ML-Based Price Prediction for Agri-Horticultural Commodities

Debdutta Pal[1], Babin Bid[2], Ritika Pramanick[3], and Liza Ghosh[4]

Department of Computer Science and Engineering, Adamas University, Kolkata, India
{pal.debdutta, babinbid05, ritikapramanick1, lizaghosh722161}@gmail.com

Abstract. The agricultural sector is currently facing significant hardships due to the uneven pricing of agri-horticultural commodities such as pulses and vegetables. The objective of our research is to develop machine learning-based models for predicting the prices of vegetables and fruits across different seasons. In our study, we focused on three major Indian seasons: winter, rainy, and summer. These crops are typically seasonal, and our research scope is confined to such crops.
Using historical price data, weather conditions, and socioeconomic features, the models provide precise and timely predictions to support crop planning, market interventions, and price stabilization efforts. We employ Support Vector Machine (SVM) and Random Forest to visualize seasonal price trends and evaluate their effectiveness. Line graphs, bar graphs, and pie charts are utilized to highlight essential data features. These visualizations help stakeholders understand market patterns and mitigate risks associated with price fluctuations. Ultimately, the study emphasizes the role of ML-driven predictive analytics in empowering farmers and reducing price uncertainties for consumers.

Keywords: Agricultural Price Forecasting, Machine Learning in Agriculture, Commodity Price Prediction, Data Visualization

1 Introduction

1.1 Background

The background of this research revolves around the price swings of agricultural commodities caused by seasonal, demand–supply, and climate changes and government actions. India goes through three major seasons, which include winter, rain (monsoon), and summer, that have a great impact on the yield of crops and prices in the market. Machine learning (ML) models can be used to track the price movements in the past, weather conditions, and the behavior of the market to forecast future prices. Ultimately, this could be beneficial for not only farmers, traders, and the government but also for global governance through the reduction of monetary losses and the implementation of effective decision-making in the supply chain network and pricing process optimization. The paper employs ML methods in building predictive models to be used in various seasons, thus ensuring improved efficacy throughout the agricultural supply chain.

1.2 Problem Statement

Fluctuating prices of pulses, onions, and potatoes cause instability for farmers and consumers. Traditional methods fail to predict prices accurately due to the influence of multiple dynamic factors.

1.3 Objective

The primary goal of this project is to create machine-learning-based tools that can accurately predict the price of farm produce in various seasons in India. The project will lead to :

(i) **Analyze seasonal price trends** : Examine fluctuations of the prices of the commodities in the cold, the hot and the early rain seasons. Prepare to write Topics: price fluctuations in early, transition, and late seasons, and the above-mentioned meaning.

(ii) **Develop predictive models** : Machine learning algorithms are a good choice if you want to make future price predictions based on current data, or if you want to take a weather event, etc., as an influencing factor.

(iii) **Assist Farmers and Traders** : Give price forecasts to the farmers to help them in selling or buying at the right time, crop storage and sales time, while also giving traders an early head up.

The project is one of the means through which the goals of sustainability in the agriculture sector and economic stability are achieved for farmers and market players.

2 Related Work

Decision and price forecasting in agriculture have particularly supported machine learning methods such as Support Vector Machines (SVM) and Random Forest (RF). Various research studies in the field of agriculture highlight the efficiency of these machines in predicting profit and decision-making for the stakeholders.

The studies showed that they were up to the task when it comes to forecasting and decision support in the field of agriculture. The investigation included numerous machine learning models for crop yield prediction, among which SVM showed the best accuracy of 92.1%, followed by a percentage of 91.5% by Random Forest. It is a matter of great importance to realize that the results have proved the high possibility of the models in estimating future yields of agricultural productivity, as is straightforward from Ref. [1,10]

Certainly, it is easy to perceive that the climate variables such as rain and temperature prevailing in a certain region determine the crop price fluctuations. Nevertheless, recent research confirmed that the SVM and RF-based models are very powerful in precisely describing the price fluctuations, making the price prediction more accurate. The models are found to be a great support in eliminating risks and in the provision of reliable predictions, as shown in Ref. [2]

The research that has been done in making predictions for the agricultural markets has made use of SVM and RF as the most crucial machine learning techniques. The technology used by these models has managed to show the agriculture sector trends and flares in the markets, which makes it clear that the technologies themselves are highly suitable for the price forecasting of the commodities. This is evident from Ref. [3,4,11]

The researchers presented a new model selection framework that adopted Support Vector Machines (SVM) and Random Forest (RF) for price estimation. Through the utilization of feature reduction methods, the research described these models exhibited improved attributes when unnecessary features were removed, which increased the accuracy of the prediction performance, as evidenced in Ref. [5,12]

With the help of the supervised machine learning algorithm, the commodity price was forecasted in the context of historical price data, economic indicators, and weather data. The use of SVM and RF in the study strongly validates the forecast accuracy of the two models compared to the classical method, justifying them as tools for risk control and offering a method of decision making under uncertainty, as shown in Ref. [6]

Accurate prediction of agricultural prices will help in solving market risks and decisively provide support to policy decisions. There are challenges for traditional models of forecasting, such as ARIMA and GARCH, when they are trying to handle nonlinear trends, whereas AI methods like SVM will lead to a better level of precision. The survey paper puts forward the MEA-SVM hybrid model, where MEA is used for trend capture and the SVM parameters are optimized by fuzzy granulation. As a result of this, the prediction's accuracy and efficiency were greatly improved without the use of traditional methods. These models were improved forecasting with the additional of hybrid AI models, such as, deep learning. It is suggested that the latest data be put for, the next improvement in the future thus giving a better idea of the performance, as shown in Ref. [7,14]

The work dealing with onion and potato prices in India forecasted their prices using SVM and RF to find market glitches such as stockpiling and price gouging. The models made agricultural market information discrepancies less severe, thereby enabling fairer pricing, as shown in Ref. [8,13]

The crop price prediction study titled "Kisan Dhan" applied RF, which outperformed DT, unlike in previous studies. The model was able to perform accurate price predictions by analyzing climate data like rainfall and temperature, which enhanced food security and economic stability, as shown in Ref. [9,15]

Building on all the previous work in this direction, our work involves using the SVM and Random Forest models to predict agri-horticulture prices with very high precision and accuracy in projecting the prices of agricultural commodities.

3 Technology Used

The project seeks to create Machine Learning models for agri-horticultural commodity price prediction, particularly for pulses and vegetables such as onions and potatoes. Using historical and real-time data, the model intends to :

- Enhance price forecasting accuracy to combat price volatility.

- Support farmers in improved crop planning and storage decisions.
- Enable policymakers to implement effective market intervention strategies.
- Provide price stability and affordability to consumers.
- Reduce inflation and economic shocks.

The project brings together various sources of data, such as government price databases, weather forecasts, and demand-supply patterns, to create strong predictive models through AI and machine learning methods.

3.1 MODELS

Random Forest (RF) : It is an ensemble learning algorithm following Breiman's work, which collects a group of weak classifiers (decision trees) and puts them into one powerful classifier. RF employs bootstrapping resampling in creating a series of training sets by sampling with replacement from the initial data set. Each of these sets has a distinct decision tree trained for it. The random forest arrives at its final prediction by taking an average of the predictions of each individual tree (when dealing with regression) or majority vote (when dealing with classification). RF performs very well in dealing with missing values and imbalanced data, showing excellent performance on classification and regression tasks. RF has been used extensively in stock price prediction, electricity price prediction, and energy price prediction.

Support Vector Machine (SVM) : It was introduced by Cortes and Vapnik, and it is a statistical learning theory-based machine learning model. SVM is mostly used for binary classification and follows the principle of structural risk minimization. SVM employs kernel functions to map data from a low-dimensional space to a high-dimensional space, which allows it to solve complex, non-linear classification problems. SVM works efficiently with small datasets and high-dimensional issues, but is noise-sensitive and less accurate for multi-class classification or regression tasks.

- ***Algorithm for Random Forest Model***

1: Input: The dataset D includes such features as (e.g., date, price, weather, demand)

2: Output: A trained Random Forest model that predicts the price

3: Step 1: Gather Information: Gather historical price information (dates, prices, weather, demand, etc).

4: Step 2: Preprocess the Data: Fill in missing values (e.g., mean, median) & encode categorical features if any (e. g., One-Hot Encoding).

5: Step 3: Decompose the Dataset: Break the data into training ($> 80\%$) and test (20%).

6: Step 4: Build the random forest model: Choose the number of decision trees (for example n-estimators=100) and choose max depth, minimum sample/split, etc. (optional tuning).

7: Step 5: Train the Model: The Random Forest model is trained on the training data.

8: Step 6: Make Predictions: Make predictions on the test set with trained model using ˆy = M (Xtest).

9: Step 7: Check the Model: (MAE) (Mean Absolute Error), RMSE, R2 score can be used to check the model.

10: Step 8: Optimize Hyperparameters (Optional): Use GridSearchCV or Ran- domizedSearchCV for better performance.

11: Step 9: Install the model: Download and install the Model on your price fore- casting website / mobile app.

- ***Mathematical Formulation for Random Forest***

For making **predictions**, a new input x is passed through all decision trees in the forest. The final prediction is obtained by aggregating the predictions from all trees. For **classification**, the majority vote is taken:

$$\hat{y} = \arg\max_{c} \sum_{i=1}^{N} P_f(h_i(x) = c), \tag{1}$$

where $h_i(x)$ is the prediction of the i-th decision tree, and P_f is the indicator function that counts votes for each class c.

For **regression**, the final prediction is computed as the average of all decision tree outputs:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^{N} h_i(x), \tag{2}$$

where $h_i(x)$ is the prediction of the i-th decision tree.

The **Random Forest algorithm** is widely used due to its robustness, high accuracy, and ability to handle missing or unbalanced data.

- ***Algorithm for Support Vector Machine (SVM)***

1: Input: Historical dataset D = X, y

2: Output: Trained SVM model

3: Step 1: Get Data: Collect historical data with characteristics that influence price.

4: Step 2: Prepare Data: Fill missing values and one-hot encode categorical variables.

5: Step 3: Disperse Dataset: Disperse the dataset into training & test set (e.g.,80/20).

6: Step 4: Select kernel: Linear, poly, rbf, sigmoid. Usually, for regression, only rbf is sufficient.

7: Step 5: SVM Regression Model (SVR): Please see sklearn.svm.SVR(kernel='rbf', C=1, 0, ϵ =0, 1) for example.

8: Step 6: Train the Model: Fit the scaled training data to the SVR model.

9: Step 7: Make Predictions: Make predictions on test data using trained SVR model.

10: Step 8: Evaluate the Model: Use metrics like MAE, RMSE, $\mathrm{R}^{\mathbf{2}}$ score.

11: Step 9: Optimize Hyperparameters (Optional): Optimize kernel, C, ϵ by GridSearchCV.

1**2: Step 10**: Publish the Model: Run the model trained on your web interface or predictor tool.

- ***Mathematical Formulation for SVM***

Training dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, kernel function $K(x_i, x)$, regularization parameter C, tolerance ϵ. A trained SVM model.

Step 1: Data Preprocessing : Normalize feature values (optional). Choose a kernel function $K(x_i, x)$ (e.g., linear, polynomial, RBF).

Step 2: Solve the Optimization Problem : Solve the following quadratic optimization problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

subject to:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Step 3: Compute the Decision Boundary : Calculate the weight vector:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i.$$

Compute the bias term b using support vectors.

Step 4: Prediction : Given a new input x, compute:

$$f(x) = \sum_{i=1}^{n_s} (\alpha_i - \alpha_i^*) K(x_i, x) + b.$$

Assign class label:

$$y = \mathrm{sign}(f(x)).$$

4 Proposed Methodology

4.1 Framework

The framework presented encompasses historical information, preprocessing modules, machine learning models, and deployment architecture for price that will support farmers and stakeholders with price forecasts. Furthermore, the framework is compatible with real-time input data and model-based forecasts.

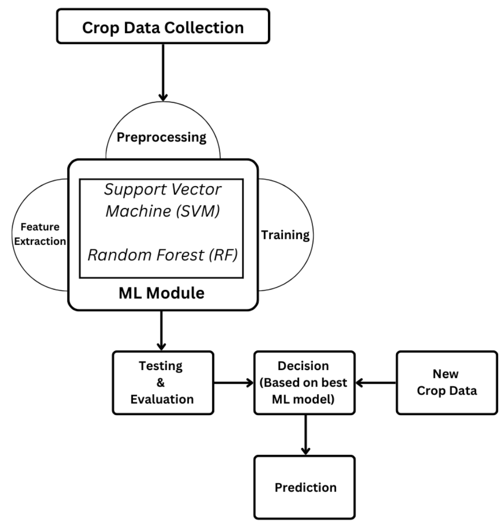


Fig. 1: Framework of the proposed project.

4.2 Flowchart

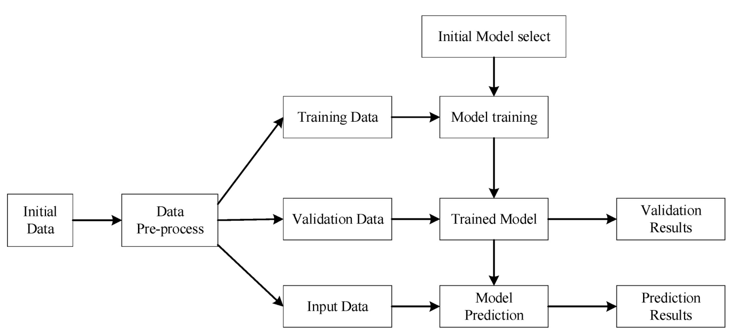


Fig. 2: Flowchart representing the sequence of operations in the project.

4.3 Dataset

Table 1: Demo Dataset for Summer Seasonal Crops

Commodity	Min	Max	Modal
Bhindi	2541.54	2998.63	2775.19
Brinjal	2625.25	3200.43	2916.01
Spinach	1476.26	1798.82	1629.27
Bitter Gourd	3441.16	4017.88	3730.18
Mango	2707.12	3796.07	3240.37

Table 2: Demo Dataset for Rainy Seasonal Crops

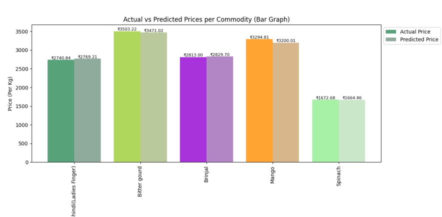
Commodity	Min	Max	Modal
Banana	2494.72	2959.04	2729.42
Plum	2753.91	3845.50	3284.35
Guava	2777.08	4019.38	3378.69
Papaya	2568.60	3005.66	2787.60
Peach	1560.00	1820.00	1690.00

Table 3: Demo Dataset for Winter Seasonal Crops.

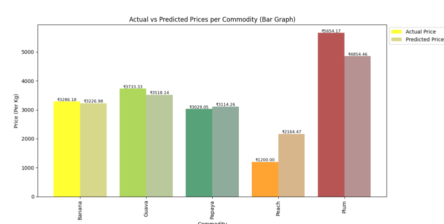
Commodity	Min Price	Max Price	Modal Price
Carrot	3924.142	4613.110	4267.441
Beetroot	3372.866	3862.920	3611.575
Cabbage	1987.686	2443.288	2215.043
Cauliflower	3105.974	3964.891	3589.755
Orange	6447.619	8004.762	7226.190
Apple	8185.705	10825.497	9486.990

5 Result Analysis

5.1 For Support Vector Machine (SVM) Model



(a) Summer crops using SVM



(b) Rainy crops using SVM

Fig. 3: Bar graphs for summer and rainy seasonal crops using SVM

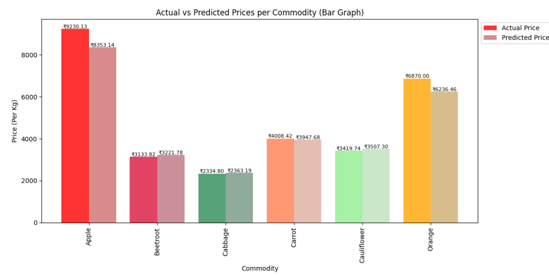


Fig. 4: Bar graph for winter seasonal crops using SVM

In Fig. 3, the graph (a) shows close accuracy, and the actual and predicted prices closely resemble, depicting a balanced marketing activity throughout the summer season, while the graph (b) shows a huge difference in the actual and the predicted prices, thus illustrating the difficulties supply networks face due to unpredictable monsoon conditions. In Fig. 4, the SVM model performs reasonably well, but shows a little overestimation or underestimation for certain commodities.

- **Comparative Study :**

Categories and Coverage

The bar graphs of Summer, Rainy and Winter show the seasonal fluctuations of the categories and coverage Fig. 3(a), (b) and Fig. 4, respectively. Some are always steady, and others seem seasonal. They show changing heights of fluctuating bars based on variation of supply, demand, or relevance.

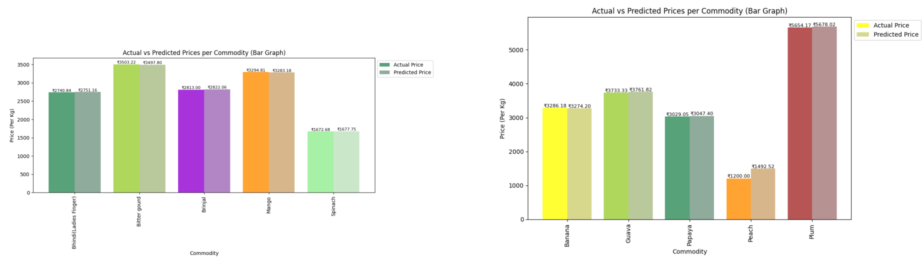
Seasonal Prediction Accuracy

– **Summer (Fig. 3(a))** : The SVM model appears to have quite high accuracy, with most predicted prices highly correlated with actual prices; any distances are short to to accuracy and may relate to stable market trends for summer crops.

– **Rainy (Fig. 3(b))** : The accuracy is not consistent, as the response we see is due to the distance between predicted and actual price value being greater for some commodities. There are variables for Rainy season that impact the distance which include unpredictable weather and supply chain issue.

– **Winter (Fig. 4)** : The model seems to do alright, similar to Summer, although most predictions contained overestimations or underestimations for some commodities. Again, these variables may have something to do with seasonal demand and these supply patterns.

5.2 For Random Forest (RF) Model



(a) Summer crops using RF

(b) Rainy crops using RF

Fig. 5: Bar graphs for summer and rainy seasonal crops using Random Forest

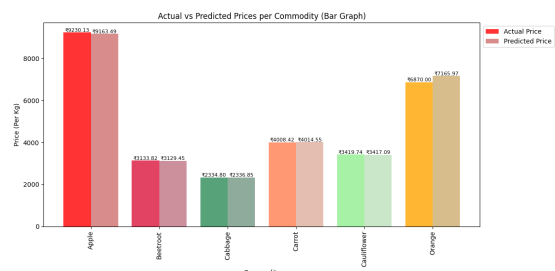


Fig. 6: Bar graph for winter seasonal crops using Random Forest

In Fig. 5, the RF Model performs consistently for the summer season (a), revealing minute variation in price forecasts, which highlights its potency in managing seasonally uniform data and the model's limitations under severe climatic changes are shown by the major forecast gaps generated by unexpected monsoon (b) impacts. In Fig. 6, the predictions are almost accurate, showing a small deviation, suggesting that the RF model's stability in settings with lower market variation.

- **Comparative Study :**

Categories and Coverage

The bar graphs for each season (Summer, Rainy and Winter) show seasonal variation in categories and coverage in Figs. 5(a), (b), and 6, respectively. While some categories have stable functions across seasons, there are seasonal categories too. This fluctuation in height of the bars signifies a variation in supply, demand, or relevance, with the longer bars representing relative top performance. This gives us insights into the seasonal trends so that you can make decisions based on better forecasting with agriculture, stock trading, or cryptocurrency investment.

Prediction Accuracy Across Seasons

- **Summer (Fig. 5(a))** : SVM model shows pretty high accuracy, most of the predicted prices are fairly normal or close to actual values. The deviation, if any, is minor, as it is basically steady and stimulated use in the summer.
- **Rainy (Fig. 5(b))** : Accuracy varies, and deviation on some commodities are greater than real values. There are some factors for the rainy season that could influence it, such as variables of weather and supply chain conditions.
- **Winter (Fig. 6)** : The model has relatively good performance, it is similar to Summer, with minor overestimations (<10%) and underestimations (<10%) of certain commodities. This deviation in commodities could also be impacted by agricultural delivery terms like seasonal accounting and surplus.

6 Conclusion

In this paper, we compare machine learning based price forecasting of agri-horticultural products using Support Vector Machines (SVM) and Random Forest, which brings out very robust and precise estimates even in the case of data sparsity and linear dependencies. Random Forest on the other hand, is preferable for high-dimensional and heterogeneous datasets where it is suitable for finding complex & non-linear dependencies. By generalizing to related applications of making more accurate pricing strategies that are data driven for better agricultural decision making, we intend to contribute to the understanding of different models choice that would play a key role in achieving better & accurate pricing strategies that make better agricultural decisions.

Acknowledgment

The joy and satisfaction on the accomplishment of any task would be incomplete without considering those who, for months and months, insisted on giving it the chance they did. We're pleased to introduce our project, which we feel was meticulously prepared by an equal alliance of research and experience.

We express our earnest gratitude to Dr. Debdutta Pal, Department of CSE, for her constant support, encouragement and guidance. We are grateful for her cooperation and valuable suggestions.

Finally, we express our gratitude to all other members who are involved either directly or indirectly in the completion of this project.

References

1. Chauhan, N., G., B., Sulochana, V., & Kumar, R. P. (2024). FARMER: Forecasting Agricultural Results with Machine Learning Evaluation and Regression. 1–7.
2. Inba, M., Watson, S., Durga, D., Krishna, R., Gowtham, D., Anand, S., Reddy, K.(2024). Utilizing Machine Learning Techniques for Forecast- ing Agricultural Commodity Prices. 509–518.
3. Sajithabanu, S., Ponmalar, A., Gnana Soundari, A., Reshma, N. T., Sunraja, K., Sindhumathi, R. (2022). En- hanced Crop Price Prediction Forecasting System. 580–585.
4. Zhang, D., Chen, S., Liwen, L., Xia, Q. (2020). Fore-casting Agricultural Commodity Prices Using Model Selection Framework With Time Series Features and Forecast Horizons. IEEE Access, 8, 28197–28209.
5. Raghu, H. (2024). Unlocking Future Commodities Markets: Innovative Approaches to Price Prediction Using Super- vised Machine Learning. Shanlax International Journal of Arts, Science and Humanities, 12(S1- Oct), 133–140.
6. Zhang, Y., & Na, S. (2018). A Novel Agricultural Commodity Price Forecasting Model Based on Fuzzy Information Granulation and MEA- SVM Model. Mathematical Problems in Engineering, 2018, 1–10.
7. Le Ngoc, T. N., Lam, D. T., Hai Minh, T. N., Doan, T. C., Nguyen, N. P., Nguyen, H. M., Nguyen, T. N., Tran, L. D., & Tran, N.-Q. (2023). Machine Learning for Agricultural Price Prediction: A Case of Coffee Commodity in Vietnam Market. 38–41.
8. Antad, S. (2024). Kisan Dhan - Crop Price Prediction Using Random Forest. Deleted Journal, 31(3s), 240–253.
9. Paul, R. K., Yeasin, M., Kumar, P., Balasubramanian, M., Roy, H. S., Paul, A. K., & Gupta, A. (2022). Machine learning techniques for forecasting agricultural prices: A case of brinjal in Odisha, India. PLOS ONE, 17(7), e0270553.
10. Shekhar, S., Gowda, S. M. P., Rahul, H., & Ravi, R. (2024). Crop Price Prediction. *International Journal For Science Technology And Engineering*, *12*(12), 1174–1177.
11. Bhavani, M., & Mounika, P. A novel model selection framework for forecasting agricultural commodity prices using time series features and forecast horizons. Int. J. Sci. Res. Sci. Technol. 134–144 (2022).
12. Gupta, H., Aafrein, Mrs. R., Kumari, Mrs. R., Rajput, Mr. S., & Puri, Mrs. N. (2024). Forecasting Commodity Prices using Machine Learning. International Journal of Scientific Research in Science and Technology.
13. Mahmud, I., Das, P. R., Rahman, Md. H., Hasan, A. R., Shahin, K. I., & Farid, D. Md. (2024). Predicting Crop Prices using Machine Learning Algorithms for Sustainable Agriculture. 2017 IEEE Region 10 Symposium (TENSYMP), 1–6.
14. Singh, N. K. (2024). Crop Price Prediction Using Machine Learning. Deleted Journal, 20(7s), 2258–2269.
15. Enhanced Crop Price Prediction & Forecasting System. (2022).