Weather based Localized Crop Prediction using Machine Learning

Sada Hussain[1], Farida Memon[2,*], Arbab Nighat[2], Fayaz Ahmed Memon[3], Majid Hussain[4],

[1]Postgraduate Student MUET, Jamshoro, Pakistan
[2]Department of Electronic Engineering, MUET, Jamshoro, Pakistan
[3]Department of Software Engineering, QUEST, Nawabshah, Pakistan
[4]Department of Electronic Engineering, QUEST, Nawabshah, Pakistan
[*]Corresponding author: farida.memon@faculty.muet.edu.pk

Abstract

Crop prediction in agriculture is critical and essentially depends upon soil and environmental conditions which include rainfall, humidity, and temperature. Accurate crop prediction results in increased crop production. Recently, machine leering techniques have been successfully employed in the agriculture field for classification and detection tasks. The fundamental goal of this research is to employ several machine learning methods to forecast the accurate crop for a land-based on soil and weather parameters. The classification algorithms employed in this study involve Logistic Regression, Naive Bayes, Random Forest, Support Vector Machines (SVM), XGBoost, and AdaBoost; with XGBoost offering the highest level of prediction accuracy and reliability. This work can greatly assist farmers and other stakeholders in making appropriate storage and business decisions to locate the crop before sowing. Moreover, a web-based application using Flask platform is developed to assist farmers in choosing which crop to cultivate to elicit the greatest return.

Keywords—Machine Learning, Crop Prediction, Logistic Regression, Support Vector Machines, Random Forest, Confusion matrix

1 Introduction

From decays, agriculture has been one of the primary tasks as well as a backbone of the economy and plays a criti-cal role in the development of every country[1]. Not only it is necessary for economic growth as well it is essential for our survival also. Crop production is the primary source of human life. Increasing crop production is regarded as an important aspect of agriculture. Accurate prediction of crops is a difficult and challenging task because it in-volves numerous factors such as soil type, temperature, humidity, and so on[2]. If it is possible to locate the crop before sowing, it will greatly assist farmers and other stakeholders in making appropriate storage and business decisions[3]. Generally, the farmers continue to grow the same crops and do not make experiments by planting new crops and they use fertilizers in random quantities without knowing the side effects of using deficient quantity and quality of fertilizers. Therefore, it has a direct impact on crop production and it also initiates soil acidification and damage to the top layer. Prior crop predictions were made based on the farmer's experience at a specific loca-tion having a lack of information about the soil nutrients including potassium, phosphorous, and nitrogen in the soil.

Machine learning has become increasingly important in recent years in all fields, including agriculture[4]. Ma-chine learning is a fast-growing technology that facilitates decision-making across all industries to provide the most useful of its capabilities. Most modern tools examine machine learning models before implementation. The main purpose is to use machine learning algorithms to enhance the throughput of the agricultural sector.

In this paper, different machine learning algorithms including Logistic Regression[5], Naive Bayes[6], Random Forest[7], Support Vector Machines (SVM)[8], XGBoost[9], and AdaBoost[10] classifiers are used for prediction and classification of crops and their performance is analyzed and compared. We use these algorithms to design a sys-tem for conducting a reliable prediction that reflects irregular patterns in

ISSN: 2523-0379 (Online), ISSN: 1605-8607 (Print)
DOI: https://doi.org/10.52584/QRJ.2002.13.

weather and soil characteristics. Applying the said machine learning algorithms allowed us to reach the fact that the XGBoost algorithm offers the highest level of accuracy. The system forecasts crops based on the collection of historical data. The information is provided using historical data on the weather, temperature, and a variety of other variables. Our application runs an algo-rithm and displays a crop that matches the inputted data. The research work in the paper may help the farmers by guiding sowing the crops by using the latest machine learning methods to reduce the financial losses that are faced by farmers due to planting of wrong crops.

The remainder of this paper is structured as follows: in section 2, the literature review is presented and in section 3, the methodology and dataset are described. Finally, in section 4, the experimental results are discussed followed by a conclu-sion and future work in section 5.

2 Literature Review

The use of machine learning in the field of agriculture is new that has attracted several researchers to develop systems that are capable of learning themselves without any need for programming. This section discusses the most recent existing machine learning algorithms in the field of agriculture for crop prediction. Leo Brieman[11] is an expert in the random forest algorithm. This algorithm firstly forms decision trees using several data samples and then predicts the data from each subset. After predicting the data, the best solution for the system is determined through voting. The data is trained in Random Forest using the bagging approach. The ran-domization must reduce the correlation p while preserving strength to increase accuracy.

Mishra et al.,[12] have presented a critical review of several machine-learning approaches in the field of agricul-ture for crop prediction. This study is proposed to evaluate the performance of these up-to-date existing machine learning techniques for crop yield prediction and their application based on various parameters in the dataset.

P.Priya et al.,[13] proposed a crop yield prediction system using a random forest classifier. To anticipate the ag-ricultural output, many factors like rainfall, temperature, and season were considered. On the datasets, no further machine-learning methods were used. Because alternative algorithms were lacking, comparison and quantifica-tion could not be done, making it impossible to give the best method. Pavan Patil et al.,[14] used decision tree and Naïve Bayes models for crop prediction, Thomas et al., [3] designed a system for crop prediction using Support Vector Machine/ Naïve Bayes machine learning models that achieve higher predicting accuracy. N et al., 2020 [4] developed a crop prediction system using supervised machine learn-ing algorithms. The developed system suggests the best suitable crop for land in terms of content and weather pa-rameters. Kalimuthu et al., [15] developed a crop prediction system using the Naive Bayes model. In this work, the most suitable parameters such as temperature, humidity, and moisture content for crop prediction are used which helps the farmers for achieving successful growth. Gowda and Reddy [16] developed a machine learning-based crop yield prediction system. The purpose of the developed system is to predict the best yield crop for a specific area. In this work, the authors use Random forest, Polynomial Regression, and Decision Tree machine learning algorithms.

Dr. G.Suresh [1] developed a machine learning-based crop yield recommendation system for digital farming. In this work, the authors designed a system utilizing a supervised machine-learning technique that suggests the right yields with higher precision and productivity. They also proposed SVM to locate the yield list. Later Venugopal et al., [2] proposed a crop yield prediction using Random Forest and Naive Bays Machine learning Algorithms. These algorithms achieved significant accuracy. According to Dr. Y. Jeevan Nagendra Kumar[17], supervised learning allows machine learning algorithms to forecast an objective or result. This study focuses on supervised learning methods for predicting agricultural yields. It must create an acceptable function using a collection of variables that may map the input variable to the desired output to obtain the outputs that are required. According to the article, crop forecasts may be made using the Random Forest ML method, which achieves the greatest accuracy value while taking into account the fewest number of models.

Abirami et al.,[18] carried out experiments on the Indian government dataset and found that the Random Forest machine learning method provides the best yield forecast accuracy. The authors observed that a simple sequential Recurrent Neural Network (RNN) model is more effective for predicting rainfall as compared LSTM model for pre-dicting temperature. For yield forecasts, they combined variables such as rainfall, temperature, season, area, etc. When all factors are considered, the results showed that Random Forest is the best classifier.

In [19], the authors employed several machine learning algorithms such as SVM, KNN, and GB trees for the pre-diction of the type of crops. The GB trees ma-

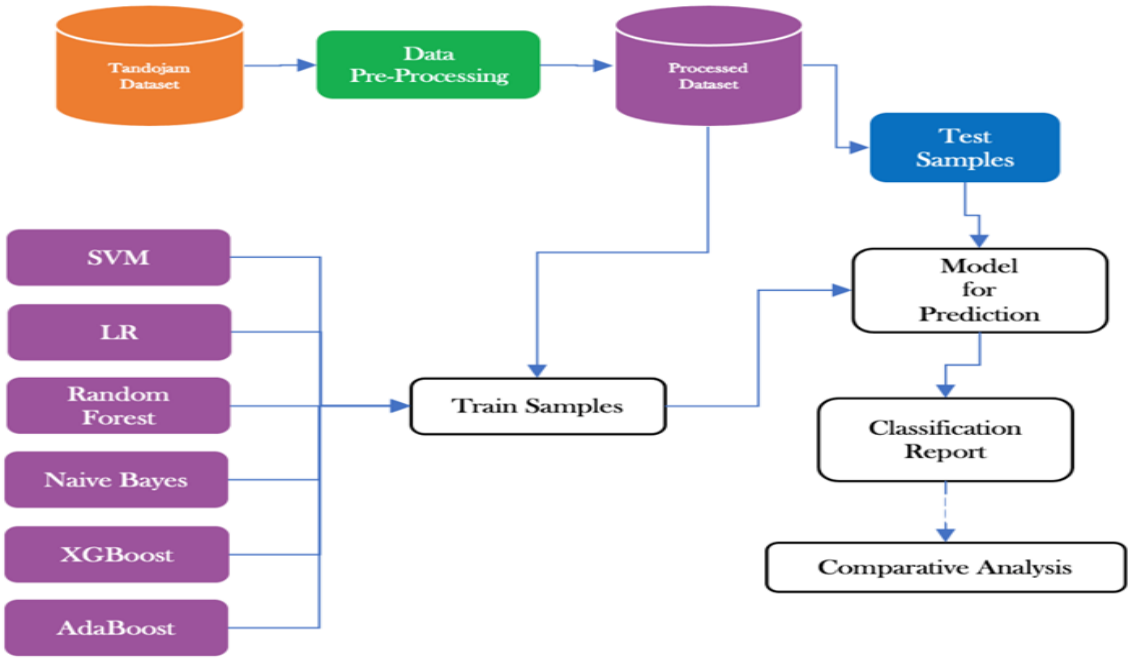


Fig. 1: Proposed Crop Prediction System

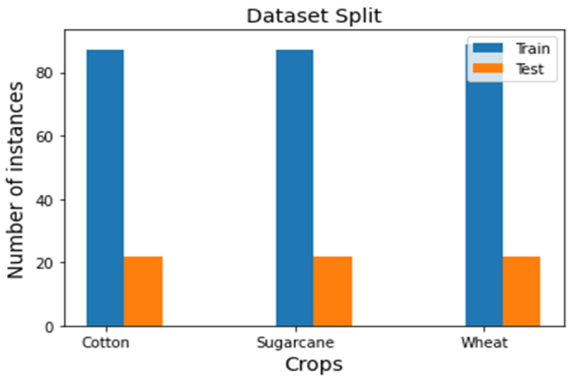


Fig. 2: Dataset Split

chine learning algorithm outperformed all the machine learning algo-rithms and achieved accuracy and F-score equal to 99.11% and 99.20% respectively.

In short, numerous machine learning-based approaches are used for the prediction of the best suitable crop for spe-cific land. These approaches achieved significant prediction accuracies and F1-scores on datasets that contain dif-ferent feature parameters. However, better prediction performance is still describable in-order to reduce the finan-cial losses that are faced by farmers due to planting of wrong crops.

3 Methodology

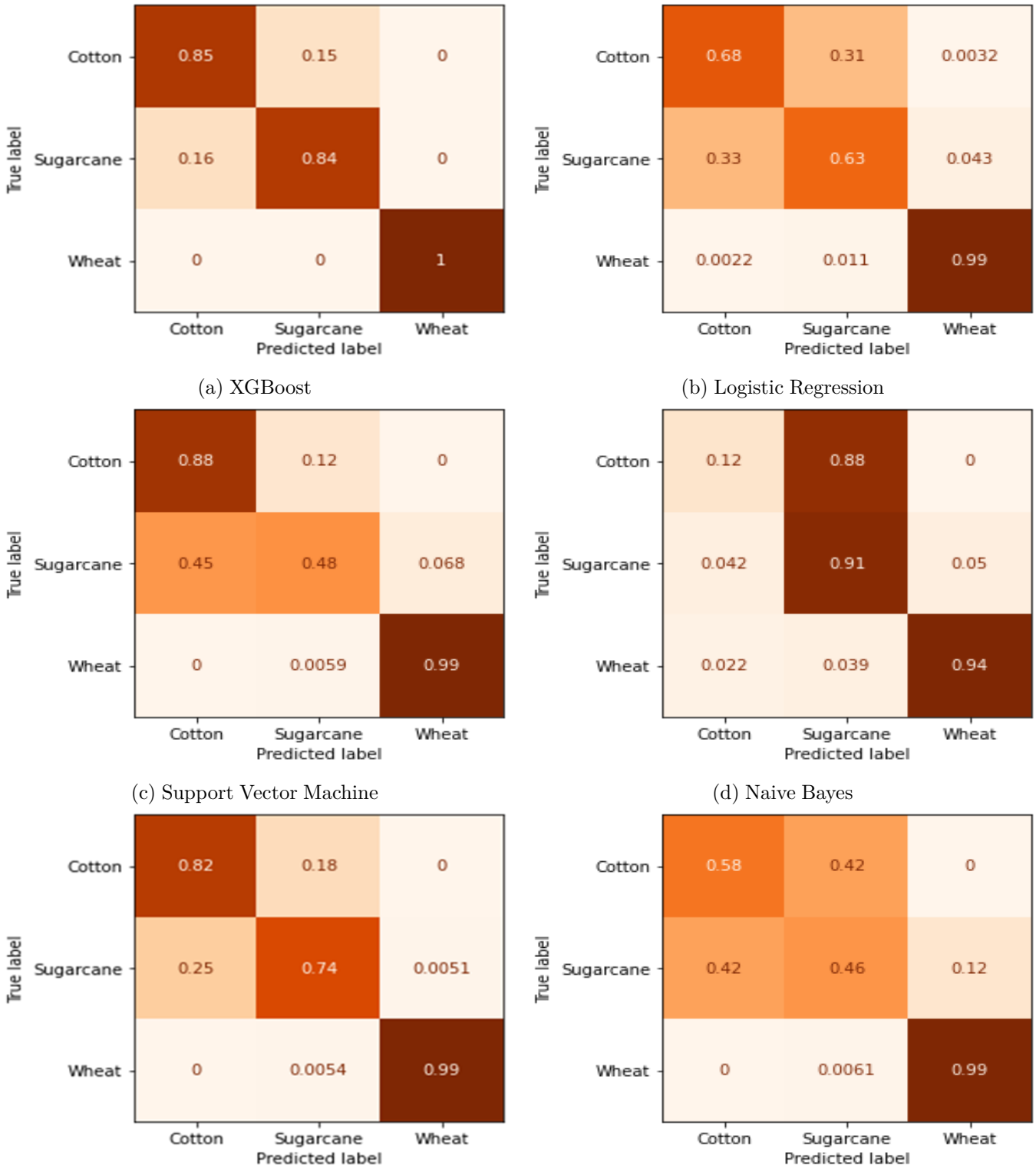
The proposed crop prediction system in this paper is shown in Fig. 1. The system first performs preprocessing to clean the dataset by removing duplicate and null entries, encoding labels, and balancing the class distribution. Af-ter preparing and distributing the dataset into training, validation, and test sets, the various machine algorithms are trained and tested and the system generated the classification report in terms of precision, recall, and f1-scores. Finally, we perform a comparative analysis of these machine learning algorithms. These steps are further ex-plained in the following sub-sections.

3.1 Data Collection

We prepared a dataset for crop prediction by collecting 329 samples which include 9 attributes such as crop name, month, temperature, rainfall, humidity, and soil features. These samples of the dataset have been obtained from Sindh Agriculture University (SAU) Tandojam field. We also considered soil qualities like pH, nitrogen, phosphorus, and potassium levels. Some of the rows of the dataset are shown in Table 1 where the last column indicates the label of the crop.

3.2 Preprocessing

A technique called data preprocessing is used to turn the original data into a clean data set. The dataset is col-lected in raw format, which makes analysis impractical. We perform various operations including removing du-plicate and null entries, encoding labels, and using the SMOTE approach[20] for increasing instances of data and balance class distribution. SMOTE



(a) XGBoost

(b) Logistic Regression

(c) Support Vector Machine

(d) Naive Bayes

(e) Random Forest

(f) Ada Boost

Fig. 3: Confusion Matrices for Different Classifiers

TABLE 1: Some Rows of Dataset

Month	Temp	Humidity	PH	Rain Fall	N	P	K	Crop Type
March	18.4	50	8.3	0	0.024	0.8	149	Sugarcane
April	26.3	45	8.4	0	0.026	0.7	237	Sugarcane
May	31.8	53	8.3	0	0.017	0.6	191	Sugarcane
December	15.5	55	8.5	0	0.023	0.5	159	Wheat
January	13.2	64	8.5	0	0.016	1.2	284	Wheat
January	16.6	53	8.4	0	0.025	0.6	176	Wheat
August	32.0	64	8.4	0	0.011	0.4	40	Cotton
August	34.2	56	8.5	0	0.020	0.8	235	Cotton
June	30.2	87	8.6	0	0.027	0.9	172	Cotton
June	30.2	80	8.4	0	0.019	0.7	247	Cotton

is a statistical technique that increases the number of instances in the dataset by generating new instances from existing smaller classes which we apply as input. The split of training and testing data is the last stage in the data preparation process. Since training the model often requires as many samples as feasible, the data is typically likely to be distributed unevenly. The initial dataset used to train ML algorithms to learn how to make correct estimates is known as the training dataset[21]. We perform the classification of three crops i.e., cotton, sugarcane, and wheat with a split ration 80:20 for training and testing as shown in Fig. 2.

3.3 Assessment of Machine Learning Algorithms

In this work, six machine learning algorithms i.e., Logistic Regression, Naive Bayes, Random Forest, Support Vec-tor Machines (SVM), XGBoost, and AdaBoost are assessed and their performance is compared on crop prediction dataset to select an algorithm that can be used for prediction of crops. These machine learning algorithms are the most widely used machine learning algorithms for crop prediction. We use various performance metrics such as accuracy, confusion matrix, ROC plot, precision, recall, and F1-scores to measure and compare the performance of these algorithms.

3.4 Web Application

To query the outcomes of machine learning analysis, a web app using the flask platform has been created. Any operating system will work with the software. This software offers an intuitive design that just needs a few clicks to acquire the needed results. The online program simply requires the location and size of the field to provide the name of the appropriate crop to grow there.

3.5 Development Tools

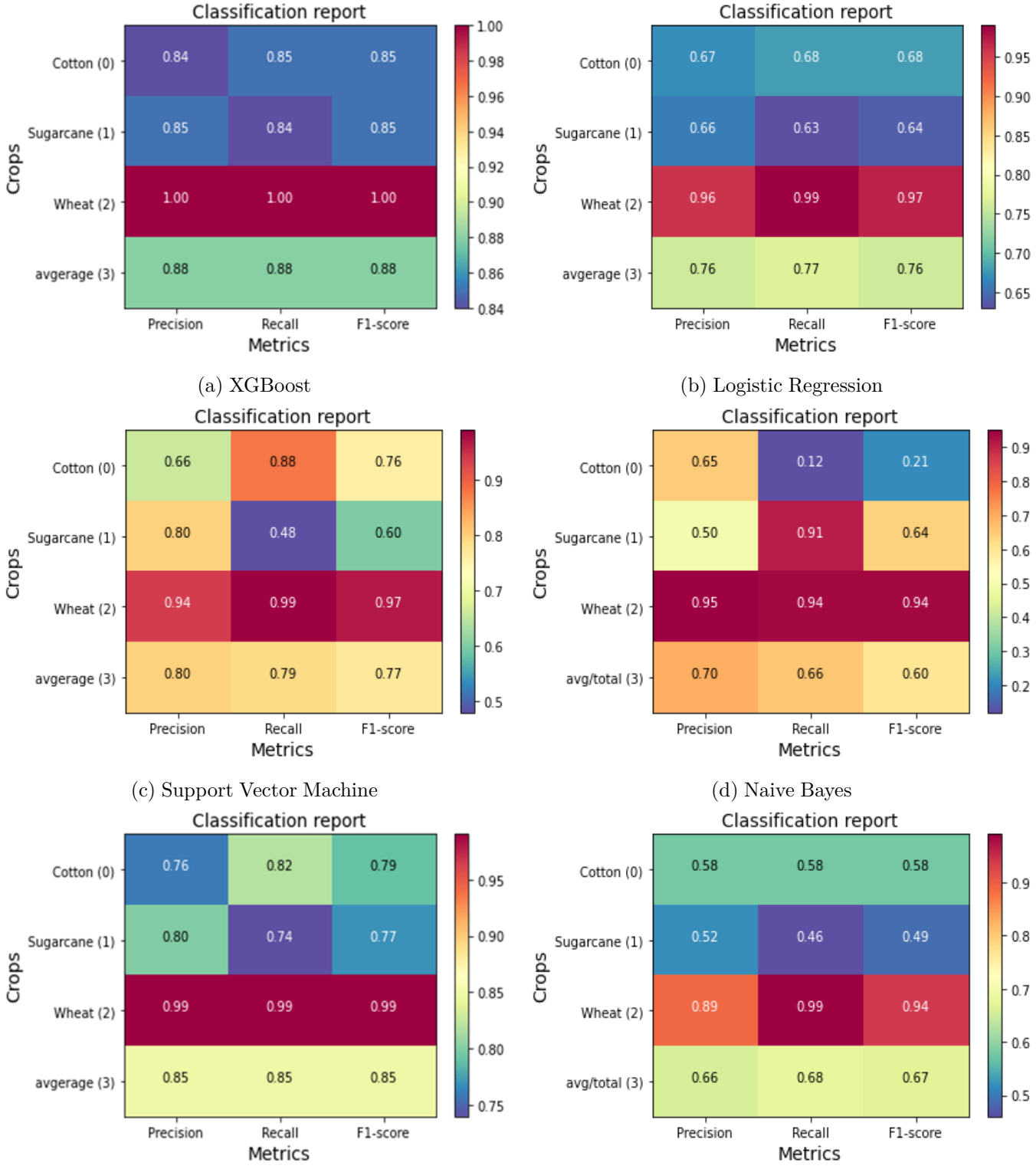
In this work, the python programming language is utilized as the foundation for machine learning analysis. Python has a little framework called Flask. The Jinja2 template engine and the WSGI tools are the foundation of Flask. The Flask is utilized in this work as the back-end framework for creating the application.

4 Results & Discussions

In this work, we train six machine learning algorithms on the dataset for crop prediction, and their performance is analyzed and compared. The machine algorithm that produces high accuracy forecasted of the correct crop is se-lected. In addition, an android application is created that shows the outcomes of the machine learning study. The crop name is shown in a web application built using Flask.

4.1 Experimental Results on Machine Learning Algorithms

Six classifiers i.e., Logistic Regression, Random Forest, Naïve Bayes, Support Vector Machine, XGBoost, and Ada-Boost are used on crop datasets and their performance is evaluated in terms of confusion matrix, precision, recall, and F1-scores. The confusion matrix illustrates the performance of a classifier by showing how many instances of each class of a test dataset are correctly classified and misclassified. The normalized confusion matrix is obtained from the confusion matrix by dividing each row of a confusion matrix with the sum of the entire row. Precision measures the ability of a classifier to classify the positive instances in the model where as recall measures the abil-ity of a classifier to detect positive instances. Fig. 3 and Fig. 4 respectively show the accuracy of each class of crop dataset using a normalized confusion matrix and their



(a) XGBoost

(b) Logistic Regression

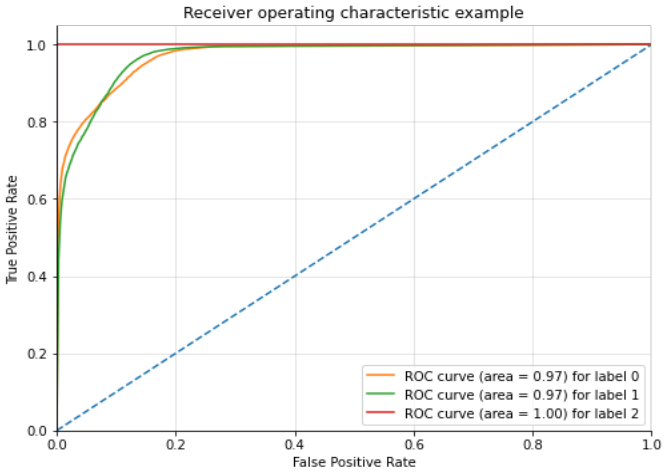
(c) Support Vector Machine

(d) Naive Bayes

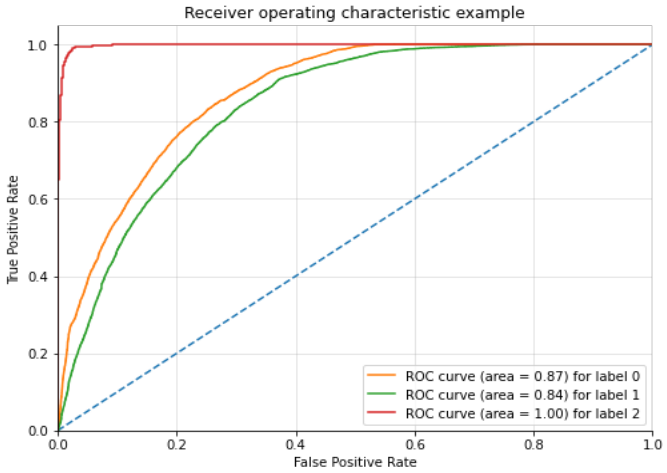
(e) Random Forest

(f) Ada Boost

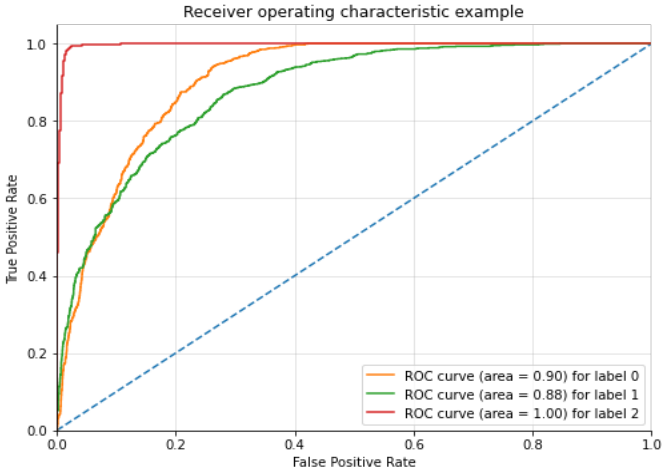
Fig. 4: Classification Reports for Different Classifiers



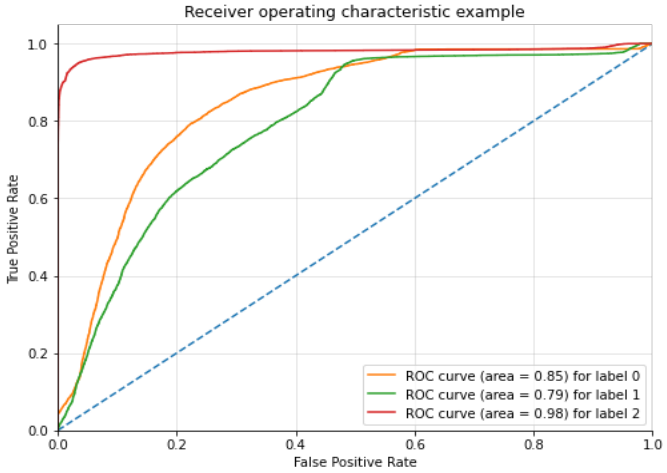
(a) XGBoost



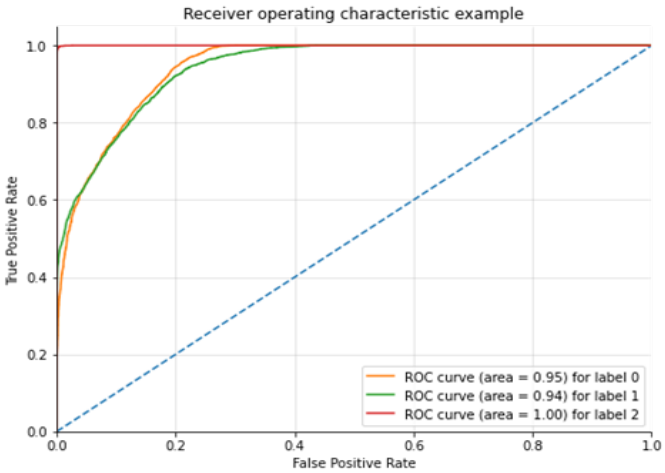
(b) Logistic Regression



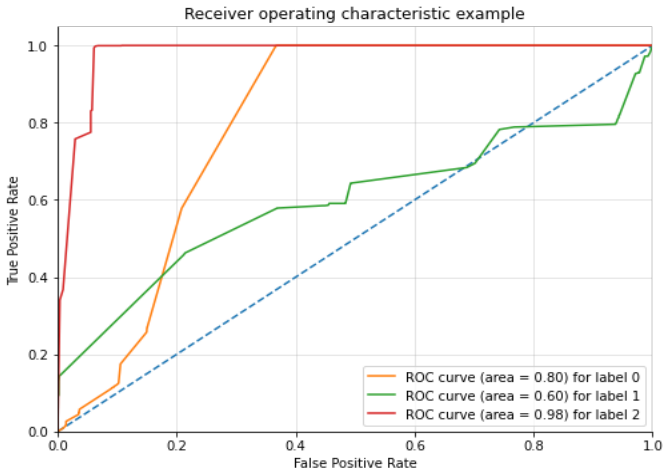
(c) Support Vector Machine



(d) Naive Bayes



(e) Random Forest



(f) Ada Boost

Fig. 5: ROC Plots for Different Classifiers

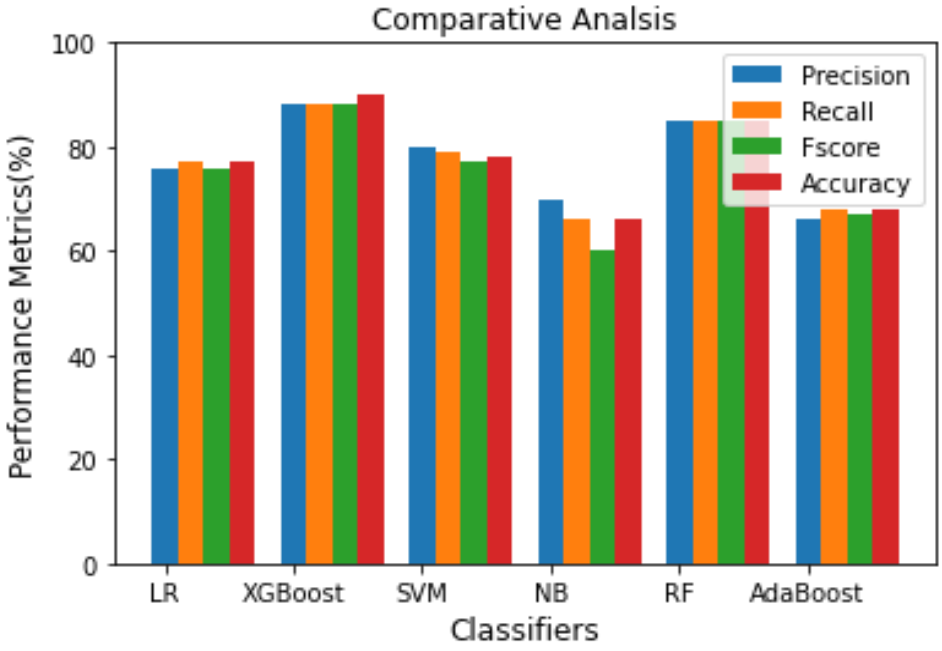


Fig. 6: Performance comparison of different classifiers

performance in terms of precision, recall, and F1 scores on these machine learning algorithms. XGBoost classifier achieved the highest crop prediction accuracy and precision, recall and F1-scores as shown in Fig. 3(a) and Fig. 4(a) respectively. Using this classifier, 85%, 84%, and 100% accuracies are achieved for cotton, sugarcane, and wheat crops respectively. The average accuracy achieved is 90%. The ability of classifiers i.e., how much classifiers are capable to predict only samples of their class is described by ROC plots as shown in Fig. 5. In this Fig (i.e., Fig. 5), the nearness of curves in ROC plot towards their top left corner indicates better performance.

4.2 Comparative Analysis

In this paper, the performance of various machine algorithms on crop prediction datasets is analyzed and com-pared. This comparative analysis is shown in Fig. 6 and also presented in Table 2. From this comparative analysis, it can be seen that XGBoost classifier achieved the highest perdition accuracy (i.e., 90%) and precision, recall, and F1-scores (i.e., 88%) as compared to other classifiers. Using the Random Forest classifier, the second-highest results are achieved. In this case, the prediction accuracy achieved is equal to 85%, and precision, recall, and F-scores are equal to 85%. According to comparative analysis, the Naive Bayes classifier achieved the lowest accuracy and precision, recall, and F1-scores as compared to other classifiers. Using this comparative analysis, a classifier that achieved the highest prediction results (i.e., in this case, XGBoost) is selected and can be used for the prediction of a crop that is to be grown on particular land based on soil and weather parameters.

4.3 Web Application

A simple web application is also created that outputs the prediction. This application assists farmers in choosing which crop to cultivate to elicit the greatest return. For this purpose, Tandojam's profusely producing crops are se-lected and their names are foreseen. Utilizing XGBoost (i.e., achieved the highest prediction accuracy), the prepro-cessed dataset is trained. Prediction is done using the immediate weather data for the chosen district that is acces-sible via API. For the chosen district, the trained model produced the appropriate crop prediction. The prediction output when the user inputs sample input data is shown in Fig. 7.

5 Conclusion and Future Work

In this work, the performance of various machine learning algorithms on a particular dataset is analyzed and compared. XGBoost and Random Forest classifiers achieved significant prediction results as compared to other classifiers. The performance of these classifiers can be improved by collecting more samples in the dataset. This study proposes an effective crop prediction framework that employs a machine learning

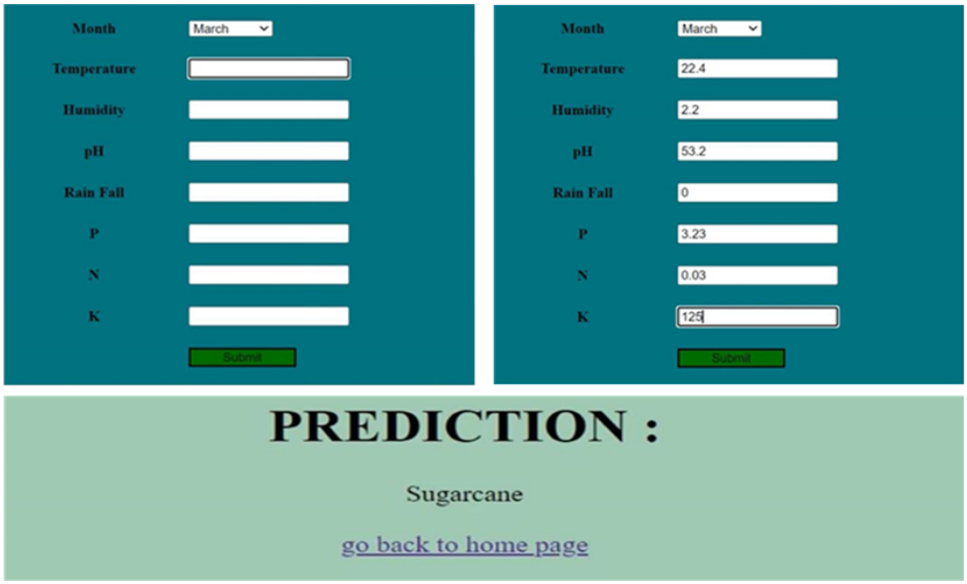


Fig. 7: App Snapshots

TABLE 2: Performance Metrics of different classifiers

ML Classifiers	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	77	76	77	76
Random Forest	85	85	85	85
Support Vector Machine	78	80	79	77
Naive Bayes	66	70	66	60
XGBoost	**90**	**88**	**88**	**88**
AdaBoost	68	66	68	67

technique to suggest ap-propriate crops based on input soil and weather parameters at Sindh Agriculture University (SAU) Tandojam field. The farmers will experience fewer financial losses as a result of planting the incorrect crops, and it will also assist farmers in discovering new crop varieties that can be grown in their region.

In the future, we collect more data samples for crops in the dataset. In addition, the dataset can also be extended by including more crop classes. To improve the precision of our prediction model, we may also employ hybrid ML and deep learning models. By making this study available throughout the country, it may be improved to a higher degree.

References

[1] Suresh, G., A. Senthil Kumar, S. Lekashri, R. Manikandan, and C. O. Head. "Efficient crop yield recommendation system using machine learning for digital farming." International Journal of Modern Agriculture 10, no. 1 (2021): 906-914.

[2] Venugopal, Anakha, S. Aparna, Jinsu Mani, Rima Mathew, and Vinu Williams. "Crop yield prediction using machine learning algorithms." International journal of engineering research & technology (IJERT) NCREIS 9, no. 13 (2021).

[3] Kalimuthu, M., P. Vaishnavi, and M. Kishore. "Crop prediction using machine learning." In 2020 third international conference on smart systems and inventive technology (ICSSIT), pp. 926-932. IEEE, 2020.

[4] Nischitha, K., Dhanush Vishwakarma, Mahendra N. Ashwini, and M. R. Manjuraju. "Crop prediction using machine learning approaches." International Journal of Engineering Research & Technology (IJERT) 9, no. 08 (2020): 23-26.

[5] Battilani, Paola, Amedeo Pietri, Carlo Barbano, Andrea Scandolara, Terenzio Bertuzzi, and Adriano Marocco. "Logistic regression modeling of cropping systems to predict fumonisin contamination in maize." Journal of Agricultural and Food Chemistry 56, no. 21 (2008): 10433-10438.

[6] Wbb., G. "Naïve Bayes." (2016), pp. 1–2.

[7] Rigatti, Steven J. "Random forest." Journal of Insurance Medicine 47, no. 1 (2017): 31-39.

[8] Hearst, Marti A., Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. "Support vector machines." IEEE Intelligent Systems and their applications 13, no. 4 (1998): 18-28.

[9] Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, and Tianyi Zhou. "Xgboost: ex-

treme gradient boosting." R package version 0.4-2 1, no. 4 (2015): 1-4.
[10] Schapire, Robert E. "Explaining adaboost." Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik (2013): 37-52.
[11] Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32.
[12] Mishra, Subhadra, Debahuti Mishra, and Gour Hari Santra. "Applications of machine learning techniques in agricultural crop production: a review paper." Indian J. Sci. Technol 9, no. 38 (2016): 1-14.
[13] Priya, P., U. Muthaiah, and M. Balamurugan. "Predicting yield of the crop using machine learning algorithm." International Journal of Engineering Sciences & Research Technology 7, no. 1 (2018): 1-7.
[14] Patil, Pavan, Virendra Panpatil, and Shrikant Kokate. "Crop prediction system using machine learning algorithms." Int. Res. J. Eng. Technol.(IRJET) 7, no. 02 (2020).
[15] Kalimuthu, M., P. Vaishnavi, and M. Kishore. "Crop prediction using machine learning." In 2020 third international conference on smart systems and inventive technology (ICSSIT), pp. 926-932. IEEE, 2020.
[16] Sangeeta, Shruthi G. "Design and implementation of crop yield prediction model in agriculture." International Journal of Scientific & Technology Research 8, no. 1 (2020): 544-549.
[17] Kumar, J. "Supervised Learning Approach for Crop Production." (2020).
[18] Abirami, G., R. Reni Hena Helan, K. Anandhan, G. Vishnuvardhan Reddy, S. Naveen Kumar, and V. Ruban Karthick. "Crop Yield Prediction Using Ensemble Algorithm." International Journal of Modern Developments in Engineering and Science 1, no. 6 (2022): 54-59.
[19] Bhuyan, Bikram Pratim, Ravi Tomar, T. P. Singh, and Amar Ramdane Cherif. "Crop Type Prediction: A Statistical and Machine Learning Approach." Sustainability 15, no. 1 (2023): 481.
[20] Fernández, Alberto, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." Journal of artificial intelligence research 61 (2018): 863-905.
[21] Tan, Jimin, Jianan Yang, Sai Wu, Gang Chen, and Jake Zhao. "A critical look at the current train/test split in machine learning." arXiv preprint arXiv:2106.04525 (2021).