Analyzing Temperature and Rainfall effects on Crop Yield using Data Mining

Mahlaka Abbas Soomro
Department of Software Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan
19mese07@students.muet.edu.pk

Sania Bhatti
Department of Software Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan
sania.bhatti@faculty.muet.edu.pk

Areej Fatemah Meghji
Department of Software Engineering
Mehran University of Engineering and Technology
Jamshoro, Pakistan
areej.fatemah@faculty.muet.edu.pk

***Abstract*—Pakistan's economy is largely influenced by agriculture. The ability to produce a high yield of crops is important. Factors such as temperature, soil type, rainfall, seed quality, precipitation, and a lack of technical facilities directly influence crop yield. As a result, new technologies are required to meet the developing need, and farmers need to work smartly through adopting new technology rather than relying on ineffective ways. This research focuses on using the Data Mining technique to create a crop yield prediction system by analyzing agricultural and climatic datasets.**

***Index Term*—**data mining, crop yield prediction, temperature, rainfall, random forest

I. INTRODUCTION

The least developed countries (LDCs), such as Pakistan, depend on agriculture for their economies. Agriculture plays a significant role in Pakistan's economy. Most farmers in Pakistan are not achieving the projected crop yield due to a variety of factors [1]. Climatic conditions have a major impact on agricultural productivity [1]. Agricultural land is typically crop production that depends on the climate and economic factors [2]. Factors that depend on agriculture include temperature, rainfall, climate, soil, fertilizers, pesticides, cultivation, irrigation, and other factors. Yield prediction plays a vital role in agricultural problems [2].

A process of exploration that searches for hidden patterns in a collection of data is called Data Mining (DM). In agriculture, data mining is an important object of research [3]. Many parameters are investigated, and crop yield predictions are made using Data Mining and Machine Learning techniques [3].

Data collection is the first stage of DM in which the required data is collected from various resources and arranged in a systematic order. Later, the collected dataset is pre-processed to get the required information [4]. The most efficient yield output will be forecasted by DM algorithms [5]. Previously, the yield was predicted based on the farmer's previous experience. However, the weather may now drastically change, and they can no longer guess the yield [6]. As a result, technology can help them in predicting agricultural yields and deciding whether to continue the crop or not. The crop and yield pattern will be understood by a DM model, which will predict the yield of the area in which he would crop based on several conditions [7].

Using the DM technique, we will investigate how temperature and rainfall effects crop yield. In our work, WEKA Tool is used for obtaining the results of Crop Yield Prediction using Random Forest (RF) Algorithm [8]. RF is the most popular algorithm for this type of problem; thus, it is selected from among the numerous available algorithms [8].

The rest of the paper is organized as follows. In Section II, Literature Review is described. Methodology is described in Section III respectively. The Experimental Results and Discussion is discussed in Section IV. Section V concludes the paper.

II. LITERATURE REVIEW

The author presents a brief study of crop yield prediction using Density-based (DB) clustering, Multiple Linear Regression (MLR) techniques for the region of India in the study [8]. This study tries to identify the precise crop yield analysis for a certain region, which is then processed using DB clustering algorithms. In [9] by combining different classifiers, the researchers examined DM techniques used to predict rice crop yield for the Kharif season in India's tropical wet and dry climatic zones. 2 DM techniques are used to complete the tasks i.e., K Nearest Neighbors (KNN), and Decision Tree (DT).

The author collected the dataset of weather over 2 years and used a DT and Naïve Bayes (NB) algorithm in the study [10].

ISBN # 978-969-23372-2-9

As a result, DT performed better than NB. In [11] the author uses Clustering algorithms i.e., Expectation-Maximization (EM), K-means, and Density-Based Spatial Clustering of Applications with Noise. (DBSCAN). K-means and DBSCAN produce similar results, while EM produces more specific values for rainfall and temperature.

The author investigated the effect of temperature and rainfall on paddy yield after pre-processing the data using the Predictive Apriori Algorithm with the DM tool (WEKA) in the study [12]. In [13] the author uses many DM techniques to predict grass grub damage.

The use of different DM techniques for rainfall prediction in Lahore is discussed in [14]. They took 2 classes one is Rain other is No -Rain. For No-Rain class techniques perform well but not for Rain class. In [15] a user-friendly interface is created by the author for the farmers that predicts production using DM techniques like Regression and Clustering based on available data in all districts of Kerala.

The crop prediction model was built using KNN and Linear Regression (LR) approaches by researchers in the study [16]. Turmeric, cotton, and sugar cane are all considered when predicting yields. They used climatic and crop data to forecast crop growth. The output is then subjected to prediction rules, which are used to calculate the crop's overall yield.

In [17] the crop yield was estimated using the XG Boost algorithm, which was developed by the researcher. In this study, the dataset used was the cultivated location, rainfall, production, and maximum and minimum temperatures. It is focused on rice crops. The proposed work compares the accuracy of LR, Support Vector Regression (SVR), DT classifier, and RF with the XG Boost method. Experiments are being carried out in Andhra Pradesh, Tamil Nadu, Karnataka, and Kerala.

In [18] an android-based yield prediction tool is presented by the author. This proposed technique forecasts the most profitable crops based on existing soil and climate conditions. The input variables are temperature, rainfall, and soil parameters. A model was created using MLR.

Different authors have used different DM techniques. Some authors have used crop yield datasets, while others have used rainfall and temperature datasets. None of the authors have worked on crop yield prediction using data mining techniques in Pakistan using WEKA. The authors of the study [14] covered Lahore, Pakistan for rainfall prediction only.

III. METHODOLOGY

This section contains details of data collection, data preprocessing, tool, algorithm, and performance evaluation. Fig.1 shows the block diagram of the Crop Yield Prediction model.

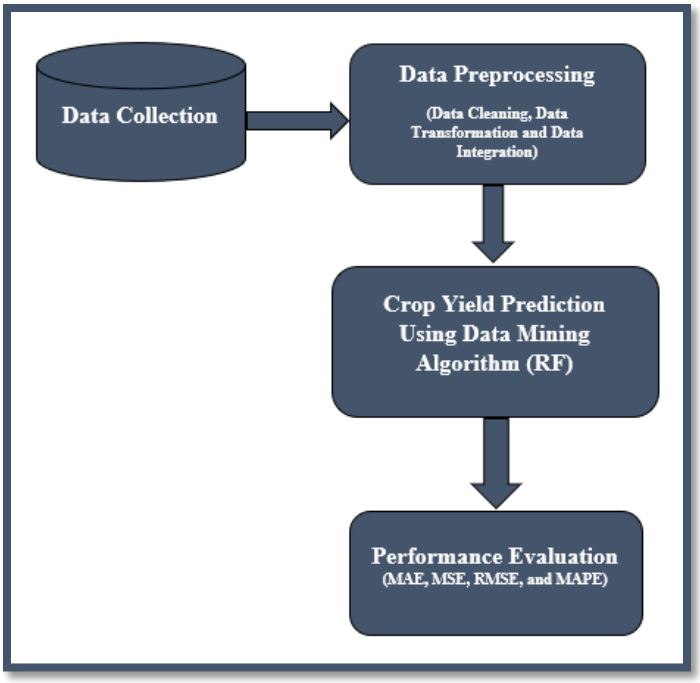

Fig. 1. Block diagram of Crop Yield Prediction model

A. Data Collection

Crop prediction models have several goals, one of which is to estimate agricultural production (i.e., Mango and Tomato yield) as a function of climatic conditions i.e., Minimum Temperature, Maximum Temperature, and Rainfall. This study is carried out in four cities (Bahawalpur, Faisalabad, Lahore, and Multan) of Punjab, Pakistan. Mango and Tomato yield is obtained from the Crop Reporting Service (Government of Punjab) [19], while climate-related data i.e., Minimum Temperature, Maximum Temperature, and Rainfall are obtained from the Meteorological Department of Pakistan [20]. The arithmetic averages of daily maximum and minimum temperatures recorded over the month are the mean maximum and mean minimum temperatures. We have taken the dataset of 10 years from 2011 to 2020.

B. Data Preprocessing

The information gathered from many sources is frequently in its raw state. It may include insufficient or inconsistent information. As a result, such redundant data should be filtered throughout this procedure. As a result, we began by preprocessing the dataset.

Data cleansing is the first and most important stage in data pre-processing. Data Transformation is the next step in which data collected was transformed into an understandable and appropriate format. In last, we integrated the dataset which is the process of merging data from several sources into a single dataset.

ISBN # 978-969-23372-2-9

The datasets were collected in Microsoft Office Excel. Each crop has the following columns: Year, Yield, Average Minimum Temperature, Average Maximum Temperature, and Average Rainfall

C. Tool

"WEKA (Waikato Environment for Knowledge Analysis)" [21] was used to analyze the entire data set. It is an open-source software written in the JAVA programming language by the University of Waikato in New Zealand. It is used to solve difficulties involving machine learning and data mining.
The data must be entered into the software, and the appropriate method must be chosen. It includes many classifiers that may be used to create models and solve issues. It has an interactive Graphical User Interface (GUI) that incorporates all the data analysis options. The dataset is saved in the arff format (Attribute Relation file format) for WEKA processing [21].

D. Algorithm

Random Forest

A RF is a classification approach that combines several decision trees with an ensemble classification model. To improve predictions, the RF model collects trained data from all tree nodes and separates the weaker node training data. The RF model is used to solve difficulties in classification and regression [20].

E. Performance Evaluation

The following are some of the factors that are used in this experiment to evaluate performance:

- *Mean Absolute Error (MAE):* The average absolute difference between the classifier's predicted and actual output.
- *Mean Squared Error (MSE):* The average of the sum of squared differences between the classifier`s predicted and actual output.
- *Root Mean Square Error (RMSE):* It is the difference between the predicted and actual values obtained by the model.
- *Mean Absolute Percentage Error (MAPE):* It's a statistic for evaluating a forecasting system's accuracy. This accuracy is calculated as a percentage using the average absolute percent error for each period minus actual values divided by actual values.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

This section shows the outcomes using the RF algorithm on a dataset of Mango and Tomato crops of various cities. We have computed 3 years' prediction of 4 cities of Punjab, Pakistan. The results of the RF algorithm for each crop yield prediction with Average Minimum. Average Maximum Temperature and Average Rainfall are presented in this section.

A. Mangoes

Fig. 2 - Fig.5 illustrate the yield of Mangoes from various cities of Punjab, Pakistan. The x-axis shows the time, and the y-axis shows yield, temperature, and rainfall. The unit of yield is Acre, the temperature unit is Celsius (°C), and the rainfall unit is Millimeter (mm).

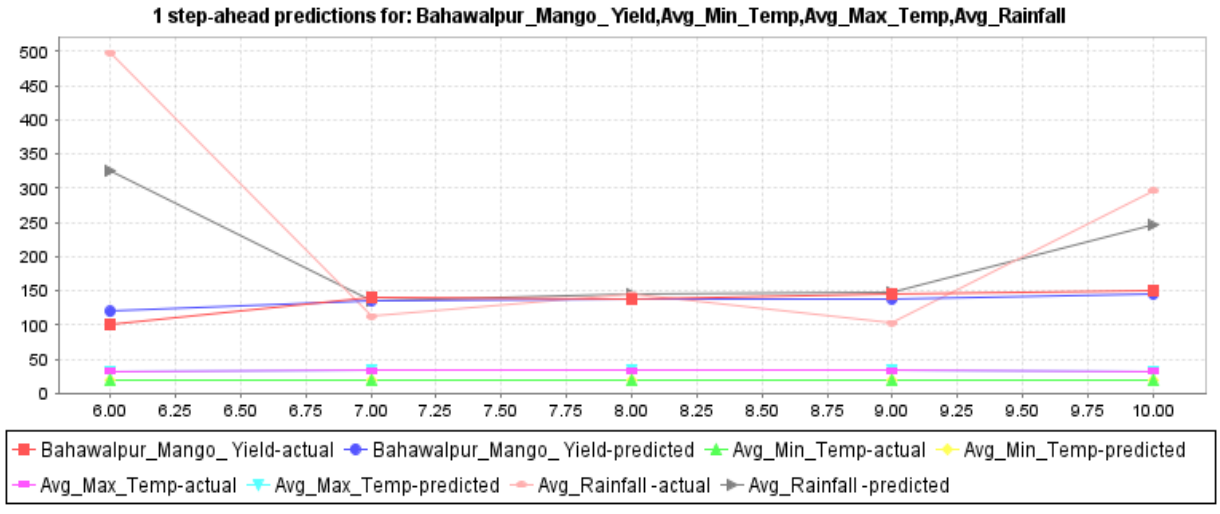

Fig. 2. Bahawalpur Mango Yield

ISBN # 978-969-23372-2-9

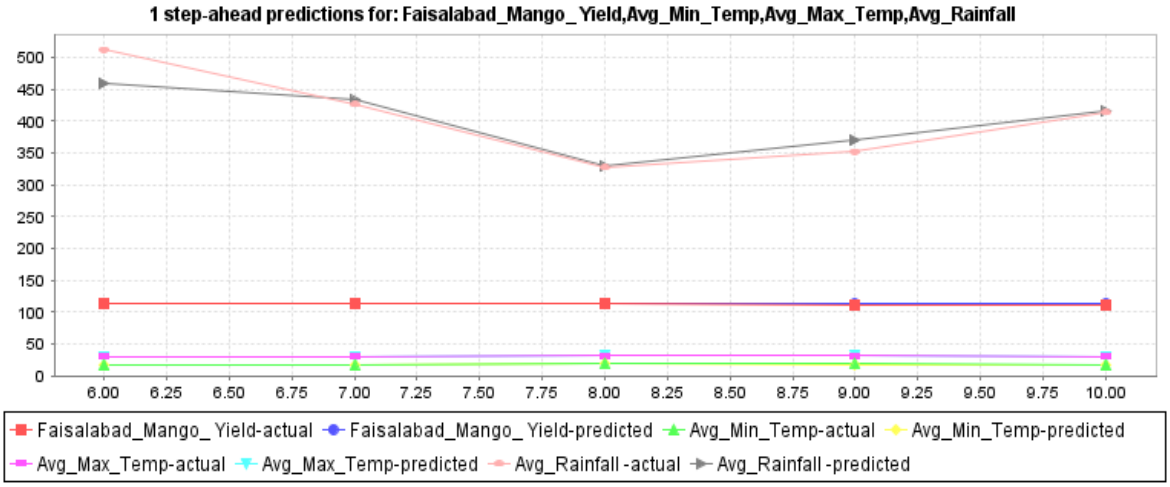

Fig. 3. Faisalabad Mango Yield

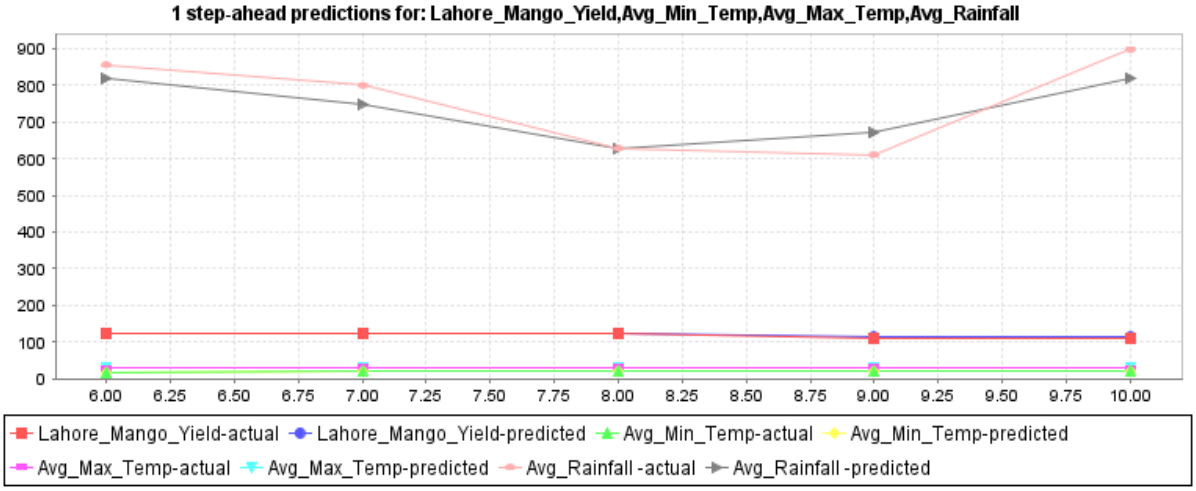

Fig. 4. Lahore Mango Yield

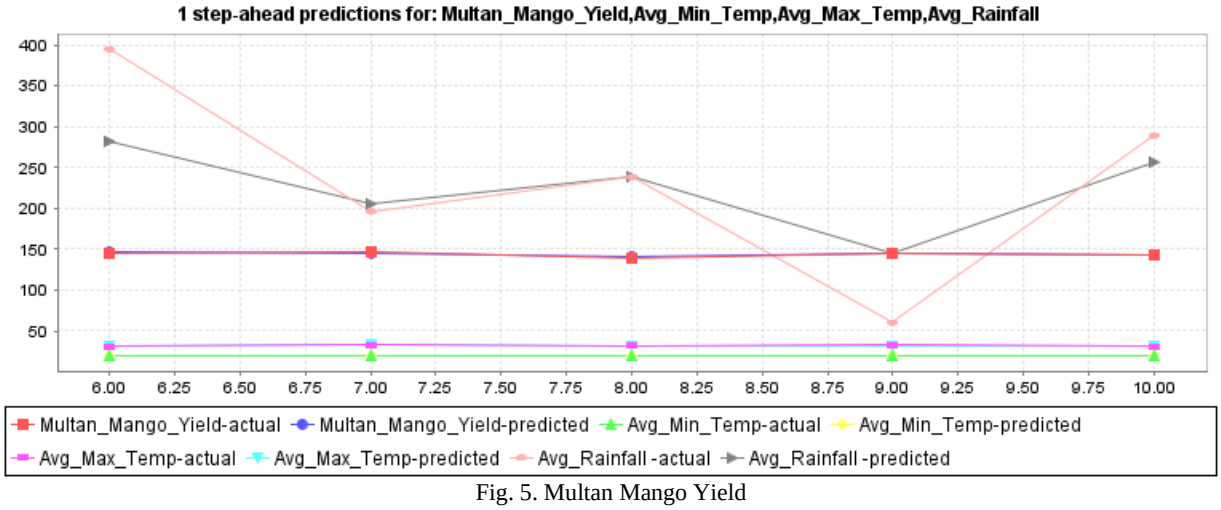

Fig. 5. Multan Mango Yield

ISBN # 978-969-23372-2-9

In Bahawalpur, mangoes will yield 131.0655 in 2021, 137.4314 in 2022, and 130.7692 in 2023, as shown in Fig. 2. As a result, the projection falls in 2021, rises in 2022, and then falls in 2023. In Faisalabad, mangoes will yield 112.7565 in 2021, 112.6228 in 2022, and 112.6853 in 2023, as shown in Fig. 3. So, based on the findings, the forecast for 2021-2023 is almost the same. In Lahore, mangoes will yield 116.6696 in 2021, 120.0213 in 2022, and 122.0016 in 2023, according to Fig. 4. As a result, the annual output of mangoes in Lahore is increasing. In Multan, mangoes will yield 143.4861 in 2021, 142.4903 in 2022, and 142.4932 in 2023, as shown in Fig. 5. As a result of the analysis, the yield of mangoes in Multan is expected to increase in 2021 but decrease in 2022 and 2023.

We have chosen four parameters i.e., MAE, MSE, RMSE, and MAPE. Performance evaluation of these prediction models has been provided below in Table-I

Table-I: Comparison of Performance Evaluation of Mangoes

Cities and Mango Crop	**Steps**	**MAE**	**MSE**	**RMSE**	**MAPE**
Bahawalpur Mango	1-step ahead	7.294	95.926	9.794	6.238
	2-steps ahead	5.223	35.618	5.968	3.588
	3-steps ahead	5.259	38.575	6.210	3.572
Faisalabad Mango	1-step ahead	0.108	0.015	0.123	0.096
	2-steps ahead	0.122	0.0191	0.138	0.108
	3-steps ahead	0.136	0.021	1.146	0.121
Lahore Mango	1-step ahead	1.826	5.011	2.238	1.599
	2-steps ahead	2.180	6.678	2.584	1.934
	3-steps ahead	2.850	10.237	3.199	2.554
Multan Mango	1-step ahead	0.674	0.554	0.744	0.467
	2-steps ahead	0.73	0.553	0.744	0.509
	3-steps ahead	1.534	2.743	1.656	1.082

According to the data obtained using the WEKA tool, as we can see from Table-I, Bahawalpur Mango MAE in the first step is 7.2944, however in the second and third steps, the error is smaller. The MSE in the first step is 95.9261, but the error in the second step is lower, and the error in the third step is somewhat higher. The RMSE in the first step is 9.7942, however, it is reduced in the second step and slightly increases the error in the third step. The MAPE in the first phase is 6.2382, but it is lower in the second and third steps. Faisalabad, Lahore, and Multan Mangoes have lower MAE, MSE, RMSE, and MAPE in the first step, but increase in the second and third steps.

B. Tomatoes

Fig. 6 - Fig. 9 shown below illustrate the yield of Tomatoes from various cities of Punjab, Pakistan. The x-axis shows the time, and the y-axis shows yield, temperature, and rainfall.

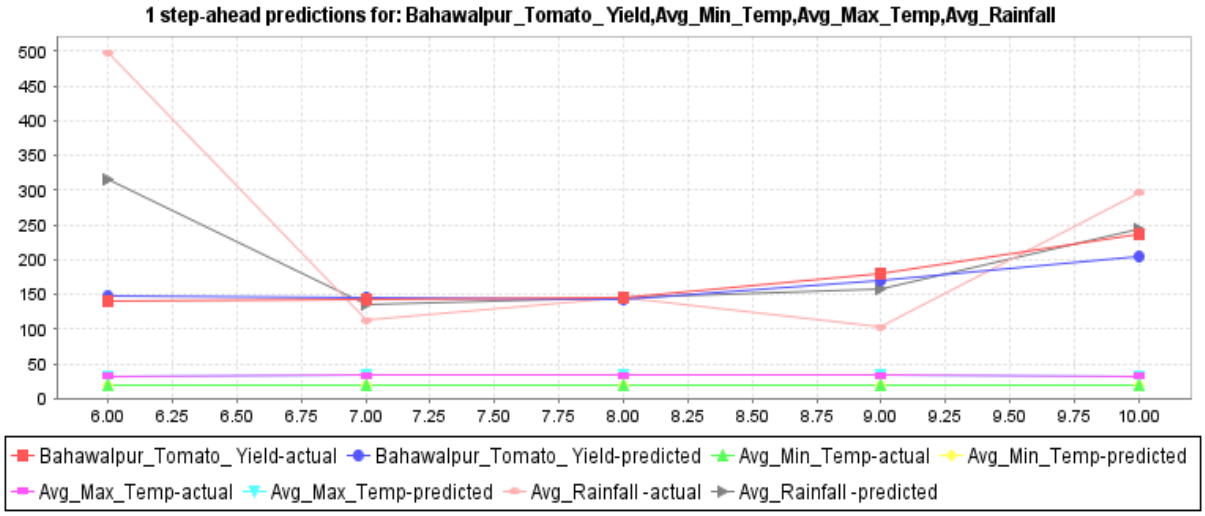

Fig. 6. Bahawalpur Tomato Yield

ISBN # 978-969-23372-2-9

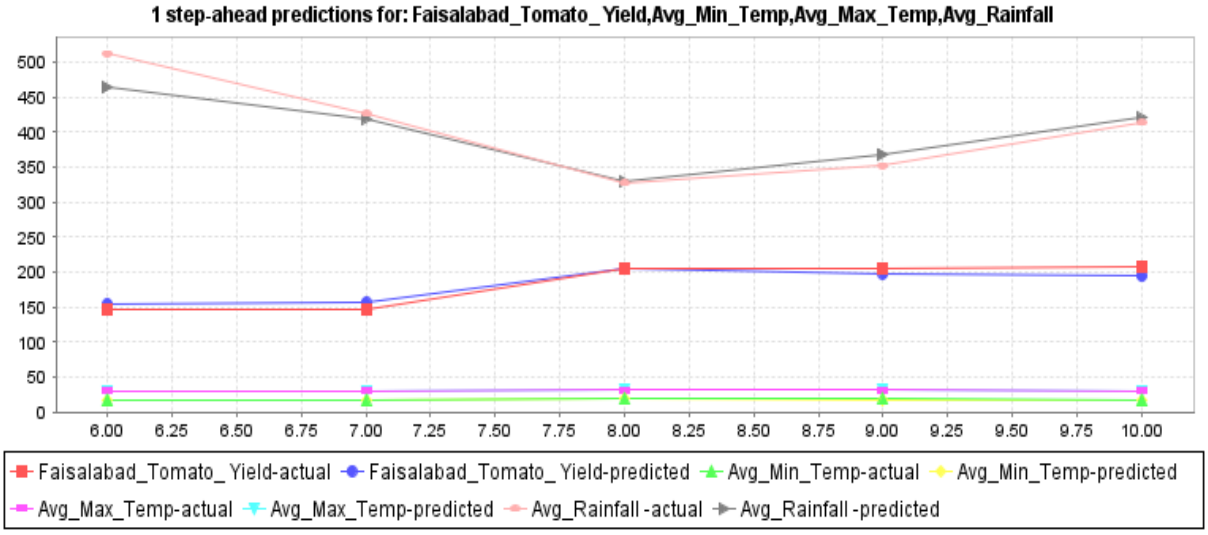

Fig. 7. Faisalabad Tomato Yield

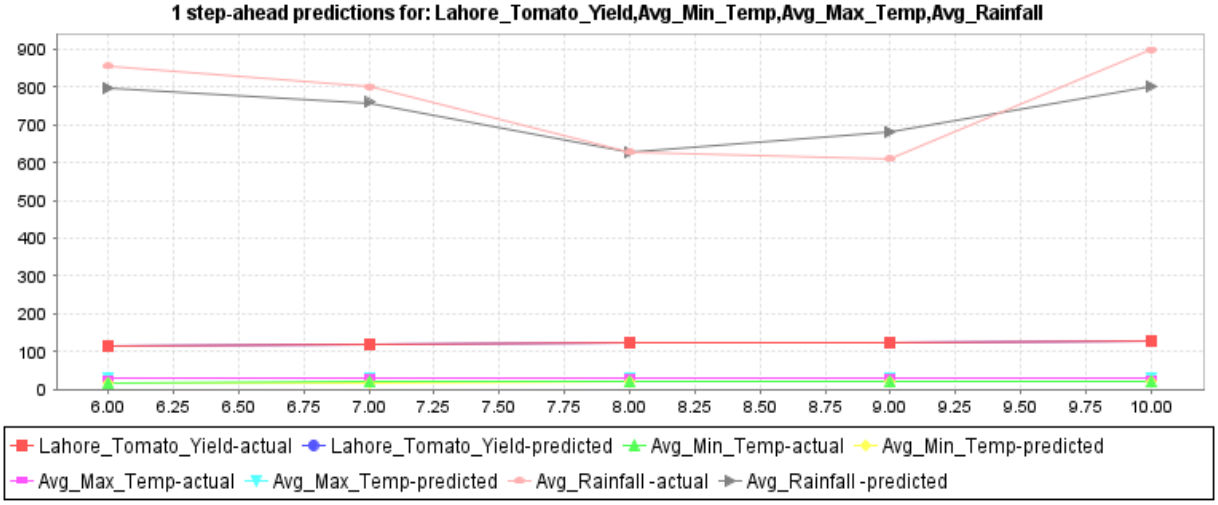

Fig. 8. Lahore Tomato Yield

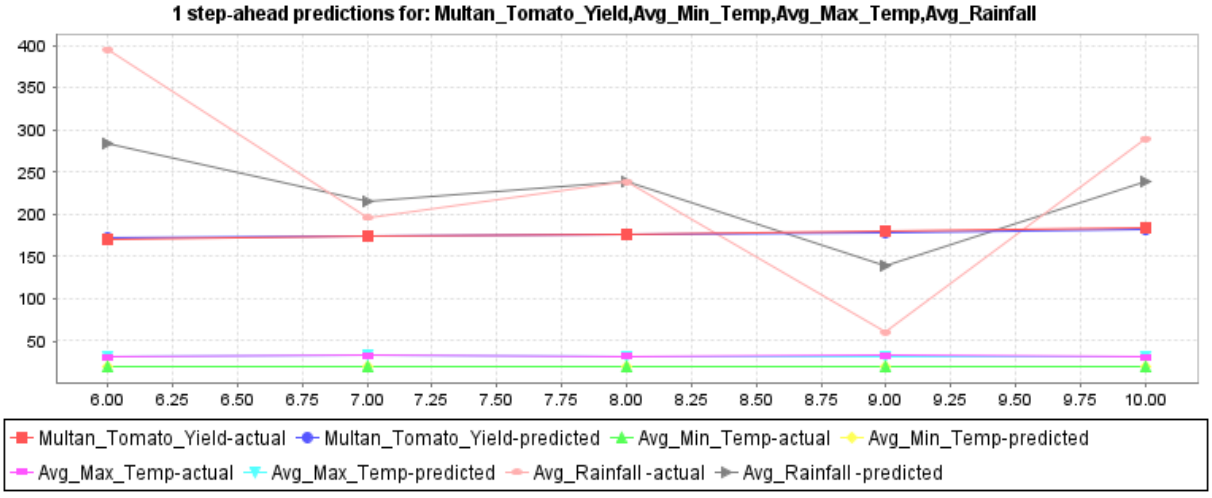

Fig. 9. Multan Tomato Yield

ISBN # 978-969-23372-2-9

In Bahawalpur, tomatoes will yield 188.2634 in 2021, 186.7706 in 2022, and 185.0823 in 2023, according to Fig. 6. As a result, the yield is expected to decline between 2021 and 2023. In Faisalabad, tomatoes will yield 184.2836 in 2021, 198.2719 in 2022, and 194.8847 in 2023, according to Fig. 7. As a result of the findings, the yield will decrease in 2021, increase in 2022, and then decrease in 2023. In Lahore, tomatoes will yield 124.7694 in 2021, 125.8506 in 2022, and 125.6036 in 2023, according to Fig. 8. As a result of the findings, yield is expected to decrease in 2021 but increase in 2022 and 2023. In Multan, tomatoes will yield 179.0197 in 2021, 179.312 in 2022, and 179.7333 in 2023, according to Fig. 9. So, based on the results, the yield will be approximately the same from 2021 to 2023.

Performance evaluation of the prediction model has been provided in table-II

Table-II: Comparison of Performance Evaluation of Tomatoes

Cities and Tomato Crop	**Steps**	**MAE**	**MSE**	**RMSE**	**MAPE**
Bahawalpur Tomato	1-step ahead	10.495	233.982	15.296	5.320
	2-steps ahead	13.684	369.598	19.224	6.690
	3-steps ahead	18.678	508.357	22.546	9.220
Faisalabad Tomato	1-step ahead	7.546	72.335	8.505	4.321
	2-steps ahead	10.765	125.347	11.195	5.908
	3-steps ahead	10.568	116.133	10.776	5.124
Lahore Tomato	1-step ahead	0.934	1.989	1.410	0.736
	2-steps ahead	1.190	3.410	1.846	0.926
	3-steps ahead	1.655	4.648	2.156	1.287
Multan Tomato	1-step ahead	1.311	2.878	1.696	0.727
	2-steps ahead	1.514	4.841	2.200	0.828
	3-steps ahead	1.807	6.337	2.517	0.983

According to the obtained results, as we can see from Table-II the MAE, MSE, RMSE, and MAPE of Bahawalpur, Faisalabad, Lahore, and Multan Tomatoes are lower in the first step, but it increases in the second and third steps.

V. CONCLUSION

Data mining is a technique for identifying patterns and relationships in data that may be utilized to address business challenges. Enterprises can predict future trends using DM techniques and tools. As a result, it will be valuable in predicting agricultural yields. Many studies have been carried out to predict various crop yields. This paper presents a brief analysis of Crop Yield Prediction using a Data Mining Technique-based Random Forest algorithm for Punjab province in Pakistan. The experimental result shows that the proposed work efficiently predicts Mango and Tomato crop yield using the WEKA tool.

In the future, this research could be extended to predict various crop yields based on agricultural factors such as soil, fertilizers, pesticides, cultivation, irrigation, and other factors.

REFERENCES

[1] Mistry, Ami, et al. “Brief Survey of Data Mining Techniques Applied to Applications of Agriculture.” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 5, no. 2, Feb. 2016, pp. 301–304., doi:10.17148/IJARCCE.2016.5263.

[2] Majumdar, Jharna, et al. “Analysis of Agriculture Data Using Data Mining Techniques: Application of Big Data.” *Journal of Big Data*, vol. 4, no. 1, 2017, doi:10.1186/s40537-017-0077-4.

[3] Jambekar, Suvidha, et al. “Prediction of Crop Production in India Using Data Mining Techniques.” *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, 2018, doi:10.1109/iccubea.2018.8697446.

[4] T, Jasper Varun, et al. “Analysis and Prediction of Crop Yield Using Data Mining and Machine Learning Algorithms.” *International Journal of Modern Agriculture*, vol. 10, no. 2, 2021, pp. 3550–3559.

[5] Ms. Fathima, et al. “ANALYSIS OF CROP YIELD PREDICTION USING DATA MINING TECHNIQUE.” *International Research Journal of Engineering and Technology*, vol. 07, no. 05, May 2020, pp. 7708–7713.

[6] C, Pooja M, et al. “Implementation of Crop Yield Forecasting Using Data Mining.” *International Research Journal of Engineering and Technology*, vol. 05, no. 04, Apr. 2018, pp. 2057–2059.

[7] Jeong, Jig Han, et al. “Random Forests for Global and Regional Crop Yield Predictions.” *PLOS ONE*, vol. 11, no. 6, 2016, doi:10.1371/journal.pone.0156571.

[8] Ramesh, D, and B Vishnu Vardhan. “Analysis of Crop Yield Prediction Using Data Mining Techniques.” *International Journal of Research in Engineering and Technology*, vol. 04, no. 01, 25 Jan. 2015, pp. 470–473., doi:10.15623/ijret.2015.0401071.

[9] Gandhi, Niketa, and Leisa J. Armstrong. “Rice Crop Yield Forecasting of Tropical Wet and Dry Climatic Zone of India Using Data Mining Techniques.” *2016 IEEE*

ISBN # 978-969-23372-2-9

International Conference on Advances in Computer Applications (ICACA), 2016, doi:10.1109/icaca.2016.7887981.

[10] Sheikh, Fahad, et al. "Analysis of Data Mining Techniques for Weather Prediction." *Indian Journal of Science and Technology*, vol. 9, no. 38, 2016, doi:10.17485/ijst/2016/v9i38/101962.

[11] Bharadi, Vinayak A., et al. "ANALYSIS AND PREDICTION IN AGRICULTURAL DATA USING DATA MINING TECHNIQUES." *International Journal of Research In Science & Engineering*, no. 7, Mar. 2017, pp. 386–393.

[12] Kaur, Kuljit, and Kanwalpreet Singh Attwal. "Effect of Temperature and Rainfall on Paddy Yield Using Data Mining." *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence*, 2017, doi:10.1109/confluence.2017.7943204.

[13] Ayub, Umair, and Syed Atif Moqurrab. "Predicting Crop Diseases Using Data Mining Approaches: Classification." *2018 1st International Conference on Power, Energy and Smart Grid (ICPESG)*, 2018, doi:10.1109/icpesg.2018.8384523.

[14] Aftab, Shabib, et al. "Rainfall Prediction in Lahore City Using Data Mining Techniques." *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 4, 2018, pp. 254–260., doi:10.14569/ijacsa.2018.090439.

[15] Kabeer, Nishiba, et al. "Prediction of Crop Yield Using Data Mining." *International Journal of Computer Science and Network*, vol. 8, no. 3, June 2019, pp. 300–304.

[16] Surya, P., and Dr. I. Laurence Aroquiaraj. "CROP YIELD PREDICTION IN AGRICULTURE USING DATA MINING PREDICTIVE ANALYTIC TECHNIQUES." *International Journal of Research and Analytical Reviews*, vol. 5, no. 4, Dec. 2018, pp. 783–787.

[17] Lata, Kusum, and Bhushan Chaudhari. "CROP YIELD PREDICTION USING DATA MINING TECHNIQUES AND MACHINE LEARNING MODELS FOR DECISION SUPPORT SYSTEM." *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 4, Apr. 2019, pp. 391–396.

[18] Palanivel, Kodimalar, and Chellammal Surianarayanan. "An Approach for Prediction of Crop Yield Using Machine Learning and Big Data Techniques." *International Journal of Computer Engineering and Technology*, vol. 10, no. 3, 2019, pp. 110–118., doi:10.34218/ijcet.10.3.2019.013.

[19] Crop Reporting Service Government of Punjab https://crs-agripunjab.punjab.gov.pk/node/165#overlay-context=reports

[20] Pakistan Meteorological Department https://www.pmd.gov.pk/en/

[21] Mishra, Shruti, et al. "Use of Data Mining in Crop Yield Prediction." *2018 2nd International Conference on Inventive Systems and Control (ICISC)*, 2018, pp. 796–802., doi:10.1109/icisc.2018.8398908.

[22] Sangeeta, and Shruthi G. "Design And Implementation Of Crop Yield Prediction Model In Agriculture." International Journal of Scientific & Technology, vol. 8, no. 01, Jan. 2020, pp. 544–549.

ISBN # 978-969-23372-2-9