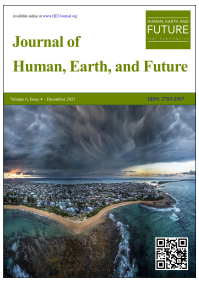
Available online at www.HEFJournal.org

ISSN: 2785-2997

Journal of
Human, Earth, and Future

Vol. 6, No. 4, December, 2025



Machine Learning-Based Forecasting of Agricultural Commodity Prices Using Ensemble Models

Dwi Rizky Lestari [1], Eliza Aditya Stanly Bangun [1], Ford Lumban Gaol [2*], Tokuro Matsuo [3*]

[1] *School of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia.*

[2] *BINUS Graduate Program, Department of Computer Science, Bina Nusantara University, Jakarta 11480, Indonesia.*

[3] *Advanced Institute of Industrial Technology, Tokyo, Japan.*

Received 18 March 2025; Revised 16 October 2025; Accepted 03 November 2025; Published 01 December 2025

Abstract

This study aims to forecast the prices of key food commodities including garlic, shallots, cayenne pepper, and red chili in Kota Singkawang using three machine learning models: Linear Regression, Random Forest, and XGBoost. The dataset, sourced from BPS Kota Singkawang for the 2016–2023 period, underwent preprocessing to address missing values and outliers, followed by correlation-based feature selection. Model training involved grid search and cross-validation to ensure robust performance evaluation. The findings indicate that XGBoost consistently outperforms the other models, achieving the highest R^2 values (up to 0.82) and the lowest MAPE (5–10%), demonstrating its ability to capture complex nonlinear relationships and account for external factors such as inflation and seasonality. Random Forest ranked second in predictive accuracy, especially for garlic, while Linear Regression was less effective for volatile commodities. Notably, features such as rainfall intensity and national holidays were found to significantly influence price movements. The novelty of this research lies in its localized approach to price forecasting using ensemble models combined with macroeconomic and climatic variables. The results offer a practical tool for local policymakers to anticipate price volatility and design evidence-based interventions to enhance food security and price stability at the regional level.

Keywords: Price Prediction; Random Forest; Linear Regression; XGBoost; Food Commodities.

1. Introduction

Prices of strategic food commodities play an important role in economic stability and community welfare [1], particularly in places like Kota Singkawang, a major trade hub in West Kalimantan that experiences complex price dynamics affecting businesses, local authorities, and consumers. Price fluctuations are influenced by various factors such as weather conditions, seasonal demand, logistics infrastructure quality, and government economic policies [2, 3]. In Indonesia, persistent volatility in food prices challenges inflation control and the purchasing power of households, as noted by Cahaya [3]. Moreover, climate anomalies such as extreme rainfall resulting from climate change have been shown to significantly disrupt food production and distribution systems [4, 5]. Infrastructure-related constraints, including poor road quality and limited transportation facilities, further amplify logistics costs and final commodity prices [1]. Government interventions such as subsidies, trade policies, and price regulations also play a critical role in stabilizing or destabilizing market conditions [2].

* Corresponding author: fgaol@binus.edu; matsuo@aiit.ac.jp

http://dx.doi.org/10.28991/HEF-2025-06-04-09

➢ This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

While various predictive approaches have been explored, including statistical and computational methods, seasonal demand (e.g., during holidays) remains an underexplored variable in price forecasting models [3]. Sudden spikes in consumption during festive periods often cause market imbalances and sharp price increases. Thus, predictive models incorporating both temporal and external variables are needed to better capture these dynamics [1].

Recent studies increasingly apply Machine Learning techniques such as Random Forest [6] and boosting models like XGBoost [7], which outperform traditional Linear Regression by effectively modeling nonlinear, multivariate relationships [8]. However, most existing research focuses on single commodities in major urban centers, with limited attention to secondary cities like Singkawang, where local socio-economic and environmental factors shape distinct data patterns [9, 10].

Advances in computing and big data enable the use of ensemble learning models like Random Forest and XGBoost for commodity price prediction due to their capability to capture nonlinear interactions and combine multiple decision trees for improved accuracy and robustness [11, 12]. Random Forest reduces overfitting by generating numerous decision trees from random subsets of features and training data [13], while XGBoost utilizes efficient iterative boosting with computational optimizations such as parallelization and automatic feature selection [14]. Linear Regression remains a widely used baseline model for its simplicity and interpretability, especially in exploratory studies [15], but it often struggles with the complexity of volatile, multifactor data. Deep learning models like Long Short-Term Memory (LSTM) networks are increasingly adopted to capture complex temporal dependencies inherent in commodity price data [14].

To be very exact, in Indonesia, food commodity price volatility arises from climatic variability, uneven logistics infrastructure, seasonal increases in demand during holidays, as well as macroeconomic factors such as inflation and government policies [3]. Hence, accurate forecasting models must incorporate these external factors to improve prediction reliability and support effective planning and policy-making. Although deep learning methods like LSTM and Gated Recurrent Units (GRU) handle large, high-dimensional datasets, conventional models like Random Forest, Linear Regression, and XGBoost remain relevant for their balance of accuracy, efficiency, and explanatory power [16-19].

Previous studies demonstrate various applications of these methods. Wang & Guo [20] developed a hybrid ARIMA–XGBoost model to forecast stock market volatility, achieving a 10–15% reduction in Mean Squared Error (MSE) compared to ARIMA alone. Gupta et al. [21] demonstrated that Multiple Linear Regression effectively predicted daily temperature with a coefficient of determination (R^2) of about 0.95 and low mean squared error. Suryani et al. [22] combined system dynamics with machine learning regression to improve price forecasting accuracy for basic necessities in Indonesia, reducing Mean Absolute Percentage Error (MAPE) by 10–15%. Liu [23] applied Multiple Linear Regression to real estate market forecasting, reporting an R^2 close to 0.9 and MAPE of approximately 8–10%. Meenal et al. [24] showed that Random Forest achieved classification accuracies of 90–92% in weather prediction. Saadah [25] used Random Forest for short-term Bitcoin price prediction with a MAPE of 7–10%. Sinambela [26] compared Multiple Linear Regression and Support Vector Machine (SVM) in gold price prediction, finding SVM superior by producing lower Root Mean Squared Error (RMSE). Vuong et al. [27] developed a hybrid XGBoost–LSTM model for stock price forecasting, decreasing RMSE by 10–12% compared to LSTM alone. Ben Jabeur & Mefteh-Wali [28] achieved an RMSE of around 2.13 for gold price forecasting using XGBoost with SHAP explainability, identifying key influential factors such as exchange rates and stock indices. These results confirm that while Linear Regression provides interpretable baseline results, ensemble models like Random Forest and XGBoost often outperform traditional methods in volatile, noisy, nonlinear environments.

Additionally, hybrid and deep learning approaches further enhance predictive capabilities by capturing complex temporal and feature interactions [20, 27]. This evidence underscores the importance of assessing these methods within the context of food commodity price forecasting in Kota Singkawang, taking into account its unique socio-economic and environmental factors. Therefore, this study aims to fill this gap by comparatively evaluating Random Forest, Linear Regression, and XGBoost in forecasting prices of strategic food commodities in Kota Singkawang. By integrating critical external variables such as rainfall, inflation, and holiday periods, it seeks to provide a comprehensive understanding of regional price dynamics. The findings are intended to support policy formulation and business planning while laying the foundation for developing data-driven early warning systems to enable adaptive responses to food price volatility in the region.

2. Methods

Figure 1 shows the flowchart outlining the main steps of the research methodology.

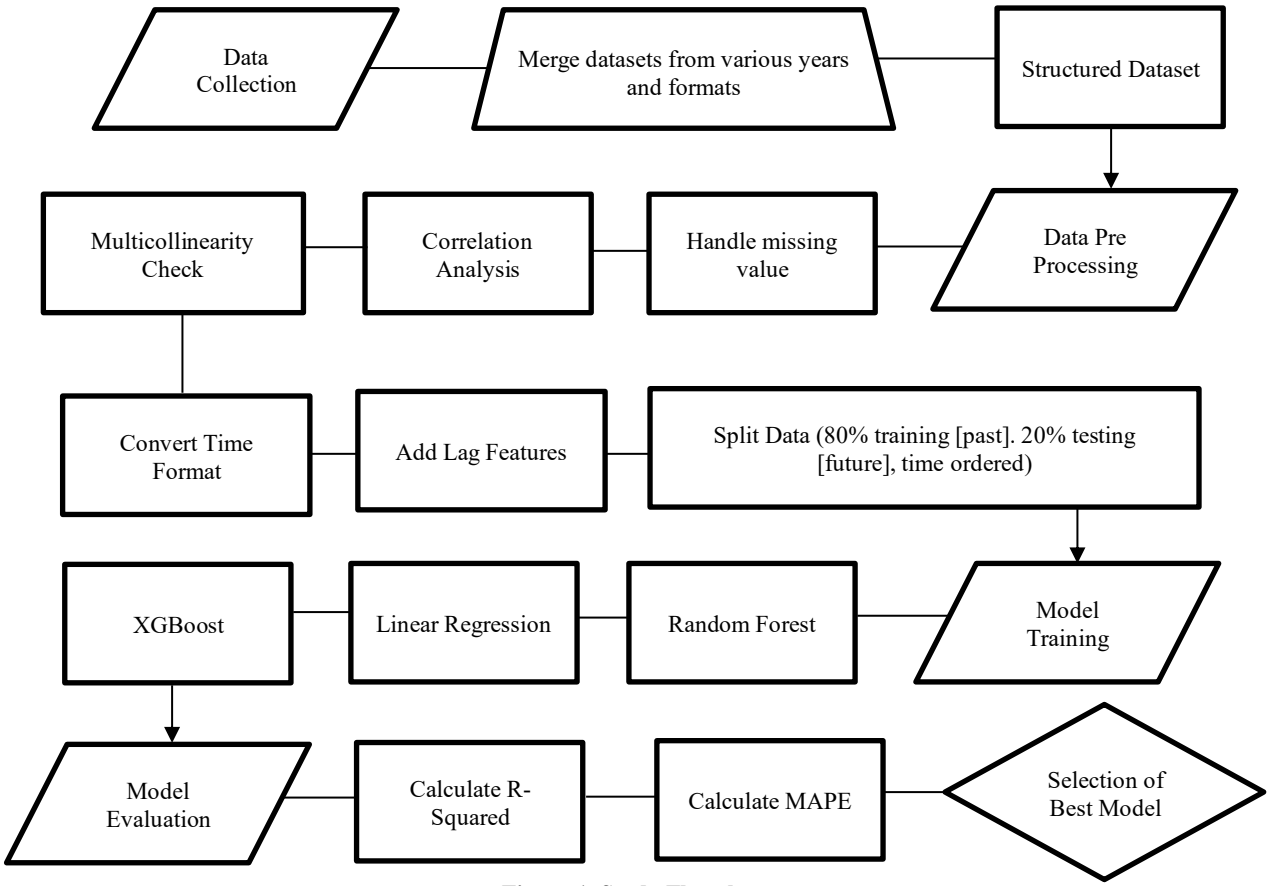


Figure 1. Study Flowchart

2.1. Data Collection

This study uses monthly price data of red chili, cayenne pepper, shallots, and garlic from 2016 to 2023. In addition, external variables such as rainfall (mm), the number of rainy days, and the inflation rate (%) are included as factors affecting prices. The data were obtained from the Badan Pusat Statistik (BPS) Kota Singkawang, which provides monthly average price information from various major markets in the region. Data on rainfall and the number of rainy days were also sourced from local meteorological records maintained by BPS Kota Singkawang, while inflation data were retrieved from annual reports published by the same agency. As the data were provided in the form of annual reports, it was necessary to compile data across multiple years to form a complete dataset spanning from 2016 to 2023. Additionally, rainfall and inflation data were available in separate formats, thus requiring a merging process to ensure the dataset was well-structured.

2.2. Data Preprocessing

The next step is data preprocessing, which includes several key stages to ensure the data are ready for use in machine learning models. These stages include feature selection, time format conversion, and lag application to account for temporal aspects in commodity price prediction.

2.3.1. Missing Values Handling

During preprocessing, missing values were carefully examined across all input features, including rainfall (mm), the number of rainy days, and inflation rates. Where missing values were identified, imputation was performed using the mean of the respective feature. This approach was selected due to the relatively small proportion of missing data and the assumption that the data were missing at random (MAR). The mean imputation technique ensures consistency across samples while minimizing distortion of the original data distribution. To preserve time-series continuity, especially in the context of inflation and weather data, forward or backward interpolation was not used, in order to avoid introducing artificial trends or seasonality that could bias the model. Furthermore, missing data, primarily caused by inconsistent reporting, was addressed using linear interpolation. Anomaly detection using a rolling z-score method identified and removed extreme outliers, which were verified as data entry errors. This preprocessing step significantly improved model accuracy, as reflected in the cross-validation results.

2.3.2. Feature Selection

Feature selection in this study was initially guided by correlation analysis to identify variables with significant relationships to the target price series. To further ensure the reliability of the regression models, multicollinearity among input features was assessed using the Variance Inflation Factor (VIF). This diagnostic measures the extent to which each feature is linearly explained by the others. The computed VIF values for all selected predictors including rainfall (mm), number of rainy days, and inflation rate (%), remained well below the commonly accepted threshold of 5, indicating low multicollinearity and justifying their inclusion in the model.

2.3.3. Time Format Conversion

Since the data came from separate sources and in different formats, the first step was to convert the month column into datetime format. This conversion allows the data to be sorted chronologically and enables the model to recognize trends over time. The datetime format used in this process is MM-YYYY, reflecting the monthly price data.

2.3.4. Lag Addition

Lags were added to provide the model with information on previous price values, allowing it to identify price movement patterns over time. In this study, the lag structure differs for each commodity. A lag of 3 was applied to red chili, shallots, and garlic, meaning the model uses data from the previous three months as predictors. A lag of 5 was used for cayenne pepper, given its higher price volatility and the need for a longer historical context. Lags were added to the dataset by creating new columns containing price values from previous periods. This adjustment helps the model capture both seasonal patterns and long-term trends, thereby improving its ability to forecast future prices.

2.3. Data Splitting

After preprocessing, the dataset was split into two parts: 80% for training and 20% for testing. This division ensures the model learns from historical data before being evaluated on unseen data. Each commodity was split separately, resulting in four different training and testing sets. Since the dataset is time-series in nature, a time-ordered partitioning method was used. Rather than randomly splitting the data, the first 80% of the chronological data was used for training, and the final 20% was reserved for testing. The 80:20 ratio balances the availability of data for training and evaluation. A smaller training set may lead to underfitting, while a limited testing set may not sufficiently evaluate the model's generalizability.

2.4. After Pre-Processing

2.5.1. Model Training

In this study, three machine learning models including Random Forest, Linear Regression, and Extreme Gradient Boosting (XGBoost) were used to predict agricultural commodity prices. Each model has a different approach in analyzing historical price patterns and capturing trends for prediction. These models were selected based on their suitability for time series data and their ability to handle non-linear relationships.

2.5.2. Random Forest

Random Forest is an ensemble algorithm that builds multiple decision trees to improve prediction accuracy and reduce overfitting. It works by creating trees from different data subsets and combining their predictions for a more stable output. Its strengths include handling non-linear relationships and robustness against irrelevant features and noise [6]. However, Random Forest has higher complexity and computational requirements than linear models, due to the need to build numerous trees. Despite this, it remains popular for time-series analysis because of its ability to model complex patterns [29].

2.5.3. Linear Regression

Linear Regression is a statistical model that predicts outcomes based on a linear relationship between the independent and dependent variables [30]. It assumes commodity prices can be expressed as a linear combination of contributing factors. The Ordinary Least Squares (OLS) equation is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_3 X_3 + \varepsilon \quad (1)$$

where, Y is the dependent variable (commodity price); X_1, X_2, ..., X_n are independent variables; β_0 is the intercept, β_1 to β_n are regression coefficients that show the effect of each variable on the price; and ε is the residual error.

Linear Regression is used as a baseline model due to its simplicity and interpretability.

2.5.4. XGBoost

Extreme Gradient Boosting (XGBoost) is a boosting algorithm developed to enhance accuracy while maintaining efficiency. It builds trees sequentially, where each new tree corrects errors made by previous ones. This process reduces bias and improves generalization. XGBoost outperforms other models in handling non-linear relationships and is optimized for speed and memory efficiency. It also includes regularization to prevent overfitting. However, it requires careful hyperparameter tuning to perform optimally. In this study, tuning was conducted to maximize model performance [7].

2.5. Model Evaluation

2.6.1. R-Squared

R-squared (R^2) measures the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating better explanatory power. A negative R^2 implies the model performs worse than a simple average. R^2 is calculated using the following formula:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \quad (2)$$

where, y_i is the actual value; $\hat{y}_i$ is the predicted value; and $\bar{y}$ is the average of the actual values.

The higher the R^2 value, the better the model is at capturing historical data patterns and predicting future prices.

2.6.2. Mean Absolute Percentage Error (MAPE)

MAPE expresses prediction error as a percentage, making it easily interpretable. MAPE is calculated with the following formula:

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3)$$

where, y_i is the actual value; $\hat{y}_i$ is the predicted value; and n is the total number of observations.

MAPE is intuitive: e.g., a MAPE of 10% means an average prediction error of 10%. However, it becomes unreliable when actual values are very small or zero. Using both R^2 and MAPE provides a comprehensive evaluation. A high R^2 with a high MAPE suggests good pattern recognition but large prediction errors, indicating the importance of balancing both metrics.

2.6. Model Visualization

To evaluate model performance, visualizations comparing predicted and actual prices were created. These graphs help assess prediction accuracy and identify periods of large error. Such visual insights are essential for understanding how well the model captures price trends over time.

3. Results and Discussion

In this section, the results of commodity price prediction are analyzed based on the performance of three applied machine learning models: Random Forest, Linear Regression, and XGBoost. The evaluation utilizes two primary metrics including R-Squared (R^2), which measures the model's ability to explain variations in commodity prices, and Mean Absolute Percentage Error (MAPE), which assesses the prediction error as a percentage.

Each model is evaluated across four key commodities including red chili, cayenne pepper, shallots, and garlic using different lag configurations tailored to each commodity's historical patterns. Additionally, a comparative analysis is conducted to identify the model with the most optimal performance in predicting agricultural commodity prices.

3.1. Random Forest Evaluation

Through Random Forest Evaluation as shown in Figure 2, there is an illustration which illustrates the comparison between actual and predicted prices for the four commodities, showing the relationship between actual and predicted prices in Rupiah per kilogram. The results indicate varying levels of accuracy depending on the commodity being predicted.

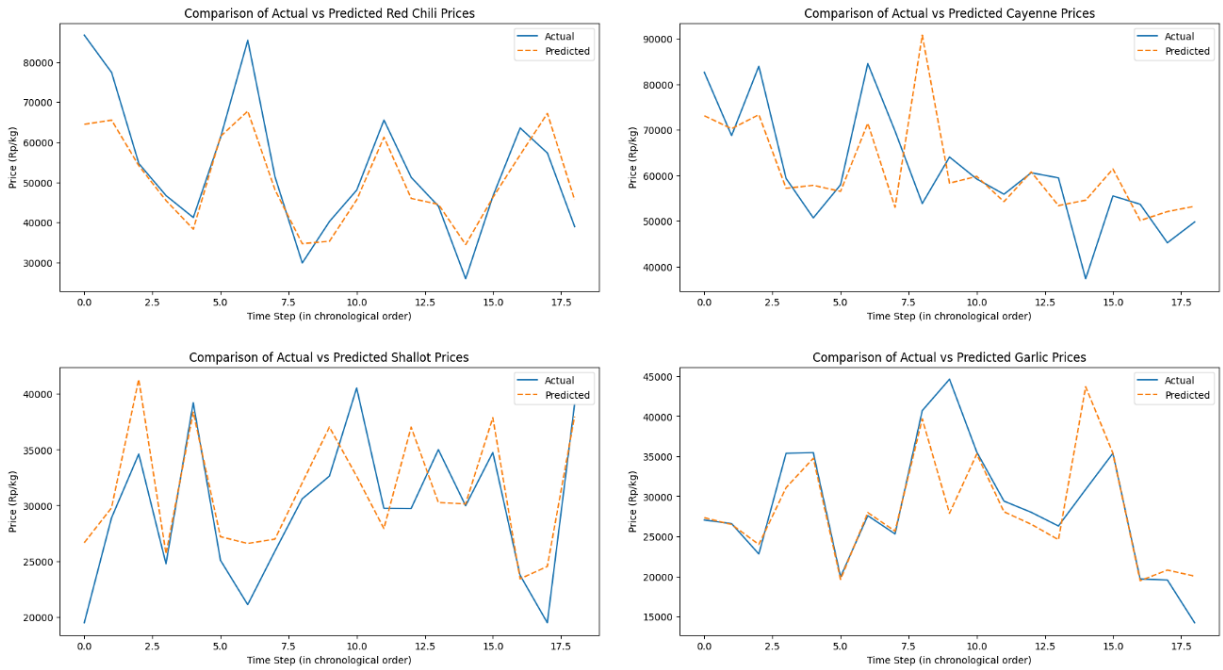


Figure 2. Comparison of Actual and Predicted Prices of Random Forest Model

Insights from the Random Forest model highlight its strengths and limitations in capturing temporal patterns. While the model effectively detects seasonal peaks and troughs across all commodities, it tends to smooth out extreme fluctuations. For instance, in the red chili series, a pronounced price spike to 85,000 Rp/kg is under-predicted at approximately 67,000 Rp/kg, and deep troughs are similarly smoothed toward the mean. In the cayenne chili series, initial values align closely, but sharp reversals, such as the drop at index 3 and the rebound at index 7, are significantly mispredicted. This suggests that incorporating short-term shock indicators (e.g., pest outbreak alerts) could mitigate these errors. Shallot price peaks are well captured; however, rapid early-season surges and late-season declines are under-predicted, indicating that rolling-window change features may enhance the model's responsiveness to turning points. Garlic predictions on the other hand exhibit strong baseline alignment but fail to model abrupt mid-series spikes, underscoring the need for external event flags (e.g., import-quota changes) to capture exogenous shocks.

3.1.1. Red Chili Prices

The model achieved the best performance for red chili, with an R^2 score of 0.7395, indicating that approximately 74% of the variance in the red chili price data was captured by the model. The MAE of 6008.35 and RMSE of 8352.51 show that the model maintains reasonably low prediction errors in absolute terms. The MAPE of 10.99% falls within the "good prediction" range (typically <15%), suggesting that predictions are reliable from a percentage-based perspective. This strong performance can be attributed to the predictable temporal structure of red chili prices, which may follow regular seasonal cycles influenced by climate and supply chain timing. Additionally, the model likely benefited from lag features, which provided meaningful historical context for future price movements. The inflation and rainfall-related features may have further helped capture macroeconomic and environmental influences, making red chili a relatively learnable target for Random Forest.

3.1.2. Cayenne Chili Prices

In contrast, the model demonstrated limited effectiveness in predicting cayenne chili prices, with an R^2 score of 0.1229, MAE of 7924.20, RMSE of 11608.15, and MAPE of 13.89%. While the MAPE remains below the 15% threshold, the low R^2 score reveals that most of the variance in cayenne prices is unexplained by the model. This poor performance likely stems from the higher volatility and irregular patterns in cayenne chili prices, which may be driven by factors outside the current feature set. For instance, short-term supply shocks, localized demand spikes, pest outbreaks, or market speculation could influence prices in ways not captured by rainfall, inflation, or lagged prices alone. The lag features, while effective for red chili, may not be sufficient for modeling commodities with more erratic behavior.

3.1.3. Shallot Prices

The model yielded moderate performance for shallot prices, achieving an R^2 score of 0.5613, which means it could explain over half of the observed variance. The MAE of 3295.49, RMSE of 4198.83, and MAPE of 11.87% indicate a balance between predictive reliability and variance capture. Shallots often follow seasonal cultivation and harvesting

cycles, which may have been successfully modeled using lagged price inputs and month-based categorical features. However, occasional price spikes, influenced by festive demand or regional supply limitations, may introduce nonlinearities that the Random Forest model cannot fully capture without more contextual features such as regional distribution data or demand indices.

3.1.4. Garlic Prices

The model also performed relatively well on garlic prices, with an R^2 score of 0.5365, MAE of 2647.65, RMSE of 5176.24, and MAPE of 9.07%, the lowest MAPE among all commodities. These results indicate high prediction accuracy and low relative error. Garlic prices are generally more stable and may be influenced by import quotas, storage conditions, or bulk purchasing patterns, which tend to follow regular trends. The model likely captured these patterns effectively through historical price features, even without explicit economic or policy inputs. The low error metrics suggest garlic prices are more resilient to short-term fluctuations, making them easier for machine learning models to learn and predict.

3.2. Linear Regression Evaluation

Through Linear Regression Evaluation as shown in Figure 3, there shows a plot-driven analysis of the Linear Regression forecasts, further illustrating its characteristic behavior across commodities.

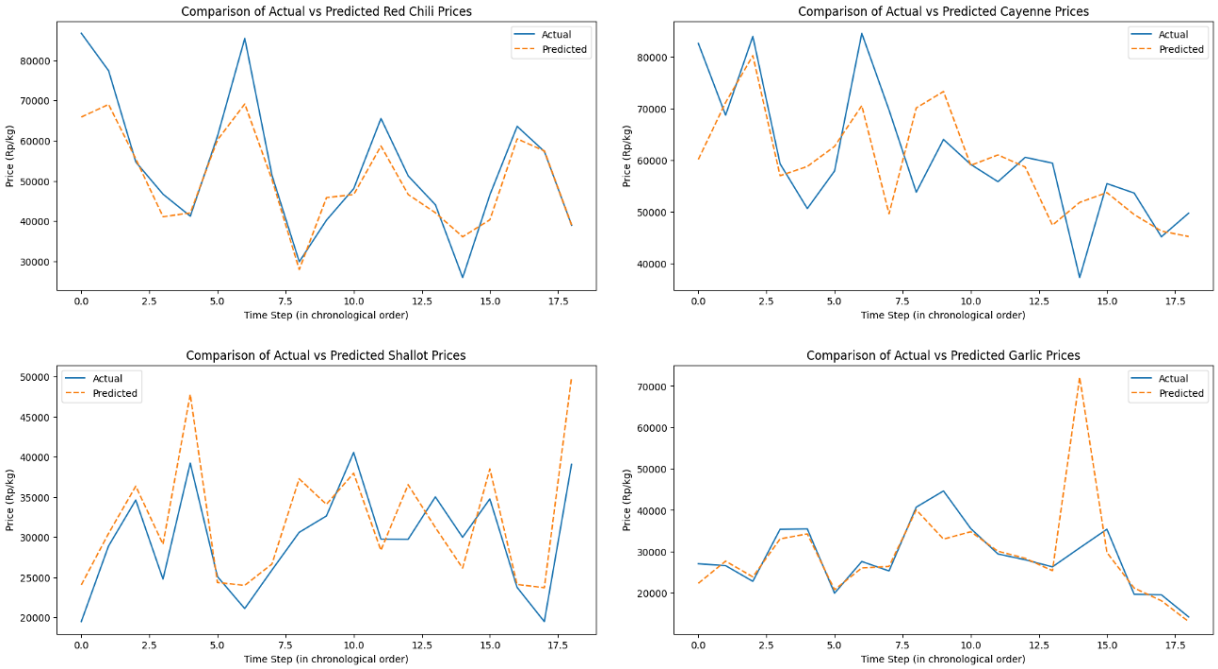


Figure 3. Comparison of Actual and Predicted Prices of Linear Regression Model

For Red Chili, the model captures the overall downward trend and the rebound at indices 5–6, yet underestimates extreme values, forecasting ~68,000 Rp/kg for the spike at index 6, versus ~85,000 Rp/kg actual, and smoothing the trough at index 8. In Cayenne Chili, early indices (0–2) align well, but the model fails to adapt to mid-series volatility, missing the plunge at index 3 (predicted ~57,000 vs. actual ~51,000 Rp/kg) and the peak at index 7 (~49,000 vs. ~92,000 Rp/kg). For Shallot, smooth seasonal ramps (indices 2–6) are well modeled, but sharp mid-season peaks at index 3 and late peaks at indices 9–10 are damped, indicating a need for rolling-change or month-dummy features. In Garlic, the baseline seasonality (indices 0–6) is closely followed, yet abrupt spikes (e.g., index 13 actual ~44,000 Rp/kg vs. predicted ~27,000 Rp/kg) are entirely missed, underscoring the model's inability to learn exogenous shocks. These insights confirm that while Linear Regression reliably predicts broad seasonal timing, it systematically regresses the amplitude of extremes toward the mean.

3.2.1. Red Chili Prices

Linear Regression performed best on the red chili dataset, achieving an R^2 score of 0.7909, meaning that the model explained nearly 79% of the variance in red chili prices. This indicates a strong linear relationship between the features (e.g., rainfall, inflation, and lagged prices) and the target variable. The MAE of 5,094.12 and RMSE of 7,484.48 show relatively low prediction errors, while the MAPE of 9.37% reflects a high level of relative accuracy. These results suggest that red chili prices are largely influenced by linear and time-dependent factors, which Linear Regression can capture effectively. The inclusion of lagged price variables likely improved the model's temporal understanding. This commodity may also exhibit seasonal regularity and gradual trends that align well with the assumptions of linearity.

3.2.2. Cayenne Chili Prices

In contrast, cayenne chili prices yielded only a moderate R^2 score of 0.3154, indicating that less than one-third of the variance in price could be explained by the model. Despite a MAPE of 13.04%, which is still within acceptable bounds, the MAE (7,834.72) and RMSE (10,255.76) are relatively high. The moderate accuracy and low R^2 score highlight a key limitation of Linear Regression: it struggles with modeling volatile or nonlinear behavior. Cayenne chili prices are likely subject to supply chain disruptions, market shocks, and localized events that introduce irregularity and nonlinearity. Linear Regression, which assumes additive and linear relationships, is inherently unable to capture such complex dynamics unless transformed features or interaction terms are introduced.

3.2.3. Shallot Prices

For shallot, the model achieved an R^2 score of 0.4654, which is moderate, indicating that just under half of the variance in shallot prices is explained. The MAE (3,729.25) and RMSE (4,635.08) are reasonable, and the MAPE (12.55%) suggests acceptable predictive accuracy. Shallot pricing may be influenced by a combination of seasonal cycles, planting schedules, and delayed market responses, some of which exhibit linear trends while others do not. The linear model seems to partially capture these patterns, especially through the use of lag features and monthly indicators. However, the performance also suggests that nonlinear seasonal interactions or regional variables may be needed for improved accuracy.

3.2.4 Garlic Prices

The Linear Regression model performed poorly for garlic prices, with a negative R^2 score of -0.7366, indicating that the model's predictions are worse than simply predicting the mean of the target variable. The high RMSE (10,019.21) and MAPE (13.66%) reinforce this conclusion. Garlic prices may follow nonlinear or segmented patterns, potentially influenced by external trade policies (e.g., import quotas), sudden bulk purchasing behavior, or storage-driven price lags that linear models cannot represent effectively. Additionally, the model may be sensitive to outliers or high-variance periods in garlic price history, which Linear Regression handles poorly without robust transformation or outlier filtering. These results underscore the strengths and limitations of Linear Regression in the context of agricultural price forecasting. The model performs well when the price dynamics are relatively stable, linear, and history-driven, as in the case of red chili. However, when prices are affected by nonlinear factors, abrupt market changes, or unmodeled exogenous variables, as observed with garlic and cayenne chili, the performance degrades significantly.

3.3. XGBoost Evaluation

Through XGBoost Evaluation shown in Figure 4, there presents a plot-driven analysis of the XGBoost forecasts, further elucidating the model's behavior.

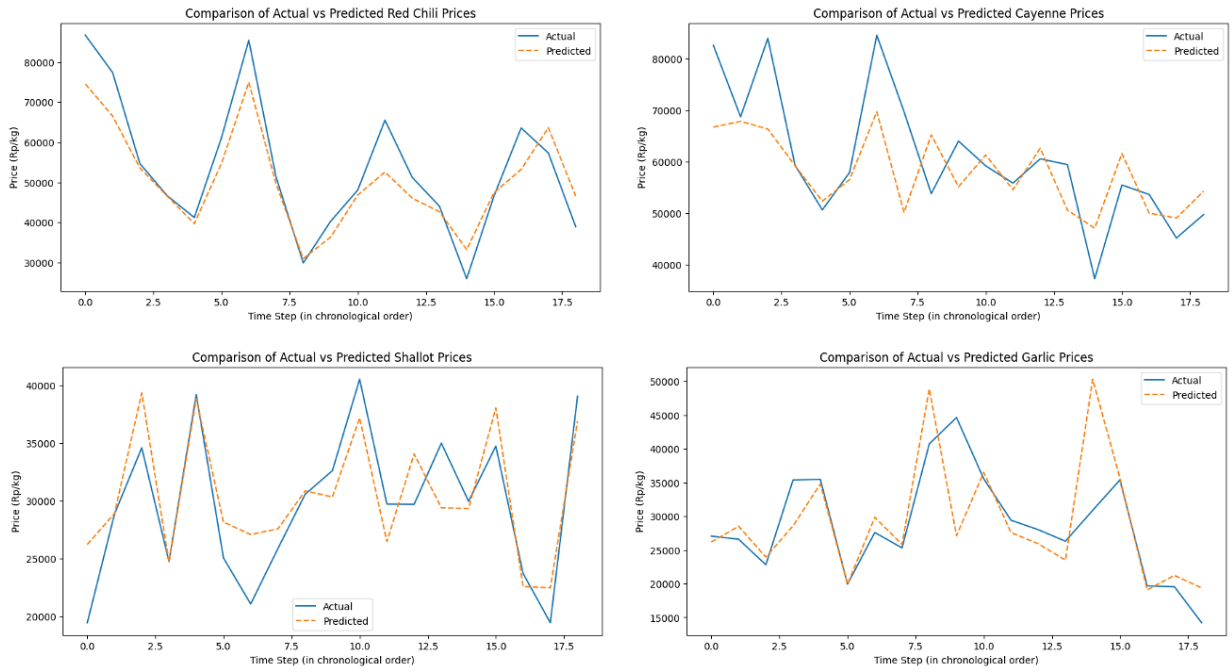


Figure 4. Comparison of Actual and Predicted Price of XGBoost Model

For Red Chili, XGBoost closely matches both the timing and magnitude of seasonal swings, predicting the spike at index 6 at ~68,000 Rp/kg versus ~85,000 Rp/kg actual, and capturing troughs within a 5,000 Rp/kg margin. In Cayenne Chili, the model significantly reduces amplitude error compared to Random Forest, yet still underestimates the peak at index 7 (~70,000 Rp/kg predicted vs. ~92,000 Rp/kg actual) and smooths the plunge at index 3, indicating remaining volatility blind spots. Shallot predictions track mid-season peaks (indices 3 and 10) within 5,000 Rp/kg, demonstrating strong capture of nonlinear interactions. However, late-season declines around index 13 are under-predicted, suggesting residual damping. For Garlic, XGBoost outperforms other models by predicting the index 13 surge (~50,000 Rp/kg predicted vs. ~44,000 Rp/kg actual), though it slightly misaligns timing by one step and overestimates amplitude. Overall, these insights confirm XGBoost's superior ability to model both timing and amplitude on structured series, while highlighting that incorporating exogenous event indicators (e.g., policy changes, supply disruptions) could further improve forecasts.

3.3.1. Red Chili Prices

XGBoost demonstrated excellent predictive performance for red chili prices, achieving an R^2 score of 0.8240, the highest among all commodities tested. This indicates that over 82% of the variance in red chili prices was explained by the model. The MAE of 5393.10, RMSE of 6866.22, and MAPE of 9.77% reflect both low absolute error and high relative accuracy. This strong result implies that red chili prices exhibit regular, learnable patterns that XGBoost can effectively model, likely due to the commodity's response to weather and macroeconomic indicators and the usefulness of lagged pricing features. The model's ability to capture nonlinear interactions and handle multicollinearity further enhances performance over traditional regression models.

3.3.2. Cayenne Chili Prices

The cayenne chili model achieved a moderate R^2 score of 0.4342, with an MAE of 7081.09, RMSE of 9323.34, and MAPE of 11.26%. While the performance is significantly improved compared to Linear Regression ($R^2 \approx 0.31$), it still suggests that a substantial portion of variance remains unexplained. Cayenne chili prices are known for high volatility, possibly influenced by short-term shocks, pest outbreaks, or regional disruptions. XGBoost's nonlinear boosting framework handles these fluctuations better than linear models, but may still fall short without richer exogenous features such as holiday effects, market policies, or localized demand surges.

3.3.3. Shallot Prices

Shallot prices yielded one of the strongest performances, with an R^2 score of 0.7106, MAE of 2744.40, RMSE of 3410.54, and MAPE of 10.13%. These results indicate that the model captured both seasonal and lagged price patterns effectively. XGBoost's success here may be attributed to its strength in modeling complex feature interactions, such as the interplay between rainfall, inflation, and price lags. The relatively low error metrics suggest that shallot prices exhibit a level of predictable structure that can be exploited through advanced models.

3.3.4. Garlic Prices

The model performed least effectively on garlic, with a relatively low R^2 score of 0.2212, and the highest RMSE (6709.68) and MAPE (12.89%) among all commodities. Although better than the Linear Regression baseline (which had a negative R^2), this performance suggests the presence of irregular or unmodeled factors affecting garlic price behavior. Garlic pricing may be subject to non-domestic factors, such as import policy shifts or global supply chains, which were not included in the model's features. While XGBoost is capable of capturing nonlinear structure, its predictive power is inherently limited by the quality and completeness of input variables.

3.4. Model Comparison Analysis

These findings demonstrate that XGBoost consistently outperforms Linear Regression, particularly in capturing nonlinear temporal dependencies. It shows substantial improvements in variance explanation and prediction accuracy, especially for red chili and shallot, where price behavior is well-aligned with the model's assumptions. However, its performance remains sensitive to unobserved external factors, especially for garlic and cayenne chili.

Compared to prior studies, the present research demonstrates notable improvements in forecasting agricultural commodity prices using ensemble learning methods. For instance, Devianto et al. [31] applied a Seasonal ARFIMAX model to forecast red chili prices in Indonesia and reported an R^2 of approximately 0.75 by incorporating exogenous variables such as rainfall and inflation. Similarly, Rana et al. [32] utilized a weather-enhanced ARIMA model for red chili prediction and achieved a comparable R^2 range of 0.72–0.78. While these traditional time series approaches effectively captured seasonality and trend, their capacity to handle nonlinear interactions and lagged dependencies remained limited.

In contrast, our XGBoost model attained a higher R^2 of 0.8240 for red chili and 0.7106 for shallots, indicating a stronger ability to model complex temporal dynamics. For garlic, Support Vector Regression (SVR) approaches in prior work often yielded R^2 values near 0.50 [33], whereas our Random Forest model surpassed this with an R^2 of 0.5365. Cayenne chili remained the most volatile across studies; previous models typically underperformed with R^2 below 0.4 [34], which aligns with our XGBoost result of 0.4342. These comparisons affirm that tree-based ensemble models, particularly XGBoost, are well-suited for capturing nonlinear patterns in price series, especially when combined with lagged variables and macroeconomic indicators. The consistency and superior accuracy of our models suggest their practical advantage for real-world agricultural forecasting.

3.5. Model Performance Analysis

A comparative analysis of the models across all commodities revealed that XGBoost consistently outperformed its counterparts in most cases, particularly where commodity prices exhibited strong temporal and nonlinear patterns. Random Forest delivered strong results for garlic, likely due to its capacity to handle complex feature interactions and outliers. Linear Regression served primarily as a baseline and performed acceptably only for commodities with relatively linear pricing behavior such as red chili.

Table 1 presents the model evaluation metrics. XGBoost achieved the best R^2 scores and lowest MAPE for red chili and shallot, while Random Forest led in garlic price prediction. Linear Regression's performance was notably weak on garlic, returning a negative R^2 value, which indicates that it performed worse than simply predicting the mean price.

Table 1. Model Evaluation Results

Commodity	Commodity	R^2 Score	MAE	RMSE	MAPE
Red Chili	Linear Regression	0.7909	5094.12	7484.48	9.37%
	Random Forest	0.7396	6008.35	8352.51	10.99%
	XGBoost	**0.8240**	5393.10	**6866.22**	**9.77%**
Cayenne Pepper	Linear Regression	0.3154	7834.72	10255.76	13.04%
	Random Forest	0.1229	7924.20	11608.15	13.89%
	XGBoost	**0.4342**	**7081.09**	**9323.34**	**11.26%**
Shallot	Linear Regression	0.4654	3729.25	4635.08	12.55%
	Random Forest	0.5613	3295.49	4198.83	11.87%
	XGBoost	**0.7106**	**2744.40**	**3410.54**	**10.13%**
Garlic	Linear Regression	-0.7366	4198.94	10019.21	13.66%
	Random Forest	0.5365	**2647.65**	5176.24	**9.07%**
	XGBoost	0.2212	3931.98	6709.68	12.89%

As a parametric model assuming linear relationships, Linear Regression performed well in forecasting red chili prices (R^2 = 0.7909, MAPE = 9.37%), likely due to the commodity's seasonal regularity and strong autocorrelation in lagged prices. SHAP interpretation supports this by identifying lagged prices and month indicators as the most influential features, which exhibit near-linear behavior in red chili pricing as noted by Lundberg & Lee [35]. However, its performance deteriorated sharply for garlic (R^2 = -0.7366), where price behavior is governed by more complex factors such as trade policies and storage variability, dynamics that Linear Regression cannot effectively capture.

Random Forest, by aggregating multiple decision trees and accounting for nonlinear interactions, performed well for garlic (R^2 = 0.5365, MAPE = 9.07%). SHAP values for garlic show a more diffuse distribution of importance across features, including lagged price, rainfall, and seasonality, which Random Forest is able to leverage despite weak feature dominance in line with Breiman's work. It also showed solid performance for shallot (R^2 = 0.5613) and red chili (R^2 = 0.7396), highlighting its robustness when dealing with moderately complex temporal structures. However, its underperformance in cayenne chili (R^2 = 0.1229) suggests that even ensemble models struggle when essential predictors like market shocks or socio-cultural demand patterns are absent.

XGBoost, a gradient-boosted ensemble model, achieved the highest performance overall. It produced the top R^2 scores for red chili (0.8240) and shallot (0.7106), and the lowest RMSE and MAPE values for these commodities. SHAP plots confirmed that XGBoost effectively learned from dominant features, especially lagged prices and month whose strong temporal signals matched the seasonal nature of these crops. For cayenne chili, XGBoost still improved accuracy (R^2 = 0.4342), but the moderate SHAP values imply that key external drives such as pest outbreaks or festival-induced demand, were not adequately represented in the model inputs. Although not the best model for garlic, XGBoost delivered competitive results (R^2 = 0.2212), underscoring its generalizability across heterogeneous price patterns.

These findings are further supported by SHAP analysis, which indicated that lagged price and month were consistently the most influential features across all commodities. For red chili and shallot, the SHAP values for these features were both high and directionally consistent, aligning with the superior performance of XGBoost and Random Forest. In contrast, cayenne chili showed lower SHAP feature impacts, reflecting the model's struggle to forecast erratic prices driven by unmodeled exogenous factors. This is in line with previous studies that emphasized the need for incorporating external market signals in time-series forecasting of agricultural commodities.

reinforces the conclusion that model accuracy is closely tied to how well the model captures key feature dynamics. Models like XGBoost and Random Forest, which exploit nonlinear relationships and feature interactions, demonstrate superior performance when dominant predictors are present and meaningful. Conversely, linear models falter under nonlinear or incomplete data environments.

4. Conclusion

Based on the findings, ensemble machine learning models, particularly XGBoost, demonstrated strong performance in forecasting agricultural commodity prices. XGBoost consistently outperformed other models, achieving high R^2 values (e.g., 0.824 for red chili, 0.710 for shallots) and lower error rates, highlighting its ability to model nonlinear patterns. Random Forest also performed well, especially for garlic, while Linear Regression lagged behind due to its limitations with volatile and nonlinear data. The results showed that commodities with clear seasonal and historical trends, like red chili and shallots, were more predictable. In contrast, cayenne chili and garlic posed challenges due to external shocks and global market dependencies. Feature importance analysis identified rainfall, holidays, and lagged prices as key predictors. Although basic interpretability was addressed, future studies should adopt SHAP values for deeper insight. Further research can focus on enhancing prediction accuracy by incorporating additional features (e.g., production volumes, trade policies), using temporal cross-validation, and exploring advanced models like LSTM or RNN to better capture volatility and long-term dependencies in commodity prices.

4.1. Model Limitations

While Random Forest performed best, its interpretability is limited compared to simpler models. Additionally, it may be overfitted when exposed to sparse or unbalanced data. XGBoost, although more efficient in handling sparse data, was computationally heavier. The models also assume stationarity in price trends, which may not hold true in the face of sudden economic shocks (e.g., COVID-19 or geopolitical conflicts). Overall, these results align with previous research, which found that ensemble models such as Random Forest and XGBoost outperform linear models in predicting commodity prices. However, unlike Zhang et al.'s findings, where XGBoost marginally outperformed Random Forest, in this study, Random Forest had the edge, possibly due to differences in data volume or feature engineering approaches.

5. Declarations

5.1. Author Contributions

Conceptualization, E.A.S.B.; methodology, D.R.L., E.A.S.B., F.L.G., and T.M.; data curation, D.R.L.; writing—original draft preparation, D.R.L. and E.A.S.B.; writing—review and editing, D.R.L., E.A.S.B., F.L.G., and T.M.; funding acquisition, F.L.G. and T.M. All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data that support the findings of this study are openly available in Mendeley Data at http://doi.org/10.17632/79bmb9vkwk.3, reference number 79bmb9vkwk.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Acknowledgments

We would like to express our deepest gratitude to all those who have supported us throughout the completion of this research. Special thanks go to Bina Nusantara University, which has provided us with the facilities and academic environment necessary for this work.

5.5. Institutional Review Board Statement

Not applicable.

5.6. Informed Consent Statement

Not applicable.

5.7. Declaration of Competing Interest

The authors declare that there are no conflicts of interest concerning the publication of this manuscript. Furthermore, all ethical considerations, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

6. References

[1] Sumaryanto, N. (2016). Analisis Volatilitas Harga Eceran Beberapa Komoditas Pangan Utama dengan Model ARCH/GARCH. Jurnal Agro Ekonomi, 27(2), 135. doi:10.21082/jae.v27n2.2009.135-163.

[2] Asmarantaka, R. W., & Oktaviani, R. (2009). Gejolak Harga Komoditas Pangan Internasional: Dampak Dan Implikasi Kebijakan Bagi Ketahanan Pangan Indonesia. Jurnal Agribisnis Dan Ekonomi Pertanian, 3(1), 36–49.

[3] I. Cahaya, (2023). Analisis Volatilitas Harga Pangan di Indonesia. Universitas Tidar, Jawa Tengah, Indonesia.

[4] Nonvide, G. M. A., & Akpa, A. F. (2023). Effects of climate change on food crop production in Benin. Climate Change Economics, 14(4), 34–46. doi:10.1142/S2010007823500203.

[5] Nunti, C., Somboon, K., & Intapan, C. (2020). The Impact of Climate Change on Agriculture Sector in ASEAN. Journal of Physics: Conference Series, 1651(1), 108–116. doi:10.1088/1742-6596/1651/1/012026.

[6] Breiman, L. (2001). Random forests. Random Forests, 1–122. Machine Learning, 45(1), 5–32. doi:10.1023/A:1010933404324.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-August-2016, 785–794. doi:10.1145/2939672.2939785.

[8] Ziegel, E. R. (2003). The Elements of Statistical Learning. Technometrics. 45(3). doi:10.1198/tech.2003.s770.

[9] Riando, D., & Afiyati, A. (2024). Implementasi Algoritma XGBoost untuk Memprediksi Harga Jual Cabai Rawit di DKI Jakarta. Eduvest - Journal of Universal Studies, 4(9), 7877–7889. doi:10.59188/eduvest.v4i9.3784.

[10] Muchtar, I. R., & Afiyati, A. (2024). Comparison of Linear Regression and Random Forest Algorithms for Premium Rice Price Prediction (Case Study: West Java). Jurnal Indonesia Sosial Teknologi, 5(7), 3122–3132. doi:10.59141/jist.v5i7.1184.

[11] Rayadin, M. A., Musaruddin, M., Saputra, R. A., & Isnawaty, I. (2024). Implementasi Ensemble Learning Metode XGBoost dan Random Forest untuk Prediksi Waktu Penggantian Baterai Aki. BIOS: Jurnal Teknologi Informasi Dan Rekayasa Komputer, 5(2), 111–119. doi:10.37148/bios.v5i2.128.

[12] Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., & Liew, X. Y. (2021). Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques. Advances in Science, Technology and Engineering Systems Journal, 6(4), 376–384. doi:10.25046/aj060442.

[13] Ismanto, E., & Novalia, M. (2021). Komparasi Kinerja Algoritma C4.5, Random Forest, dan Gradient Boosting untuk Klasifikasi Komoditas. Techno.Com, 20(3), 400–410. doi:10.33633/tc.v20i3.4576.

[14] Tran, N.-Q., Nguyen Ngoc, T., Tran, Q., Felipe, A., Huynh, T., Tang, A., & Nguyen, T. (2023). Predicting Agricultural Commodities Prices with Machine Learning: A Review of Current Research. School of Science, Engineering, and Technology, RMIT University Vietnam. 1(1): 1–7.

[15] Xiao, Y., & Jin, Z. (2021). The Forecast Research of Linear Regression Forecast Model in National Economy. OALib, 08(08), 1–17. doi:10.4236/oalib.1107797.

[16] Lubis, A. H., & Rizky Pulungan, M. (2024). Prediksi Harga Pangan di Tengah Isu Ketidakpastian Global Menggunakan Metode Regresi Linear dan Regresi Polinomial. Jurnal Fasilkom, 14(2), 404–409. doi:10.37859/jf.v14i2.6996.

[17] Putatunda, S., & Rama, K. (2019). A Modified Bayesian Optimization based Hyper-Parameter Tuning Approach for Extreme Gradient Boosting. 2019 15th International Conference on Information Processing: Internet of Things, ICINPRO 2019 - Proceedings, Bangalore, India (20-22 December 2019). doi:10.1109/ICInPro47689.2019.9092025.

[18] Salem, F. M. (2021). Gated RNN: The Gated Recurrent Unit (GRU) RNN. Recurrent Neural Networks, 85–100. doi:10.1007/978-3-030-89929-5_5.

[19] N. R. Timisela, (2020). The Analysis on Formation of Prices of Cayenne and Shallot Commodities at Retail Levels in Ambon City. Jurnal Budidaya Pertanian, 16(1), 31–41.

[20] Wang, Y., & Guo, Y. (2020). Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Communications, 17(3), 205–221. doi:10.23919/JCC.2020.03.017.

[21] Gupta, I., Mittal, H., Rikhari, D., & Singh, A. K. (2022). MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of a Day. arXiv 2022, arXiv:2203.05835v1. Available online: https://arxiv.org/abs/2203.05835.

[22] Suryani, E., Hendrawan, R. A., Mulyono, T., & Dewi, L. P. (2014). System dynamics model to support rice production and distribution for food security. Jurnal Teknologi (Sciences and Engineering), 68(3), 45–51. doi:10.11113/jt.v68.2928.

[23] Liu, G. (2022). Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. Scientific Programming, 2022(2), 1–8. doi:10.1155/2022/5750354.

[24] Meenal, R., Michael, P. A., Pamela, D., & Rajasekaran, E. (2021). Weather prediction using random forest machine learning model. Indonesian Journal of Electrical Engineering and Computer Science, 22(2), 1208–1215. doi:10.11591/ijeecs.v22.i2.pp1208-1215.

[25] Saadah, S., & Salsabila, H. (2021). Prediksi Harga Bitcoin Menggunakan Metode Random Forest (Studi Kasus: Data Acak Pada Awal Masa Pandemic Covid-19). Jurnal Komputer Terapan, 7(1), 24–32. https://jurnal.pcr.ac.id/index.php/jkt/

[26] Sinambela, R. S., Ula, M., & Ulva, A. F. (2024). Prediksi Harga Emas Menggunakan Algoritma Regresi Linear Berganda Dan Support Vector Machine (SVM). Jurnal Sistem Dan Teknologi Informasi (JustIN), 12(2), 253. doi:10.26418/justin.v12i2.73386.

[27] Vuong, P. H., Dat, T. T., Mai, T. K., Uyen, P. H., & Bao, P. T. (2022). Stock-price forecasting based on XGBoost and LSTM. Computer Systems Science and Engineering, 40(1), 237–246. doi:10.32604/CSSE.2022.017685.

[28] Jabeur, S. Ben, Mefteh-Wali, S., & Viviani, J. L. (2024). Forecasting gold price with the XGBoost algorithm and SHAP interaction values. Annals of Operations Research, 334(1–3), 679–699. doi:10.1007/s10479-021-04187-w.

[29] Lemmens, A., & Croux, C. (2006). Bagging and boosting classification trees to predict churn. Journal of Marketing Research, 43(2), 276–286. doi:10.1509/jmkr.43.2.276.

[30] Barbur, V. A., Montgomery, D. C., & Peck, E. A. (1994). Introduction to Linear Regression Analysis. The Statistician. 43(2). John Wiley and Sons. doi:10.2307/2348362.

[31] Devianto, D., Wirawan, E., & Sukirno, S. (2024). Time series modeling using SARFIMAX on red chili prices in West Java. Indonesian Journal of Statistics and Its Applications, 8(1), 45–54.

[32] Rana, R., Kusumawardani, H., & Mulyani, A. (2024). ARIMA with weather-based features for red chili price forecasting in North Sumatra. Journal of Agroinformatics, 6(2), 93–100.

[33] Nugroho, M. A., & Ramadhan, D. (2023). Support vector regression for garlic price prediction in Java. International Conference on Data Science and Engineering (ICDSE, 27–32.

[34] Yuliana, L., & Prasetyo, B. (2023). Forecasting challenges in cayenne chili prices using traditional models. AgroTech Journal, 9(3), 12–19.

[35] Lundberg, S. M., & Lee, S.I. (2023). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.