Crop Yield Prediction Using Machine Learning: A Pragmatic Approach

Rajswee Surana

suranarajswee17@gmail.com

Student-MCA, IIS (Deemed To Be University)

Ritu Khandelwal

IIS University

Research Article

Keywords: Crop Prediction, Data Mining, Machine Learning, Algorithms

Posted Date: July 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4575893/v1

Additional Declarations: No competing interests reported.

Crop Yield Prediction Using Machine Learning: A Pragmatic Approach

Rajaswee Surana[1], Ms. Ritu Khandelwal[2]

[1] Student-MCA, IIS (Deemed To Be University), Jaipur
[2]Assistant Professor, International School of Informatics & Management

Abstract: The agricultural sector is the major driver of revenues in India. organic, economic, and seasonal factors all have got a bearing on an agricultural producer`s manufacturing.
Accurate crop yield prediction (CYP) is required due to the agricultural industry`s rapid innovation and liberalized market economy. Accurate prediction is greatly aided by the chosen characteristics and machine learning (ML) techniques Any ML Algorithm's performance may be enhanced by using a unique set of features from the same training dataset. This study assesses the key characteristics of an accurate crop yield prediction. For greater accuracy, it uses machine learning (ML) methods like Random Forest (RF), Adaboost, Gradient Boost, and Support Vector Machine (SVM). The agriculture dataset has 2201 instances in it. 80% of them are randomly chosen for the model`s training, while 20% are used to test the model`s predictive power. The results show that the Random Forest approach gets the highest level of accuracy. The goal of this paper is to predict crop yields. Using different Machine Learning Algorithms so that farmers can make their yields higher.

Keywords: Crop Prediction, Data Mining, Machine Learning, Algorithms

1. INTRODUCTION

Horticulture is vital to the economic growth of a nation such as India. Given that it supplies an avenue of money and work for the rural poor. It's an essential resource of nourishment to people worldwide. Production of crop needs is additionally rising with time. It makes up around 20% of India's GDP. They have accumulated understanding through the years. Due to obsolete techniques, producers are ignorant of their stake in the contemporary horticulture business. As a result of this, agricultural businesses endure. Crop selection is an important aspect of farming planning. Producers' expenditures decrease when they're given the ability to obtain reliable data on the best products for their land and season. Developing a model for prediction entails collecting knowledge from trustworthy sources, as this impacts the model's predictability Artificial intelligence (AI) has a subfield called machine learning that enables computers to learn from data without explicit programming. Because it can find patterns and correlations in

enormous volumes of data and base forecasts on these relationships, it is the perfect tool for predicting agricultural productivity.

The use of machine learning enables entities to generate educated decisions through the examination of future trends and features. The dichotomy is established by recognizing, assessing, and organizing items and opinions into particular categories or "subpopulations." Using pre-categorized instructional datasets, machine learning programs employ various methods to recognize upcoming data. Classification algorithms in machine learning utilize input data to foresee the chance that the subsequent information will fall within one of the predetermined classes.

The suggested model is more accurate than the current ones, as evidenced by the fact that the forecast accuracy of earlier accuracy models was lower than that of the model now being presented in this paper.

2. LITERATURE REVIEW

The Multiple Linear Regression strategies and the Density-based clustering technique were employed in the analysis phase of the region-specific crop production study, as stated by the researcher in the study's publication [1]. even though these designs have been evaluated in all of Andhra Pradesh's districts, only the East Godavari district was consulted in the evaluation process. Concerning the East Godavari District sample data, the exact amount as well as the associated calculated value were determined via the Multiple Linear Regression technique.

Data analysis has numerous applications in farming. The scientists [2] looked at the efficacy of information in the classification algorithms J48, Random Forest, and AdaBoost, and Using a variety of rating criteria, such as the false positive rate, actual positive rate, precision, ROC area, and F-Measure, Naive Bayes is used. It facilitates the creation of an effective crop forecast system.. The results show that J48 is the best prediction technique for the decision tree, with an accuracy of 89.33% on the data that has been provided set, while Random Forest and Adaboost yield reliability with values of 85.66% and 82.66%, correspondingly.

Researchers [3] uncovered that owing to the rapid expansion of science and technology and associated fields, quasi-technological fields have become simpler. Machine learning and agriculture are two of the most stimulating study fields that are gaining popularity among researchers. As a consequence, limitations on machine learning applications for farming are also tackled. This limitation causes the identification of more potential accomplishments for the field in question's swift development.

Researchers [4] aimed to discover further the random forest method's handling of region-specific agricultural yield predictions. In that particular endeavour , 80 percent of the data serves for instruction, whereas the rest, or twenty percent, is utilized for assessment. Following effective testing and training, a model's preciseness gets assessed.

Employing prior statistics, the authors [5] constructed a robust harvest projection framework. Datasets undergo evaluation utilizing Xarray functions to figure out the crop depending on geography and seasonality utilizing a trend-matching strategy. Producers may employ the advised methodology for choosing crops that are suitable depending on the time of the year and region. Ranchers are going to succeed by minimizing waste and broadening the net production of crops.

Considering these scientists [6], all of the machine learning (ML) methods beneath evaluation could potentially utilize for assessing the productivity of agriculture. Naive Bayes models had the poorest precision, 72.33%, whereas the kernel neural network while random forest modelling had the highest, 88.67% and 94.13%, correspondingly. Regarding exactness, ANN anticipated the highest benefit—99.94%—while Logistic Regression anticipated the weakest—24.17%. Apart for the naive Bayes model, each classification system considered predicts a recalled admire of more than 90%. Logistic regression had the highest rate of false positives and the smallest actual negative percentage, while Naive Bayes got the overall greatest incorrect negative incidence. ANN and KNN had the greatest f-scores, having characteristics around 99.78% & 80.72%, correspondingly, along with the strongest clarity.

Linear Regression surpassed K Nearest Neighbour concerning predictive power and accuracy, as reported by the researchers [7]. The regression technique's capacity utilization for foreseeing yields was investigated. The data were utilized as model inputs. fortunately substitute elements including the environment, used for farming practices, or soil characteristics were taken into consideration in the equation's creation, specifications like as year, crop, area, and quantity produced (in tonnes) might have offered deeper estimation influence, as well as linear regression techniques continued to offer appropriate forecasting preciseness.

3. PROPOSED FRAMEWORK

The advocated mechanism in Fig.1 intends to help farmers nurture harvests for higher volumes. The crops designated under the endeavour were essential in the stipulated locale. Crop preference is an important aspect of agrarian strategy. Farmers' declines diminish when they have access to precise details on the best harvest for their land and season. throughout the past few decades, a yield from agriculture data set has been compiled from a variety of evidence.

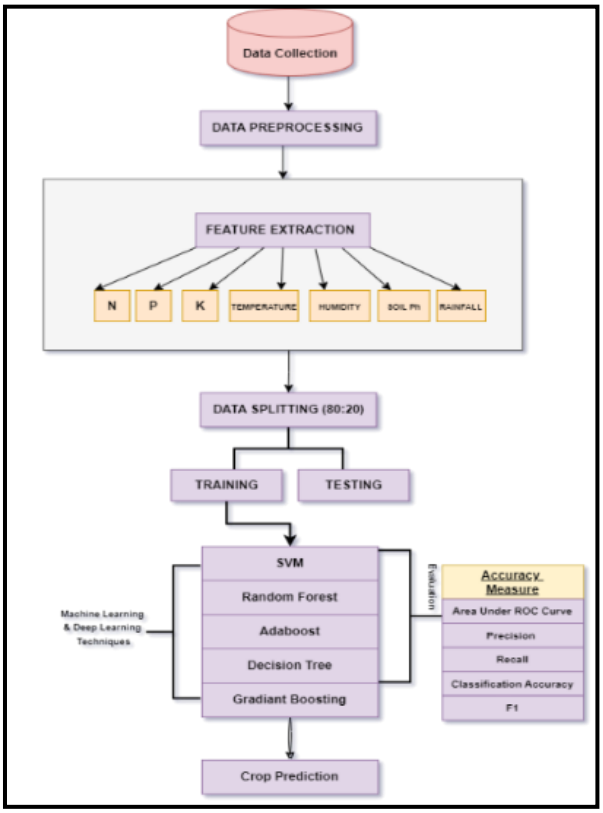

Fig. 1. The Proposed Model's Architecture

3.1 Dataset Description

Data for this study was gathered from Kaggle.com. The target label with the name crop is present in the relevant dataset. Rice, apple, maize, chickpea, and other foods are values for the same. It also contains the factors affecting the growth of the crops such as the ph. of the soil, rainfall, temperature, and humidity. The dataset has 2201 instances.

3.2 Applied Algorithms

Harvest success is exceedingly crucial data for farmers. Understanding the profit margin is fairly advantageous since it ultimately results in an offset in depreciation. In earlier times, qualified farmers speculated the yields of crops. The idea put forward works in the same way. It forecasts future yields based on existing information.

Decision Tree

A decision tree is an overarching learning technique that can be used for both classification and regression problems, however, it is most frequently used for classification. It is a tree-structured classification where each leaf node denotes the outcome and nodes within it contain data properties, decision-making rules, and pathways.
The expected value of each result can be determined as follows:
Expected value (EV) = (First possible outcome x Likelihood of outcome) + (Second possible outcome x Likelihood of outcome) - Cost

SVM Algorithm (Support Vector Machine)

The procedure known as SVM is aimed at finding the ideal line or threshold for categorizing space in n dimensions, so that we may only add additional information in the appropriate category in the future. The best suitable barrier is a higher dimensional space. SVM chooses the highest and lowest observations point/vectors that will help create a hyperplane.

The hyperplane can be represented by the equation:

$$w.x + b = 0$$

The distance between the hyperplane and the closest point to it in each class is calculated by :

$$\text{distance} = y(w.x + b)/\|w\|$$

Random Forest Algorithm

The Random Forest Approach is a technique that utilizes an assortment of varying decision trees as shown in Fig.3 that use various elements of the information at hand and adopt the mean to improve the forecasting accuracy of that dataset.

How Does Random Forest Works?

Step 1 − Start by choosing random samples from a pre-existing dataset.

Step 2 − A decision tree will then be built by this method for each sample. The forecast outcome from each decision tree will then be obtained.

Step 3 − Voting will be done in this phase for each expected outcome.

Step 4 − Finally, choose the prediction result that received the most votes as the final forecast result. The average of all the trees' forecasts forms the final forecast:

$$\text{y\_hat} = (1/T) * \text{sum(y\_hat\_t)}$$

This formula computes the average of the predictions y_hat_t from all the trees in the ensemble, effectively smoothing out the noise and reducing the variance of the final prediction.

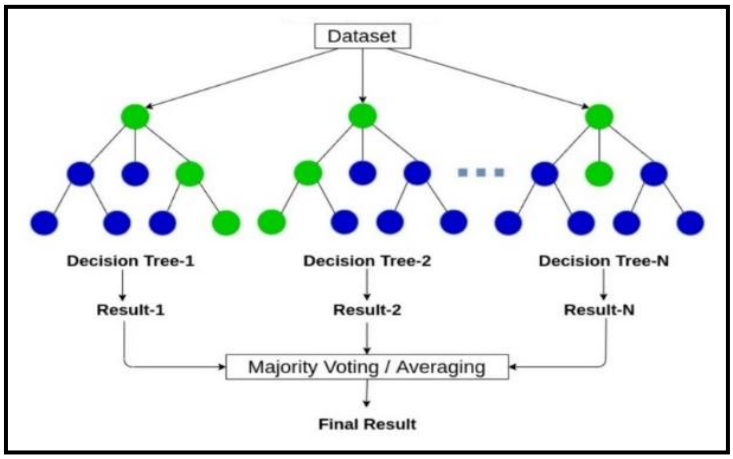

Fig 3 . The procedure of Random Forest Algorithm

Adaboost Algorithm

It integrates multiple classifiers to increase the precision of classifiers. AdaBoost is an incremental collection composition approach. The AdaBoost classifier builds an efficient predictor by combining many negligible-performing classification algorithms, resulting in a hot-accuracy strong classifier.

Gradient Boosting Algorithm

The gradient boosting process can forecast both continuous and categorical target variables (as a Regressor and a Classifier).

The gradient Boosting Model pertains to a genre of machine learning models that incorporates additional algorithms like random forest models and neural networks, among others. If the data is non-linear, sparsely occupied has a poor refill rate, or merely when regress fails to produce the expected conclusions, it could potentially be used as opposed to regression.

3.3 Performance Evaluation

The evaluation method commences with the acquisition of a promptly collected crop dataset. Data in the pre-processing stage is the next crucial step in this set after the import. The information is then divided into sets for training and testing. Following that, a framework appears in which the requisite AI algorithms serve to distill the most promising crop that ought to be fostered on one particular piece of terrain.

To calculate the Accuracy, Precision, and Recall we have used the following formulas

$$\text{Accuracy} = (TP + TN) / (TP + FP + TN + FN)$$

One of the most used performance metrics is accuracy. It is the proportion of observations that were successfully predicted to all of the observations.

$$\text{Precision} = TP / (TP + FP)$$

The percentage of correctly identified positive cases among all projected positive cases is the implicit definition of precision. It concerns how well your model can forecast the actual positives. It emphasises Type I mistake.

$$\text{Recall} = TP / (TP + FN)$$

It's the proportion of accurately detected positive cases among all positive instances.

wherein TP is interpreted as when each instance is presupposed to be advantageous but attains the age of out to be optimistic. When scenarios show up to be either advantageous or detrimental in reality, FP emerges. TN develops when events are predicted to be unfavorable but inevitably turn out to be unfavorable FN arises when domains that were portrayed to be negative ultimately prove to be affirmative.

4. RESULT & DISCUSSION

Based on experimental findings, it is possible to predict agricultural yield using all of the ML algorithms now being researched. Adaboost had the lowest accuracy prediction at 99.03 percent, while random forest had the highest accuracy of 99.99%. In terms of accuracy, Decision Tree predicted the lowest value (97.01%), while Random Forest predicted the greatest value of 98.98%. For this experimental analysis, the orange tool has been employed.

4.1 Findings

A. When we assessed the data set via a cross-validation model with 5 folds of numbers as shown in Fig.5, the precision, AUC, F1, CA, and recall were as follows:

MODEL	AUC	CA	F1	PRECISION	RECALL
AdaBoost	0.993	0.987	0.987	0.987	0.987
Gradient Boosting	0.999	0.988	0.988	0.988	0.988
Random Forest	0.999	0.993	0.993	0.993	0.999
SVM	0.999	0.981	0.981	0.981	0.981
Tree	0.99	0.976	0.976	0.976	0.976

Fig. 5. Reviewed Data set using Cross Validation Model

B. When we reviewed the dataset using a random sampling model as shown in Fig.6, the AUC, F1, CA, and Recall were as follows:

MODEL	AUC	CA	F1	PRECISION	RECALL
AdaBoost	0.991	0.981	0.981	0.981	0.981
Gradient Boosting	0.999	0.981	0.981	0.981	0.981
Random Forest	0.999	0.988	0.988	0.988	0.988
SVM	0.999	0.977	0.977	0.977	0.977
Tree	0.986	0.971	0.971	0.971	0.971

Fig.6. Reviewed Data set using Random Sampling Model

C. the values of different parameters in the dataset , how they are related to each other are shown below:

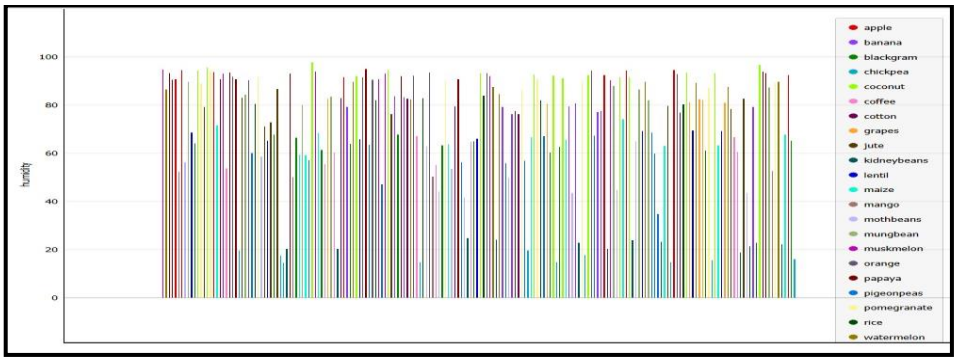

Fig.7.: Bar Plot Showcasing values of Humidity

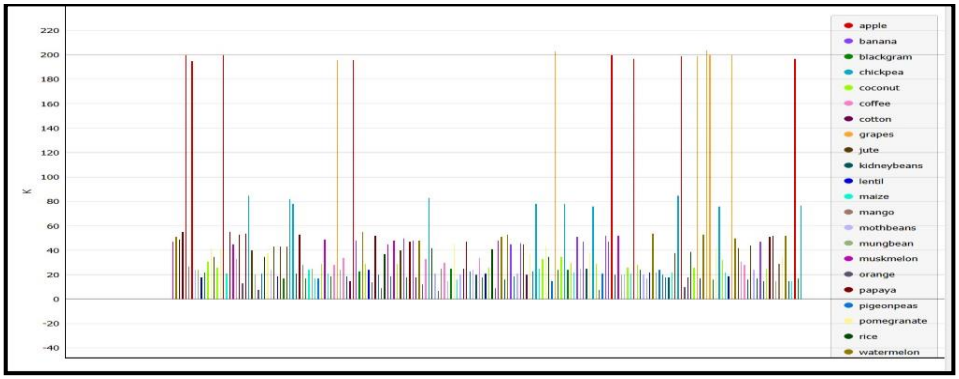

Fig.8.: Bar Plot Showcasing values of K

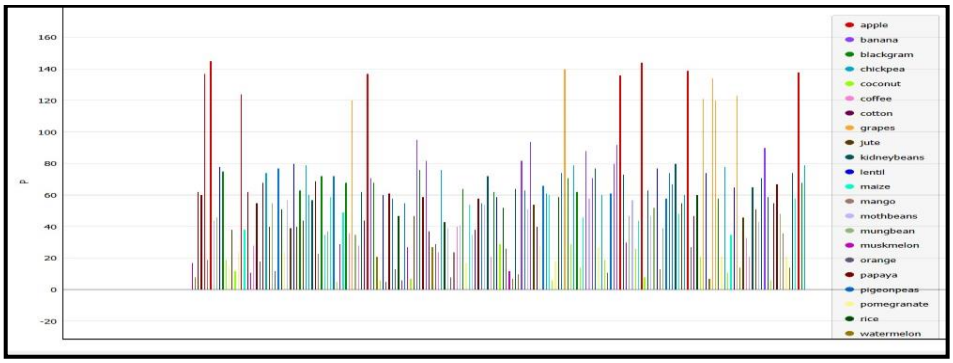

Fig.9.: Bar Plot Showcasing values of P

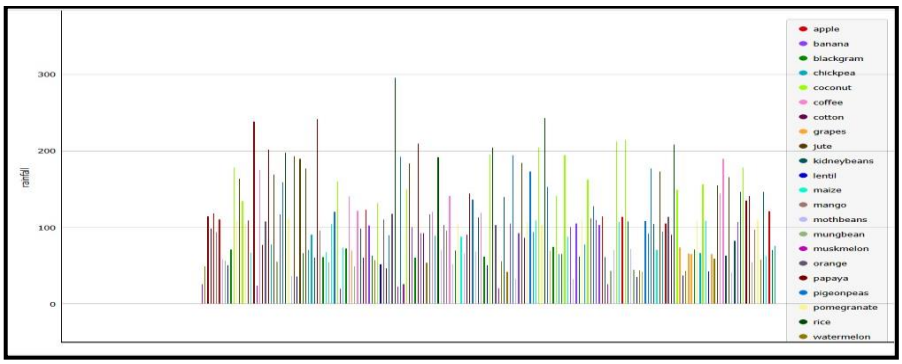

Fig.10.: Bar Plot Showcasing values of Rainfall

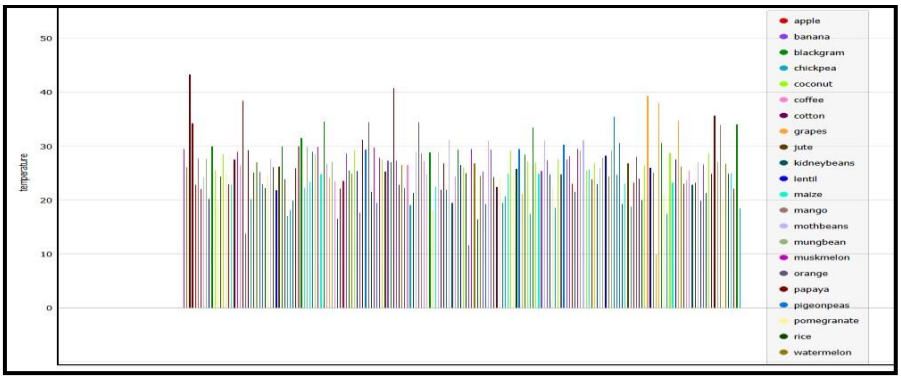

Fig.11.: Bar Plot Showcasing values of Temperature

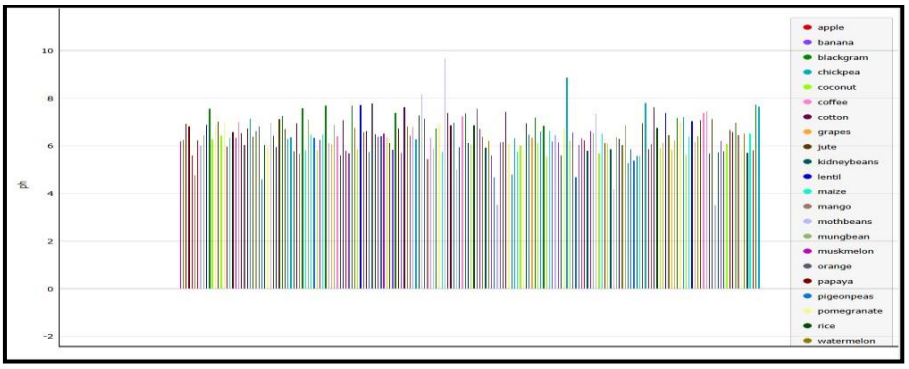

Fig.12.: Bar Plot Showcasing values of Ph

4.2 Performance Analysis

The data gathered was examined as shown in Fig.7 according to a number of defined parameters such as temperature, rainfall, soil ,ph ,P ,K, N, Humidity. The aforementioned are traits that are employed to delve into the raw data and derive the ideal yield. Crops are germinated with the mission to boost production upon an isolated tract of the area. In Fig.8 the Accuracy of the models employed in the experiment is shown.

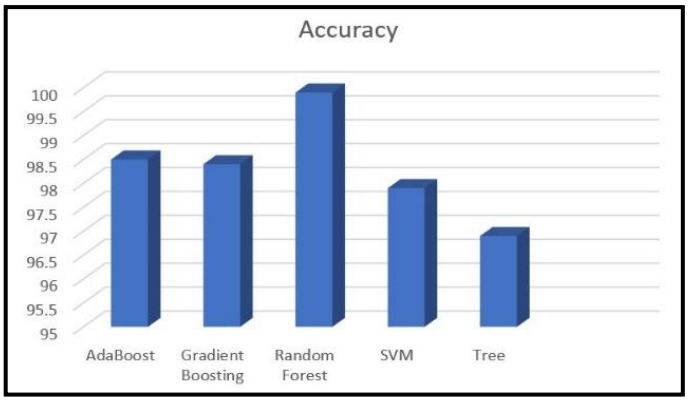

Fig. 13. Bar Graph showing the Accuracy of the models

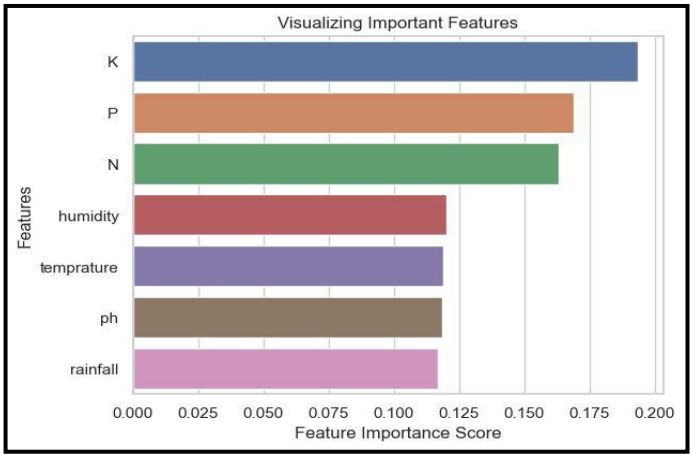

Fig. 14. Features Importance using Random Fores

5. CONCLUSION

The suggested structure is being created using machine learning to help ranchers analyze agricultural losses brought on by a lack of knowledge about farms under varied soil and weather conditions. In our study, we used a range of assessment metrics, such as false positive rate, true positive rate, recall, precision, ROC area, and F-1 score, to compare the performance of the classification algorithms such as Decision Tree, Adaboost, SVM, Gradient Boost, and Random Forest.

It supports the creation of an effective crop forecasting system. The generated findings show that Random Forest is the best predictor method, with a 99.93% accuracy rate on the supplied data. An extensive dataset, 5-fold cross-validation, and a Random Sampling model are utilized in this comparative crop prediction study to demonstrate the predictive ability of the employed data mining techniques. The exactness has been verified to be 99%. It has plenty of room for expansion and can be put into practice and coupled with a diversified and multifaceted application. Ranchers should be enlightened. The improvement in the field of agriculture is going to be extremely helpful, which will assist producers in producing crops.

REFERENCES

[1] Ramesh, D., & Vardhan, B. V. (2015). Analysis of crop yield prediction using data mining techniques. *International Journal of Research in Engineering and Technology*, *4*(1), 47-473.

[2] Patel, H., & Patel, D. (2016). A comparative study on various data mining algorithms with special reference to crop yield prediction. *Indian Journal of Science and Technology*, *9*(22), 18.

[3] Madhubhashini, D. A. P. Data Mining for Smart Agriculture *research*, *64*(2), 394-398.

[4] Kamath, P., Patil, P., Shrilatha, S., & Sowmya, S. (2021). Crop yield forecasting using data mining. *Global Transitions Proceedings*, *2*(2), 402-407.

[5] Kedlaya, A., Sana, A., Bhat, B. A., Kumar, S., & Bhat, N. (2021). An efficient algorithm for predicting crops using historical data and pattern-matching techniques *Global Transitions Proceedings*, *2*(2), 294-298.

[6] Pandith, V., Kour, H., Singh, S., Minhas, J., & Sharma, V.(2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. *Journal of scientific research*, *64*(2), 394-398.

[7] Van Kloppenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709

[8] Reddy, D. J., & Kumar, M. R. (2021, May). Crop yield prediction using a machine learning algorithm.

In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1466-1470). IEEE

[9] Mishra, S., Paygude, P., Chaudhary, S., & Idate, S. (2018, January). Use of data mining in crop yield prediction. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)* (pp. 796-802). IEEE.

[10] Surya, P., & Aroquiaraj, I. L. (2018). Crop yield prediction in agriculture using data mining predictive analytic techniques. *International Journal of Research and Analytical Reviews*, *5*(4), 783-787.

[11] Bondre, D. A., & Mahagaonkar, S. (2019). Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *International Journal of Engineering Applied Sciences and Technology*, *4*(5), 371-376.

[12] Pant, J., Pant, R. P., Singh, M. K., Singh, D. P., & Pant, H. (2021). Analysis of agricultural crop yield prediction using statistical techniques of machine learning. *Materials Today: Proceedings*, *46*, 10922-10926.

[13] Kale, S. S., & Patil, P. S. (2019, December). A machine learning approach to predict crop yield and success rate. In *2019 IEEE Pune Section International Conference (PuneCon)* (pp. 1-5). IEEE.

[14] Oikonomidis, A., Catal, C., & Kassahun, A. (2023). Deep learning for crop yield prediction: a systematic literature review. *New Zealand Journal of Crop and Horticultural Science*, *51*(1), 1-26.

[15] Wu, J., & Zhao, F. (2023). Machine learning: An effective technical method for future use in assessing the effectiveness of phosphorus-dissolving microbial agroremediation. *Frontiers in Bioengineering and Biotechnology*, *11*.