Prediction of Crop Recommendation Using Machine Learning Method

Rohit Kumar Pradhan
Email ID: rkp2023@gift.edu.in

Dr. Satya Ranjan Pattanaik
Email ID: drsatyaranjan@gift.edu.in

Abstract- The proposed crop recommendation system leverages machine learning techniques to support farmers in making informed crop selection decisions. It analyzes essential parameters such as soil pH, nutrient content, rainfall, and climate conditions to determine the most suitable crops for cultivation in a given region. By processing this data, the system improves agricultural planning, promotes water-efficient practices, and helps reduce the risk of crop failure caused by climatic uncertainty. This research assesses various machine learning algorithms and performance metrics, comparing different crop recommendation approaches using a publicly available dataset from www.kaggle.com. In this study, multiple machine learning algorithms were compared using a publicly available dataset containing 22 crop types. Among the tested models, the Random Forest classifier demonstrated the highest prediction accuracy of 99.31%, making it a promising tool for practical agricultural applications.

Keywords- Crop Recommendation, Ensemble Learning, Decision Tree, Machine Learning, Random Forest classifier.

1. INTRODUCTION

Agriculture is a cornerstone of economic growth and food security, particularly in countries like India, where a significant portion of the population depends on farming. Despite its importance, many farmers face challenges when deciding which crops to grow, as they must consider factors such as soil health, weather conditions, and resource availability. [1] These decisions are often made without the aid of reliable data or advanced tools, leading to inconsistent yields and inefficient practices. With recent advancements in technology, machine learning (ML) has emerged as a valuable solution for improving agricultural decision-making. By analyzing historical data related to soil nutrients, rainfall, temperature, and previous crop yields, ML models can identify patterns that help in recommending the most appropriate crops for specific regions. [2] Among various algorithms, the Random Forest Classifier has shown strong performance in handling complex agricultural datasets. Its ability to combine multiple decision trees enhances the model's accuracy and reduces overfitting. This study explores the use of different machine learning algorithms for crop recommendation and evaluates their effectiveness using real-world agricultural data. The goal is to build a system that empowers farmers with data-driven insights to optimize their crop planning and improve sustainability in agriculture.

1.1 Machine learning methods

Computers can emulate human-like learning and behavior through machine learning algorithms, which can be enhanced by supplying data and knowledge from observations. Machine learning, a subset of artificial intelligence, allows computers to learn from data to improve their performance without explicit coding. This process involves algorithms that detect patterns, make predictions, and automate various tasks. In supervised learning, labeled data is utilized, whereas unsupervised learning seeks to uncover hidden patterns in unlabeled data. Numerous machine learning methods exist, and this paper focuses on using the random forest classifier model for analysis.

1.1.1 Logistic Regression

Logistic Regression is a simple yet effective algorithm, especially useful in binary classification tasks. In the context of crop recommendation, it helps decide whether a particular crop is suitable based on various input features like soil nutrients, moisture, and temperature. The model works by mapping input features to probabilities using a logistic function. Prior to training, the data undergoes preprocessing such as normalization and handling of missing values.

1.1.2 Support Vector Machine

Support Vector Machines are powerful tools for classification tasks. In crop prediction, SVM helps classify crops by finding the optimal boundary that separates data points of different crop types based on features like soil pH, climate, and land type. It is especially effective in dealing with non-linear decision boundaries using kernel functions. This model will predict which crops are most suitable to grow in that particular context. [3]

1.1.3 K-Nearest Neighbors

KNN classifies a new data point based on the majority class among its closest neighbors in the feature space. It is intuitive and works well for datasets where similar conditions (e.g., soil type, weather) yield similar crop outcomes. The distance between points is often calculated using metrics such as Euclidean or Manhattan distance.

1.1.4 Decision Tree

Decision Trees break down a dataset into smaller subsets while developing an associated tree structure. Each node represents a decision based on an attribute (like rainfall or temperature), and the final leaf nodes represent crop recommendations. This model is easy to understand and interpret.

1.1.5 Random Forest

Random Forest is an ensemble learning method that operates by constructing multiple decision trees and combining their predictions to achieve a more accurate and stable outcome. Each tree in the Random Forest is built using a randomly selected subset of the training data and a randomly chosen subset of features at each decision node. By introducing randomness in both data sampling and feature selection, Random Forest helps mitigate overfitting and enhances generalization. This approach fosters diversity among the trees, which contributes to the model's robustness and ability to generalize effectively.

1.1.6 Bagging

Bagging, short for bootstrap aggregating, is a widely utilized ensemble learning technique in ML that enhances model accuracy.

By training multiple versions of a model on different subsets of the data and then aggregating the results, bagging reduces variance. This is particularly useful in stabilizing weak learners like decision trees. Random Forest and Random Subspace are upgraded versions of decision trees that use the bagging approach to improve the pre- dictions of the decision tree's base classifier.

1.1.7 AdaBoost

AdaBoost, short for Adaptive Boosting, is an ensemble learning method designed to improve the performance of weak classifiers. By sequentially combining multiple weak classifiers, it constructs a robust classifier. Subsequent models place greater emphasis on instances that prior models misclassified. This approach enhances overall prediction accuracy by concentrating on challenging cases.

1.1.8 Gradient Boosting

Gradient Boosting is a robust ensemble learning methodology, extensively utilized in machine learning for classification and regression tasks. This method constructs models in a stage-wise fashion, where each subsequent model is trained to rectify the errors of its predecessor by minimizing a specified loss function. Consequently, each tree is trained to correct the residual errors of its predecessor, thereby incrementally enhancing the model's overall predictive accuracy. The core principle of Gradient Boosting involves iteratively optimizing a loss function by minimizing residual errors, enabling the model to progressively improve its performance.

1.1.9 Extra Trees

Extra Trees is an ensemble learning method that enhances accuracy and robustness by constructing multiple decision trees. This method reduces variance and bolsters model robustness. In Extra Trees, each tree contributes to the final prediction through majority voting for classification tasks or averaging for regression tasks, resulting in an effective ensemble model. A potential trade-off for this added randomness is a decrease in accuracy for datasets characterized by distinct patterns. Nevertheless, Extra Trees can be a highly efficient and robust approach for managing large and high-dimensional datasets, rendering it a versatile tool across healthcare, finance, and marketing domains.

2. LITERATURE SURVEY

The system uses Decision Tree, and KNN to predict suitable crops based on soil and environmental factors, trained on datasets of soil properties [1]. The author [2] utilizes SVM and Decision Tree algorithms to recommend crops based on soil nutrients, pH, rainfall, humidity, and temperature. [6] The study by Prabhu et al. (2020) presents a soil analysis and crop prediction model that evaluates soil fertility using parameters like NPK and rainfall to recommend suitable crops. It compares classification algorithms such as SOM and K-means, highlighting SOM's superior accuracy for soil type classification. The study [3] used Pearson correlation to select relevant atmospheric features and applied three machine learning models MLR, Random Forest and XGBoost. The study [5] proposes a cloud-based crop recommendation platform that leverages machine learning algorithms to support precision farming through real-time, data-driven decisions. It compares five ML models—KNN, DT, RF, XGBoost, and SVM—to identify the most effective approach for optimizing crop selection. The author [4] utilizes Machine Learning algorithms for classification and prediction and Big Data Analytics to process environmental and soil data.

Kaggle Dataset: Crop Recommendation Dataset Available at: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

3. METHODOLOGY

This research aims to develop a crop recommendation system by leveraging various machine learning algorithms. The proposed system includes data acquisition and preprocessing, and crop prediction.

3.1 DATA COLLECTION

This study utilizes a dataset obtained from the Kaggle archives, curated by the Food and Agriculture Council of India, consisting of 2,200 data points across 22 different crops, namely Rice, Maize, Jute, Cotton, Coconut, Papaya, Orange, Apple, Muskmelon, Watermelon, Grapes, Mango, Banana, Pomegranate, Lentil, Blackgram, Mungbean, Mothbeans, Pigeon Peas, Kidney Beans, Chickpea, and Coffee.

The dataset includes variables related to Nitrogen, Phosphorus, Potassium, fertilizers, soil pH, and climatic factors such as rainfall, temperature, and humidity. [4] Initially, the data undergoes importation, followed by a preliminary evaluation to detect null or duplicate entries. Subsequently, each crop is labeled using one-hot encoding and organized into a dictionary. Following this, the Data Distribution Testing and Scaling function is trained using MinMaxScaler, and the Data Training Model is implemented. The dataset exhibits high quality due to its incorporation of diverse geographical conditions and crops, underscoring its potential applicability across regions worldwide with similar environmental conditions.

A detailed description of the dataset and proposed system employed in this study is presented in Table 3.1. and Fig 3.1.

The dataset encompasses information regarding various attributes pertinent to agricultural conditions. The description of each attribute is outlined below:

N: This attribute exhibits a range of (0–139) kg/ha, indicative of the quantity of nitrogen in the soil, measured in kilograms per hectare.

P: This attribute ranges from (5–145) kg/ha, representing the amount of phosphorus in the soil, measured in kilograms per hectare.

K: With a range of (5–205) kg/ha, this attribute denotes the quantity of potassium in the soil, measured similarly to N and P.

Temperature: Ranging from (10.78–43.36) K, this attribute is provided in Kelvin value, reflecting the temperature conditions.

Humidity: With a range of (14.69–98.80) F, this attribute can be expressed in Fahrenheit or Celsius, indicating humidity.

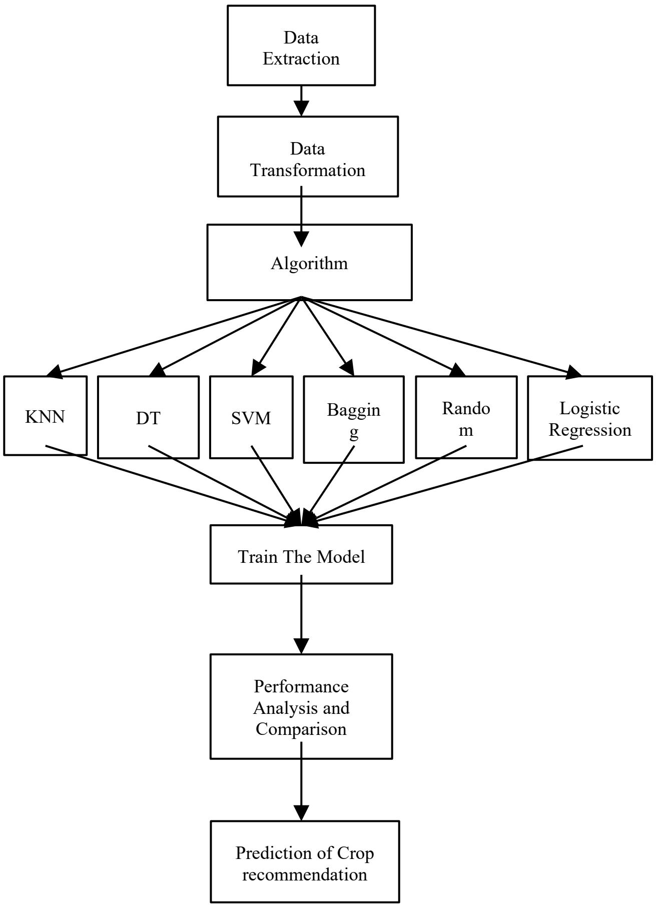
pH: Spanning from (3.55–7.45), the pH attribute typically operates on a scale from 0 to 14, measuring the acidity or alkalinity of a substance, thus reflecting soil conditions.

Rainfall: This attribute signifies the volume of rainfall in millimeter. exhibiting a range of (20.21–291.29) mm, providing insight into precipitation levels.

Table 3.1 Dataset Description

Attributes	Range
N	(0–139) kg/ha
P	(5–145) kg/ha
k	(5–205) kg/ha
Temperature	(10.78–43.36) K
Humidity	(14.69–98.80) F
Ph	(3.55–7.45)
Rainfall	(20.21–291.29) mm

Fig.3.1. Proposed System



3.2 Training and Testing

Handling the challenge of imbalanced datasets is a critical aspect of developing reliable machine learning models, as an uneven distribution of class labels can lead to biased outcomes, particularly against minority classes. To address this, a down-sampling technique was implemented to reduce the dominance of the majority class and achieve a more balanced class representation. Furthermore, to enhance the model's generalization capabilities and avoid overfitting, early stopping was incorporated during the training process, allowing the model to halt once performance ceased to improve on validation data. Several machine learning algorithms were then applied to accurately predict the most suitable crop cultivation strategy based on the input features. For performance assessment, the dataset was divided into training and testing subsets, with 80% allocated for model learning and the remaining 20% reserved for validation. This split enabled a thorough evaluation of the model's ability to generalize and make accurate predictions on unseen data.

3.3 Performance Metrics

The assessment was carried out using multiple evaluation metrics to ensure a thorough comparison of model effectiveness. To measure model performance, standard classification metrics were employed, including accuracy, precision, recall, F1-score, as well as Confusion Matrix and ROC (Receiver Operating Characteristic) Curve analysis.

- **Accuracy** indicates how often the model predicts the correct crop across all test cases.
- **Precision** reflects the proportion of correctly predicted positive instances out of all predicted positives, helping assess how reliable the model is when it suggests a crop.
- **Recall** measures the proportion of actual positives that were correctly identified by the model, which is crucial for detecting all suitable crop options.
- **F1-score** balances precision and recall, providing a more comprehensive view of the model's effectiveness, especially in cases where class distribution is uneven.
- **Confusion Matrix** offers a visual breakdown of correct and incorrect predictions, showing true positives, false positives, false negatives, and true negatives for each class.
- **ROC Curve** helps visualize the trade-off between the true positive rate and false positive rate at various classification thresholds.

The study trained and evaluated nine different ML algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Bagging (BG), AdaBoost (AB), Gradient Boosting (GB), and Extra Trees (ET). The dataset was split, with 80% used for training and 20% for testing. Among the models tested, Random Forest achieved the highest accuracy of 99.31%, making it the most reliable choice for crop prediction in this experiment. [5]

4. RESULT

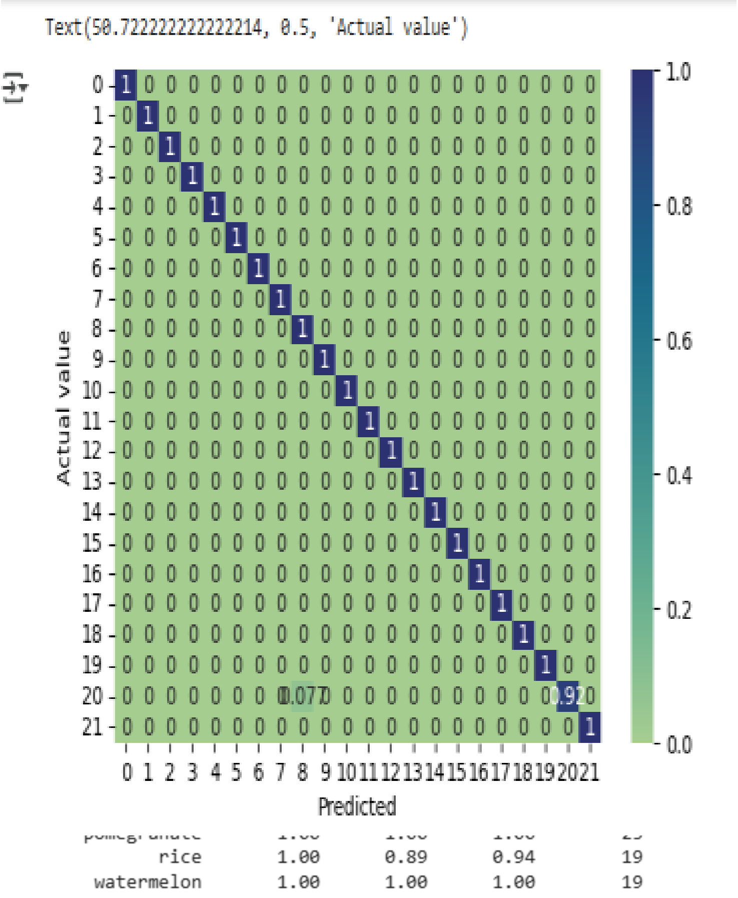
This section provides an in-depth analysis of the performance metrics for all Machine Learning algorithms used in the proposed crop recommendation system. It includes detailed evaluations of accuracy, precision, recall, F1 score, and, confusion matrix analysis. These results enable a comprehensive assessment of each algorithm’s effectiveness and appropriateness for the task. The study aimed to recommend crops based on multiple factors, employing nine ML algorithms, including LR, SVM, KNN, DT, RF, BG, AB, GB, and ET. [6], [7] Among the models tested, **Random Forest achieved the highest accuracy of 99.31%**, making it the most reliable choice for crop prediction in this experiment. Other models also performed well, with SVM and Bagging models demonstrating strong consistency, while AdaBoost showed relatively lower accuracy. The performance of each algorithm is summarized in the following comparison table. These models underwent training and optimization with specific parameters outlined in the methodology section. A comparative analysis of all models is presented in Table 4.3. Here Random forest classifier gives the confusion matrix in Fig 4.3 and F1 score of Random Forest model is presented below in fig 4.4.

Table 4.3 Comparative Analysis of ML Algorithms

Name of classifiers	Precision (%)	Recall (%)	F1-score (%)	Accuracy (%)
LR	88	87	88	96.36
SVM	88	88	88	96.82
KNN	68	69	68	96
DT	72	72	72	98.2
RF	92	93	92	99.31
BG	92	92	92	98.9
AB	78	78	76	14.1
GB	85	83	83	98.2
ET	96	96	96	91

Fig 4.3 Confusion matrix of Random Forest classifier

Fig 4.4 F1 score of RF Model



```
Text(50.72222222222214, 0.5, 'Actual value')
```

```
 pomegranate       1.00      1.00      1.00        2
        rice       1.00      0.89      0.94       19
  watermelon       1.00      1.00      1.00       19
```

5. CONCLUSION

Agriculture plays a pivotal role in the global economy, providing numerous employment opportunities and contributing significantly to the GDP. It is essential for supporting livelihoods and ensuring food production on a global scale. However, optimizing crop yields remains a persistent challenge for farmers due to climate unpredictability, soil heterogeneity, and limited access to data-driven technologies. This study aimed to mitigate these issues by developing a machine learning-based crop recommendation system. [8][9]. By analyzing critical agricultural factors, such as temperature, rainfall, humidity, soil pH, and nutrient composition, the system offers personalized crop suggestions tailored to local environmental conditions. Nine machine learning models were implemented and rigorously evaluated, including Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Decision Tree, Random Forest, Bagging, AdaBoost, Gradient Boosting, and Extra Trees. Among these, the Random Forest classifier demonstrated the highest accuracy, achieving a prediction accuracy of 99.31%. The ensemble approach of Random Forest proved highly effective in managing diverse datasets and producing reliable results. This system has the potential to assist farmers in making well-informed decisions, enhancing productivity, and fostering sustainable agricultural practices. Despite the model's strong performance, some limitations persist.

6. FUTURE SCOPE

The future of crop recommendation systems is poised for remarkable advancements with the integration of machine learning models. Among the models evaluated in this study, the Random Forest algorithm demonstrated superior performance with an accuracy of 99.31%. Its robustness in handling various types of agricultural data including soil nutrients, weather conditions, and past crop performance makes it an excellent foundation for future enhancements. Looking ahead, integrating Internet of Things (IoT) devices, remote sensing, and real-time weather data can further refine the prediction process. [10]These advancements would allow the system to adjust recommendations dynamically based on current environmental conditions. Additionally, the use of deep learning models and hyper parameter tuning could improve precision for more complex datasets. Cloud based platforms may also play a key role by offering farmers easy access to crop insights through mobile apps or web dashboards. Moreover, incorporating local farmer knowledge and feedback into the system could make the recommendations more practical and region-specific. Overall, with continued research and innovation, smart crop recommendation tools can greatly enhance agricultural productivity, reduce input waste, and support more climate-resilient farming worldwide. Future improvements may involve integrating real-time data streams, expanding the feature set, and incorporating expert insights to enhance the system's adaptability and precision.

REFERENCES

[1] Wang SW, Lee WK, Son Y. An assessment of climate change impacts and adaptation in South Asian agriculture. Int J Clim Chang Strateg Manag. 2017;9:517–34. https://doi.org/10.1108/ IJCCSM-05-2016-0069.

[2] Bouguettaya A, Zarzour H, Kechida A, Taberkit AM. Deep learning techniques to classify agricultural crops through UAV imagery: a review. Neural Computer Appl. 2022;34:9511–36. https://doi.org/10.1007/S00521-022-07104-9.20223412.

[3] Gathala MK, Timsina J, Islam MS, Rahman MM, Hossain MI, Harun-Ar-Rashid M, McDonald A. Conservation agriculture based tillage and crop establishment options can maintain farmers' yields and increase profits in South Asia's rice-maize systems: evidence from Bangladesh. Field Crops Res. 2015;172:85–98.

[4] Fei S, Hassan MA, Xiao Y, Su X, Chen Z, Cheng Q, Duan F, Chen R, Ma Y. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. Precis Agric. 2022. https://doi.org/10.1007/s11119-022-09938-8.

[5] Paudel D, Boogaard H, de Wit A, van der Velde M, Claverie M, Nisini L, Janssen S, Osinga S, Athanasiadis IN. Machine learning for regional crop yield forecasting in Europe. Field Crops Res. 2022. https://doi.org/10.1016/J.FCR.2021.108377.

[6] Burdett H, Wellen C. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precis Agric. 2022;23:1553–74. https://doi.org/10.1007/ s11119-022-09897-0.

[7] Duke OP, Alabi T, Neeti N, Adewopo J. Comparison of UAV and SAR performance for Crop type classification using machine learning algorithms: a case study of humid forest ecology experimental research site of West Africa. Int J Remote Sens.2022;43:4259–86. https://doi.org/10.1080/01431161.2022.2109444.

[8] Too EC, Yujian L, Njuki S, Yingchun L. A comparative study of fine-tuning deep learning models for plant disease identification. Compute Electron Agric. 2019;161:272–9. https://doi.org/10. 1016/j.compag.2018.03.032.

[9] Chu H, Zhang C, Wang M, Gouda M, Wei X, He Y, Liu Y. Hyper- spectral imaging with shallow convolutional neural networks (SCNN) predicts the early herbicide stress in wheat cultivars. J Hazard Mater. 2022. https://doi.org/10.1016/j.jhazmat.2021. 126706.

[10]Dey B, Masum Ul Haque M, Khatun R, Ahmed R. Comparative performance of four CNN-based deep learning variants in detecting His pest, two fungal diseases, and NPK deficiency symptoms of rice (Oryza sativa) Comput. Electron Agric. 2022. https://doi.org/10.1016/j.compag.2022.107340.

[11]Sai Sankar PR, Ramakrishna SDPS, Venkata Rakesh MM, Raja P, Hoang VT, Szczepanski C. Intelligent health assessment system for paddy crop using CNN, 2021 3rd. Int Conf Signal Process Commun. ICPSC. 2021;2021:382–7. https://doi.org/10.1109/ ICSPC51351.2021.9451644.

[12]Z. Doshi S, Nadkarni R, Agrawal N. Shah, Agro-consultant: intelligent crop recommendation system using machine learning algorithms. In: Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, 2018, pp. 1–6. https://doi.org/10.1109/ICCUBEA.2018.8697349.

[13]ChouguleVKA, Mukhopadhyay D. Crop suitability and fertilizers recommendation using data mining techniques, in: Advances in Intelligent Systems and Computing, Vol. 714, Springer Verlag, 2019, pp. 205–213. https://doi.org/10.1007/978-981-13-0224-419.

[14]Kulkarni NH, Srinivasan GN, Sagar BM, Cauvery NK. Improving crop productivity through A crop recommendation system using ensembling technique. Proc 2018 3rd Int Conf Comput Syst Inf Technol Sustain Solut CSITSS. 2018. https://doi.org/10.1109/ CSITSS.2018.8768790.

[15]Modi D, Sutagundar AV, Yalavigi V, Aravatagimath A. Crop recommendation using machine learning algorithm. 2021 5th Int Conf Inf Syst Comput Networks ISCON. 2021.

[16]Prabhu S, Revandekar P, Shirdhankar S, Paygude S. Soil analysis and crop prediction. Int J Sci Res Sci Technol. 2020;7(4):117–23.

[17]Gosai D, Raval C, Nayak R, Jayswal H, Patel A. Crop recommendation system using machine learning. Int J Sci Res Comput Sci Eng Inf Technol. 2021;7(3):558–69.

[18]Viviliya B, Vaidhehi V. The design of hybrid crop recommenda- tion system using machine learning algorithms. Int J Innov Tech- nol Explor Eng (IJITEE). 2019;9:4305–11.

[19]Abrougui K, Gabsi K, Mercatoris B, Khemis C, Amami R, Che- haibi S. Prediction of organic potato yield using tillage systems and soil properties by artificial neural network (ANN) and multiple linear regressions (MLR). Soil Tillage Res. 2019;190:202–8.

[20]Villanueva MB, Salenga MLM. Bitter melon crop yield prediction using machine learning algorithm. Int J Adv Comput Sci Appl. 2018;9:1–6.

[21] Kaggle. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset. Accessed date: 12 January 2024.

[22]Suresh G, Kumar DAS, Lekashri DS, Manikandan DR, Head C-O. Efficient crop yield recommendation system using machine learn- ing for digital farming. Int J Mod Agric. 2021;10:906–14.

[23]Hua Y, Li F, Yang S. Application of support vector machine model based on machine learning in art teaching. Wireless Communication Mobile Computing 2022. https://doi.org/10.1155/2022/7954589.

[24]Dey B, Abir KAM, Ahmed R, Salam MA, Redowan M, Miah MD, Iqbal MA. Monitoring groundwater potential dynamics of north-eastern Bengal Basin in Bangladesh using AHP-Machine learning approaches. Ecol Indicat. 2023. https://doi.org/10.1016/j.ecolind.2023.110886.

[25]Charoen-Ung P, Mittrapiyanuruk P. Sugarcane yield grade pre- diction using random forest with forward feature selection and hyper-parameter tuning. Adv Intell Syst Comput. 2019;769:33-42. https://doi.org/10.1007/978-3-319-93692-5_4.

[26]Sridevi V, Chellamuthu V. Impact of weather on rice—a review. Int J Appl Res. 2015;1:825–31.

[27]Swathi T, Sudha S. Crop classification and prediction based on soil nutrition using machine learning methods. Int J Inf Technol. 2023;15:2951–60. https://doi.org/10.1007/S41870-023-01345-0.

[28]Singh Jatav Sri Karan H, Jena J, Maitra S, Hossain A, Pramanick B, Gitari HI, Praharaj S, Shankar T, Bharati Palai J, Rathore A, Kumar Mandal T, Singh Jatav H. Role of legumes in cropping system for soil ecosystem improvement improvement of cropping system view project precision agriculture view project role of legumes in cropping system for soil ecosystem improvement. Nova Science Publishers, Inc. 2022.

[29]Thilakarathne NN, Bakar MSA, Abas PE, Yassin H. A cloud enabled crop recommendation platform for machine learning-driven precision farming. Sensors. 2022. https://doi.org/10.3390/s2216 6299.

[30]JainS, Ramesh D. Machine Learning convergence for weather- based crop selection, IEEE International Students' Conference on Electrical, Electronics and Computer Science (2020). https:// doi.org/10.1109/SCEECS48394.2020.75.

.