DOI: https://doi.org/10.34123/jurnalasks.v18i1.973

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 18 No. 1, June 2026
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

Ensemble Boosting Models for Forecasting Rice Prices in Indonesia

Muhammad Jimmy Saputra[1], Yeni Rahkmawati[2*], Selvi Annisa[3], Anne Mudya Yolanda[4]

[1,2,3]Univeristas Lambung Mangkurat, Banjarbaru, Indonesia, [4]University of Leeds, Leeds, United Kingdom
**Corresponding Author: E-mail address: yeni.rahkmawati@ulm.ac.id*

ARTICLE INFO

Article history:
Received 6 May, 2026
Revised 15 June, 2026
Accepted 25 June, 2026
Published 30 June, 2026

Keywords:
Rice Price; Ensemble Boosting; GBM; LightGBM; Forecasting

Abstract

Introduction/Main Objectives: Rice is a key staple commodity influencing food security and inflation in Indonesia, making accurate price forecasting essential. In this study, we aim to compare ensemble boosting models and identify the best-performing model for rice price prediction. **Background Problems**: Notably, rice prices exhibit non-linear patterns over time, while classical statistical methods have limitations in capturing such complexities, resulting in suboptimal forecasting performance. **Novelty**: This study proposes a lag-based approach that uses lag variables as the only predictors, arranged across multiple input schemes to flexibly capture historical patterns without external variables. **Research Methods**: Daily national medium rice price data (Jan 2021–Jan 2026) from the National Food Agency are modeled using Gradient Boosting Machine (GBM) and LightGBM, with hyperparameter tuning via Optuna. The forecasting framework relies exclusively on significant lag variables without incorporating exogenous factors. Model performance is evaluated using RMSE, MAE, and MAPE. **Findings/Results**: LightGBM with optimized hyperparameters achieves the best performance (RMSE = 66.389; MAE = 50.213; MAPE = 0.362%). Furthermore, forecasts for the next 89 days indicate stable prices around Rp13,360–Rp13,395/kg, with no significant fluctuations.

1. Introduction

Rice is a fundamental food source for a majority of global population. Its consumption are found is concentrated in Asia, Sub-Saharan Africa, and South America, where it plays a crucial role in meeting dietary energy and carbohydrate requirements. Indonesia ranks fourth globally among rice-consuming countries, with consumption reaching 35.367 million tons in 2020 [1]. Moreover, data from the National Socioeconomic Survey (Susenas) conducted in March 2025 indicate that rice, including local rice, premium-quality rice, and imported rice, has the highest consumption participation rate in Indonesia at 99.07%, highlighting the strong reliance of Indonesian society on rice as a daily staple food [2].

Rice is a strategic food commodity whose prices are classified within the volatile food group and are highly sensitive to supply disruption and market dynamics, thereby directly influencing year-on-year inflation in Indonesia [3]. This relationship is illustrated in Figure 1, where rice consistently appears as one of the main contributors to inflation, although its impact may fluctuate overtime. Typically, increases in rice prices are driven by imbalances between supply and demand, which are caused by factors such as El Niño, rising non-subsidized fertilizer prices, limited availability of subsidized fertilizers, and irrigation damage [4]. In response to these challenges and to maintain price stability and food security, the government implements the second mission of Asta Cita through strategies such as agrarian reform, fertilizer assurance, irrigation improvement, distribution efficiency, import policies, and food reserve management [5].

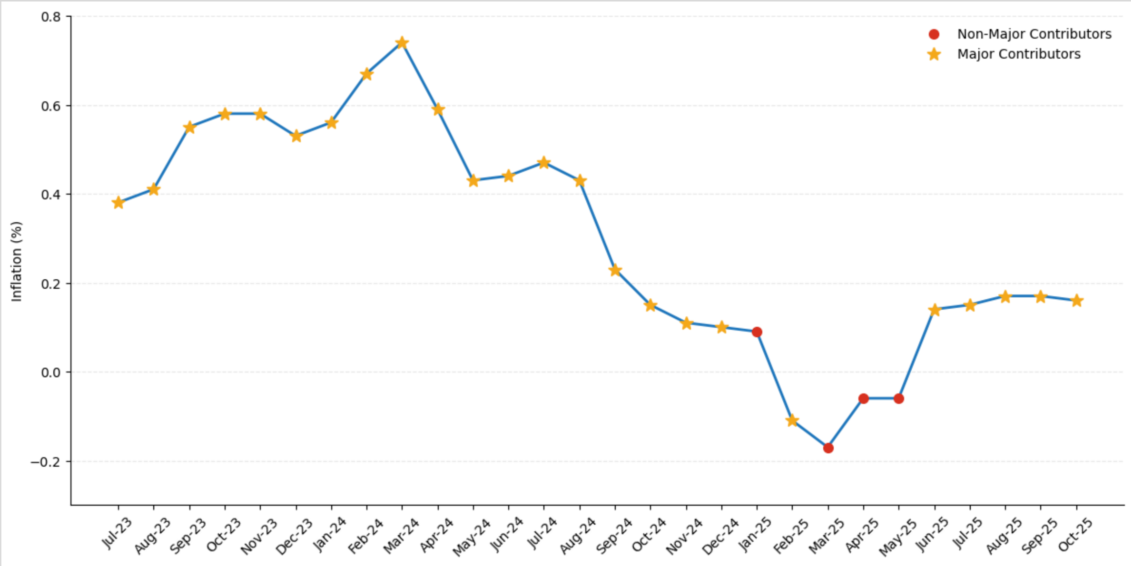

Figure 1. Contribution of rice commodity to year-on-year inflation

Government efforts to maintain long-term price stability can be supported by rice price forecasting, which enables the anticipation of future trends and the formulation of more effective policy responses [6]. Previous studies on rice price forecasting have employed various traditional statistical models, including Double Exponential Smoothing [7], Linear Regression [8], ARIMA [9], and SARIMA [10]. However, advancements in science and technology have led to the increasing adoption of modern statistical approaches, particularly machine learning, due to their ability to capture non-linear relationships and complex interactions among variables that are often difficult to handle using traditional methods [11]. Several machine learning models applied in forecasting studies include Random Forest, Support Vector Machine (SVM) [12], Decision Tree (DT), Gradient Boosting Machine (GBM) [13], and Extreme Gradient Boosting (XGBoost) [14].

One of the most widely developed machine learning approaches is ensemble boosting, a sequential learning technique that combines multiple weak learners to construct a strong predictive model, where each successive model focuses on correcting the errors of its predecessor [15]. Common implementations of this method include Gradient Boosting Machine (GBM) and Light Gradient Boosting Machine (LightGBM). GBM builds decision trees iteratively to improve accuracy [16], but its performance becomes less efficient when applied to large datasets due to the need to processes the entire dataset at each iteration. As an advancement of GBM, LightGBM enhances computational efficiency by focusing only on important data and combining certain features without sacrificing accuracy [17].

Various studies have explored the capability of ensemble boosting models across different fields. For instance, GBM has been shown to produce wind power generation predictions with the lowest nRMSE among boosting models [18]. Similarly, in the agricultural commodity sector, LightGBM has consistently outperformed six other models, namely ARIMA, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Artificial Neural Network (ANN), in predicting agricultural product prices [19]. Taken together, these findings justify the selection of GBM and LightGBM as a suitable methods for forecasting rice prices, which exhibit substantial volatility.

This study adopts a distinct approach by utilizing lag variables as the sole predictors, which are combined into multiple input schemes to represent historical information. Each scheme is evaluated to examine its effect on model performance in capturing rice price movement patterns. To further enhance model performance, hyperparameter tuning is performed using Optuna. This framework allow for flexible exploration of temporal relationships without relying on a single variable combination. This study intends to support the government in formulating food self-sufficiency policies through more accurate rice price forecasting.

2. Material and Methods

2.1. Data Source

The data used in this study are secondary data obtained from the official website of the National Food Agency. The dataset is a daily time series covering the period from January 2021 to January 2026, with a total of 1,857 observations. The variable used is the national medium rice price at the consumer level, measured in Indonesian Rupiah per kilogram (IDR/kg), which is subsequently modeled and forecasted. Medium rice refers to rice with a moderate quality level that meets the standards set by the National Food Agency Regulation No. 2 of 2023. Its main characteristics include a minimum milling degree of 95%, a maximum moisture content of 14%, a maximum of 2% broken fragments (menir), a maximum of 25% broken grains, a maximum of 4% other rice grains, a maximum of 1 unhusked grain per 100 grams, and a maximum of 0.05% foreign matter [20] The selection of medium rice is based on its role as the most widely consumed type of rice among low- to middle-income populations, particularly households slightly above the poverty line [21]. Therefore, fluctuations in medium rice prices can reflect changes in society's purchasing power and are directly linked with economic stability and food policy.

2.2. Research Framework

This study aims to develop a rice price forecasting model to understand price movement patterns and support food policy decision-making. The research stages include data collection, data preprocessing, construction of input variable, variable selection, model development, and evaluation and interpretation of results. Visualization, analysis, and modeling were conducted using Python 3.12.13 in the Google Colab environment, utilizing libraries such as NumPy, pandas, scikit-learn, LightGBM, statsmodels, Optuna, and Matplotlib. The expected outcome is the best-performing model for rice price forecasting, along with prediction results that illustrate future price trends, as presented in the research framework in Figure 2.

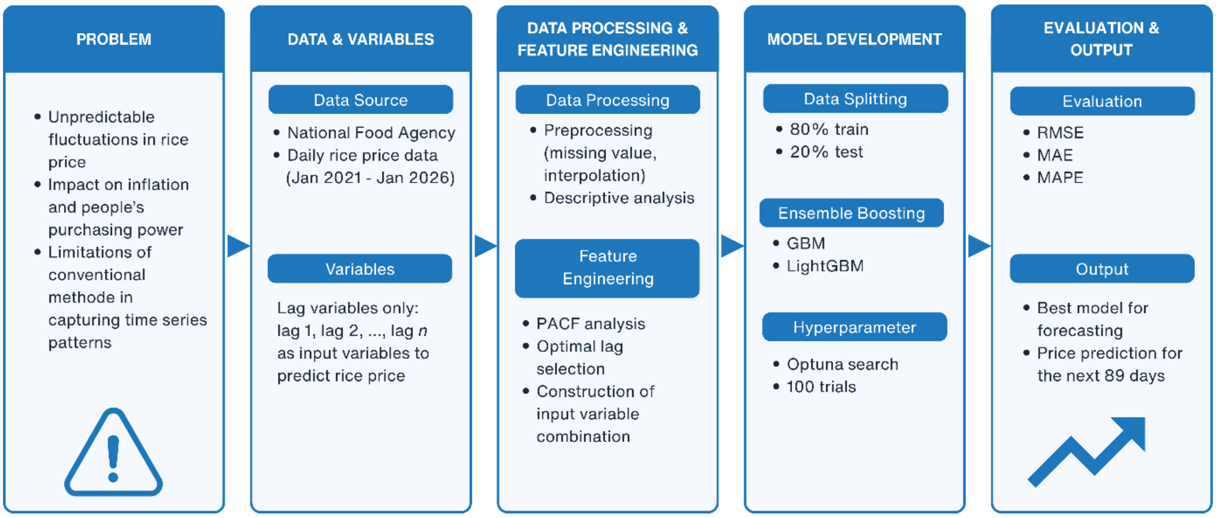

Figure 2. The research framework

2.3. Linear Interpolation

Linear interpolation is a commonly used method for handling missing data, with its primary advantage being its ability to preserve the underlying trend without introducing unnecessary fluctuations [22]. By applying this approach, data consistency can be maintained while reducing potential bias that may affect subsequent analysis, particularly in time series contexts where continuity between observations is essential. The method operates by constructing a straight line connecting two known data points surrounding the missing value [23], allowing the unknown value to be estimated as a proportion between them. Despite its simplicity, this approach is effective because it assumes a linear relationship between the observations before and after the missing interval.

$$y_t = y_{t-1} + \frac{y_{t+1} - y_{t-1}}{x_{t+1} - x_{t-1}}(x_t - x_{t-1}) \tag{1}$$

In Equation (1), y_t represents the estimated value at time t, y_{t-1} and y_{t+1} denote the observed values at the previous and subsequent time points, respectively, while x_t, x_{t-1}, and x_{t+1} correspond to the time indices of the missing, previous, and subsequent observations.

2.4. Partial Autocorrelation Function

The Partial Autocorrelation Function (PACF) measures the correlation between observations Y_t and Y_{t+k} after removing the linear influence of all intermediate observations, namely $Y_{t+1}, Y_{t+2}, \dots, Y_{t+k-1}$. In other words, PACF captures the direct correlation between two data points separated by lag k, while controlling for the linear dependencies of the intervening variables [24]. Mathematically, this function is expressed as the following conditional correlation:

$$\phi_{kk} = \text{corr}(Y_t, Y_{t+k} \mid Y_{t+1}, \dots, Y_{t+k-1}) \tag{2}$$

Here, ϕ_{kk} denotes the PACF coefficient at lag k, Y_t represents the time series value at time t, and Y_{t+k} is the value at time $t+k$, while k indicates the lag and t represents time. The sample PACF can be mathematically defined as:

$$\hat{\phi}_{kk} = \frac{\hat{\rho}_k - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_{k-j}}{1 - \sum_{j=1}^{k-1} \hat{\phi}_{k-1,j}\hat{\rho}_j} \tag{3}$$

with the recursive relation:

$$\hat{\phi}_{kj} = \hat{\phi}_{k-1,j} - \hat{\phi}_{kk}\hat{\phi}_{k-1,k-j}, j = 1,2, \dots, k-1. \tag{4}$$

In these expressions, $\hat{\phi}_{kk}$ represents the estimated PACF value at lag k, while $\hat{\phi}_{k-1,j}$ denotes the estimated PACF at lag j from lag $k-1$. Furthermore, $\hat{\rho}_k$ is the estimated ACF at lag k, $\hat{\rho}_{k-j}$ is the estimated ACF at lag $k-j$, and $\hat{\rho}_j$ is the estimated ACF at lag j. Here, k indicates the lag, and j is the lag index.

The selection of input variables is crucial for forecasting model performance. Before training, an initial set of input must be determined. The PACF is used to identify these initial inputs, where variables at specific lags are considered potential predictors if their PACF values fall outside the confidence interval bounds, indicating statistical significance. The corresponding lags highlight significant autocorrelation, which can be visualized with the PACF plot [25]. By leveraging PACF and considering other relevant information, an initial set of input variables for different subseries can be approximated [26].

2.5. Gradient Boosting Machine

Gradient Boosting Machine (GBM) is a supervised learning algorithm based on ensemble boosting that builds models sequentially to optimize a specific objective function. At each iteration, the model improves previous prediction errors by fitting weak learners, typically decision trees, to the residual errors. A weak learner is a simple model with predictive performance slightly better than random guessing, but through sequential boosting its performance can be significantly improved [27]. The main objective of GBM is to minimize the loss function by combining weak learners in an additive manner [28].

Figure 3 illustrates that GBM constructs a new weak learner by following the gradient direction to reduce residuals. This process uses multiple regression trees as weak learners and applies a forward stagewise learning approach to iteratively minimize residuals from the previous iteration. At the same time, the algorithm adjusts sample weights across iterations to generate subsequent weak learners, allowing the model to better capture patterns in the data and improve prediction accuracy at each step. To obtain more accurate predictions, all weak learners produced in each iteration are then combined linearly through an additive model [29].

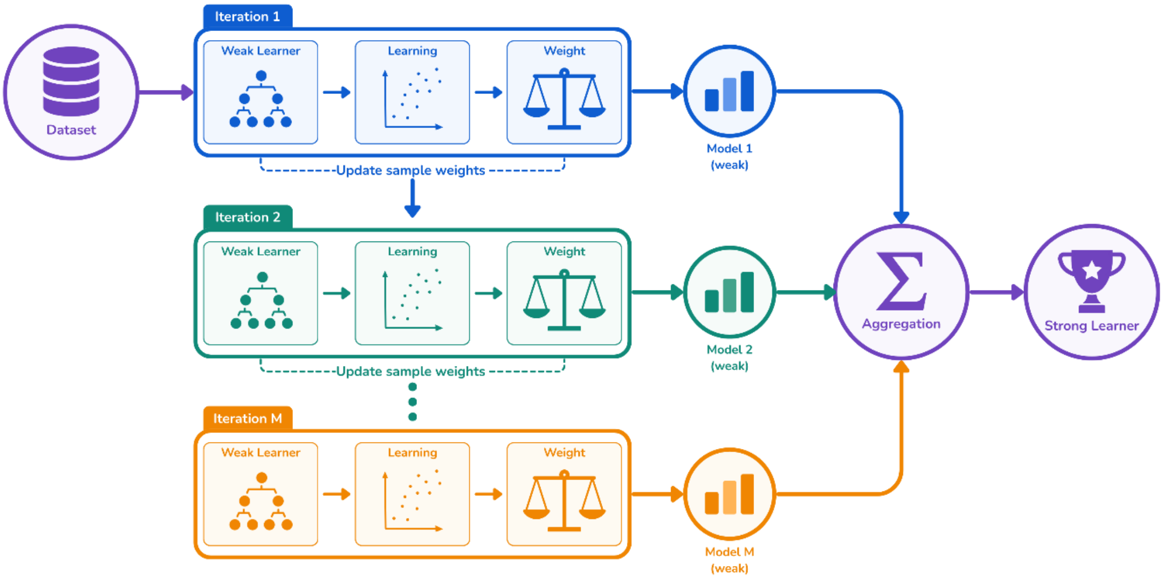

Figure 3. GBM workflow

The process begins with the initialization of a base model that minimizes error on the training data [30], followed by the computation of pseudo-residuals to construct weak learners at each iteration. The model is updated iteratively by adding the contribution of each new weak learner until a stable final model is obtained in the form of an additive regression function:

$$\hat{y}_t = F_0(x_t) + \sum_{m=1}^{M} \nu \cdot h_m(x_t) \tag{5}$$

In equation (5), $\hat{y}_t$ represents the predicted value for the t-th observation, $F_0(x)$ is the initial prediction before boosting begins, ν is the learning rate that controls the contribution of each weak learner, $h_m(x_t)$ denotes represents the weak learner at iteration m evaluated using the input features at time t, and M is total number of boosting iterations used to construct the model.

2.6. *Light Gradient Boosting Machine*

Light Gradient Boosting Machine (LightGBM) is a decision tree-based model that extends the ensemble boosting approach with high computational efficiency. Its main advantage lies in the use of a *leaf-wise tree growth* strategy, which expands the node that yields the largest error reduction in a greedy manner, allowing faster training compared to conventional methods [31]. In contrast to *level-wise tree growth*, which grows the tree layer by layer where the number of nodes at depth D reaches 2^D, the leaf-wise approach focuses on error optimization, leading to improved speed and model performance [32].

During the learning process, each tree is constructed to minimize the loss function while controlling model complexity through regularization, ensuring a balance between accuracy and model stability. This approach also utilizes gradient and Hessian information to improve the quality of model updates at each iteration. In addition, LightGBM incorporates several acceleration techniques such as histogram-based algorithms, Gradient-Based One-Side Sampling (GOSS), and Exclusive Feature Bundling (EFB), which enable efficient processing of large-scale data without degrading model performance [33].

In general, the final prediction of LightGBM is obtained by summing the contributions of all decision trees built sequentially, as follows:

$$\hat{y} = \sum_{i=1}^{I} T_i\,(X, \theta_i) \tag{6}$$

In this equation, $\hat{y}$ represents the final predicted value, T_i denotes the i-th decision tree, X is the set of input variables, θ_i represents the parameters learned by the i-th tree, and I is the total number of trees in the model.

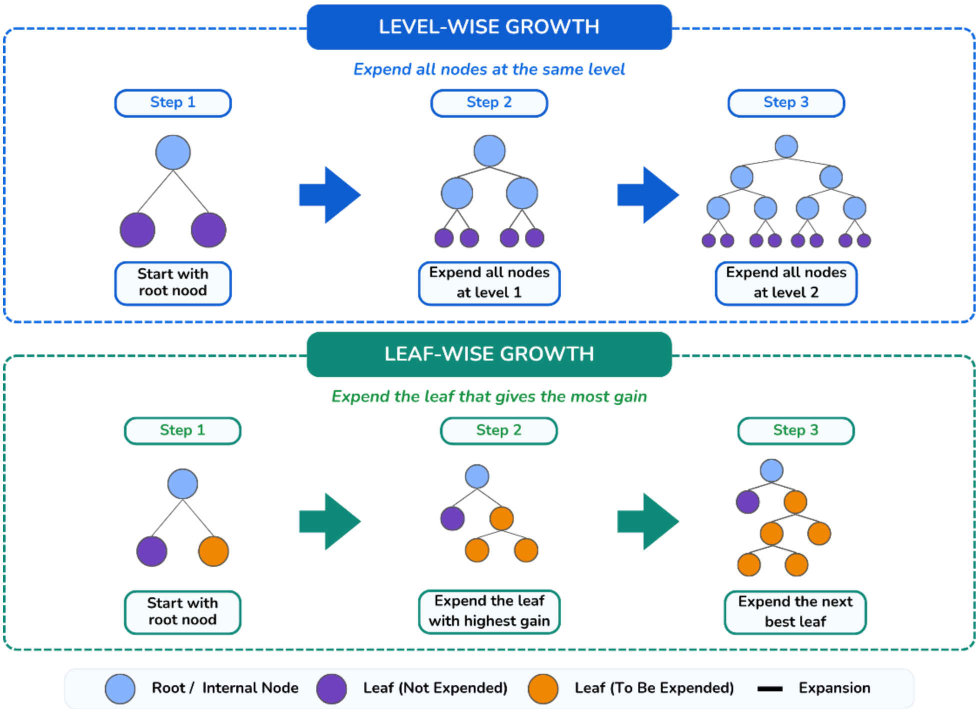

Figure 4. Level-wise and leaf-wise growth mechanism

2.7. Optuna Search

Optuna is an automated framework for optimization hyperparameters in machine learning models. This library enables a flexible hyperparameter search process through a define-by-run approach, allowing the search space to be dynamically defined within the code according to the desired conditions. Hyperparameters themselves are parameters that are set prior to the model training process and remain constant throughout training. These parameters play a crucial role in determining the model's learning capacity, generalization ability, and computational cost. Compared to traditional optimization methods such as grid search and random search, Optuna offers a more flexible and efficient approach to exploring complex search spaces [34].

In its optimization process, Optuna commonly employs the Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian optimization method operating within the Sequential Model-Based Optimization (SMBO) framework. This algorithm divides the parameter space into two probability distributions: one representing trials with good performance and another representing the remaining trials. Candidate parameters are then selected by maximizing the ratio between these two distributions, guiding the search toward more promising regions of the parameter space. This approach enables more efficient exploration, particularly in high-dimensional optimization problems [35]. Furthermore, Optuna provides several advantages that enhance optimization efficiency. The Bayesian optimization algorithm leverages results from previous trials to guide the search process, allowing optimal solutions to be identified with fewer trials compared to conventional methods. Optuna also incorporates a pruning mechanism that monitors the performance of each trial in real time and terminates underperforming trials at an early stage. As a result, computational resources can be focused on the most promising hyperparameter combinations, making the optimization process more efficient [36].

2.8. Model Evaluation

The evaluation methods used in this study include Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). MAE measures the average magnitude of errors between predicted and actual values without considering their direction, making it easy to interpret since it has the same unit as the original data [37]. RMSE is the square root of the average squared differences between predicted and observed values, making it more sensitive to large errors and useful for detecting high deviations. Meanwhile, MAPE expresses the error in percentage terms, which facilitates interpretation as it reflects the average relative error compared to actual values [38]. The smaller the values of these three metrics, the better the model performance.

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|\hat{y}_t - y_t| \tag{7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t - y_t)^2} \tag{8}$$

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \times 100\% \tag{9}$$

In these equations, y_t represents the actual value at time t, $\hat{y}_t$ is the predicted value at time t, and n denotes the total number of observations. Lewis's benchmark is commonly used to interpret MAPE values, where values below 10% indicate highly accurate forecasts, values between 10% and 20% indicate good forecasting performance, and values between 20% and 50% suggest reasonable predictive accuracy [39].

It should be noted that the forecasting models developed in this study rely exclusively on significant lag variables derived from historical rice price data. Consequently, the models do not account for exogenous factors that may influence price movements. While this univariate approach enables a focused assessment of the predictive capability of lag-based information, it may limit the models' ability to capture influences beyond the observed historical patterns.

3. Results and Discussion

3.1. Data Exploration

After data pre-processing, including the handling of missing values using linear interpolation, an exploratory data analysis was conducted to examine the characteristics of rice prices in Indonesia, including trends, variability, and price dynamics overtime.

Table 1. Descriptive statistics of rice prices (IDR/kg)

Mean	Median	Standard Deviation	Minimum	Maximum
12,308.71	11,985	1,383.67	7,850	14,595

Table 1 shows that the average rice price is IDR 12,308.71/kg, while the median value is IDR 11,985/kg. The relatively small difference between meand and media suggest that the price distribution is approximately symmetric. The standard deviation of IDR 1,383.67 indicates a moderate level of variability around the mean, reflecting price fluctuations over time. The minimum rice value of IDR 7,850/kg was recorded on March 12, 2021, which is associated with incomplete interprovincial data reporting during that period. In contrast, the maximum price of IDR 14,595/kg occurred on April 11, 2024, two days after Eid al-Fitr 1445 H, likely driven by increased demand around religious holiday. These statistics highlight the sensitivity of rice price to both data condition and seasonal demand factors.

Figure 5 illustrates the movement of rice prices in Indonesia over the study period. The prices series exhibits distinct phases, beginning with pronounced short-term fluctuation in early 2021, which are likely associated with incomplete regional data reporting during that period. From mid-2021 to mid-2022, rice prices display a relatively stable pattern, indicating balanced market conditions with limited votality. However, starting in late 2022, a persistent upward trend becomes evident, reflecting increasing price pressure over time. A notable price peak is observed in early 2024, influenced by production disruptions caused by extreme weather conditions and the El Niño phenomenon, which suppressed supply amid high demand. This increase was also driven by rising non-subsidized fertilizer prices, reduced subsidies, land conversion, and global factors such as increased rice imports by Indonesia and export restrictions by India [40]. After a temporary decline, prices rose again and reached another peak in mid-2025, before correcting in September 2025, which coincided with government intervention aimed at stabilizing rice prices [41].

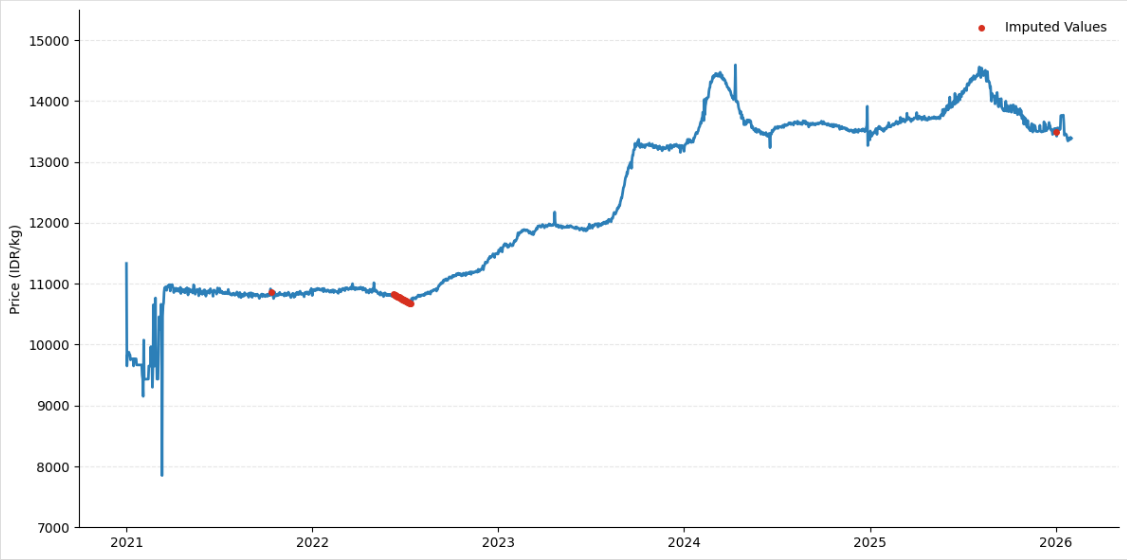

Figure 5. Daily rice price movement

3.2. Input Variable Selection

Variable selection in this study was conducted using Partial Autocorrelation Function (PACF) analysis to identify lags that have a direct influence on rice price movements. The analysis was carried out up to lag 30, corresponding to a monthly cycle, to capture short-term temporal relationships without being affected by indirect dependencies across lags. Significant lags are those whose PACF values fall outside the confidence interval, indicating a meaningful contribution in explaining rice price variation.

As illustrated in Figure 6, the presence of significant positive autocorrelation at early lags, particularly lags 1 to 4, indicates that current prices are strongly influenced by prices from recent periods. Autocorrelation values tend to decrease with increasing lag, indicating a weakening temporal dependence. At higher lags, the pattern becomes unstable, alternating between positive and negative values. Although some lags up to lag 30 remain significant, their influence is relatively small compared to early lags, thus contributing less to the model.

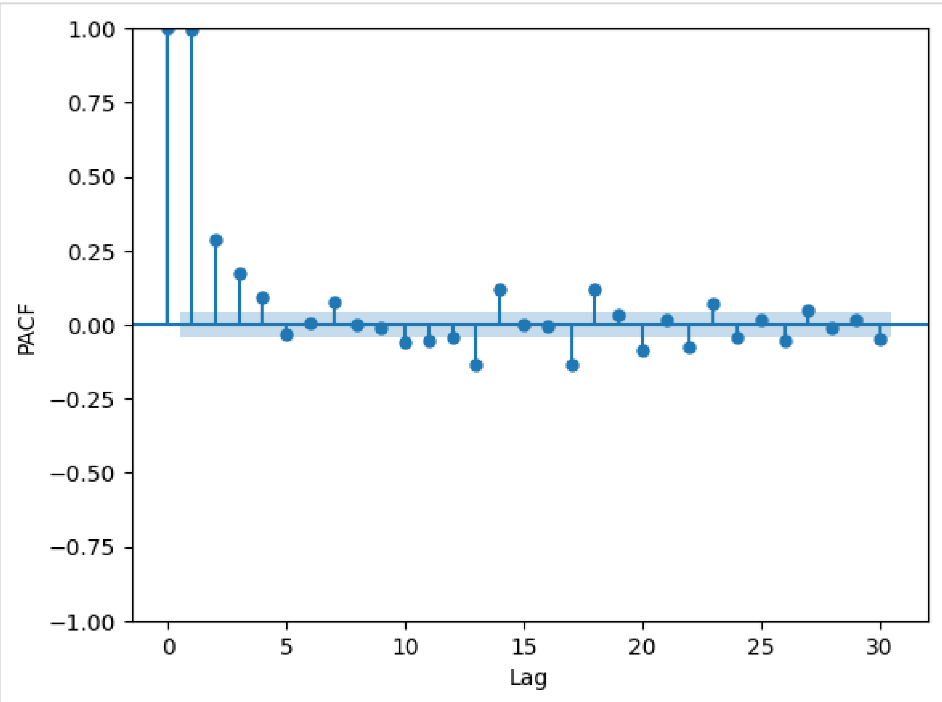

Figure 6. PACF plot of rice price data

Based on PACF analysis, 18 significant lags were identified. However, to maintain modeling efficiency, further selection was performed by choosing the 10 lags with the highest PACF values, namely lags 1, 2, 3, 4, 7, 13, 14, 17, 18, and 20. This selection aims to retain important information while reducing model complexity.

Table 2. Input variable schemes based on lag combinations

Scheme	Lag Variables
Scheme 1	1
Scheme 2	2
Scheme 3	3
…	…
Scheme 1021	1, 3, 4, 7, 13, 14, 17, 18, 20
Scheme 1022	2, 3, 4, 7, 13, 14, 17, 18, 20
Scheme 1023	1, 2, 3, 4, 7, 13, 14, 17, 18, 20

The selected lag variables were then used to form various combinations of input variables in the modeling process. With 10 lag variables, the total number of possible combinations is 1,023 schemes, derived from all possible subsets excluding the empty set ($2^{10} - 1$). Each scheme represents a different combination of lag variables used as model input. This approach aims to evaluate various input structures to obtain the most optimal variable combination without assuming specific relationships between lags.

3.3. *Model Development*

Model development was carried out using two ensemble boosting approaches, namely Gradient Boosting Machine (GBM) and Light Gradient Boosting Machine (LightGBM), by utilizing combinations of selected lag variables. Model performance was evaluated using Expanding Window Cross-Validation with five folds and the Root Mean Square Error (RMSE) metric to maintain the temporal structure of the data. The evaluation results indicate that both models achieve their best performance with lag combinations dominated by short-term lags. A summary of the ten best schemes for each model is presented in Table 3.

Table 3. Top ten schemes based on RMSE CV for GBM and LightGBM

GBM			LightGBM		
Scheme	Lag Variables	RMSE CV	Scheme	Lag Variables	RMSE CV
392	1,2,3,7,13	291.7	177	1,2,3,7	333.619
183	1,2,4,7	293.379	386	1,2,3,4,7	333.74
638	1,2,3,4,7,13	293.676	204	1,3,4,7	334.214
58	1,2,7	293.775	58	1,2,7	334.368
393	1,2,3,7,14	294.092	183	1,2,4,7	334.449
407	1,2,4,7,13	294.162	71	1,4,7	334.656
215	1,3,13,14	294.169	393	1,2,3,7,14	334.689
60	1,2,14	294.214	65	1,3,7	334.971
397	1,2,3,13,14	294.499	179	1,2,3,14	335.286
653	1,2,3,7,13,14	294.517	176	1,2,3,4	335.725

Table 3 shows that the GBM model produces lower RMSE CV values compared to LightGBM. The best-performing GBM scheme is obtained using a combination of lags 1, 2, 3, 7, and 13, achieving an RMSE CV of 291.70, while the optimal LightGBM scheme uses lags 1, 2, 3, and 7 with an RMSE CV of 333.62. The dominance of lags 1 and 2 in both models indicates that the most recent historical information plays a key role in modeling rice prices. In addition, the inclusion of lag 7 suggests the presence of a weekly periodic pattern in rice price movement.

Table 4. Hyperparameter search space for GBM and LightGBM

Model	Hyperparameter	Default Value	Candidate Range
GBM	n estimators	100	100–500 (step of 50)
	learning rate	0.1	0.005–0.2
	max depth	3	3–10
LightGBM	n estimators	100	100–500 (step of 50)
	learning rate	0.1	0.005–0.2
	max depth	-1	3–10

To improve model accuracy, hyperparameter optimization was performed using Optuna with the search space presented in Table 4. The parameter n_estimators controls the number of trees in the boosting process, learning_rate controls the contribution of each tree to the final model, and max_depth controls model complexity. The optimization process was conducted over 100 trials to obtain the parameter combination with the lowest RMSE CV.

Table 5. Optimal hyperparameter configuration and RMSE CV Results for GBM and LightGBM

Model	n estimators	learning rate	max depth	RMSE CV
GBM	350	0.1351	4	285.73
LightGBM	100	0.0953	10	334.05

These results show that hyperparameter optimization in GBM significantly improves model performance by reducing the RMSE CV compared to the initial configuration. On the other hand, optimization in LightGBM does not lead to performance improvement, indicating that its default configuration is already sufficient in capturing the main data patterns. Results presented in Table 5 indicated that the GBM model outperforms LightGBM in modeling rice prices.

3.4. Evaluation and Forecasting

Model performance evaluation was conducted to compare the performance of GBM and LightGBM under both default settings and optimized hyperparameter. The evaluation employed RMSE, MAE, and MAPE metrics on training and testing datasets, while also accounting for computational time efficiency.

Table 6. Model performance evaluation on training and testing data

Model	Training Data			Testing Data			Runtime (sec)
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	
GBM Default	42.750	22.004	0.190%	70.404	51.276	0.369%	0.393
GBM Optuna	**12.460**	**9.136**	**0.077%**	73.651	53.870	0.388%	1.751
LightGBM Default	83.524	25.845	0.233%	67.161	50.774	0.366%	0.056
LightGBM Optuna	83.978	26.492	0.239%	**66.389**	**50.213**	**0.362%**	**0.045**

As demonstrated in Table 6, the GBM Optuna model achieves the best performance on the training data, with substantially lower RMSE, MAE, and MAPE values than all other models. This significant reduction in error demonstrates that hyperparameter optimization effectively enhances the model's ability to learn and fit the underlying patterns in the training data. The improvement is particularly notable compared to the GBM default configuration, suggesting that parameter tuning plays a crucial role in maximizing the potential of the GBM algorithm. On the other hand, LightGBM shows relatively stable performance before and after optimization, with only marginal changes in evaluation metrics. This suggests that the default configuration of LightGBM is already well-suited to the data structure and does not benefit significantly from further tuning in this case. Based on Lewis's benchmark, all models can be categorized as having highly accurate forecasting performance, as all testing MAPE values are below the 10% threshold. From a computational efficiency perspective, LightGBM clearly outperforms GBM. The training time required by LightGBM is considerably shorter, even after optimization, highlighting its advantage for efficiently handling large-scale data. In contrast, optimized GBM model exhibits longer training time, reflecting the computational cost of their iterative parameter refinement and exhaustive learning strategy.

However, a different pattern emerges when evaluating the models on the test data. The LightGBM model, especially with Optuna optimization, produces the lowest RMSE, MAE, and MAPE, indicating superior generalization performance. This means that LightGBM is better at capturing the true underlying patterns in the data and applying them effectively to unseen observations. In contrast, the GBM model experiences a noticeable decline in performance when moving from training to testing data. Despite achieving very low errors during training, its higher error values on the test data suggest that the model may have overfitted the training data, learned noise or overly specific patterns that do not generalize well. This interpretation is further supported by the Time Series Cross-Validation results, which show that validation RMSE values are higher than those in training. This discrepancy highlights the importance of generalization in a univariate forecasting framework based solely on significant lag variables. Models that perform exceptionally well during training may not necessarily maintain the same level of accuracy on unseen data. The stronger performance of LightGBM with Optuna optimization on the test data suggests that it was better able to produce consistent predictions beyond the training period.

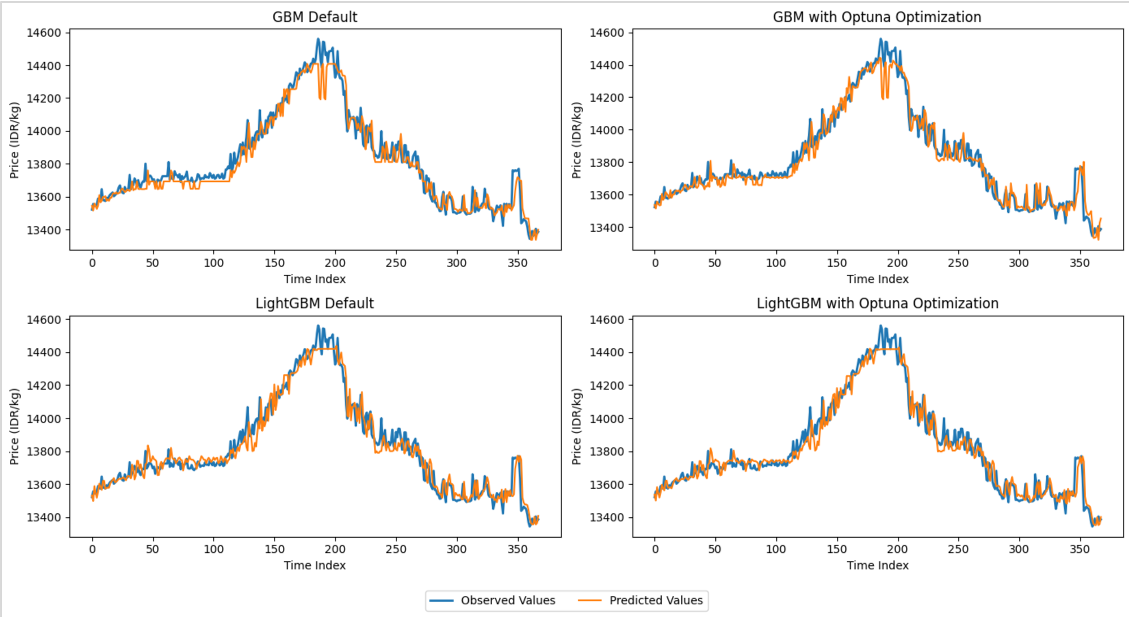

Figure 7. Comparison of GBM and LightGBM performance on testing data

The visualization in Figure 7 provides a clearer picture representation of how each model follows the movement of rice prices over time. Overall, both models are able to capture the general direction of price changes, particularly during periods characterized by gradual trend. However, differences between models become apparent during episodes of sharper price movements, such as sudden increase decreases. LightGBM tends to generate smoother and more stable over time. Its estimated values remain close to the actual data not only under normal conditions but also around turning points, including periods when prices peak and subsequently decline. This shows that LightGBM is able to adjust its predictions more consistently when the underlying price of the data shifts. Conversely, GBM shows larger gaps between predicted and actual values during certain periods. This is most visible after the price reaches a peak, where the model reacts more slowly to the downward trend. As a result, the predictions tend to lag behind actual price movements, creating wider prediction errors. This indicates that GBM faces greater difficulty in adapting to rapid changes in trend direction.

Based on the overall evaluation results, the LightGBM model with Optuna optimization is selected as the best-performing model. While GBM achieves very low error values on the training data, the decline in its performance on the testing data highlights the importance of generalization when evaluating forecasting models. LightGBM, in contrast, maintains more consistent performance between training and testing data. Moreover, it requires less computation time, making it more efficient. These factors indicate that LightGBM is more reliable for practical forecasting of rice prices.

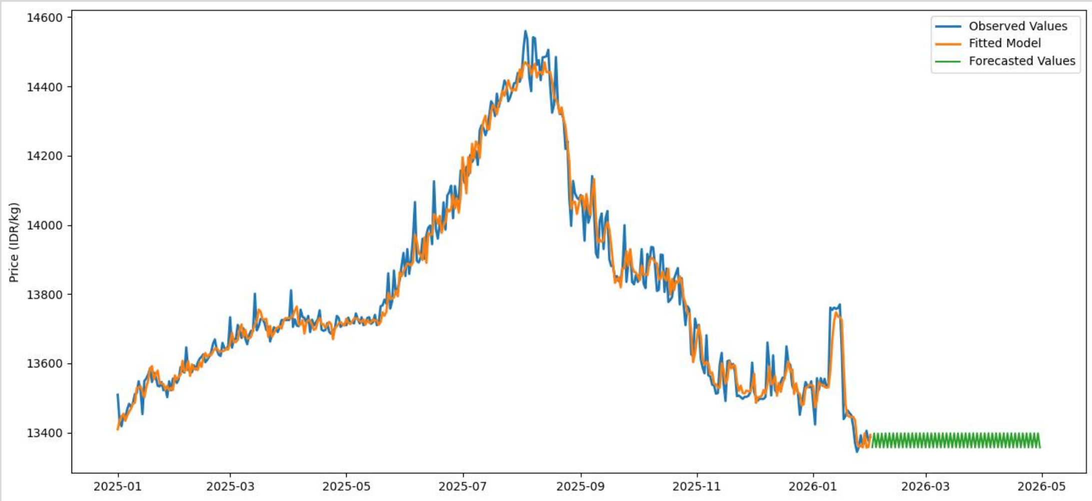

Figure 8. Model fitting and rice price forecasting results

Figure 8 illustrates the comparison between actual values, fitted values, and forecast results. The model demonstrates a strong ability to replicate historical patterns, as indicated by the close alignment between fitted values and actual observations. In the forecasting period, the predicted values show a relatively stable trend without extreme fluctuations. This suggests that, in the short term, rice prices are expected to remain relatively stable with only minor variations. The absence of sharp increases or decreases indicates that no strong signals of significant market disruption are detected based on historical patterns learned by the model.

4. Conclusion

The evaluation results show that the Gradient Boosting Machine (GBM) model achieves very low error values on the training data, particularly after hyperparameter optimization using Optuna. However, this strong in sample performance does not translate well to the testing data, indicating a tendency toward overfitting. In contrast, the Light Gradient Boosting Machine (LightGBM) model demonstrates more stable and consistent performance on out-of-sample data, reflecting superior generalization capability.

Based on the evaluation results on the testing data, the LightGBM model with hyperparameter optimization achieves the best performance, with an RMSE of 66.389, MAE of 50.213, and MAPE of 0.362%. In addition to higher accuracy, LightGBM also exhibits greater computational efficiency, with a significantly faster training time compared to the GBM model. Therefore, the LightGBM Optuna model is selected as the best model for forecasting rice commodity prices in Indonesia. Rice price forecasting using the LightGBM Optuna model for the 89-day period (February to April 2026) indicates that prices are expected to remain relatively stable in the short term, ranging from IDR 13,360/kg to IDR 13,395/kg. This suggests that over the next three months, rice prices are predicted to move within a narrow and controlled range, with no indication of extreme price changes based on the historical patterns learned by the model.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data from National Food Agency.

Credit Authorship

Muhammad Jimmy Saputra: Conceptualization, Methodology, Software, Writing, Editing, Visualization. **Yeni Rahkmawati:** Conceptualization, Reviewing, Validation, Supervision. **Selvi Annisa:** Methodology, Reviewing, Editing, Validation. **Anne Mudya Yolanda:** Reviewing, Validation, Editing.

References

[1] U.S. Department of Agriculture, "Rice - Rice Sector at a Glance," Economic Research Service. Accessed: Aug. 05, 2025. [Online]. Available: https://www.ers.usda.gov/topics/crops/rice/rice-sector-at-a-glance#Global

[2] Badan Pusat Statistik, *Ringkasan Eksekutif Pengeluaran dan Konsumsi Penduduk Indonesia Maret 2025*. Jakarta, 2025.

[3] A. Fauzi, B. Firmansyah, and D. P. Sari, "Analisis Volatilitas Harga Beras dan Segmentasi Wilayah Berbasis Data Digital Pemanfaatan data crawling dari Panel Harga Pangan Nasional," in *Seminar Nasional Official Statistics*, 2025, pp. 277–286. doi: 10.34123/semnasoffstat.v2025i1.2420.

[4] A. M. Ginting, "Strategi Kebijakan Mencegah Kenaikan Harga Beras di Indonesia," 2024.

[5] P. Subianto and G. R. Raka, "Visi, Misi dan Program Prabowo - Gibran," 2023.

[6] H. Santoso, L. Hakim, and H. Magdalena, "Sosialisasi Dampak Kenaikan Beras dengan Prediksi Kebutuhan Beras Masyarakat di Pasar Induk Cipinang dengan Kerjasama Badan Pangan Nasional," *J. Abdidas*, vol. 5, no. 2, pp. 90–96, 2024, doi: 10.31004/abdidas.v5i2.901.

[7] N. A. Muzakir and M. Z. Yahya, "Analisis Perbandingan Model Double Exponential Smoothing dan ARIMA untuk Prediksi Harga Beras di Indonesia," *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 7, no. 1, pp. 7–20, 2025, doi: 10.35580/variansiunm349.

[8] V. Arinal and M. Azhari, "Penerapan Regresi Linear untuk Prediksi Harga Beras di Indonesia," *J. Sains dan Teknol.*, vol. 5, no. 1, pp. 341–346, 2023, doi: 10.55338/saintek.v5i1.1417.

[9] E. D. B. Tarigan, M. F. Balqis, T. A. Hutapea, and D. I. Sihombing, "Peramalan Harga Beras di Indonesia Dengan ARIMA," *Sepren*, vol. 5, no. 02, pp. 117–126, 2024, doi: 10.36655/sepren.v5i02.1508.

[10] D. A. Fajari, M. F. Abyantara, and H. A. Lingga, "Peramalan Rata-Rata Harga Beras Pada Tingkat Perdagangan Besar Atau Grosir Indonesia Dengan Metode Sarima (Seasonal Arima)," *J. Agribisnis Terpadu*, vol. 14, no. 1, p. 88, 2021, doi: 10.33512/jat.v14i1.11460.

[11] B. Agan, S. Celik, O. I. Damak, and B. Miba'am, *Evaluating the machine learning-based models for predicting carbon neutrality in Sub-Saharan African Nations*, no. 0123456789. Springer Netherlands, 2025. doi: 10.1007/s10668-025-06289-y.

[12] D. Saputra, D. R. Trinadi, and D. Agustina, "Perbandingan Metode Random Forest , Linier Regression , SVM Untuk Memprediksi Harga Beras Premium," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, 2025, pp. 154–160. doi: 10.47701/yqx4ss53.

[13] M. Andriyani, S. Nurwilda, D. Zatusiva Haq, and D. Candra Rini Novitasari, "Prediksi Harga Beras Premium Tahun 2024 Menggunakan Metode Gradient Boosted Trees Regression," *J.*

Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform., vol. 18, no. 2, pp. 75–84, 2024, doi: doi.org/10.47111/JTI.
[14] I. Sabilirrasyad, N. A. Prasetyo, and M. Hermansyad, "Modeling and Predicting Indonesia Rice Prices Using Hyperparameter Optimization XGBoost," in *International Conference On Economics , Business and Information Technology*, Institut Teknologi dan Sains Mandala, 2024, pp. 459–470. doi: 10.31967/prmandala.v5i0.1226.
[15] R. De Bin and V. G. Stikbakke, "A Boosting First-Hitting-Time Model for Survival Analysis in High-Dimensional Settings," *Lifetime Data Anal.*, vol. 29, no. 2, pp. 420–440, 2023, doi: 10.1007/s10985-022-09553-9.
[16] I. B. Mustapha *et al.*, "Comparative Analysis of Gradient - Boosting Ensembles for Estimation of Compressive Strength of Quaternary Blend Concrete," *Int. J. Concr. Struct. Mater.*, 2024, doi: 10.1186/s40069-023-00653-w.
[17] G. Ke *et al.*, "LightGBM : A Highly Efficient Gradient Boosting Decision Tree," in *31st Conference on Neural Information Processing Systems*, California, 2017, pp. 1–9. doi: 10.5555/3294996.3295074.
[18] M. Kopyt, P. Piotrowski, and D. Baczyński, "Short-Term Energy Generation Forecasts at a Wind Farm—A Multi-Variant Comparison of the Effectiveness and Performance of Various Gradient-Boosted Decision Tree Models," *Energies*, vol. 17, no. 23, 2024, doi: 10.3390/en17236194.
[19] N. Zhang, Q. An, S. Zhang, and H. Ma, "Price Prediction for Fresh Agricultural Products Based on a Boosting Ensemble Algorithm," *Mathematics*, vol. 13, no. 1, 2025, doi: 10.3390/math13010071.
[20] Badan Pangan Nasional, "Panduan Implementasi Persyaratan Mutu Dan Label Beras," 2023. [Online]. Available: https://badanpangan.go.id/storage/app/media/Panduan Implementasi Persyaratan Mutu dan Label Beras_compressed.pdf
[21] M. Alam *et al.*, "Improving Rice Grain Quality Through Ecotype Breeding for Enhancing Food and Nutritional Security in Asia–Pacific Region," *Rice*, vol. 17, no. 1, 2024, doi: 10.1186/s12284-024-00725-9.
[22] H. F. Fiqa, A. R. Dewi, and R. Pandiya, "Perbandingan Metode ARIMA dan Prophet dalam Prediksi Harga Cabai Rawit di Provinsi Jawa Timur," *Pros. Semin. Nas. Sains Data*, vol. 4, no. 1, pp. 850–862, 2024, doi: 10.33005/senada.v4i1.350.
[23] M. S. Buntara, H. Napitupulu, and N. Gusriani, "Pemrograman Python Untuk Peramalan Data Deret Waktu Menggunakan Metode Seasonal Autoregressive Integrated Moving Average (Sarima)," *In Search*, vol. 22, no. 2, pp. 354–362, 2023, doi: 10.37278/insearch.v22i2.774.
[24] W. W. S. Wei, *Time Series Analysis Univariate and Multivariate Methods*. Boston: Pearson-Addison Wesley, 2006.
[25] S. Annisa, Y. Rahkmawati, and H. Hafid, "Peramalan Harga Minyak Mentah Indonesia Menggunakan Algoritma Random Forest," *J. Gaussian*, vol. 13, no. 2, pp. 472–478, 2025, doi: 10.14710/j.gauss.13.2.472-478.
[26] W. J. Niu, Z. K. Feng, S. S. Li, H. J. Wu, and J. Y. Wang, "Short-Term Electricity Load Time Series Prediction by Machine Learning Model via Feature Selection and Parameter Optimization Using Hybrid Cooperation Search Algorithm," *Environ. Res. Lett.*, vol. 16, no. 5, 2021, doi: 10.1088/1748-9326/abeeb1.
[27] M. M. Høgsgaard, K. G. Larsen, and M. E. Mathiasen, "The Many Faces of Optimal Weak-to-Strong Learning," in *38th Conference on Neural Information Processing Systems*, 2024, pp. 51885–51904. doi: 10.5555/3737916.3739560.
[28] G. Airlangga and A. Liu, "A Hybrid Gradient Boosting and Neural Network Model for Predicting Urban Happiness: Integrating Ensemble Learning with Deep Representation for Enhanced Accuracy," *Mach. Learn. Knowl. Extr.*, vol. 7, no. 1, pp. 1–23, 2025, doi: 10.3390/make7010004.
[29] E. Q. Shehab, F. F. Taha, S. H. Muhodir, H. Imran, K. A. Ostrowski, and M. Piechaczek, "Gradient Boosting Regression Tree Optimized with Slime Mould Algorithm to Predict the Higher Heating Value of Municipal Solid Waste," *Energies*, vol. 17, no. 17, 2024, doi: 10.3390/en17174213.
[30] M. Saied, S. Guirguis, and M. Madbouly, "A Comparative Study of Using Boosting-Based Machine Learning Algorithms for IoT Network Intrusion Detection," *Int. J. Comput. Intell. Syst.*, vol. 16, no. 177, 2023, doi: 10.1007/s44196-023-00355-x.
[31] M. Gan, S. Pan, Y. Chen, C. Cheng, H. Pan, and X. Zhu, "Application of the Machine Learning LightGBM Model to the Prediction of the Water Levels of the Lower Columbia River," *J. Mar. Sci. Eng.*, vol. 9, no. 5, 2021, doi: 10.3390/jmse9050496.
[32] D. Zhao, Z. Hu, Y. Yang, and Q. Chen, "Energy Conservation for Indoor Attractions Based on

NRBO-LightGBM," *Sustain.*, vol. 14, no. 19, 2022, doi: 10.3390/su141911997.

[33] J. Zhao, J. He, J. Wan, and K. Liu, "Energy Consumption Prediction for Electric Buses Based on Traction Modeling and LightGBM," *World Electr. Veh. J.*, vol. 16, no. 3, p. 159, 2025, doi: 10.3390/wevj16030159.

[34] A. K. Karakutuk, O. Ozdemir, and S. Senturk, "Optuna-Optimized Pythagorean Fuzzy Deep Neural Network : A Novel Framework for Uncertainty-Aware Image Classification," *Appl. Sci.*, vol. 15, no. 20, p. 11097, 2025, doi: 10.3390/app152011097.

[35] L. Wang, C. Zhang, W. Wang, T. Deng, T. Ma, and P. Shuai, "Slope Stability Assessment Using an Optuna-TPE-Optimized CatBoost Model," *Eng*, vol. 6, no. 8, p. 185, 2025, doi: 10.3390/eng6080185.

[36] Q. Qin and L. Li, "A VMD-Based Four-Stage Hybrid Forecasting Model with Error Correction for Complex Coal Price Series," *Mathematics*, vol. 13, no. 18, p. 2912, 2025, doi: 10.3390/math13182912.

[37] M. Alharithi, E. M. Almetwally, O. Alotaibi, M. M. Eid, E. S. M. El-kenawy, and A. A. Elnazer, "A Comparative Study of Statistical and Intelligent Classification Models for Predicting Airlines Passenger Management Satisfaction," *Alexandria Eng. J.*, vol. 119, pp. 99–110, 2025, doi: 10.1016/j.aej.2025.01.109.

[38] X. Wen, M. Jaxa-Rozen, and E. Trutnevyte, "Accuracy Indicators for Evaluating Retrospective Performance of Energy System Models," *Appl. Energy*, vol. 325, p. 119906, 2022, doi: 10.1016/j.apenergy.2022.119906.

[39] A. Fayyazbakhsh, T. Kienberger, and J. Vopava-wrienz, "Comparative Analysis of Load Profile Forecasting : LSTM , SVR , and Ensemble Approaches for Singular and Cumulative Load Categories," *Smart Cities*, vol. 8, no. 2, p. 65, 2025, doi: 10.3390/smartcities8020065.

[40] Kementerian Pertanian, "Analisis Kinerja Perdagangan Beras," 2024, [Online]. Available: https://satudata.pertanian.go.id/assets/docs/publikasi/1A_Analisis_Kinerja_Perdagangan_Beras_2024_-_publish.pdf

[41] Institute for Development of Economics and Finance, "Dinamika Pangan dan Energi: Dari Gejolak Harga Beras hingga Perbaikan Tata Kelola Impor BBM," 2025.