

Intelligent Crop Selection: Leveraging Machine Learning for Soil-Informed Agriculture

Kranti Sapkal, Avinash Kadam

*Department of Computer Science & Engineering, Sant Gadge Baba Amravati University,
Amravati, Maharashtra, India*

Email: krantisapkal2006@gmail.com

Crop recommendation is a crucial aspect of modern agriculture, enabling farmers to make informed decisions for optimal crop selection and improved yields. Technology driven agricultural process is gaining the boom as it is proving more beneficial in terms of in-creased yield production and reduced manual efforts. This research focuses on the devel-opment of a crop recommendation system based on soil data using Machine Learning algo-rithms. The methodology involves data collection, analysis, and the implementation of the various machine learning algorithms such as k-means clustering, Decision tree, Random Forest and Logistic regression to predict suitable crops for given soil conditions. The accu-racy of the model is evaluated, and the main findings of the research are presented.

Keywords: Agriculture, Machine Learning, Random Forest, Decision tree.

1. Introduction

Agriculture is the primary source to feed the growing population of the world and the selection of appropriate crops is vital for agricultural productivity and sustainability.

Climate change and natural disasters are main environmental factors limiting the crop yield production and affecting the overall ratio of production and need [1]. Choice of the crop for cultivation should base on the soil contents, soil type and climate. To gain more yield productivity and farmers should aware these crucial things.

Several Crop recommendation systems leveraging soil data and machine learning algorithms have gained prominence for assisting farmers in making data-driven choices [2]. Recent technological development using Machine Learning and AI in agriculture field is the key part to enhance the farming practices. Machine Learning techniques are data driven, intelligent and

having capability to learn itself without being explicitly programmed. It is classified as Supervised, Unsupervised and Reinforcement Learning [3].

This research aims to build a robust crop recommendation system using the various machine learning algorithms such as k-means clustering, Decision tree, Random Forest and Logistic regression to predict suitable crops for given soil conditions. Soil data, including pH, EC, Organic Carbon(OC), Available Nitrogen(N), Available Phosphorus(P), Potassium(K), Sulphur(S), Zinc(Zn), Boron(B), Iron(Fe), Manganese(Mn), Copper(Cu) are collected from various agricultural regions from Pune District, Maharashtra, India. Collected data is labeled using K-Means Clustering as it is initially unlabeled. Model is developed in Python and trained using various Classification algorithms such as Decision Tree, Random Forest, Logistic Regression. Model performance is evaluated and results are discussed.

2 Literature Review:

A comprehensive review of existing studies related to crop recommendation systems, soil data analysis, and the application of machine learning algorithms in agriculture is presented. This review provides valuable insights into the methodologies used in similar research and lays the foundation for the proposed approach. A Crop Recommender System model is presented by Pande, Shilpa Mangesh et al.[3], using Machine Learning algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), Multivariate Linear Regression (MLR), and K-Nearest Neighbour (KNN). Random Forest showed the best results with 95% accuracy and updating the datasets from time to time to produce accurate predictions are key outcomes. S.Pudumalar, E.Ramanujamet et al. [4] proposed the crop recommendation system for precision agriculture using ensemble model with majority voting technique using Random tree, CHAID, K-Nearest Neighbor and Naive Bayes as learners to recommend a crop for the site specific parameters with high accuracy and efficiency. Y. Jia et al. [5] presented a study of Temporal-Spatial Soil Moisture Estimation using machine learning (ML) regression aided by a preclassification strategy. The total observations are classified by land types and corresponding subsets are built for constructing ML regression submodels. Ten-fold cross-validation technique is adopted. This approach has been shown to be effective for different ML algorithms, and the estimated CYGNSS SM achieved a satisfactory performance in daily and seasonal predictions. Gao, H.[6] performed Agricultural Soil Data Analysis Using Spatial Clustering Data Mining Techniques, study summarizes the characteristics of soil data and existing spatial data clustering methods, and compares spatial clustering algorithms employed in four categories of agricultural applications for soil data analysis. Raja, S. K. S.et al. [7] used sliding window non-linear regression technique to predict crop yield and price that a farmer can obtain from his land based on different factors affecting agricultural production such as rainfall, temperature, market prices, area of land and past yield of a crop. The analysis is done for several districts of the state of Tamilnadu, India. Kumar, A., Sarkar, S et al. [8] applied SVM classification

algorithm, Decision Tree algorithm and Logistic Regression algorithm for Recommendation System for Crop Identification and Pest Control Technique in Agriculture And found that SVM classification model gives the better accuracy as compared to other algorithms.

Kiruthiga, C., & Dharmarajan, K. [9] used data mining and machine learning algorithms like support vector machines (SVMs), Naive Bayes (NBs), decision trees (DTs), and linear discriminant analyses (LDAs) to make predictions and draw conclusions from agricultural data about soil-borne diseases and Crop yielding.

Parvez, R., Ahmed, T. et al.[10] applied a Multinomial Logit Model (MLM) and predictive models including Random Forest (RF), Gradient Boosting (GB), and Light Gradient Boosting Machine (LightGBM) to assess their impact on rice and maize production. Their findings indicate that nitrogen, rainfall, and humidity significantly enhance rice yields, whereas temperature and soil pH negatively affect it. For maize, nitrogen is beneficial, while potassium, temperature, rainfall, and soil pH are detrimental. Mahmud, T., Datta, N. et al. [11] proposed approach to employ a hybrid methodology, where a Genetic Algorithm is utilized to optimize the hyperparameters of the model, enhancing its performance and robustness. Authors use a Random Forest classifier, a powerful ensemble learning technique, to classify the class labels associated with 22 different types of crops giving a remarkable accuracy rate of 99.3%.

Burri, S. R., Agarwal, D. K. et al.[12] developed a model and evaluated and trained using data from the “Smart Irrigation System Dataset” made publicly available by the University of California, Irvine. A transfer-learned ResNet50 model is evaluated using various classification measures like accuracy, recall, precision, and area under the ROC curve (AUC).

3 Methodology

The prime objective of the research is to give proper recommendation regarding the crop on the basis of properties of the soil using machine learning techniques. To achieve the objective of the current research, experiments have been carried out using Python. Dataset used for this work is unlabeled. K-means clustering method of unsupervised learning is used to get the clusters of various crops. Further supervised machine learning methods are used to train the model (see Fig.1). The performance of all the trained models has been measured using performance measures namely accuracy, recall, precision, specificity and F-score. On the basis of the experimental results, comparative analysis of all the trained models has been carried out to reveal the most accurate technique for crop recommendation.

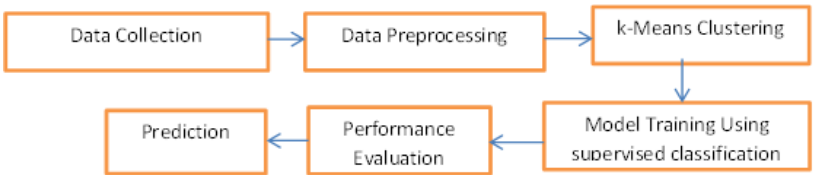


Fig. 1. Workflow of the Model

3.1 Data Collection

For current research problem, soil samples from various places of Pune districts of Maharashtra, India have been collected from Soil Testing Lab, Krushi Vikas Kendra (KVK), Malegoan, Baramati, India. This dataset consists of 10396 rows with 12 input parameters representing soil nutrient status of Pune region. Dataset does not contain output value, it is

unlabeled. Parameters of dataset are pH, EC, Organic Carbon(OC),Nitrogen(N),Phosphorus(P) , Potassium(K),Sulphur(S),Zinc(Zn),Boron(B), Iron(Fe),Manganese(Mn) ,Copper(Cu) representing the soil properties. Fig.2 shows the sample records from the dataset..

	pH	EC	Organic Carbon(OC)	N	P	K	Sulphur(S)	Zinc(Zn)	Boron(B)	Iron(Fe)	Manganese(Mn)	Copper(Cu)
3563	8.26	0.34	0.8	192.00	7.25	192.0	7.01	0.08	0.23	1.09	0.52	0.29
7431	8.18	0.63	0.9	187.00	5.87	187.0	8.1	0.67	0.46	1.6	0.46	0.66
6268	8.19	0.52	0.43	80.06	7.19	515.0	9.25	0.61	0.09	0.84	0.55	0.47
9208	8.02	0.29	0.77	111.00	9.99	188.0	9.92	0.88	0.89	1.49	0.89	0.29
4584	7.16	0.52	0.42	182.00	8.29	171.0	6.26	0.95	0.10	0.35	0.9	0.35

Fig. 2. Sample data

Contents of the each soil paramer is described in the Fig.3

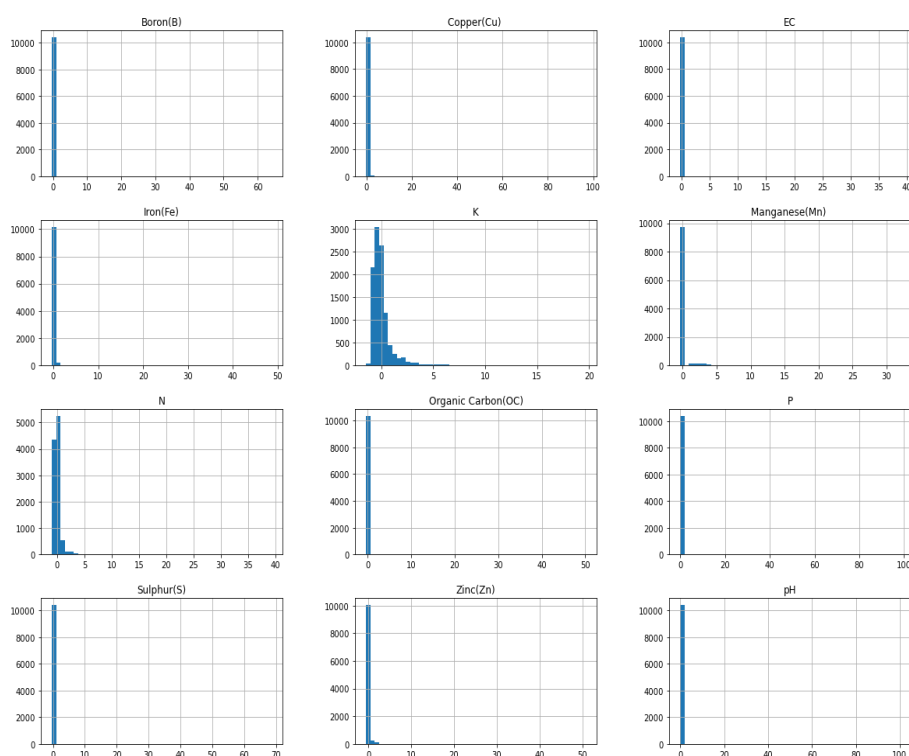


Fig. 3. Soil parameter contents

3.2 Data Preprocessing

Data preprocessing plays a key role in data analysis and machine learning phase [13].

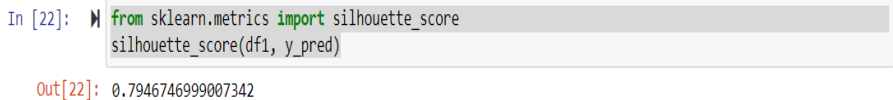
In this phase dataset is checked for missing values and missing data is handled either removing it or imputing the values [14]. Duplicate records and noise from data is removed to clean the data for further processing. Afterwards data transformation is performed to scale the values in similar range using the standard scalar of sklearn in Python

3.3 K- Means Clustering

K-means clustering is commonly used clustering technique in which the algorithm attempts to divide observations into k groups, with each group having roughly equal variance [15]. Cluster centroids represents the each group of the data. This algorithm works in two steps. First it assigns each data point to the closest centroid. In the second step it reassigns the centroid of the group by taking mean of the data points that are assigned to the group. The algorithm stops when further formed clusters make no longer changes [16]. For the experimental study, K-Means Clustering has been implemented to group the data into clusters. Optimal value of k is chosen as 3 and silhouette score metric used to evaluate the clustering technique. Its value ranges from -1 to 1 [17].

$$\text{Silhouette Score} = (b-a) / \max(a,b) \quad (1)$$

In equation (1) a is average distance between each point within a cluster and b is average distance between all clusters. For the experimental data using k=3 the silhouette score is 0.7946 as shown in Fig. 4.



```
In [22]: from sklearn.metrics import silhouette_score
         silhouette_score(df1, y_pred)

Out[22]: 0.7946746999007342
```

Fig. 4. Silhouette Score

3.4 Model Training

Decision Tree

Decision trees are commonly employed algorithms for both classification and regression purposes. In essence, they construct a hierarchy of if/else inquiries to arrive at a final decision [18]. It operates by recursively partitioning the data into subsets based on the features that best separate the target variable. At each step, the algorithm selects the feature that optimally splits the data, typically by maximizing information gain (for classification) or minimizing variance (for regression). This process continues until a stopping criterion is met, such as reaching a maximum tree depth or a minimum number of data points in each leaf node [19].

Random Forest Classifier

A random forest is a collection of decision trees, where each tree is slightly different from the others. The idea behind random forests is that each tree might do a relatively good job of predicting, but will likely overfit on part of the data [20]. If we build many trees, all of which work well and overfit in different ways, we can reduce the amount of overfitting by averaging their results

Multinomial Logistic Regression.

Multinomial logistic regression, allows to predict the probability that an observation is of a certain class. For multiclass classification, multinomial logistic regression is used where target variable can have more than two values [21]. It is implemented using Softmax function [22]. If dataset is with highly imbalanced classes and have not addressed it during preprocessing, we have the option of using the `class_weight` parameter to weight the classes to make certain we have a balanced mix of each class. Specifically, the `balanced` argument will automatically weigh classes inversely proportional to their frequency [23].

$$w_j = n/kn_j \quad (2)$$

Where w_j is the weight to class j , n is the number of observations, n_j is the number of observations in class j , and k is the total number of classes.

Support Vector Machine.

Support Vector Machines excel at classification by identifying the optimal decision boundary, which creates the widest separation in terms of hyperplane between data points belonging to different classes [24]. In our dataset, we have data containing 12 features (i.e., 12 dimensions) and a target vector with the class of each observation. Importantly, the classes are assigned such that they are linearly inseparable. That is, there is no straight line we can draw that will divide the three classes.

K Nearest Neighbor

The K-Nearest Neighbors (KNN) classifier is a straightforward yet widely used technique in supervised machine learning [25]. It is often referred to as a "lazy learner" because it doesn't build a model during the training process [26]. Instead, when it needs to make a prediction, it examines the ' k ' nearest data points and predicts the new observation's class based on the most common class among those neighbors. The algorithm uses various distance measures such as Euclidian distance measure, Manhattan distance measure, Murkowski distance measure [27].

3.5 Performance Evaluation

Performance evaluation of the model is carried out by using the metrics accuracy, precision, recall and f1 score.

Accuracy is the number of correct predictions divided by the total number of predictions [28][29].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Precision identify the number of positive predictions which are relatively correct. It is calculated as the ratio of true positives to the sum of true and false positives for each class [30][31].

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall is the capability of a classifier to discover all positive cases from the confusion matrix. It is calculated as the ratio of true positives to the sum of true positives and false negatives for each class[32].

$$Recall=TP/(TP+FN) \tag{3}$$

F1 score is a weighted harmonic mean of precision and recall, with 0.0 being the worst and 1.0 being the best. Since precision and recall are used in the computation, F1 scores are often lower than accuracy measurements [33].

$$F1\ Score = \frac{2*PR}{(P+R)} \tag{4}$$

Fig.5 and Fig. 6 depicts the performance of the models with evaluation metrics.

	Model	accuracy	precision	recall	f1score
0	Decesion Tree	0.996922	0.9792	0.9792	0.979200
1	Random Forest	0.996152	0.9644	0.9644	0.973490
2	Logistic Regression	0.996152	0.9644	0.9644	0.973490
3	SVM	0.999615	0.9950	0.9950	0.997387
4	KNN	0.999615	0.9950	0.9950	0.997387

Fig. 5. Model evaluation with all features

	Model	accuracy	precision	recall	f1score
0	Decesion Tree	0.995383	0.9592	0.9592	0.968188
1	Random Forest	0.995383	0.9544	0.9544	0.967873
2	Logistic Regression	0.995383	0.9544	0.9544	0.967873
3	SVM	0.998846	0.9850	0.9850	0.992086
4	KNN	0.998846	0.9850	0.9850	0.992086

Fig.6. Model evaluation with important features

4 Results and Discussions

The study carried out using machine learning (ML) for intelligent crop selection based on soil properties yielded promising results. We collected soil data from KVK Baramati in Pune district. This data fueled the training and testing of an ML model, specifically a Decision tree, Random Forest, Logistic regression, SVM and KNN algorithms. The model exhibited high accuracy in predicting the most suitable crops for various soil conditions with and without selecting important features. Overall classification accuracy reached 99%, demonstrating a strong ability to identify appropriate crops based on soil analysis. Furthermore, the model performed well in terms of precision, recall, F1-score for different crop categories.

5 Conclusion

The research highlighted the machine learning models such as K-means clustering, Decision tree, Random Forest, Logistic regression, SVM and KNN in facilitating intelligent crop selection tailored to soil characteristics. The combination of soil data with these machine learning algorithms can greatly improve agricultural efficiency and sustainability, offering essential decision support for farmers.

References

1. Jain, S., & Ramesh, D. (2020, February). Machine Learning convergence for weather based crop selection. In 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) (pp. 1-6). IEEE.
2. Priyadharshini, A., Chakraborty, S., Kumar, A., & Pooniwala, O. R. (2021, April). In-telligent crop recommendation system using machine learning. In 2021 5th international conference on computing methodologies and communication (ICCMC) (pp. 843-848). IEEE.
3. Pande, S. M., Ramesh, P. K., ANMOL, A., Aishwarya, B. R., ROHILLA, K., & SHAURYA, K. (2021, April). Crop recommender system using machine learning ap-proach. In 2021 5th international conference on computing methodologies and com-munication (ICCMC) (pp. 1066-1071). IEEE.
4. Pudumalar, S., Ramanujam, E., Rajashree, R. H., Kavya, C., Kiruthika, T., & Nisha, J. (2017, January). Crop recommendation system for precision agriculture. In 2016 Eighth International Conference on Advanced Computing (ICoAC) (pp. 32-36). IEEE.
5. Jia, Y., Jin, S., Chen, H., Yan, Q., Savi, P., Jin, Y., & Yuan, Y. (2021). Temporal-spatial soil moisture estimation from CYGNSS using machine learning regression with a preclassification approach. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 4879-4893.
6. Gao, H. (2021, January). Agricultural Soil Data Analysis Using Spatial Clustering Data Mining Techniques. In 2021 IEEE 13th International Conference on Computer Re-search and Development (ICCRD) (pp. 83-90). IEEE.
7. Raja, S. K. S., Rishi, R., Sundaresan, E., & Srijit, V. (2017, April). Demand based crop recommender system for farmers. In 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR) (pp. 194-199). IEEE.
8. Kumar, A., Sarkar, S., & Pradhan, C. (2019, April). Recommendation system for crop identification and pest control technique in agriculture. In 2019 International Confer-ence on Communication and Signal Processing (ICCSP) (pp. 0185-0189). IEEE
9. Kiruthiga, C., & Dharmarajan, K. (2023, January). Machine learning in soil borne dis-eases, soil data analysis & crop yielding: a review. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE) (pp. 702-706). IEEE.
10. Parvez, R., Ahmed, T., Ahsan, M., Arefin, S., Chowdhury, N. H. K., Sumaiya, F., & Hasan, M. (2024, May). Integrating Multinomial Logit and Machine Learning Algo-rithms to Detect Crop Choice Decision Making. In 2024 IEEE International Conference on Electro Information Technology (eIT) (pp. 525-531). IEEE.
11. Mahmud, T., Datta, N., Chakma, R., Das, U. K., Aziz, M. T., Islam, M., ... & Anders-son, K. (2024). An approach for crop prediction in agriculture: Integrating genetic algo-rithms and machine learning. *IEEE Access*.
12. Burri, S. R., Agarwal, D. K., Vyas, N., & Duggar, R. (2023, July). Optimizing Irrigation Efficiency with IoT and Machine Learning: A Transfer Learning Approach for Accu-rate Soil Moisture Prediction. In 2023 World Conference on Communication & Compu-ting (WCONF) (pp. 1-6). IEEE.
13. Kumar, Y. J. N., Spandana, V., Vaishnavi, V. S., Neha, K., & Devi, V. G. R. R. (2020, June). Supervised machine learning approach for crop yield prediction in agriculture sector. In 2020 5th International Conference on Communication and Electronics Sys-tems (ICCES) (pp. 736-741). IEEE.
14. Sood, K., Shamsher, S., Thirumalaisamy, M., Tyagi, P., & Suvarna, N. (2022, April). A novel methodology based soil characteristic analysis using machine learning tech-niques. In 2022 2nd International conference on advance computing and innovative technologies in engineering (ICACITE) (pp. 965-973). IEEE.
15. Khaledian, Y., & Miller, B. A. (2020). Selecting appropriate machine learning methods for digital soil mapping. *Applied Mathematical Modelling*, 81, 401-418.
16. Huang, Y., Srivastava, R., Ngo, C., Gao, J., Wu, J., & Chiao, S. (2023). Data-Driven Soil Analysis and Evaluation for Smart Farming Using Machine Learning Approaches. *Agriculture*, 13(9), 1777.
17. Modi, D., Sutagundar, A. V., Yalavigi, V., & Aravatagimath, A. (2021, October). Crop recommendation using machine learning algorithm. In 2021 5th International Confer-ence on Information Systems and Computer Networks (ISCON) (pp. 1-5). IEEE.
18. Kudale, G. A., & Rajpoot, S. S. (2023, December). Enhancing Academic Performance Prediction Through K-Means Clustering and Comparative Evaluation of Machine Learning Algorithms: A Case Study on Student Dataset. In *International Conference on Business Data Analytics* (pp. 272-286). Cham: Springer Nature Switzerland.

19. Deng, P., Gao, Y., Mu, L., Hu, X., Yu, F., Jia, Y., ... & Xing, B. (2023). Development potential of nanoenabled agriculture projected using machine learning. *Proceedings of the National Academy of Sciences*, 120(25), e2301885120.
20. Motwani, A., Patil, P., Nagaria, V., Verma, S., & Ghane, S. (2022, January). Soil analysis and crop recommendation using machine learning. In *2022 International Conference for Advancement in Technology (ICONAT)* (pp. 1-7). IEEE.
21. Pandith, V., Kour, H., Singh, S., Manhas, J., & Sharma, V. (2020). Performance evaluation of machine learning techniques for mustard crop yield prediction from soil analysis. *Journal of scientific research*, 64(2), 394-398.
22. Attaluri, S. S., Batcha, N. K., & Mafas, R. (2020). Crop plantation recommendation using feature extraction and machine learning techniques. *Journal of Applied Technology and innovation*, 4(4), 1.
23. Gosai, D., Raval, C., Nayak, R., Jayswal, H., & Patel, A. (2021). Crop recommendation system using machine learning. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 7(3), 558-569.
24. Albon, C. (2018). *Machine learning with Python cookbook: Practical solutions from preprocessing to deep learning*. O'Reilly Media
25. Suchithra, M. S., & Pai, M. L. (2020). Improving the prediction accuracy of soil nutrient classification by optimizing extreme learning machine parameters. *Information processing in Agriculture*, 7(1), 72-82.
26. Hengl, T., Miller, M. A., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., ... & Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific reports*, 11(1), 6130.
27. Pham, V., Weindorf, D. C., & Dang, T. (2021). Soil profile analysis using interactive visualizations, machine learning, and deep learning. *Computers and Electronics in Agriculture*, 191, 106539.
28. Singh, V., Sarwar, A., & Sharma, V. (2017). Analysis of soil and prediction of crop yield (Rice) using Machine Learning approach. *International Journal of Advanced Research in Computer Science*, 8(5).
29. Li, Y., Rahardjo, H., Satyanaga, A., Rangarajan, S., & Lee, D. T. T. (2022). Soil database development with the application of machine learning methods in soil properties prediction. *Engineering Geology*, 306, 106769.
30. Parvez, R., Ahmed, T., Ahsan, M., Arefin, S., Chowdhury, N. H. K., Sumaiya, F., & Hasan, M. (2024, May). Integrating Multinomial Logit and Machine Learning Algorithms to Detect Crop Choice Decision Making. In *2024 IEEE International Conference on Electro Information Technology (eIT)* (pp. 525-531). IEEE.
31. Ahmed, I. A., Talukdar, S., Baig, M. R. I., Ramana, G. V., & Rahman, A. (2024). Quantifying soil erosion and influential factors in Guwahati's urban watershed using statistical analysis, machine and deep learning. *Remote Sensing Applications: Society and Environment*, 33, 101088.
32. Jain, S., Sethia, D., & Tiwari, K. C. (2024). A critical systematic review on spectral-based soil nutrient prediction using machine learning. *Environmental Monitoring and Assessment*, 196(8), 699.
33. Roopashree, S., Anitha, J., Challa, S., Mahesh, T. R., Venkatesan, V. K., & Guluwadi, S. (2024). Mapping of soil suitability for medicinal plants using machine learning methods. *Scientific Reports*, 14(1), 3741.