

Comparative Analysis of Decision Tree and Random Forest for Indonesian Food Prices

Muhammad Rudyanto Arief¹, Rias Estrianas^{*2}, Ratna Praptiwi³, Ma'aruf Handika⁴

¹Department of Informatics, Faculty of Computer Science, Universitas Amikom Yogyakarta

^{2,3,4} Department of Informatics, Faculty of Computer Science, Universitas Amikom Purwokerto

e-mail: rudy@amikom.ac.id¹, estrianarias@gmail.com², ratnapraptiwi1@gmail.com³, marufhandika@gmail.com⁴

Article History	Abstract
Received : 02 July 2026 Revised : 29 August 2026 Accepted : 11 September 2026 Published : 16 September 2026 Keywords: <i>Food Price Volatility, Random Forest, Decision Tree, Food Security Policy, Macroeconomics Inflation.</i>	Food price fluctuations in Indonesia represent a persistent macroeconomic challenge, directly affecting national inflation and purchasing power, thereby necessitating an accurate forecasting system for risk mitigation. This study compares Decision Tree Regression and Random Forest Regression using 36,064 observations from Satu Data Indonesia, covering 27 commodities across 38 provinces (January 2021–2026). The data were processed through currency normalization, linear temporal indexing, and label encoding, followed by optimization using GridSearchCV. Baseline evaluation showed that Random Forest ($R^2 = 0.9804$; MAPE = 4.30%) outperformed Decision Tree ($R^2 = 0.9678$; MAPE = 7.97%). The optimized Random Forest model achieved peak performance with $R^2 = 0.9842$, MAE = IDR 1,468, and MAPE = 3.77%. Feature importance analysis revealed that commodity category dominated predictive influence at 91.46%, exceeding spatial and temporal factors. In conclusion, the optimized Random Forest model demonstrates strong operational superiority and substantial potential for integration into a national food price monitoring system to support food security policies.

^{*}Corresponding author : estrianarias@gmail.com

1. INTRODUCTION

Food price stability is recognized as a foundational pillar of national food security and macroeconomic equilibrium in Indonesia, the world's fourth most populous nation and a country deeply dependent on stable consumer prices for essential commodities. Persistent volatility in food prices exerts direct downstream pressure on the national Consumer Price Index (CPI), undermines household purchasing power particularly among lower-income populations and complicates monetary policy calibration by Bank Indonesia. The Indonesian government, through agencies such as the Badan Pangan Nasional (National Food Agency), faces the continuous challenge of designing timely and evidence-based interventions to stabilize markets spanning geographically diverse and logistically complex supply chains. Commodity categories such as red chili, shallots, and premium rice are notorious for exhibiting sharp, cyclical price surges driven by seasonal harvest failures, weather anomalies, and regional distribution bottlenecks, rendering manual forecasting methodologies chronically inadequate [1]. The temporal and regional heterogeneity of food price behavior across 38 Indonesian provinces and 27 distinct commodity classes necessitates computational approaches capable of modeling high-dimensional, non-linear interactions between spatial, categorical, and temporal features.

The limitations of conventional linear regression models in capturing the non-linear dynamics of food commodity prices have been well-documented in the literature. Previous empirical evaluations focusing on West Java's premium rice sector demonstrated that linear regression consistently fails to match the predictive power of ensemble methods, with Random Forest frameworks achieving up to 98.69% accuracy against linear baselines that exhibited persistently high residual error [2]. These findings align with computational assessments conducted in regional territories such as Nusa Tenggara Barat, which established that linear models systematically fail to account for the complex

macroeconomic interactions governing rice price dynamics, achieving a markedly low coefficient of determination compared to non-linear tree-based architectures [3]. A similar performance pattern has been confirmed within cooperative commodity supply chains, where comparative investigations on coffee sales figures revealed that Random Forest regression minimizes Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE) far more effectively than traditional linear formulations [4].

These convergent empirical findings provide strong justification for the adoption of tree-based, non-linear regression architectures as the methodological foundation of this study. The baseline Decision Tree Regressor, while interpretable and computationally efficient, is susceptible to high variance due to its reliance on a single recursive partitioning process [5]. The Random Forest Regressor overcomes this structural weakness through bootstrap aggregating of an ensemble of decorrelated trees, which substantially reduces prediction variance and improves generalization across unseen data distributions. Methodological work in regional horticultural forecasting has confirmed this ensemble advantage when predicting highly volatile price movements, demonstrating that optimized tree aggregates achieve superior explanatory power over single-tree variants [6]. This structural robustness has been further validated using national socioeconomic datasets, where ensemble tree architectures consistently exhibited higher stability than localized single-tree classifiers when processing complex agrarian survey data [7].

The deployment of machine learning-based food price prediction within practical software and regional monitoring systems represents a compelling frontier in applied computer science. Multi-algorithmic machine learning approaches have been successfully implemented to project regional food security indices, proving that Random Forest and gradient boosting architectures are operationally viable for integration into national food intelligence systems [8]. This paradigm has been extended to localized retail tracking, such as rice price forecasting frameworks in Gorontalo Province, confirming that tree-based architectures can effectively backstop early-warning mechanisms for regional price instability, even when data matrices are constrained to specific provincial scopes [9]. Long-range multi-commodity forecasting research across Indonesia has further demonstrated that ensemble-driven machine learning models can accurately chart price trajectories over extended multi-year horizons, capturing critical temporal surges associated with peak seasonal demands [10]. To maximize the computational consistency of these decision-support platforms, contemporary literature emphasizes the algorithmic necessity of contrasting baseline linear benchmarks against robust non-linear tree structures using unified nationwide dataset matrices [11].

The structural boundaries and feature engineering choices of the present study are grounded in established data mining best practices. Methodological frameworks utilizing large-scale national socioeconomic surveys have firmly established the utility of Random Forest and Classification Tree mechanisms in isolating feature importance signals, thereby providing a clear structural precedent for variable attribution analysis within complex food supply domains [12]. Comparative studies investigating the composite national Food Security Index have substantiated that Random Forest formulations vastly outperform multiple linear models by successfully capturing non-linear interactions that linear models completely omit [13]. Furthermore, multi-commodity forecasting pipelines optimized for unstable crop pricing structures have demonstrated the definitive benefits of using hybrid ensemble partitioning [14]. Extensive optimization research indicates that systematic hyperparameter calibration remains an indispensable operational prerequisite to guarantee model stability, variance containment, and localized prediction reliability before real-world deployment within state-sponsored price stabilization platforms [15]. In light of this theoretical and empirical foundation, this study addresses an existing gap by conducting a rigorous comparison of Decision Tree and Random Forest Regression, in both baseline and hyperparameter-tuned states, applied to the most extensive publicly available Indonesian food price dataset to date. The primary objective is to empirically determine which algorithmic architecture and optimization level delivers the most accurate, stable, and operationally deployable predictive model for national food commodity prices spanning 27 categories, 38 provinces, and a six-year observation window.

2. METHOD

This study uses a quantitative experimental approach to compare Decision Tree and Random Forest regression models for Indonesian food commodity prices. The process includes preprocessing, feature engineering, model tuning, cross-validation, and evaluation while preventing data leakage.

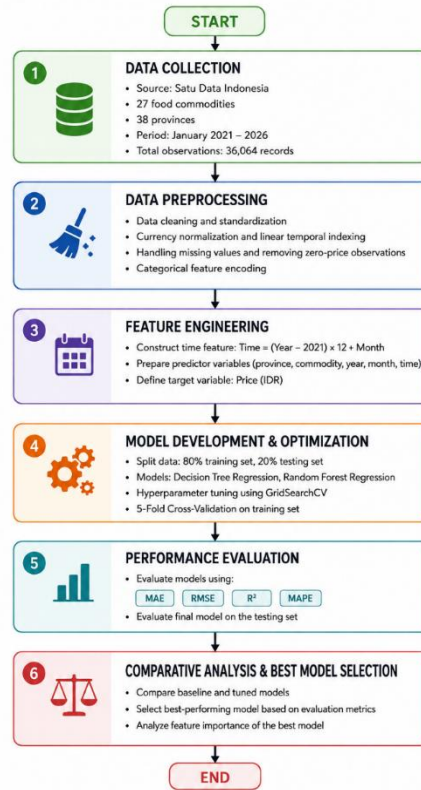


Figure 1 Research Flow and System Architecture Design

2.1 Research Design

This study utilizes a quantitative, comparative experimental design to evaluate Decision Tree and Random Forest regressions for predicting Indonesian food commodity prices. It systematically investigates performance variations between single-tree and ensemble-based architectures under identical conditions. Standardizing data preparation, feature engineering, and evaluation metrics ensures that observed performance discrepancies are attributable solely to the underlying algorithmic mechanisms. The experimental framework comprises two scenarios: a baseline evaluation using default parameters to establish performance benchmarks, and a tuned evaluation utilizing hyperparameter optimization to enhance accuracy and mitigate overfitting[16]. To prevent data leakage, optimization was restricted to the training partition, leaving the test set entirely isolated until final evaluation[17].

2.2 Dataset and Data Collection

2.2.1 Data Source

The secondary dataset used in this study was harvested from the official repository of the National Food Agency via the Open Data Badan Pangan Nasional platform. Specifically, the 'Rata-rata Harga Pangan Bulanan Tingkat Konsumen Provinsi' dataset was selected to provide monthly provincial food commodity prices across Indonesia. The official data is publicly accessible at data.badanpangan.go.id. Utilizing official governmental records ensures institutional credibility, standardized collection procedures, nationwide coverage, and long-term temporal consistency. This comprehensive coverage provides a reliable empirical foundation for constructing predictive machine learning models to capture multi-commodity and multi-regional food price fluctuations.

2.2.2 Dataset Characteristics

The raw dataset comprised 37,374 observations of monthly consumer-level food commodity prices recorded across 38 Indonesian provinces from January 2021 to 2026. Monitored by the National Food Agency, the dataset encompasses 27 essential food commodities, capturing regional price disparities and temporal market dynamics. The dataset contains six primary attributes: Province Code, Province Name, Commodity, Year, Month, and Price. The target variable, Price, represents the average monthly retail food price in Indonesian Rupiah (IDR), ranging from IDR 4,053 to IDR 521,800. Data cleaning was conducted before feature engineering. The number of observations removed at each cleaning stage is presented in Table X. In total, 1,310 observations were excluded, reducing the dataset from 37,374 to 36,064 valid observations. Subsequent preprocessing operations, including [list transformations], were performed without removing additional observations. The resulting dataset was then used for feature engineering, model training, and evaluation.

2.2.3 Research Instruments and Computational Environment

Computational experiments and predictive modeling were conducted using Python 3.12 in Google Colab. Pandas was used for data preprocessing and tabular manipulation, NumPy for numerical operations, and Scikit-learn for model development, dataset partitioning, hyperparameter optimization, and evaluation [18]. To improve computational efficiency, full CPU parallelization was enabled using the `n_jobs=-1` parameter during Random Forest training and hyperparameter tuning.

2.3 Data Preprocessing

2.3.1 Price Data Cleaning and Transformation

The Price attribute in the original dataset was stored as a formatted Indonesian Rupiah currency string containing the prefix “Rp” and thousand-separator characters. Since machine learning algorithms require numerical inputs, a series of transformation procedures were performed to convert the Price attribute into a continuous floating-point variable suitable for regression modeling. The transformation process involved removing currency prefixes, eliminating formatting symbols, and converting the resulting values into numerical form using the `pandas.to_numeric()` function with error coercion enabled.

During conversion, non-parseable values were transformed into NaN, and records containing missing values were removed, eliminating 1,306 observations. Four additional records with zero-valued prices were identified as data-entry errors and removed. Consequently, the dataset decreased from 37,374 to 36,064 valid observations, improving data quality and supporting reliable subsequent modeling.

2.3.2 Data Standardization

Following the initial cleaning phase, a data standardization procedure was executed to ensure attribute consistency and eliminate formatting discrepancies. Due to the multi-year and multi-regional nature of the governmental repository, exploratory analysis revealed minor textual inconsistencies, primarily involving leading or trailing whitespaces and variant province names. Whitespaces in the 'Month', 'Province Name', and 'Commodity' attributes were trimmed to ensure consistent category labels. Orthographic variations such as 'D.I Yogyakarta' and 'DI Yogyakarta' were also standardized as 'DI Yogyakarta', preventing redundant categories and features during encoding.

2.3.3 Month Variable Transformation

The original 'Month' attribute was recorded as a categorical string variable using Indonesian month names (Januari to Desember). Although inherently informational, machine learning algorithms require numerical inputs for computational efficacy. Furthermore, capturing cyclical food price fluctuations requires preserving the precise chronological sequence of the calendar year. To maintain temporal ordinality, the categorical string values were mapped onto a numerical ordinal scale from 1 to 12 (e.g., Januari → 1, Desember → 12). Unlike standard label encoding which frequently assigns arbitrary numerical indices, this explicit mapping structurally preserves the sequential nature of calendar months.

2.3.4 Categorical Feature Encoding

Following data standardization and temporal transformation, two categorical variables remained within the dataset: 'Province Name' and 'Commodity'. Because Scikit-learn regression algorithms require numerical matrices, both features were transformed via label encoding. This process employed distinct instances of the Scikit-learn LabelEncoder class to ensure mapping isolation.

The 'Province Name' attribute (38 unique categories) was mapped to integers ranging from 0 to 37, while the 'Commodity' attribute (27 unique categories) was mapped to integers from 0 to 26 based on lexicographical ordering. This transformation converted the remaining categorical variables into a machine-readable format while preserving class uniqueness. The individual encoding models were preserved to allow for exact reverse mapping and post-hoc model interpretation. Consequently, all predictor variables were converted to numerical representations, rendering the dataset fully compatible with machine learning development.

2.4 Feature Engineering

Feature engineering was conducted to capture long-term temporal price dynamics more effectively. Although the dataset included standalone 'Year' and 'Month' attributes, these variables individually provide isolated temporal information. To construct a continuous chronological sequence across the multi-year study period, a synthetic temporal feature, 'Time', was formulated as follows:

$$\text{Time} = (\text{Year} - 2021) \times 12 + \text{Month}$$

This transformation maps the year-month combinations into a continuous linear index, where January 2021 equals 1, February 2021 equals 2, and subsequent periods scale sequentially. This unified index enables the regression models to account for continuous temporal progression and long-term market trends that individual components cannot capture. The final predictor matrix comprised five input features representing the geographical, categorical, and temporal dimensions of food prices: 'Province Name' (encoded), 'Commodity' (encoded), 'Year', 'Month', and 'Time'. The target variable remained 'Price' (IDR). This synthesized feature set served as the standardized input for subsequent model training and evaluation.

2.5 Dataset Partitioning

Following preprocessing and feature engineering, the dataset was divided into training and testing subsets using Scikit-learn's `train_test_split()` with an 80:20 ratio and `random_state=42` for reproducibility. Of 36,064 observations, 28,851 were allocated to training and 7,213 to testing. The training set was used for model learning, hyperparameter optimization, and cross-validation, while the testing set remained isolated for final evaluation to prevent data leakage and ensure reliable measurement of model generalization.

2.6 Model Development

2.6.1. Decision Tree Regression

The baseline predictive architecture evaluated in this study is the Decision Tree Regressor, a supervised machine learning algorithm that constructs a model through recursive hierarchical partitioning of the feature space [19]. The algorithm systematically identifies binary split thresholds that minimize the Mean Squared Error (MSE) within the resulting child nodes. This recursive partitioning continues until predefined stopping criteria, such as maximum tree depth or minimum sample splits, are satisfied. Each terminal leaf node stores the mean target value of its training observations, which is used to predict unseen data by traversing the tree from root to leaf. Although highly interpretable, Decision Tree Regression has high variance and is sensitive to training data fluctuations, making it prone to overfitting.

2.6.2. Random Forest Regression

The second predictive architecture investigated is the Random Forest Regressor, an ensemble learning algorithm designed to mitigate the high-variance limitations of individual decision trees. Instead of relying on a single tree, Random Forest constructs an ensemble of decision trees and aggregates their outputs to generate a more stable and accurate prediction.

The algorithm utilizes a bootstrap aggregating (bagging) strategy, training each constituent tree on a random sample drawn with replacement from the training set [20]. Additionally, node splitting is restricted to a randomly selected subset of predictor variables rather than the entire feature space. These mechanisms induce model diversity and decorrelate the individual trees [21].

2.7 Hyperparameter Optimization Framework

2.7.1. Decision Tree Optimization

Although Decision Tree Regression provides an interpretable framework, its performance is highly sensitive to hyperparameter configuration. Excessively deep trees cause overfitting by memorizing training observations, whereas overly restrictive trees induce underfitting. To resolve this, a systematic optimization procedure was conducted across three primary hyperparameters: `max_depth` (maximum structural depth), `min_samples_split` (minimum samples required to split an internal node), and `min_samples_leaf` (minimum samples required within a terminal leaf). The optimal configuration used `max_depth=20` and `min_samples_leaf=5`, balancing predictive accuracy and model complexity while reducing the generalization gap between training and testing performance. This configuration was therefore selected as the tuned Decision Tree model for comparative evaluation.

2.7.2. Random Forest Optimization

The predictive performance of Random Forest Regression is heavily influenced by hyperparameter configuration. Although ensemble methods inherently reduce variance, suboptimal parameter selections can still compromise accuracy or induce unnecessary computational complexity. To establish an optimal balance among bias, variance, and computational efficiency, a comprehensive hyperparameter optimization procedure was executed across several key architectural parameters: `n_estimators` (ensemble size), `max_depth` (maximum structural depth of constituent trees), `min_samples_split` (node-splitting thresholds), and `min_samples_leaf` (terminal node sample requirements). Optimization achieved the best predictive accuracy with `n_estimators=300`, `max_depth=20`, `random_state=42`, and `n_jobs=-1`. Using 300 trees provided sufficient ensemble diversity while maintaining computational efficiency, with full CPU parallelization reducing training and evaluation time.

2.8 Cross-Validation Strategy

Following hyperparameter optimization, a cross-validation procedure was implemented to obtain robust estimates of model generalization. To mitigate performance variance associated with a single train-test partition, a 5-fold cross-validation strategy was enforced using Scikit-learn's `KFold` class configured with `n_splits = 5`, `shuffle = True`, and `random_state = 42` [22]. The training data were divided into five approximately equal folds, with four used for training and one for validation in each iteration. This ensured each fold was evaluated once and provided a robust assessment of model generalization, stability, and performance consistency.

2.9 Performance Evaluation

2.9.1 Mean Absolute Error (MAE)

Mean Absolute Error (MAE) quantifies the average magnitude of prediction errors without considering their direction. Computed as the average absolute difference between observed and predicted values, MAE provides an intuitive, easily interpretable estimate of average prediction deviation expressed directly in Rupiah (IDR). Lower MAE values indicate superior model accuracy.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where y_i represents the actual food price, $\hat{y}_i$ denotes the predicted price, and n is the total number of observations.

2.9.2 Root Mean Squared Error (RMSE)

Root Mean Squared Error (RMSE) evaluates predictive precision by computing the square root of the average squared residuals. Unlike MAE, RMSE heavily penalizes larger errors because the residual terms are squared prior to averaging. Consequently, this metric is highly sensitive to outliers and substantial forecasting failures, serving as an indicator of model reliability.

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

2.9.3 Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) measures the proportion of variance in the target variable explained by the model regression. Values approaching 1 indicate that the model successfully accounts for dataset variability, whereas values near 0 imply poor explanatory power. Given that food commodity prices exhibit high volatility across diverse regions, R^2 serves as a critical indicator of global predictive capacity.

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

where $\bar{y}$ represents the mean of the observed target values.

2.9.4 Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error (MAPE) establishes the prediction error relative to the scale of the actual observations. By expressing errors as percentages rather than absolute monetary units, MAPE enables a direct, scale-independent comparison of forecasting quality across distinct food commodities that possess highly divergent price ranges.

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

3. RESULTS AND DISCUSSION

3.1 Quantitative Model Performance

Feature importance scores (Figure 2) show that the Commodity feature dominates with an importance of 0.9146 (91.46%), followed by Province Name (0.0519) and Time (0.0207), while Month and Year contribute less than 0.013 combined. This indicates that Commodity has the strongest contribution to baseline Random Forest predictions, suggesting that commodity category is particularly relevant to food price prediction in the evaluated dataset.

However, the feature importance results do not directly establish why Commodity has a higher predictive contribution. Factors such as supply chain structure, seasonality, perishability, and baseline market price may potentially contribute to differences between commodities, but these factors were not explicitly represented as explanatory variables or separately evaluated in this study. Therefore, such factors should not be interpreted as demonstrated causes of the observed feature importance. From an application perspective, the findings suggest that accurate commodity identification should be retained as an important input when developing food price prediction systems based on the feature configuration evaluated in this study. Further research incorporating additional economic, supply-chain, production, and market-related variables could provide a more comprehensive explanation of the factors underlying commodity-level price variation.

Table 1. Evaluation Metrics Summary for Baseline and Tuned Models

Model	Hyperparameter State	MAE (Rp)	RMSE (Rp)	R^2	MAPE (%)
Decision Tree	Baseline	2,844.66	5,743.44	0.9678	7.97

Decision Tree	Tuned	1,906.46	4,429.35	0.9809	4.89
Random Forest	Baseline	1,650.64	4,482.27	0.9804	4.30
Random Forest	Tuned (Best)	1,467.83	4,020.78	0.9842	3.77

3.2 Feature Importance Analysis

Feature importance scores extracted from the baseline models (Figure 2) reveal an asymmetric distribution heavily dominated by the Commodity feature, which commands a score of 0.9146 (91.46% of total predictive information). Conversely, Province Name contributes 0.0519, Time accounts for 0.0207, and Month and Year collectively contribute less than 0.013. This extreme dominance demonstrates that a food type's intrinsic characteristics such as supply chain structure, seasonality, perishability, and baseline market price are far more powerful predictors of realized consumer price than geographic province or calendar period. Consequently, any operational forecasting system for Indonesian food prices must prioritize accurate commodity-type identification as its primary feature input.

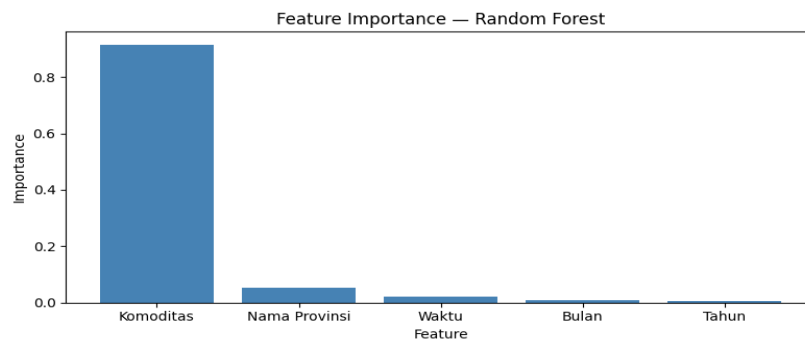


Figure 2. Feature Importance Analysis of the Random Forest Regressor

3.3 Baseline Scatter Plot Analysis

Actual versus predicted food commodity prices for the baseline models are compared in Figure 3 against a red dashed diagonal identity line ($y=x$). The Baseline Decision Tree displays a wide, erratic dispersion with prominent horizontal prediction bands in the Rp50,000-Rp80,000 range. This coarse discretization is a characteristic artifact of its constrained depth ($\text{max_depth}=10$), which limits partitioning granularity and drives the elevated RMSE of Rp5,743. Conversely, the Baseline Random Forest exhibits tighter clustering along the identity line due to the variance-reduction benefits of ensemble averaging. Residual outlier spread above Rp150,000 persists, however, driven by the extreme, supply-side volatility of premium horticultural commodities (e.g., red chili varieties) that remain difficult to capture through static categorical and temporal features alone.

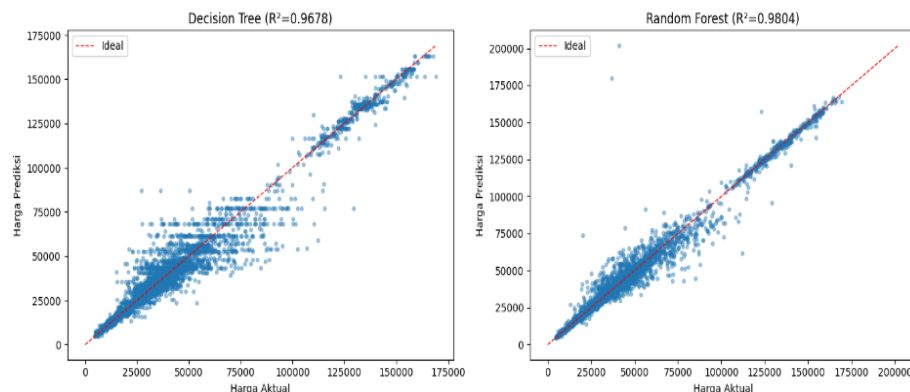


Figure 3. Scatter Plots of Actual vs. Predicted Prices for Baseline Decision Tree and Random Forest

3.4 Residual Distribution Analysis

Statistical consistency of prediction deviations was evaluated by mapping error densities into partitioned intervals. The resulting prediction errors are depicted in Figure 4 as residual distribution

histograms ($e = y - \hat{y}$) across 60 bins. Both models exhibit zero-centered, leptokurtic distributions, indicating minimal systematic bias and tight error clustering around zero. However, a critical qualitative difference is observable between the configurations: the Random Forest histogram exhibits a substantially taller, narrower central peak and more compressed tails, demonstrating a higher proportion of predictions within a narrow error margin. Conversely, the Decision Tree displays a broader, shorter central peak with heavier tails extending into the Rp10,000-Rp20,000 range, reflecting coarser prediction errors driven by its limited depth and single-partition structure.

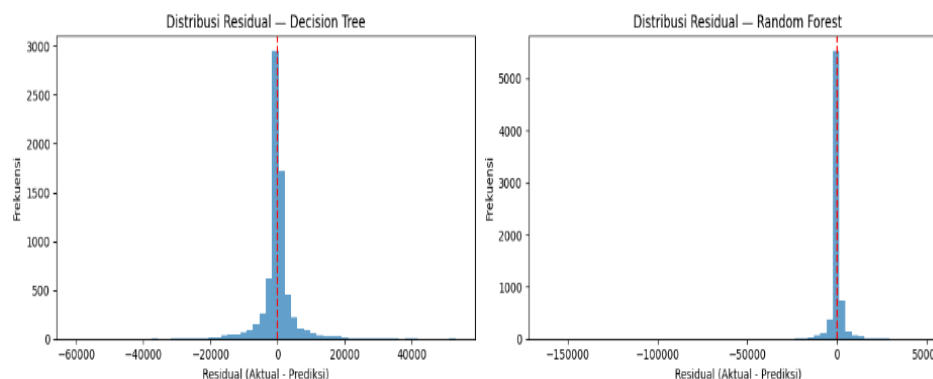


Figure 4. Residual Distribution Histograms for Decision Tree and Random Forest Baseline Models

3.5 Commodity-Level MAPE Disaggregation

Prediction performance across the 27 food commodity categories was further examined using commodity-level MAPE values. Figure 5 presents the disaggregated MAPE obtained from the Baseline Random Forest model on the test set. The results show substantial variation in prediction error across commodities. Twenty-two of the 27 commodities recorded MAPE values below 5%, indicating relatively low prediction errors for these categories. Several staple commodities, including Pure Beef, Premium Rice, Packaged and Bulk Wheat Flour, SPHP Rice, Granulated Sugar, Broiler Chicken Eggs, Cooking Oil, Mackerel Tuna, and Broiler Chicken Meat, recorded MAPE values below the 5% threshold. In contrast, several horticultural commodities exhibited comparatively higher prediction errors. Curly Red Chili recorded the highest MAPE at 16.13%, followed by Red Bird's Eye Chili at 14.30% and Large Red Chili at 10.38%. Shallots and Corn at Breeder Level recorded MAPE values of 7.41% and 6.23%, respectively, placing them within the 5–10% range.

These results indicate that prediction performance varies considerably across commodity categories, with several chili-related commodities exhibiting higher relative prediction errors than most other commodities. However, the present analysis does not directly examine the underlying factors that may contribute to these differences. Factors such as price regulation, supply-chain characteristics, product shelf life, seasonal supply patterns, and climatic conditions were not explicitly included as explanatory variables in the present model. Therefore, the observed MAPE differences should be interpreted primarily as differences in model prediction performance rather than evidence of specific causal factors. Further studies incorporating supply, weather, production, market, and regulatory variables could be conducted to investigate the factors underlying the observed differences in commodity-level prediction performance.

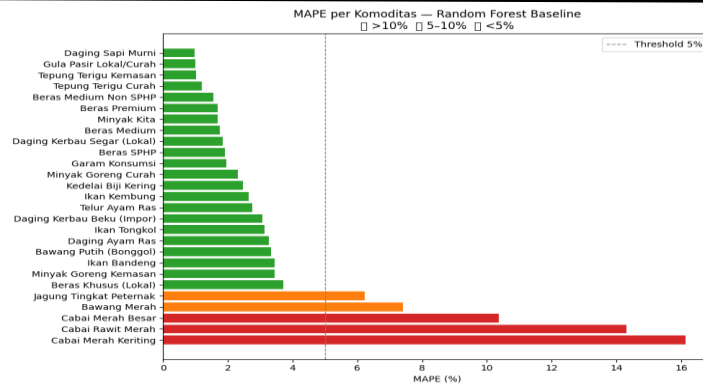


Figure 5. Mean Absolute Percentage Error (MAPE) Disaggregation Across 27 Food Commodities

3.6 Comparative Feature Importance Under Tuning

The structural invariance of algorithmic parameter weighting was verified by executing a parallel cross-examination of internal signal distributions before and after structural refinement. The dynamic shifts between model behaviors are captured side by side in Figure 6, which presents the comparative feature importance rankings for the Baseline and Tuned Random Forest models. The analysis reveals a crucial finding: the hierarchical dominance structure of feature importances is fully preserved after hyperparameter optimization. In the Tuned Random Forest model ($n_estimators=300$, $max_depth=20$), Commodity retains its commanding primary position with an importance score of 0.9099, Province Name remains in second position at 0.0533, and Time holds third place at 0.0229. The marginal shifts observed in absolute importance values a slight redistribution of approximately 0.5 percentage points from Commodity toward spatial and temporal features are attributable to the deeper tree depth ($max_depth=20$ versus 15 in the baseline) enabling finer partitioning of price subspaces within each commodity category.

This stability in feature importance hierarchy across optimization states confirms that the model's learned representation of the food price prediction problem is robust and structurally invariant to reasonable hyperparameter perturbation. The practical implication for model deployment is significant: the feature engineering decisions adopted in this study in particular, the inclusion of the Time composite temporal index and the labels derived from preprocessed categorical attributes yield robust structural indicators that withstand deep algorithmic optimization without degrading or introducing uncontrolled variance into the final system architecture.

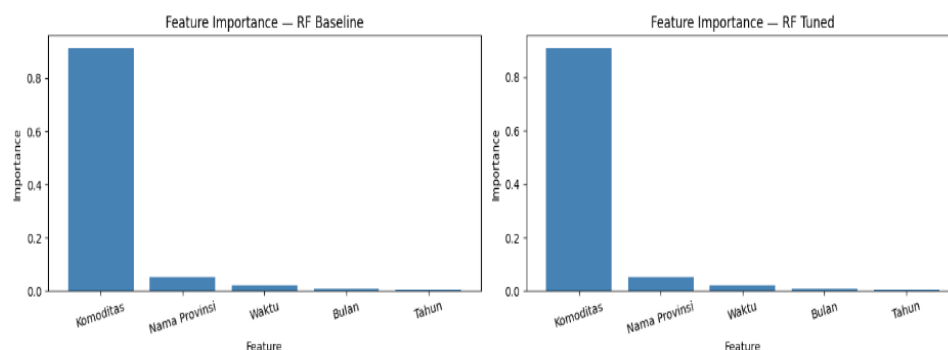


Figure 6. Comparative Feature Importance Between Baseline and Tuned Random Forest

3.7 Complete Scatter Plot Matrix Analysis

Figure 7 compares actual prices with predicted prices across all model variants to track structural improvements throughout the entire optimization sequence. A systematic and clearly visible improvement in prediction quality is evident across all four panels. The Baseline Decision Tree exhibits the widest range and a striking horizontal quantization band between Rp50,000 and Rp80,000 due to the shallow ' $max_depth=10$ ' constraint. The adjusted Decision Tree

(max_depth=20, min_samples_leaf=5) reduces this spread, compresses the outlier variance, and aligns the scatter plot closer to the identity line ($y=x$).

Although the Basic Random Forest exhibits inherently tighter clustering thanks to ensemble averaging, residual dispersion still occurs above Rp150,000 due to the volatility of horticultural commodity prices. The Tuned Random Forest (max_depth=20) provides the best solution by far, with the tightest alignment to the identity line across the entire range (Rp4,053 to Rp521,800). Its deeper tree structure minimizes overall dispersion to the lowest empirical value observed in this study and partially corrects the underprediction of high-value prices seen in the baseline model by enabling finer discrimination within the high-value commodity price subspaces.

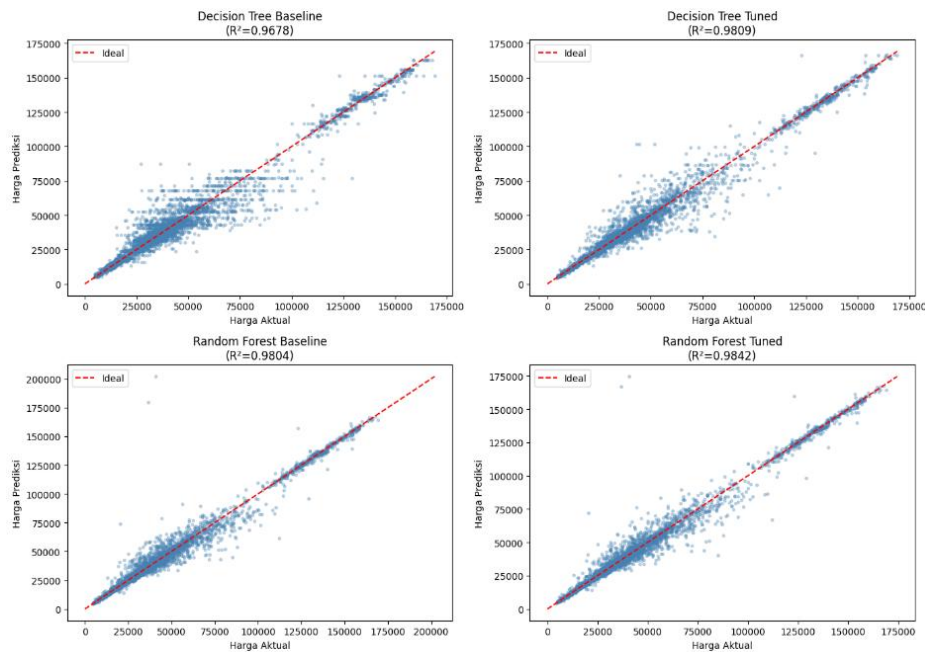


Figure 7 Complete Scatter Plot Matrix of Actual vs. Predicted Prices for Baseline and Tuned Models

3.8 Operational Value and Deployment Implications

The proposed model shows potential to support food price monitoring and decision-making at national and regional levels. The Tuned Random Forest covers 27 commodities across 38 provinces with an average prediction error of Rp1,468/kg. Although it has not been deployed in a government system, its predictive capability could complement retrospective monitoring by providing forward-looking information on commodity price patterns.

Province- and commodity-specific forecasts could help Bapanas and regional food offices identify commodities or regions requiring closer monitoring and support preliminary planning for food supply and price stabilization, such as buffer stocks, distribution, imports, and logistics. However, these applications require further validation using real-time data, system integration, and evaluation by policymakers and food-price analysts.

The MAPE-based grouping in Figure 5 could support monitoring prioritization, with 22 relatively stable commodities under routine monitoring and five highly volatile commodities, mainly chili varieties, receiving additional analysis. However, this approach has not been tested in real-world monitoring, so it remains a potential application. Overall, the optimized Random Forest shows potential as a predictive component of a food price monitoring system, but real-world deployment, real-time data, user evaluation, and policy-impact assessment are still needed.

4. CONCLUSION

This study conducted a rigorous comparative analysis of the baseline and tuned configurations of Decision Tree and Random Forest Regression, using a nationally representative dataset of 36,064

records across 27 commodities and 38 Indonesian provinces (2021-2026). Empirical results identify the Tuned Random Forest (`n_estimators=300`, `max_depth=20`) as the optimal model, achieving peak performance metrics (R^2 , MAE, RMSE, MAPE) alongside a cross-validation mean that confirms robust generalization. The Tuned Decision Tree (`max_depth=20`, `min_samples_leaf=5`) also showed significant improvement, raising its R^2 to 0.9809 to match the Baseline Random Forest, demonstrating that hyperparameter optimization is critical regardless of the underlying tree algorithm. Feature importance analysis revealed that “Komoditas” (commodity type) is the dominant predictor with an importance score exceeding 91%, far surpassing spatial (province: ~5%) and temporal (Time, Month, Year: <3%) inputs across all states. This proves that intrinsic commodity dynamics influence Indonesian food markets more than regional or seasonal factors. However, a key limitation persists for five highly volatile horticultural commodities that exceeded the 5% MAPE threshold, led by Curly Red Chili (16.13%), Red Bird’s Eye Chili (14.30%), and Large Red Chili (10.38%), whose biological and weather-driven supply shocks cannot be fully captured by categorical and calendar features alone. Future research should integrate exogenous meteorological variables, supply-chain indicators (BPS harvest areas and import volumes), and deep learning architectures such as LSTM or TCN networks optimized for temporal sequence modeling to improve predictions of high volatility. Operationally, deploying this pipeline within a real-time national price intelligence dashboard via the Satu Data Indonesia API represents a highly feasible next step toward evidence-based food price governance.

ACKNOWLEDGMENTS

The authors sincerely thank the Faculty of Computer Science, Amikom Purwokerto University, for its academic and institutional support. Appreciation is also extended to Satu Data Indonesia (data.go.id) for providing the dataset used in this study. The authors also thank Prof. Dr. Eng. Ir. Imam Tahyudin, M.M., for his valuable guidance and academic feedback.

BIBLIOGRAPHY

- [1] M. A. Erissafilla, M. Yosep, and T. Sulistyono, “Data-Driven Model for Predicting Essential Commodity Demand: A Comparative Study of Weekend – Weekday Patterns in Chili Commodities Using Decision Tree and Random Forest Regression,” *J. Appl. Informatics Comput.*, vol. 10, no. 2, pp. 1971–1981, 2026, [Online]. Available: <https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/12378/>
 - [2] I. R. Muchtar, “Comparison of Linear Regression and Random Forest Algorithms for Premium Rice Price Prediction (Case Study : West Java),” *J. Indones. Sos. Teknol.*, vol. 5, no. 7, pp. 3122–3132, 2024, doi: 10.59141/jist.v5i7.1184.
 - [3] U. Wiharja, S. Hartanto, and A. K. Albahar, “Rice Price Prediction in Province Nusa Tenggara Barat (NTB) Using Comparison of Linear Regression and Random Forest Algorithms 2 . Technical Work Preparation,” *Open Int. J. Informatics*, vol. 13, no. 2, pp. 64–74, 2025, doi: 10.1113/oiji2025.13n2.342.
 - [4] N. Syadzwin, D. Hamdhana, A. Razi, and R. Forest, “Comparison of Coffee Bean Sales Predictions at the Ketiara Coffee Traders Cooperative (KOPEPI) Using Linear Regression and Random Forest Methods,” *J. Adv. Comput. Knowl. Algorithms*, vol. 3, no. 2, pp. 82–89, 2026, doi: [doi:doi.org/10.29103/jacka.v3i2.25974](https://doi.org/10.29103/jacka.v3i2.25974).
 - [5] A. Theofilou, S. A. Nastis, A. Michailidis, and T. Bournaris, “Predicting Prices of Staple Crops Using Machine Learning: A Systematic Review of Studies on Wheat , Corn , and Rice,” *Sustainability*, vol. 17, no. 11, pp. 1–34, 2025, doi: 10.3390/su17115456.
 - [6] D. R. Lestari, E. Aditya, S. Bangun, F. L. Gaol, and T. Matsuo, “Machine Learning-Based Forecasting of Agricultural Commodity Prices Using Ensemble Models,” *J. Human, Earth, Futur.*, vol. 6, no. 4, pp. 887–899, 2025, doi: 10.28991/HEF-2025-06-04-09.
 - [7] K. N. Khikmah, B. Sartono, B. Susetyo, and G. A. Dito, “Performance Comparative Study of Machine Learning Classification Algorithms for Food Insecurity Experience by Households in West Java,” *JOIN (Journal Inf. Syst. Eng. Bus. Intell. / J. Online Inform.*, vol. 9, no. 1, pp. 128–137, 2024, doi: 10.15575/join.v9i1.1012.
-

-
- [8] F. M. Saragih and W. C. Wibowo, "Analysis of Food Security Index Predictions in Indonesia Using Machine Learning Approach," *Agro Bali Agric. J.*, vol. 8, no. 2, pp. 377–393, 2025, doi: 10.37637/ab.v8i2.2302.
 - [9] J. D. Giu, A. F. Gaib, and A. Rasyid, "Forecasting Medium Rice ' s Retail Price with Machine Learning in Gorontalo Province," *CogITO Smart J.*, vol. 11, no. 2, pp. 462–477, 2025, doi: 10.31154/cogito.v11i2.930.462-477.
 - [10] D. Kobi, "Multi-Commodity Food Price Forecasting Ahead of Eid Al-Fitr in Indonesia (2026 – 2030): A Comparative Study of Machine Learning Algorithms Using Time-Series Data," *MALCOM Indones. J. Mach. Learn. Comput.*, vol. 6, no. January, pp. 404–413, 2026, doi: 10.57152/malcom.v6i1.2572.
 - [11] S. Aji *et al.*, "Comparison of Multiple Linear Regression and Random Forest Methods for Predicting National Rice Production in Indonesia," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3479–3489, 2025, [Online]. Available: <https://jurnal.polibatam.ac.id/index.php/JAIC/article/view/11398>
 - [12] E. Ramadhani, B. Sartono, A. F. Hadi, S. 'Ufa, and T. Akhdansyah, "Comparison of main characteristics of food insecurity using classification Tree and Random," *Sink. J. dan Penelit. Tek. Inform.*, vol. 6, no. 4, pp. 2486–2497, 2022, doi: <https://doi.org/10.33395/sinkron.v7i4.11852> e-ISSN.
 - [13] V. A. Handayani, S. Rahmiati, B. Varischa, and A. Arrafi, "Comparative Analysis : Multiple Regression and Random Forest Regression in Predicting Food Security Index in Indonesia," *J. Math. Comput. Stat.*, vol. 8, no. 2, pp. 389–400, 2025, doi: 10.35580/jmathcos.v8i2.8593.
 - [14] I. W. Kintara and A. Putra, "A Hybrid Approach to Chili Price Classification Using Ensemble Methods," vol. 7, no. September, pp. 245–254, 2024, [Online]. Available: <https://www.infoteks.org/journals/index.php/jsikti/article/view/241/268>
 - [15] M. Naufal and R. Ibrahim, "Price Forecasting of Shallots Using the Machine Learning Approach of Random Forest Regression Supporting Price Stabilization," *JTEP J. Keteknikan Pertanian*, vol. 13, no. 3, pp. 449–461, 2025, doi: 10.19028/jtep.013.3.449-461.
 - [16] V. Nicolet, J. Dodds, I. Lynce, and R. Martins, "Hypergraph-Guided Regex Filter Synthesis for Event-Based Anomaly Detection," *arXiv Prepr.*, vol. 1, no. 1, 2025.
 - [17] E. Ruthra and A. R. Bevi, "Automated deep U-Net model for ischemic stroke lesion segmentation in the sub-acute phase," *Sci. Rep.*, vol. 15, pp. 1–13, 2025, doi: 10.1038/s41598-025-17337-y.
 - [18] G. Nguyen *et al.*, "Machine Learning and Deep Learning frameworks and libraries for large-scale data mining : a survey," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 77–124, 2019, doi: 10.1007/s10462-018-09679-z.
 - [19] G. H. Dinh *et al.*, "Leveraging Artificial Intelligence to Enhance Port Operation Efficiency," *Polish Marit. Res.*, vol. 31, no. 122, pp. 140–155, 2024.
 - [20] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian J. Mach. Learn.*, pp. 69–79, 2024, doi: <https://doi.org/10.58496/BJML/2024/007>.
 - [21] Y. Dai, P. Dai, D. K. Nguyen, and W. Zhou, "Multiscale risk spillovers and external driving factors: Evidence from the global futures and spot markets of staple foods," *arXiv Prepr.*, pp. 1–32, 2025, doi: 10.48550/arXiv.2501.15173.
 - [22] M. Wójcikowski, P. J. Ballester, and P. Siedlecki, "Performance of machine-learning scoring functions in structure-based virtual screening," *Nat. Publ. Gr.*, vol. 7, no. December 2016, pp. 1–10, 2017, doi: 10.1038/srep46710.
-