Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:05:2022
DOI 10.17605/OSF.IO/AN5V9

COMPARE MACHINE LEARNING VALIDATION TECHNIQUES AND ESTIMATE EVALUATION PERFORMANCE USING SOIL ENZYME ACTIVITY AND SUGGESTED CROPS

YOGESH SHAHARE*
Department of Information Technology, MGMCET, Navi Mumbai, Maharashtra, India,
(*Corresponding Author)

MUKUND PRATAP SINGH
Department of Computer Science and Engineering, CUET, Chitkara University, Punjab, India,

VINAY GAUTAM
Department of Computer Science and Engineering, CUET, Chitkara University, Punjab, India

SANJAY B. WAYKAR
Department of Information Technology, MGMCET, Navi Mumbai, Maharashtra, India,

N.P.KARLEKAR
Department of Computer Engineering, MGMCET, Navi Mumbai, Maharashtra, India

Abstract

The aim of this study was to compare the three machine learning validation approach like Holdout, K-fold, and stratified for predicting each soil enzyme activities such as Acid phosphatase, Alkaline Phosphatase, Cellulase, Dehydrogenase, Invertase, N-acetyl-glucosaminidase, Phosphatase, Protease, Urease with physical soil features like sand, silt, clay, depth, and chemical soil properties are available nitrogen, available phosphorus, soil organic carbon, soil organic matter, and other components like PH value, soil fertility level such as low, medium and high. This study used different machine learning algorithms random forest, extra tree, AdaBoost, support vector machine, logistic, ridge, k-nearest, and decision tree algorithm to predict the soil enzyme activity. Compare all the machine learning algorithms and artificial neural networks for calculating better accuracy using classifier algorithm, and also calculate to measure the optimum error using evaluation techniques like means squared error(MSE), root means squared error(RMSE), and mean absolute error(MAE) by regressor algorithm. Suggest the specific crops based on soil properties using a k-means unsupervised machine learning algorithm. In this study, for cellulose, N-acetyl-glucosaminidase enzyme activity by RF, Extra tree, and Adaboost algorithm was better accuracy (99%) using holdout, and K-fold, and stratified validation approach. N-acetyl-glucosaminidase, MSE, RMSE, and MAE measure the optimum error like random forest regressor (RFR) is 0.0094, 0.0712, 0.0155 multiple linear regression (MLR) is 0.005,0.0712, 0.2265 and Decision tree regressor (DTR) is 0.0103,0,0712, 0.0103.

Keywords: Machine Learning algorithm, artificial neural network, soil enzyme activity, soil chemical properties, soil fertility

1. **Introduction**

The activity of soil enzymes is influenced by soil microbial characteristics. Each characteristic takes into account a variety of factors, including soil texture, soil organic matter content, composition, and soil microbial activity [1], [2]. Because of their sensitivity to heavy metal pollution and direct relationship with soil functions related to the C, N, P, and S cycles, soil enzyme activity (SEAs) is the performance of soil quality and health [3], [4]. Soil enzyme activity can indicate the direction and intensity of soil biogeochemical cycling processes and is used as a biological indicator to assess soil fertility and quality. The enzymatic action of hydrolases and oxidoreductases determines the conversion of various organic and inorganic nutrients, as well as the mineralization rate of available nutrients in the soil [5]. The mixed activity of microbial community and environment may be responsible for the increased soil enzyme activity found in most soils. Furthermore, soil quality is measured by evaluating a variety of enzymatic activities as a substitute for soil microorganisms [6]. Soil microorganism activities are aided in part by enzymes inferred from microbes and are critical for the decomposition of many insoluble organic substances, thereby activating nutrient cycling. Extracellular enzymes obtained by soil microbe reducing and transform polymeric residues into commonly available nutrients that can be inculcated by plants and microorganisms [7]. These extracellular enzymes are involved for the mineralization and cycling of geological nitrogen (N), phosphate (P), and carbon (C) and can be classified as such, though some enzymes may participate in more than one cycle [8],[9]. Catalyse is one kind of the enzyme which is involved the hydrogen peroxide (H_2O_2) with water H_2O and Oxygen (O_2). C cyclicling of enzyme are involved the different enzyme activity include β-D-cellobiohydrolase, dehydrogenase, and β-D-glucosidase. Based on soil depth, the effects of soil pH, organic carbon (C), and available nitrogen (N) on hydrogen peroxidase, dehydrogenase, and alkaline phosphatase activities were significant (0-10 cm). Furthermore, even at depths of 10–20 cm, the pH value has a positive impact on all soil enzyme activities such as hydrogen peroxidase, dehydrogenase, and alkaline phosphatase[10][11]. Important biological soil fertility indicators included hydrogen peroxidase, dehydrogenase, and alkaline phosphatase. Urease, for example, is a key player in the degradation and transformation of nitrogen in the soil ecosystem, hydrolyzing urea into ammonia or amino salts via carbon-nitrogen bonds acting on organic matter [12]. Urease has a direct impact on the nitrogen supply rate in soil, which is commonly employed as a measure of nitrogen deficiency. As a result, estimating urease activities can help investigators better understand the biological mechanisms of carbon and nitrogen transformation, as well as provide guidelines for assessing soil quality in specific areas.

The goal of the study was to discover the optimal model for predicting soil enzyme activity utilizing several machine learning validation techniques to calculate achievable accuracy and MSE, RMSE, and MAE. Different soil enzyme activity, as well as soil physical and chemical parameters, were utilized in this study. For classification and regression, various machine learning techniques were employed, including Random forest, Decision tree, K-nearest neighbor, logistic regression, Ridge regression, Support vector machine, Adaptive boosting, Extra tree algorithm, and Artificial neural network. We worked on additional aspects in this research, such as soil fertility and soil enzyme activity, to get a good result, so that this research will be beneficial to farmers in growing a decent number of crops.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:55 Issue:02:2022
DOI 10.17605/OSF.IO/UWJK5

2 Materials and Methods

*2.*1 Study Area and Soil Sampling

The research was conducted in several villages in Maharashtra's Bhandara district, Sakoli taluka. Bhandara is a prominent district in Maharashtra, located at 21°10'N 79°39'E in the Nagpur division. The district is bordered on the north by the Madhya Pradesh districts of Balaghat and Chandrapur, on the south by the Chandrapur district of Madhya Pradesh, on the southeast by Gadchiroli, and on the east and west by Gondia and Nagpur [13]. **Figure 1.** Depicts the location of the research area.

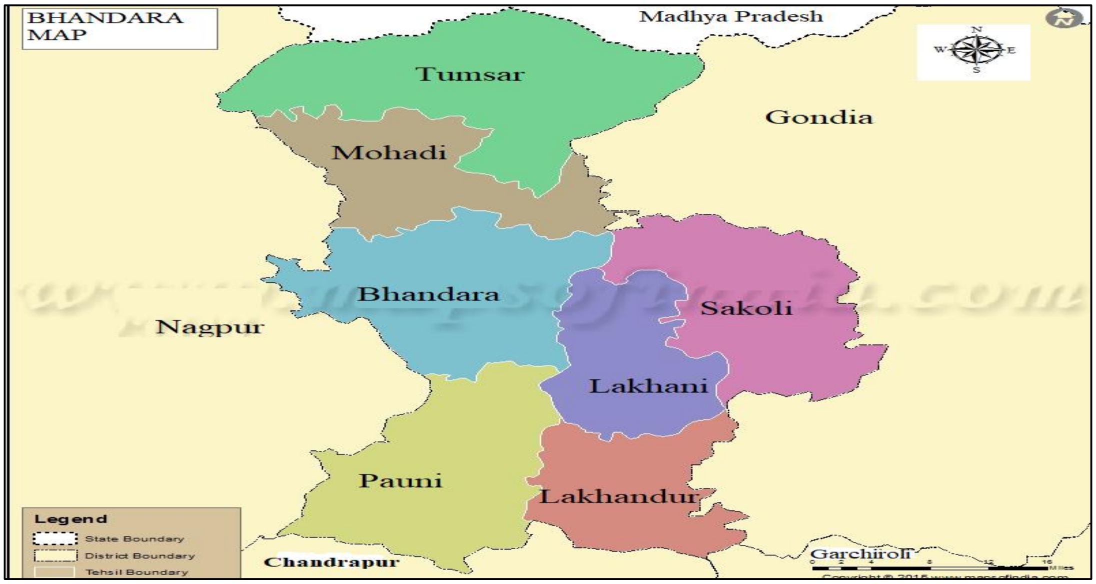

Figure 1. Location of the research area

2.2 Determination of soil properties

Soil samples intended for the analysis of the soil properties were collected from Maharashtra, Bhandara district from different villages. Total soil samples were collected from 5000 from each village of specific farmer's land data for analyzing the soil physical properties, chemical and biological properties [14], [15]. Soil physical properties consist of sand, silt, clay, and depth of soil (10, 15, and 30 cm) for collecting the soil sample. PH value rating is represented in **Figure 2.**

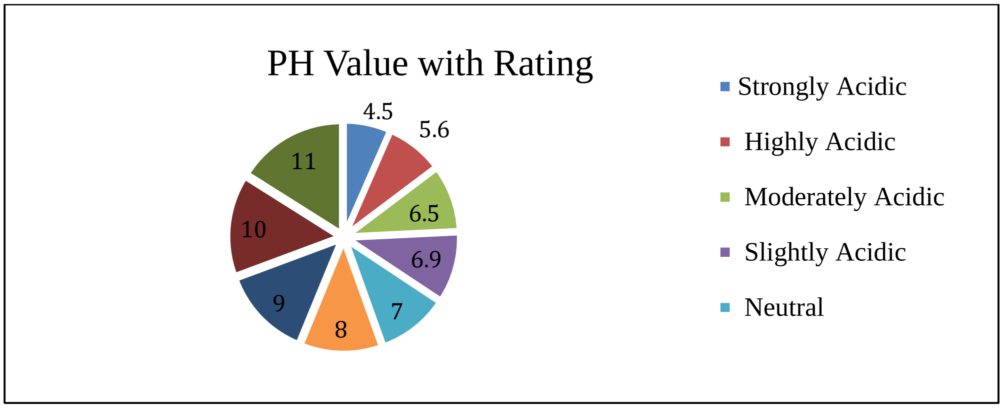

Figure 2. PH classification with a rating

2.3 Proposed Methodology

This study used various machine learning algorithms such as random forest, extra tree, AdaBoost, support vector machine, logistic algorithm, ridge algorithm, k-nearest algorithm, and decision tree algorithm for calculating the better accuracy using three validation approaches like holdout validation, K-Fold validation, and stratified validation

Algorithm 1: Validation approach for predicting soil enzyme activity (VPSEA)

Input: Read all the soil features (Sf_i) sand, silt, clay, depth, PH, av_N, Av_P, SOM, SOC, fertility level
Output: Predict each soil enzyme activity (SEA_i)

```
Step 1:  Apply pre-process methods
         If check $Sf_i$ = = null, missing value
         Then remove null values or missing value
         If check $Sf_i$ = = categorical value
         Then convert the numerical value
         If $Sf_i$ ! = = Scalar
         Then apply the scalar method (MinScalar)
Step 2:  Take 80% data for training and 20% for testing using train_test_split method
Step 3:  Apply Machine learning algorithms (MLA) like RF, Extra Tree, Adaboost, SVM, Logistic,
         Ridge, Decision, and KNN
Step 4.  Apply Artificial neural network (ANN) with batch size=32, epoch=100
Step 5.  Apply three validation techniques like HoldOut, K-Fold, and Stratified validation methods
Step 6.  Calculate the Mean accuracy of all MLA
Step 7.  Measure the optimum error using MSE, RMSE, and MAE
Step 8.  Predict $SEA_i$
Step 9.  End
```

Algorithm 2: Proposed a specific crop (PSC)

```
           Input:  Read all the soil cluster data ($SC_i$) sand, silt, clay, depth, PH, av_N,
           Av_P, SOM, SOC, fertility level, Soil enzyme activity
           Output:  Estimate the specific crops
Step 1:    Apply preprocess methods
           If check $SC_i$ = = null, missing value
           Then remove null values or missing value
           If check $SC_i$ = = categorical value
           Then convert the numerical value
Step 2:    Take X = DV (dependent variable)
Step 3     Remove Y = IDV (independent variable)
Step 4:    Apply K-Means algorithm
Step 5:    Find the K value using elbow visualization graph
Step 6:    Select k= n where n = 1,2, 3, 4, --------n
Step 7:    K= n
Step 8:    $SC_i$ = n (keep all features data cluster wise)
Step 9:    Copy k value in $SC_i$ with crops data (IDV)
Step 10:   Find out k = crops (find specific crops with cluster no and $SC_i$ data)
Step 11:   End
```

2.4. Machine Learning Models:

2.4.1. Random Forest Algorithm (RF)

Random forest algorithm is a supervised machine learning algorithm which is used ensemble techniques. The set of decision trees consists of a forest and some changes should randomly depend on the dataset. Random forest algorithm is the best result algorithm of supervised machine learning algorithm[16], [17]. Mathematical notation of Random Forest: if each $h_k\,(x)$ is a decision tree, then the ensemble is a RF. We defined the parameters of the decision tree for classifier $h_k\,(x)$,

$$\theta_k\ = (\theta_{k1}\,, \theta_{k2}\,, \theta_{k\;3,}\ldots\ldots\theta_{kn}\,) \qquad (1)$$

These are all parameters include the structure of the tree, which variables are split in which node, etc.). $h_k\,(x) = \ h(x \mid \theta_k\ \ )$ (2)

Where θ is the feature or variables of the kth tree which is split into a different variable for making the decision tree.

2.4.2. Extra Tree Algorithm

It's an ensemble technique that generates a classification result by combining the results of several de-correlated decision trees collected in a "tree." It is conceptually identical to a Random Forest Classifier, with the exception of how the decision trees in the forest are constructed[18], [19]. The Gini index is used for constructing the forest for getting the result. Randomized decision tree features are collected for creating the forest from a subsample of the dataset and finding out the important feature by using Gini index methods. For the classification problem calculate the Gini impurity and entropy. For regression problem calculate to measure the means squared error (MSE), mean absolute error (MAE)

2.4.3. Adaptive Boosting Algorithm (Adaboost)

AdaBoost, also known as Adaptive Boosting, is a Machine Learning approach that is employed as part of an Ensemble Method. Decision trees with one level, or Decision trees with only one split, are the most popular algorithm used with AdaBoost [20], [21]. Decision Stumps is another name for these trees. Consider there are three stumps are present along with weight which is derived a prediction p1, p2, and p3.

Adaboost prediction = $\sum_{i=1}^{n} stump_i * p_i$ (3)

1. Assume the dataset with Ns_i is the number of samples, we initialize the weight of each feature data point with $we_i = \frac{1}{Ns_i}$
2. For m = 1 to M:
   (a) The sample dataset of weight is we_i to obtain the training sample x_i
   (b) Fit a classifier k_m using all the training sample x_i
   (c) Calculate € = $\frac{\sum_{y_i \neq k_m(x_i)} {we_i}^{(m)}}{\sum_{y_i} {we_i}^{(m)}}$ (4)
   Where y_i is the target variable, ${we_i}^{(m)}$ is the weight of the sample of i and iteration of m.
   (d) Calculate α_m = $\frac{1}{2}\ln\frac{1-€}{€}$, where α_m= confidence predictive power of stump (5)
   (e) Updated all the sample weights ${we_i}^{(m+1)}$ = ${we_i}^{(m)}\ e^{-\alpha_m y}\ k_{m^{(x)}}$ (6)
   (f) New prediction calculated is $k(x) = sign\ [\ \sum_{m=1}^{m} \alpha_m\ k_m\ (x)\ ]$ (7)

2.4.4. Support Vector Machine (SVM)

Support vector machine is a supervised machine learning algorithm. It is used for both classification and regression problems. SVM consists of support vectors, hyperplane, and marginal distance, linear separable, and nonlinear separable technique [22]. Maximize marginal distance and support vector are passing through the marginal plane which is calculated nearest (+ve) and (–ve) point. Mathematical notation of SVM as follows:

$w^T x_1$ + b = -1, $w^T x_2$ + b = 1 (8)

$w^T (x_2$ -x_1) = 2, we get $\frac{w^T}{||w||}\ (x2 - x1) = \frac{2}{||w||}$ (9)

$(w^*, b^*)\ \max \frac{2}{||w||}$ By optimization function (10)

$$y_i \begin{cases} -1, & w^T x_1 + \mathrm{b}\ \leq -1 \\ 1\ , & w^T x_1 + \mathrm{b} \geq 1 \end{cases} \qquad (11)$$

$y_i\ =\ w^T x_i\ +\ b_i\ \geq 1$ (12)

2.4.5. Logistic Regression

Logistic regression solve by classification problem under supervised machine learning algorithm. Logistic classification is derived from two categories which are binary classification and multiclass classification to solve the classification problem and find the efficient accuracy using the logistic function or sigmoid function [23], [24].

Logistic function or sigmoid: $p = \frac{1}{1+ e^{-(\beta_0+\beta_1 x)}}$ => $y = \frac{1}{1+ e^{-x}}$ (13)

Assume that $y = (\beta_0 + \beta_1 x)$ => $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x + \in$ (14)

Where β_0 *is the intercept, and* $\beta_1 x$ *is the coeficient and* $\in$ *is error*

2.4.6. Ridge Regression

Ridge regression is used to find the best fit line of regression and calculate the cost function for fit the regression. Whenever the overfitting data are available then more chances are occurred for error so need to reduce the sum of means square error using the regularization method. Ridge regression is used as a regularization technique for creating the generalized model [25], [26] . Ridge regression is defined as the L2 regularization method for calculating optimum error.

Ridge = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{m} we_j \times x_{ij})^2 + \lambda \sum_{j=0}^{m} {we_j}^2$ (15)

2.4.7. K-Nearest Algorithm

K-nearest algorithm is the most powerful classification and regression supervised machine learning algorithm. To measure the k parameter like a number of the nearest data point. The technique works by calculating the distance between these points' mathematical values. It finds the chance of the points being similar to the test data by computing the distance between each data point and the test data [27], [28]. The probabilities of which points share the highest probabilities are used to classify based on the dataset. Calculate the distance metrics by using e Euclidean metric formula.

$d\,(x, x') = \sqrt{(x_1 - {x'}_1)^2 + \cdots + (x_n - {x'}_n)^2}$ (16)

$p\,(y = j | X = x) = \frac{1}{k} \sum_{i \in A} I(y^{(i)} = j)$ (17)

2.4.8. Decision Tree

A decision tree was used to solve the problem using both techniques classification and regression supervised machine learning algorithm. A decision tree algorithm is used to create a tree where the nodes indicate features (attributes), and branch has noted a decision (rule) and leaf nodes indicate the outcomes (discrete and continuous). To measure the purity of split node using entropy, information gain is to compute the average of entropy of each attribute or property [29], [30]. Assume that $p_{(+ve)}$ is the percentage of positive class and $p_{(-ve)}$ is the negative class.

$H(s) = -p_{(+ve)} log_2\,(p_{+ve}) - p_{(-ve)}\, log_2\,(p_{-ve})$ (18)

$Gain\,(S, A) = H(S) - \sum_{v \in val} \frac{|SV|}{|S|} H(SV)$ (19)

Higher information gain is used to construct the decision tree.

Gini Impurity (GI) = $1 - \sum_{i=1}^{n}(p)^2$ => $1 - [(p_{+ve})^2 + (p_{-ve})^2]$ (20)

2.4.9. K-Means Algorithm

The K–means clustering algorithm is a simple unsupervised machine learning approach for generating a number of clusters based on a dataset. This approach allocates data points repeatedly for the creation of k clusters based on how close the data point is to the

cluster centroid [31], [32]. Determine the k number of cluster centroids and data points that belong to the cluster.
Using Euclidian formula

$$d\ (r,s) = \sqrt{(x_1 - x'_1)^2 + \cdots + (x_n - x'_n)^2} \qquad (21)$$

Cluster centroid is denoted by C_i then each data point of x is allocated to the cluster

$$\arg min_{ci \in C}\ dist\ (c_i, x)^2 \qquad (22)$$

Find new cluster centroids $c_i = \frac{1}{|s_{i|}} \sum_{x_i \in S_i} x_i$ (23)

Where S_i is the set of all points assigned in the i^{th} cluster

3. Result and Analysis

In this paper, for experiment analysis collected the 5000 total soil sample datasets including soil physical properties, chemical properties, and biological properties along with crop dataset. Collected the soil laboratory testing dataset from Bhandara district, Maharashtra to find out whether the soil nutrient is balanced or not in the available soil and what amount of nutrients is present and how much nutrients amount is required for growing the crops, and which crops are suitable for increasing the crop productivity [33]. In this paper, analyze the soil's different soil enzyme activities such as Acid phosphatase, Alkaline phosphatase, Cellulase, Dehydrogenase, Invertase, N-acetyl-glucosaminidase, Phosphatase, Protease, Urease [34], [35]. For each enzyme activities find out the nutrients in available soil and how much proportion is required for balancing the nutrients. To predict each soil enzyme activity based on all soil properties and fertility levels.

Figure 3. Represents the quantity of each soil property.

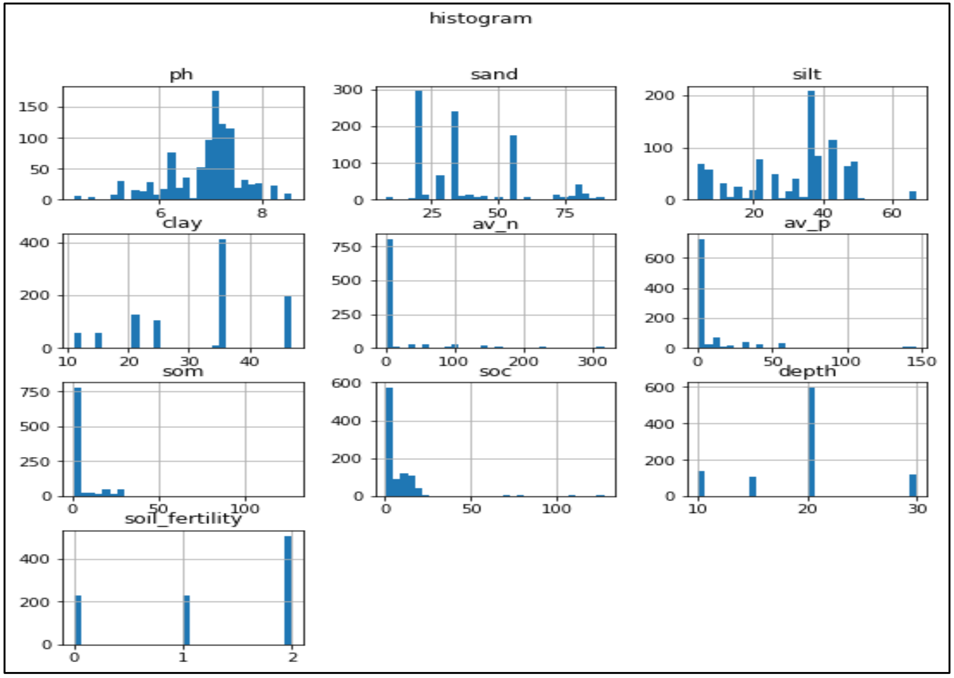

Figure 3. Histogram graph of **s**oil properties quantity available in soil dataset

This paper used an artificial neural network for calculating the better accuracy and optimum loss for predicting each soil enzyme activity is shown in **Table 1.** For this analysis we have taken batch size is 32 and 100 epochs for each soil enzyme activity. Artificial neural networks represent a good accuracy of N-acetyl-glucosaminidase enzyme is (99 %) and Cellulase is also given (99 %) as compared to other enzyme activities and loss is for both given an optimum value is 0.0059, and 0.0129.

Table 1. Calculate loss and accuracy using ANN

Soil Enzyme Activities	**Batch size**	**Epochs**	**loss**	**Accuracy**
Acid phosphatase	32	100	0.3859	0.8192
Alkaline phosphatase	32	100	0.2403	0.909
Cellulase	32	100	0.0129	0.994
Dehydrogenase	32	100	0.2232	0.9168
Invertase	32	100	0.308	0.8752
N-acetyl-glucosaminidase	32	100	0.0059	0.9961
Phosphatase	32	100	0.1345	0.9597
Protease	32	100	0.2229	0.9129
Urease	32	100	0.5454	0.7191

3.1. Result of Validation Techniques

3.5.1. HoldOut Validation

Holdout techniques for randomly splitting the training and testing data from the unseen dataset for calculating better accuracy. This paper used different machine learning algorithms like Random forest, Extra tree, Adaboost, SVM, KNN, Logistic, Ridge, and Decision tree and find the best performance model using the holdout validation approach. Holdout techniques find a better performance model for predicting each soil enzyme activity. **Figure 4.** Depicts the holdout methods of all machine learning algorithms with accuracy. In this analysis, Cellulase, and N-acetyl-glucosaminidase enzyme activities represent a better accuracy (99 %) using Random forest algorithm, Extra Tree, Adaboost, and Decision tree all these algorithms given best performance for predicting the activity.

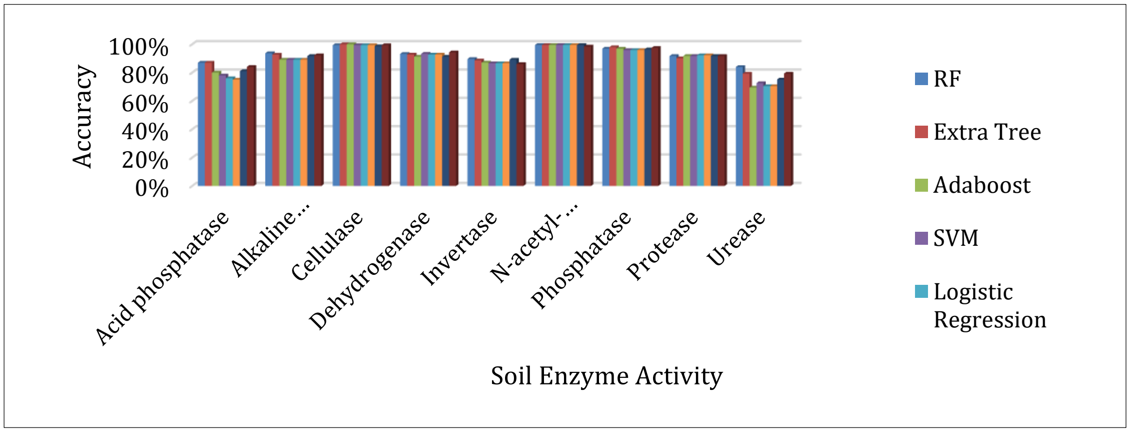

Figure 4. Accuracy of all Machine Learning Algorithm using HoldOut Validation Approach

3.5.2. K-Fold Validation

K-fold validation techniques used the k fold method which is a randomly folded dataset of training and testing depending on the k value. **Figure 5**. Shows a K-fold validation approach for finding a better performance model for predicting each soil enzyme activity. This approach for predicting the Cellulase enzyme activity finds the (99 %) accuracy showing with Random forest algorithm, Extra Tree, Adaboost and Decision tree, and N-acetyl-glucosaminidase enzyme activities represent a better accuracy (99.5 %) showing with all algorithms.

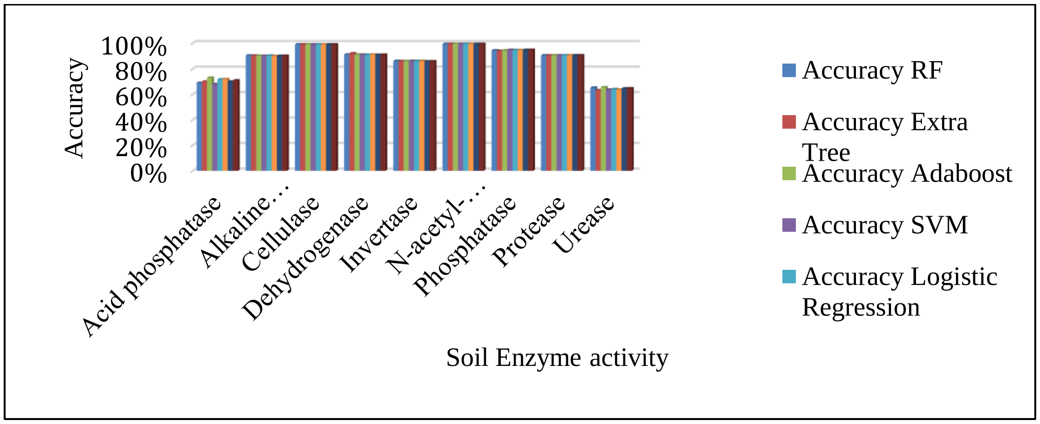

Figure 5. Accuracy of all Machine Learning Algorithm using K-Fold Validation Approach

3.5.3. Stratified Validation

Stratified validation technique use to find out the better accuracy from the mean value of taking same proportion group of training and testing dataset from the original dataset. It implements cross-validation techniques to find out the better performance model by using different machine algorithms. **Figure 6**. Shows the Stratified validation technique for calculating the efficient accuracy. In this approach, In this approach for predicting the

Cellulase enzyme activity finds the (99 %) accuracy showing with Random forest algorithm, Extra Tree, Adaboost and Decision tree, and N-acetyl-glucosaminidase enzyme activities represent a better accuracy (99.5 %) showing with all algorithms.

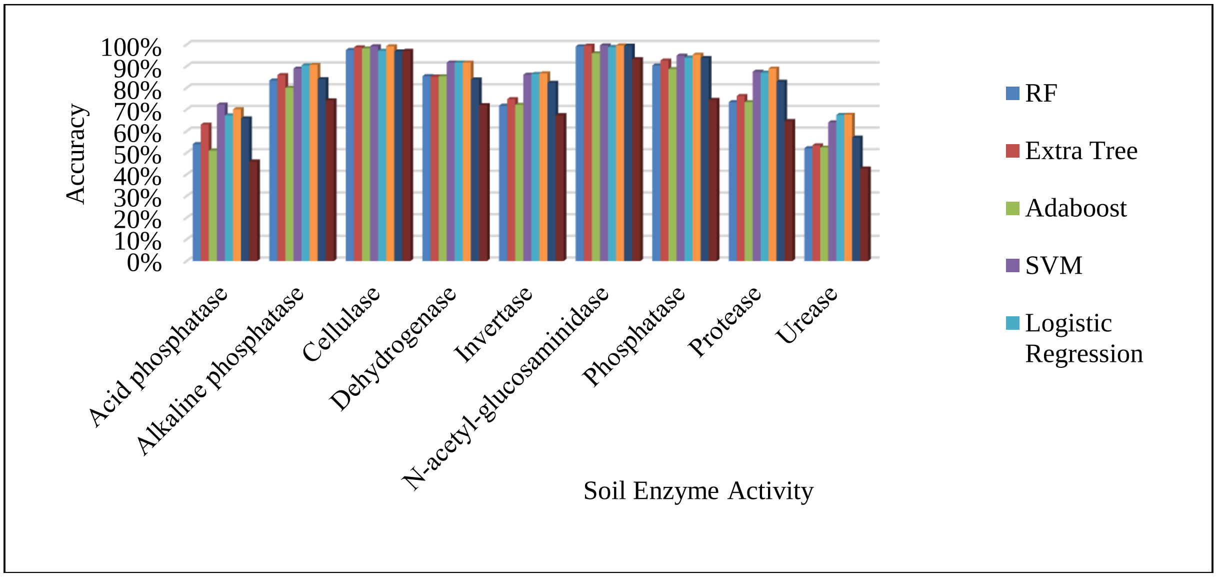

Figure 6. Accuracy of all Machine Learning Algorithm using Stratified Validation Approach

In this analysis, measure the optimum error using means squared error evaluation methods to predict each soil enzyme activity as shown in **Figure 7.** According to this analysis, the Decision tree regressor (DTR) is the best optimum error model for predicting the cellulose, and N-acetyl-glucosaminidase activity. Compare both activities using Random forest regressor (RFR), Multiple linear regressors (MLR), and Decision tree regressor (DTR) for obtaining a better result. N-acetyl-glucosaminidase activity predicts the best solution using all these algorithms (0.0094, 0.005, 0.0103).

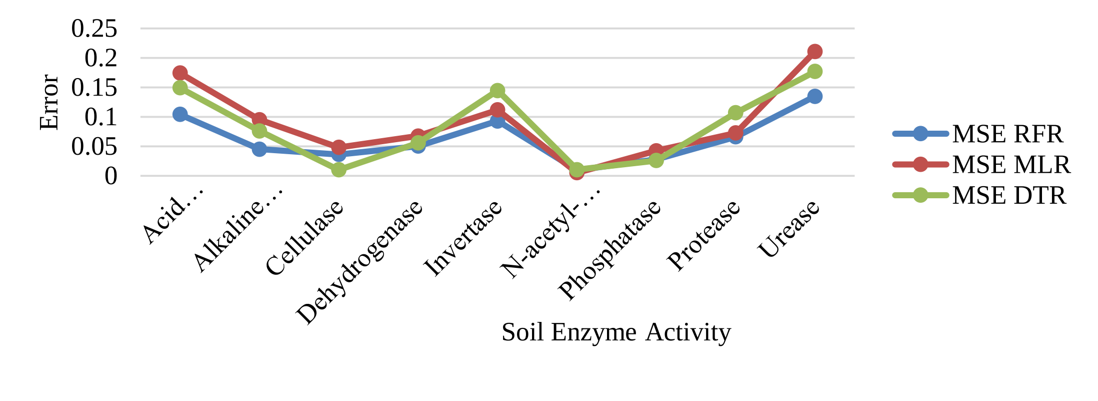

Figure 7. Means Squared Error for soil enzyme activity with regressor algorithms

Root means the squared error is the second evaluation technique for calculating less error using RFR, MLR, and DTR regressor algorithm shown in Figure **8**. RFR is the best model for predicting the cellulose, and N-acetyl-glucosaminidase enzyme activity as compared to another regressor algorithm. For cellulose, RMSE is 0.0698, and N-acetyl-glucosaminidase RMSE is 0.0712. Mean absolute error is the third evaluation technique for measuring the less error for predicting each soil enzyme activity shown in **Figure 9.** Both enzyme activities have given the best performance results using all regressor problems. For cellulose enzyme activity using DTR of MAE is 0.01, MLR of MAE is 0.0178, and RFR of MAE is 0.006. For N-acetyl-glucosaminidase using DTR of MAE is 0.0103, MLR of MAE is 0.0155, and RFR of MAE is 0.0155.

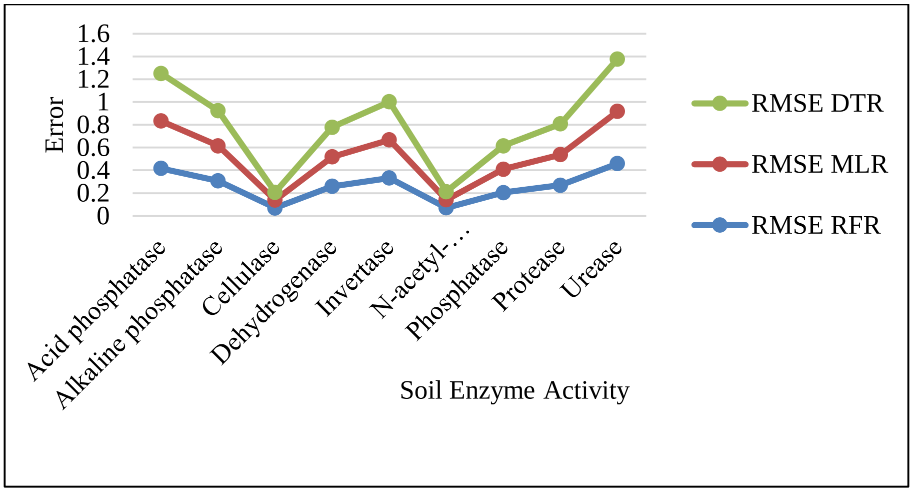

Figure 8. Root Means Squared Error for soil enzyme activity with regressor algorithms

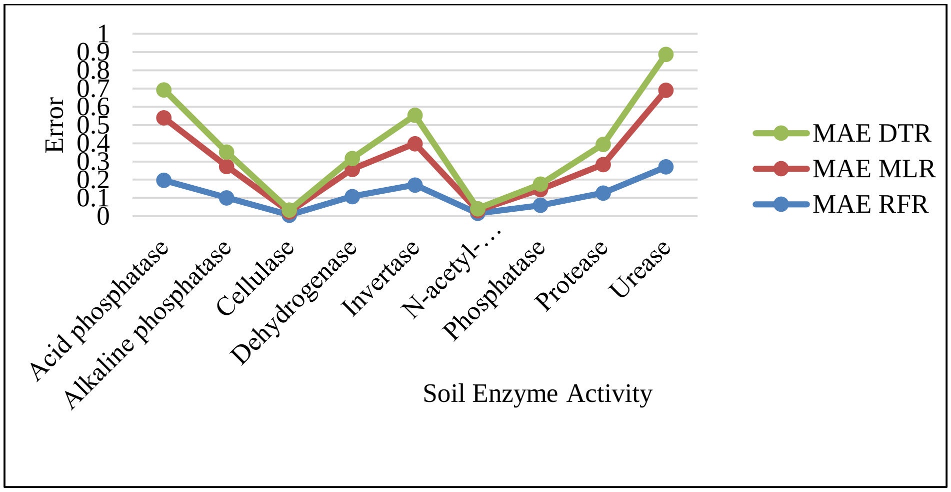

Figure 9. Mean Absolute Error for soil enzyme activity with regressor algorithms

This study used a K-means clustering unsupervised machine learning algorithm to suggest the specific crops by analysing the dependent variable noted as X variable which is including PH value, available nitrogen, available phosphorus, fertility level (low, medium, and high). K-mean clustering finds the cluster denoted as K and all the properties are derived cluster wise based on K value. K = n where n is number of cluster like (k =1, 2, 3, 4----n). Based on the k-means analysis found the k value is 5 using the elbow visualization graph shown in **Figure 10.** K-means cluster created 5 cluster consist of (k =0, 1, 2, 3, 4, 5).

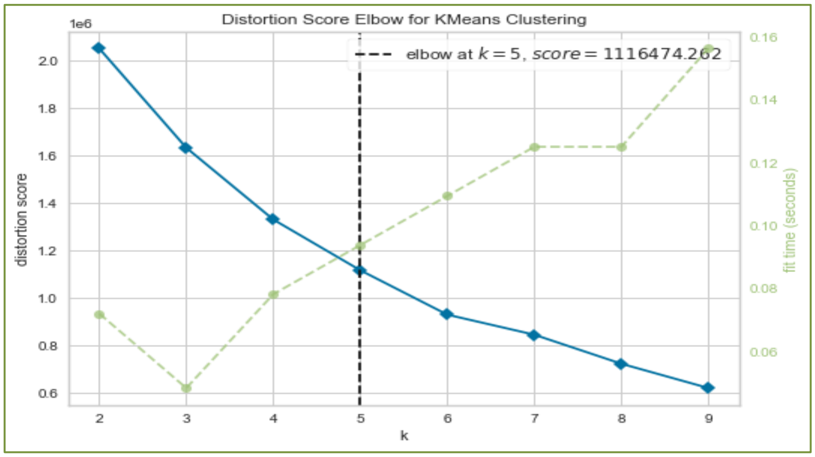

Figure 10. K Means Clustering with Elbow visualization graph

Cluster wise crops are recommended for example cluster 0: [wheat, sorghum, maize, bean, cabbage, pea, Banana, pineapple, grape], cluster 1: [Cotton, Soybean, Ragi (naachnni), Chickooo, Dates Mung Beans], cluster 2: [Tomato, Potato, Onion, Dry Beans, Cowpeas, Green Onion, Bottle Gourd, Capsium], cluster 3:[Sugarcane, Brinjal, Gilki], cluster 4 : [Rice, Black gram, Coriander, Bajara, rapeseed (Mohri)]. **Table 2**.Shows the 10 sample records of crop recommendation based on available nitrogen content, available phosphorus content, PH value, and Fertility level (low, medium, and high). In this study, the K-means clustering algorithm was used for recommending the specific crops.

Table 2. Sample records of crop recommendation

Sr. No	Crops	Available Nitrogen	Available Phosphorus	PH	Fertility Level	Cluster No.
1	Rice	274.40	15.50	6.20	2	4
2	Wheat	338.69	10.32	6.61	2	0
3	Sorghum	301.06	19.78	7.37	1	0
4	Maize	301.06	19.78	7.37	1	0
5	Sugarcane	225.79	24.08	8.06	2	3
6	Cotton	150.53	11.47	7.35	1	1
7	Soybean	150.53	11.47	7.35	2	1
8	Tomato	163.07	18.64	8.06	0	2
9	Potato	163.07	18.64	8.06	2	2
10	Onion	163.07	18.64	8.06	1	2

4. Conclusion

This research assessed the different soil enzyme activity using both soil physical properties and chemical components including sand, silt, clay, depth, available nitrogen, available phosphorus, soil organic matter, soil organic carbon, PH value, and soil fertility level such as low, medium, and high. The effect of each soil enzyme activity based on selected soil properties was assessed in Bhandara district Maharashtra. The following conclusions of the conducted research are as follows:

1. Collected the laboratory testing soil sample dataset from Maharashtra state. 5000 soil samples are collected to predict each soil enzyme's activities including both soil's physical as well as chemical properties. Apply various machine learning algorithms such as random forest, extra tree, AdaBoost, support vector, K-nearest, logistic, ridge, and decision tree algorithms for calculating the efficient accuracy using three validation techniques such as HoldOut, K-Fold, and Stratified validation techniques.

2. Compare all the machine learning algorithms and artificial neural networks for calculating better accuracy using classifier algorithm, and also calculate to measure the optimum error using evaluation techniques like means squared error(MSE), root means squared error(RMSE), and mean absolute error(MAE) by regressor algorithm. For cellulose, N-acetyl-glucosaminidase enzyme activity using RF, Extra tree, and Adaboost algorithm was better accuracy (99%) using holdout, and K-fold, and stratified validation approach. Suggested the specific crops based on soil properties using a k-means unsupervised machine learning algorithm.

References

[1] S. K. Jha and Z. Ahmad, “Soil microbial dynamics prediction using machine learning regression methods,” *Comput. Electron. Agric.*, vol. 147, no. October 2017, pp. 158–165, 2018, doi: 10.1016/j.compag.2018.02.024.

[2] H. Aponte *et al.*, “Meta-analysis of heavy metal effects on soil enzyme activities,” *Sci. Total Environ.*, vol. 737, p. 139744, 2020, doi: 10.1016/j.scitotenv.2020.139744.

[3] H. Zhang, P. Wu, A. Yin, X. Yang, M. Zhang, and C. Gao, “Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model,” *Sci. Total Environ.*, vol. 592, pp. 704–713, 2017, doi: 10.1016/j.scitotenv.2017.02.146.

[4] S. N. Kivlin and K. K. Treseder, “Soil extracellular enzyme activities correspond with abiotic factors more than fungal community composition,” *Biogeochemistry*, vol. 117, no. 1, pp. 23–37, 2014, doi: 10.1007/s10533-013-9852-2.

[5] C. Liu *et al.*, “Soil Enzyme Activities and Their Relationships With Soil C, N, and P in Peatlands From Different Types of Permafrost Regions, Northeast China,” *Front. Environ. Sci.*, vol. 9, no. May, pp. 1–12, 2021, doi: 10.3389/fenvs.2021.670769.

[6] K. Bogati and M. Walczak, “The Impact of Drought Stress on Soil Microbial Community, Enzyme Activities and Plants,” *Agronomy*, vol. 12, no. 1, pp. 1–26, 2022, doi: 10.3390/agronomy12010189.

[7] M. Ebrahimi, M. R. Sarikhani, J. Shiri, and F. Shahbazi, “Modeling soil enzyme activity using easily measured variables: Heuristic alternatives,” *Appl. Soil Ecol.*, vol. 157, no. December 2019, p. 103753, 2021, doi: 10.1016/j.apsoil.2020.103753.

[8] D. W. Bergstrom, C. M. Monreal, A. D. Tomlin, and J. J. Miller, “Interpretation of soil enzyme activities in a comparison of tillage practices along a topographic and textural gradient,” *Can. J. Soil Sci.*, vol. 80, no. 1, pp. 71–79, 2000, doi: 10.4141/S99-034.

[9] M. Szostek, E. Szpunar-Krok, R. Pawlak, J. Stanek-Tarkowska, and A. Ilek, “Effect of Different Tillage Systems on Soil Organic Carbon and Enzymatic Activity,” *Agronomy*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.3390/agronomy12010208.

[10] H. Zhang *et al.*, “Machine learning-based source identification and spatial prediction of heavy metals in soil in a rapid urbanization area, eastern China,” *J. Clean. Prod.*, vol. 273, p. 122858, 2020, doi: 10.1016/j.jclepro.2020.122858.

[11] N. S. Zungu, S. O. Egbewale, A. O. Olaniran, M. Pérez-Fernández, and A. Magadlela, “Soil nutrition, microbial composition and associated soil enzyme activities in KwaZulu-Natal grasslands and savannah ecosystems soils,” *Appl. Soil Ecol.*, vol. 155, no. November 2019, p. 103663, 2020, doi: 10.1016/j.apsoil.2020.103663.

[12] S. Tajik, S. Ayoubi, and N. Lorenz, “Soil microbial communities affected by vegetation, topography and soil properties in a forest ecosystem,” *Appl. Soil Ecol.*, vol. 149, no. January 2020, p. 103514, 2020, doi: 10.1016/j.apsoil.2020.103514.

[13] S. R. Kashiwar, D. M. C. Kundu, and U. R. Dongarwar, “Soil fertility appraisal of Bhandara block of Maharashtra using geospatial techniques,” *Int. J. Chem. Stud.*, vol. 8, no. 2, pp. 2570–2576, 2020, doi: 10.22271/chemi.2020.v8.i2am.9136.

[14] R. Recena, V. M. Fernández-Cabanás, and A. Delgado, “Soil fertility assessment by Vis-NIR spectroscopy: Predicting soil functioning rather than availability indices,” *Geoderma*, vol. 337, no.

March 2018, pp. 368–374, 2019, doi: 10.1016/j.geoderma.2018.09.049.

[15] G. A. Helfer, J. Luis, V. Barbosa, R. Santos, and A. Ben, “A computational model for soil fertility prediction in ubiquitous agriculture,” *Comput. Electron. Agric.*, vol. 175, no. June, p. 105602, 2020, doi: 10.1016/j.compag.2020.105602.

[16] K. K. T. G, C. Shubha, and S. A. Sushma, “Random Forest Algorithm for Soil Fertility Prediction and Grading Using Machine Learning,” no. 1, pp. 1301–1304, 2019, doi: 10.35940/ijitee.L3609.119119.

[17] K. Karthigadevi, “Random Forest Classification Algorithm for Agricultural Data Analysis in Tirunelveli District,” vol. XII, no. Viii, pp. 418–432, 2020.

[18] S. Dharumarajan, R. Hegde, and S. K. Singh, “Geoderma Regional Spatial prediction of major soil properties using Random Forest techniques - A case study in semi-arid tropics of South India,” *Geoderma Reg.*, vol. 10, no. April, pp. 154–162, 2017, doi: 10.1016/j.geodrs.2017.07.005.

[19] H. P. Sahragard and M. R. Pahlavan-rad, “Prediction of Soil Properties Using Random Forest with Sparse Data in a Semi-Active Volcanic Mountain,” vol. 53, no. 9, pp. 1222–1233, 2020, doi: 10.1134/S1064229320090136.

[20] G. T. Prasanna, M. Vijayasanthi, and J. Sabeena, “Agriculture soil classification and fertilizer recommendation using Adaboost and Bagging approaches.”

[21] K. Sukhadia and M. B. Chaudhari, “A Survey on Rice Crop Yield Prediction in India Using Improved Classification Technique,” vol. 5, no. 1, pp. 501–507, 2019.

[22] M. A. Pukalchik, A. M. Katrutsa, D. Shadrin, V. A. Terekhova, and I. V. Oseledets, “Machine learning methods for estimation the indicators of phosphogypsum influence in soil,” *J. Soils Sediments*, vol. 19, no. 5, pp. 2265–2276, 2019, doi: 10.1007/s11368-019-02253-2.

[23] S. Rose, S. Nickolas, and S. Sangeetha, “Machine learning and statistical approaches used in estimating parameters that affect the soil fertility status: A survey,” *Proc. 2nd Int. Conf. Green Comput. Internet Things, ICGCIoT 2018*, pp. 381–385, 2018, doi: 10.1109/ICGCIoT.2018.8753025.

[24] S. Chakraborty *et al.*, “Rapid estimation of compost enzymatic activity by spectral analysis method combined with machine learning,” *Waste Manag.*, vol. 34, no. 3, pp. 623–631, 2014, doi: 10.1016/j.wasman.2013.12.010.

[25] X. Xie *et al.*, “Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land,” *Ecol. Indic.*, vol. 120, no. September 2020, p. 106925, 2021, doi: 10.1016/j.ecolind.2020.106925.

[26] H. S. K. Pinheiro, W. De Carvalho Junior, C. D. S. Chagas, L. H. C. Dos Anjos, and P. R. Owens, “Prediction of topsoil texture through regression trees and multiple linear regressions,” *Rev. Bras. Cienc. do Solo*, vol. 42, pp. 1–21, 2018, doi: 10.1590/18069657rbcs20170167.

[27] R. Chaudhari, S. Chaudhari, A. Shaikh, and R. Chiloba, “Soil Fertility Prediction Using Data Mining Techniques,” vol. 21, no. 01, pp. 20–27, 2020.

[28] R. G. Devi, “Improved classification techniques by combining KNN and Random Forest with Naive Bayesian Classifier,” no. March, pp. 1–4, 2015.

[29] J. R and S. D. M, “Predictive model construction for prediction of soil fertility using decision tree machine learning algorithm,” *Kongunadu Res. J.*, vol. 8, no. 1, pp. 30–35, 2021, doi: 10.26524/krj.2021.5.

[30] V. Bhuyar, “Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District,” *Int. J. Emerg. Trends e Technol. Comput. Sci.*, vol. 3, no. 2, pp. 200–203, 2014, [Online]. Available: http://www.ijettcs.org/Volume3Issue2/IJETTCS-2014-04-23-111.pdf.

[31] Priya, Muthaiah, and Balamurugan, “Predicting yield of the crop using machine learning algorithms,” *Int. J. Eng. Sci. reseach Technol.*, vol. 7, no. 4, pp. 1–7, 2018.

[32] V. Pandith, H. Kour, S. Singh, J. Manhas, and V. Sharma, “Performance Evaluation of Machine Learning Techniques for Mustard Crop Yield Prediction from Soil Analysis,” *J. Sci. Res.*, vol. 64, no. 02, pp. 394–398, 2020, doi: 10.37398/jsr.2020.640254.

[33] S. Tajik, S. Ayoubi, and F. Nourbakhsh, “Prediction of soil enzymes activity by digital terrain analysis: Comparing artificial neural network and multiple linear regression models,” *Environ. Eng. Sci.*, vol. 29, no. 8, pp. 798–806, 2012, doi: 10.1089/ees.2011.0313.

[34] A. Piotrowska-Długosz, M. Kobierski, and J. Długosz, “Enzymatic activity and physicochemical properties of soil profiles of luvisols,” *Materials (Basel).*, vol. 14, no. 21, 2021, doi: 10.3390/ma14216364.

[35] S. Jian *et al.*, “Soil extracellular enzyme activities, soil carbon and nitrogen storage under nitrogen fertilization: A meta-analysis,” *Soil Biol. Biochem.*, vol. 101, pp. 32–43, 2016, doi: 10.1016/j.soilbio.2016.07.003.