



Chili Price Prediction Model in Banyumas and Cilacap Regencies Using Comparative Machine Learning Algorithm

Fima Rizqina Mubaroka¹ , Maulida Ayu Fitriani², Elindra Ambar Pambudi³, Dimara Kusuma Hakim⁴

¹⁻⁴ Department of Informatics Engineering, Universitas Muhammadiyah Purwokerto, Indonesia, 153174

 rizqinafima@gmail.com

 <https://doi.org/10.37339/e-komtek.v10i1.3003>

Published by Politeknik Piki Ganesha Indonesia

Abstract

Artikel Info

Submitted:

13-01-2026

Revised:

14-01-2026

Accepted:

14-01-2026

Online first :

19-06-2026

Fluctuations in chili prices, as a strategic commodity, frequently pose challenges to food price stability, particularly in Jawa Tengah regions such as Banyumas and Cilacap Regencies. This study aimed to explore and compare the performance of three machine learning algorithms: Ridge Regression, Lasso Regression, and Random Forest Regressor. The dataset comprised historical daily price data from January 2018 to June 2025, which were used to identify market patterns and trends. In addition, the study generated chili price forecasts for the subsequent 365 days. The results consistently indicated that Lasso Regression achieved the best predictive performance in both regions, attaining the highest accuracy in Banyumas Regency (MAPE = 2.75%) and Cilacap Regency (MAPE = 5.08%). However, visual analysis of long-term forecasts revealed that Ridge Regression produced more realistic prediction patterns compared to Lasso Regression. In contrast, the Random Forest Regressor failed to capture long-term trends, as it generated predictions that tended to be overly stable and flat. The forecasting results were subsequently visualized through an interactive dashboard developed using the Flask framework.

Keywords: Price Prediction; Chilli Pepper; Machine Learning.

Abstrak

Fluktuasi harga cabai sebagai komoditas strategis sering kali menjadi tantangan bagi stabilitas harga pangan, terutama di wilayah Jawa Tengah seperti Kabupaten Banyumas dan Kabupaten Cilacap. Penelitian ini bertujuan untuk mengeksplorasi dan membandingkan performa tiga algoritma machine learning seperti Ridge Regression, Lasso Regression, dan Random Forest Regressor. Data yang digunakan mencakup data historis harian dalam periode Januari 2018 hingga Juni 2025 untuk mengidentifikasi pola dan tren pasar. Selain itu, penelitian ini memprediksi harga cabai untuk 365 hari ke depan. Secara konsisten, Lasso Regression menunjukkan performa terbaik di kedua wilayah. Model tersebut mencapai akurasi tertinggi di Kabupaten Banyumas (MAPE 2,75%) dan Kabupaten Cilacap (MAPE 5,08%). Namun, analisis visual pada prediksi jangka panjang mengungkapkan bahwa Ridge Regression menghasilkan grafik yang lebih realistis dibandingkan Lasso Regression. Sebaliknya, Random Forest Regressor gagal menangkap tren jangka panjang karena menghasilkan prediksi yang cenderung stabil (datar). Hasil prediksi tersebut selanjutnya divisualisasikan ke dalam dashboard interaktif berbasis framework Flask.

Kata-kata kunci: Prediksi Harga; Cabai; Machine Learning.



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

1. Introduction

Chili peppers are a strategic commodity in traditional market trade. Central Java Province is one of Indonesia's red chili pepper producing provinces. One of the contributing regions is Banyumas and Cilacap Regencies, which have a climate that supports chili pepper production. Chili pepper crops play a crucial role in the economic sector in influencing food commodity prices throughout Indonesia. As a leading commodity, chili prices are highly susceptible to production volumes and weather disturbances, causing chili prices to fluctuate constantly [1]. This condition poses a challenge for the government as the guardian of food price stability. Therefore, accurate and precise price prediction strategies are needed to deal with high demand and price spikes.

Several previous studies have been conducted to test the performance of machine learning (ML) algorithms in predicting chili prices using data sourced from the National Strategic Food Price Center website. For example, study [2] entitled "Chili Price Prediction One Year Ahead Using the Gated Recurrent Unit Method" shows that GRU provides excellent prediction results with a MAPE value below 10% for four types of chili peppers and hyperparameter combinations. Research [3] found that in a limited dataset, the Multilayer Perceptron (MLP) algorithm provided more accurate performance (MAPE 3.79%) than the Recurrent Neural Network (RNN). In addition, the conventional statistical approach in research [4] using the ARIMA (4,1,1) method produced an MSE of 13,839.57. Although several methods have been applied, the use of supervised learning algorithms such as Ridge Regression, Lasso Regression, and Random Forest Regressor in the field of machine learning is still limited. In fact, these three algorithms involve different but more efficient and simpler approaches, do not require large computations, and produce consistent prediction results. Specifically, Ridge Regression and Lasso Regression have the ability to optimize linear relationship patterns, while Random Forest Regressor has the flexibility to model non-linear patterns based on decision trees.

The research focused on a smaller area, namely the regencies of Banyumas and Cilacap. The selection of the area was based on considerations of the quality and completeness of the available data. The purpose of this research is to explore the Ridge Regression, Lasso Regression, and Random Forest Regressor algorithms. Furthermore, a comparative analysis of each algorithm will be conducted to determine the best performance. Performance comparison is measured using evaluation metrics, such as Mean Absolute Percentage Error (MAPE).

In the final stage of the research, the algorithm implementation will be presented in the form of an interactive dashboard. This dashboard aims to facilitate understanding of the analysis and decision making. The visualization will display the prediction results graphically, including a comparison plot between actual data and test data, as well as a prediction graph for the next 365 days. The prediction results are visualized with Flask. Flask is an efficient Python-based backend website framework for presenting data analysis.

2. Method

The research was conducted using a quantitative approach. The data used was time series data, which is a set of observations within a certain time interval [5]. However, in this study, historical data was converted into input features used in supervised learning algorithm training. Supervised learning is a type of machine learning that focuses on identifying data to predict future values using labeled data. This transformation was carried out using a sliding window technique with a 60-day duration as to predict the value for the following day. The algorithm implementation stages can be seen in Figure 1.

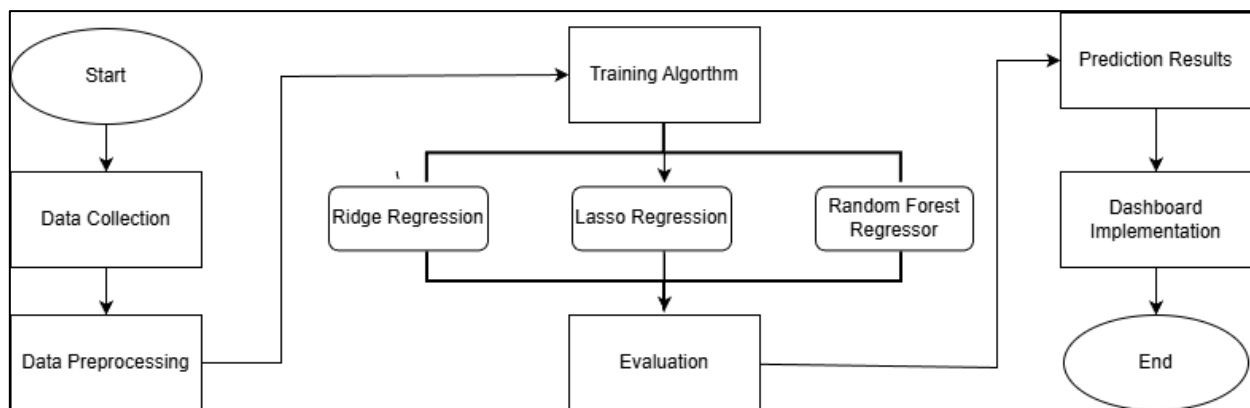


Figure 1 Algorithm Implementation

2.1. Data Collection

Data was taken from the National Strategic Food Price Center (PIHPS Nasional) website. This website provides a list of food prices for several commodities and specific regions, including the Banyumas and Cilacap regencies. The data used is the price of red chili peppers. The time range taken is based on all daily data from January 2018 to June 2025.

2.2. Data Pre-processing

Pre-processing is an important step in improving data quality and model performance [6]. The initial pre-processing stage was carried out by adding temporal variables, handling missing

values, and ensuring there were no duplicates. Data with empty values were filled in using interpolation. Interpolation is a way of handling missing values based on historical data available between empty data [7]. This stage was conducted to improve data quality.

The next stage involves data normalization to improve effectiveness during the training process and increase accuracy. Data normalization is necessary to adjust the scale of each variable in the dataset to be smaller and more balanced [8]. Min-Max Scaling was selected for data normalization because of its ability to unify feature scales while preserving the original data distribution. The equation (1) used in the price normalization process is as follows:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Description:

X_{scaled} = Data values after normalization

X = Original data values

X_{min} = Minimum value of all data

X_{max} = Maximum value of all data

The final stage is the division of the dataset into training data and test data at a ratio of 80% and 20%. The training data will be used to train the model to recognize and learn patterns. After that, the test data, which the model has never seen before, is used to test and evaluate its performance. The division of data aims to ensure that the evaluation results are objective.

2.3. Algorithms

In the modeling stage, the research focused on the implementation and comparison of three machine learning algorithms. The three models are Ridge Regression, Lasso Regression, and Random Forest Regressor.

a. Ridge Regression

Ridge Regression is a model that can overcome multicollinearity by adding a penalty (L2) to the error calculation formula [9]. The model works by inserting a bias value into the regression equation to reduce the variance of its parameter estimator [10]. Thus, this L2 penalty will minimize the coefficients (w) of all variables to produce a stable and slightly biased model. Equation (2) is the formula used in calculating the Ridge Regression model:

$$J(\mathbf{w}) = \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \alpha \|\mathbf{w}\|_2^2 \quad (2)$$

Description:

- $J(w)$ = Minimized cost function
- w = Vector weight (coefficient of the model)
- X = Input data feature matrix
- y = Target value vector (actual)
- α = hyperparameter (*alpha*)
- $\|w\|_2^2$ = L2 *Penalty Term* (sum of square weights)

b. Lasso Regression

Lasso Regression has the same concept as Ridge Regression, which is to add a penalty (L1) to the calculation formula. The uniqueness of L1 is its ability to suppress coefficients (w) to zero. This step is considered more effective for feature selection in variables that are considered informative [11]. Equation (3) is the formula used to calculate the Lasso Regression model:

$$J(w) = \frac{1}{n} \|Xw - y\|_2^2 + \alpha \|w\|_1 \quad (3)$$

Description:

- $\|w\|_1$ = L2 *Penalty Term* (sum of the absolute value of weight)

c. Random Forest Regressor

Random Forest Regressor has the ability to adapt to high-dimensional and highly varied data [12]. The output is calculated based on the average value of the decision trees. Equation (4) is the formula used to calculate the Random Forest Regressor model:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^B T_b(x) \quad (4)$$

Description:

- \hat{y} = final prediction result for a single sample x
- B = Total number decision trees built
- $T_b(x)$ = Prediction function of the second decision tree b

2.4. Evaluation

Several evaluation metrics are used to calculate algorithm performance. The evaluation metric used in this study is Mean Absolute Percentage Error (MAPE). MAPE is the absolute error percentage of the average difference between the actual value and the forecast value for a certain period. The lower the MAPE value, the higher the accuracy of a prediction [13]. The method for calculating MAPE is shown in equation (5).

$$MAPE = \left(\frac{100\%}{n}\right) \sum_{t=1}^n \frac{|x_t - F_t|}{x_t} \quad (5)$$

The criteria for interpreting MAPE values are shown in **Table 2.1**.

Table 2.1 Criteria of MAPE Value

MAPE Values	Criteria
< 10%	Excellent
10% - 20%	Good
20% - 50%	Enough
>50%	Bad

2.5. Dashboard Implementation

The final result of this research will display a Flask-based graph visualization. Flask is a Python-based micro website framework with strong compatibility and high scalability [14]. One of the advantages of this framework is its flexibility in structuring code, so that the interface design can be developed separately from the underlying mathematical code [15]. The website pages use a simple HTML structure and data sources in the form of CSV files containing prediction results from training data, actual data, and predictions for the next 365 days in the period July 2025–June 2026.

3. Results and Discussion

3.1 Data Pre-Processing

This study contains 1,956 data in the period from January 2018 to June 2025. The data structure is compiled in a univariate, namely by using the main variables in the form of historical price data from both regions (hargaBms and hargaClp) and additional variables in the form of temporal such as day, week and month indices (hari_ke, minggu_ke and bulan_ke). The addition of these variables aims to enable the model to learn patterns and prevent overfitting. To ensure that the data is ready to be used during the training process, duplicate values and missing values

are checked. The handling of missing value is carried out by the interpolation method. The dataset can be seen in [Figure 2](#) and [Figure 3](#).

Data columns (total 5 columns):				
#	Column	Non-Null	Count	Dtype
0	Date	1956	non-null	datetime64[ns]
1	hargaBms	1956	non-null	int64
2	hari_ke	1956	non-null	int8
3	minggu_ke	1956	non-null	int8
4	bulan_ke	1956	non-null	int8

Figure 2. Banyumas Regency Dataset

Data columns (total 5 columns):				
#	Column	Non-Null	Count	Dtype
0	Date	1956	non-null	datetime64[ns]
1	hargaClp	1956	non-null	int64
2	hari_ke	1956	non-null	int8
3	minggu_ke	1956	non-null	int8
4	bulan_ke	1956	non-null	int8

Figure 3. Dataset of Banyumas Regency

3.2 Algorithm Comparison

a. Ridge Regression

The implementation of the Ridge Regression algorithm uses the parameter of the model regulation value = Ridge(alpha=1.0). The use of this parameter aims to activate the L2 penalty to reduce the value of the coefficient w. The graph of the prediction results can be seen in [Figures 4](#) and [Figure 5](#).

This algorithm shows high accuracy between the test data lines and the actual data that are squeezed together. Meanwhile, in the future prediction line, there is a strong seasonal trend, fluctuating and capturing high price volatility. Trends suggest that chili prices will rise by January.

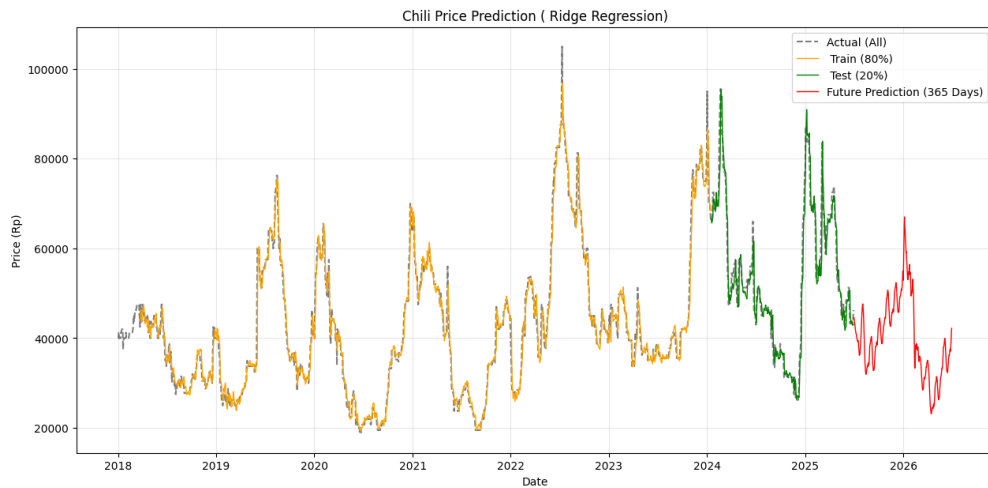


Figure 4. Graph of Banyumas Regency Ridge Regression Prediction Results

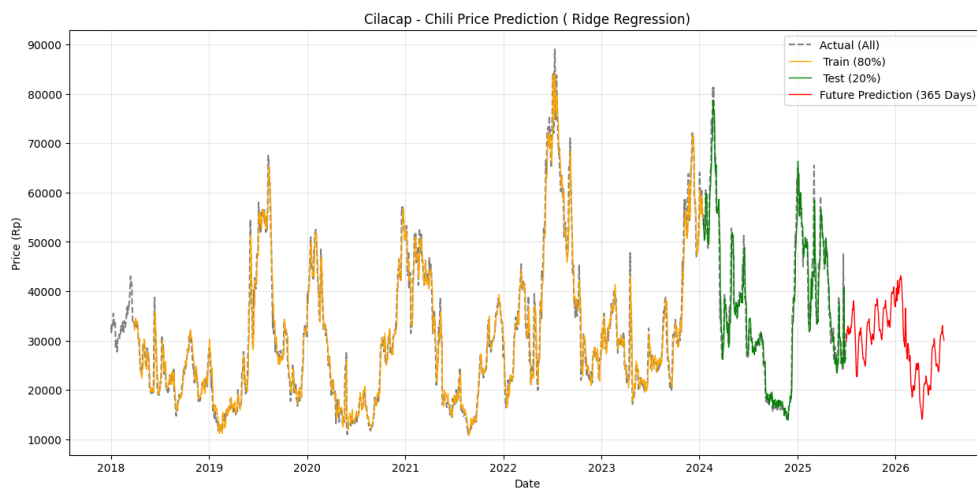


Figure 5. Graph of Cilacap Regency Ridge Regression Prediction Results

b. Lasso Regression

Unlike *Ridge Regression*, the implementation of *Lasso Regression* is focused on its ability to select features through *L2 penalties*. With the model parameter = $\text{Lasso}(\alpha=0.0001)$, the *Lasso* will trim the irrelevant coefficients close to zero. The results of the prediction can be seen in **Figures 6** and **Figure 7**. The Lasso Regression algorithm displays actual data and test data more precisely than the Ridge Regression algorithm. Seasonal trends can still be captured with this algorithm, although the future prediction chart does not show high and fluctuating price spikes.

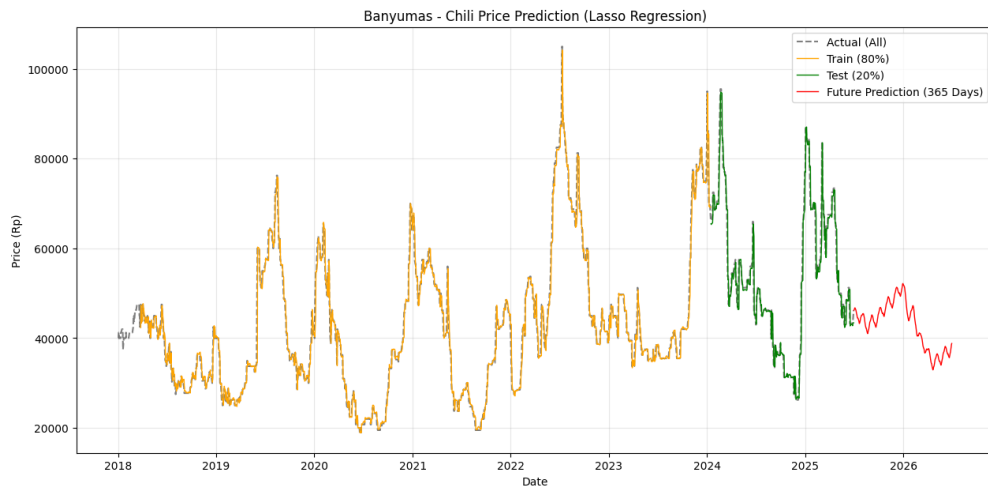


Figure 6 Graph of Lasso Regression Prediction Results of Banyumas Regency

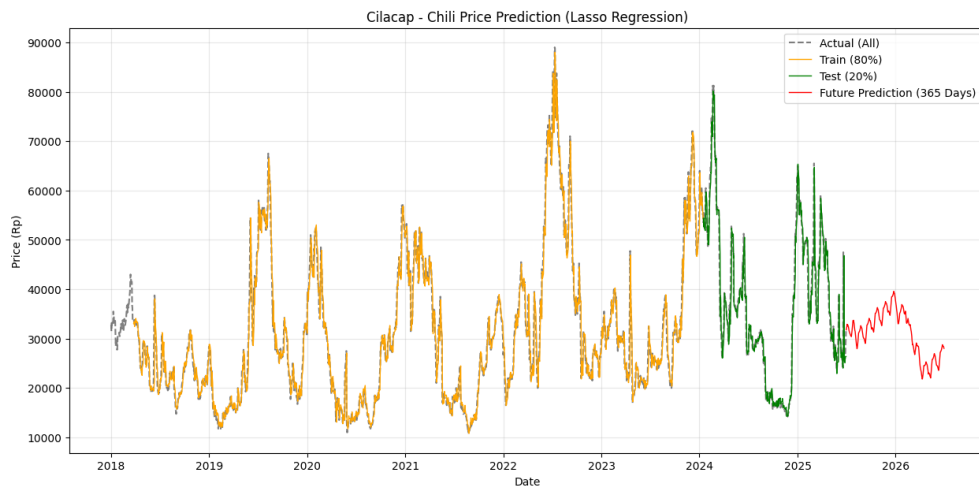


Figure 7 Graph of Cilacap Regency Lasso Regression Prediction Results

c. Random Forest Regressor

The Random Forest Regressor algorithm is trained with parameters $n_estimators=100$ and $max_depth=10$ to capture non-linear relationships that are not reached on the Ridge Regression and Lasso Regression algorithms. Visually, the prediction results can be seen in [Figures 8](#) and [Figure 9](#). The actual data graph and test data are quite good, but in future predictions there are lines that tend to be stable. This indicates that the algorithm cannot capture trends and price fluctuations for long-term predictions due to the nature of the Random Forest Regressor which cannot predict values outside the range of the training data and provide predictions based on historical average values.

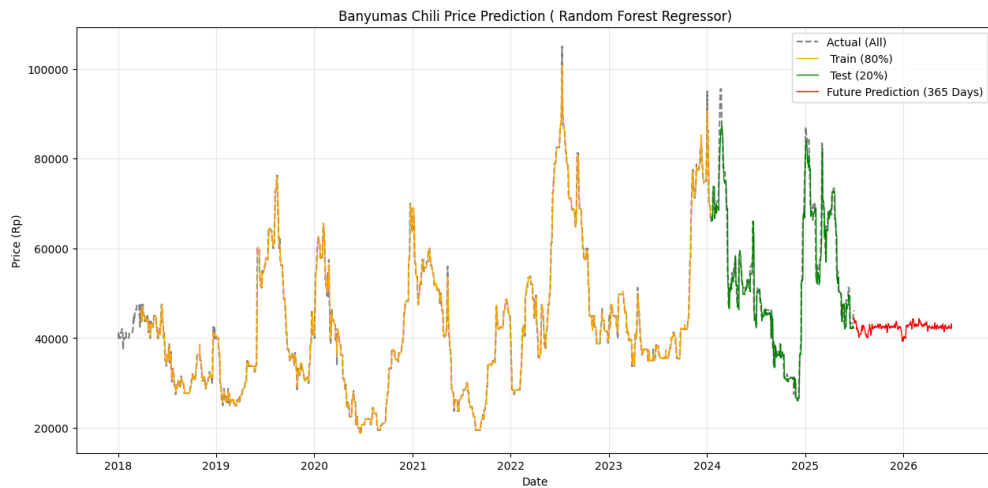


Figure 8 Graph of Random Forest Regressor Prediction Results in Banyumas Regency

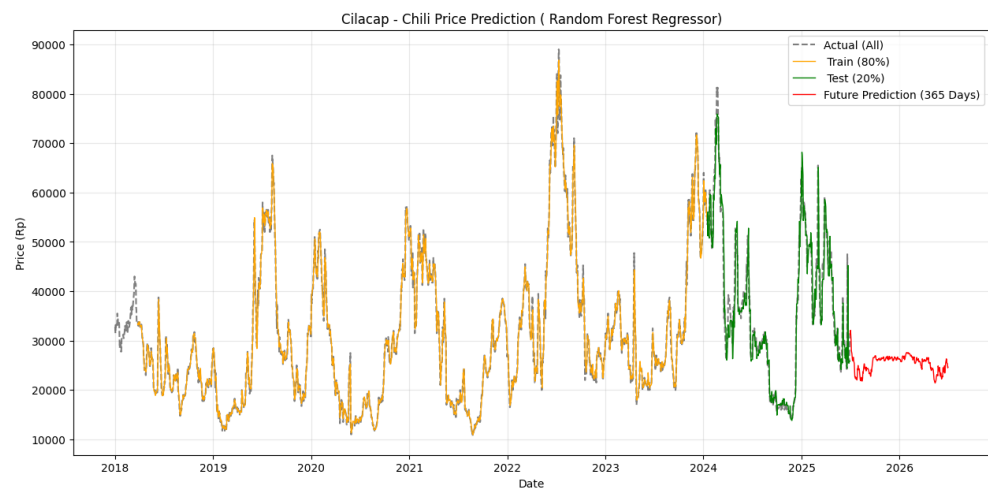


Figure 9 Graph of Random Forest Regressor Prediction Results in Cilacap Regency

3.3 Comparison of Algorithm Performance Evaluation

The performance of the three algorithms was evaluated on 20% of the test data by comparing the MAPE values. Based on the three regions, the order of algorithms with the lowest MAPE were Lasso Regression, Random Forest Regressor, and Ridge Regression. Lasso Regression consistently showed the best performance in both regions. In Banyumas Regency, Lasso Regression obtained a MAPE value of 2,75%. Meanwhile, in Cilacap Regency, it obtained MAPE value of 5,08%. The comparison results can be seen in [Table 3.1](#).

Table 3. 1 Matrix Evaluation Comparison

Location	Algoritma	MAPE (%)
Banyumas	Ridge Regression	4,03
Banyumas	Lasso Regression	2,75

Banyumas	Random Forest Regressor	3,44
Cilacap	Ridge Regression	5,96
Cilacap	Lasso Regression	5,08
Cilacap	Random Forest Regressor	5,90

3.4 Dashboard Implementation

As the final stage of the research, three machine learning algorithms were implemented into an interactive dashboard shown in [Figure 10](#). At the top, there is a navigation feature to filter data based on the 'Location' and 'Algorithm' parameters to be displayed. The visualization results are presented in the form of graphs consisting of actual data (dotted lines), training data results (green lines), and 365-day prediction results (red lines). Color information for the graphs is also available in the information labels at the top of the graphs.

This website has a data inspection feature, whereby when the user hovers the cursor over any point on the graph line, the system will display a tooltip containing complete information in the form of the date and price value at that point. This feature makes it easier for users to observe data with precision. In addition, users can also select 365-day predictions on a specific date.

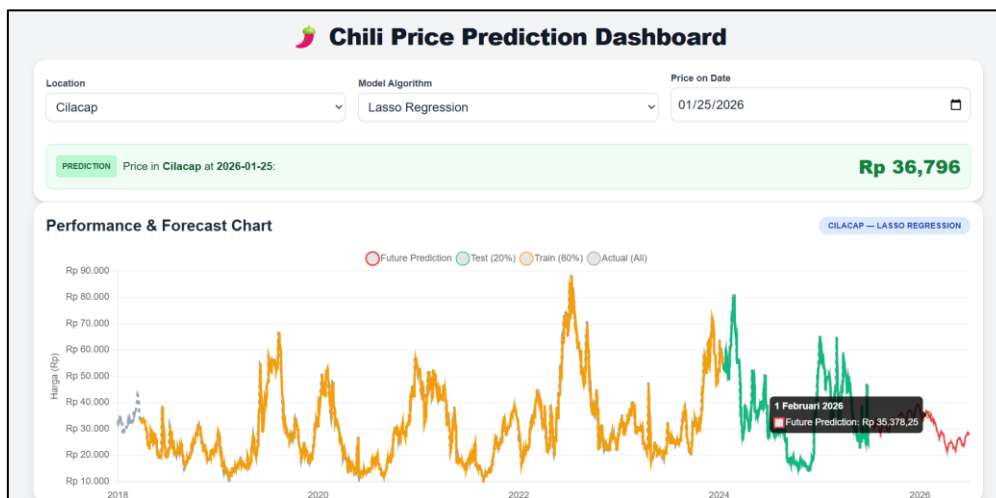


Figure 10 Chili Price Prediction Dashboard in Banyumas and Cilacap Regencies

4. Conclusion

This study has successfully explored and compared three machine learning algorithms (Ridge Regression, Lasso Regression, and Random Forest Regressor) to predict chili prices in Banyumas and Cilacap Regencies. The algorithm evaluation results show that Lasso Regression performs very well and is the most accurate in both regions. Specifically, Lasso Regression

achieved the highest accuracy rate in Banyumas with MAPE value of 2.75% and in Cilacap with MAPE value of 5.08%.

However, the visualization results for long-term predictions (365 days) for the Lasso Regression algorithm showed poor and flat results. In contrast, Ridge Regression produced long-term prediction graphs that looked good and realistic. These findings show that Lasso Regression (L1) is more accurate for short-term predictions, while Ridge Regression (L2) is better for long-term predictions.

The practical implications of these outcomes highlight the importance of using different analytical tools for local food price policies. Instead of relying on one model, local governments like Banyumas and Cilacap should use hybrid system: Lasso Regression as a tactical tool for high-accuracy short-term market interventions, and Ridge Regression as a strategic tool to project annual trends more realistically. The Random Forest Regressor was unable to capture long-term prediction patterns. For future research, model accuracy can be improved by integrating more complex external variables, such as precipitation and temperature data, to mitigate the impact of weather anomalies on chili production volumes. Furthermore, exploring hybrid methods that combine the strengths of L1 and L2 regularization could provide a solution for generating more dynamic and reliable long-term predictions.

References

- [1] S. K. Zahra et al., "Analisis Fluktuasi Harga Terhadap Faktor-Faktor Yang Mempengaruhi Produktivitas Usaha Tani Cabai Merah di Indonesia," *J. Ris. dan Inov. Manaj.*, vol. 2, no. 2, pp. 162–172, 2024, doi: <https://doi.org/10.59581/jrim-widyakarya.v2i2.3526>.
- [2] C. Yustitia and N. Charibaldi, "Chili Price Prediction One Year Ahead Using the Gated Recurrent Unit Method," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 10, no. 1, pp. 1–7, 2024, doi: [10.25139/inform.v10i1.8269](https://doi.org/10.25139/inform.v10i1.8269).
- [3] K. H. Suradiradja, "Algoritme Machine Learning Multi-Layer Perceptron dan Recurrent Neural Network untuk Prediksi Harga Cabai Merah Besar di Kota Tangerang," vol. 14, no. 4, pp. 194–205, 2021, doi: [10.30998/faktorexacta.v14i4.10376](https://doi.org/10.30998/faktorexacta.v14i4.10376).
- [4] A. A. Asriadi et al., "PERAMALAN CABAI BESAR DI KOTA MAKASSAR DENGAN METODE ARIMA," *J. Pemikir. Masy. Ilm. Berwawasan Agribisnis*, vol. 9, no. 1, pp. 24–39, 2023.
- [5] Fungsi Wahyu and Billy Hendrik, "Perbandingan Algoritma Time Series Dan Fuzzy Inference System Dalam Analisis Data Deret Waktu," *J. Penelit. Teknol. Inf. dan Sains*, vol. 1, no. 3, pp. 16–24, 2023, doi: [10.54066/jptis.v1i3.711](https://doi.org/10.54066/jptis.v1i3.711).

- [6] K. Maharana et al., "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: 10.1016/j.gltip.2022.04.020.
- [7] N. M. Sakti et al., "Water Pollution Forecasting in Cengklik Reservoir Using the Triple Exponential Smoothing Method," *J. E-Komtek*, vol. 9, no. 1, pp. 313–324, 2025.
- [8] P. P. Allorerung et al., "Analisis Performa Normalisasi Data untuk Klasifikasi K-Nearest Neighbor pada Dataset Penyakit," vol. 9, no. 3, pp. 178–191, 2024.
- [9] Z. Zeng, "Using Linear Regression , Ridge Regression , Lasso Regression , and Elastic Net Regression for Predicting Real Estate Price," no. Icdse, pp. 509–515, 2025, doi: 10.5220/0013700300004670.
- [10] M. P. Rajan, "An Efficient Ridge Regression Algorithm with Parameter Estimation for Data Analysis in Machine Learning," *SN Comput. Sci.*, vol. 3, no. 2, pp. 1–16, 2022, doi: 10.1007/s42979-022-01051-x.
- [11] M. Xu, "Sales Prediction Based on Lasso Regression," vol. 88, pp. 343–349, 2024.
- [12] B. Roger et al., "Information Processing in Agriculture Random forest regressor applied in prediction of percentages of calibers in mango production," *Inf. Process. Agric. J.*, vol. 12, no. December 2024, pp. 370–383, 2025, doi: 10.1016/j.inpa.2024.12.002.
- [13] Y. W. P. Setiawan et al., "Wind Speed Forecast Using the Triple Exponential Smoothing Method in Pangandaran Beach," *J. E-Komtek*, vol. 8, no. 2, pp. 583–594, 2024.
- [14] Z. Liang et al., "Data Analysis and Visualization Platform Design for Batteries Using Flask-Based Python Web Service," *World Electr. Veh. J.*, vol. 12, no. 4, 2021, doi: 10.3390/wevj12040187.
- [15] M. S. Bonney et al., "Development of a digital twin operational platform using Python Flask," *Data-Centric Eng.*, vol. 3, no. 1, pp. 1–14, 2022, doi: 10.1017/dce.2022.1.