Plant Archives Vol. 26, Supplement 1, 2026 pp. 1930-1937 e-ISSN:2581-6063 (online), ISSN:0972-5210

Plant Archives

Journal homepage: http://www.plantarchives.org
DOI Url : https://doi.org/10.51470/PLANTARCHIVES.2026.v26.supplement-1.258

A DECISION SUPPORT SYSTEM FOR SUSTAINABLE CROP RECOMMENDATION IN SAURASHTRA FROM DIVERSE SOIL CHARACTERS BASED ON RANDOM FOREST, DECISION TREE AND LOGISTIC REGRESSION

Chovatiya P.V.[1*], Patel D.V.[1], Shitap M.S.[1] and Sakarvadia H.L.[2]
[1]Department of Agricultural Statistics, College of Agriculture, Junagadh Agricultural University, Junagadh, Gujarat (India).
[2]Department of Soil Science, College of Agriculture, Junagadh Agricultural University, Junagadh, Gujarat (India)
*Corresponding author E-mail: pruthvip19618@gmail.com

(Date of Receiving : 08-10-2025; Date of Acceptance : 17-12-2025)

ABSTRACT

The process of analysing data and extracting useful information from it is known as data mining. Numerous industries, including banking, retail, healthcare, and agriculture, use data mining. In agriculture, data mining is utilized to analyse different biotic and abiotic aspects. India's economy and employment are heavily dependent on agriculture. Indian farmers frequently struggle with selecting the appropriate crop based on the needs of their soil. They experience a significant decline in productivity as a result. Precision agriculture has been used to solve this issue for farmers. Precision agriculture is a modern farming method that makes recommendations to farmers on the appropriate crop based on site-specific criteria using research data on soil properties and crop yield data gathering. This reduces the wrong choice on a crop and increase in productivity. In order to recommend a crop for the site-specific parameters with high accuracy and efficiency, this study proposes an ensemble model with majority voting technique that uses Random Tree, Decision Tree, and Logistic Regression as learners. This report provides a thorough analysis of these potential crops, discussing their significance, opportunities, difficulties, and prospects for the future. The system aims to increase agricultural output, lower risk, and boost farmer profitability through the use of historical data, soil quality evaluations, and crop performance models. The objective variable is the crop recommendation based on its nutrient content, whereas the feature variables are the several soil nutrients OC (Organic Carbon), S, Fe, K_2O, P_2O_5, Zn, etc. After preprocessing the dataset and applying different classification algorithms for crop recommendation, we discovered that Random Forest had the greatest accuracy score.

***Keywords*:** Crop recommendation, Random Forest, Machine learning, Decision Tree, Logistic Regression.

Introduction

For a nation, one of the most important aspects of its growth revolves around its potential to produce food. For generations, the production of essential food crops has been correlated with agriculture. In reality, however, the rapid pace of population growth has, by far, been the single biggest preoccupation of our society. In doing so, the scope of agriculture has been greatly undermined, particularly in terms of land use and fertility. Given that the area of land under cultivation in this era of urbanization and globalization is unlikely to increase, the focus will have to be on making the most of what there is. In agriculture, crop prediction is a key factor. Although recent research has opened up statistical information on agriculture, few studies have investigated crop prediction based on historical data. However, the unregulated use of fertilizers containing micronutrients, potassium, and nitrogen makes crop selection challenging. The vast datasets obtained can be used for crop prediction on a massive scale. Because of the nature of the issues, new machine learning techniques for cultivating arable land and optimizing the use of limited land resources are required. Numerous forecasting techniques have been tested by agricultural researchers in an effort to determine which crop would be most suited for a certain plot of land.

Random Forests, based on the ensembles of classification, Decision Trees and Logistic Regression, have become a widely used classification approach in various fields, including decision making process like crop selection based on soil parameters and environmental factor, etc. It is computationally efficient and quite simple to implement in a range of software packages (e.g., Python and R Statistics). The latter is particularly pertinent nowadays since picture categorization frequently uses high-dimensional data sets from many sources that are freely accessible. Reducing the model data load to the fewest number of inputs with maximum prediction accuracy is frequently desirable since not all data sets and predictor variables give the classifier pertinent information. Reducing model data load can reduce processing times and storage requirements, and can also be used to inform long-term analyses, as attention can focus on variables that provide relevant information to a given classification problem. Furthermore, it has also been demonstrated that with very high dimensional data sets, results can be noisier than models where only the most important variables are used. Random forest is a versatile machine learning algorithm with numerous applications across various industries. Let's see into some of the key sectors where random forest is widely used like banking, land use, medicine, marketing, agriculture, etc. In agriculture, random forest method applicable on Yield prediction, Disease &Weed Detection, Crop Recognition, Crop Quality, Category associated to soil management aspects & soil protection and management of Animal Welfare & Livestock Production.

Sahu *et al.* (2017) studied An efficient analysis of crop yield prediction using hadoop framework based on random forest approach at Bilaspur and stated that it included various parameters from soil like pH, N, P, Cu and atmosphere like temperature, CO_2 to analyze the crop by processing it in Hadoop framework. Random forest framework described works faster and gives better accuracy 91.43 percent in prediction than the current system to predict the suitable crop for the field. Keerthan Kumar *et al.* (2019) worked on Random Forest algorithm for soil fertility prediction and grading using machine learning at Andra Pradesh and started to propose a machine learning based solution for the analysis of the important soil properties based on the grading of the soil and prediction of crops suitable to the land. Soil attributes like pH, EC, OC, S, K, Zn, Mn, B, and soil type were taken as feature variables. The results of these algorithms were compared for crop recommendation and found that random forest has the highest accuracy score 72.74 percent. Rani *et al.* (2023) carried out Machine learning-based optimal crop selection system in smart agriculture at Telengana and concluded that the random forest classifier showed 97.24 percent accuracy for crop selection, 96.44 percent accuracy in predicting resource dependency, and 97.65 accuracies in giving the appropriate sowing time for the crop. The model construction time taken with a random forest classifier using mentioned data size was 5.34s. The model construction time for random forest classifer was also given which is not done in any of the previous ML-based crop selection models.

Materials and Methods

Study Area

The aim of proposed system is to help farmers to cultivate crop for better yield. The crops selected in this work are based on important crops from selected location. The selected crops are Groundnut and Cotton. Total number of observations 990 in which Groundnut 468 (47.30%) and Cotton 522 (57.70%) data (Figure-1). The dataset of crop yield is collected from department of soil science, college of agriculture, JAU, Junagadh. Soil classification can be done using soil nutrients data. Also all data collected from different district like Amreli, Bhavnagar, Jamnagar, Junagadh, Porbandar, Rajkot and Surendranagar. Based on all district each district has different yield and price of market was collected from DMI (2025). Three Machine learning algorithms used for soil classification are Random Forest, Decision Tree and Logistic Regression. The three algorithms will classify, and display confusion matrix, Precision, Recall, f1-score and average values, and at the end accuracy in percentage as output. Crop Prediction: Crop Prediction can be done using crop yield data, nutrients and location data. These inputs are passed to Random Forest, Decision Tree and Logistic Regression algorithms. These algorithms will predict crop based on present inputs.

Logistic Regression

One technique for examining functional correlations between variables is regression analysis. The link between the response or dependant variable and one or more explanatory or predictive factors is stated as an equation or model. The standard theory of multiple linear regression (MLR) analysis is valid when the response variable is quantitative. However, there are also many instances in statistical applications when the answer variable is qualitative, or more precisely, binary. The binary logistic regression model is the statistical model of choice for analyzing such binary (dichotomous) responses Pudumalar *et al.* (2017). It is a type of regression that uses a combination of continuous and discrete predictors to predict discrete variables. Without making any distributional assumptions about the predictors, it tackles the same issues as multiple regression and discriminant function analysis. It is not necessary for the predictors in a logistic regression model to be normally distributed, for the response-predictor connection to be linear, or for the variance of the observations in each group to be equal, among other requirements (LaValley, 2008).

An S-shaped curve is similar to the logistic response function. In this case, when X increases, the probability π first rises slowly before accelerating and eventually stabilizing, but it does not rise over 1.

If the probabilities can be represented using just one predictor variable, the S-curve's shape can be replicated as follows:

$$\pi= P(Y=1|X= x) = 1/(1+e^{-z})$$

where $z = \beta_0 + \beta_1 x$, and e is the base of the natural logarithm. Thus for more than one (say r) explanatory variables, the probability π is modeled as

$$\pi= P(Y=1|X_1 = x \ ...X_r = x_r)=1/(1+e^{-z})$$

where $z= \beta_0 +\beta_x +...+\beta_r x_r$

The equation in (2.2) is called the logistic regression equation. It is nonlinear in the parameters β_0, $\beta_1 \ldots \beta_r$. Modeling the response probabilities by the logistic distribution and estimating the parameters of the model given in (2.2) constitutes fitting a logistic regression. The method of estimation generally used is the maximum likelihood estimation method.

Let's look at the mathematical form that the logistic model is based on in order to understand why logistic regression is so popular. This function, f (z), is given by

$$f (z) = 1/ (1+e^{-z}), -\infty < z < \infty$$

Now when $z = -\infty$, $f (z) =0$ and when $z = \infty$, $f (z) =1$. Thus the range of f (z) is 0 to1. So the logistic model is popular because the logistic function, on which the model is based, provides. Estimates that lie in the range between zero and one. Secondly, an attractive S-shaped depiction of the cumulative impact of multiple explanatory factors on the likelihood of an occurrence.

Decision Tree

Decision tree is supervised learning technique that can be used for both classifier and regression problems but mostly it is preferred for solving classification problem. It is a tree structure classifier (Figure-2) where internal node represents the features of the dataset, branches represent the decision rule and each node represents the outcome. Decision or test performed on the basis of features of the given dataset. Decision trees are a non-parametric classification method that doesn't require any statistical presumptions about data distribution. It is one of the most effective and widely used tools for prediction and classification (Breiman, 2001).

The following steps can be used to demonstrate the working process of Decision Tree:

Step-1: Begin the tree with the root node, says S, which contains the complete dataset

Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM)

Step-3: Divide the S into subsets that contains possible values for the best attributes

Step-4: Generate the decision tree node, which contains the best attribute

Step-5: Recursively make new decision trees using the subsets of the dataset created in step-3 Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node

Random Forest

Ensemble learning techniques can significantly improve the performance of decision tree models by enhancing accuracy and resilience through the combination of predictions from multiple models. The goal is to address errors or biases in individual models by harnessing the collective intelligence of the ensemble. Bagging, also known as bootstrap aggregation, is a widely used ensemble learning method that helps reduce variance in noisy datasets. In bagging, a random sample of data is selected with replacement from the training set, allowing individual data points to be chosen multiple times. Random Forest is a powerful and versatile ensemble technique used for both classification and regression tasks. It is based on the bagging (bootstrap aggregation) technique over decision trees (Yariyan *et al.*, 2020). Bagging effectively minimises the variance of the base algorithms, particularly when they exhibit weak correlations. This supervised machine learning algorithm is renowned for its flexibility and user-friendly nature. The operational premise of RF begins by creating multiple bootstrap samples (randomly selected subsets with replacement) from the original dataset. These subsets serve as the training data for each decision tree in the forest (Figure-3). For each tree, a random subset of features (predictors) is selected for training. This random selection helps in do away with correlations among the trees and introduces diversity in the forest. Each decision tree is grown using the bootstrap sample and the randomly selected features. The trees are typically grown to their maximum depth, without pruning. When making predictions, each tree in the forest predicts an outcome based on its individual features. Then, the predictions are averaged to obtain the final output. RF often provides better accuracy than individual decision trees due to the ensemble effect and feature randomisation. Moreover, as it averages predictions from each of the multiple trees, it reduces over fitting compared to a single decision tree.

The final result from a Random Forest regressor can be obtained as:

$$\text{Entropy} = -\sum_{i=1}^{c} p_i * log_2(p_i)$$

Where, Entropy is the coefficient, c is the number of classes and p_i is the proportion of instances of class i in the node.

$$\text{Information gain (D, Feature)} = \text{Entropy(D)} - \sum_{i=1}^{c} p_i * Entropy\text{(Feature)}$$

Where, Information gain is the coefficient value at D(target variable) and Feature value (independent variable), c is the number of classes of feature, p_i is the proportion of instances of class i in the feature dataset.

The following steps can be used to demonstrate the working process of Random Forest:

Step 1: Pick M data points at random from the training set

Step 2: Create decision trees for your chosen data points (Subsets)

Step 3: Each decision tree will produce a result and analyze it

Step 4: For classification and regression, accordingly, the final output is based on majority voting or averaging, accordingly

Model evaluation and diagnostics

A confusion matrix [Susmaga (2004)] is a matrix that summarizes the performance of a machine learning model on a set of test data. It is a means of displaying the number of accurate and inaccurate instances based on the model's predictions. It is often used to measure the performance of classification models, which aim to predict a categorical label for each input instance.

The matrix displays (Figure-4) the number of instances produced by the model on the test data.

True Positive (TP): The model correctly predicted a positive outcome (the actual outcome was positive).

True Negative (TN): The model correctly predicted a negative outcome (the actual outcome was negative).

False Positive (FP): It is the proportion of predicted event responses that were observed as nonevents Also known as a Type I error.

False Negative (FN): It is the proportion of predicted nonevent responses that were observed as events. Also known as a Type II error.

Accuracy measures the overall correctness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the accuracy of the positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall measures the ability of the model to capture all the positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

Random Forest provides a feature importance score for each feature, indicating the contribution of each feature to the model's predictive performance.

Result and Discussion

The suggested approach makes use of many techniques to increase crop prediction accuracy. We split the 990 total observations into two datasets, 80:20 for training and testing. Thus, we have 792 training observations and 198 testing observations. In this study, crops were predicted using three models: logistic regression, decision tree, and random forest. Prediction of crops based on soil property, area of sowing, market price, and yield of crops. There are areas of crops based on different districts, like Amreli, Bhavnagar, Jamnagar, Junagadh, Porbandar, Rajkot, and Surendranagar. Soil property is categorized by organic carbon, S (sulphur), Fe (iron), K_2O (potassium), P_2O_5 (phosphorus) and Zn (zinc). Also, data is collected on different crop market prices and yield per hectare for the calculation of total output from the farm. Organic carbon (0.09-1.20 %), P_2O_5 (1.08-98.75 kg/ha), K_2O (192-1164 kg/ha), S (0.78-54.18 ppm), Zn (0.18-4.73 ppm) and Fe (1.08-22.06 ppm) was divide into five categories: Low, Moderate low, Moderate, Moderate high and high.

After Pre-processing the dataset categories each data using label encode code using the python. All dataset divides into 80 percent data as training and 20 percent data as testing. In which our dependent data was crop category and explanatory data was organic carbon, S, Fe, K_2O, P_2O, Zn, profite and District. After collecting all dataset, we train logistic regression, decision tree and random forest models. After applying models, we test all model and record the result. Result saw in Table-1 that saw the testing actual data of groundnut and cotton 97 and 101, respectively. Logistic regression model predicted 56 for groundnut and 86 for cotton. Decision tree model predicted 85 for groundnut and 94 for cotton. Random forest model predicted 94 for groundnut and 99 for cotton.

As per mentioned Table-2, Logistic Regression model given 0.6588, 0.6747 and 0.6667 value of Precision, Recall and F1-score, respectively. Decision Tree model given precision, Recall and F1-score 0.8947, 0.9043 and 0.9040 value, respectively. So, Decision tree model was more accurate than Logistic Regression as per models result. Result of Random Forest model given 0.9691, 0.9792 and 0.9741 precision, Recall and F1-score, respectively. Overall comparison between all the model highly precise and accurate model was Random Forest (97.47 %) compare to Logistic Regression (71.72 %) and Decision Tree (90.40%) that easily sawn in Figure-5. For crop recommendation based on Random Forest model, total output value and District have more impotance than Organic carbon, S, Fe, K_2O, P_2O_5 and Zn (Table-3).

Conclusion

India is a country where agriculture is very important. The nation grows when farmers are wealthy. Agriculture is critical to ensuring global food security and financial stability, but faces significant challenges such as climate change, resource scarcity and population growth. To address these issues, crop recommendation systems have proven to be valuable tools to help farmers decide which crops to plant. In order to boost productivity and profit from this method, our work would assist farmers in planting the appropriate seed depending on soil conditions. As a result, farmers can plant the appropriate crop to increase both their own yield and the nation's total productivity. In addition to implementing yield prediction, our next study will focus on an enhanced data set with numerous features. The system provides the best results based on confusion matrix analysis and employs supervised machine learning methods such as Random Forest Classifier, Decision Tree, and Logistic Regression Multi-Variate. The Random Forest Classifier, which produces the best and most accurate results, will be selected once the outcomes of various algorithms are compared. Therefore, this system will help reduce the struggle faced by the farmers. We are concerned with the grading of the soil and the prediction of crops that are appropriate for the land after analyzing the key characteristics of the soil. This will serve as an easy way to give farmers the knowledge they need to maximize their surplus and achieve high yields, which will reduce their challenges.

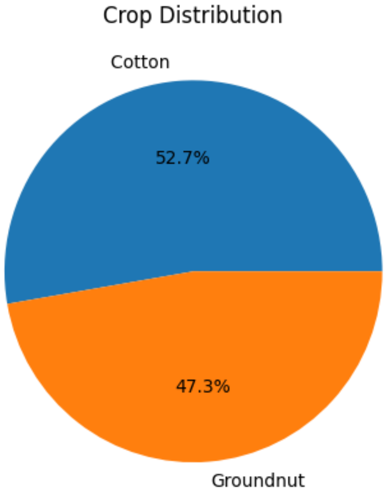


Fig. 1: Analysis of dataset with respect to crops

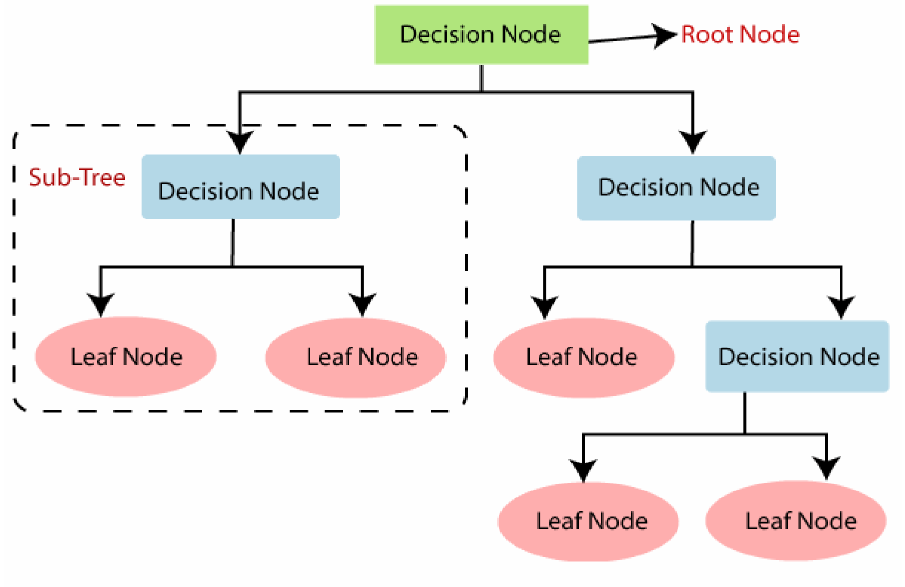


Fig. 2 : Diagram explains the working of the Decision Tree algorithm

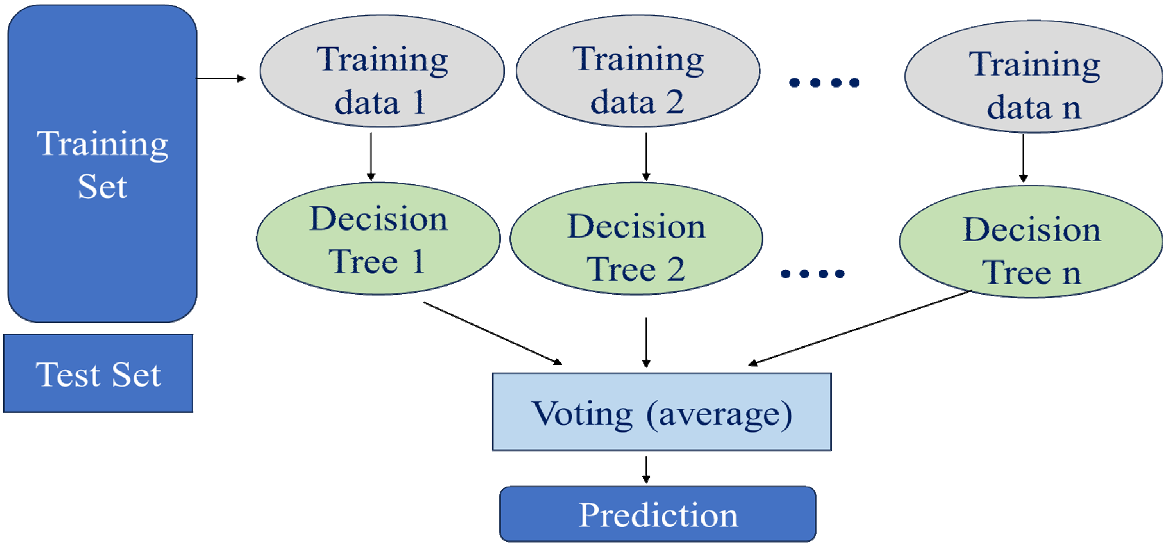
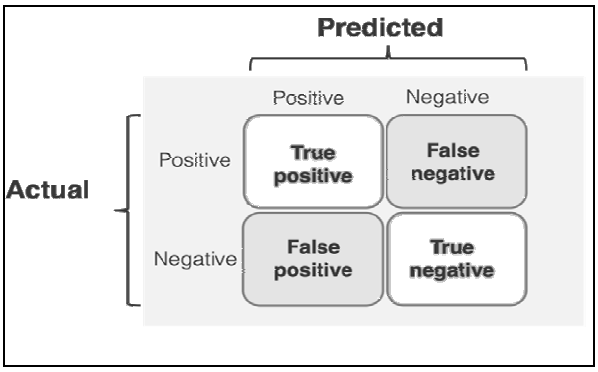


Fig. 3 : Diagram explains the working of the Random Forest algorithm



		Predicted	
		Positive	Negative
Actual	Positive	True positive	False negative
	Negative	False positive	True negative

Fig. 4 : Confusion matrix

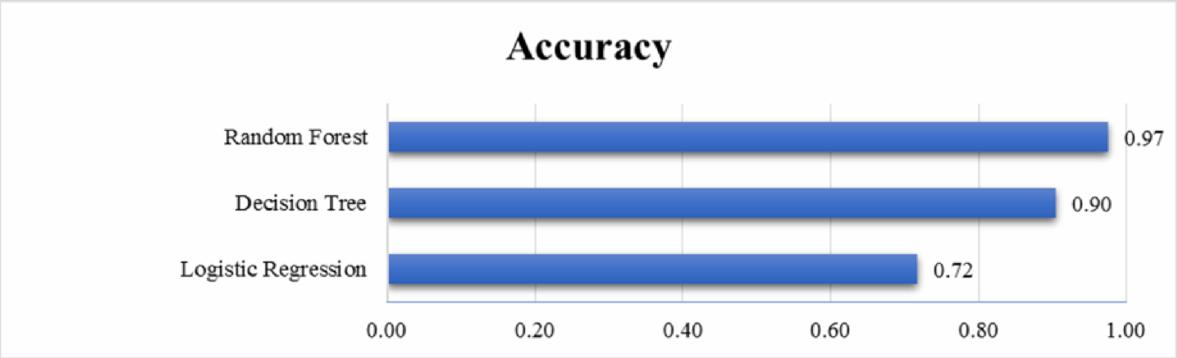


Fig. 5 : Proposed logistic regression, Decision tree and Random Forest accuracy among groundnut and cotton dataset

Table 1 : Actual crop vs Predicted crop based on different model

Crops	**Testing Actual Dataset**	**Crop Recommendation**		
		Logistic Regression	**Decision Tree**	**Random Forest**
Groundnut	97	56	85	94
Cotton	101	86	94	99

Table 2: Classification report of all the models

Model	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Rank**
Logistic Regression	0.6588	0.6747	0.6667	0.7172	3
Decision Tree	0.8947	0.9043	0.8995	0.9040	2
Random Forest	0.9691	0.9792	0.9741	0.9747	1

Table 3: Feature Importance for Random Forest

	Feature	**Importance**
1	Total output value	0.6594
2	District	0.1466
3	Organic carbon	0.0416
4	S	0.0412
5	Fe	0.0356
6	K_20	0.0312
7	P_2O_5	0.0244
8	Zn	0.0200

References

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1): 5-32.

DMI, (2025). Directorate of Marketing and Inspection (DMI), Ministry of Agriculture and Farmers Welfare, Government of India, Available at https://agmarknet.gov.in/home, Accessed on 26th September, 2025.

Keerthan Kumar, T. G., Shubha, C. A. and Sushma, S. A. (2019). Random forest algorithm for soil fertility prediction and grading using machine learning. *International Journal Innovative Technology Exploring Engineering*, **9**: 1301-1304.

LaValley, M. P. (2008). Logistic regression. *Circulation*, **117**(18): 2395-2399.

Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T. and Nisha, J. (2017, January). Crop recommendation system for precision agriculture. In *2016 eighth international conference on advanced computing* (ICoAC) (pp. 32-36). IEEE.

Rani, S., Mishra, A. K., Kataria, A., Mallik, S. and Qin, H. (2023). Machine learning-based optimal crop selection system in smart agriculture. *Scientific Reports*, **13**(1): 15997.

Sahu, S., Chawla, M. and Khare, N. (2017, May). An efficient analysis of crop yield prediction using Hadoop framework based on random forest approach. In 2017 *international conference on computing, communication and automation (ICCCA)* (pp. 53-57). IEEE.

Susmaga, R. (2004, May). Confusion matrix visualization. In *Intelligent information processing and web mining: proceedings of the* *international IIS: IIPWM '04 conference held in zakopane, Poland, may 17–20, 2004* (pp. 107-116). Berlin, Heidelberg: Springer Berlin Heidelberg.

Yariyan, P., Janizadeh, S., Van Phong, T., Nguyen, H. D., Costache, R., Van Le, H. and Tiefenbacher, J. P. (2020). Improvement of best first decision trees using bagging and dagging ensembles for flood probability mapping. *Water Resources Management*, **34**: 3037-3053.