Forecasting Medium Rice's Retail Price with Machine Learning in Gorontalo Province

Jamal Darusalam Giu*[1], Amalan Fadil Gaib[2], Abdul Rasyid[3]
[1] Software Engineering Technology, Universitas Negeri Gorontalo, Indonesia
[2,3] Industrial Engineering, Universitas Negeri Gorontalo, Indonesia
e-mail: *[1]jamaldarusalam@ung.ac.id, [2]amalanfadil@gmail.com, [3]abdul.rasyid@ung.ac.id

Abstract

The stability of rice prices is essential for food security in Indonesia, particularly in Gorontalo Province where volatility has increased in recent years. This study develops a machine learning-based forecasting framework using Decision Tree, Random Forest, and K-Nearest Neighbors (KNN) to estimate next-day retail prices. A harvest-season indicator was incorporated to capture agricultural seasonal patterns. Data preprocessing included feature engineering, data cleaning, exploratory analysis, and chronological splitting to maintain temporal order. Model performance was assessed using RMSE and MAPE. The optimized KNN model achieved the highest accuracy, with an RMSE of 96.76 and a MAPE of 0.4%, demonstrating its strength in capturing short-term price fluctuations. The integration of seasonal indicators further improved predictive performance compared to univariate approaches, offering practical value for supporting timely policy interventions. This study is limited by its narrow feature set and the absence of external drivers such as weather conditions, production shocks, and distribution disruptions. Future research may incorporate additional exogenous variables or explore deep learning and hybrid ensemble methods to enhance robustness and generalizability.

Keywords— *Forecasting, Harvest Season, Machine Learning, Retail Rice Price, Time Series.*

1. INTRODUCTION

As a major agricultural country in the world, Indonesia has a significant level of agricultural production. This high production needs to be balanced with the ability to meet the basic food needs of the community. Among various food crops, rice occupies a strategic position because it is the main focus of farmers. This is due to its role as a source of rice, a staple food commodity whose consumption level is more dominant than other commodities such as corn, soybeans, to livestock products and vegetables. Rice is even the most consumed food by Indonesians, surpassing the consumption of sweet potatoes, eggs, milk, and vegetables [1].

Rice is a vital staple food in Indonesia, central to national food security. Rising population continues to drive demand, exerting upward pressure on prices. Globally, rice accounts for over 20% of caloric intake, with Asia producing and consuming around 90% of the world's supply [2]. In Indonesia, retail rice prices have risen over the past decade and remain highly volatile, reducing household purchasing power and frequently requiring government intervention [3]. Rice is the most significant commodity influencing regional price instability and inflation [4]. In Gorontalo, fluctuations in rice and other essential food prices play a key role in shaping local inflation and highlight the region's vulnerability to food price shocks [5]. Historically, such volatility has been linked to heightened food insecurity during global crises [6].

Rising price volatility particularly affects low-income households, making it harder for them to meet basic food needs. Understanding the drivers of this volatility requires examining the underlying supply chain structure. A commodity's supply chain consists of product, information, and cash flows, each shaping the final price paid by consumers. In the case of rice, product flow traces the movement from farmers to wholesalers, retailers, and consumers, while information flow relates to price transparency and demand forecasting. Cash flow reflects profit distribution

along the chain, and together these components influence overall price stability [7].

Gorontalo Province experienced increasingly complex rice demand dynamics between 2021 and 2024. Annual demand grew by approximately 4.2 percent, driven by 1.8 percent population growth and shifting consumption patterns. Although rice production increased by 11.3 thousand tons of GKG (4.7 percent) in 2023, this growth remained insufficient to match rising demand. In 2024, rice demand reached 32,450 tons, while local production in the January–April subround stood at only 42.20 thousand tons [8]. This persistent supply–demand imbalance places continued pressure on market stability and heightens the likelihood of price fluctuations, highlighting the need for forecasting methods capable of capturing both structural and short-term dynamics.

During 2021–2024, rice prices in Gorontalo Province exhibited a consistent upward trend, with volatility becoming more pronounced from 2023 to early 2024. Price instability intensified in 2022, highlighted by a sharp increase to IDR 14,200 per kilogram in April, driven by Ramadan related demand and higher distribution costs [9]. These fluctuations underscore the sensitivity of regional prices to seasonal and logistical pressures and the need for predictive tools to anticipate market disruptions. By January 2024, retail prices reached IDR 18,500 per kilogram, the highest nationwide, prompting concerns from local legislators about potential supply-tightening by major traders and calls for greater transparency in stock and distribution practices [10].

Previous studies on rice price forecasting in Indonesia have predominantly used time-series models, typically relying on monthly data and focusing on national or wholesale markets. Methods such as ARIMAX, SARIMAX, and Prophet have been widely applied to analyze historical trends [2], [11]. while machine learning approaches including Decision Tree, Random Forest, and K-Nearest Neighbors (KNN) have proven effective for capturing non-linear patterns [12]. Prior works demonstrate this diversity: Fajari et al. [2] applied SARIMA for national wholesale forecasts, and Adjie Setyadji et al. [13] used Neural Networks to predict prices in East Kalimantan.

Yulianti et al. [14] improved accuracy using SARIMAX with exogenous variables, though still at a low-frequency scale. Other studies, such as Ilmani et al. [15] with EEMD–ARIMA and Anggraeni et al. [16], with ANN–ARIMAX, offer methodological enhancements but remain constrained by aggregated datasets and assumptions of stable seasonality. Notably, none have examined daily retail medium rice prices in Gorontalo Province or incorporated harvest-season indicators alongside lag features to capture its short-term and highly volatile price behavior. Machine learning has been widely applied across various domains. For instance, Kawengian et al. [17] used Apriori and OCVR to optimize product layouts based on consumer purchase patterns. Although in a different context, this highlights the flexibility of data-driven models in supporting strategic decisions relevant to this study's goal of informing policy through predictive analytics.

Accordingly, this study aims to develop and evaluate machine learning models for forecasting daily medium rice retail prices in Gorontalo Province by integrating lag-based features and a harvest-season indicator. This approach addresses limitations in previous research that relied on monthly data and lacked localized contextual variables. Model performance is compared using RMSE and MAPE to identify the most accurate method for high-frequency price prediction, with the goal of providing actionable insights for policymakers and regional stakeholders to enhance market monitoring and strengthen food security strategies. The study also employs Optuna to systematically optimize model hyperparameters and ensure optimal performance.

2. RESEARCH METHODS

The study focuses on building a machine learning-based model to predict medium rice retail prices in Gorontalo Province by capturing their complex, non-linear fluctuations. In contrast to traditional time series models, machine learning methods do not require assumptions of stationarity and can flexibly learn patterns from data without prior specification of trend or seasonality. The research focuses on the application of three machine learning regression models: Decision Tree Regressor, Random Forest Regressor, and K-Nearest Neighbors Regressor (KNN).

These models were selected for their respective strengths in interpretability, ensemble learning, and locality-based prediction.

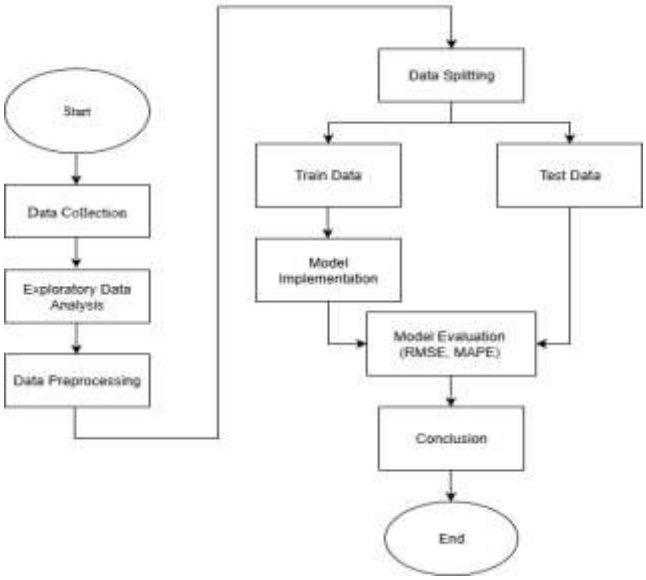

Figure 1. Research Flow

Based on the Figure 1, this research framework consists of several main phases: (1) Data Collection and Exploratory Data Analysis (EDA), (2) Data Preprocessing, (3) Model Implementation, and (4) Model Performance Evaluation using RMSE and MAPE. The overall goal is to determine which model yields the most accurate prediction results and to provide practical recommendations for rice price monitoring and policy support in the region.

2.1. Data Collection

This research makes use of a dataset that consists of daily rice retail price records from Gorontalo Province, spanning the period from March 2021 to December 2024. The primary data source is the National Food Agency's Price Panel website [18], which publishes officially verified and regularly updated price data across Indonesian provinces.

The collected dataset comprises 1,387 observations, capturing the daily fluctuations in rice retail prices over nearly four years. Each observation initially includes two attributes:

- Medium Rice Retail Price: The daily retail price of medium-quality rice (measured in Indonesian Rupiah per kilogram),
- Date: The corresponding date of observation.

In addition, this study enriches the dataset by constructing a Harvest Season Indicator, a binary variable (1 = harvest period, 0 = non-harvest period) based on Gorontalo's two main harvest seasons occurring in March-April and September-October each year. Although the Price Panel provides data on multiple rice categories (premium and medium, retail and wholesale), this study focuses exclusively on medium-quality retail prices, as this category represents the most consumed rice type and is more price-sensitive to seasonal variations.

2.2. Exploratory Data Analysis & Data Preprocessing

Exploratory Data Analysis (EDA) and preprocessing confirmed that the daily rice price dataset contained no missing values or anomalies, eliminating the need for imputation or outlier removal using techniques such as MAD or interpolation [19]. Statistical summaries, time-series plots, ACF/PACF analysis, and multiplicative decomposition were employed to examine trends,

price dynamics, and seasonal patterns [20]. The decomposition revealed consistent annual seasonality with increasing amplitude, highlighting the need for seasonality-related features [21].

Although formal stationarity tests (e.g., ADF) were not performed since the selected machine learning models do not require stationarity short-term autocorrelation was captured via a Lag-1 feature, and a binary harvest season indicator was added to reflect supply changes during peak harvest periods. Following EDA and PACF insights, the data were split chronologically into training and test sets (with a validation portion considered during development) to maintain temporal order and prevent data leakage [22][23]. These steps ensured the dataset was thoroughly validated and appropriately prepared for feature engineering and subsequent modeling.

2.3. Model Implementation

At this stage, Decision Tree, Random Forest, and KNN Regressors from scikit-learn were applied to predict next-day rice prices using Lag-1 and Harvest Season Indicator features. Models were initially trained with default settings, followed by hyperparameter optimization. Performance was evaluated on a time-respecting test set by comparing predictions with actual prices.

2.3.1. Machine Learning

Machine learning (ML) was used in this study because of its ability to capture non-linear patterns and short-term dependencies commonly found in daily commodity price movements. Unlike traditional time series models such as AR, MA, or ARIMA, ML methods do not require strict assumptions of stationarity and can flexibly learn relationships directly from the data [24][25][26]. Additionally, the chosen models are computationally lightweight, easy to reproduce, and suitable for practical deployment in resource-constrained environments. This aligns with the study's objective of producing a forecasting framework that can be applied by government agencies and local stakeholders monitoring rice price dynamics.

2.3.2. K-Nearest Neighbor

The K-Nearest Neighbors (KNN) Regressor was employed due to its ability to capture short-term local patterns in time series data without relying on strong parametric assumptions. KNN predicts future values by identifying the most similar historical observations based on feature proximity and averaging their outcomes [27][28]. This non-parametric mechanism makes KNN well-suited for datasets with complex or irregular distributions [29]. In this study, KNN was first trained using default settings to establish baseline performance, after which key parameters such as *n_neighbors*, *weights*, *algorithm*, *leaf_size*, and *p* were optimized to enhance its responsiveness to daily rice price fluctuations.

2.3.3. Decision Tree

The Decision Tree Regressor was used in this study due to its interpretability and ability to model non-linear relationships in the data. A decision tree predicts outcomes by recursively splitting the feature space into regions that minimize prediction error, ultimately forming leaf nodes that represent final predictions [30][31]. After establishing baseline performance using default configurations, key parameters such as *max_depth*, *min_samples_split*, *min_samples_leaf*, and *max_features* were optimized to balance model complexity and generalization. This enabled the model to better capture the structural variability present in daily rice price movements.

2.3.4. Random Forest

The Random Forest Regressor was employed as an ensemble extension of the Decision Tree model to improve robustness and reduce variance. Random Forest constructs multiple trees

using random subsets of both the training data and feature space, and aggregates their predictions through averaging to produce a more stable output [30]. This ensemble mechanism enables the model to capture complex interactions while reducing the risk of overfitting commonly observed in single-tree models [31]. After establishing a baseline model, key hyperparameters such as *n_estimators*, *max_depth*, *min_samples_split*, *min_samples_leaf*, *max_features*, and *bootstrap* were optimized to enhance predictive performance across varying market conditions.

2.3.5. Hyperparameter Optimization Using Optuna

Hyperparameter tuning was performed using Optuna's Tree-structured Parzen Estimator (TPE) algorithm [32], which systematically identified configurations that minimized validation RMSE more efficiently than grid or random search. The resulting optimal hyperparameters were used to retrain each model prior to final testing.

2.4. Model Evaluation

Forecast accuracy was measured using Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). RMSE penalizes larger errors more heavily, making it suitable for volatile price periods, while MAPE offers percentage-based interpretability. These widely used metrics were selected for their complementarity and relevance in food-price forecasting, enabling robust model comparison on the hold-out test set.

2.4.1. Root Mean Squared Error

Root Mean Squared Error (RMSE) is a commonly applied evaluation metric in both statistical modeling and machine learning, used to quantify the average size of errors in model predictions.It accounts for both variance and bias, making it useful for evaluating model accuracy [33]. The Root Mean Squared Error (RMSE) is calculated by taking the square root of the average of the squared differences between predicted and observed values [34]. The squaring process prevents error cancellation, while the square root restores the unit of measurement, improving interpretability [35]. Equation (1) presents the formula used to compute the Root Mean Squared Error (RMSE) as follows:

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

In the RMSE equation y_i represents the actual value, $\hat{y}_i$ indicates the predicted value**,** and n represents the total number of data points. The formula calculates the average of the squared deviations between predictions and actuals, then applies a square root to express the result in the original measurement unit, as presented in Equation (1). In the context of retail rice prices, RMSE is particularly relevant because it amplifies the impact of large deviations that may occur during supply shocks, seasonal peaks, or extreme price surges. This characteristic makes RMSE a suitable indicator for evaluating how well the models capture sudden market fluctuations, which are critical for policymakers monitoring price stability.

2.4.2. Mean Absolute Percentage Error

Mean Absolute Percentage Error (MAPE) is a commonly used metric for evaluating forecasting performance by computing the average absolute error as a percentage of actual values. A smaller MAPE value signifies better predictive accuracy [36]. Unlike absolute error metrics, MAPE standardizes the magnitude of errors by dividing them by actual observations, thereby offering a relative assessment of model performance [37]. This proves especially useful when

comparing forecasts across datasets with varying magnitudes [38]. The Equation (2) for calculating MAPE can be expressed as follows.

$$\boldsymbol{MAPE} = \mathbf{100\%} * \sum_{\mathbf{1=1}}^{\boldsymbol{n}} \left| \frac{\boldsymbol{A_t} - \mathbf{F_t}}{\boldsymbol{A_t}} \right| \quad (2)$$

where, $\boldsymbol{A_t}$ denotes the actual value at time $\boldsymbol{t}$, $\boldsymbol{F_t}$ is the forecasted value, and \$n\$ is the number of observations. Although MAPE can be undefined when actual values are zero and may produce excessively large errors for near-zero values, these limitations do not apply in this study because all observed rice prices are positive and remain within a stable numerical range. MAPE was chosen for its intuitive percentage-based interpretation, which is particularly useful for government officials and stakeholders who need easily understandable metrics to assess forecasting accuracy in rice price movements.

3. RESULT AND DISCUSSION

This section outlines the results derived from the data analysis and model development processes. It commences with an examination of dataset characteristics through exploratory data analysis (EDA), proceeds with data preprocessing steps, and continues with the construction and evaluation of multiple machine learning models aimed at forecasting retail rice prices in Gorontalo Province.

3.1. Data Collection

This study utilizes reliable daily market price data for medium-quality rice, sourced from the National Food Agency of Indonesia's official Price Panel. A manually constructed Harvest Season Indicator, a binary variable (1 or 0), was added to capture peak harvest periods, based on official agricultural calendars, Gorontalo Provincial Department of Agriculture reports, and validated news articles. This combination of time series price data and the exogenous seasonality indicator enhances the dataset, enabling supervised machine learning models to account for agricultural cycle-driven fluctuations. The dataset supports temporal learning through lag features and context-aware adjustments during seasonal periods, improving predictive modeling accuracy.

Table 1 Sample Dataset of Daily Rice's Retail Prices

Date (YYYY-MM-DD)	Rice Retail Price in Rupiah	Harvest Season Indicator
2021-03-10	10,500	1
2021-03-11	10,500	1
2021-03-12	7,000	1
2021-03-13	10,500	1
2021-03-14	10,560	1
…	…	…
2024-12-21	12870	0
2024-12-22	12850	0
2024-12-23	12840	0
2024-12-24	12980	0
2024-12-25	13130	0

Table 1 illustrates the multi-year rice price series and the role of the harvest season indicator. In March 2021 (harvest season = 1), prices fluctuated between IDR 7,000 and IDR 10,560, with an anomalously low value of IDR 7,000 on 12 March 2021 retained from the original National Food Agency records to maintain authenticity. By contrast, December 2024 (harvest season = 0) recorded markedly higher prices between IDR 12,840 and IDR 13,130, reflecting the long-term upward trend and typical scarcity outside harvest periods. This clear contrast between harvest and non-harvest phases highlights the substantial influence of seasonal agricultural cycles on retail prices. The dataset's inclusion of both low-volatility harvest periods and high-price non-harvest periods, along with occasional anomalies and shocks, provides a realistic and diverse training environment, enhancing the robustness and generalizability of the forecasting models.

3.2. Exploratory Data Analysis

In order to explore the dataset's statistical properties and detect hidden patterns, an EDA process was implemented. This step is essential to identify statistical characteristics, trends, and potential anomalies before proceeding to model development.

Table 2. Descriptive Statistics of Daily Rice's Retail Prices Dataset

Statistic	Rice Retail Price in Rupiah	Harvest Season
Count	1,387	1,387
Mean	11,750.86	0.345
Standard Deviation	1,589.48	0.476
Minimum	7,000	0
25^{th} Percentile	10,270	0
Median	11,250	0
75^{th} Percentile	13,070	1
Maximum	16,390	1

The descriptive statistics, presented in Table 2, indicate that the average rice price during the observed period was approximately IDR 11,750 per kilogram, with the lowest observed price recorded at IDR 7,000 and the highest reaching IDR 16,390. The relatively high standard deviation of IDR 1,589.48 suggests substantial variability in daily prices. Additionally, the median price of IDR 11,250 slightly below the mean indicating a slight positive skew in the price distribution. The harvest season indicator shows that approximately 34.5% of the data points fall within major harvest periods, with most low-price observations concentrated in these intervals.

Figure 2 identifies two significant structural breaks in Indonesian rice prices from 2021–2024. The first occurred at the beginning of 2023, with prices jumping from IDR 10,956 in December 2022 to IDR 11,235 in January 2023 (+2.5% MoM) and rising 19.3% year-on-year (from an average of IDR 10,364 in 2022 to IDR 12,360 in 2023). A much sharper break followed in early 2024, when prices surged 23.2% from IDR 12,980 in January to IDR 16,002 in March, triggered by intense El Niño impacts, production shortfalls, and supply-chain disruptions. Prices then declined steadily to IDR 13,265 by June 2024 due to government interventions and the onset of the harvest season. These pronounced non-linear shifts highlight the importance of incorporating external shocks and regime changes in rice price modeling.

Figure 2 Retail Rice Price Trend (March 2021 – December 2024)

To better understand the drivers of price movements, a seasonal decomposition was applied, separating the time series into trend, seasonal, and residual components. This approach provides clearer insights into long-term direction, recurring cyclical patterns, and irregular fluctuations.

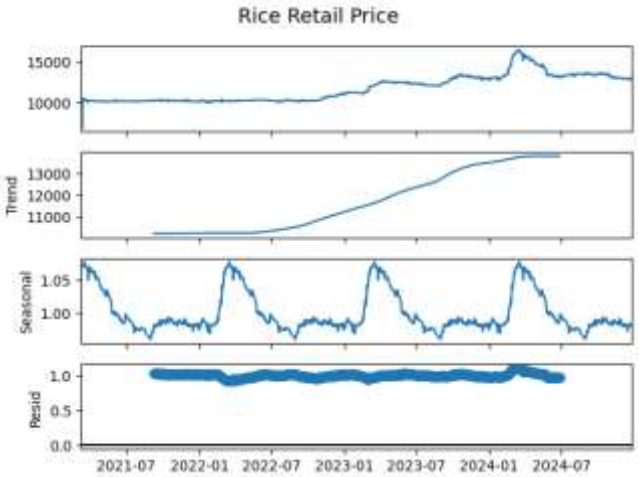

Figure 3 Seasonal Decomposition of Rice Retail Prices Dataset

Figure 3's seasonal decomposition shows a strong annual cycle in rice prices with an amplitude of ~IDR 2,700 (peak: IDR 16,002 in March 2024; trough: IDR 13,265 in June 2024), representing approximately 17% of the average price. This confirms that a large share of price variability stems from predictable, harvest-related seasonal patterns rather than random shocks. The clear seasonality explains why the K-Nearest Neighbors model outperformed others: it effectively captures recurring local patterns via historical similarity, while the stricter splitting rules of Decision Tree and Random Forest models suppress seasonal amplitude. ACF and PACF plots were also examined to identify temporal dependencies and guide lag selection, offering additional insight into the short-term dynamics of the price series.

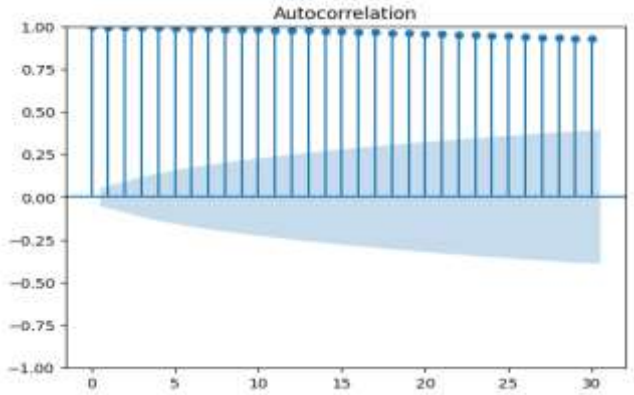

Figure 4. Autocorrelation Function (ACF) of Rice Prices Dataset

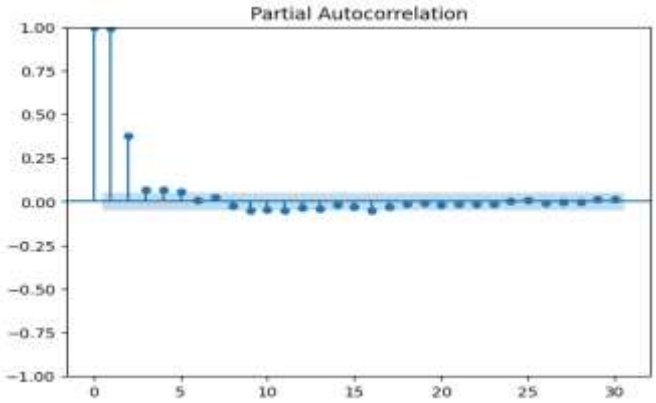

Figure 5. Partial Autocorrelation Function (PACF) of Rice Prices Dataset

Although the PACF plot in the Figure 5 indicates that both lag-1 and lag-2 exhibit significant partial autocorrelations, the very high persistence observed in the ACF suggests that lag-1 already carries the dominant share of temporal information in the series. As the autocorrelation at lag-1 is nearly perfect and the decay across subsequent lags remains extremely slow, incorporating only the immediate past value is sufficient to represent the underlying price dynamics. During model experimentation, lag-1 also provided the most stable and accurate predictions, while the addition of lag-2 or higher lags introduced redundant information without improving error metrics. This reinforces the decision to prioritize lag-1 as the primary temporal feature for machine learning models, despite the AR(2) signature suggested by the PACF.

To explore the linear relationships between variables in the dataset, Pearson correlation analysis was performed to assess the strength and direction of the linear relationship between medium rice retail prices and the harvest season indicator.

Table 3. The Pearson's Correlation Matrix of Rice's Retail Prices Dataset

Variable	Retail Rice Price	Harvest Season
Retail Rice Price	1.000	0.092
Harvest Season	0.092	1.000

Although the harvest season indicator exhibited a weak Pearson correlation (r = 0.092) with retail rice prices, time-series decomposition (Figure 3) confirmed large recurring seasonal declines. Such non-linear seasonal effects are well-suited to machine learning models, justifying the retention of the indicator as an informative exogenous feature despite its low linear correlation.

3.3. Data Preprocessing

Following exploratory data analysis, the daily rice price series was preprocessed to enable supervised machine learning forecasting. The series displayed strong first-order temporal dependence, with ACF showing near-unity autocorrelation at lag-1 and slow decay, while PACF confirmed the dominance of lag-1. Accordingly, only the lag-1 price was retained as the primary temporal feature, as machine learning models can flexibly capture additional patterns without requiring multiple consecutive lags.

The forecasting task was framed as a regression problem by shifting the original price series forward by one day to create the target variable. A binary exogenous Harvest Season Indicator was added to account for the non-linear seasonal patterns observed in the decomposition analysis. The final feature set thus comprised two variables: Price(t-1) and Harvest Season Indicator.

The steps of data preprocessing are outlined as follows:

- Lag Feature Engineering: A lag-1 feature was generated by shifting the rice price values by one time step backward.
- Target Variable Construction: The prediction target was established by shifting the rice price series one time step forward.
- Feature Selection: The final feature set consisted of Lag1 and Harvest Season Indicator.
- Chronological Train-Test Split:
    - Training set: March 10, 2021 – April 30, 2024.
    - Testing set: May 1, 2024 – December 25, 2024.

The chronological split ensures that the models are trained only on past observations to predict future prices, maintaining the causality structure necessary for time series forecasting. A sample overview of the preprocessed data is shown in Figure 6.

Tanggal	Medium Rice Retail Price	Harvest Season	lag1	target
2021-03-11	10500	1	10500.0	10500
2021-03-12	7000	1	10500.0	7000
2021-03-13	10500	1	7000.0	10500
2021-03-14	10560	1	10500.0	10560
2021-03-15	10380	1	10560.0	10380

Figure 6. Preprocessed Data in Python 3.10

The train–test split was deliberately set with the boundary at the end of April 2024 to ensure the training period included the sharp rice price surge observed in early 2024. This approach allowed models to learn both normal and high volatility price behaviors, yielding more robust and representative performance. No separate validation set was used; instead, a strict time-based train–test framework without shuffling was adopted to prevent data leakage and maintain true temporal separation. Consequently, the test set covering the period after April 2024 remained completely unseen during model development and hyperparameter optimization, providing a realistic out-of-sample evaluation of forecasting accuracy.

3.4. Model Implementation & Evaluation

After completing the data preprocessing steps, three machine learning regression models were implemented to forecast retail rice prices, namely the Decision Tree Regressor, Random Forest Regressor, and K-Nearest Neighbors (KNN) Regressor. All models were developed in Python using the scikit-learn library. The training phase utilized two predictor variables, specifically the Lag-1 price and the Harvest Season Indicator, which together capture short-term price dynamics and seasonal effects.

To enhance predictive performance, each model underwent hyperparameter optimization using Optuna. The objective function minimized the Root Mean Squared Error (RMSE), enabling the optimization process to identify parameter sets that yielded the most accurate and stable results. Table 4 summarizes the initial default parameters of the models we used in this research.

Table 4. Configuration of Machine Learning Models

Model	Decision Tree Regressor	Random Forest Regressor	K-Nearest Neighbors
Default Parameters	max_depth=None min_samples_split=2 min_samples_leaf=1 max_features=None random_state=42	n_estimators=100, max_depth=None min_samples_split=2 min_samples_leaf=1 max_features=1.0 bootstrap=True random_state=42	n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2

Each model was trained to predict the rice price for the next day based on the historical patterns captured by the lagged feature and the seasonal context provided by the harvest season indicator. These models were then evaluated using the testing dataset, the results of which are discussed in the subsequent section.

The predictive performance of the three base parameter machine learning models was assessed using two widely adopted forecasting evaluation metrics: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). The evaluation results are presented in Table 5.

Table 5. Evaluation results of base models

Lags	Decision Tree		Random Forest		KNN	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
1	169.45	0.95	122.59	0.66	104.57	0.53
2	192.94	0.95	126.07	0.61	115.07	0.54
3	262.71	1.38	161.02	0.83	144.77	0.64
4	224.68	1.17	187.47	0.95	163.18	0.72
5	264.43	1.22	213.48	1.09	187.20	0.82

6	216.83	1.12	255.58	1.31	199.91	0.82
7	252.70	1.18	258.53	1.32	218.19	0.88

Table 5 presents the baseline performance of all three machine learning models across seven different lag configurations prior to any hyperparameter tuning. The results indicate that lower lag values consistently yield better predictive accuracy, with Lag-1 producing the lowest RMSE and MAPE across all models. Among the baseline models, KNN with Lag-1 achieved the best overall performance, followed by Random Forest and Decision Tree. Increasing the lag order generally led to higher prediction errors, suggesting diminishing contributions from more distant historical observations in this dataset.

Given that Lag-1 demonstrated the most stable and accurate performance, this configuration was selected for the subsequent hyperparameter optimization stage. The Table 6 below summarizes the model results after tuning using Optuna.

Table 6. Configuration of Optimized Machine Learning Models

Model	Decision Tree	Random Forest Regressor	K-Nearest Neighbors
Search Space	max_depth=range(2 – 101), min_samples_split=range(2-100), min_samples_leaf= range(1 – 50), max_features= range(0,1 – 1,0)	n_estimators=range(10 – 10000), max_depth=range(2 – 101), min_samples_split=range(2 – 100), min_samples_leaf=range(1 – 50), max_features=range(0,1 – 1,0) bootstrap=boolean(True, False)	n_neighbors=range(1 – 32), weights=categorical(uniform, distance), algorithm=categorical(auto, ball_tree, kd_tree, brute), p=range(1 – 5)
Optimized Hyperparameters	max_depth=71, min_samples_split=3, min_samples_leaf=2, max_features=0.999 random_state=42	n_estimators=8125, max_depth=39, min_samples_split=3, min_samples_leaf=3, max_features=0.105, bootstrap=True random_state=42	n_neighbors=15, weights='uniform', algorithm='brute', leaf_size=44, p=1

The hyperparameter optimization process was performed using Optuna, where each model was evaluated over a predefined search space as summarized in Table 6. The search space was designed to cover a wide range of parameter values, enabling Optuna to explore both shallow and deep trees for Decision Tree and Random Forest, as well as various neighborhood sizes and distance metrics for KNN. Through iterative trials and RMSE-based objective evaluation, Optuna identified the optimal parameter configurations for each model.

Table 7 reveals that, even after hyperparameter optimization, K-Nearest Neighbors (KNN) delivers the highest forecasting accuracy. This result aligns with the nature of daily rice price movements, which exhibit gradual, short-term patterns that KNN excels at capturing through similarity with neighboring observations.

Table 7. Evaluation results of optimized models

Model	RMSE (IDR)	MAPE (%)
K-Nearest Neighbors	96.76	0.47
Random Forest Regressor	135.99	0.74
Decision Tree Regressor	180.75	0.87

In contrast, both Random Forest and Decision Tree models showed reduced performance post-tuning. With only Lag-1 and a binary harvest season indicator as features, tree-based models lack sufficient richness for meaningful splitting, leading to overfitting of noise rather than genuine patterns. Consequently, in this low-dimensional, locally dependent setting, instance-based methods like KNN clearly outperform hierarchical ensemble approaches.

4. CONCLUSION

This study addresses the research gap in machine-learning-based retail rice price forecasting at the provincial level in Indonesia, particularly the lack of seasonal indicators in prior models. Focusing on Gorontalo Province, the authors incorporate lag features and an exogenous harvest-season indicator to improve upon traditional univariate approaches that ignore agricultural seasonality.The results demonstrate that the K-Nearest Neighbors (KNN) model outperformed other methods, achieving the lowest RMSE of 96.76 and MAPE of 0.47%. These metrics highlight KNN's effectiveness in capturing short-term dependencies and rapid price fluctuations through local-learning techniques. The inclusion of a domain-specific seasonal indicator significantly enhanced forecast reliability across harvest and non-harvest periods.

Overall, the study proposes a practical provincial-level forecasting framework that combines machine learning with agricultural seasonality, offering direct benefits for supply-chain management, market interventions, and volatility monitoring. The authors suggest future extensions involving additional exogenous variables (e.g., weather, production costs, policy changes), hybrid/ensemble methods, deep learning models, or cross-regional comparisons to further improve robustness and generalizability of rice price forecasting in Indonesia.

5. ACKNOWLEDGMENTS

The authors would like to express their gratitude to the National Food Agency of Indonesia for providing access to the rice price dataset used in this study. Special thanks are also extended to the Department of Industrial Engineering, Universitas Negeri Gorontalo, for the academic and technical support throughout the research process.

REFERENCES

[1] F. Abdullah, S. Imran, and A. Rauf, "Analisis Ketersediaan Beras Di Kabupaten Gorontalo Selang Tahun 2021-2030," *AGRINESIA J. Ilm. Agribisnis*, vol. 6, no. 3, pp. 187–197, Aug. 2022, doi: 10.37046/agr.v6i3.16138.

[2] 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.

[3] A. Nurina Aulia, D. Indra Wardhana, M. Afifutoyyiba, T. Putra, and U. J. Muhammadiyah Jember Karimata, "Hubungan Volatilitas Harga Konsumsi Beras Terhadap Ketahanan Pangan di Jawa Timur," *Mimb. Agribisnis J. Pemikir. Masy. Ilm. Berwawasan Agribisnis*, vol. 11, no. 2, pp. 2612–2620, Jul. 2025, doi: 10.25157/MA.V11I2.18373.

[4] C. Faradilla, E. Marsudi, and A. Baihaqi, "Analisis Statistik Ketahanan Pangan Terhadap Perubahan Harga Komoditas Pangan Strategis Di Indonesia," *J. Agrisep*, vol. 22, no. 1, pp. 53–62, Jul. 2021, doi: 10.17969/agrisep.v22i1.21497.

[5] E. Podungge, F. Z. Olilingo, and I. R. Santoso, "Pengaruh Harga Komoditas Sembilan Bahan Pokok Terhadap Tingkat Inflasi di Provinsi Gorontalo Tahun 2020--2024," *Ekopedia J. Ilm. Ekon.*, vol. 1, no. 2, pp. 275–285, Jun. 2025, doi: 10.63822/qv4tdx45.

[6] A. W. Putra, J. Supriatna, R. H. Koestoer, and T. E. B. Soesilo, "Differences in local rice price volatility, climate, and macroeconomic determinants in the indonesian market," *Sustain.*, vol. 13, no. 8, 2021, doi: 10.3390/su13084465.

[7] A. Rasyid, "Value Chain Analysis of Corn Commodity Supply Chain," *Budapest Int. Res. Critics Institute-Journal*, vol. 4, no. 4, pp. 9715–9726, Nov. 2021, doi: 10.33258/BIRCI.V4I4.2994.

[8] R. Indriani, S. Imran, and M. Mukhlis, "Struktur dan Efisiensi Kinerja Rantai Pasok Beras di Provinsi Gorontalo, Indonesia," *Agro Bali Agric. J.*, vol. 7, no. 2, pp. 542–558, 2024, doi: 10.37637/ab.v7i2.1648.

[9] S. Lahay, "Produksi Padi dan Beras di Gorontalo Naik, Kenapa Harganya Mahal?," *Benua Indonesia*, 2024. https://benua.id/produksi-padi-dan-beras-di-gorontalo-naik-kenapa-harganya-mahal/ (accessed Jun. 28, 2025).

[10] Arfandi Ibrahim, "Harga Beras Gorontalo Termahal Se-Indonesia, DPRD Soroti Langkah Pemerintah," *LIPUTAN6*, 2024. https://www.liputan6.com/regional/read/5540763/produksi-padi-dan-beras-di-gorontalo-naik-kenapa-harganya-mahal (accessed Jun. 29, 2025).

[11] C. Chandra and S. Budi, "Analisis Komparatif ARIMA dan Prophet dengan Studi Kasus Dataset Pendaftaran Mahasiswa Baru," *J. Tek. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 278–287, 2020, doi: 10.28932/jutisi.v6i2.2676.

[12] E. Mardiani *et al.*, "Komparasi Metode Knn, Naive Bayes, Decision Tree, Ensemble, Linear Regression Terhadap Analisis Performa Pelajar Sma," *Innov. J. Soc. Sci. Res.*, vol. 3, no. 2, pp. 13880–13892, Jun. 2023, Accessed: Nov. 20, 2025. [Online]. Available: https://j-innovative.org/index.php/Innovative/article/view/1949

[13] M. Adjie Setyadji, A. Faqih, and Y. Arie Wijaya, "Peramalan Harga Komoditas Beras Di Kalimantan Timur Menggunakan Algoritma Neural Network," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 320–324, 2023, doi: 10.36040/jati.v7i1.6327.

[14] R. Yulianti, N. T. Amanda, K. A. Notodiputro, Y. Angraini, and L. N. A. Mualifah, "Comparison Of SARIMA And SARIMAX Methods For Forecasting Harvested Dry Grain Prices In Indonesia," *Barekeng*, vol. 19, no. 1, pp. 319–330, Jan. 2025, doi: 10.30598/barekengvol19iss1pp319-330.

[15] E. A. Ilmani, I. M. Sumertajaya, and A. Fitrianto, "Rice Price Forecasting for All

Provinces in Indonesia Using The Time Series Clustering Approach and Ensemble Empirical Mode Decomposition," *Sci. J. Informatics*, vol. 12, no. 1, pp. 119–132, May 2025, doi: 10.15294/SJI.V12I1.23536.

[16] W. Anggraeni, F. Mahananto, A. Q. Sari, Z. Zaini, K. B. Andri, and Sumaryanto, "Forecasting the Price of Indonesia's Rice Using Hybrid Artificial Neural Network and Autoregressive Integrated Moving Average (Hybrid NNs-ARIMAX) with Exogenous Variables," *Procedia Comput. Sci.*, vol. 161, pp. 677–686, Jan. 2019, doi: 10.1016/J.PROCS.2019.11.171.

[17] A. R. Kawengian, I. H. Lahay, and J. D. Giu, "Redesain Tata Letak Produk Berdasarkan Market Basket Analysis," *JISI J. Integr. Sist. Ind.*, vol. 12, no. 1, pp. 49–58, Feb. 2025, doi: 10.24853/JISI.12.1.49-58.

[18] Badan Pangan Nasional, "Panel Harga Pangan: Harga Beras Medium," *Direktorat, Stabilisasi Pasokan dan Harga Pangan Kedeputian Bidang Ketersediaan Pangan dan Stabilisasi Pangan*, 2025. https://panelharga.badanpangan.go.id/beranda (accessed Apr. 12, 2025).

[19] S. Meisenbacher *et al.*, "Review of automated time series forecasting pipelines," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 12, no. 6, 2022, doi: 10.1002/widm.1475.

[20] K. Yemets, I. Izonin, and I. Dronyuk, "Time Series Forecasting Model Based on the Adapted Transformer Neural Network and FFT-Based Features Extraction," *Sensors*, vol. 25, no. 3, pp. 1–19, 2025, doi: 10.3390/s25030652.

[21] M. Usmani, Z. A. Memon, A. Zulfiqar, and R. Qureshi, "Preptimize: Automation of Time Series Data Preprocessing and Forecasting," *Algorithms*, vol. 17, no. 8, 2024, doi: 10.3390/a17080332.

[22] R. P. Masini, M. C. Medeiros, and E. F. Mendes, "Machine learning advances for time series forecasting," *J. Econ. Surv.*, vol. 37, no. 1, pp. 76–111, 2023, doi: 10.1111/joes.12429.

[23] Abhishek Gautam, Aditya Prakash, and Gariyas Kaushal, "Artificial Intelligence in Cybersecurity," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 607–610, 2024, doi: 10.48175/ijarsct-17681.

[24] R. N. Ramadhon, A. Ogi, A. P. Agung, R. Putra, S. S. Febrihartina, and U. Firdaus, "Implementasi Algoritma Decision Tree untuk Klasifikasi Pelanggan Aktif atau Tidak Aktif pada Data Bank," *Karimah Tauhid*, vol. 3, no. 2, pp. 1860–1874, 2024, doi: 10.30997/karimahtauhid.v3i2.11952.

[25] S. Elsayed, D. Thyssens, A. Rashed, H. Samer Jomaa, and L. Schmidt-Thieme, "Do We Really Need Deep Learning Models for Time Series Forecasting?," *arXiv e-prints*, p. arXiv:2101.02118, Jan. 2021, doi: 10.48550/arXiv.2101.02118.

[26] G. Skenderi, C. Joppi, M. Denitto, and M. Cristani, "On the Use of Learning-Based Forecasting Methods for Ameliorating Fashion Business Processes: A Position Paper," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2023, vol. 13644 LNCS, pp. 647–659. doi: 10.1007/978-3-031-37742-6_50.

[27] V. Saha, "Predicting Future Cryptocurrency Prices Using Machine Learning Algorithms," *J. Data Anal. Inf. Process.*, vol. 11, no. 04, pp. 400–419, 2023, doi: 10.4236/jdaip.2023.114021.

[28] A. D. P. Pacheco, J. A. D. S. Junior, A. M. Ruiz-Armenteros, and R. F. F. Henriques, "Assessment of k-nearest neighbor and random forest classifiers for mapping forest fire areas in central portugal using landsat-8, sentinel-2, and terra imagery," *Remote Sens.*, vol. 13, no. 7, pp. 1–25, 2021, doi: 10.3390/rs13071345.

[29] E. Ozturk Kiyak, B. Ghasemkhani, and D. Birant, "High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis," *Electronics*, vol. 12, no. 18, 2023, doi: 10.3390/electronics12183828.

[30] J. Singh Kushwah, A. Kumar, S. Patel, R. Soni, A. Gawande, and S. Gupta, "Comparative study of regressor and classifier with decision tree using modern tools," *Mater. Today Proc.*, vol. 56, pp. 3571–3576, 2022, doi: 10.1016/j.matpr.2021.11.635.

[31] E. Pekel, "Estimation of soil moisture using decision tree regression," *Theor. Appl. Climatol.*, vol. 139, no. 3–4, pp. 1111–1119, 2020, doi: 10.1007/s00704-019-03048-8.

[32] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, and P. Networks, "Optuna : A Next - Generation Hyperparameter Optimization Framework," pp. 1–10, 2019.

[33] H. Sharma, H. Harsora, and B. Ogunleye, "An Optimal House Price Prediction Algorithm: XGBoost," *Analytics*, vol. 3, no. 1, pp. 30–45, 2024, doi: 10.3390/analytics3010003.

[34] K. Sako, B. N. Mpinda, and P. C. Rodrigues, "Neural Networks for Financial Time Series Forecasting," *Entropy*, vol. 24, no. 5, 2022, doi: 10.3390/e24050657.

[35] Ł. Ledziński and G. Grześk, "Artificial Intelligence Technologies in Cardiology," *J. Cardiovasc. Dev. Dis.*, vol. 10, no. 5, 2023, doi: 10.3390/jcdd10050202.

[36] S. I. N. Suwandi, Raras Tyasnurita, and Hanifan Muhayat, "Peramalan Emisi Karbon Menggunakan Metode SARIMA dan LSTM," *J. Comput. Sci. Informatics Eng.*, vol. 6, no. 1, pp. 73–80, 2022, doi: 10.29303/jcosine.v6i1.436.

[37] M. H. L. Lee *et al.*, "A Comparative Study of Forecasting Electricity Consumption Using Machine Learning Models," *Mathematics*, vol. 10, no. 8, 2022, doi: 10.3390/math10081329.

[38] L. Dague, N. Badaracco, T. DeLeire, J. Sydnor, A. S. Tilhou, and D. Friedsam, "Trends in Medicaid Enrollment and Disenrollment During the Early Phase of the COVID-19 Pandemic in Wisconsin," *JAMA Heal. Forum*, vol. 3, no. 2, pp. e214752–e214752, Feb. 2022, doi: 10.1001/jamahealthforum.2021.4752.