iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70844

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Machine Learning Based Crop Prediction and Recommendation System

Dik Sharma[1], Hritik Raj[2], Chithra H N[3]
Research Scholar, BANGALORE, INDIA

Abstract: ***Over the past few years, the convergence of machine learning (ML) and artificial intelligent computing has transformed numerous industries, with agriculture being one of the major beneficiaries. This paper explores the usage of ML algorithms in crop yield prediction and the optimization of agricultural operations, with a focus on the Indian context, where agriculture is an important sector. Crop yield prediction is greatly dependent on environmental factors like soil content, humidity, rainfall, and other cultivation parameters. But conventional approaches, like historical averages, may not be able to capture the dynamic behaviour of these factors, and hence may not provide correct predictions. In order to overcome this limitation, the present research uses a variety of supervise learning models, viz., Random Forest, Naïve Bayes Classifier, and Decision Trees, to forecast the yield of crops and determine the appropriate crops for a given terrain under given weather conditions and soil characteristics. Results indicate that Naïve Bayes Classifier in crop classification gives the best predictions of crop yield with an accuracy of 99.74%, while. In addition, through the incorporation of IoT sensor inputs, the system provides farmers with data-driven information about crop locations, optimal planting, watering, and fertilizing strategies. The research proves that decision support systems based on ML can equip farmers with the means to maximize crop yields, minimize waste, and reduce environmental footprint, ultimately leading to more sustainable and resilient agricultural practices.***
Keywords: Crop Prediction, Machine Learning, Naïve Bayes, Artificial Intelligence, Random Forest, SVM.

I. INTRODUCTION

The primary source of revenue in India is agriculture providing jobs to a large population and contributing heavily to country's GDP. With more than 60% of India's geographical area used for agriculture to sustain a population of more than 1.3 billion, the industry is under increasing pressure to enhance productivity and sustainability [1]. Historically, farmers have made decisions about what to plant and when by using personal judgment, past patterns, and local custom. These approaches usually ignore critical environmental considerations like soil nutrient levels, weather patterns, and moisture levels. Consequently, improper crop selection, overapplication or underapplication of fertilizers, and insufficient crop rotation have caused lower yields and greater soil degradation, including soil acidification and loss of fertility. Emerging developments in data science and artificial intelligence, specifically in machine learning (ML), have now presented new possibilities for solving these issues. ML gives agriculture a powerful, data-intensive method of decision-making that is able to handle complex, multi-dimensional data and draw meaningful predictions from them. Through the use of ML algorithms can predict agricultural yields based on information like rainfall, humidity, temperature, soil composition, and previous crop yields. precisely and recommend the most appropriate crop for a given field. Not only does this assist in increasing agricultural productivity, but it also encourages sustainable methods of farming. Many supervised learning models like DT, K-NN, RF, & NB have also produced encouraging outcomes for predicting both the crop and the yield it should give. Such models have the ability to revolutionize conventional agriculture by delivering accurate recommendations according to unique environmental and agricultural settings. Combining Internet of Things (IoT) devices improves this capacity by making it possible to receive real-time information from farms, enhancing predictive model responsiveness and accuracy. In this research work, we intend to present a machine learning-driven system that guides farmers in choosing the most suitable crop to be planted and predicts the possible yield on the basis of environmental variables. The intention is to give farmers accurate, useful insights. to make rational decisions, leading to enhanced profitability, minimized resource wastage, and soil conservation for future farm operations.

II. OBJECTIVE OF THE PROPOSED METHOD

The main aim of this paper is to show the practical application of machine learning methods in modern crop production and precision agriculture.

- To examine the effect of changing soil conditions on crop growth and development.

- To forecast the best crops for a specific soil type based on data-driven models.
- To maximize crop yield through natural environmental conditions and predictive analytics.

III. REVIEW OF LITERATURE

Numerous studies have been conducted on crop prediction and crop production increase using innovative machine learning algorithms, according to the literature study. and IoT devices. Among these, Anakha Venugopal et al. [1] came up with a mobile application that can predict the crop type and estimate its yield. The data used in their research included meteorological factors like temperature, wind speed, and humidity, and crop production information. The dataset, however, did not contain any soil-related information, limiting the analysis of soil-based variables. The research used classification models such as RF, NB, and LR, the RF model had the highest accuracy (92.81%), followed by NB (91.50%) and Logistic Regression (87.80%). [30].

Sonal Agarwal et al. [3] also suggested a crop yield forecast model that combines DL & ML methods. The accuracy of the enhanced model was 97%, higher than the current method's 93% accuracy. SVM were used in the machine learning component, and RNN and LSTM networks are used in the deep learning component. However, trade-offs between other hybrid models & their computational intensity are not fully examined in this work..

van Klompenburg et al. [4] (2020) utilized machine learning techniques to anticipate agricultural yields in a systematic literature review (SLR). They stated that models based on neural networks, such as CNN, LSTM, and DNN, are most frequently employed for this job. In some instances, yield prediction depends more on item detection and counting than on conventional tabular data, and the amount of input features employed varies from research to study, they added. In image processing, Hani et al. [5] (2020) examined fruit counting and fruit detection in apple orchards using semi-supervised and DL-based methods. According to their findings, deep learning models like CNN, Faster R-CNN, and U-Net performed worse in yield mapping tasks than conventional methods like Gaussian Mixture Models. Likewise, Koirala et al. [6] (2019) described how deep learning is used in fruit counting and yield estimation, highlighting the methods' resilience in feature extraction and suggesting CNN-based detectors, deep regression models, along with LSTM networks to effectively forecast the fruit load..

Chlingaryan et al. [7] (2018) carried out a survey on the use of machine learning algorithms for estimating nitrogen status and agricultural production prediction. According to their findings, advances in machine learning, especially deep learning, could provide inexpensive and comprehensive solutions. Further, they pointed out the future significance of hybrid systems integrating various ML approaches. P. Mishra et al. [8] employed GBR to enhance maize yield forecasting across various regions in France. The model outperformed AdaBoost, KNN, LR, or RF with an $R^2$of 0.51. However, the study only looked at maize yield in France; it did not look at other crops or areas. A few other studies have also used GBR to improve yield forecast accuracy [9,10,11]. In a related study, V. Latha Jothi et al. [12] have used the data mining model, i.e., KNN, for predictive modeling of crop production for future based on historical parameters like rain, temperature, and water table levels. It also wanted to assist agriculture planning by considering previous and projecting the future trends in groundwater levels. A major limitation of the present study is the difficulty in predicting rainfall, which is a crucial parameter used in crop yield estimation. Similar studies with KNN for crop yield estimation have been done in [13,14,15,16].

IV. METHODOLOGY

A. Dataset Description

The Crop Recommendation Dataset is a publicly accessible dataset on Kaggle. [17] having details of optimal crop selection depending upon soil and climatic factors. The data set has 2,200 observations with 8 independent variables: Temperature, humidity, pH, rainfall, soil phosphorus, nitrogen, and potassium levels. The variable of interest is the crop type to be recommended for cultivation, which has 22 different classes of crops.

After cleaning and pre-processing the dataset using the Interquartile Range (IQR), that is one of the statistical methods for detecting outliers in a dataset. It is obtained by splitting the data into quartiles, calculating the IQR (75th-25th percentile difference) and subsequently forming the upper and lower "fences" to identify values which are outside the usual range, we are left with 1846 rows only which then are being split into two parts randomly using the hold-out technique.

The first section contains 1458 observations, or 79% of the training data. The second component uses 388 observations, or 21% of the dataset, to demonstrate testing.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

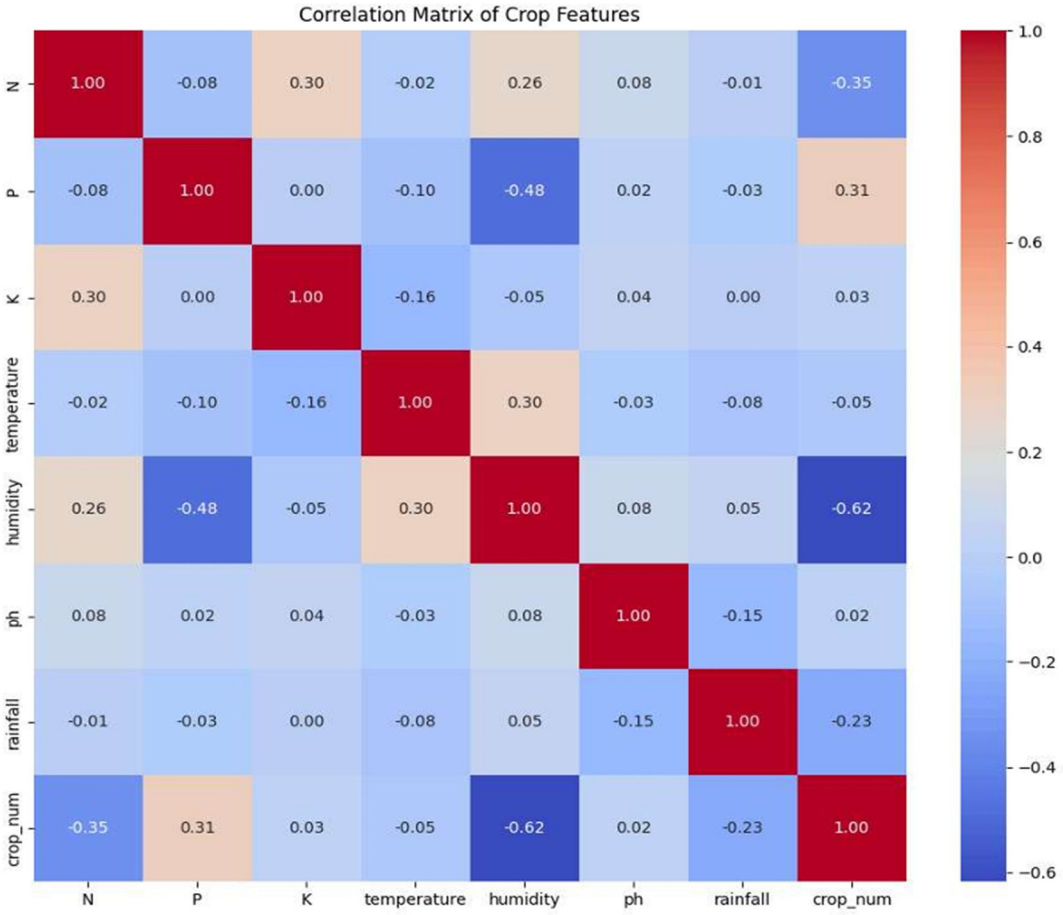


Figure 1: Dataset Correlation matrix

A. Logistic Regression

The statistical technique known as logistic regression is employed to estimate binary results—a yes or a no—based on past experience. It quantifies the influence of one or more independent factors on a dependent factor in an effort to predict [20]. The model predicts the probability of occurrence of a given class and applies the sigmoid function to convert it into a binary value (0 or 1) [21]:

$$f(x) = \frac{1}{1 + e^{-x}}$$

B. Naïve Bayes

Using the premise that every feature is conditionally independent given the class label, NB is a supervised learning classifier based on Bayes' Theorem. While this "naive" assumption is actually made, the algorithm performs remarkably well on various classification problems like spam filtering, document classification, and sentiment detection. Its usage is popular because it is very simple, easy to implement, and efficient on large datasets [22].

$$P(\mathrm{A}|\mathrm{B}) = \frac{P(\mathrm{B}|A) \cdot P(A)}{P(B)}$$

C. Support Vector Machines

SVM is a versatile method for problems involving regression or classification. Its capacity to handle linear as well as non-linear issues makes it a suitable tool for a broad set of real-world applications [24]. The fundamental idea of SVM is to build a decision boundary, i.e., a line or hyperplane, that optimally classifies various classes in the data. With the kernel trick, SVM re-maps the data to a space where it can determine the best boundary and thus is effective at classifying sophisticated problems.

$$F(x, xj) = sum(x.xj)$$

Here, information requiring being classified is represented as x, xj. [19].

D. K-Nearest Neighbor

KNN is an extremely well-used non-parametric classification method. Its basic principle is to order known instances within a feature space, and classifying an instance is achieved with respect to closeness. A new instance is classified by dropping it into the most frequent class among its k nearest instances [21]. Instance closeness is usually computed with the help of distance measures such as the Euclidean distance:

$$D_{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

E. Decision Tree

DT divide the data into smaller subsets according to the most useful criteria in order to forecast a target value. A test or condition is symbolized by every internal node, while the branches stand for it possible outcomes of these tests. The splitting continues recursively until no further improvement is possible or a stopping condition—such as a specified maximum tree depth—is reached [23].

Impurity	Task	Formula	Description
Gini impurity	Classification	$\sum_{i=1}^{C} f_i(1-f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label i at a node and C is the number of unique labels.
Variance / Mean Square Error (MSE)	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$
Variance / Mean Absolute Error (MAE) (Scikit-learn only)	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i-\mu\rvert$	y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N}\sum_{i=1}^{N} y_i$

Figure 2: Impurity Formulas for Scikit-Learn and Spark.

F. Random Forest

RF, another name for this type of ensemble learning technique is random choice forests. It creates many decision trees during the training phase to address various issues, including regression, classification, and others. The number will be returned to the class with the highest selection by the trees during classification. [18].

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

The formula computes the Gini on each branch to find out which branch from a node, class, or probability has a higher chance of happening. In the dataset, c indicates the number on classes, along with pi for the relative frequency in a given class. [19].

G. Bagging

Bagging, or an ensemble learning technique called bootstrap aggregation is primarily used to lower variance in noisy datasets. Interestingly, this technique is expanded by the RF algorithm, which creates a collection of highly uncorrelated DT by combining bagging with feature randomness. [25].

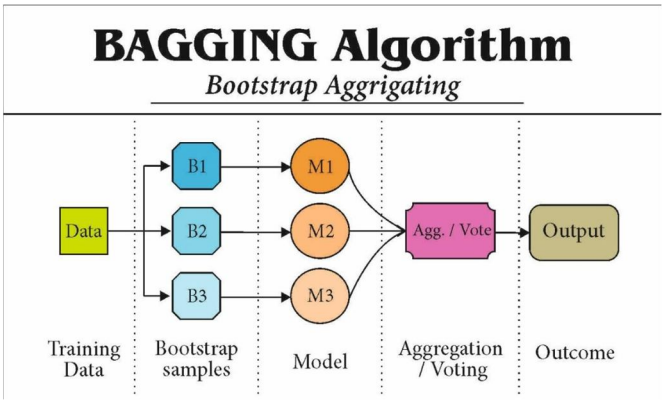


Figure 3: Bagging Model Algorithm

H. Gradient Boosting

GB is a method for ensemble learning that builds a series of DT, with each tree being built to fix mistakes made by earlier ones. While AdaBoost employs shallow trees as base learners, Gradient Boosting employs deeper trees as base learners. Instead of learning directly from the initial output values, each tree is learned from the residual errors—differences between predicted and true values [26,27].

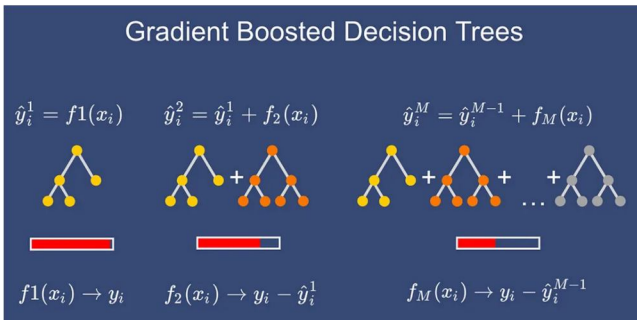


Figure 4: Gradient Boosting Algorithm

I. Extra Trees

An ensemble learning technique called Extremely Randomized Trees, or Extra Trees, creates several decision trees in a manner similar to Random Forest. However, it introduces greater randomness by selecting split points entirely at random, rather than determining the best split at every node. Additionally, unlike Random Forest which uses bootstrap sampling, Extra Trees generally trains each tree on the entire dataset, relying on randomly chosen feature splits to maintain model diversity.. [28].

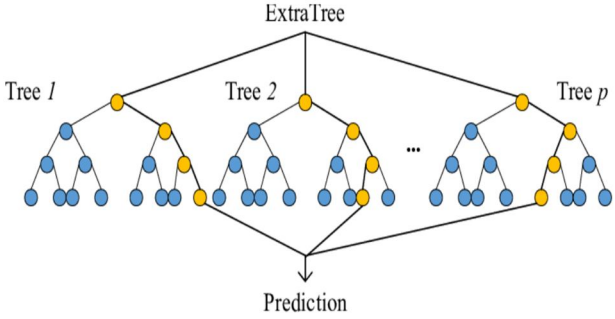


Figure 5: Extra tree Algorithm

V. RESULTS & CONCLUSION

In this study, the efficiency of many machine learning algorithms in identifying the best crop is compared given environmental and soil conditions. The models used for comparison are LR, NB, SVM, KNN, DT, RF, BG, GB, & ET.

As the first objective, we examined which environmental variables contribute most to accurately predicting the crop. According to the ranking of feature importance, rainfall along with humidity were found to be the leading predictors. Nutrient levels in the soil, i.e., Potassium (K), Phosphorus (P), and Nitrogen (N), were also found to be strongly contributing, reflecting that crop yield and compatibility are largely determined by weather as well as soil health. Temperature and pH were relatively weaker in their contribution but still important for a complete picture of compatibility between soil and crops.

Table 1: Feature Importance based on ML model

Feature	**Importance**
Rainfall	0.223189
Humidity	0.217193
K (Potassium)	0.171237
P (Phosphorus)	0.125898
N (Nitrogen)	0.119951
Temperature	0.083575
pH	0.058958

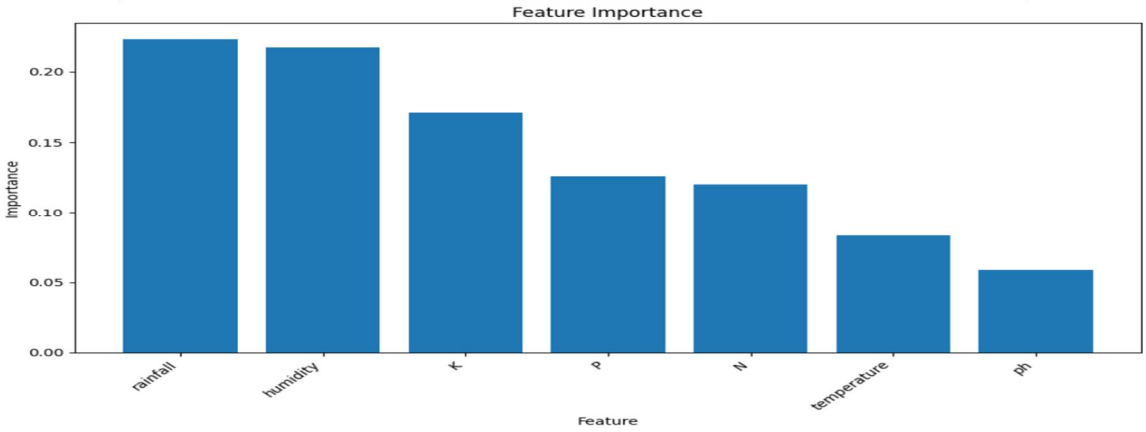


Figure 6: Feature importance based on ML analysis

The second objective was to analyze the precision of many models for ML in crop recommendation. Out of all the models, Naive Bayes and Random Forest worked most efficiently with 99.74% and 99.48% accuracy, respectively. These were then followed by Gradient Boosting and Bagging with 98.96% accuracy. However, the Extra Trees model, while still relatively effective, had the lowest accuracy (90.72%), possibly due to over-randomization or less effective feature splits.

Table 2: Accuracy of different ML algorithms

Model Name	Accuracy
Logistic Regression	0.9794
Naive Bayes	0.9974
Support Vector Machine	0.9845
K-Nearest Neighbors	0.9716
Decision Tree	0.9923
Random Forest	0.9948
Bagging	0.9897
Gradient Boosting	0.9897
Extra Trees	0.9072

These results validate the effectiveness of probabilistic (Naive Bayes) along ensemble tree-based models (RF, GB, BG) in agricultural sector, particularly in making crop suggestions against key environmental and soil characteristics.

This paper compares different ML algorithms for endorsing crops according to climatic and soil characteristics. Among all the tested models, Naive Bayes and Random Forest showed the greatest accuracy, with 99.74% for Naive Bayes and 99.48% for Random Forest, ascertaining their great potential for crop prediction tasks. Feature importance analysis identified rainfall, humidity, and potassium content as the most impactful factors influencing the suitability of crops. These results indicate that employing correct classifiers combined with important environmental characteristics can improve decision-making in precision agriculture substantially.

VI. FUTURE ENHANCEMENT

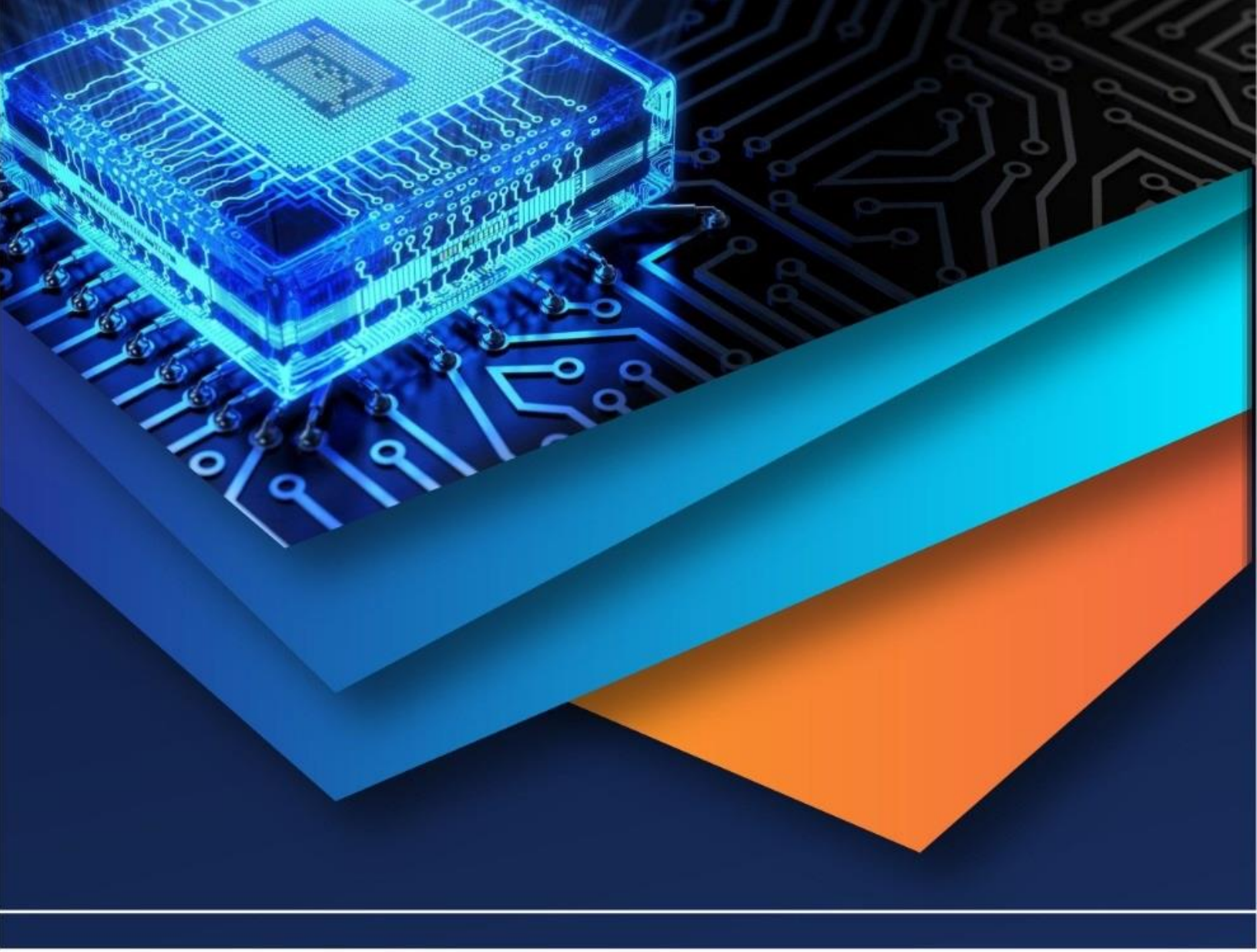
1) IoT Integration: Use IoT sensors for data collection (e.g., moisture, pH, temperature) to facilitate dynamic crop suggestions.
2) Increased Dataset Coverage: Incorporate varied datasets from various geographies for improved model generalization.
3) Multi-Crop Yield Prediction: Forecast yields for multiple crops at once to facilitate informed decision-making.
4) Climate Adaptability: Integrate models considering climate change trends for a sustainable crop plan.
5) Mobile Application Rollout: Build an app in local languages so that farmers have easy access to it.
6) Market-Based Suggestions: Incorporate prices and demand patterns of crops to provide economically optimum crop recommendations.

This will further aid the Crop production in providing higher yields, and thereby improving boosting farming.

REFERENCES

[1] Nischitha, K., Vishwakarma, D., Mahendra, N., Ashwini & Manjuraju, M.R., 2020. Crop Prediction using Machine Learning Approaches. International Journal of Engineering Research & Technology (IJERT), 9(08). Available at: http://www.ijert.org [Accessed 6 April 2025].
[2] Venugopal, A., Aparna, S., Mani, J., Mathew, R. & Williams, V., 2021. Crop Yield Prediction using Machine Learning Algorithms. IJERT 2021, 9. Available online: https://ieeexplore.ieee.org/abstract/document/8985951 [Accessed on 19 April 2025].
[3] Agarwal S., and Tarar, S., 2021. A hybrid approach for crop yield prediction using machine learning and deep learning algorithms. In Journal of Physics: Conference Series (Vol. 1714, No. 1, p. 012012). IOP Publishing.
[4] van Klompenburg, T., Kassahun, A. and Catal, C., 2020. Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, p.105709. https://doi.org/10.1016/j.compag.2020.105709
[5] Hani, N., Roy, P. and Isler, V., 2020. A comparative study of fruit detection and counting methods for yield mapping in apple orchards. Journal of Field Robotics, 37(2), pp.263–282. https://doi.org/10.1002/rob.21902
[6] Koirala, A., Walsh, K.B., Wang, Z. and McCarthy, C., 2019. Deep learning – Method overview and review of use for fruit detection and yield estimation. Computers and Electronics in Agriculture, 162, pp.219–234.
[7] Chlingaryan, A., Sukkarieh, S. and Whelan, B., 2018. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, pp.61–69. https://doi.org/10.1016/j.compag.2018.05.012
[8] Mishra, P., Khan, R. and Baranidharan, D.B., 2020. Crop yield prediction using gradient boosting regression. International Journal of Innovative Technology and Exploring Engineering, 9, pp.2293–2297.
[9] Lamos-Díaz, H., Puentes-Garzón, D.E. and Zarate-Caicedo, D.A., 2019. Comparison between machine learning models for yield forecast in cocoa crops in Santander, Colombia. Revista Facultad de Ingeniería, 29, p. e10853.
[10] Pradeep, G., Rayen, T.D.V., Pushpalatha, A. and Rani, P.K., 2023. Effective crop yield prediction using gradient boosting to improve agricultural outcomes. In: Proceedings of the 2023 International Conference on Networking and Communications (ICNWC), Chennai, India, 5–6 April 2023, pp.1–6.
[11] Yasaswy, M.K., Manimegalai, T. and Somasundaram, J., 2022. Crop yield prediction in agriculture using gradient boosting algorithm compared with random forest. In: Proceedings of the 2022 International Conference on Cyber Resilience (ICCR), Dubai, UAE, 6–7 October 2022, pp.1–4.
[12] Jothi, V.L., Neelambigai, A., Sabari, N.S. and Santhosh, K., 2020. Crop yield prediction using KNN model. International Journal of Engineering Research & Technology (IJERT), 8.
[13] Pavani, S. and Beulet, P.A.S., 2022. Prediction of Jowar crop yield using K-Nearest Neighbor and Support Vector Machine algorithms. In: Proceedings of the International Conference on Futuristic Communication and Network Technologies, Niagara Falls, ON, Canada, 9–11 August 2022.
[14] Sundari, M., Rekha, G., Krishna, V.S.R., Naveen, S. and Bharathi, G., 2023. Crop recommendation system using K-Nearest Neighbors algorithm. In: Proceedings of the 6th International Conference on Recent Trends in Computing, Chennai Campus, India, 14–15 December 2023, pp.581–589.
[15] Karn, R.K. and Suresh, A., 2023. Prediction of crops based on a machine learning algorithm. In: Proceedings of the 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 23–25 January 2023, pp.1–8.
[16] Cheong, L.R.N., Kwong, K.F.N.K. and Du Preez, C.C., 2009. Effects of sugar cane (Saccharum hybrid sp.) cropping on soil acidity and exchangeable base status in Mauritius. South African Journal of Plant and Soil, 26, pp.9–17.
[17] Atharva Ingle, 2021. Crop Recommendation Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset [Accessed 12 March 2025].
[18] M. Schott, "Random Forest Algorithm for Machine Learning," Capital One Tech, Apr. 25, 2019. [Online]. Available: https://medium.com/capital-one-tech/randomforest-algorithm-for-machine-learning-c4b2c8cc9feb. [Accessed: Jul. 5, 2024].

[19] Alamma, B.H., Sharma, D., Chithra, H.N., Bhat, S., Suhana, B.V.A., Raj, A. and Ashok, G., 2024. Enhancing Lung Cancer Early Detection: A Hybrid Ensemble Model. Journal of Electrical Systems, 20(Special Issue 10), pp.01–06.
[20] Chakrabarty, N., Chowdhury, S. & Rana, S., 2020. A statistical approach to graduate admissions. Chance Prediction, pp.145–154.
[21] Sharma, D., 2023. Machine Learning Classifiers for Breast Cancer Diagnosis. International Journal of Engineering Research & Technology (IJERT), NCRTCA - 2023. Available at: https://www.ijert.org/research/NCRTCA-PID-098.pdf [Accessed 14 April 2024].
[22] GeeksforGeeks. (n.d.). Naive Bayes Classifiers. [online] Available at: https://www.geeksforgeeks.org/naive-bayes-classifiers/ [Accessed 15 February 2025].
[23] S. Ronaghan, "The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark," Towards Data Science, May 12, 2018. [Online]. Available: https://towardsdatascience.com/the-mathematicsof-decision-trees-random-forest-and-feature-importance-inscikit-learn-and-spark-f2861df67e3# . [Accessed: 5 July 2024].
[24] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning,46(1-3),389-422.
[25] IBM, n.d. Bagging. [online] Available at: https://www.ibm.com/think/topics/bagging [Accessed 2 May 2025].
[26] Baladram, S., 2024. Gradient Boosting Regressor, Explained: A Visual Guide with Code Examples. [online] Medium. Available at:https://medium.com/data-science/gradient-boosting-regressor-explained-a-visual-guide-with-code-examples-c098d1ae425c [Accessed 3 May 2025].
[27] Singh, A., 2025. Gradient Boosting Explained: Turning Weak Models into Winners. [online] Medium. Available at: https://medium.com/@abhaysingh71711/gradient-boosting-explained-turning-weak-models-into-winners-c5d145dca9ab [Accessed 6 May 2025].
[28] Baladram, S., 2024. *Extra Trees, explained: A visual guide with code examples – Setting Random Forest ablaze with more randomness*. [online] Medium. Available at: https://medium.com/@samybaladram/extra-trees-explained-a-visual-guide-with-code-examples-4c2967cedc75 [Accessed 4 May 2025].
[29] Chu, Z., Yu, J. and Hamdulla, A., 2020. *Throughput prediction based on ExtraTree for stream processing tasks*. [online] Available at: https://www.researchgate.net/figure/The-structure-of-ExtraTree_fig1_346995264 [Accessed 1 May 2025].
[30] Uppugunduri, V.N., Pandiyan, A.M., Raja, S.P. and Stamenkovic, Z., 2024. Machine learning-based crop yield prediction in South India: Performance analysis of various models. Computers, 13(6), p.137. https://doi.org/10.3390/computers13060137



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY