Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 6, 2026

MACHINE LEARNING-BASED CLASSIFICATION OF AGRICULTURAL COMMODITY PRICES: A COMPARATIVE STUDY OF RANDOM FOREST, LOGISTIC REGRESSION AND SUPPORT VECTOR MACHINE USING PAKISTANI MARKET DATA

Haroon Khan[*1], Muhammad Ismail[2]

[*1,2]*MS Climate Change and Environmental Informatics, National University of Technology, Islamabad- Pakistan*

[*1]haroonkhan@aup.edu.pk, [2]abbasiismail08@gmail.com

DOI: https://doi.org/10.5281/zenodo.20773051

Keywords
Agricultural Economics, Crop Price Classification, Machine Learning, Random Forest, Support Vector Machine, Logistic Regression, Pakistan.

Article History
Received: 22 April 2026
Accepted: 04 June 2026
Published: 20 June 2026

Corresponding Author: *
Haroon Khan

Abstract

The price of crops and farm products depends on the time, market conditions, the volume of food produced, and the nature of the food purchased. An understanding of these prices is valuable for farmers, crop buyers and sellers, and the government and other students of farm economics when making decisions. The agricultural commodity prices are indeed significant. What farmers, traders, policy makers and agricultural economists need to know are these prices? The idea of this study is to determine agricultural commodity prices into 3 classes (Low, Medium and High) using machine learning technique considering the historical price data of the market in Pakistan. The data set includes over 411,000 valid observations from the Mango and Apple (Golden) markets in 138 cities over a 15-year period (2007-2022). First, continuous price values were converted into three balanced classes by using a quantile-based approach in order to formulate the classification task. Feature engineering techniques were used to create temporal features (year, month, and season) and numerical features for categorical features (crop type and city). Three supervised machine learning algorithms were trained and tested, namely: Random Forest, Logistic Regression and Support Vector Machine (SVM). The data set was split into 80% training and 20% testing. The performance of the model was evaluated based on the standard evaluation metrics such as accuracy, precision, recall, F1-score and confusion matrix analysis. The results indicated that the Random Forest model performed better than the other models, with an accuracy of 81.78% and an F1 score of 0.8189. SVM and Logistic Regression had comparatively low predictive accuracy. The results show that using ensemble learning techniques to capture temporal and spatial changes in agricultural market data is more suitable. The agriculture sector can greatly benefit from machine learning applications in price analysis and market prediction, as shown in this study that demonstrates its applications in agriculture as proof of concept. The suggested framework offers valuable lessons in the designing of Data-driven Decision Support Systems (DDSS) to enhance the monitoring of crop prices, market intelligence and agricultural policy making in Pakistan.

INTRODUCTION

In Pakistan Agriculture is one of the most significant sectors of economy as it plays a major role in generation of employment, food security and national income. The agricultural markets play a crucial role in determining the profitability of agricultural operations and the agricultural price changes directly affect the farmers, traders, consumers, and the government policy [15, 16]. Therefore, accurate prediction and categorization of agricultural commodity prices are essential to promote transparency in the marketplace and improve decision making.

The fruit crops, mango and apple are significant components of the horticultural crop sector in Pakistan. Mango is one of the most important fruit export commodities in the country and the apple industry contributes to the agricultural economy of the northern parts of the country. These commodities are seasonal in production, climatic conditions, transportation costs, market demand and supply differences in regions affect prices. All those concerned in production planning, storage management and marketing strategies are interested in these price dynamics.

The conventional statistical methods have been widely applied for the analysis of trends in agricultural prices.

In recent years, however, with the growing availability of datasets at a larger scale, machine learning techniques including the ability to find complex relationships between variables have become more popular [4, 5]. Non-linear interactions and hidden patterns in the data can be effectively captured with the use of machine learning algorithms, and they may be more predictive than traditional methods [6], [16]. In recent years, applications of machine learning in agriculture have also been found to be effective, such as predicting the yield of crop, identifying the disease, measuring the fertility of soil, controlling the irrigation and forecasting the market [4]–[8].

Amongst those methods, Random Forest, Logistic Regression and Support Vector Machine have become popular algorithms for their robustness, interpretability and prediction strength [1] [2] [18].

The aim of this research is to create a machine learning-based structure for pricing classification of agricultural crops based on historical market data throughout Pakistan. To specifically achieve, the study is going to:

1. Create a well-balanced multi-class crop price classification dataset.
2. Identify temporal and spatial characteristics of agricultural markets.
3. Compare the performance of models – Random Forest, Logistic Regression, and SVM.
4. 4. Determine the best machine learning method used for classification of prices in agriculture.

This study yields insights into market behavior and will be useful for a wider range of studies on the use of machine learning in agricultural economics.

RELATED WORK

In agriculture, the use of machine learning has grown significantly in the last decade, including tasks such as forecasting yields, diagnosis of diseases, determination of soil fertility, irrigation control, and forecasts of agricultural markets [4], [8]. With the growing availability of such large scale agricultural data, researchers have been tempted to apply data driven methods that could capture the meaningful patterns of complex and dynamic agricultural systems.

Numerous efforts have been made to develop models for forecasting agricultural commodity prices for its importance for farmers, traders, and policy makers and supply-chain stakeholders. Many traditional statistical models have been used to analyze commodity prices and predict commodity prices such as Autoregressive Integrated Moving Average (ARIMA) models and exponential smoothing models. Such approaches work reasonably well for linear time-series data but are not effective with agricultural markets' nonlinear relationships and complex interactions [15].

Machine learning techniques are being used more often to solve agricultural price prediction and categorization issues in order to get around these restrictions. Among the most popular ensemble learning algorithms was first described by Breiman [1] and is known as the Random Forest algorithm. This is because it is robust and can process large datasets, is not prone to overfitting, and can

capture nonlinear relationships. Random Forest has been proven effective in various agricultural applications such as yield prediction, commodity price prediction, agricultural risk assessment and market intelligence [7], [9], [13].

The SVM algorithm [2] proposed by Cortes and Vapnik has been the other popular model used in agricultural analytics. SVM has been proved to be effective in a variety of applications such as high dimensional feature space crop classification, disease detection, yield estimation and market analysis [14, 18]. The capability to model nonlinear patterns is further improved with the use of kernel functions, typical in agricultural data. Logistic Regression is still one of the most used base classifiers due to its simplicity, ease of computation and interpretability [18]. Logistic Regression is not as effective for very non-linear data, but is a good baseline for comparison to more sophisticated machine learning methods.

In recent years, there is a growing realization of the importance of feature engineering in agricultural market analysis. Apart from the spatial variation in agricultural markets, spatial variation in location, transportation infrastructure, and regional supply and demand conditions may also influence prices and contribute to the spatial variation of prices [12, 13, 16]. Temporal variables, such as year, month, season, and long-term market trends, also have significant impacts on commodity prices. Incorporating such features into machine learning models has been demonstrated to boost the predictive efficiency and to make models more interpretable.

While significant advances have been made in agricultural price forecasting, most current research is limited to short-term prediction, a single commodity or a relatively small number of prices. Moreover, there is little research on how to perform a large-scale, multi-class classification of the prices of agricultural commodities based on long time series data covering multiple regions. To address this gap, this study proposes to design a machine learning approach to determine the agriculture commodity prices classification based on over 411,000 observations from Mango and Apple markets of 138 cities in Pakistan for a 15-year period. The systematic comparison of the performance of Random Forest, Logistic Regression and Support Vector Machine indicates the optimal solution for using these models in large-scale agricultural market intelligence applications.

DATASET DESCRIPTION

Data Source

In this study, the data was collected from Agricultural Market Price records of all over Pakistan. Two fruits which are economically important, namely Mango and Apple (Golden) were selected for study as they play an important role in the horticultural sector. The data came from around 803,325 data points gathered from 138 cities over a 15-year period (2007–2022). The data presented in this study has been summarized in Table 1.

For each record, there were four major attributes:

- City Name
- Date of Observation
- Crop Type
- Market Price (PKR)

The data is temporally and geographically comprehensive, and can therefore be used to analyze the price trends of the crops in various regions and seasons.

Table 1. Dataset Summary

Attribute	Value
Crops	Mango, Apple (Golden)
Cities	138
Period	2007–2022
Original Records	803,325
Clean Records	411,359
Features	5
Classes	Low, Medium, High

Data Cleaning

At first looking, it was established that there were records that contained 0 price points. The records were excluded from the analysis because they did not reflect any actual market transactions and were not in the form of such values. Following data cleaning, 411,359 valid observations were left for further modeling and evaluation.

Feature Engineering

The original data was expanded and several additional features were created to enhance model performance.

Temporal Features

- Year
- Month
- Season

Months were divided into seasons as follows:

- Winter (December–February)
- Spring (March–May)
- Summer (June–August)
- Autumn (September–November)
- Categorical Features
- Crop Type
- City

We applied Label Encoding for both categorical variables.

Price Class Generation

Crop prices were converted to three equally-balanced categories using quantile based discretization as the classification algorithms used machine learning techniques that necessitate the presence of categorical target variables. The class distribution in the quantile-based categorization was almost evenly distributed as shown in Table 2 and Fig.1.

The target classes were determined as:

- Low Price
- Medium Price
- High Price

Table 2. Class Distribution

Class	Records	Percentage
Low	138,219	33.6%
Medium	136,925	33.3%
High	136,215	33.1%

This is a well-balanced distribution to avoid classification bias and enhance model reliability.

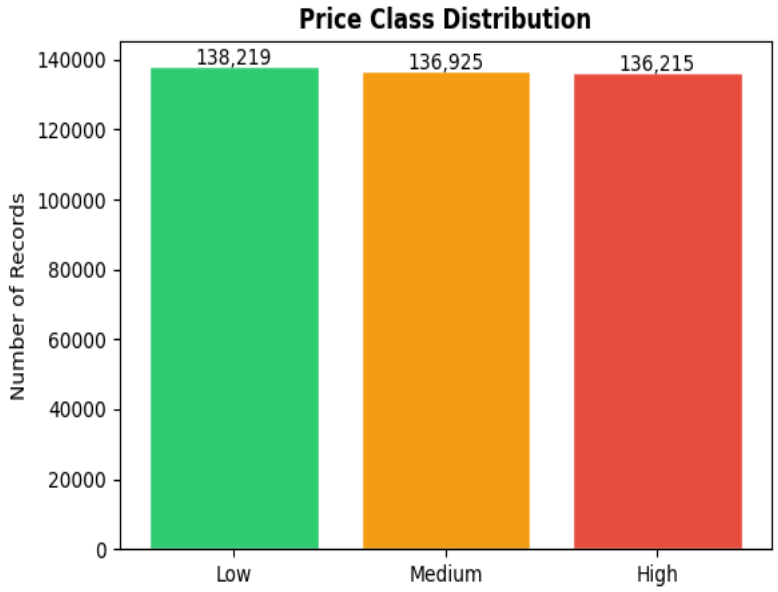

Fig. 1. Price Class Distribution

METHODOLOGY

The general approach adopted was a typical machine learning workflow as shown in Fig.2.

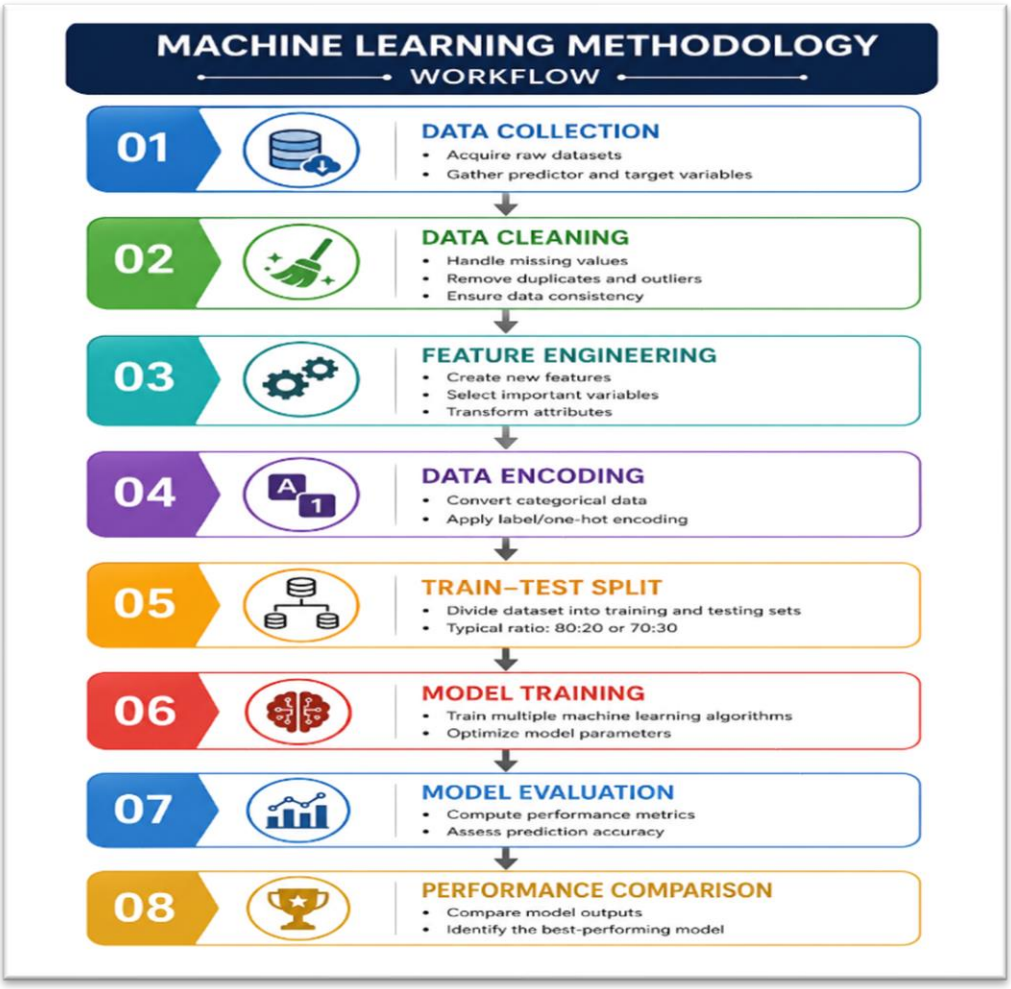

Fig. 2. Machine Learning Workflow

Data Splitting

The data set was split in the ratio of 80:20 for training and testing sets, respectively.

Training samples: 329,087
Testing samples: 82,272

Stratified sampling was used to make sure that the class size distributions in both subsets were representative.

Feature Scaling

Numerical features were standardized using Standard Scaler as it is sensitive to the magnitude of features for Support Vector Machine and Logistic Regression.

The features were standardized to become:

- Mean = 0
- Standard Deviation = 1
- Random Forest

Random Forest model was used as the primary model in this study. Random Forest ensemble learning was introduced by Breiman [1] to boost the accuracy of predictions and reduce overfitting by using multiple decision trees.

Advantages include:

- Robustness against overfitting
- Model nonlinear relationships (equationically and graphically) using examples.
- High predictive accuracy

- Feature importance estimation

The model was trained using:

- Number of trees = 100
- Maximum depth = 15
- Random state = 42

Logistic Regression

Logistic Regression was used as a baseline classifier because it is an easy-to-understand and clear classification model when there are multiple classes [18].

Model parameters:

- Maximum iterations = 500
- Random state = 42

Support Vector Machine

To find the nonlinear decision surfaces, Support Vector Machine (SVM) was used, as proposed by Cortes and Vapnik [2] with a Radial Basis Function (RBF) kernel.

Model parameters:

- Kernel = RBF
- C = 1.0
- Gamma = Scale

The SVM will try to find the most separating decision boundaries in feature space.

Evaluation Metrics

The models are tested with:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

These metrics are all comprehensive measures of classification performance.

EXPERIMENTS AND RESULTS

Overall Model Performance

Table 3 & Fig. 3 shows the comparison of the performance of the machine learning models evaluated.

Table 3. Model Performance

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.8178	0.8208	0.8178	0.8189
Logistic Regression	0.6208	0.6150	0.6208	0.6170
SVM	0.6601	0.6766	0.6601	0.6654

The outcome depicts that the performance of the Random Forest model is better than the other two models, namely Logistic Regression and SVM in all the evaluation metrics.

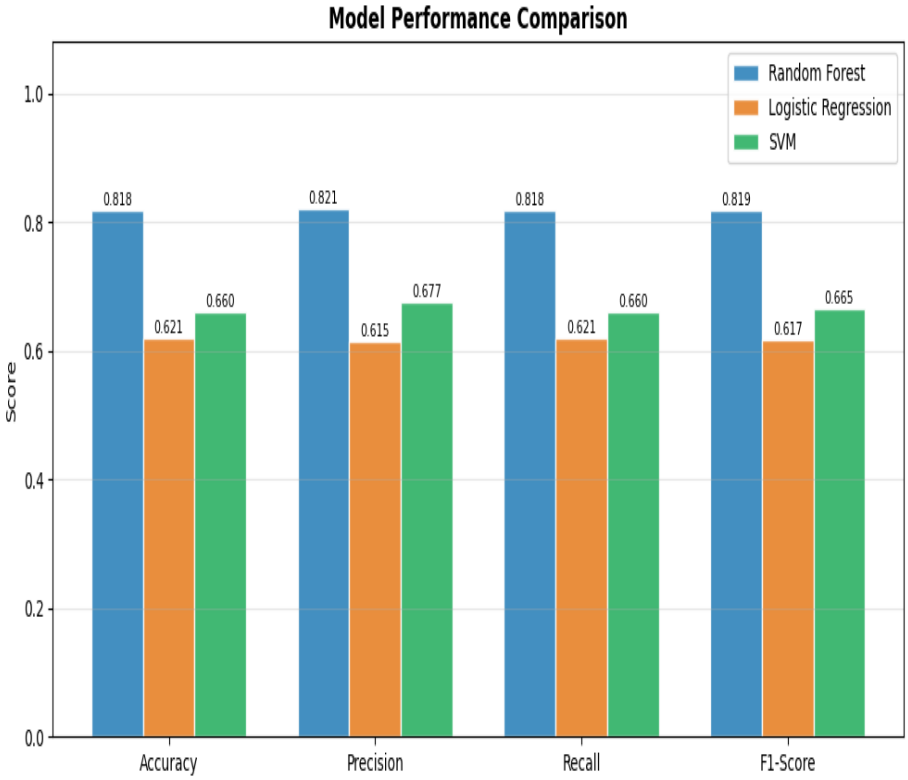

Fig. 3 Model Performance Comparison

Random Forest Performance

The random forest model (81.78%) showed the highest accuracy, highlighting its strong ability to recognize patterns in agricultural market data that are temporal and spatial in nature. The class-based performance was strong and predictive at all three price levels.

Logistic Regression Performance

The overall accuracy of Logistic Regression was 62.08%. Model failed to learn the nonlinear relationship that exists in the data and hence its classification performance was less than that of the Random Forest model.

Support Vector Machine Performance

SVM had an accuracy of 66.01%, while Logistic Regression had an accuracy of 64.49%, which is better than SVM, but still inferior to Random Forest. Overall, the RBF kernel demonstrated good performance in nonlinear modeling, but its performance was constrained by both computational requirements and the complexity of the datasets.

Confusion Matrix Analysis

The results of the confusion matrices showed that Random Forest gave the highest number of correct classifications for all the classes. Most of the misclassifications were between adjacent price categories (Low-Medium, and Medium-High), as this is to be expected, as the categories correspond to neighboring quantile bands.

Feature Importance Analysis

The Random Forest feature importance scores are shown in Table 4 % Fig. 4.

Table 4. Feature Importance (Random Forest)

Feature	Importance
Year	0.626
City Encoding	0.229
Month	0.060
Crop Encoding	0.058
Season	0.027

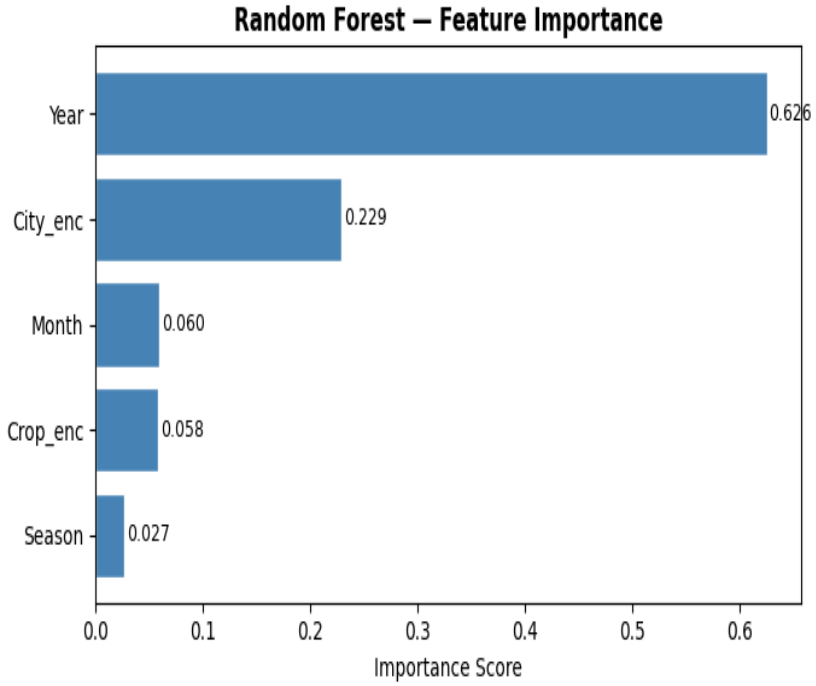

Fig. 4. Feature Importance (Random Forest)

The results indicate that long-term temporal trends represented by the Year variable were the most influential factors affecting crop price classification.

DISCUSSION

Findings from this research indicate the ability of machine learning approaches to categorize agricultural commodities pricing based on large amounts of data from historical market trends. Some of the machine learning algorithms used include Logistic Regression, Support Vector Machine, and Random Forest, which had accuracy levels of 81.78% and F1 scores of 81.89%. The accuracy level was considerably higher than the one achieved by Support Vector Machine with a score of 66.01% and the one realized by Logistic Regression algorithm, which was 62.08%. This shows that ensemble learning approaches can be employed for the analysis of agricultural markets. Random Forest has been stated to be highly effective because of its learning method through the use of ensemble decision tree predictions, lowering variance and avoiding overfitting [1, 9, 11]. Unlike Logistic Regression, which assumes that there are linear relationships between variables, Random Forest effectively captures complicated relationships between temporal attributes such as year, month, and seasons, as well as geographic attributes such as the location of cities. The ability to capture complicated relationships is vital for Ag markets since the pricing of agricultural products depends on many variables, all related to one another.

The findings are consistent with the earlier research which found Random Forest to have been successful in its application in agriculture, particularly in crop prediction, estimating crop yield, and market intelligence [6, 7, 13]. As is evident from agricultural datasets, which usually contain diverse and large data, ensemble learning models outshine other conventional machine learning/statistical methods in accuracy. The results of the current study further substantiate these insights through a lengthy dataset consisting of over 411,000 records for different cities in Pakistan.

SVM (Support Vector Machine) was better than Logistic Regression, but not as good as Random Forest. The SVM algorithm might not have been able to capture the complete nonlinear pattern in the data due to the size and complexity of the data set.The size and complexity of the data set may have limited the ability of the SVM algorithm to capture the complete nonlinear pattern in the data. This is related to observations in comparative machine learning studies using large scale agri-data that often shows that Random Forest is more scalable and more robust to prediction [14; 16].

The variable Year has about 62.6% of the total prediction importance, followed by City with 22.9% of prediction importance. In summary, from the findings, it can be concluded that commodity prices in Pakistan are largely affected by long-term time trend effects as well as market conditions in their region. The results also imply that since the seasonal factors do not have much effect on price levels, then general economic/market conditions might play more of an important role in determining prices than seasonality.

From a pragmatic perspective, the suggested framework can offer a scalable approach to agricultural market intelligence and commodity price prediction. The processes of strategic planning, market regulations, storage control, and market-oriented decision making can be aided by appropriate classification of crop prices into low, medium, and high levels. These systems can help reduce uncertainties and improve economic outcomes for agricultural participants. Despite all the successes achieved by this study, there are still several drawbacks in its methodology. Most importantly, the independent variables used in this system are only related to time and geography and do not include other possible variables such as cost of fuel, currency rate, transportation costs, world market trend, or the level of inflation. These variables can prove helpful in predicting and analyzing behavior of commodity prices. It would also be helpful to continue researching the area with the use of machine learning and deep learning approaches and hybrid forecasting models.

CONCLUSION AND FUTURE WORK

In this paper, a machine learning-based framework is proposed for categorizing the prices of agricultural commodities based on a huge dataset of more than 411,000 records of the Mango and Apple markets in 138 cities in Pakistan between 2007 and 2022. The approach used to divide the continuously valued variable into distinct classes and including time and space information helped in successfully capturing the complicated nature of prices in different regions. Among the three classifiers used in this analysis which are Random Forest, Logistic Regression, and Support Vector Machine, the performance of ensemble models proved to be much better when modeling the non-linearity present in agricultural data, and Random Forest performed best with an accuracy of 81.78% and an F1-score of 0.8189.

In addition to the predictive ability of the models, the implications for data analytics and its ability to increase market transparency and effective planning and risk management practices are highlighted through this study. The other important implication regarding time and space being crucial factors affecting prices is that there is much to be learned about the structure of Pakistan's agricultural market system. There are many ways through which this model could be further enhanced, including incorporating more variables such as climatic information (rainfall, temperature, and humidity), macroeconomic information (inflation rates, exchange rates, and energy costs), logistics information, and commodity prices on the international scene. Further research is also required in the use of deep

learning algorithms, ensemble hybrid models, and time series analysis techniques.

REFERENCES

[1] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[2] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] A. Kamilaris and F. X. Prenafeta-Boldú, "Deep Learning in Agriculture: A Survey," Computers and Electronics in Agriculture, vol. 147, pp. 70–90, 2018.

[5] M. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine Learning in Agriculture: A Review," Sensors, vol. 18, no. 8, Article 2674, 2018.

[6] S. Benos, D. Tagarakis, G. Dolias, N. Berruto, D. Kateris, and D. Bochtis, "Machine Learning in Agriculture: A Comprehensive Updated Review," Sensors, vol. 21, no. 11, Article 3758, 2021.

[7] M. Shahhosseini, G. Hu, and S. V. Archontoulis, "Forecasting Corn Yield with Machine Learning Ensembles," Scientific Reports, vol. 10, Article 9958, 2020.

[8] A. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," Frontiers in Plant Science, vol. 10, Article 621, 2019.

[9] S. Wager, J. Athey, and J. Tibshirani, "Generalized Random Forests," Annals of Statistics, vol. 47, no. 2, pp. 1148–1178, 2019.

[10] J. M. Klusowski, "Sharp Analysis of a Simple Model for Random Forests," arXiv preprint arXiv:1805.02587, 2018.

[11] G. Biau, E. Scornet, and J. Welbl, "Neural Random Forests," Journal of Machine Learning Research, vol. 20, no. 1, pp. 1–25, 2019.

[12] M. Rahman, S. Islam, and T. Hossain, "Data-Driven Agricultural Price Classification Using Machine Learning Models," Computers and Electronics in Agriculture, vol. 210, Article 107924, 2024.

[13] Y. Zhang, H. Wang, and X. Chen, "Ensemble Learning Approaches for Agricultural Market Intelligence Systems," Expert Systems with Applications, vol. 216, Article 119135, 2023.

[14] A. Kumar, P. Singh, and R. Sharma, "Comparative Analysis of Machine Learning Algorithms for Agricultural Commodity Price Prediction," Procedia Computer Science, vol. 218, pp. 1245–1254, 2023.

[15] M. A. Awan, M. Irfan, and S. Ahmed, "Agricultural Price Forecasting Using Machine Learning Techniques: Evidence from Developing Economies," International Journal of Agricultural Management, vol. 11, no. 2, pp. 45–58, 2022.

[16] S. Jeong, J. Kim, and H. Lee, "Machine Learning Approaches for Agricultural Commodity Price Prediction: A Review," Computers and Electronics in Agriculture, vol. 187, Article 106294, 2021.

[17] S. Madeh Piryonesi and T. E. El-Diraby, "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling," Journal of Infrastructure Systems, vol. 27, no. 2, 2021.

[18] H. H. Huang, T. Xu, and J. Yang, "Comparing Logistic Regression, Support Vector Machines and Classification Methods in Predictive Analytics," BMC Proceedings, vol. 8, Suppl. 1, 2014.