residuals of the preceding ensemble, with the final prediction being the sum of all base learners. The model aimed to minimize the objective function shown in Eq. (6):

$$L^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_i(x_i)) + \Omega(f_i) \quad (6)$$

where $l(\cdot)$ is the loss function and $\Omega(f_i)$ is the regularization term to penalize model complexity (Chen and Guestrin 2016). Initial processing parameters and values for the XGBoost model's parameters are shown in Table 1.

The ANN model, which is modeled after biological neurons akin to those in the human brain (Sharma and Kaur 2018), comprises three layers: the input layer, the hidden layer and the output layer. The input layer has nodes and the number of independent predictors determines the number of nodes in the input layer. The hidden layer's neurons use an activation function to convert a neuron's activation level into an output signal. The primary processes occur in the hidden layer and the outputs are then passed to the output layer. The output (h_i) of neuron i in the hidden layer is described by Eq. (7) (Wang, 2003):

$$h_i = \sigma \left(\sum_{j=1}^n V_{ij} x_j + T_i^{\text{hid}} \right) \quad (7)$$

where σ is the activation function, N is the quantity of input neurons, V_{ij} is the weights, x_j is the neuron inputs and T_i^{hid} is the hidden neurons' threshold.

Following model construction, prediction errors were averaged across iterations to assess performance. The optimizer was used to perform hyperparameter tuning to determine the best fit for both algorithms.

Model evaluation

The model was evaluated by contrasting the predicted sugar recovery values with the observed values, using two statistical error metrics: mean squared error (MSE) and mean absolute error (MAE). The mean of the squared differences between the actual and predicted values was computed to perform the MSE evaluation, as defined in Eq. (8).

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (8)$$

where n is the number of observations, y_i is the actual value and $\hat{y}_i$ is the predicted value. In contrast, the average magnitude of mistakes in a series of forecasts was represented by MAE (Shetty et al. 2021), without considering their direction, using Eq. (9):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

The mean absolute percentage error (MAPE) was used to assess the model's relative accuracy (de Myttenaere et al., 2016) and was calculated based on averaging the absolute

percentage discrepancies between the actual and anticipated values for each observation, as shown in Eq. (10):

$$\text{MAPE} = \frac{1}{n} \left(\sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \times 100 \right| \right) \quad (10)$$

The percentage of the response variable's variation to that the independent variables in the regression model could account for was measured using the coefficient of determination R^2 . R^2 is a number between 0 and 1, where higher values indicate that the model accounts for a larger percentage of the variance (Colin Cameron and Windmeijer 1997). The computation of R^2 is given in Eq. (11):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (11)$$

Results and Discussion

Data preprocessing

Data on attributes influencing the final sugar recovery were collected and subjected to a data preprocessing stage. The dataset was obtained from daily quality control reports of the installation and fabrication divisions at the sugar mill during milling days from 2020 to 2024. The data consisted of daily observations of Brix, Pol, apparent purity and other

relevant variables across various processing stages, including milling (mill.), clarification (clar.), evaporation (eva.), crystallization (crys.) and centrifugation (centr.) stations. The number of milling days changed annually depending on the factory's operating preparedness and raw material availability.

Before developing the predictive model, the raw dataset was preprocessed to guarantee data validity. The results of the first data exploration are shown in Table 2, which guided the subsequent imputation and removal of missing values, while considering the target and independent variables. Outlier detection was done in addition to missing value analysis to assess the distribution of the data.

The interquartile range approach was used to eliminate outliers to guarantee data quality and lessen the impact of extreme values. Six variables (winter recovery, apparent purity of Pan A, clarification efficiency, BHR, Pol of filter cake, apparent purity of clear juice) were subjected to this procedure after they were found to contain possible outliers. These variables exhibited extreme right-skewness or irregular distributions that could negatively impact model training and evaluation.

To address multicollinearity, Pol of extracted juice and total Brix extraction yield were excluded due to their high correlations with Brix of extracted juice and milling potential efficiency, as indicated by the Pearson correlation values ($|r| > 0.96$). Descriptive statistics of the dataset after preprocessing are presented in Table 3.

Table 2 Descriptive statistics of sugar recovery data before pre-processing.

Attribute name	Unit	Minimum	Maximum	SD	N	Missing
Pol in cane (material)	%	4.30	25.36	0.94	664	35
Brix of extracted juice (mill.)	%	13.06	19.06	1.22	666	33
Pol of extracted juice (mill.)	%	10.15	15.20	0.99	666	33
Milling potential efficiency (mill.)	%	46.73	119.96	3.34	666	33
Total of Brix extraction yield (mill.)	%	61.04	134.86	3.13	664	35
Cane juice concentration (mill.)	%	16.49	249.03	10.11	665	34
Equivalent purity Ratio (mill.)	%	10.11	100.53	1.76	666	33
Winter recovery (mill.)	%	-129.02	3811.18	150.90	659	48
Pol of bagasse (mill.)	%	1.85	3.28	0.33	665	34
Boiling house recovery (BHR)	%	0.00	198.30	16.17	643	56
Brix of raw juice (clar.)	%	9.42	13.93	0.77	666	33
Pol in filter cake (clar.)	%	0.00	3.56	0.27	659	40
Apparent purity of clear juice (clar.)	%	0.00	81.32	5.44	663	36
Clarification efficiency (clar.)	%	0.00	100.00	5.25	664	35
Brix of fine syrup (eva.)	%	0.00	68.15	4.71	661	38
Apparent purity of pan A (crys.)	%	0.00	12400	1908	665	34
Apparent purity of molasses (centr.)	%	0.00	46.58	2.90	647	52
Final sugar recovery	%	0.04	19.65	1.49	655	44

mill. = milling stage; clar. = clarification stage; eva. = evaporation stage; crys. = crystallization stage; centr. = centrifugation stage.

Table 3 Descriptive statistics of sugar recovery data after pre-processing.

Attribute name	Unit	Minimum	Maximum	SD	N	Missing
Pol in cane	%	7.83	11.48	0.58	481	0
Brix of extracted juice	%	13.14	19.06	1.17	481	0
Milling potential efficiency	%	84.56	98.60	2.36	481	0
Cane juice concentration	%	57.39	92.87	6.90	481	0
Equivalent purity ratio	%	87.24	99.23	1.48	481	0
Winter recovery	%	71.21	115.41	7.86	481	0
Pol of bagasse	%	1.85	3.24	0.32	481	0
Boiling house recovery	%	60.14	100.42	7.37	481	0
Brix of raw juice	%	9.81	13.93	0.74	481	0
Pol in filter cake	%	1.91	2.68	0.15	481	0
Apparent purity of clear juice	%	73.79	80.49	1.39	481	0
Clarification efficiency	%	1.21	17.22	2.97	481	0
Brix of fine syrup	%	46.26	68.15	1.61	481	0
Apparent purity of Pan A	%	3850	9300	1170.55	481	0
Apparent purity of molasses	%	26.65	42.93	2.39	481	0
Final sugar recovery	%	5.30	9.66	0.58	481	0

Model performance

Predictive modeling was performed using the RF, XGBoost and ANN algorithms, with 20% of the dataset put aside for testing and the remaining 80% used for training. A five-fold cross-validation process was used on the training dataset to guarantee equal representation and lessen bias between the training and testing subsets and the entire dataset. A baseline model with relatively low accuracy was obtained after constructing the initial model using default parameters (see Table 1). The MSE value was 1.6801 with an R^2 of 13.06% and an MAPE of 8.52% in the RF model. A hyperparameter adjustment was required to optimize the model under these conditions. For better model performance, hyperparameter tuning was carried out utilizing the random search technique. The optimal hyperparameters for the RF and XGBoost models were obtained using random search, see Table 4, while ANN used Adam optimization.

Following hyperparameter tuning, model validation was conducted using the previously separated testing dataset.

Table 4 Optimal hyperparameters used.

Algorithm	Parameter	Value
Random forest	Number of trees (n_estimator)	100
	Maximum depth	None
	Max_features	Auto
	Minimum samples leaf	1
	Minimum samples split	2
Extreme gradient boosting	Number of trees (n_estimator)	100
	Maximum depth	3
	Learning rate	0.2
	Subsample	0.6
	Scale pos weight	0.8
	Gamma	0

Singla, Garg and Dubey (2020) studied the performance evaluation used to estimate final sugar recovery, utilizing performance criteria such as MSE, MAE, MAPE and R^2 . Table 5 displays the best evaluation results for the current models and the parameters aiding in sugar recovery prediction.

Based on 15 chosen features, the XGBoost model had the best predicted final sugar recovery (MSE = 0.038 and MAE = 0.116). In addition, the model produced an R^2 value of 86.84% and MAPE of 1.81%, indicating high accuracy and explained variance.

Fig. 2 compares the predicted and actual sugar recovery values. Visual examination revealed that the predicted values closely matched the actual values, suggesting that this model could be practical for predicting the final sugar recovery under the right input feature and parameter settings. Furthermore, Fig. 3 compares the actual and predicted daily sugar recovery values for the 2024 milling season. There was little deviation and strong alignment between the actual trend and the predicted values. On day 43, there was just a slight deviation. Overall, the model was reliable and accurate at forecasting sugar recovery.

Table 5 Performance evaluation statistics for machine learning models.

	Final sugar recovery		
	RF	XGB	ANN
MSE	0.059	0.038	0.0182
MAE	0.1512	0.116	0.8485
MAPE	2.16%	1.81%	1.12%
R^2	0.808	0.8684	0.8989

RF = random forest method; XGB = extreme gradient boosting; ANN = artificial neural network; MSE = mean squared error; MAE = mean absolute error; MAPE = mean absolute percentage error; R^2 = coefficient of determination.

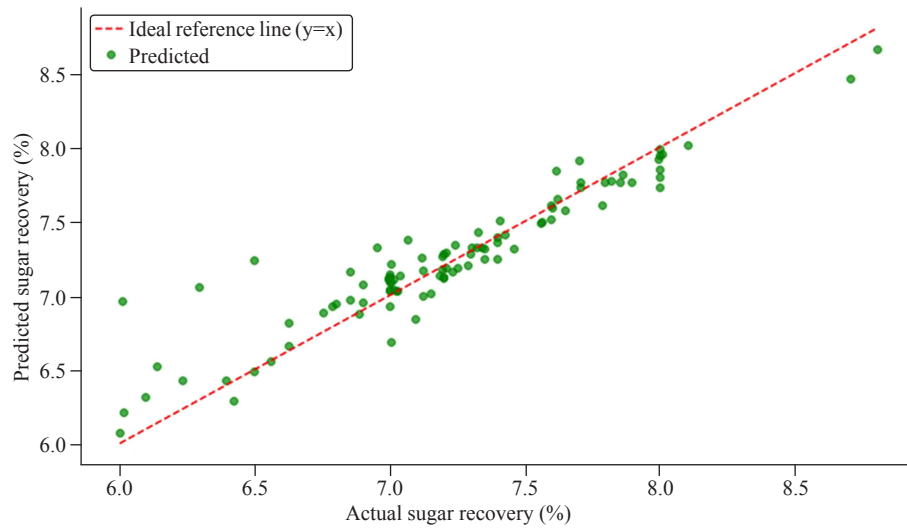


Fig. 2 Actual versus prediction values for sugar recovery.

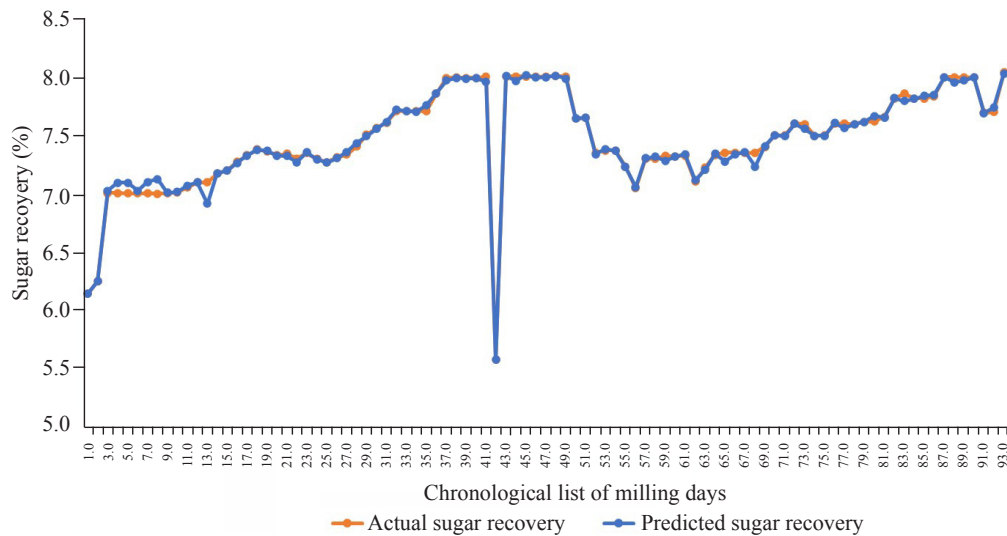


Fig. 3 Actual versus prediction values for daily sugar recovery in 2024.

Feature importance results

The RF and XGBoost models showed a consistent pattern in the feature importance analysis results, highlighting a set of dominant variables that greatly influenced final sugar recovery predictions (Fig. 4). The RF model identified the top-three contributing features (with markedly higher importance scores than the remaining variables) as: BHR, representing the recovery efficiency at the boiler house stage and computed as the ratio of sugar solids in syrup to those in mixed juice, independent of the final sugar output; WR; and HPG. These findings highlight how important raw material properties and downstream process efficiency influence the ultimate sugar recovery. Since they play a part in intermediate processing stages, other variables, such as the apparent purity

of Pan A and Pol in cane, were also determined to be of moderate importance.

The XGBoost model (Fig. 4) had a comparable feature ranking to that of the RF model, with BHR, WR and HPG having the greatest influence. The consistency between the two algorithms supported the robustness of these variables as important predictors. Clarification efficiency and Pol of filter cake were also given considerable weight by XGBoost, indicating that minute changes in juice purity and by-product recovery could further influence model performance. The selected features' validity was reinforced by the congruence of the data-driven insights and the qualitative expert judgment, which facilitated their incorporation into the predictive systems for sugar mill operational optimization.

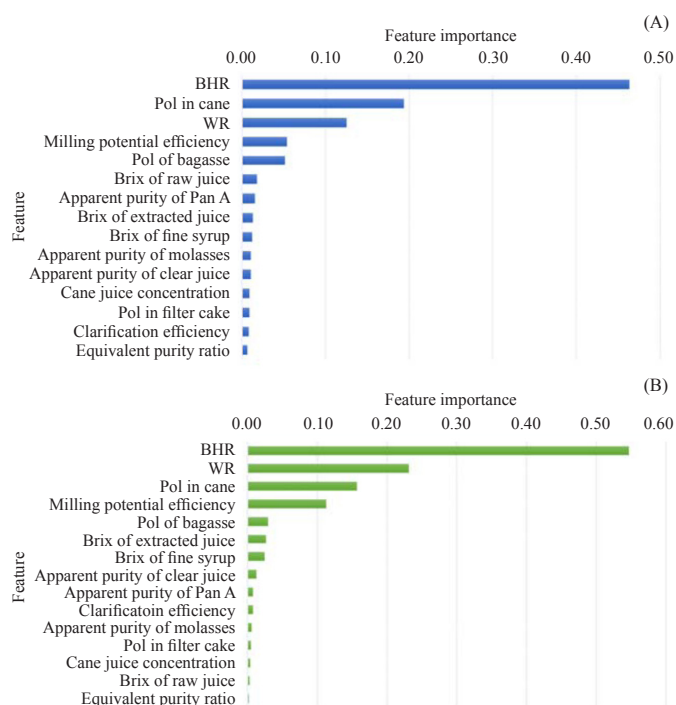


Fig. 4 Feature importance ranking of input variables used in predictive model for sugar recovery based on: (A) random forest; and (B) extreme gradient boosting.

Implications for sugarcane production

The developed prediction model has the potential to improve decision-making for sugarcane production as sugar mills could predict expected yield outcomes more precisely using model outputs during the milling season. This would allow for more efficient planning of operational tasks such as milling schedules, equipment readiness, labor allocation and raw material logistics. Proactive corrective actions, such as process modification or material segregation, could be informed by early detection of low-recovery batches, decreasing production inefficiencies and increasing throughput overall.

Furthermore, real-time sugar recovery optimization would be possible by dynamically adjusting process parameters based on integrating predictive modeling into production workflows. For example, targeted interventions could be implemented at the clarification, evaporation or crystallization stations to mitigate yield loss if model predictions showed a lower expected recovery because of decreased apparent purity or suboptimal BHR. Such predictive insights could also help with harvest scheduling and raw material procurement planning in the larger framework of supply chain optimization, guaranteeing that premium cane was given priority and processed under ideal

circumstances. Ultimately, this data-driven strategy would help the sugarcane industry transition to precision agro-industrial management.

Conclusion and Recommendation

A predictive model was successfully developed and validated, using adjusted and validated datasets to ensure the predictive model's dependability for real-world applications. The study showed that the XGBoost algorithm was effective at predicting the final sugar recovery based on key process-related variables. The model was developed using 16 selected features, refined through expert validation and optimized using hyperparameter tuning via the random search method. Based on important process-related variables, the study demonstrated that the XGBoost algorithm could provide accurate forecasts of the final sugar recovery. In total, 16 chosen features were used to build the model, which was improved based on expert validation and hyperparameter tuning using the random search technique. The model performed excellently, with an MSE of 0.03852, an MAE of 0.116 and an exceptionally low MAPE of 1.81%. These results demonstrated the model's high predictive accuracy. The R^2 value of 86.84% confirmed the strong fit between the actual and predicted values. Boiling house recovery, winter recovery, milling potential efficiency and Pol in cane were the most important variables.

The optimized model should be incorporated into the factory's operational and logistics systems to facilitate real-time decision-making and process optimization. Facilitation could include making digital dashboards, modules for planning raw materials and tools for scheduling harvests. Real-time data streaming from production systems and in-field sensors would be facilitated if future research could focus on automating the predictive framework through Internet of Things integration. This would allow for continuous monitoring and predictive analytics, improving operational efficiency and scalability in sugar mills.

Conflict of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This research was conducted as part of a collaborative research project funded by the BIMA Applied Research Grant from the Indonesian Ministry of Education, Culture, Research and Technology and was facilitated by IPB University. Sugar Factory XYZ provided the operational data and quality control measures essential to this research.

References

- Akbarian, S., Xu, C., Wang, W., Ginns, S., Lim, S. 2022. Sugarcane yields prediction at the row level using a novel cross-validation approach to multi-year multispectral images. *Comput. Electron. Agric.* 198: 107024. doi.org/10.1016/j.compag.2022.107024
- Alfarisi, A.R., Tjandrasa, H., Ariesianti, I. 2013. Perbandingan performa antara imputasi metode konvensional dan imputasi dengan algoritma mutual nearest neighbor. *J. Tek. Pomits.* 2: 73–76. [in Indonesian]
- Anshul. 2024. “An Introduction to Random Forest Algorithm for Beginners.” *Analytics Vidhya*. <https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/#:~:text=To calculate feature importance using,the average of these numbers.%0A>.
- Asrol, M., Marimin, Machfud, Yani, M., Taira, E. 2020. Supply chain fair profit allocation based on risk and value added for sugarcane agro-industry. *Oper. Supply Chain Manag.* 13: 150–165. doi.org/10.31387/OSCM0410259
- Asrol, M., Marimin, Machfud, Yani, M., Taira, E. 2021. Risk management for improving supply chain performance of sugarcane agroindustry. *Ind. Eng. Manag. Syst.* 20: 9–26. doi.org/10.7232/iems.2021.20.1.9
- Asrol, M., Marimin, Machfud, Yani, M. 2020. Business intelligence model construction to improve sugarcane yield for a sustainable sugar industry. *J. Adv. Res. Dyn. Control Syst.* 12: 109–118. doi.org/10.5373/JARDCS/V12SP6/SP20201013
- Bantacut, T., Sukardi, SPatma, I.A. 2012. Kehilangan gula dalam sistem tebang muat angkut di Pabrik Gula Sindang Laut dan Tersana Baru, Cirebon. *J. Teknol. Ind. Pertan.* 13: 199–206. [in Indonesian]
- Cahyani, W.K.D., Marimin, M., Sukardi, S. 2017. Model produktivitas bagi hasil agroindustri gula tebu dalam kemitraan antara petani dan perusahaan: Studi kasus di PG Kremboong, Sidoarjo. *J. Teknol. Ind. Pertan.* 27: 114–24. doi.org/10.24961/j.tek.ind.pert.2017.27.2.114. [in Indonesian]
- Chen, T., Guestrin, C. 2016. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM. pp. 785–794.
- Cameron, A.C., Windmeijer, F.A.G. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econom.* 77: 329–342. doi.org/10.1016/S0304-4076(96)01818-0
- Elith, J., Leathwick, J.R., Hastie, T. 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77: 802–813. doi.org/10.1111/j.1365-2656.2008.01390.x
- Maldaner, L.F., de Paula Corrêdo, L., Canata, T.F., Molin, J.P. 2021. Predicting the sugarcane yield in real-time by harvester engine parameters and machine learning approaches. *Comput. Electron. Agric.* 181: 105945. doi.org/10.1016/j.compag.2020.105945
- Haifid, H. 2023. Penerapan K-fold cross validation untuk menganalisis kinerja algoritma K-nearest neighbor pada data kasus Covid-19 di Indonesia. *J. Math. Comput. Stat.* 6: 161–168. [in Indonesian]
- Istifanus, A.B., Ezekiel, A.M., Ezeanya, N.C., Gagare, H.U. 2018. Prediction of sugar yield from sugar cane using process modeling. *Int. J. Res. Eng. Technol.* 07: 57–74. doi.org/10.15623/ijret.2018.0706009
- Jaelani, T., Yamin, M., Mahandari, C.P. 2022. Machine learning untuk prediksi produksi gula nasional. *J. Mater. Proses Manufaktur.* 6: 31–36. doi.org/10.18196/jmpm.v6i1.14897. [in Indonesian]
- Karaogul, E., Sark, G., Ögütü, A.S. 2025. Bio-hybrid films from chirich tuber starch: A sustainable approach with machine learning-driven optimization. *Sustainability.* 17: 1935. doi.org/10.3390/su17051935
- Kumar, S., Pant, M., Nagar, A. 2024. Forecasting the sugarcane yields based on meteorological data through ensemble learning. *IEEE Access.* 12: 176539–176553. doi.org/10.1109/ACCESS.2024.3502547
- Kurniawanda, M.R., Tobing, F.A.T. 2022. Analysis sentiment cyberbullying in Instagram comments with XGBoost method. *Int. J. New Media Technol.* 9: 28–34. doi.org/10.31937/ijnmt.v9i1.2670
- Mishra, P., Nehru, J., Vishwa, K., et al. 2022. Modeling and forecasting of sugarcane production in South Asian countries. *Curr. Appl. Sci. Technol.* 23: 1–15. doi.org/10.55003/cast.2022.01.23.002
- de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F. 2016. Mean absolute percentage error for regression models. *Neurocomputing.* 192: 38–48. doi.org/10.1016/j.neucom.2015.12.114
- Naphtaleni, Asrol, M. 2025. Design of a sugarcane yield and productivity prediction model. *Int. J. Technol.* 16: 1484–1500. doi.org/10.14716/ijtech.v16i5.7233
- Nihar, A., Patel, N.R., Danodia, A. 2022. Machine-learning-based regional yield forecasting for sugarcane crop in Uttar Pradesh, India. *J. Indian Soc. Remote Sens.* 50: 1519–1530. doi.org/10.1007/s12524-022-01549-0
- Paidipati, K.K., Banik, A., Shah, B., Sangwa, N.R. 2022. Forecasting of sugarcane productivity estimation in India - a comprehensive study with advanced non-parametric regression models. *J. Algebr. Stat.* 13: 760–778.
- Prasetyo, A. 2023. Teknologi jadi jalan keluar defisit gula nasional. *KOMPAS*, Jakarta. Available Source: <https://www.kompas.id/baca/ekonomi/2023/12/13/teknologi-jadi-jalan-keluar-defisit-gula-nasional>. [in Indonesian]
- Pusat Data dan Sistem Informasi Pertanian Kementerian Pertanian [Pusdatin Kementan]. 2021. Buletin tebu/gula triwulan II-2021. Jakarta. [in Indonesian]
- Ramadhan, A.J., Priya, S.R.K., Pavithra, V., et al. 2024. Machine learning techniques for sugarcane yield prediction using weather variables. In: Aldahan, N., Ramadhan, A.J. (Eds.). *BIO Web Conf.* 97: 00157. doi.org/10.1051/bioconf/20249700157
- Satpathi, A., Chand, N., Setiya, P., et al. 2025. Evaluating statistical and machine learning techniques for sugarcane yield forecasting in the Tarai region of North India. *Comput. Electron. Agric.* 229: 109667. doi.org/10.1016/j.compag.2024.109667
- Sharma, S., Kaur, R. 2018. An ANN based approach for software fault prediction using object oriented metrics. Independently Published. Punjab, India.

- Shetty, S.A., Padmashree, T., Sagar, B.M., Cauvery, N.K. 2021. Performance analysis on machine learning algorithms with deep learning model for crop yield prediction. In: *Proceedings of the International Conference on Artificial Intelligence and Evolutionary Computations in Engineering Systems*. Chennai, India. pp. 739–750. doi.org/10.1007/978-981-15-8530-2_58
- Silalahi, A.V. 2024. Kebijakan pengembangan tebu menuju swasembada gula konsumsi. *J. Agric. Dev. Plan.* 1: 75–86. [in Indonesian]
- Singla, S.K., Garg, R.D., Dubey, O.P. 2020. Ensemble machine learning methods to estimate the sugarcane yield based on remote sensing information. *Rev. Intell. Artif.* 34: 731–743.
- Subiyanto. 2017. Analisis keragaman parameter penentu rendemen gula kristal putih pada pabrik gula BUMN. *Maj. Ilm. Pengkaj. Ind.* 11: 1–10. [in Indonesian]
- Todd, J., Dufrene, E., Waguespack, H., Kimbeng, C., Pontif, M., Boykin, D. 2021. Data mining sugarcane breeding yield data for ratoon yield prediction. *Euphytica* 217: 2786. doi.org/10.1007/s10681-021-02786-z
- Wang, S.-C. 2003. Artificial neural network. In: *Interdisciplinary computing in Java programming*. Springer, Boston, MA, USA, pp. 81–100. doi.org/10.1007/978-1-4615-0377-4_5
- Xu, J.X., Ma, J., Tang, Y.N., et al. 2020. Estimation of sugarcane yield using a machine learning approach based on UAV-LiDAR data. *Remote Sens.* 12: 2823. doi.org/10.3390/rs12172823