ICEBIT
The 5th International Conference On Economics , Business and Information Technology
E-ISSN 3025-4892
August 24, 2024 pp. 459-470

MODELING AND PREDICTING INDONESIA RICE PRICES USING HYPERPARAMETER OPTIMIZATION XGBOOST

Iqbal Sabilirrasyad
Institut Teknologi dan Sains Mandala
Jember, Indonesia
5171243269, +628
iqbal@itsm.ac.id

Nur Andita Prasetyo
Institut Teknologi dan Sains Mandala
Jember, Indonesia
2329201769, +628
nurandita.prasetyo69@itsm.ac.id

Mas'ud Hermansyad
Institut Teknologi dan Sains Mandala
Jember, Indonesia
1330466463. +628
masudhermansyah@gmail.com

ABSTRACT

This study explores the application of the XGBoost model, fine-tuned through hyperparameter optimization and cross-validation, for forecasting rice prices in Indonesia for the two years following July 2024. Utilizing a comprehensive dataset spanning from January 2010 to July 2024, the research emphasizes the importance of detailed feature engineering, including temporal and cyclical patterns, to enhance the model's predictive accuracy. A 4-fold cross-validation approach was employed, resulting in a rigorous evaluation process that involved over 26,000 model fits. The study highlights the effectiveness of XGBoost in capturing complex patterns within the data, yielding highly accurate predictions. Additionally, the integration of external factors—such as climate conditions, government policies, global market trends, economic indicators, and technological advancements—is recommended to further refine the model and ensure adaptability to real-world conditions. The findings suggest that this approach provides valuable insights for stakeholders, including policymakers and market analysts, facilitating informed decision-making regarding production, pricing strategies, and food security. This research underscores the potential of advanced machine learning techniques in improving time series forecasting within dynamic and complex markets like rice pricing.

Keywords : XGBoost, Rice Price, Forecasting

1. INTRODUCTION

Rice is the cornerstone of Indonesia's food security, serving as a staple for the vast majority of the population. The cultural, social, and economic significance of rice in Indonesia cannot be overstated. It is more than just a dietary preference; it is an integral part of the nation's identity and a symbol of sustenance. The phrase "belum makan kalau belum makan nasi" (which means "you haven't eaten unless you've eaten rice") is a common expression that highlights the centrality of rice in daily life. This cultural attachment underscores the critical importance of rice availability and affordability, which are deeply intertwined with the country's food resilience.

Rice occupies a unique position in Indonesia, not only as the primary source of calories for the population but also as a major economic commodity. Indonesia is the world's fourth-most populous country, and with over 270 million people, the demand for rice is immense. On average, Indonesians consume around 100-120 kilograms of rice per person annually, making it one of the highest per capita rice consumption rates globally. This heavy reliance on rice as a staple means that any fluctuation in its availability or price can have a widespread impact, influencing not just the cost of living but also social stability and economic growth. In rural areas, where a significant portion of the population is engaged in agriculture, rice farming is a primary livelihood. Approximately 29% of Indonesia's labor force is employed in agriculture, with a substantial number involved in rice cultivation. Thus, rice farming is not just a means of sustenance but also a vital economic activity that supports millions of households. The interconnectedness of rice with the broader economy makes its price volatility a crucial concern for the government, farmers, consumers, and other stakeholders.

Food resilience refers to the capacity of a food system to endure, adapt, and recover from various shocks and stresses, whether they are environmental, economic, or social in nature. In Indonesia, a country prone to natural disasters such as earthquakes, volcanic eruptions, and floods, food resilience is particularly critical. The ability of the agricultural system to bounce back from these disruptions is essential for maintaining food security and ensuring that all citizens have access to adequate nutrition. The concept of food resilience also encompasses the sustainability of food production practices. Sustainable agricultural practices are vital for ensuring that rice can continue to be produced in sufficient quantities without degrading the environment. This includes managing water resources effectively, as rice is a water-intensive crop, and addressing the challenges posed by climate change, which is expected to alter rainfall patterns and increase the frequency of extreme weather events in Indonesia. In addition to environmental factors, food resilience is also influenced by economic policies, infrastructure, and the efficiency of supply chains. For instance, the ability to store, transport, and distribute rice efficiently across the archipelago is crucial for ensuring that even the most remote regions have access to affordable rice. This aspect of food resilience becomes even more important during times of crisis, such as during the COVID-19 pandemic, when supply chains were disrupted, highlighting the vulnerabilities in the food system.

Given the significant impact that rice prices have on both food security and economic stability in Indonesia, the ability to accurately predict future rice prices is of paramount importance. The ability to predict not only the price of rice but also various aspects of it, such as supply chain dynamics, market demand, and regional production trends,

can facilitate the generation of insights into prospective arrangements and governance structures. This foresight enables stakeholders to make informed decisions that could enhance market stability, ensure food security, and optimize resource allocation. Additionally, researchers have conducted forecasting and prediction research in other fields, such as climate science, healthcare, and finance, where advanced predictive models have been instrumental in anticipating challenges and opportunities, thereby improving decision-making processes and policy formulation (Azadi et al., 2023; Fraher et al., 2024; Subhra et al., 2023). Predicting rice prices can help policymakers, farmers, and other stakeholders make informed decisions to mitigate the adverse effects of price fluctuations. For instance, accurate forecasts can enable the government to implement timely interventions, such as adjusting import quotas or releasing stocks from the national reserve, to stabilize prices and ensure food security. For farmers, price predictions can guide planting decisions and financial planning, allowing them to maximize their income while minimizing risks. In the broader economy, accurate price forecasts can help businesses in the food industry manage their supply chains more effectively, reducing costs and improving efficiency. Ultimately, the ability to predict rice prices contributes to greater food resilience by enabling stakeholders to prepare for and respond to potential shocks in the market. However, predicting rice prices is a complex task, given the multitude of factors that influence price movements. These factors include domestic production levels, global market trends, weather conditions, government policies, and even geopolitical events. Traditional methods of price forecasting, which often rely on historical data and simple statistical models, may not be sufficient to capture the complexity and dynamism of the rice market.

Rice prices in Indonesia have fluctuated significantly over the years, influenced by a variety of factors including domestic production levels, global market trends, weather conditions, and government interventions. These fluctuations can have profound implications, not just for consumers but for the entire economy. One of the primary drivers of rice price fluctuations is the level of domestic production. Indonesia produces a substantial amount of rice domestically, with key rice-producing regions including Java, Sumatra, and Sulawesi. However, the country's rice production is heavily dependent on seasonal rainfall and is vulnerable to climatic variations. For example, the El Niño phenomenon, which leads to prolonged dry spells, has historically caused significant reductions in rice yields, leading to spikes in rice prices. Conversely, La Niña, which brings above-average rainfall, can lead to better harvests but also increase the risk of flooding, which can damage crops (Trenberth, 1997).

In years when domestic production falls short of demand, Indonesia has to rely on imports to fill the gap. The global rice market is highly volatile, and prices can be influenced by a range of factors, including production levels in major rice-exporting countries like Thailand, Vietnam, and India, as well as global demand and supply dynamics. For instance, during the global food crisis of 2007-2008, rice prices surged dramatically, leading to increased food insecurity in many importing countries, including Indonesia. The reliance on imports during such periods exposes Indonesia to global market fluctuations, which can exacerbate domestic price volatility. Government policies also play a significant role in shaping rice prices. The Indonesian government has historically implemented various measures to stabilize rice prices, including setting floor and ceiling prices, providing subsidies to farmers, and maintaining a strategic rice reserve through the state logistics agency, Bulog. These interventions are aimed at protecting both consumers and producers from the adverse effects of price fluctuations. However, the effectiveness of these policies has been mixed. For example, while price controls can prevent extreme price spikes, they can also lead to unintended consequences, such as discouraging farmers from planting rice if the prices are perceived to be too low to cover their production costs.

The impact of rising rice prices is particularly severe for low-income households, who spend a significant portion of their income on food, with rice being a major component of their diet. When rice prices rise sharply, these households may be forced to reduce their food intake or switch to less nutritious alternatives, leading to an increase in food insecurity and malnutrition. This has broader social implications, as inadequate nutrition can affect health outcomes, particularly for children, and reduce productivity, perpetuating the cycle of poverty (Lestari et al., 2024; Pandiangan et al., 2024). On the other hand, falling rice prices, while beneficial to consumers in the short term, can have negative consequences for farmers. Rice farming in Indonesia is often characterized by small-scale operations with limited access to capital and technology. When prices fall below production costs, farmers' incomes are squeezed, leading to financial stress and, in some cases, forcing them to abandon rice farming altogether. This not only threatens their livelihoods but also jeopardizes future rice production, as fewer farmers may be willing or able to plant rice in subsequent seasons.

To understand the implications of rice price fluctuations, it is important to examine historical trends. Over the past few decades, Indonesia has seen significant variations in rice prices, influenced by a combination of domestic and international factors. For example, during the 1997-1998 Asian financial crisis, Indonesia experienced a severe economic downturn, which led to a sharp depreciation of the rupiah and a corresponding increase in the price of imported goods, including rice. The crisis also caused disruptions in domestic rice production, as farmers faced difficulties in accessing credit and inputs. As a result, rice prices surged, leading to widespread food insecurity and social unrest. More recently, the COVID-19 pandemic in 2020-2021 brought about new challenges. The pandemic disrupted supply chains and led to logistical challenges in transporting rice across the archipelago. Additionally, the economic slowdown caused by the pandemic reduced household incomes, making it more difficult for many Indonesians to afford rice. In response, the government implemented various measures to stabilize prices and ensure food security, including providing cash assistance to low-income households and releasing rice from the national reserve.

Analyzing rice prices on a monthly basis over several years reveals patterns that can be linked to specific events or trends. For instance, rice prices typically rise during the lean season, when stocks from the previous harvest are depleted, and fall after the main harvest season, when supply increases. However, these seasonal patterns can be disrupted by unexpected events, such as natural disasters or global market shocks.

The fluctuations in rice prices have far-reaching socio-economic implications for Indonesia. High rice prices can exacerbate poverty and inequality, as low-income households are disproportionately affected. Food insecurity, which arises from an inability to afford sufficient food, can lead to various negative outcomes, including poor health, reduced educational attainment, and lower productivity. These outcomes, in turn, can hinder economic development and perpetuate the cycle of poverty. Moreover, food insecurity can lead to social unrest, as seen during the 1997-1998 crisis when rising rice prices contributed to widespread protests and riots. Ensuring that rice remains affordable and accessible is therefore crucial for maintaining social stability.

On the other hand, low rice prices, while beneficial for consumers, can have adverse effects on rural communities. In Indonesia, where a large proportion of the population is engaged in agriculture, rice farming is a key source of income. When rice prices fall, farmers' incomes decline, leading to financial stress and, in some cases, forcing them to take on debt or sell assets to make ends meet. This can result in a cycle of poverty and debt, reducing farmers' ability to invest in their farms and improve productivity. In the long term, low rice prices can discourage farmers from planting rice, leading to a decline in domestic production and increased reliance on imports. This, in turn, can make the country more vulnerable to global market fluctuations and reduce food resilience.

Given the importance of rice to Indonesia's food security, the government has implemented various policies to stabilize rice prices and support both consumers and producers. One of the key institutions involved in this effort is Bulog, the state logistics agency. Bulog is responsible for maintaining the national rice reserve, which is used to stabilize prices and ensure food security during times of crisis. Bulog's interventions include purchasing rice from farmers at a guaranteed price, releasing rice from the reserve to stabilize prices, and distributing rice to low-income households through various social assistance programs. These interventions are aimed at ensuring that rice remains affordable for consumers while also providing a stable income for farmers. In addition to Bulog's interventions, the government has implemented various other measures to support rice production and stabilize prices. These include providing subsidies for seeds, fertilizers, and other inputs, improving irrigation infrastructure, and promoting the adoption of modern farming techniques to increase productivity. However, the effectiveness of these policies has been a subject of debate. While they have helped to stabilize rice prices to some extent, there have been instances where they have led to unintended consequences. For example, price controls can sometimes discourage farmers from planting rice if they perceive the prices to be too low to cover their production costs. Additionally, government interventions in the rice market can create distortions that affect the efficiency of the market and lead to resource misallocation.

Looking ahead, Indonesia faces several challenges in maintaining food resilience and ensuring that rice remains affordable and accessible for all. One of the key challenges is climate change, which is expected to have a significant impact on rice production. Changes in rainfall patterns, increased frequency of extreme weather events, and rising temperatures are likely to affect rice yields and increase the volatility of rice prices. To address these challenges, it is important for Indonesia to adopt more sustainable agricultural practices that can help mitigate the impacts of climate change. This includes promoting the use of drought-resistant rice varieties, improving water management practices, and reducing the environmental impact of rice farming. Another challenge is the need to improve the efficiency of the rice supply chain. This includes reducing post-harvest losses, improving transportation and storage infrastructure, and enhancing the distribution of rice across the archipelago. By improving the efficiency of the supply chain, Indonesia can reduce the risk of price fluctuations and ensure that rice remains affordable for all. Finally, there is a need for continued investment in research and development to improve rice productivity. This includes developing new rice varieties that are more resilient to climate change, improving farming techniques, and promoting the use of modern technology in agriculture.

In response to these challenges, researchers are increasingly turning to advanced machine learning techniques that can handle large datasets and complex patterns to predict rice prices. One such technique is XGBoost (Extreme Gradient Boosting), a powerful machine learning algorithm known for its high accuracy and efficiency in predictive modeling. XGBoost is particularly well-suited for time series forecasting, as it can model non-linear relationships and interactions between multiple variables. Studies have demonstrated the effectiveness of XGBoost in forecasting various types of time-series data, such as enhancing the prediction on student perfomance (Asselman et al., 2023), showcasing its versatility and reliability in complex predictive tasks. The research aims to forecast rice prices two years into the future using XGBoost, with a specific focus on fine-tuning the model's hyperparameters and optimizing the number of folds in cross-validation. Hyperparameter tuning is a critical step in building machine learning models, as it involves selecting the best set of parameters that control the learning process of the algorithm. Fine-tuning these parameters, such as the learning rate, maximum depth of the trees, and the number of boosting rounds, can significantly improve the model's performance and predictive accuracy. In addition to hyperparameter tuning, the research will employ k-fold cross-validation, a robust technique for evaluating the model's performance. Cross-validation involves dividing the dataset into k subsets, or "folds," and training the model k times, each time using a different fold as the validation set and the remaining folds as the training set. By averaging the results across all folds, cross-validation provides a more reliable estimate of the model's performance and helps prevent overfitting, where the model performs well on the training data but poorly on unseen data. The combination of hyperparameter tuning and cross-validation in XGBoost is expected to yield a highly accurate model for forecasting rice prices. By leveraging these advanced techniques, the research aims to provide stakeholders with reliable predictions that can inform decision-making and contribute to greater food resilience in Indonesia. By leveraging these advanced methods, Indonesia can build a more resilient food system that can withstand the challenges of the 21st century, ensuring that rice remains affordable and accessible for all.

2. -Literature Overview

2.1 XGboost

XGBoost, short for Extreme Gradient Boosting, is an implementation of gradient boosting that was designed to be highly efficient, flexible, and portable. It was developed by Tianqi Chen and has become one of the most popular machine learning algorithms due to its performance and speed (Chen & Guestrin, 2016). XGBoost is particularly well-suited for structured or tabular data and has been used to win numerous machine learning competitions. Gradient boosting is an ensemble learning technique that builds models sequentially, where each new model attempts to correct the errors made by the previous models. The final prediction is a weighted sum of the predictions made by all individual models (typically decision trees).

The key to understanding XGBoost is to look at its objective function, which it seeks to minimize. The objective function in XGBoost consists of two parts:

1. **Loss Function (L)**: Measures how well the model fits the training data. Common loss functions include mean squared error (for regression) or log loss (for classification).
2. **Regularization Term (Ω)**: Adds a penalty for complexity to prevent overfitting.

The objective function can be written as:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

- $L(y_i, \hat{y}_i)$ is the loss function, which measures the difference between the predicted value $\hat{y}_i$ and the actual value y_i
- $\Omega(f_k)$ is the regularization term for the model f_k, which controls the complexity of the model.

XGBoost builds the model in an additive manner, where the prediction is updated in each iteration by adding a new decision tree. The prediction at the *t*-th iteration is given by:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i)$$

Here, $f_t(x_i)$ is the new decision tree added at the *t*-th iteration. The goal is to minimize the objective function by adding this new tree.

XGBoost uses a second-order Taylor series expansion to approximate the objective function, making the optimization process more efficient. The expansion helps in optimizing the loss function by considering both the first and second derivatives (gradients and Hessians) of the loss function.

The approximation of the objective function at the *t*-th step is:

$$\text{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t(x_i)^2 \right] + \Omega(f_t)$$

- $g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$ is the first derivative (gradient) of the loss function.
- $h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}^2$ is the second derivative (Hessian) of the loss function.

XGBoost uses decision trees as the base learners. In each iteration, it builds a new tree to minimize the objective function. The score for each leaf in the tree is computed by:

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

- w_j is the weight assigned to leaf *j*.
- $I_j$represents the set of data points that fall into leaf *j*.
- $\lambda$is a regularization parameter that controls the leaf weights.

The score for the entire tree is:

$$\text{Score} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T$$

- T is the total number of leaves in the tree.
- γ is a regularization parameter that penalizes the number of leaves.

XGBoost incorporates regularization terms λ and γ into the objective function to control the complexity of the model:

- λ penalizes the L2 norm of the leaf weights, preventing the model from assigning too high weights to any particular leaf.
- γ penalizes the addition of new leaves in the tree, ensuring that the model doesn't grow too complex.

Some of that hyperparameter that used in XGboost :

Table 1 Hyperparametes of XGBoost

Hyperparameters	Contribution in XGBoost Model
max_depth	This hyperparameter controls the maximum depth of each decision tree in the model.
subsample	This hyperparameter determines the fraction of the training data to be used for growing each tree.
colsample_bytree	This hyperparameter specifies the fraction of features (columns) to be used when building each tree.
gamma	This hyperparameter specifies the minimum loss reduction required to make a further partition on a leaf node of the tree.
min_child_weight	This hyperparameter sets the minimum sum of instance weight (or hessian) needed in a child node.
alpha	This hyperparameter controls L1 regularization on leaf weights, which can help make the model more robust by penalizing large coefficients.
lambda	This hyperparameter controls L2 regularization on leaf weights, which penalizes large weights and can help in controlling the model complexity.
eta	This hyperparameter controls the step size shrinkage used to prevent the model from overfitting by scaling down the contribution of each tree.

2.2 Cross-Validation and Fold Optimization

Cross-validation is a technique used to evaluate the model's performance by partitioning the dataset into k subsets (folds). XGBoost often uses k-fold cross-validation to ensure that the model generalizes well to unseen data. Fold optimization involves determining the optimal number of folds and balancing the bias-variance tradeoff (Browne, 2000).

2.3 R-squared

R-squared (R^2) is a statistical measure that represents the proportion of the variance in the dependent variable (the variable you're trying to predict) that is explained by the independent variables (the predictors) in the model. It's also known as the coefficient of determination (Armstrong, 2001).

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- SS_{res} Sum of squares of residuals, or the difference between the observed and predicted values.
- SS_{tot} Total sum of squares, or the difference between the observed values and the mean of the observed values.

The R-squared value ranges from **0 to 1**.

- $R^2 = 0$: This indicates that the model explains none of the variance in the dependent variable. In other words, the independent variables do not explain any of the variation in the dependent variable.
- $R^2 = 1$: This indicates that the model explains all the variance in the dependent variable. The independent variables perfectly predict the dependent variable.

3. Methodology

Given the complexity of time-series forecasting and the influence of various external factors on rice prices, it is essential to adopt a robust and comprehensive methodology. This approach ensures that the model not only captures the intricate patterns within the data but also generalizes well to unseen future observations. By leveraging XGBoost's capabilities and refining its performance through systematic experimentation, the study aims to produce reliable and accurate forecasts that can inform stakeholders and guide decision-making processes.

Figure 2 Reaserch Methodology

The following is the explanation of the flowchart:

1. **Data Collection** : The data was collceted from Badan Pusat Statistic (BPS) Indonesia from January of 2010 until Jully of 2024. The data collected was the average rice price every month in Indonesia.
2. **Pre-Processing Data** : XGboost requires additional parameters in order to make the forecast. From the data obtained, only the average price of each month from the time mentioned is obtained. The addition of new features is necessary to improve the accuracy of the model. Therefore, in this pre-processing, new features are created from the existing data as follows
    a. Year, Month, Day, Weekday, Week of Year, Quarter, Day of Year:
        - Purpose: These features decompose the 'Date' column into its constituent parts. Each of these components can capture different aspects of seasonality or trends in the data.
        - Impact:
            - Year: Helps the model understand long-term trends (e.g., a general increase or decrease over the years).
            - Month: Captures seasonality within a year (e.g., certain months might consistently show higher or lower values).
            - Day: Captures patterns related to specific days of the month.
            - Weekday: Useful for identifying weekly patterns (e.g., prices might fluctuate on weekends versus weekdays).
            - Week of Year: Helps in capturing seasonal patterns across weeks of the year.
            - Quarter: Captures seasonality within quarters, which is often useful in economic or financial data.
            - Day of Year: Can capture annual patterns that might not be evident from month or week alone.
    b. Is Month Start, Is Month End:
        - Purpose: These binary features indicate whether a given date is at the start or end of a month.
        - Impact: These can capture any spikes or drops that typically occur at the beginning or end of months, which can be due to economic cycles, billing cycles, or supply chain activities.
    c. Lag Features (Lag 1, Lag 7, Lag 30):
        - Purpose: Lag features provide the model with past values of the target variable ('Value'). They help capture temporal dependencies by giving the model access to recent past data points. These features are crucial because they help the model understand the temporal relationship between consecutive observations, which is fundamental in time series forecasting.
        - Impact:
            - Lag 1: Captures the immediate past value, which is often a strong predictor of the next value in time series data.
            - Lag 7: Captures the value from the same day in the previous week, useful for weekly patterns.
            - Lag 30: Captures the value from approximately a month ago, useful for monthly seasonality.
    d. Rolling Window Features (Rolling Mean 7, Rolling Std 7):
        - Purpose**:** These features compute the rolling mean and standard deviation over a 7-day window. Rolling features help the model to understand recent trends and the stability of the series over short periods, which are often predictive of future behavior.
        - Impact:
            - Rolling Mean 7: Provides a smoothed version of the recent values, helping to capture trends over the past week.
            - Rolling Std 7: Measures the volatility or variability over the past week, which can help in predicting sudden spikes or drops.
    e. Differencing Features (Diff 1, Diff 7):
        - Purpose: Differencing is a technique used to make a time series stationary by subtracting the previous value from the current one. These features help in stabilizing the time series by removing trends and seasonality, allowing the model to focus on predicting the differences (changes), which are often more stationary.
        - Impact:
            - Diff 1: Captures the day-to-day change in value, which can be crucial for models trying to understand short-term changes.
            - Diff 7: Captures the week-to-week change, helping the model understand weekly fluctuations.
    f. Cyclical Features (Month Sin, Month Cos, Day Sin, Day Cos):
        - Purpose: Time series data often has cyclical patterns (e.g., monthly or daily cycles). These features transform the linear month and day variables into cyclical representations. By using sine and cosine transformations, the model can better understand the cyclicality inherent in time series data, which linear representations fail to capture.
        - Impact:

- ♦ Month Sin/Cos: These encode the cyclical nature of months, ensuring that December and January are considered closer together than December and June.
- ♦ Day Sin/Cos: Similarly, these encode the cyclical nature of days within a month.

g. Interaction Term (Month Year Interaction):

- Purpose: Interaction terms capture the combined effect of two features that might interact with each other. Interaction terms allow the model to learn more complex patterns that might arise from the combination of two or more features, thereby enhancing the predictive power of the model.
- Impact: This specific interaction term (Month * Year) can help capture any trend that is specific to a particular year-month combination (e.g., certain months being particularly strong or weak in specific years).

3. **Find Best Hyperparameter and Fold** : Using cross validation to get the best hyperparameter and folds values. This process is done by taking the r-square value that is closest to the value of 1.
4. **Deploying Best Model**: XGBoost model with the best hyperparameters and folds will be deployed, preparation for forcasting
5. **Forecasting Data** : From the model that has been deployed, forecasting will be carried out for the next 2 years (24 months) to find out the average price of rice.

4. Result and Discussion

The research was conducted using average rice price data from January 2010 to July 2024. The following plots present the analysis of the average rice prices over this period. This data serves as the foundation for exploring trends, seasonality, and other patterns critical to understanding price dynamics in the rice market. The visualization of this data provides a comprehensive overview that supports the subsequent modeling and forecasting efforts within this study.

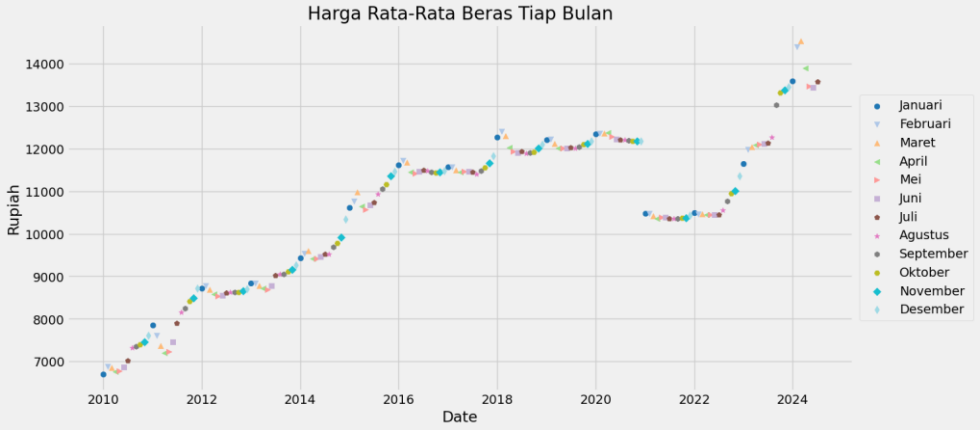

Figure 2 Monthly Average Rice Price Plot

The analysis of the plotted data reveals that from 2010 to December 2021, the price of rice underwent a substantial increase, rising by approximately 70% from its initial level. This significant growth reflects various factors, such as inflation, changes in supply and demand, and possibly the impact of government policies during that period. However, at the beginning of 2021, the rice prices experienced a sharp and notable decline, with the average price dropping to Rp. 10,474. This drastic reduction could be attributed to shifts in market dynamics, including increased production, changes in import policies, or economic conditions influencing consumer purchasing power.

In August 2022, the price trajectory reversed once again, with rice prices climbing steadily to reach their highest recorded level during this period at Rp. 14,528. This peak in prices may have been driven by factors such as supply chain disruptions, changes in agricultural productivity, or other macroeconomic variables. By the end of July 2024, however, the prices began to decline, settling at Rp. 13,571. The overall trend across the years has predominantly been one of increasing prices, although punctuated by periods of decline, highlighting the volatility and sensitivity of the rice market to various internal and external factors.

The subsequent phase of the research involves the creation of new features derived from the existing data. This feature engineering is a critical step designed to enhance the predictive performance and accuracy of the model that will be employed to forecast rice prices over the next two years. By carefully crafting these additional features, the model can more effectively capture underlying patterns and relationships within the data, thereby improving its forecasting capability. The outcomes of this feature engineering process, carried out during the pre-processing stage, are detailed as follows:

Feature Headers and First Few Rows of Data

	Date	Value	Month	Year	day	weekday	weekof year	quarter	dayof year	is_month_ start	is_month_end	lag_1	lag_7	lag_30	rolling_ mean_7	rolling _std_7	diff_1	diff_7	month_ sin	month_ cos	day_sin	day_cos	month_year _interaction
0	2010-01-01 00:00:00	6702.49	1	2010	1	4	53	1	1	1	0	6702.49	6702.49	6702.49	6702.49	131.06	0	0	0.5	0.87	0.2	0.98	2010
1	2010-02-01 00:00:00	6887.83	2	2010	1	0	5	1	32	1	0	6702.49	6702.49	6702.49	6702.49	131.06	185.34	0	0.87	0.5	0.2	0.98	4020
2	2010-03-01 00:00:00	6853.78	3	2010	1	0	9	1	60	1	0	6887.83	6702.49	6702.49	6795.16	131.06	-34.05	0	1	0	0.2	0.98	6030
3	2010-04-01 00:00:00	6761.49	4	2010	1	3	13	2	91	1	0	6853.78	6702.49	6702.49	6814.7	98.66	-92.29	0	0.87	-0.5	0.2	0.98	8040
4	2010-05-01 00:00:00	6772.46	5	2010	1	5	17	2	121	1	0	6761.49	6702.49	6702.49	6801.4	84.83	10.97	0	0.5	-0.87	0.2	0.98	10050

Figure 3 Transformed Data with New Feature

In Figure 3, represents a portion of the data after feature engineering, which will be utilized in the modeling process. In this dataset, the average price of rice for each month is designated as the target variable to be predicted, while the other features will serve as inputs to the model. The original Date and Value columns have been supplemented with a variety of temporal features that capture different aspects of the data's underlying patterns. These include basic temporal features such as Month, Year, day, weekday, weekofyear, quarter, and dayofyear, which allow the model to recognize seasonal trends, weekly cycles, and long-term trends over the years. Additionally, binary indicators like is_month_start and is_month_end are included to capture any specific effects that might occur at the beginning or end of a month, such as price adjustments related to inventory cycles or other market factors.

The dataset also includes several lagged features, such as lag_1, lag_7, and lag_30, which represent the price of rice from 1 day, 7 days, and 30 days prior, respectively. These features are critical for capturing temporal dependencies, allowing the model to understand how recent price changes influence current prices. The rolling statistics (rolling_mean_7, rolling_std_7) further enrich the model by providing a view of short-term trends and volatility over the past week, which can be key indicators of future price movements. Differencing features like diff_1 and diff_7 capture the momentum or rate of change in prices, offering additional predictive power by highlighting recent shifts in the market.

To account for cyclical patterns inherent in the data, sine and cosine transformations of the month (month_sin, month_cos) and day (day_sin, day_cos) have been included. These transformations enable the model to recognize repeating cycles in a way that respects their continuous nature, particularly useful for understanding seasonal variations. Finally, the month_year_interaction feature captures the combined effects of specific months within specific years, adding another layer of detail that might not be captured by the individual month or year features alone. Overall, this comprehensive set of features equips the XGBoost model with the necessary tools to learn and predict the complex patterns in rice prices, improving its accuracy and reliability in forecasting future trends.

The subsequent step involves preparing the hyperparameters that will be fine-tuned for the XGBoost model, along with the division of the data into training and testing sets. Cross-validation will be conducted to evaluate the model's performance based on the chosen hyperparameters and the train-test split ratio, using k-fold cross-validation. The specific hyperparameters that will be utilized are outlined in Table 2, with the values for each hyperparameter detailed as follows:

Table 2 List of Hyperparametes

Hyperparameters	List of Hyperparamter Value
max_depth	[3, 5, 7]
subsample	[0.5, 0.7, 1.0]
colsample_bytree	[0.5, 0.7, 1.0]
gamma	[0, 0.1, 0.2]
min_child_weight	[1, 5, 10]
alpha	[0.1, 0.5, 0.1]
lambda	[1.0, 1.5, 2.0]
eta	[0.01, 0.05, 0.1]

In this research, a 4-fold cross-validation was also utilized to evaluate the performance of the XGBoost model. The decision to use 4 folds in the k-fold cross-validation process was made to ensure a balance between computational efficiency and the robustness of the model evaluation. Specifically, with the chosen set of hyperparameters and the 4-fold configuration, each cross-validation run involved conducting 6,561 individual model fits. When scaled to the full 4-fold cross-validation process, this amounted to a total of 26,244 model fits.

Given the extensive nature of the hyperparameter search, the process required significant computational resources. On average, it took approximately 21 minutes to complete the entire cross-validation process for a single set of hyperparameters, despite being conducted on a low-end laptop. This duration highlights the computational intensity of the process, which underscores the importance of optimizing both the hyperparameters and the number of folds to balance model performance with practical feasibility.

The search for the optimal combination of hyperparameters and fold configuration was driven by the R-squared (R^2) metric, which was employed as the primary criterion for evaluating model performance. The top five configurations,

which demonstrated superior performance across all iterations, were then highlighted as the most effective parameter settings for the XGBoost model.

The comprehensive search for the best hyperparameters and fold number not only ensured that the model was fine-tuned to achieve optimal accuracy but also provided insights into the sensitivity of the model to various configurations. The following section presents the results of this exhaustive tuning process, detailing the top-performing hyperparameter settings identified for the XGBoost model, which are expected to deliver the most accurate predictions based on the R^2 metric.

Table 3 R^2 Value on Each Folds

Number of Folds	R^2 Value
Fold 1	0.9951
Fold 2	0.9842
Fold 3	0.9936
Fold 4	0.9914

Table 4 Hyperparametes of XGBoost

Rangking	R^2	MAE	Parameters
1	0.9911 (+/- 0.0042)	95.7395 (+/- 19.9290)	{'alpha': 0.1, 'colsample_bytree': 0.7, 'eta': 0.1, 'gamma': 0.2, 'lambda': 1.0, 'max_depth': 7, 'min_child_weight': 1, 'subsample': 0.5}
2	0.9911 (+/- 0.0042)	95.7395 (+/- 19.9290)	{'alpha': 0.1, 'colsample_bytree': 0.7, 'eta': 0.1, 'gamma': 0.2, 'lambda': 1.0, 'max_depth': 7, 'min_child_weight': 1, 'subsample': 0.5}
3	0.9911 (+/- 0.0042)	95.7399 (+/- 19.9292)	{'alpha': 0.1, 'colsample_bytree': 0.7, 'eta': 0.1, 'gamma': 0, 'lambda': 1.0, 'max_depth': 7, 'min_child_weight': 1, 'subsample': 0.5}
4	0.9911 (+/- 0.0042)	95.7399 (+/- 19.9292)	{'alpha': 0.1, 'colsample_bytree': 0.7, 'eta': 0.1, 'gamma': 0, 'lambda': 1.0, 'max_depth': 7, 'min_child_weight': 1, 'subsample': 0.5}
5	0.9911 (+/- 0.0042)	95.7408 (+/- 19.9305)	{'alpha': 0.1, 'colsample_bytree': 0.7, 'eta': 0.1, 'gamma': 0.1, 'lambda': 1.0, 'max_depth': 7, 'min_child_weight': 1, 'subsample': 0.5}

The tables 4 presents the results of a model evaluation, showing that the models exhibit highly consistent performance across different hyperparameter settings. The R-squared (R^2) values are consistently high, around 0.9911, with minimal variation, indicating that the models explain nearly all of the variance in the target variable, suggesting an excellent fit to the data. Similarly, the Mean Absolute Error (MAE) values are also stable, around 95.7395 to 95.7408, reflecting that the models' predictions are very close to the actual values on average. The consistency in both R^2 and MAE values across the top-ranked models indicates that the model's predictive accuracy is robust, with minimal sensitivity to changes in the hyperparameters, particularly `gamma`, which varies slightly across the models but does not significantly impact performance. The other hyperparameters remain identical across the top models, suggesting that the chosen configuration is near optimal for this dataset. Overall, these results imply that the model is well-tuned, reliable, and suitable for deployment, with any of the top configurations likely to perform similarly well in practice.

While the best parameter results are found, several other analyses are performed to get a deeper insight into the best model. Feature importance analysis is performed to better understand which features are highly influenced in the XGBoost modeling process on the provided data. A residuals analysis is then performed to determine the performance of the obtained model.

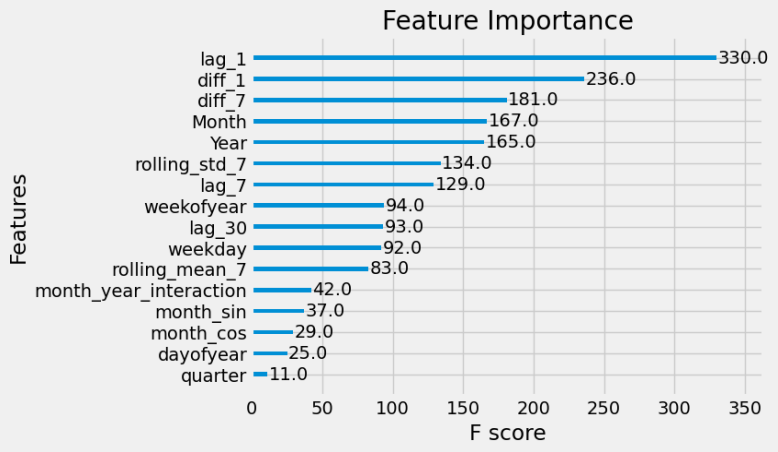

Figure 4 Feature Importance in XGBoost Best Model

From the Figure 4 we know lag_1 with F score of 330 has the highest importance, meaning it is the most frequently used by the model. lag_1 typically represents the value of the target variable from the previous time step, which makes sense as recent values in time series data are often strong predictors of future values. With feature diff_1 (F score = 236) the difference between the current value and the previous value is also very important, capturing the recent trend in the time series. For diff_7, Month, Year, rolling_std_7, These features also have high importance scores, indicating that the model finds them useful in capturing patterns that predict the target variable.

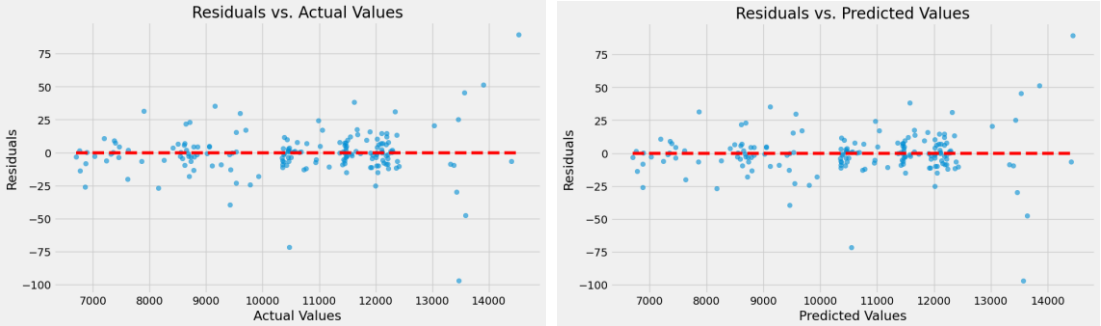

Figure 5 Residual vs. Actual (Left) and Predicted (Right) Values

The residual plots (Figure 5) provide valuable insights into the performance of the regression model. The first plot, which shows residuals versus predicted values, indicates that the residuals are generally centered around zero, suggesting that the model's predictions are reasonably accurate on average. The random distribution of residuals around the zero line is a positive sign, as it implies that there is no strong systematic error in the model's predictions. However, the plot also reveals some points with larger residuals, particularly at higher predicted values, indicating that the model may struggle to predict accurately at these higher levels. This could be a sign of heteroscedasticity, where the variance of the errors is not constant across all predicted values.

Similarly, the second plot, which shows residuals versus actual values, reinforces these observations. The residuals are again centered around zero with no clear pattern, suggesting that the model is not systematically biased and is capturing the general relationship between the features and the target variable. However, the increasing spread of residuals at higher actual values indicates that the model's accuracy might vary depending on the value it is trying to predict. This variability in prediction errors across different ranges of the target variable could also point to potential heteroscedasticity issues.

Overall, while the model appears to perform well, as evidenced by the lack of strong patterns in the residuals, there are areas for improvement, particularly in addressing the variance in errors at higher values. This could involve further refinement of the model or exploring techniques to handle heteroscedasticity more effectively.

Following the identification of the optimal XGBoost model through rigorous hyperparameter tuning and cross-validation, the subsequent step involves deploying the model to forecast rice prices for the upcoming two years, extending beyond July 2024. In this phase, it is essential to carefully handle the initialization of feature values required for the forecasting process, particularly those features that depend on historical data for their computation (lag_1, lag_7, lag_30, diff_1, etc.). To ensure accurate predictions, the last observed value of the rice price, specifically the price recorded in July 2024, is utilized as the starting point for these lagged features. This approach allows the model to maintain continuity and consistency in its predictions by leveraging the most recent data available. Once the model has generated the forecasts for the two-year period, these results are systematically plotted against the original historical data.

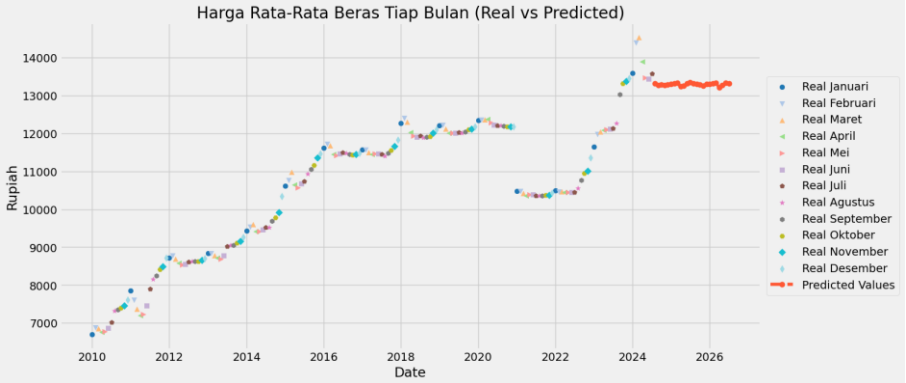

Figure 6 Actual and Predicted Data

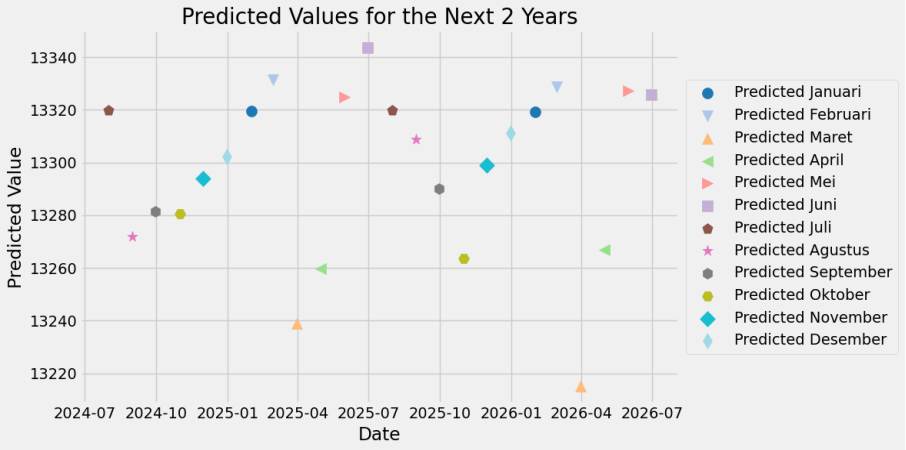

Figure 7 Predicted Data for 24 Month

Figure 6 illustrates the comparison between the actual rice prices in Rupiah over time and the predictions generated by the XGBoost model using the best hyperparameters. The actual data points are color-coded by month, while the predicted values are represented by a solid red line. The close alignment between the predicted and actual values indicates that the model has effectively captured the overall trend and seasonal variations in rice prices over the years, demonstrating high accuracy. The seasonal trends, as shown by the various colors for each month, are well reflected in the model's predictions, suggesting that the feature engineering process successfully incorporated these seasonal patterns. Toward the end of the timeline, where the actual data ends, the model predicts a continuation of the upward trend in prices, though with some flattening, indicating an expectation of price stabilization in the near future.

The second plot (Figures 7) expands in on the model's predictions for the next 24 months, showing the predicted rice prices for each month without the actual historical data for context. The plot reveals relatively small fluctuations in predicted prices from month to month, suggesting that the model expects moderate variation but overall stability in rice prices over the two-year forecast period. The clustering of predicted values within a narrow range indicates that the model does not foresee significant changes, implying stable market conditions or consistent seasonal factors based on historical patterns.

Overall, the combination of both plots demonstrates that the XGBoost model, with well-tuned hyperparameters, performs effectively in capturing historical trends and forecasting future prices. The model's predictions for the next two years show a stable trend with expected seasonal fluctuations, offering reliable forecasts that could be valuable for stakeholders in planning and decision-making related to rice price stability and market behavior.

5. Conclusion

This study demonstrates the effectiveness of the XGBoost model, fine-tuned through hyperparameter selection and 4-fold cross-validation, in forecasting rice prices in Indonesia for the two years following July 2024. Using data from January 2010 to July 2024, the research employed detailed feature engineering to capture temporal and cyclical patterns, resulting in highly accurate predictions. In addition to internal features derived from historical prices, it is essential to consider external factors such as climate patterns, government policies, global market trends, economic indicators, and technological advancements for a more comprehensive forecasting model. These external variables can significantly influence rice prices and should be integrated into future models to improve predictive accuracy. The results indicate that a well-tuned XGBoost model, combined with relevant external factors, is a powerful tool for forecasting rice prices. This approach offers valuable insights for stakeholders, aiding in informed decision-making related to production, pricing strategies, and food security planning. Overall, this study highlights the potential of advanced machine learning techniques in addressing the complexities of time series forecasting in dynamic markets..

6. REFERENCES

Armstrong, J. S. (2001). Evaluating Forecasting Methods. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 443–472). Springer US. https://doi.org/10.1007/978-0-306-47630-3_20

Asselman, A., Khaldi, M., & Aammou, S. (2023). Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments*, *31*(6), 3360–3379. https://doi.org/10.1080/10494820.2021.1928235

Azadi, M., Yousefi, S., Farzipoor Saen, R., Shabanpour, H., & Jabeen, F. (2023). Forecasting sustainability of healthcare supply chains using deep learning and network data envelopment analysis. *Journal of Business Research*, *154*, 113357. https://doi.org/10.1016/j.jbusres.2022.113357

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Fraher, E., Knapton, A., McCartha, E., & Leslie, L. K. (2024). Forecasting the Future Supply of Pediatric Subspecialists in the United States: 2020–2040. *Pediatrics*, *153*(Supplement 2), e2023063678C. https://doi.org/10.1542/peds.2023-063678C

Lestari, A. D., Erlikasna, E., Simbolon, R. C., Breta, I., & Daniyal, M. (2024). Dampak Fluktuasi Harga Beras, Bawang Merah, Cabai Terhadap Inflasi. *Jurnal Sosial Ekonomi Pertanian*, *20*(2), Article 2. https://doi.org/10.20956/jsep.v20i2.35057

Pandiangan, T. M., Simbolon, A. P., Sihite, S., Siregar, R., & Yunita, S. (2024). Analisis Dampak Kenaikan Harga Beras terhadap Kehidupan Masyarakat Kelas Ekonomi ke Bawah: Kiat Pemerintah Jaga Kebutuhan Beras di Indonesia. *Jurnal Pendidikan Tambusai*, *8*(2), 23959–23966.

Subhra, S., Mishra, S., Alkhayyat, A., Sharma, V., & Kukreja, V. (2023). Climatic Temperature Forecasting with Regression Approach. *2023 4th International Conference on Intelligent Engineering and Management (ICIEM)*, 1–5. https://doi.org/10.1109/ICIEM59379.2023.10166883

Trenberth, K. E. (1997). The Definition of El Niño. *Bulletin of the American Meteorological Society*, *78*(12), 2771–2778. https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2