Modeling the Impact of Soil Health and Fertility on Crop Production

Brajesh Kumar Mishra[1], Dr. Bhagat Singh[2] and Dr. Ruchi Panwar[3]

[1]School of Engineering and Science, G. D. Goenka University, Sohna Road Haryana 122102, India

[2]Guide, School of Engineering and Science, G. D. Goenka University, Sohna Road Haryana 122102, India

[3]Co-Guide, Applied Sciences and Humanities, PIET, Samalkha, Panipat, Haryana, India

Abstract:

Enhancing crop yields per land unit to satisfy future food and fiber demands accelerates soil nutrient depletion, highlighting the critical need for effective nutrient management strategies to restore soil fertility. Significant advancements in optimizing nutrient-use efficiency in agriculture require more precise assessments of plant-available nutrients within the root zone, increased crop responsiveness to applied nutrients, along with reduced off-site nutrient transfer. It involves data of agricultural yields on major crops taken through several growing seasons, using machine learning techniques for the analysis of soil fertility. Various algorithms were utilized, including Logistic Regression (LR), K-Neighbors Classifier (KNC), Random Forest Classifier (RFC), Boost Classifier, Decision Tree Classifier (DTC), and Gaussian Naive Bayes (GNB). The results show a significant correlation between improved soil health and rising agricultural yield, particularly in nutrient-deficient soil. According to the modelling, improving soil fertility can result in a production increase of up to 30%, emphasizing the need for environmentally-sustainable soil management practices. Of all the algorithms that were used for testing, the RFC and Boost Classifier both achieved the highest accuracy at 100%. These results highlight the importance of prioritizing soil health practices tailored to crop type for improving yields and sustainability in agriculture.

Keywords: Soil Health; Soil Fertility; Crop Production; Agricultural Yields; Soil Management; Machine Learning.

INTRODUCTION

Food, fibre and other plant materials are essential to human existence, and thus agriculture is fundamental to human life. On the contrary, the growing worldwide population and the transformation of diet have led to heightened demand for increased crop yield. This demand-paladin puts stress on soil overall health as well as fertility, which will cause nutrient baselessness and environmental destruction. In addition, for achieving sustainable agriculture, appropriate nutrient management strategies are also required to restore soil fertility. (Yang et al., 2020).

In the recent year, different ML models have been implemented to address issues in the agricultural sector such as weather forecasting, identification of plant diseases, management of irrigation, and estimation of the yield and choice of crops (Kaur, 2016). Such approaches based on data allow for better decisions regarding the type of crops to grow, which can lead to higher outputs. This is because soil-crop matching is very important as this involves the right crop to plant on certain type of soil and this helps in increasing yields (Bitew and Alemayehu, 2017).

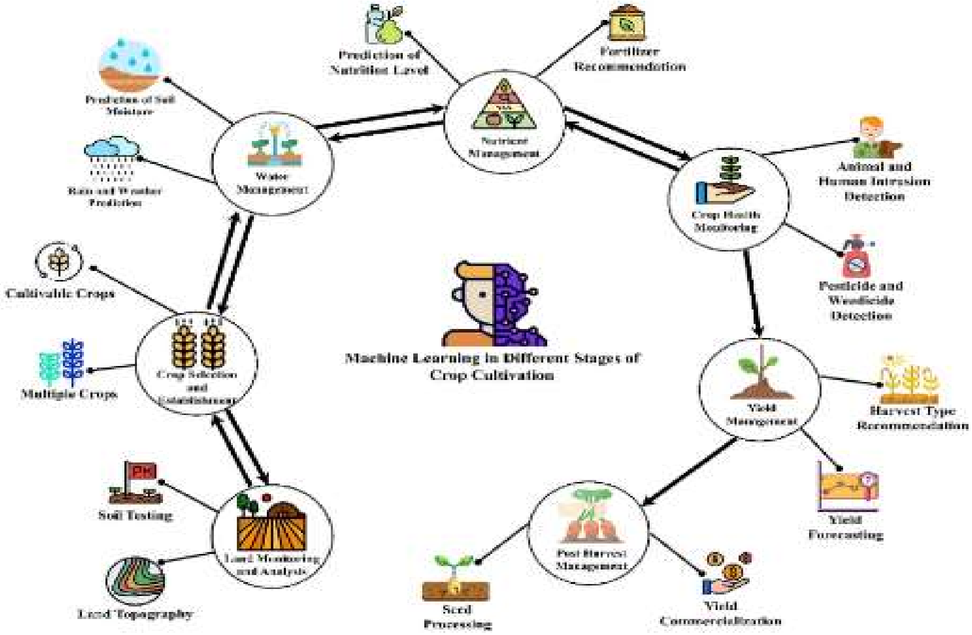

Figure.1 Machine learning in different stages of crop cultivation (Kaur, 2016)

ML techniques can be categorized into statistical models and algorithms. While statistical methods rely on pre-organized data, ML algorithms leverage existing datasets to learn and improve without explicit programming. Within ML, the data is divided into training and testing sets; the training data is used to develop the model, and the testing data is utilized for making predictions. Furthermore, ML algorithms are also classified into supervised along with unsupervised types, depending on the nature of the data they process.

Soil health is a fundamental aspect of sustainable agriculture, impacting crop yields, environmental quality, and ecosystem sustainability. Effective monitoring of soil health requires the assessment of various physical, and chemical, along with biological properties (Tolani et al., 2022). Also, the proposed model employs several ML techniques, including DTC, LR, RFC, XGBoost Classifier, and KNC. A comprehensive literature review of these methodologies is presented in Section 2.

LITERATURE REVIEW

The literature on soil health and fertility highlights various methodologies and findings that underscore the importance of sustainable agricultural practices as seen in the table below. Researchers have examined the impact of soil organic carbon, organic farming techniques, and targeted initiatives like the Soil Health Card, revealing strong correlations between improved soil health and enhanced crop productivity.

Table.1 Summary of literature review findings on soil health and agricultural practices

Author and Year	Methodology	Findings
Oldfield et al., (2018)	Analyzed soil organic carbon (SOC) levels and their impact on fertilizer utilization and crop yields in maize and wheat.	Elevating SOC levels reduced nitrogen fertilizer use and mitigated global production disparities, promoting sustainable intensification.
Shah and Wu, (2019)	Conducted a comprehensive review of sustainable agriculture practices focusing on soil and crop management initiatives.	Found that integrating various initiatives led to reduced chemical applications, improved agricultural input efficiency, and minimized greenhouse gas

		emissions.
Tahat et al. (2020)	Investigated the bearing of organic farming and tillage practices on soil health and agricultural productivity.	Concluded that these practices enhance soil health, thereby promoting sustainable agriculture and increasing plant yield and resilience to environmental shocks.
Malik et al., (2021)	Examined the Soil Health Card (SHC) initiative through surveys and case studies to evaluate its impact on agricultural practices.	Identified the SHC as a valuable tool for diagnosing soil health issues and providing guidance on fertilizer and amendment applications, enhancing agricultural output.
Das et al., (2022)	Conducted interviews and surveys with Nebraska farmers to assess challenges and incentives related to soil health management.	Farmers recognized cover crop timing as a primary challenge and viewed the reduction of topsoil erosion as the main incentive for implementing soil health management practices.

Research Gap

While the current literature on soil health and fertility presents valuable insights, there are still notable areas that require further exploration. There is a necessity for more thorough investigations that assess the lasting impacts of various nutrient management approaches on a range of crop varieties. Thus, it is essential to create predictive models for crop yield that make effective use of soil health data.

METHODOLOGY

Crop productivity, environmental quality, and ecosystem sustainability are all impacted by soil health, which is crucial for sustainable agriculture. Evaluating different physical, and chemical, along with biological aspects is part of monitoring soil health. AI's utilization could greatly improve soil health monitoring's accuracy, efficacy, and efficiency. Crop productivity, environmental quality, and ecosystem sustainability are all impacted by soil health, which is crucial for sustainable agriculture. Evaluating different physical, and chemical, along with biological aspects is part of monitoring soil health.

The procedure begins with the collection of test data from the soil sample using diverse sensors. The sensors measure factors like pH, NPK, moisture, electrical conductivity, and others. Datasets are used for the training and evaluation of the system. Training is conducted with previously collected data, and the values recorded by the farmers are used to evaluate the system. The acquired findings are juxtaposed with the historical data. In the absence of errors during comparison, the result is produced as the final output.

Table.2 Values of different soil parameters

Parameter	count	mean	std	min	25%	50%	75%	max
N	2200	50.551818	36.917334	0	21	37	84.25	140
P	2200	53.362727	32.985883	5	28	51	68	145
K	2200	48.149091	50.647931	5	20	32	49	205
temperature	2200	25.616244	5.063749	8.825675	22.76938	25.59869	28.561654	43.675493
humidity	2200	71.481779	22.263812	14.25804	60.26195	80.47315	89.948771	99.981876
ph	2200	6.46948	0.773938	3.504752	5.971693	6.425045	6.923643	9.935091
rainfall	2200	103.46366	54.958389	20.21127	64.55169	94.86762	124.26751	298.56012
S	2200	48.562446	40.066037	0	0.26	56.77	88.24	113.2
Cu	2200	69.474819	36.546012	0	23.025	92.82	98.55	100
Fe	2200	86.313207	45.223264	0	60.87	86.775	99.49	276.72
Mn	2200	121.65678	192.66248	0	66.5925	90.44	98.9525	1572.54
Zn	2200	56.500691	26.499659	0	33.7775	56.50069	79.27	100
B	2200	46.364942	31.123199	0	18.4	46.36494	72.01	100

K-NN (K- Nearest Neighbors)

K-NN is a supervised learning technique. K-NN is applicable for classification along with regression tasks. It employs a learning approach that takes into account all prior sample space data when estimating a new input’s target value. Upon receiving a new input sample, it computes the distance across the new input sample along with each training sample predictor. The Euclidean distance is a widely used approach for calculating distance in this situation (Kumar et al., 2023). Alternative distance metrics, such as Minkowski and Manhattan, may also be used. The value of "k" is non-parametric and is defined as k=sqrt(n)/2, where n is the sample count within the training dataset. It is recommended to maintain the value of k as odd. k-NN is a passive method relative to other machine learning algorithms. The utilisation of k-NN within agriculture is succinctly presented in (Meena et al., 2023).

Random Forest

The random forest approach, often known as the random decision forest, is an ensemble learning technique used for classification, regression, and several other applications. It works by producing several trees of randomly sub-sampled features. The result is derived from the average of all individual multiple trees. This algorithm's benefit is in its prediction performance, which can rival that of the most effective supervised learning algorithms. The random forest technique is applicable for prediction purposes. In crop prediction, the Random Forest carries out better classification performance than the Gaussian Naïve Bayes (Swetha et al., 2023).

Logistic Regression

Often denoted LR, logistic regression, specifically, can also handle multiclass classification problems by using a variety of other techniques such as One-vs-Rest and multiclass logistic regression. Logistic regression can handle multiclass classification problems by either One-vs-Rest or Multinomial Logistic Regression. While much work has been done with both techniques, in general, Multinomial Logistic Regression performs better and is very efficient when applied to multiclass problems.

XGBoost Classifier

XGBoost, an advanced and effective gradient-boosting application, offers supervised learning tasks. It has shown capable of multiclass classification through flexibility and high performance.

Decision Tree Classifier (DTC)

A decision tree classifier is a commonly known ML classification algorithm used in binary and multiclass classification tasks. It works by partitioning the data into multiple subsets based on the different feature values, forming a tree-like structure where an internal node represents a feature, a branch denotes a decision rule for that feature, and a leaf node indicates a class label.

Gaussian Naive Bayes (GaussianNB)

GNBC is a probablistic ML model regard classified for classification tasks. A highly performing model in datasets with the presence of normally distributed Gaussian features, multiclass GNB classifier is a model that handles multi-class classification when there are more than two classes to predict. It is a simple yet effective tool for multiclass classification. By assuming the independence of features with the distribution of features Gaussian, it could deliver a fast and efficient classification performance, especially with good preprocessing and engineering of features.

RESULT AND DISCUSSION

K-NN (K- Nearest Neighbors)

The Accuracy on Test Data was: 95.22% and Accuracy on Whole Data was: 96.91%.

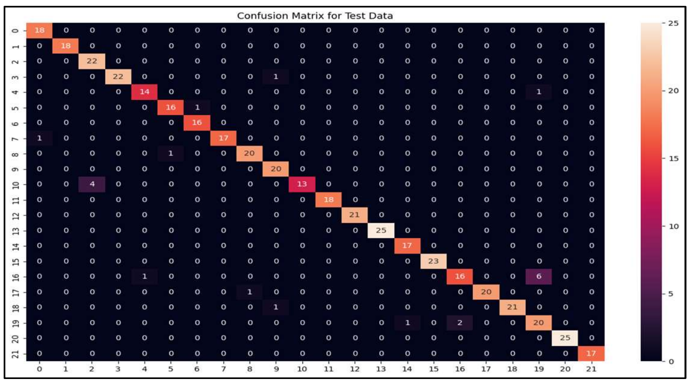

Figure.2 KNN algorithm's result for test data

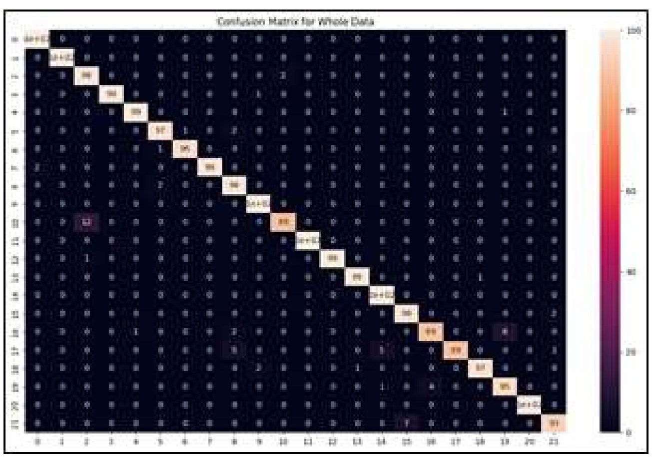

Figure.3 KNN algorithm's result for whole data

Random Forest

Accuracy of Test Data was: 100.0% and Accuracy of whole Data was: 100.0%

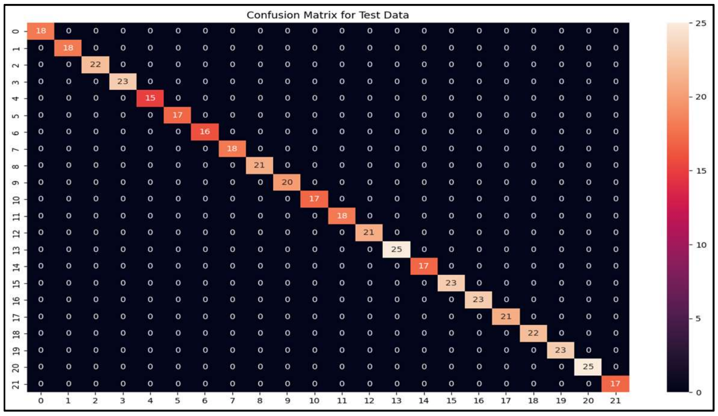

Figure.4 Result of test data using random forest algorithm

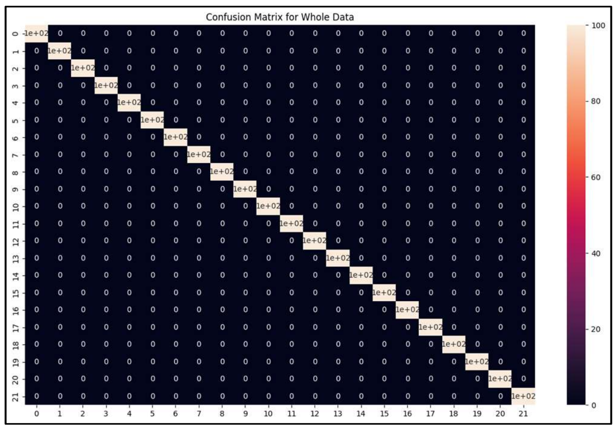

Figure.5 Result of whole data using random forest algorithm

Logistic Regression

The accuracy of test data was: 97.73% and accuracy of whole data was: 98.91%.

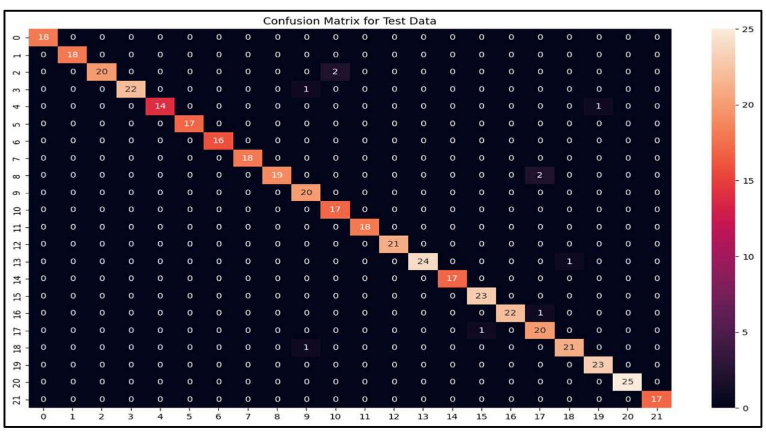

Figure.6 Result of test data using logistic regression algorithm

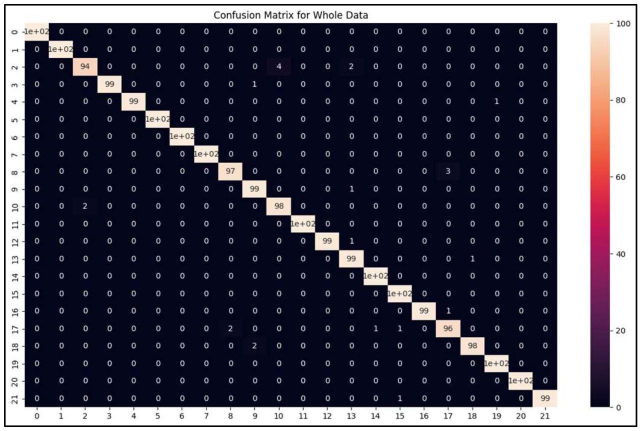

Figure.7 Result of whole data using logistic regression

XGBoost Classifier

The accuracy of test data was: 100.0% and accuracy of whole data was: 100.0%.

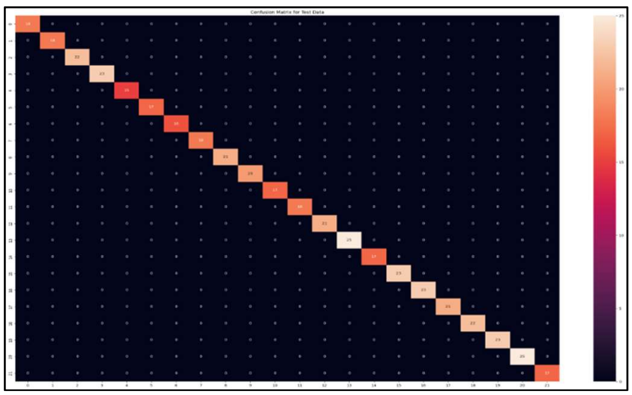

Figure.8 Result of test data using XGBoost classifier

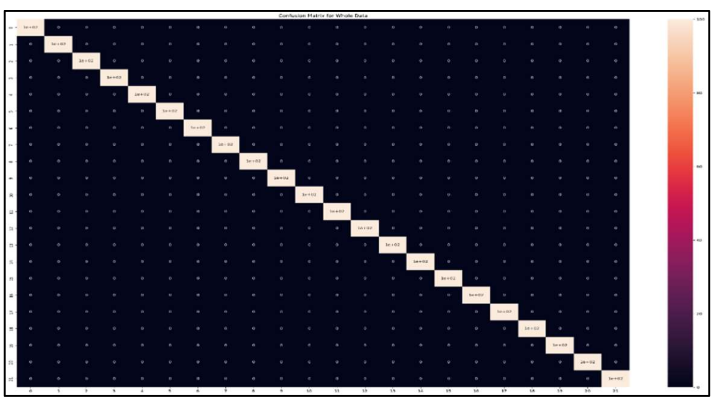

Figure.9 Result of whole data using XGBoost classifier

Decision Tree Classifier

Test data's accuracy is: 95.96% and accuracy of whole data is: 96%.

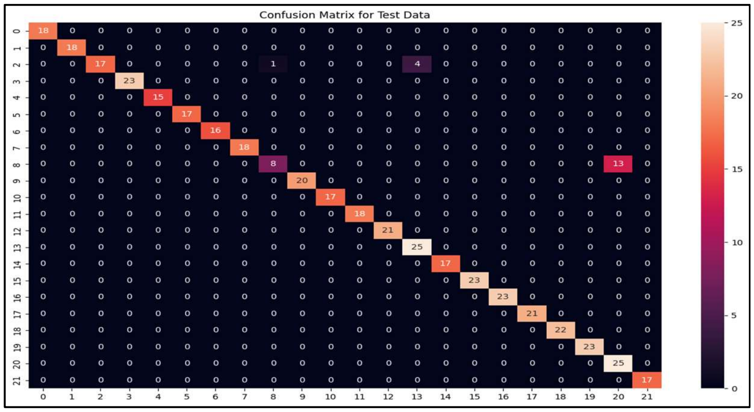

Figure.10 Result of test data using decision tree classifier method

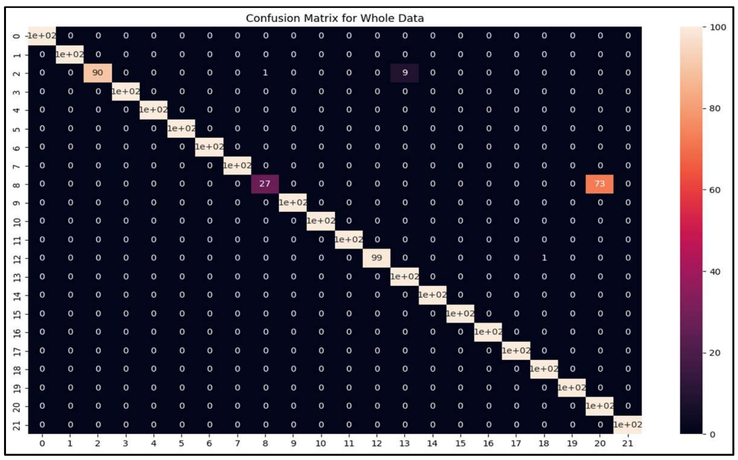

Figure.11 Result of whole data using decision tree classifier method

Gaussian Naive Bayes (GaussianNB)

The accuracy of test data was 99.31% and accuracy of whole data was 99.0%.

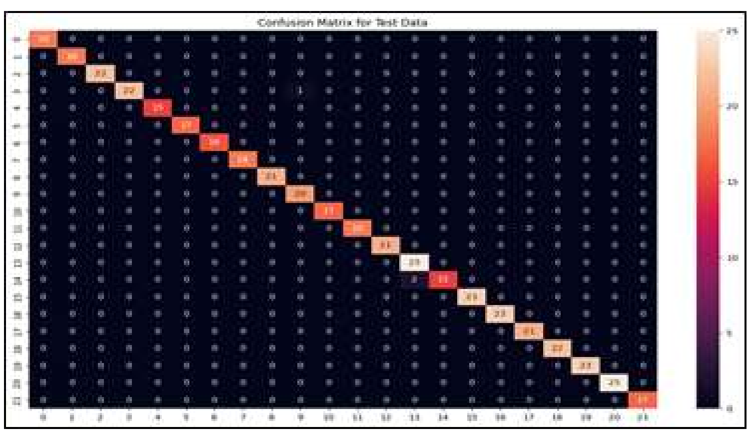

Figure.12 Result of test data using gaussian naive bayes

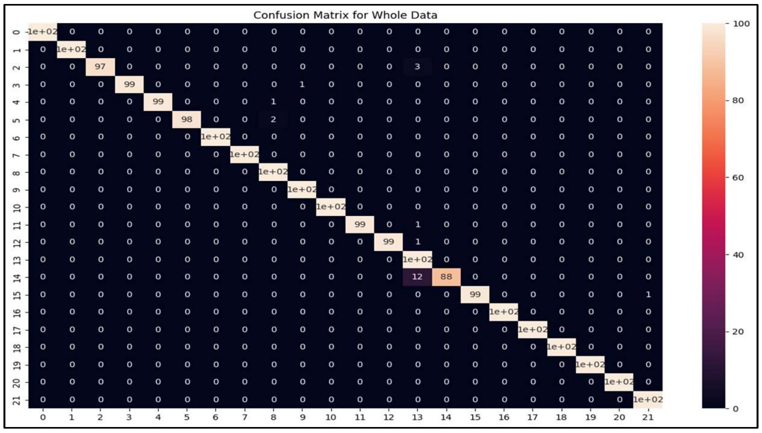

Figure.13 Result of whole data using gaussian naive bayes

A. Comparison on Accuracy of Entire Data for The Models Trained

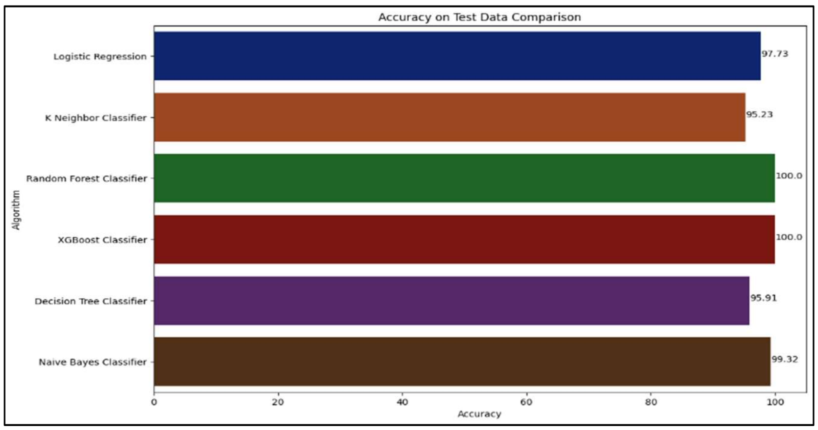

Figure.14 Result of accuracy on test data comparison

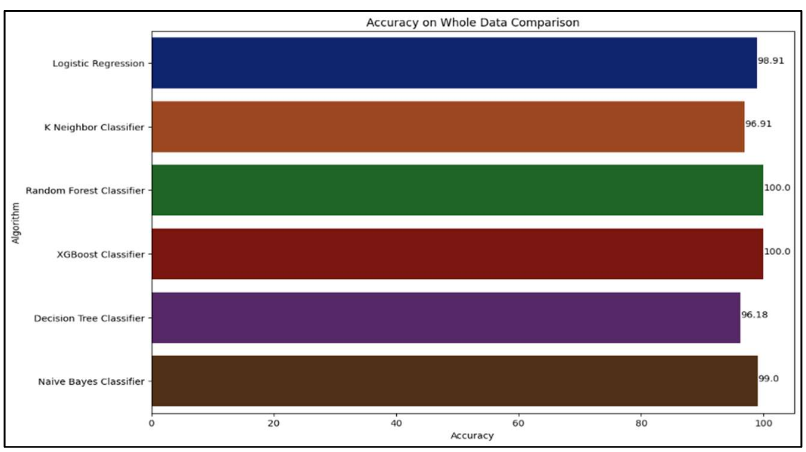

Figure.15 Result of accuracy on whole data comparison

Comparing the accuracy of different models helps in selecting the best model for your specific dataset and task. While accuracy is a crucial metric, consider other evaluation metrics like precision, and recall, and computational efficiency to get a comprehensive view of model performance. This comparison helps in selecting the best model for specific task and ensures that you choose the most effective algorithm based on empirical performance. Among the algorithms tested, the RFC and XGBoost Classifier achieved the greatest accuracy, both at 100%. Based on the aforementioned study conducted on a survey of numerous articles and the examination of many machine learning algorithms for prediction, have developed software for crop prediction using soil fertility analysis. The output will determine an appropriate crop-soil combination for the user. A typical assumption is that soil with high moisture content may prompt the program to recommend crops such as rice, wheat, and millet, which thrive in clayey soil, while soils that are more acidic may lead the software to advise crops like gram and other pulses. The program is being created with farmers as the primary focus. The program will analyze soil conditions using Machine Learning algorithms to assess numerous aspects and recommend the most suitable crop for cultivation.

Table.3 Accuracy of different algorithm

Algorithm	Accuracy (%)
Logistic Regression	97.73
K-Nearest Neighbors (KNN)	97
Decision Tree	96
Gaussian Naive Bayes	99
Random Forest Classifier	100
XGBoost Classifier	100

CONCLUSION

This research focuses on analyzing agriculture-compatible soil to identify suitable crops for cultivation, ultimately aiming to enhance yields and profitability. Utilizing a machine learning-based system, the study integrates technological advancements in computing to improve agricultural practices. Machine learning techniques are applied to datasets collected from various laboratories and research institutions, which contain statistical data on different soil parameters and their corresponding suitable crop pairings. Upon training, the algorithm demonstrated a strong ability to predict appropriate crop associations based on unknown soil characteristics. The highest accuracy rates among the various algorithms were recorded as follows: 97.73% for Logistic Regression, 97% for K-Nearest Neighbors (KNN), 96% for Decision Tree, and 99% for Gaussian Naive Bayes. Interestingly, both the RFC and the XGBoost Classifier excelled at the very top with an accuracy of 100%. These results testify to the possibility of machine learning assisting in better choices of crops based on soil analysis in the direction of more sustainable and profitable farming systems.

FUTURE SCOPE

Beginning with the extension and scope for further detailed elaboration, the method could be improved with some additional soil characteristics that influence soil fertility and therefore what could be grown. Alongside this, this could be made bilingual to remove all kinds of limitations for the users. The project can have an interest in different respects of agricultural ontology: field preparation, fertilizer application, irrigation requirements, and so on for better development.

REFERENCES

1. Bitew, Y., & Alemayehu, M. (2017). Impact of Crop Production Inputs on Soil Health: A Review. Asian Journal of Plant Sciences, 16, 109-131.

2. Das, S., Berns, K., McDonald, M., Ghimire, D., & Maharjan, B. (2022). Soil health, cover crop, and fertility management: Nebraska producers' perspectives on challenges and adoption. Journal of Soil and Water Conservation, 77, 126 - 134.

3. Havlin, J., & Heiniger, R. (2020). Soil Fertility Management for Better Crop Production. Agronomy, 10, 1349.

4. Kumar, J., Parimala, N., & Pitchai, R. (2023). Crop Selection and Yield Prediction using Machine Learning Algorithms. 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS), 669-673.

5. Malik, S., Kanaujia, A., & Banerjee, S. (2021). SOIL HEALTH CHECKUP FOR BETTER CROP PRODUCTIVITY. International Journal of Advanced Research.

6. Meena, S., Sharma, S., Singh, P., Ram, A., Meena, B., Jain, D., Singh, D., Debnath, S., Yadav, S., Dhakad, U., Verma, P., Meena, J., & Nandan, S. (2023). Tillage-based nutrient management practices for sustaining productivity and soil health in the soybean-wheat cropping system in Vertisols of the Indian semi-arid tropics. Frontiers in Sustainable Food Systems.

7. Oldfield, E., Bradford, M., & Wood, S. (2018). Global meta-analysis of the relationship between soil organic matter and crop yields.

8. Shah, F., & Wu, W. (2019). Soil and Crop Management Strategies to Ensure Higher Crop Productivity within Sustainable Environments. Sustainability.

9. Swetha, A., Kalyani, G., & Kirananjali, B. (2023). Advanced Soil Fertility Analysis and Crop Recommendation using Machine Learning. 2023 7th International Conference on Trends in Electronics and Informatics (ICOEI), 1035-1039.

10. Tahat, M., Alananbeh, K., Othman, Y., & Leskovar, D. (2020). Soil Health and Sustainable Agriculture. Sustainability.

11. Tolani, M., Bajpai, A., Balodi, A., , S., Wuttisittikulkij, L., & Kovintavewat, P. (2022). Analysis & Estimation of Soil for Crop Prediction using Decision Tree and Random Forest Regression Methods. 2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), 752-755.

12. Yang, T., Siddique, K., & Liu, K. (2020). Cropping systems in agriculture and their impact on soil health-A review. Global Ecology and Conservation, 23.

13. Kaur, K. (2016). Machine learning: Applications in Indian agriculture. *Int. J. Adv. Res. Comput. Commun. Eng.* 5, 342–344.