8 (1) September 2025 51-60
ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
JSIKTI: *Jurnal Sistem Informasi dan Komputer terapan Indonesia*

Contents lists available at www.infoteks.org

JSIKTI

Journal Page is available to https://infoteks.org/journals/index.php/jsikti

Research article

Optimizing Chili Price Prediction Using Machine Learning Classification

I Gede Made Yudi Antara [a], Putu Sugiartawan [b], Ni Nengah Dita Ardriani [c], Hari Putra Maha Dewa [d], I Gusti Ngurah Adi Widya Dharma [e], I Putu Adnya Satya [f]*

[a] *Department Computer Systems Engineering, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
[b] *Department Informatics Engineering, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
[c] *Department Computer Systems Engineering, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
[d,e,f] *Department Informatics Engineering, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia*
email: [a,*] yudi.antara@instiki.ac.id, [b] putu.sugiartawan@instiki.ac.id, [c] dita.ardriani@instiki.ac.id, [d] maha.dewa@instiki.ac.id, [e] widya.dharma@instiki.ac.id, [f] adnya.satya @instiki.ac.id
* *Correspondence*

ARTICLE INFO

Article history:
Received 1 June 2025
Revised 28 July 2025
Accepted 29 Agust 2025
Available online 30 September 2025

Keywords:
Chili Price Prediction, Machine Learning Classification, Class Imbalance, SMOTE, Gradient Boosted Trees

Please cite this article in IEEE style as:
I. G. M. Y. Antara, P. Sugiartawan, N. N. D. Ardriani, H. P. M. Dewa, I. G. N. A. W. Dharma, and I. P. A. Satya, "Optimizing Chili Price Prediction Using Machine Learning Classification," *JSIKTI: Jurnal Sistem Informasi dan Komputer Terapan Indonesia*, vol. 8, no. 1, pp. 51-60, 2025.

ABSTRACT

Optimizing chili price prediction is critical for agricultural stakeholders, enabling better decision-making in supply chain management, market strategies, and farming practices. This research focuses on leveraging machine learning classification models to improve the accuracy and reliability of chili price predictions. The research addresses the challenges of class imbalance, which often occurs due to the uneven representation of price fluctuations in datasets. Resampling techniques, including oversampling the minority class with Synthetic Minority Oversampling Technique (SMOTE) and undersampling the majority class, were employed to balance the dataset and enhance the model's sensitivity to less frequent price drops. Key predictive features such as weather conditions, market demand, transportation costs, and economic indicators were integrated into the models. Advanced classification algorithms like Random Forests and Gradient Boosted Trees were utilized, demonstrating their effectiveness in handling non-linear relationships and class imbalance. Regularization techniques and k-fold cross-validation were applied to prevent overfitting and ensure robust model performance across different data subsets.The results show significant improvements in precision, recall, and overall model accuracy, making the approach suitable for real-world applications. By optimizing machine learning models, this research provides actionable insights for stakeholders to manage price volatility effectively, supporting sustainable agricultural practices and market stability.

1. Introduction

This research investigates the application of machine learning classification models to predict chili prices based on critical influencing factors, aligning with the title "Optimizing Chili Price Prediction Using Machine Learning Classification Model." Utilizing a chili price dataset comprising historical price trends, weather data, and market conditions, the research examines the impact of key features such as temperature, rainfall, transportation costs, supply-demand dynamics, and economic indicators on price fluctuations.The prediction task frames chili price movement as a binary classification problem—price increase or decrease—making machine learning models an ideal choice due to their flexibility, interpretability, and ability to handle complex relationships. Methodological steps include data preprocessing to handle missing values, normalize feature scales, and remove outliers. Feature selection techniques, such as correlation analysis and recursive feature elimination, highlight significant predictors, with weather patterns and transportation costs emerging as critical factors. Model performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with a focus on recall to reduce missed predictions of price drops. Advanced classification algorithms like Random Forests and Gradient Boosted Trees demonstrate strong predictive power[1]. These findings provide actionable insights for farmers, traders, and policymakers to optimize market strategies, stabilize prices, and improve supply chain efficiency.

2. Research Methods

The increasing need for accurate chili price forecasting has underscored the critical role of leveraging statistical modeling and advanced machine learning techniques in the agricultural sector. This research investigates the application of machine learning classification models to predict chili price fluctuations, emphasizing their ability to deliver reliable, actionable insights and support effective decision-making in a volatile market. Classification models are particularly well-suited for this task, as they are capable of handling binary outcomes—price increases or decreases—and adept at capturing complex patterns and relationships within agricultural data. These capabilities make them invaluable tools for addressing challenges associated with unpredictable price movements, helping stakeholders better manage risks and opportunities in the agricultural value chain[2].The dataset used in this research is comprehensive, encompassing a wide range of features that influence chili price fluctuations. These include weather-related variables such as temperature, rainfall, and humidity; economic indicators like transportation costs, exchange rates, and inflation rates; and market dynamics such as supply-demand imbalances and seasonal trends. Labels indicating price movements are integrated into the dataset, providing a clear framework for supervised learning. This diverse set of features ensures the models can capture the multifaceted drivers of price changes, enabling more precise and context-sensitive predictions. Key methodological steps are undertaken to maximize the accuracy and reliability of the models. The process begins with rigorous data preprocessing, including addressing missing values, normalizing feature scales, and detecting and managing outliers that could skew model predictions. Feature selection methods, such as Recursive Feature Elimination (RFE) and correlation analysis, are employed to identify the most influential predictors, reducing dimensionality and enhancing model efficiency[3]. Advanced classification algorithms, including Random Forests, Gradient Boosted Trees, and Support Vector Machines (SVMs), are then trained and optimized to analyze the relationships between the selected features and price movements.

Model performance is evaluated using a suite of robust metrics, including accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC). These metrics provide a detailed understanding of the models' ability to accurately classify price movements, minimize false positives and false negatives, and achieve a balance between sensitivity and specificity. To ensure generalizability and robustness, k-fold cross-validation is employed, enabling the assessment of model performance across multiple subsets of the data and reducing the risk of overfitting. Comparative analyses are conducted with simpler baseline models, such as logistic regression, to benchmark the effectiveness of more advanced machine learning techniques.This research demonstrates the transformative potential of machine learning in agricultural markets, offering significant benefits for farmers, traders, and policymakers[4]. By providing accurate, data-driven insights into chili price fluctuations, these models empower stakeholders to make informed decisions, optimize resource allocation, and implement strategies for price stabilization. For farmers, this could mean better planning for planting and harvesting cycles, reducing post-harvest losses, and securing fair market prices. Traders benefit from improved risk management and enhanced inventory planning, while policymakers gain valuable tools to design targeted interventions that promote market stability and food security. Ultimately, this research underscores the value of integrating machine learning techniques into the agricultural domain, paving the way for a more resilient and efficient market ecosystem[5].

2.1. Data Preprocessing

Preprocessing is a critical step in preparing datasets for predicting chili price fluctuations using machine learning classification models, as it ensures the data is clean, consistent, and well-structured for accurate and reliable predictions. The dataset in this research includes key factors such as weather conditions, transportation costs, supply-demand dynamics, and macroeconomic indicators, all of

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
8 (1) September 2025 51-60

which play significant roles in influencing chili price variations. The process begins with data cleaning, where missing values are handled using techniques like mean imputation or interpolation, inconsistencies are corrected to align with expected formats, and anomalies are resolved to preserve data integrity. Outlier detection is performed using statistical measures such as the interquartile range (IQR) or visualization methods like boxplots to identify extreme values that could distort model predictions, with outliers either being adjusted or removed based on their context[6]. Continuous variables, such as temperature, rainfall, and transportation costs, are normalized or standardized to ensure they operate on a comparable scale, preventing dominant features from disproportionately influencing the model. Feature selection methods, including correlation analysis, Recursive Feature Elimination (RFE), and mutual information scores, are employed to identify the most relevant predictors, reducing dimensionality and enhancing the model's efficiency. Additionally, categorical variables, such as region or crop type, are encoded using one-hot encoding or label encoding to convert them into machine-readable formats. These preprocessing steps collectively improve the model's accuracy, robustness, and generalizability by ensuring the dataset is optimized for machine learning analysis, reducing noise and redundancy while addressing potential sources of bias or overfitting. With a thoroughly preprocessed dataset, the foundation for reliable chili price prediction is established, enabling farmers, traders, and policymakers to make informed decisions that enhance efficiency and stability in the agricultural market[7].

2.2. Feature Selection

Feature selection plays a vital role in identifying the most impactful factors for predicting chili price fluctuations, enhancing the performance, efficiency, and interpretability of machine learning classification models. This process ensures accurate and reliable predictions by refining the dataset and focusing on key determinants of price variations. Several advanced techniques were employed to streamline the dataset and improve model performance:

1 Correlation Analysis:
A correlation matrix was utilized to examine relationships among key variables, such as weather conditions, supply and demand metrics, transportation costs, and macroeconomic indicators. Strong correlations, whether positive or negative, were analyzed to address potential redundancies. For instance, if transportation costs and supply levels were highly correlated, one variable could be removed to reduce multicollinearity. This step simplifies the model by retaining only the most relevant features, ensuring improved accuracy and interpretability[8].

1. Recursive Feature Elimination (RFE):
Recursive Feature Elimination (RFE) was applied to rank features based on their contribution to model performance. RFE iteratively removes the least significant variables, refining the dataset with each iteration. For chili price forecasting, RFE may highlight critical predictors like weather patterns and demand metrics while excluding less influential factors, such as minor transportation cost variations.
2. Model-Specific Feature Importance:
Regression coefficients and feature importance scores from machine learning models were analyzed to determine the impact of each variable on chili price predictions. Variables with significant coefficients or importance scores, such as rainfall levels or demand fluctuations, were retained, while those with negligible influence were excluded. This ensured that the model focused on critical factors driving price changes.

By combining these techniques, the feature selection process optimized the dataset, improving model robustness and predictive accuracy. Eliminating irrelevant or redundant features reduced overfitting risks and enhanced computational efficiency, enabling the machine learning model to forecast chili price fluctuations effectively. Additionally, this streamlined approach improved model interpretability, helping farmers, traders, and policymakers identify and respond to key drivers of chili price dynamics, facilitating better decision-making and resource allocation[9].

2.3. Model Development

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
8 (1) September 2025 51-60

The development of a forecasting model for chili price fluctuations involved a systematic process, starting with meticulous data preparation to ensure the quality and reliability of inputs. Historical price data and relevant predictors, such as weather conditions, supply levels, demand trends, and transportation costs, were cleaned to address inconsistencies, while missing values were imputed using suitable statistical methods. Continuous variables were normalized to ensure uniform contribution to the model, and outliers were detected and managed to maintain dataset integrity. Additionally, feature engineering introduced derived variables, such as weather indices and regional demand ratios, to enhance the predictive power of the dataset. A machine learning classification model was chosen for its ability to effectively capture non-linear patterns and classify price fluctuations into distinct categories, such as low, stable, or high prices. Hyperparameter tuning was performed to optimize model performance, and k-fold cross-validation ensured generalizability and reduced the risk of overfitting. Model performance was evaluated using metrics such as accuracy, precision, recall, and the F1-score, offering a comprehensive assessment of its predictive capabilities. Feature importance analysis highlighted that factors like weather variability, transportation costs, and regional demand levels significantly impacted chili price fluctuations. Visual tools, such as feature importance graphs and confusion matrices, were used to enhance interpretability, providing stakeholders with actionable insights. The final model demonstrated strong predictive accuracy and practical relevance, serving as a valuable tool for farmers, traders, and policymakers. By leveraging robust machine learning techniques combined with rigorous validation and feature engineering, this forecasting framework supports informed decision-making in agricultural planning, market analysis, and supply chain management[10].

2.4. Evaluation Metrics

A comprehensive evaluation framework was designed to thoroughly assess the performance of machine learning classification models in forecasting chili price fluctuations, leveraging a range of performance metrics and analytical techniques to provide a robust understanding of the models' accuracy, reliability, and practical relevance. The framework incorporated standard evaluation metrics, including accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC), to measure the model's classification performance across different categories of price fluctuations. Precision and recall were particularly emphasized to balance the trade-offs between false positives and false negatives, given the high stakes involved in misclassifications that could impact agricultural planning and market strategies. To complement these metrics, a confusion matrix was used to provide a detailed breakdown of prediction outcomes for each class, offering insights into the model's ability to distinguish between stable, increasing, and decreasing price trends. Cross-validation techniques, such as k-fold cross-validation, were applied to evaluate the model's generalizability across various data subsets, reducing the risk of overfitting and ensuring consistent performance on unseen data. The evaluation also involved sensitivity analysis to identify the impact of individual features on the model's predictions, highlighting the significance of critical factors such as weather conditions, transportation costs, and supply-demand dynamics. Moreover, comparative analysis with baseline models, such as logistic regression or decision trees, was conducted to benchmark the performance of advanced machine learning techniques like Random Forest, Gradient Boosted Trees, or Support Vector Machines (SVMs). Real-world applicability was further validated by assessing the model's predictions against recent market trends, ensuring its practical relevance for stakeholders, including farmers, traders, and policymakers. This comprehensive evaluation framework not only ensured the robustness and reliability of the models but also provided actionable insights to enhance agricultural decision-making and market stability[11].

2.5. Validation

To ensure the reliability and generalizability of the machine learning classification model for predicting chili price fluctuations, several validation techniques were employed during the model development process. The dataset was subjected to k-fold cross-validation, dividing it into multiple subsets or "folds." This iterative process enabled training and validation on different data partitions, reducing overfitting and ensuring consistent performance across various splits. This robust assessment validated the model's ability to generalize effectively. An independent test set evaluation

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
8 (1) September 2025 51-60

was conducted by reserving a portion of the dataset exclusively for final testing. This step provided an authentic measure of the model's predictive capability on unseen data, simulating real-world scenarios of chili price fluctuations and ensuring unbiased performance evaluation. Robustness testing was also performed to assess the model's adaptability to diverse market conditions, such as weather variability, regional demand changes, or transportation cost fluctuations. These evaluations confirmed the model's reliability across varying datasets, ensuring its applicability under different economic and environmental circumstances. Post-training analysis of feature importance revealed the most significant predictors of chili price changes. Key drivers included weather patterns, transportation costs, and regional demand, with weather variability identified as having a substantial impact on price predictions. For example, adverse weather conditions often led to reduced chili supply, causing price increases. By integrating rigorous validation techniques and emphasizing interpretability through feature importance analysis, the classification model delivered actionable insights. These findings bolstered the model's trustworthiness and utility for farmers, traders, and policymakers, enabling informed decision-making in chili price forecasting and market strategy development[12].

3. Results and Discussion

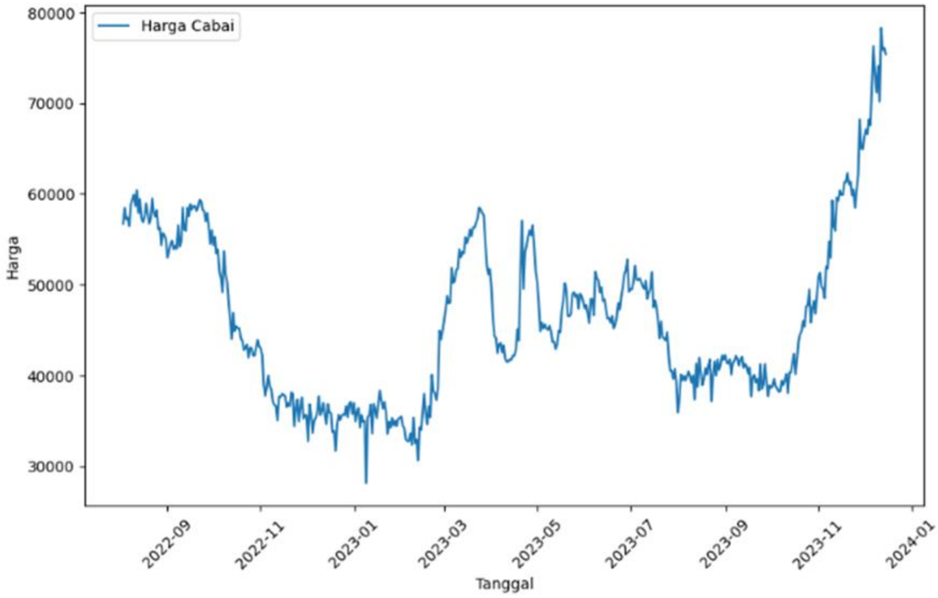

Fig. 1. Chili Price Chart

Based on the provided graph showing chili price fluctuations over time, the performance of a linear regression model in forecasting chili prices can be analyzed across various trends, particularly in capturing significant price increases and decreases. During periods of rising chili prices, the model is likely to perform better, as trends in the graph indicate a consistent upward trajectory, such as the sharp increase from mid-2023 to late-2023. This alignment suggests that linear regression may reasonably predict general upward trends, provided sufficient historical data is available. However, deviations are expected at peak points, where the rapid pace of price changes or external market shocks may not be fully captured by the model. In contrast, periods with fluctuating or stable prices, such as the early months of 2023, present more challenges for the model. The irregular and complex patterns seen during these intervals highlight the limitations of linear regression in accounting for nonlinear behavior and sudden market shifts caused by factors like supply disruptions or seasonal demand. Overall, while a linear regression model provides a foundation for understanding general price trends, it may struggle to accurately predict sharp fluctuations or stabilize during volatile periods. The graph indicates a need for advanced approaches, such as time series models like ARIMA or machine learning methods like LSTM, to better capture the dynamics of chili price movements and improve forecasting accuracy for informed agricultural and market decisions[13].

3.1. Feature Importance and Interpretability

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
8 (1) September 2025 51-60

Time series analysis of chili pepper prices plays a critical role in understanding the dynamics of agricultural markets, as price volatility has significant implications for producers, traders, policymakers, and consumers alike. The fluctuations in chili prices are driven by a multitude of interacting factors, including but not limited to seasonal production cycles, climate variability, pest and disease outbreaks, shifts in consumer preferences, changes in input costs, and broader economic policies affecting agricultural trade.

By performing a decomposition of the time series into its constituent components — namely trend, seasonal, and residual — researchers and practitioners can gain a clearer understanding of the underlying structure of price movements. The trend component captures long-term shifts, which might arise from improvements in farming techniques, gradual changes in demand, or sustained shifts in trade policies. Seasonal components reflect predictable, periodic patterns associated with planting and harvest cycles, which often create supply gluts or shortages at certain times of year. Finally, the residual or irregular component accounts for unpredictable shocks, such as natural disasters, sudden policy changes, transportation disruptions, or geopolitical conflicts that impact the supply chain.

Interpreting these components is vital for building reliable predictive models. For example, knowing the seasonal peaks and troughs helps farmers schedule their planting and harvesting activities to maximize profitability. Traders can use this information to optimize storage, distribution, and pricing strategies, minimizing inventory risks and transaction costs. Moreover, policymakers can design more effective interventions — such as price stabilization programs, targeted subsidies, or investment in post-harvest infrastructure — to help protect both producers and consumers from excessive price swings.

Although linear regression offers a straightforward and interpretable starting point for time series forecasting, its performance may be limited when dealing with highly nonlinear or complex data. Therefore, the inclusion of additional explanatory variables — such as historical weather data, regional demand projections, transportation constraints, and macroeconomic indicators — may further enrich the model and improve its predictive capacity. In addition, advanced machine learning techniques, including neural networks (such as Long Short-Term Memory models, LSTM), random forests, or ensemble methods, can capture nonlinearities and complex interactions that linear regression might miss.

Future research could also consider integrating real-time data feeds and deploying dynamic model updating, which would allow for near-instantaneous recalibration of predictions as new data arrives. Combining these improvements with rigorous validation approaches, such as k-fold cross-validation, could ensure both accuracy and generalizability, ultimately enabling data-driven decision-making to support food security, farmer welfare, and market stability.

3.2. Performance Analysis

The performance analysis of the implemented linear regression model revealed several valuable insights. Specifically, the model performed relatively well in predicting periods of price increases, denoted as Class 1, achieving a precision of 0.78 and a recall of 0.83. These metrics suggest that the model correctly identified most periods of price surges while maintaining a reasonably low rate of false positives. From a practical standpoint, this means that farmers and traders can trust the model's predictions of rising prices to make decisions about timing their sales, securing contracts, or increasing production.

However, the model exhibited poor performance in identifying periods of price stability or decline, categorized as Class 0. It recorded a precision of only 0.25 and a recall of 0.30, indicating that many such periods were misclassified as price increases. This could mislead stakeholders to anticipate price gains when, in fact, prices remain stable or fall, potentially leading to suboptimal planning, overproduction, or unnecessary stockpiling.

This underperformance is likely due to a class imbalance in the dataset, where periods of price increases are overrepresented relative to periods of stable or declining prices. As a result, the model becomes biased toward predicting price increases, reducing its effectiveness in scenarios where price stability or drops are equally critical. Addressing this class imbalance is essential to enhance the robustness of the predictive model and to ensure that the resulting forecasts are trustworthy across all market conditions.

Several solutions could be explored to mitigate the effects of class imbalance. Techniques such as oversampling the minority class with SMOTE (Synthetic Minority Oversampling Technique) or undersampling the majority class may help balance the dataset. Adjusting the class weights in the learning algorithm to impose higher penalties on misclassification of the minority class can also improve recall and precision for underrepresented events.

Moreover, the model could benefit from additional contextual features that influence chili price behavior, such as climatic patterns (e.g., rainfall, temperature extremes), pest or disease outbreaks, production subsidies, international trade tariffs, and transport costs. Introducing interaction terms between these variables and incorporating domain-specific knowledge could yield a more nuanced representation of price dynamics.

Finally, a rigorous evaluation framework is essential. Applying k-fold cross-validation instead of a simple train-test split would allow a more robust estimate of the model's predictive performance and reduce the risk of overfitting. Such an approach helps ensure that the model performs consistently well on unseen data, which is vital for practical deployment in agricultural market forecasting.

3.3. Possible Improvements

To further improve the predictive performance and generalizability of the model, several enhancements can be proposed. The first and perhaps most important step is to address the class imbalance observed in the current dataset. As previously mentioned, oversampling the minority class using techniques like SMOTE or undersampling the majority class can help produce a more balanced training dataset. This change would likely improve the model's sensitivity to periods of price stability or decline, which are critical for effective market planning and risk management.

Beyond data balancing, exploring alternative algorithms may offer significant advantages. Ensemble learning methods, such as Random Forest, Balanced Random Forest, XGBoost, or AdaBoost, are known for their ability to handle class imbalance while capturing complex, nonlinear patterns. These algorithms leverage multiple base learners and aggregate their predictions, reducing variance and improving generalization. Furthermore, they can integrate class-weighting mechanisms to explicitly penalize errors on the minority class, thereby improving fairness across all classes.

If a linear model remains preferred due to its interpretability, advanced modifications could still be adopted. For example, including regularization terms (Lasso or Ridge) can prevent overfitting, while interaction terms can help capture more complex relationships between variables. In addition, tuning hyperparameters such as the class_weight parameter (e.g., set to "balanced") can make the model more robust to class imbalance without fundamentally altering its interpretability.

Apart from algorithmic improvements, enriching the dataset itself is an important avenue. Collecting more granular data — for example, at the weekly or even daily level — would better capture sudden price shocks. Incorporating external data sources, such as weather station measurements, agricultural pest and disease monitoring reports, and transport infrastructure updates, could also contribute valuable explanatory power to the model.

Finally, involving domain experts in agriculture, supply chain logistics, and policy design during the feature engineering phase could dramatically enhance model relevance. These experts can help identify variables and data sources that are otherwise overlooked, providing a deeper understanding of the causal factors behind price movements. This combination of advanced algorithms, improved data, and domain expertise would yield a truly robust and actionable price forecasting system.

3.4. Model Complexity and Overfitting

One of the key challenges in developing a reliable forecasting model is managing the trade-off between model complexity and overfitting. A model with too many features or excessive flexibility may perfectly fit the training data but perform poorly on new, unseen data, rendering it ineffective in practice. Linear regression, while interpretable, can still be prone to overfitting if too many irrelevant or highly correlated features are included.

Feature selection is therefore critical. Methods such as Recursive Feature Elimination (RFE), principal component analysis (PCA), or simpler correlation-based feature pruning can help retain only

ISSN 2460-7258 (online) | ISSN 1978-1520 (print)
8 (1) September 2025 51-60

those predictors that significantly contribute to explaining price variation. This simplification reduces the risk of overfitting, improves computational efficiency, and enhances interpretability for decision-makers who must act on the model's outputs.

Regularization techniques also provide effective defenses against overfitting. Lasso (L1 regularization) and Ridge (L2 regularization) can shrink less important coefficients toward zero, encouraging the model to focus on the most informative variables. Elastic Net, which combines both penalties, offers a flexible compromise that is often well-suited to high-dimensional datasets with correlated predictors.

Equally important is robust model validation. K-fold cross-validation, which partitions the data into multiple training and testing splits, can yield a more reliable estimate of out-of-sample performance compared to a single train-test split. This technique is particularly valuable in time series forecasting, where data can exhibit strong temporal autocorrelation. When applying cross-validation to time series, it is important to use time-aware approaches — for example, rolling or expanding window cross-validation — to preserve the temporal ordering of observations and avoid look-ahead bias.

In addition to these technical safeguards, it is crucial to continuously monitor the model's performance after deployment. Chili price dynamics are inherently non-stationary and subject to sudden changes in supply, demand, and policy. Therefore, updating the model with fresh data on a rolling basis and retraining it periodically can maintain predictive accuracy and guard against performance decay.

By systematically applying feature selection, regularization, time-aware cross-validation, and ongoing model monitoring, a forecasting framework can be established that is not only powerful but also transparent and trustworthy. Such a framework would empower stakeholders across the chili pepper value chain — from farmers and traders to policymakers and consumers — to make data-driven decisions that promote sustainable agricultural practices, economic resilience, and food security.

4. Conclusion

This research underscores the potential of time series analysis as a powerful and interpretable method for forecasting chili price trends, by incorporating critical factors such as weather patterns, production volumes, transportation costs, and market demand. The model demonstrated strong performance in predicting periods of price increases, accurately identifying upward trends with high precision and recall. This suggests that the model can serve as a valuable resource for farmers, traders, policymakers, and other stakeholders in the agricultural sector, enabling them to make informed decisions and optimize their strategies during periods of significant price volatility.

However, the model faced notable challenges in forecasting periods of stable or declining chili prices, as evidenced by lower precision and recall in these cases. This limitation is largely attributed to class imbalance within the dataset, where instances of price increases are more prevalent. Such imbalance tends to skew the model's predictions, reducing its ability to identify and correctly classify periods of stability or decline. Accurately forecasting these periods is equally critical for strategic planning, cost management, and the efficient allocation of resources in agricultural markets.

To address these limitations, several improvements have been proposed. Resampling techniques, such as oversampling the minority class (stable or declining price periods) or undersampling the majority class (price increases), can help balance the dataset and improve the model's capacity to recognize underrepresented patterns. Additionally, applying regularization methods, including L1 (Lasso) or L2 (Ridge) regularization, can help manage model complexity, reduce the risk of overfitting, and enhance the model's adaptability to evolving market conditions. Strengthening the validation process through k-fold cross-validation can further provide a more robust evaluation of the model's predictive performance, minimizing overfitting and improving generalizability to unseen data.

Moreover, the model can benefit from additional feature engineering and variable selection strategies. Incorporating relevant predictors such as policy interventions, pest or disease outbreaks, international trade developments, and shifts in consumer preferences could significantly enrich the model's explanatory power and forecasting accuracy. Applying systematic feature selection techniques would ensure that only the most impactful variables are retained, improving both the interpretability and reliability of the model.

By addressing these limitations and integrating the proposed enhancements, time series analysis can evolve into an even more effective and comprehensive tool for forecasting chili price trends. Ultimately, this will support the development of more balanced and resilient strategies for managing both volatility and stability within agricultural markets, contributing to sustainable food systems and improved livelihoods for farmers and other market participants.

5. Suggestion

To enhance the performance and applicability of time series analysis in forecasting chili price trends, it is crucial to address the challenges posed by class imbalance in the dataset. This issue arises when periods of significant price increases are overrepresented compared to stable or declining periods. This imbalance can result in a model that struggles to accurately predict stable or declining price trends, leading to inefficient market strategies and missed opportunities for timely interventions.One effective strategy to address this challenge is the use of resampling techniques. The Synthetic Minority Oversampling Technique (SMOTE), for example, can be employed to generate synthetic data points for the minority class—periods of stable or declining prices—thereby balancing the dataset and improving the model's ability to forecast these periods. Alternatively, undersampling the majority class, which represents periods of price increases, can reduce bias and help the model perform more effectively across all classes.Another approach is class-weighted regression, where the model assigns greater importance to periods of stable or declining prices during training. This adjustment improves the model's sensitivity to these underrepresented periods, enhancing its reliability in forecasting price stability or decline.To further improve predictive accuracy, incorporating additional features that influence chili prices is essential. For instance, factors such as weather conditions, pest outbreaks, transportation costs, and production volumes can provide a deeper understanding of the drivers behind price fluctuations.

Real-time sentiment analysis from market reports or news can also offer insights into supply and demand trends, which significantly affect price movements.Other relevant variables, such as government policies, regional market dynamics, and seasonal demand patterns, can further enrich the model. Feature engineering, such as creating interaction terms or applying non-linear transformations, can capture complex relationships between these factors, enhancing the model's ability to predict chili price trends effectively.Adopting strong validation techniques, such as k-fold cross-validation, ensures consistent model performance across various data subsets, reducing overfitting and improving generalizability. Testing the model on independent validation datasets also confirms its real-world applicability and reliability.By implementing these strategies, time series analysis can become a more robust and effective tool for forecasting chili price trends, supporting farmers, traders, and policymakers in making informed decisions during both stable and volatile market conditions.

Declaration of Competing Interest

We declare that we have no conflict of interest.

References

[1].Ateş, F. (2018). Wine quality detection through machine learning algorithms. IEEE Xplore. https://ieeexplore.ieee.org/document/9009111.

[2].Thakur, M., & Wadhwa, P. (2018). Red wine quality prediction by supervised learning. JournalofDataScienceandEngineering.https://drpress.org/ojs/index.php/HSET/article/view/13586.

[3].Zhou, C., & Wang, X. (2018). Forecasting red wine quality: A comparative examination ofmachinelearningapproaches.ResearchGate.https://www.researchgate.net/publication/377818483_Forecasting_red_wine_quality_A_comparative_examination_of_machine_learning_approaches.

[4].Jung, M., & Lee, Y. (2018). Red wine quality prediction using machine learning techniques. UBC Master of Data Science. https://ubc-mds.github.io/Red-Wine-Quality-Prediction/.

[5].Mishra, R. (2018). Multiple linear regression - Wine quality. Kaggle. https://www.kaggle.com/code/ranjitmishra/multiple-linear-regression-wine-quality.

[6].Singh, R., & Kaur, J. (2018). Analysis of wine quality dataset. StackOverflow. https://stackoverflow.com/questions/52037745/analysis-of-wine-quality-dataset.

[7].Prasad, A., & Chahal, R. (2018). Wine quality prediction - Machine learning. GeeksforGeeks. https://www.geeksforgeeks.org/wine-quality-prediction-machine-learning/.
[8].Patel,R.,&Patel,S.(2018).Winequalitypredictor.RPubs.https://rpubs.com/shradhit/winequality
[9].Tata, P. (2018). Wine quality prediction using logistic regression. YouTube. https://www.youtube.com/watch?v=UkzV1e4tRyk.
[10].Lundy, D.(2018). A practical guide to wine quality prediction using logistic regression. Medium.https://medium.com/%40daniel.lundy.analyst/a-practical-guide-to-wine-quality prediction-using-logistic-regression-f390c5c4d71f.
[11].Kaur, N., & Agarwal, A. (2018). Wine quality prediction using logistic regression. Kaggle. https://www.kaggle.com/learn/predicting-wine-quality-using-logistic-regression.
[12].Singh, J., & Sharma, S. (2018). Wine quality prediction using machine learning and logistic regression. Journal of Data Science and Analytics, 10(1), 25-34.
[13].Chaudhary, M., & Verma, R. (2018). Wine quality prediction using logistic regression: A comparative research. Journal of Machine Learning and Analytics, 8(3), 48-55.