JSCE (Journal of System and Computer Engineering)
Vol.7, No.4, October 2026, pp. 485-497
ISSN (online): 2723-1240
DOI : https://doi.org/10.61628/jsce.v7i4.2903

Research Article Open Access (CC–BY-SA) ■

Stability and Significance Analysis of Daily Food Price Forecasting Models

Gus Nanang Syaifuddiin [1,a,*]; Imam Junaedi [2,b]; Angger Binuko Paksi [3,c]; Hendrik Kusbandono [4,d]

[1,2,3,4] *Politeknik Negeri Madiun, Jl. Serayu No. 84, Pandean, Madiun 63133, Indonesia*
[a] *gus.nanang@pnm.ac.id;* [b] *imam.junaedi@pnm.ac.id;* [c] *angger.binuko@pnm.ac.id;* [d] *h3ndrik57@pnm.ac.id*
** Corresponding author*

Abstract

Food price forecasting has been widely studied using machine learning, yet most published work reports accuracy from a single data split without a baseline comparison or a significance test. This study evaluates the stability and statistical significance of such comparisons. Eleven models, comprising four baselines, two statistical models, and five machine learning algorithms, were tested on eight strategic food commodities in East Java Province using 2,380 daily observations from April 2017 to September 2026, obtained from the Strategic Food Price Information Centre of Bank Indonesia. Evaluation employed six-fold expanding window validation with a 30-step horizon, using a scaled error and a skill score relative to a naive benchmark as the primary metrics, followed by the Friedman and Nemenyi tests. No model was significantly superior to the naive benchmark, which ranked third among the eleven models. Rankings were highly unstable: the standard deviation across folds reached 35 to 106 percent of the mean, the best model changed for half of the commodities after only 0.1 percent of the observations were cleaned, and the mean absolute percentage error correlated negatively with the skill score. These findings indicate that claims of model superiority require baseline comparison, multi-fold validation, and significance testing.

Keywords— *forecast evaluation, naive benchmark, ranking stability, Nemenyi test, food price*

1. Introduction

Food price stability is a principal determinant of inflation in Indonesia. Consumer Price Index inflation reached 2.92% in 2025, up from 1.57% in 2024, with the volatile food group as the largest contributor at 6.21%, compared with core inflation of 2.38% and administered price inflation of 1.93% (Bank Indonesia, 2026). This volatility is asymmetric across commodities. Chili and shallot prices can swing by tens of percent within weeks because of supply disruptions and weather anomalies, whereas rice prices are governed by a government supply and price stabilisation framework (Badan Pangan Nasional, 2022) and therefore move slowly and discretely. Anticipating these movements is thus an operational requirement for the national and regional inflation control teams.

This requirement has motivated a growing body of research applying statistical and machine learning methods, ranging from ARIMA and exponential smoothing to tree-based models and deep learning, to Indonesian food price data (Ericko et al., 2023; Fauzani & Rahmi, 2023; Immanuel & Handhayani, 2025; Pertiwi & Harani, 2025). Most of these studies report a mean absolute percentage error (MAPE) below 10% and conclude that a particular algorithm is superior

for the commodity studied (Amirudin et al., 2025; Gunawan et al., 2025; Ishaq et al., 2026; Santoso et al., 2024).

A review of 23 empirical studies published between 2021 and 2026 shows that five applied linear regression or least squares (Anjelie et al., 2024; Arinal & Azhari, 2023; Hasibuan & Musthofa, 2022; Karbala & Ali, 2023; Setiawan et al., 2025), five applied fuzzy time series (Lenisa et al., 2025; Nabila, 2023; Nurhasan et al., 2021; Sarbaini et al., 2023; Siswanti et al., 2023), five applied neural networks or deep learning (Ericko et al., 2023; Gunawan et al., 2025; Santoso et al., 2024; Setyadi et al., 2023; Veri et al., 2022), two applied distance-based algorithms (Bukit & Kurniawan, 2023; Sudrazad & Rizky, 2026), two applied ARIMA (Amirudin et al., 2025; Fauzani & Rahmi, 2023), and one applied Monte Carlo simulation (Pasaribu & Karo, 2024). Nineteen of these studies applied a single method without comparison, three compared several algorithms (Chandra et al., 2025; Immanuel & Handhayani, 2025; Permana et al., 2026), and one proposed a hybrid approach (Amirudin et al., 2025)

However, three methodological concerns remain underexamined. First, most studies omit a baseline comparison, although on series that approximate a random walk a low MAPE reflects the stability of the series rather than the quality of the model (Hyndman & Koehler, 2006; Makridakis et al., 2018). Second, evaluation typically relies on a single train–test split, so it remains unknown whether the best model retains its position in other periods (Bergmeir & Benítez, 2012; Tashman, 2000). Third, comparisons rarely include significance testing, so small differences are interpreted as superiority even though they may fall within random variation (Demšar, 2006). Because adopting a forecasting model entails infrastructure and personnel costs, verifying such claims is a prerequisite for applying the results in practice.

This study addresses four research questions: (RQ1) whether machine learning models statistically outperform a naive benchmark in forecasting daily strategic food prices over a 30-step horizon; (RQ2) how stable model rankings are across evaluation periods; (RQ3) how stable these rankings are under changes in feature configuration and small-scale data cleaning; and (RQ4) whether the commonly reported MAPE reflects the actual benefit of modelling. Eleven models were tested on eight commodities in East Java Province using 2,380 daily observations from April 2017 to September 2026, with six-fold expanding window validation, a scaled error and a skill score relative to the naive benchmark, and the Friedman and Nemenyi tests.

The contribution of this study is two-fold. Methodologically, it offers an evaluation framework that treats the naive benchmark as a mandatory reference, employs multi-fold validation, applies non-parametric significance testing, and assesses robustness to data treatment. Empirically, it documents the limits of forecasting performance on Indonesian daily food price data.

2. Method

2.1 Data Source and Scope

This study used daily traditional market prices from the Strategic Food Price Information Centre of Bank Indonesia for East Java Province, one of the largest food-producing and food-consuming provinces in Indonesia, covering the period from 3 April 2017 to 11 September 2026.

Eight commodities were selected purposively to represent diverse price characteristics: medium-quality rice, fresh broiler chicken, fresh chicken eggs, medium-sized shallot, curly red chili, red bird's eye chili, packaged cooking oil, and local granulated sugar, hereafter referred to as rice, broiler chicken, chicken eggs, shallot, curly red chili, red bird's eye chili, cooking oil, and sugar. The selection criteria were contribution to the inflation basket, diversity in volatility, and data completeness above 90%, the last of which is verified in Table 1. The overall research workflow, from data collection to robustness analysis, is summarised in Figure 1.

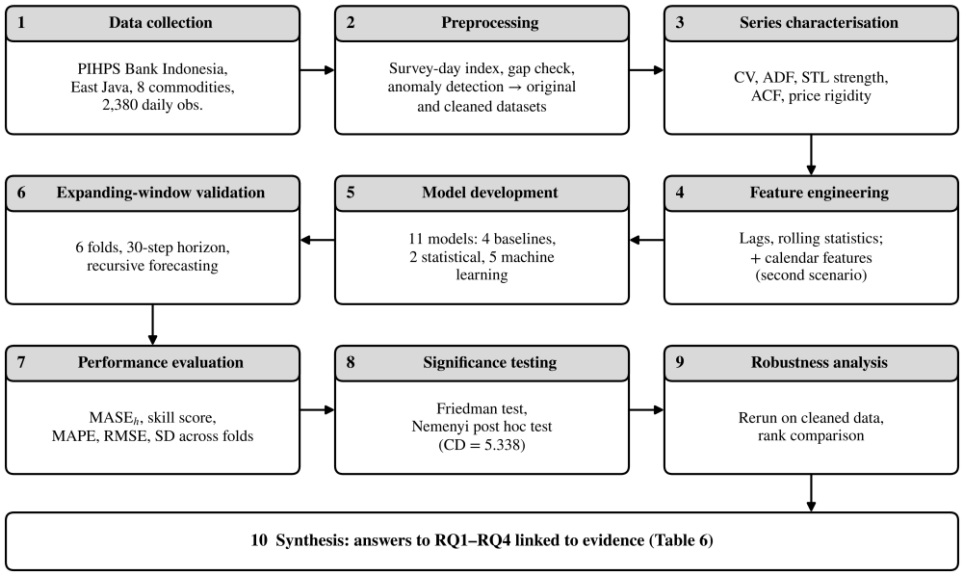


Figure 1. Research flow of the study, from data collection to synthesis of the research questions.

2.2 Survey Frequency Characteristics

Price surveys were conducted only on working days. Of the 3,449 calendar days in the observation period, 2,380 carried records; the recording rate was 95.3% to 98.0% from Monday to Friday and zero at weekends, and a further 85 working days without records coincided with national holidays and collective leave.

Weekends and holidays were therefore treated as days without a survey rather than as missing values. The series were kept on the survey-day scale without reindexing to the full calendar, so the weekly seasonal period was set to five rather than seven, and one forecast step corresponds to one survey day. Linear interpolation was applied only to genuinely missing working days, at most four per commodity, and to the anomalous observations identified in Section 2.3.

2.3 Time Series Characterisation

Volatility was measured using the coefficient of variation (CV) and the standard deviation of daily returns, stationarity using the augmented Dickey–Fuller (ADF) test (Dickey & Fuller, 1979) with lag selection by the Akaike information criterion (AIC) at the 5% significance level, and temporal dependence using the autocorrelation function (ACF) at lags 1 and 5.

The strengths of the seasonal and trend components were computed from a seasonal-trend decomposition using Loess (STL; Cleveland et al., 1990), following the formulation of Hyndman and Athanasopoulos (2021) given in Equations (1) and (2).

$$F_S = max\left(0, 1 - \frac{Var(R_t)}{Var(S_t + R_t)}\right) \quad (1)$$

$$F_T = max\left(0, 1 - \frac{Var(R_t)}{Var(T_t + R_t)}\right) \quad (2)$$

where F_S and F_T are the seasonal and trend strengths, and S_t, T_t, and R_t denote the seasonal, trend, and remainder components, respectively. The decomposition used periods of 5, 21, and 253 survey days for the weekly, monthly, and annual cycles, respectively.

Single-point anomalies were detected by comparing each observation with the mean of its two neighbours; an observation was flagged when its deviation exceeded four times the standard deviation of daily returns, with a minimum of 4%, while the two neighbours remained close to each other. The threshold was scaled because volatility differs across commodities by up to a

factor of ten. All experiments were run on both the original and the cleaned data as a robustness check.

Price rigidity was measured by the frequency of price changes, defined as the proportion of observations that differ from the preceding one, and by the maximum run length of identical prices.

2.4 Models Evaluated

Eleven models were evaluated in three groups. The baselines were the naive model, which projects the last observed value; the seasonal naive model, which repeats the last five survey days; the drift model, which extrapolates the slope between the first and last training observations; and the mean window model, which uses the mean of the last 63 observations.

The statistical models were ARIMA, with the order selected by the lowest AIC among (1,1,1), (2,1,2), (0,1,1), (1,1,0), and (3,1,1), and exponential smoothing with an additive trend and no seasonal component.

The machine learning models were ridge regression, Huber regression (Huber, 1964), random forest (Breiman, 2001), extremely randomised trees, hereafter extra trees (Geurts et al., 2006), and histogram-based gradient boosting (J. H. Friedman, 2001), all with default hyperparameters except for a maximum tree depth of 12 and 100 estimators in the ensemble models. Hyperparameters were deliberately left untuned so that the results reflect typical performance in practice.

2.5 Feature Engineering

For the machine learning models, the features comprised lags 1, 2, 3, 4, 5, 10, and 21, together with rolling means and standard deviations over windows of 5 and 21 observations, all shifted by one period to prevent information leakage. Multi-step forecasts were produced recursively.

Two scenarios were run under identical settings: autoregressive features only, and autoregressive features plus exogenous calendar features comprising indicators for the Ramadan and pre-Eid periods from 2017 to 2026, two pairs of annual Fourier terms, and a day-of-week indicator.

2.6 Evaluation Design

Evaluation used expanding window validation with six folds and a horizon of 30 steps, equivalent to approximately six calendar weeks. In each fold, the test set covered the 30 observations after the training cutoff, which then advanced by 30 observations. This scheme prevents future information from entering the training data and allows performance stability to be assessed across periods (Bergmeir & Benítez, 2012).

The MAPE and the root mean squared error (RMSE) are reported for comparability with the literature, but the primary metric is a horizon-adjusted mean absolute scaled error ($MASE_h$), adapted from Hyndman and Koehler (2006), whose denominator is the mean naive forecast error at the same horizon, as given in Equation (3).

$$MASE_h = \frac{\frac{1}{m}\sum_{t=1}^{m}|y_t - \hat{y}_t|}{\frac{1}{n-h}\sum_{t=1}^{n-h}|y_{t+h} - y_t|} \tag{3}$$

where h is the horizon length, m is the number of test observations, n is the number of training observations, y_t is the observed value, and $\hat{y}_t$ is the forecast. Using the same horizon in the denominator keeps the numerator and denominator commensurate, unlike the conventional one-step denominator.

The skill score (SS), which expresses the relative improvement over the naive benchmark, is given in Equation (4).

$$SS = \frac{MASE_{h,naive} - MASE_{h,model}}{MASE_{h,naive}} \times 100\% \quad (4)$$

Positive values indicate that a model outperforms the naive benchmark. The standard deviation (SD) of $MASE_h$ across folds was used as a measure of stability.

2.7 Significance Testing

Following Demšar (2006), the Friedman test (M. Friedman, 1937) was applied to the model ranks across the eight commodities, followed by the Nemenyi post hoc test, whose critical difference (CD) is given in Equation (5).

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \quad (5)$$

where k is the number of models, N is the number of commodities, and q_α is the critical value of the Studentised range statistic at the 5% significance level. Two models differ significantly when their average ranks differ by more than the CD.

2.8 Implementation

The analyses were implemented in Python 3 using pandas and NumPy for data processing, statsmodels (Seabold & Perktold, 2010) for ARIMA, exponential smoothing, STL, and the ADF test, scikit-learn (Pedregosa et al., 2011) for the machine learning models, and SciPy for the Friedman test, with a fixed random state of 42.

3. Results and Discussion

3.1 Data Characteristics

Data completeness was high for all commodities, as shown in Table 1. Six commodities had all 2,380 observations, whereas curly red chili and red bird's eye chili lacked four and two observations, respectively, giving completeness of 96.39% to 96.55%. No zero values or prices below IDR 1,000 per kilogram were found.

Table 1. Data quality inspection by commodity.

Commodity	Records	Completeness (%)	Zero Values	Invalid Values	Longest Flat Run	Flat Records (%)
Medium-Quality Rice	2,380	96.55	0	0	46	75.63
Fresh Broiler Chicken	2,380	96.55	0	0	7	1.13
Fresh Chicken Eggs	2,380	96.55	0	0	8	1.85
Medium-Sized Shallot	2,380	96.55	0	0	7	1.05
Curly Red Chili	2,376	96.39	0	0	8	0.88
Red Bird's Eye Chili	2,378	96.47	0	0	7	0.76
Packaged Cooking Oil	2,380	96.55	0	0	62	66.01
Local Granulated Sugar	2,380	96.55	0	0	47	57.77

The single-point inspection identified 21 anomalous observations (0.1% of the total), 15 of which coincided with national holidays or collective leave, against a random expectation of 2.2. This pattern indicates incomplete survey coverage on holidays rather than reporting errors. The association was strongest for rice (five of five), broiler chicken (three of three), and cooking oil (three of three); for the rigid rice and cooking oil series, a single deviation stands out particularly clearly. These observations were interpolated, and their effect is assessed in Section 3.6.

Table 2 shows clear diversity across commodities. The CV ranged from 0.0992 for broiler chicken to 0.5248 for red bird's eye chili, whose price ranged from IDR 10,250 to IDR 110,800 per kilogram. Rice, cooking oil, and sugar were non-stationary, with ADF p-values of 0.9152,

0.7950, and 0.5136, respectively, whereas the other five commodities were stationary at the 5% level.

Table 2. Statistical characteristics of the commodities.

Commodity	Mean (IDR)	CV	ADF p-value	Stationary	Price Change Frequency (%)	ACF (Lag 1)
Medium-Quality Rice	11,932	0.1455	0.9152	No	20.0	0.998
Fresh Broiler Chicken	33,023	0.0992	0.0000	Yes	88.8	0.983
Fresh Chicken Eggs	24,287	0.1376	0.0064	Yes	83.2	0.994
Medium-Sized Shallot	30,830	0.2920	0.0001	Yes	89.5	0.996
Curly Red Chili	34,395	0.4064	0.0010	Yes	92.8	0.993
Red Bird's Eye Chili	40,872	0.5248	0.0000	Yes	95.0	0.990
Packaged Cooking Oil	17,084	0.2092	0.7950	No	24.5	0.999
Local Granulated Sugar	13,796	0.1713	0.5136	No	31.0	0.998

Figure 2 shows the price trajectories before anomaly cleaning. Rice, cooking oil, and sugar exhibit long plateaus with discrete level shifts, whereas the chili and shallot series oscillate continuously with high amplitude. Cooking oil shows a structural break in early 2022 during the domestic supply crisis, and rice shows a sustained upward shift from mid-2023. The single vertical spikes in the rice, cooking oil, and sugar series are the anomalies described above.

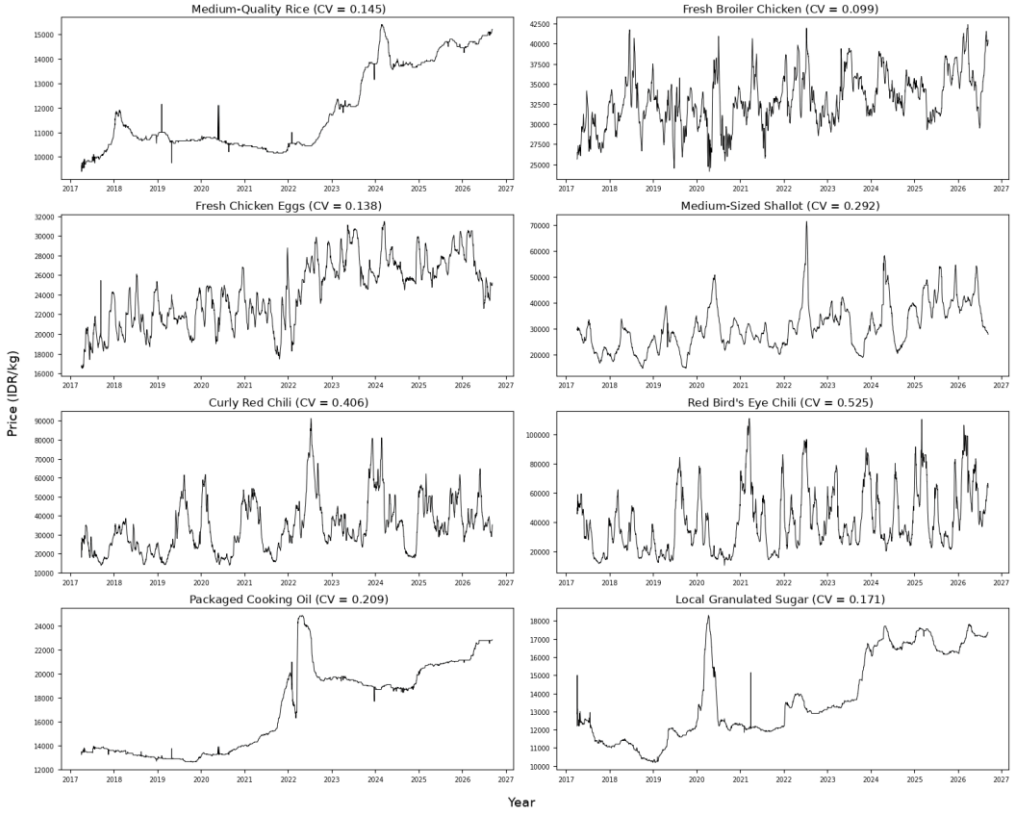


Figure 2. Daily price trajectories of eight strategic food commodities in traditional markets of East Java Province, April 2017 to September 2026.

Price rigidity does not align with volatility. Rice, cooking oil, and sugar changed price in only 20.0%, 24.5%, and 31.0% of observations, with runs of identical prices of up to 46, 62, and 47 observations, respectively, whereas the other five commodities changed price in 83.2% to 95.0% of observations. Broiler chicken had the lowest CV yet changed price in 88.8% of observations, indicating frequent movements of small amplitude.

The weekly seasonal strength was zero for six of the eight commodities and was non-zero only for packaged cooking oil (0.297) and rice (0.036). The monthly strength did not exceed 0.107 (broiler chicken). The annual seasonal strength was highest for chicken eggs (0.496), shallot (0.422), and broiler chicken (0.357), whereas curly red chili recorded only 0.044 despite having the second-highest CV, indicating idiosyncratic rather than calendar-driven fluctuations.

Most consequentially, the lag-1 autocorrelation ranged from 0.983 to 0.999 across all commodities, with trend strengths of 0.933 to 0.998. Together with the absence of weekly seasonality, this indicates that all series approximate a random walk, so the last observed value is highly informative and leaves limited room for improvement by more complex models.

3.2 Model Performance Comparison

The results of the six-fold expanding window validation are presented in Table 3. The $MASE_h$ of the best model ranged from 0.3105 for cooking oil to 0.8089 for broiler chicken, and the skill score relative to the naive benchmark ranged from 5.37% for red bird's eye chili to 31.70% for shallot.

Table 3. Performance of the best model for each commodity.

Commodity	Calendar Features	Best Model	$MASE_h$	SD across Folds	MAPE (%)	Skill Score (%)	Models Beating Naive
Medium-Quality Rice	Without	Random Forest	0.3945	0.2938	0.515	15.88	3 of 11
Fresh Broiler Chicken	Both	Exponential Smoothing	0.8089	0.5651	6.302	15.55	5 of 11
Fresh Chicken Eggs	With	Histogram-Based Gradient Boosting	0.4290	0.1491	3.256	20.44	3 of 11
Medium-Sized Shallot	With	Random Forest	0.3540	0.2298	5.585	31.70	4 of 11
Curly Red Chili	Without	Ridge Regression	0.4130	0.1784	9.990	13.85	5 of 11
Red Bird's Eye Chili	With	Extra Trees	0.5176	0.1934	16.898	5.37	2 of 11
Packaged Cooking Oil	Both	Exponential Smoothing	0.3105	0.3295	0.555	26.96	2 of 11
Local Granulated Sugar	With	Random Forest	0.3836	0.2242	0.939	11.81	2 of 11

The MAPE and the skill score diverge fundamentally. Cooking oil recorded a MAPE of 0.555% with a skill score of 26.96%, whereas red bird's eye chili recorded a MAPE of 16.898% with a skill score of only 5.37%. Across the eight commodities, the Pearson correlation between the two metrics was −0.514 (Spearman's $\rho = -0.476$), so the MAPE ranks commodities roughly in the reverse order of the actual benefit of modelling, because its magnitude reflects series volatility more than model quality (Hyndman & Koehler, 2006). This result addresses RQ4.

3.3 Significance of Differences between Models (RQ1)

The average ranks across the eight commodities are presented in Table 4. Huber regression ranked best (3.12), followed by ARIMA (3.88) and the naive benchmark (4.00), which outperformed all tree-based models and all other baselines.

The Friedman test rejected the hypothesis of equal performance ($\chi^2 = 37.114$, $p = 0.000054$), but the Nemenyi test, with a CD of 5.338, showed that this rejection stemmed from the weakest models. Only the mean window model differed significantly from the naive benchmark, performing worse by 6.88 ranks, whereas the largest favourable difference was 0.88 for Huber regression. Hence, no model was significantly better than the naive benchmark.

ARIMA and Huber regression outperformed the naive benchmark on 75% and 62.5% of the commodities, respectively, without reaching significance; random forest did so on exactly 50%, and the remaining seven models on fewer than half.

Table 4. Average model ranks, Nemenyi test results, and ranks before data cleaning.

Model	Average Rank (Cleaned)	Average Rank (Original)	Difference from Naive	Significance	Wins over Naive (%)
Huber Regression	3.12	3.50	−0.88	Not significant	62.5
ARIMA	3.88	4.12	−0.12	Not significant	75.0
Naive	4.00	3.38	0.00	Reference	—
Random Forest	4.62	4.62	0.62	Not significant	50.0
Drift	4.88	5.00	0.88	Not significant	12.5
Ridge Regression	5.75	5.88	1.75	Not significant	37.5
Extra Trees	6.38	5.62	2.38	Not significant	25.0
Seasonal Naive	7.00	6.88	3.00	Not significant	0.0
Exponential Smoothing	7.75	8.38	3.75	Not significant	25.0
Histogram-Based Gradient Boosting	7.75	7.75	3.75	Not significant	12.5
Mean Window	10.88	10.88	6.88	Significantly worse	0.0

3.4 Stability of Results across Periods (RQ2)

As reported in Table 3, the SD across folds ranged from 0.35 to 1.06 times the mean $MASE_h$. For cooking oil, the SD (0.3295) exceeded the mean $MASE_h$ (0.3105), so performance in one period is a poor guide to performance in another.

The best model also changed with the feature configuration for four of the eight commodities: rice from random forest to drift, chicken eggs from ARIMA to histogram-based gradient boosting, shallot from Huber regression to random forest, and curly red chili from ridge regression to Huber regression. Because the differences among the top models are small, such changes show that the ranking is not robust, and studies that identify a best model from a single split and configuration risk reporting results that do not generalise.

3.5 Effect of Exogenous Calendar Features (RQ3)

As shown in Table 5, the calendar features improved accuracy for four commodities, namely shallot (25.58%), chicken eggs (15.13%), sugar (5.05%), and red bird's eye chili (0.35%); reduced it for curly red chili (14.33%) and rice (13.13%); and left two commodities unchanged because their best models did not use calendar features.

At a 30-step horizon, the autoregressive signal probably dominates the calendar signal, and the additional features may raise the variance of the fitted models without a proportionate gain in information.

Table 5. Effect of adding calendar features on $MASE_h$.

Commodity	Without Calendar Features	With Calendar Features	Change (%)
Medium-Sized Shallot	0.4757	0.3540	+25.58
Fresh Chicken Eggs	0.5055	0.4290	+15.13
Local Granulated Sugar	0.4040	0.3836	+5.05
Red Bird's Eye Chili	0.5194	0.5176	+0.35
Fresh Broiler Chicken	0.8089	0.8089	0.00
Packaged Cooking Oil	0.3105	0.3105	0.00
Medium-Quality Rice	0.3945	0.4463	−13.13
Curly Red Chili	0.4130	0.4722	−14.33

3.6 Robustness to Data Cleaning (RQ3)

The full experiment was rerun on the data cleaned of the 21 anomalous observations, which represent only 0.1% of the total and all fall within the training portion.

The effect was not negligible. Table 4 also reports the average ranks before cleaning. The Spearman correlation between the two rankings was 0.961, driven largely by stability in the lower ranks, but the top of the ranking was rearranged: the naive benchmark fell from first place (3.38) to third (4.00), whereas Huber regression and ARIMA each moved up one place.

At the commodity level, the best model changed for four of the eight commodities. The skill score of cooking oil rose from 1.80% to 26.96% and that of chicken eggs from 4.94% to 20.44%, whereas that of sugar fell from 23.37% to 11.81%; the remaining four commodities changed by at most 15.3% in relative terms.

The naive benchmark uses only the last observation and is therefore unaffected by spikes elsewhere in the training data, whereas the other models are fitted to the full history and are distorted by them. Removing the spikes thus improved these models and lowered the relative position of the naive benchmark, most visibly for cooking oil.

The significance result, however, was unchanged: on both datasets, no model was significantly superior to the naive benchmark, and only the mean window model was significantly worse. The absence of a significant advantage is therefore robust, whereas the specific ranking is not.

Cleaning also ruled out price rigidity as an explanation for the benefit of modelling. The mean skill score of rigid versus flexible commodities was 8.7% versus 15.4% on the original data but 21.4% versus 17.4% after cleaning; a reversal caused by a 0.1% change in the data indicates noise rather than structure.

3.7 Discussion

The findings support two conclusions. The first is robust: no model statistically outperformed the naive benchmark on either dataset. The second is that rankings are unstable, as shown by SDs across folds of 35% to 106% of the mean, changes in the best model for four of the eight commodities under both feature modification and 0.1% data cleaning, and a reordering of the top three ranks.

With a lag-1 autocorrelation above 0.98 and no weekly seasonality, daily food prices approximate a random walk, leaving complex models with little additional signal to learn. This finding is consistent with large-scale forecasting competitions in which simple statistical methods have proved competitive with machine learning (Makridakis et al., 2018, 2020).

These results call for a correction of reporting practice rather than a rejection of machine learning. Claims of superiority should be supported by a baseline comparison, multi-fold validation, significance testing, a robustness check against data treatment, and metrics scaled against a baseline rather than the MAPE alone.

For policy operators, a naive rule is an adequate operational reference for a horizon of about six weeks on daily data, and accuracy gains are more likely to come from fundamental variables such as supply, rainfall, buffer stocks, and producer prices than from more complex algorithms. Table 6 summarises the evidence for each research question. These findings apply to eight commodities in one province, one 30-step horizon, and models with largely untuned hyperparameters, and they should not be generalised beyond these conditions.

Table 6. Research question, evidence, and main finding.

Research question	Evidence	Main finding
RQ1. Do machine learning models significantly outperform the naive benchmark?	Friedman test and Nemenyi test (CD = 5.338); Table 4	No. No model was significantly better than the naive benchmark; the largest favourable rank difference was 0.88 (Huber regression).

Research question	Evidence	Main finding
RQ2. How stable are model rankings across evaluation periods?	SD across folds equal to 35%–106% of the mean $MASE_h$; Table 3	Unstable. Performance differs markedly between periods, so a single split is unreliable.
RQ3. How robust are rankings to feature configuration and data cleaning?	Best model changed for 4 of 8 commodities under each change; top-three ranks reordered; skill scores shifted (e.g., cooking oil 1.80% to 26.96%); Tables 4 and 5	Rankings are not robust, but the absence of significant superiority is robust on both datasets.
RQ4. Does the MAPE reflect the benefit of modelling?	Pearson $r = -0.514$, Spearman $\rho = -0.476$ between MAPE and skill score; Table 3	No. The MAPE mainly reflects series volatility and should be accompanied by a benchmark-relative metric.

4. Conclusions

This study evaluated the stability and statistical significance of comparisons among forecasting models for daily food prices over a 30-step horizon, using 2,380 observations of eight commodities in East Java Province. None of the eleven models statistically outperformed the naive benchmark, which ranked third on average (4.00); although the Friedman test rejected equal performance ($\chi^2 = 37.114$, $p = 0.000054$), the Nemenyi test, with a CD of 5.338, attributed this result to the poorly performing models, the largest favourable difference being 0.88 for Huber regression. Model rankings proved unstable, with SDs across folds of 35% to 106% of the mean and changes in the best model for four of the eight commodities under both feature modification and the cleaning of only 21 observations (0.1% of the data). Calendar features improved accuracy for four commodities and reduced it for two, and the MAPE correlated negatively with the skill score (−0.514). These results reflect series that approximate a random walk, with lag-1 autocorrelations of 0.983 to 0.999; therefore, claims of model superiority should be verified through baseline comparison, multi-fold validation, significance testing, and robustness checks, whereas a naive rule suffices as a short-term operational reference. The main strength of this study is its verification framework; its limitations are the small number of commodities, which lowers statistical power, the coverage of a single province and a single horizon, and the use of untuned hyperparameters. Future work should broaden commodity and regional coverage, examine additional horizons, and assess fundamental variables such as supply, rainfall, and buffer stocks.

Acknowledgements

The authors thank Bank Indonesia for providing open access to the Strategic Food Price Information Centre data used in this study.

References

Amirudin, A., Prasetyo, S. A., & Triyono, G. (2025). Implementasi hybrid ARIMA-Prophet untuk prediksi harga beras nasional. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10*(4), 3837–3850. https://doi.org/10.29100/jipi.v10i4.8689

Anjelie, M. K., Arisandi, D., & Sutrisno, T. (2024). Penerapan metode least square untuk prediksi harga komoditas pangan Kota Singkawang. *Progresif: Jurnal Ilmiah Komputer, 20*(1), 53–64. https://doi.org/10.35889/progresif.v20i1.1293

Arinal, V., & Azhari, M. (2023). Penerapan regresi linear untuk prediksi harga beras di Indonesia. *Jurnal Sains dan Teknologi, 5*(1), 341–346.

Badan Pangan Nasional. (2022). *Peraturan Badan Pangan Nasional Nomor 15 Tahun 2022 tentang stabilisasi pasokan dan harga beras, jagung, dan kedelai di tingkat konsumen*. https://peraturan.bpk.go.id/Details/307571

Bank Indonesia. (2026). *Analisis inflasi Desember 2025*. https://www.bi.go.id/id/publikasi/laporan/Pages/Analisis-Inflasi-TPIP-Desember-2025.aspx

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bukit, S. J. A., & Kurniawan, R. (2023). Prediksi harga tandan buah segar dengan algoritma K-nearest neighbor. *Jurnal Sistem Komputer dan Informatika (JSON), 5*(1), 92–101. https://doi.org/10.30865/json.v5i1.6818

Chandra, R., Maulana, F., & Sofi, N. (2025). Perbandingan regresi linear dan Prophet untuk prediksi harga beras grosir di Indonesia. *Jurnal Ilmiah Komputasi, 24*(3), 439–446. https://doi.org/10.32409/jikstik.24.3.3857

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–73.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research, 7*, 1–30. https://jmlr.org/papers/v7/demsar06a.html

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*(366), 427–431. https://doi.org/10.2307/2286348

Ericko, T., Lauro, M. D., & Handhayani, T. (2023). Prediksi harga pangan di pasar tradisional Kota Surabaya dengan metode LSTM. *Jurnal Ilmu Komputer dan Sistem Informasi, 11*(2). https://doi.org/10.24912/jiksi.v11i2.26012

Fauzani, S. P., & Rahmi, D. (2023). Penerapan metode ARIMA dalam peramalan harga produksi karet di Provinsi Riau. *Jurnal Teknologi dan Manajemen Industri Terapan, 2*(4), 269–277. https://doi.org/10.55826/tmit.v2i4.283

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association, 32*(200), 675–701. https://doi.org/10.2307/2279372

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Gunawan, G., Andriani, W., & Ujianto, N. T. (2025). Prediksi harga cabai musiman menggunakan model LSTM di Jawa Tengah. *Infomatek, 27*(2), 243–254. https://doi.org/10.23969/infomatek.v27i2.26460

Hasibuan, L. H., & Musthofa, S. (2022). Penerapan metode regresi linear sederhana untuk prediksi harga beras di Kota Padang. *JOSTECH Journal of Science and Technology, 2*(1), 85–95. https://doi.org/10.15548/jostech.v2i1.3802

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Immanuel, J. P., & Handhayani, T. (2025). Prediksi harga pangan Jayapura menggunakan ELM, LSTM, LightGBM, dan GB. *Jurnal Ilmu Komputer dan Sistem Informasi, 13*(2). https://doi.org/10.24912/jiksi.v13i2.35122

Ishaq, K. F., Afriyansyah, R., & Wathan, M. H. (2026). Prediksi harga pangan di Kota Pangkalpinang menggunakan algoritma XGBoost. *Jurnal Inovasi Teknologi Terapan, 4*(2), 333–340. https://doi.org/10.33504/jitt.v4i2.420

Karbala, S., & Ali, I. (2023). Memprediksi harga beras eceran menggunakan algoritma regresi linier. *JATI (Jurnal Mahasiswa Teknik Informatika), 7*(3), 1554–1559. https://doi.org/10.36040/jati.v7i3.6901

Lenisa, U., Sugianti, S., & Astuti, I. P. (2025). Prediksi harga cabai menggunakan fuzzy time series model Chen. *Jurnal Rekayasa Teknologi dan Komputasi, 1*(1), 1–16. https://doi.org/10.24269/jrtekom.v1i1.6430

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS ONE, 13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting, 36*(1), 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014

Nabila, P. I. (2023). Prediksi harga minyak goreng kelapa sawit menggunakan average-based fuzzy time series Lee. *Jurnal Riset Matematika, 3*(2), 157–166. https://doi.org/10.29313/jrm.v3i2.2835

Nurhasan, U., Fatmawati, A. D., & Harijanto, B. (2021). Implementasi metode fuzzy time series Markov chain untuk prediksi harga telur puyuh. *Informatika Mulawarman: Jurnal Ilmiah Ilmu Komputer, 16*(2), 80–87. https://doi.org/10.30872/jim.v16i2.5251

Pasaribu, Y. P. P., & Karo, I. M. K. (2024). Analisis prediksi harga beras di Indonesia dengan metode Monte Carlo. *JATI (Jurnal Mahasiswa Teknik Informatika), 8*(4), 4418–4423. https://doi.org/10.36040/jati.v8i4.9962

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Permana, J., Sujana, R. N., Komala, K. P., & Dwiyanto, B. (2026). Prediksi harga cabai rawit menggunakan LGBM, random forest, dan XGBoost. *Jurnal Ilmu Komputer dan Sistem Informasi, 14*(1). https://doi.org/10.24912/7nbfrp18

Pertiwi, A. A., & Harani, N. H. (2025). Model terpopuler untuk prediksi harga bahan pokok: Sebuah tinjauan literatur. *JATI (Jurnal Mahasiswa Teknik Informatika), 9*(5), 7865–7873. https://doi.org/10.36040/jati.v9i5.14927

Santoso, A., Purnamasari, A. I., & Ali, I. (2024). Prediksi harga beras menggunakan metode recurrent neural network dan long short-term memory. *PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer, 11*(1), 128–136. https://doi.org/10.30656/prosisko.v11i1.7921

Sarbaini, S., Yanti, D., & Nazaruddin, N. (2023). Prediksi harga beras belida di Kota Pekanbaru menggunakan fuzzy time series Cheng. *Jurnal Teknologi dan Manajemen Industri Terapan, 2*(3), 234–241. https://doi.org/10.55826/tmit.v2i3.183

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 92–96). https://doi.org/10.25080/Majora-92bf1922-011

Setiawan, R., Ajuirai, A., Azka, R., Hasibuan, A., & Fhonna, R. P. (2025). Prediksi harga daging sapi menggunakan regresi linear berganda. *Jurnal Ilmu Komputer dan Sistem Informasi, 4*(3), 290–304. https://doi.org/10.70340/jirsi.v4i3.223

Setyadi, M. A., Faqih, A., & Wijaya, Y. A. (2023). Peramalan harga komoditas beras di Kalimantan Timur menggunakan algoritma neural network. *JATI (Jurnal Mahasiswa Teknik Informatika), 7*(1), 320–324. https://doi.org/10.36040/jati.v7i1.6327

Siswanti, S., Nugroho, A., Widada, B., & Kusumaningrum, A. (2023). Fuzzy time series untuk prediksi harga gabah kering panen. *Techno.Com, 22*(2), 430–439. https://doi.org/10.33633/tc.v22i2.7956

Sudrazad, A., & Rizky, F. (2026). Prediksi harga cabai merah di Pasaraya MMTC Medan menggunakan K-nearest neighbor. *Portal Riset dan Inovasi Sistem Perangkat Lunak, 4*(2), 191–196. https://doi.org/10.59696/prinsip.v4i2.200

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Veri, J., Surmayanti, S., & Guslendra, G. (2022). Prediksi harga minyak mentah menggunakan jaringan syaraf tiruan. *MATRIK: Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer, 21*(3), 503–512. https://doi.org/10.30812/matrik.v21i3.1382