IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Smart Soil Health Prediction and Agricultural Recommendation System Using Machine Learning

Advika Pophali[1], Piraji Sanap[1], Sakshi Shelke[1] , Dr Smita Bhanap[1*]
Student, Department of Data Science, Fergusson College (Autonomous), Pune, India[1]
Head Coordinator, Department of Data Science, Fergusson College (Autonomous), Pune, India[1*]

Abstract: *Soil health is one of the most important aspects in terms of agricultural productivity. Soil testing is a tedious job that is not always accurate. This leads to inefficient use of chemicals such as fertilizers. This study introduces a machine learning-based method for predicting the soil's health, suggesting the type of crops to be grown, and the type of fertilizers to be applied.*
The system uses parameters such as Nitrogen (N), Phosphorus (P), Potassium (K), pH, and Organic Carbon to compute the Soil Health Index (SHI). Random forest is a machine learning algorithm that uses multiple classifiers to predict the soil's health using the parameters mentioned above. Soil quality is classified into categories such as Poor, Moderate, Good, etc. Two classification algorithms were also applied to classify crops according to the soil's quality and to determine the type of fertilizers to be applied.
The integration of region-specific data helps in increasing the accuracy and relevance of predictions. The results show that the proposed methodology can efficiently analyze the characteristics of the soil and provide valid recommendations to enable farmers to maximize the usage of fertilizers, promote crop production, and encourage the use of environment-friendly practices.

Keywords: Soil Health, Sustainable Agriculture, Random Forest, NPK, Crop recommendation and Fertilizer classification.

I. INTRODUCTION

Machine learning in agriculture is a powerful approach to address challenges with regards to important aspects such as food security and sustainable farming. India is a country where agriculture is a major source of livelihood, so efficient soil health monitoring and accurate crop recommendation systems are essential to improve productivity and reduce risks. Data-driven techniques help us intelligent analysis of soil characteristics which support better decision-making.

But soil analysis methods that are currently used are time-consuming, expensive, and often lack actionable insights. In conventional approaches we collect soil samples and test them in laboratories, which can take a lot of time, even weeks, and even then, they do not always provide recommendations for suitable crops or fertilizers (Hossain et al., 2023; Naik et al., 2023). That is why there is a significant gap between soil testing and its practical use in farming. Moreover, inadequate soil management and lack of awareness about crop suitability are associated to serious socio-economic issues among farmers (Patil et al., 2019).

Machine learning techniques have shown promising results in tackling these issues. Random Forest Algorithms have shown high accuracy in soil fertility prediction and crop recommendation as it has ability to model complex relationships between soil parameters (M. R. et al., 2023; Kumar et al., 2019). Moreover, recent studies have highlighted the importance of artificial intelligence in enabling smart and sustainable agricultural practices (Kamilaris and Prenafeta-Boldú, 2018).

In this paper, a Smart Soil Health Prediction and Agricultural Recommendation System is proposed using machine learning techniques. The system computes a Soil Health Index (SHI), classifies soil quality, and recommends suitable

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

crops and fertilizers. By incorporating region-specific data, the proposed approach aims to assist farmers in making informed decisions, improving crop yield, and promoting sustainable agriculture.

II. LITERATURE RIEVIEW

Soil health is a multi-dimensional property that influences crop productivity, system resilience, and long-term agricultural sustainability. Key indicators such as soil pH and macronutrients—nitrogen (N), phosphorus (P), and potassium (K)—play an important role in nutrient availability and plant growth. Additionally, factors such as soil organic carbon and structure are essential for maintaining long-term soil fertility.

In practical applications, decision-support systems often prioritize easily measurable chemical properties like pH and NPK due to their cost-effectiveness and direct applicability. Recent studies have shown that combining these parameters with climatic variables such as temperature, humidity, and rainfall can effectively support crop recommendation systems [1].

For sustainable and informed agriculture, machine learning techniques are widely being used to transform soil and environmental data into actionable insights. Supervised learning models are widely applied for crop recommendation and soil classification. Among these, ensemble methods such as Random Forest have demonstrated high accuracy as their ability to capture non-linear relationships is well known(S. Ganorkar et al., 2025). Support Vector Machines (SVM) are effective for smaller datasets, while Naïve Bayes serves as a simple and efficient baseline model (R. S. Nair et al., 2025).

Recent research explores advanced approaches such as Gradient Boosting, neural networks, and deep learning models to improve prediction accuracy and scalability ("Advanced ML Models for Precision Agriculture," 2025). However, a common limitation across many studies is poor generalization, where models trained on region-specific datasets fail to perform well under different environmental conditions ("Generalization Challenges in Agricultural ML Models," 2025).

Furthermore, recent studies emphasize the importance of model interpretability in agricultural applications. Techniques such as feature importance analysis and explainable AI methods are used to enhance transparency and user trust in ML-based recommendations (H. Afzal et al., 2025).

Overall, existing research highlights the effectiveness of machine learning in soil analysis and crop recommendation. However, challenges related to generalization, interpretability, and practical applicability remain. Many existing systems do not incorporate region-specific adaptability or integrated fertilizer recommendation, limiting their real-world usability. These gaps motivate the development of a more robust and practical soil health prediction and recommendation system, as proposed in this study.

III. METHODOLOGY

In this section, the methodology used for developing the proposed Smart Soil Health Prediction and Agricultural Recommendation System has been discussed extensively. It describes the process of data collection, preprocessing, Soil Health Index (SHI) computation, machine learning model development, and the implementation of crop and fertilizer recommendation modules.

1. Data Collection: The data utilized in this study was sourced by publicly available open repositories. It contains important parameters of soil like Nitrogen (N), Phosphorus (P), Potassium (K) in (ppm), pH, and environmental parameters like Temperature (Celcius), Humidity and Rainfall, and variable that includes crops like Rice, Maize, Chickpea, Kidney Beans and many more.

There are 2,200 samples in the dataset with 22 types of crops i.e 100 samples of each crop type. These parameters are vital parameters of soil fertility and are extensively utilized in the agricultural decision-making systems.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

2. Data Preprocessing

Before any model implementation is carried out , it is necessary to first analyse and assert that the chosen data does not have any unwanted/missing values. Such values, if not handled well, can cause prediction errors and unwanted bias. Therefore, the process of data preprocessing is of paramount importance.

The collected dataset was pre-processed to ensure quality and consistency. Missing values were handled appropriately, and irrelevant or redundant features were removed.

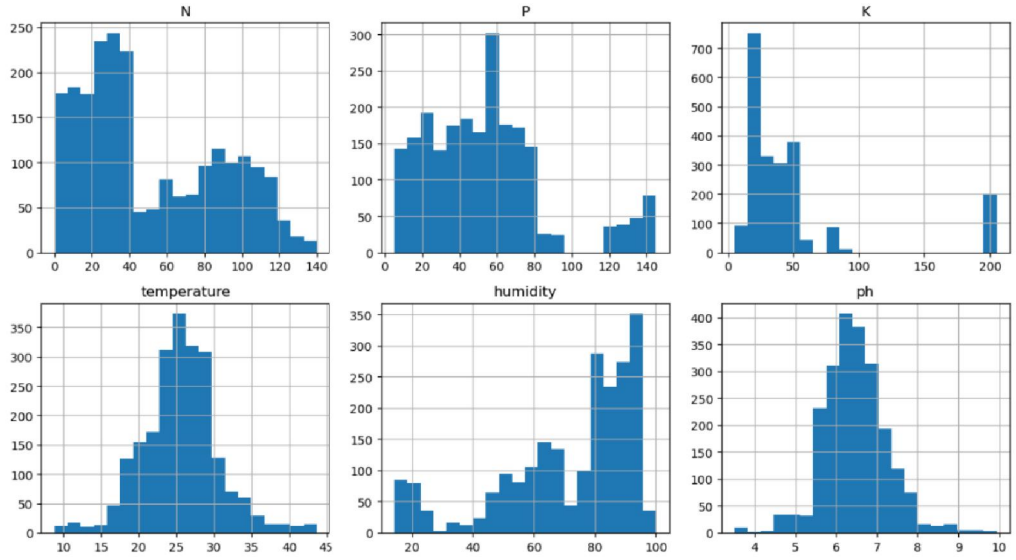

Fig1: Histogram to understand distribution of different key features

The histograms show how the values of different features are distributed in the dataset. Most features have values concentrated in a specific range, indicating common conditions, while a few values are spread toward higher or lower ends. Some features also show slight skewness, meaning the data is not perfectly symmetrical. Overall, the histograms indicate that the data is reasonably well distributed with some variation and few extreme values.

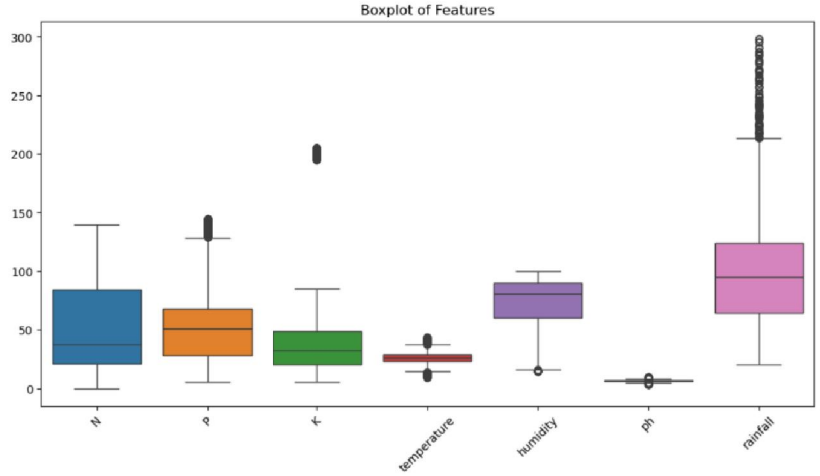

Fig2: Boxplot of features

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

The boxplot summarizes the distribution of all features (N, P, K, temperature, humidity, pH, rainfall). It shows median, quartiles, and outliers for each variable.

Most of the features in the dataset show the presence of outliers, particularly rainfall and potassium, with rainfall showing the highest variation among all features. In contrast, pH demonstrates the least variation, indicating relatively stable values across samples. Overall, while the dataset includes some extreme values, it remains generally well distributed.

2.1 Normalization of numerical attributes:

The N, P, and K values were normalized using min-max scaling, a well-known technique that transforms feature values into a fixed range between 0 and 1 by scaling them relative to their minimum and maximum values. The formula is mentioned below. This process of normalization ensures that all features contribute equally to the model and prevents bias due to differences in scale

$$XNorm = \frac{X - Xmin}{\text{Xmax} - \text{Xmin}}$$

where X represents the original value of the soil parameter (N, P, or K), Xmin and Xmax denote the minimum and maximum values of the respective parameters in the dataset.

Similarly, the pH values were normalized based on their deviation from the ideal value (7.0), using a scaling approach that assigns higher scores to values closer to neutral pH. This transformation reduces the influence of extreme pH values and reflects optimal soil conditions for most crops. The formula is mentioned below.

$$pHnorm = 1\frac{|pH - pHideal|}{\max(|pH - pHideal|)} -$$

where pH is the observed soil pH value, and pHideal represents the optimal pH value (taken as 7.0).

2.2 Encoding categorical attribute:

The target variable, which consisted of various crops, was converted into numerical form using Label Encoding. Machine learning algorithms require numerical input; therefore, categorical crop names were transformed into integer values. Label Encoding assigns a unique integer to each distinct crop category, enabling efficient model training and prediction.

In this approach, each unique category is mapped to a corresponding integer value. If C={c1,c2,...,cn} represents the set of crop labels, and mapping each category ci to an integer is done using the encoding function; i.e. ci∈{0,1,...,n−1}.

3. Soil Health Index (SHI) Calculation:

Now that all the numerical and categorical attributes necessary for model implementation were pre-processed, the next step is to calculate the Soil Health Index (SHI), which was required to predict the type of soil.

Soil Health Index (SHI) was computed to provide a unified measure of soil quality based on the normalized parameters: Nitrogen (N), Phosphorus (P), Potassium (K), and pH. After normalization, each parameter was assigned an equal weight to ensure balanced contribution.

$$SHI = \frac{1}{4}(Nnorm + Pnorm + Knorm + pHnorm)$$

The normalized SHI values were then scaled to a range of 0–100 to obtain a more interpretable soil health score:

Soil Health Score=(SHI*100)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Based on the computed score, the samples were categorized into three classes:

- Poor (Score < 40)
- Moderate (40 ⩽ Score < 70)
- Good (Score ⩾ 70)

This classification simplifies interpretation and enables users to quickly assess soil quality for decision-making.

3 Soil Health Prediction Model

A Random Forest algorithm was employed to predict soil health categories based on the input soil parameters. It is an ensemble learning technique that constructs multiple decision trees and aggregates their predictions to improve accuracy and robustness. Its ability to model complex, non-linear relationships makes it well-suited for soil data analysis.

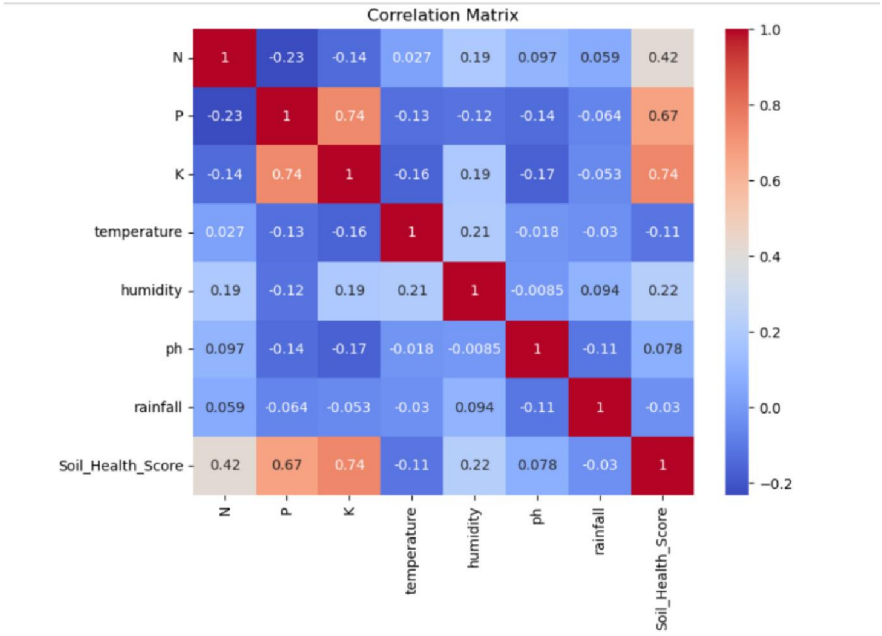

Fig3: Correlation heatmap of features

The heatmap shows the correlation between soil parameters like N, P, K, temperature, humidity, pH, rainfall, and Soil Health Score.

K (0.74) and P (0.67) have a strong positive impact on soil health. N(0.42) shows a moderate positive effect. Humidity (0.22) has a weak influence. Temperature (-0.11) and rainfall (-0.03) show very little or negative impact.

Overall, soil nutrients are the main factors affecting soil health, while environmental factors have minimal effect.

4. Crop Recommendation System

A Random Forest Classifier model was implemented to recommend suitable crops based on soil characteristics. The model takes soil parameters Nitrogen (N), Phosphorus (P), Potassium (K), and pH as input features, while the encoded crop labels were used as the target variable.

Label Encoding was applied to transform categorical crop names into numerical values, enabling compatibility with machine learning algorithms. The trained classifier predicts the most suitable crop for given soil conditions, thereby assisting farmers in making informed cultivation decisions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

Techniques commonly applied before model fitting, called the Exploratory data analysis, including correlation heatmaps and pairwise plots, were used to understand relationships between soil parameters and crop suitability.

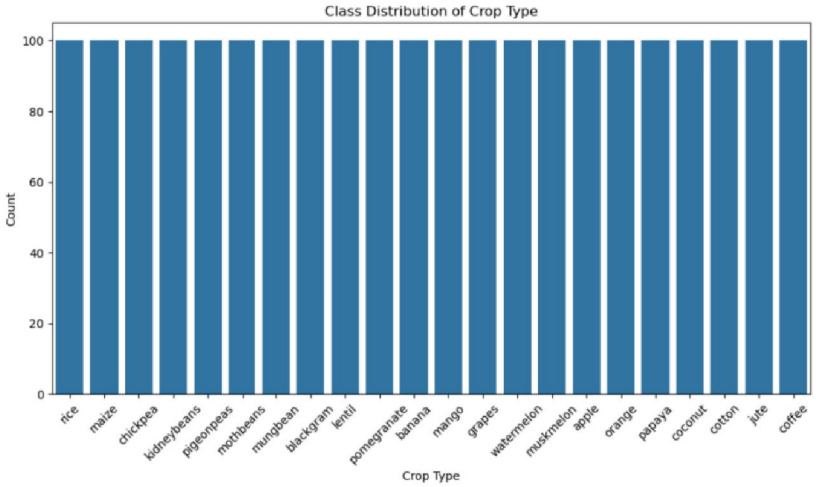

Fig4: Class Distribution of all crop types in the data

The graph represents the distribution of different crop types in the dataset. The x-axis shows various crop categories such as rice, maize, chickpea, and others, while the y-axis represents the number of samples for each crop type. Each crop appears to have an equal count of observations.

The graph indicates that the dataset is balanced, as all crop types have nearly the same number of samples. This ensures that no single crop dominates the dataset, reducing bias during model training. As a result, the machine learning model can make fair and accurate predictions for all crop categories.

5. Fertilizer Recommendation System

Similar to the crop recommendation module, a Random Forest classifier was utilized due to its robustness and ability to capture complex interactions among soil parameters. Categorical fertilizer labels were encoded into numerical form for model training.

The model analyses nutrient deficiencies in soil parameters and predicts the most suitable fertilizer to improve soil fertility. This enables targeted fertilizer usage, reducing wastage and promoting sustainable agricultural practices.

Correlation analysis was performed to examine the relationships between soil nutrients and fertilizer recommendations.

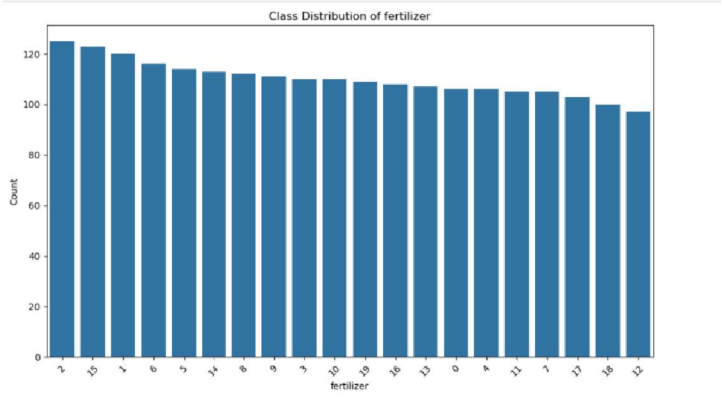

Fig5: Class Distribution of all fertilizer types in the data

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

The bar chart shows the distribution of different fertilizer classes. The x-axis represents fertilizer types, and the y-axis shows their count (frequency). Each bar indicates how many samples belong to a particular fertilizer class. The dataset is balanced, as most fertilizer classes have similar counts. No class is extremely high or low, so class imbalance is minimal. This balanced distribution is good for training machine learning models, as it avoids bias toward any specific fertilizer class.

IV. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed system, including soil health prediction, crop recommendation, and fertilizer recommendation. The applied models were verified using appropriate evaluation metrics and visualizations to analyze their accuracy, reliability, and practical applicability. The model implemented for soil health prediction was evaluated using regression metrics, namely the coefficient of determination R^2 and Root Mean Square Error (RMSE).

The R^2score represents the proportion of variance in the dependent variable that is explained by the model, with values closer to 1 indicating better performance. In this study, the model achieved an R^2 score of 0.993, indicating that it explains approximately 99.3% of the variability in soil health scores.

RMSE measures the average magnitude of prediction error, with lower values indicating higher accuracy. The obtained RMSE value of 1.06 suggests that the predicted soil health scores are very close to the actual values, demonstrating minimal error.

The crop recommendation model, implemented using a Random Forest classifier, achieved a test accuracy of 99.55%, indicating excellent predictive performance. In addition to accuracy, the model was evaluated using precision, recall, and F1-score. The following table gives the classification report achieved for the crop recommendation model

Table1: Classification Report for Crop Recommendation model

Metric	Values
Accuracy	0.99
Precision (macro avg)	0.98
Recall (macro Avg)	0.98
F1 score	0.97

The high accuracy is attributed to the ability of Random Forest to effectively model complex relationships between soil parameters and crop suitability, resulting in reliable and accurate recommendations.

The fertilizer recommendation model achieved an overall accuracy of 54.54%, which is significantly lower compared to the crop recommendation model. To further evaluate model performance, a classification report was analyzed, including precision, recall, and F1-score metrics.

Table2: Classification Report for Fertilizer Recommendation model

Metric	Values
Accuracy	0.55
Precision (macro avg)	0.56
Recall (macro Avg)	0.56
F1 score	0.55

The results indicate moderate performance across most classes, with a macro-average F1-score of approximately 0.55. While some classes achieved high precision and recall, others showed relatively lower performance, suggesting class imbalance and overlapping feature distributions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, June 2026 Impact Factor: 8.2

The comparatively lower accuracy can be attributed to the complexity of fertilizer recommendation, where multiple fertilizers may be suitable for similar soil conditions. Additionally, the dataset may contain overlapping nutrient patterns, making classification more challenging.

The proposed system effectively integrates soil health prediction, crop recommendation, and fertilizer advisory using machine learning techniques. With regards to the lower prediction score in fertilizer classification, future work involves incorporating additional features, handling class imbalance, and exploring advanced models to enhance prediction accuracy.

4.1 Deployment Using Streamlit

The developed system was deployed as an interactive web application using the Streamlit framework, providing a user-friendly interface for practical usage. The application allows users to input essential details such as farmer information (e.g., name and location) along with key soil and environmental parameters, including Nitrogen (N), Phosphorus (P), Potassium (K), temperature, humidity, pH, and rainfall.

Based on these inputs, the system performs three core functions:

- Predicts the soil health status (Good, Moderate, or Poor) using the trained Random Forest model and the computed. Soil Health Index (SHI).
- Recommends suitable crops that can be cultivated under the given soil and environmental conditions using the crop recommendation model.
- Suggests appropriate fertilizers based on nutrient deficiencies identified in the soil.

The integration of these modules into a single platform enables real-time analysis and decision support. The application presents results in a clear and interpretable format, making it accessible even to users with limited technical knowledge. Overall, the system demonstrates the practical applicability of machine learning in agriculture by assisting farmers in optimizing crop selection, improving soil management, and enhancing productivity.

V. CONCLUSION

This study presented a machine learning-based approach for soil health prediction and agricultural recommendation by integrating soil analysis with crop and fertilizer advisory systems. A Soil Health Index (SHI) was developed using key soil parameters, enabling effective classification of soil quality into interpretable categories.

The Random Forest model demonstrated excellent performance in soil health prediction, achieving a high R^2 score and low RMSE, indicating accurate and reliable predictions. Similarly, the crop recommendation system achieved very high classification accuracy, supported by strong precision, recall, and F1-score values, highlighting its effectiveness in identifying suitable crops based on soil conditions.

The fertilizer recommendation model, while achieving moderate accuracy, provided valuable insights into nutrient-based recommendations and highlighted the inherent complexity of fertilizer prediction tasks. This suggests opportunities for further improvement through enhanced feature engineering and more diverse datasets.

VI. FUTURE SCOPE

Further work can be focused on:

- Integration with mobile-based applications for easy farmer access
- Use of IoT sensors for real-time soil monitoring
- Inclusion of weather and climate data
- Implementation of deep learning techniques for improved accuracy

Furthermore, a comparative study by implementing multiple machine learning models can be conducted to evaluate their performance and identify the most effective approach for soil health prediction and agricultural recommendation.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Dr. Smita Bhanap, Head Coordinator of the Data Science program and co-author of this study, for her constant encouragement, valuable guidance, and support throughout the development of this project. Her insights and mentorship helped in shaping the direction and successful completion of this work.

The authors also extend their thanks to the Department of Data Science, Fergusson College (Autonomous), Pune, for providing the necessary resources and academic environment to carry out this research. Special appreciation is given to all faculty members for their guidance and support during this project.

Finally, the authors would like to acknowledge the use of open-source datasets and tools that contributed significantly to the successful implementation of this system.

REFERENCES

[1]. H. Afzal et al., "Incorporating soil information with machine learning for crop recommendation," Scientific Reports, 2025.
[2]. R. Sahu, V. Lowanshi, and N. Khare, "Soil health assessment and crop recommendation using deep neural networks," 2024.
[3]. H. L. Swetha et al., "Soil based crop recommendation system using machine learning," IJARCCE, 2025.
[4]. G. Kulkarni et al., "Soil fertility prediction and crop recommendation using ML algorithm," IJIERM, 2024.
[5]. S. Shariff et al., "Crop recommendation using machine learning techniques," IJERT, 2022.
[6]. A. Verma et al., "Optimized crop recommendation system using machine learning for soil analysis," 2025.
[7]. B. Dey et al., "Machine learning based recommendation of agricultural crops using NPK and climate data," Heliyon, 2024.
[8]. N. Vadivelan and K. Prasad, "Soil health analysis and classification of micronutrients for crop suggestion," 2024.
[9]. S. Bhargavi and S. Jagannathan, "Crop recommendation system using machine learning," IJERT, 2023.
[10]. Y. Mali et al., "A comparative analysis of machine learning models for soil health prediction and crop selection," 2023.

Appendix A: Implementation Details

```
In [7]: # Normalizing NPK Values
        for col in ['N', 'P', 'K']:
            col_min = dc[col].min()
            col_max = dc[col].max()

            if col_max - col_min == 0:
                dc[col + '_norm'] = 0.5   # neutral contribution
            else:
                dc[col + '_norm'] = (dc[col] - col_min) / (col_max - col_min)
```

```
In [8]: # Normalize pH values to an ideal range
        ideal_pH = 7.0
        max_dev = max(abs(dc['ph'] - ideal_pH))

        dc['pH_norm'] = 1 - (abs(dc['ph'] - ideal_pH) / max_dev)
```

Fig6: Normalization of Numerical columns

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

```
11]: # Soil health category
def soil_category(score):
    if score < 40:
        return 'Poor'
    elif score < 70:
        return 'Moderate'
    else:
        return 'Good'

dc['Soil_Category'] = dc['Soil_Health_Score'].apply(soil_category)

dc.head(20)
```

Fig7 Creating and calculating Soil Health Index

```
In [34]: # Encode target
from sklearn.preprocessing import LabelEncoder
import joblib
X = df.drop("label", axis=1)
y = df["label"]


label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)
```

Fig8 Categorization based on calculated soil health score

```
In [34]: # Encode target
from sklearn.preprocessing import LabelEncoder
import joblib
X = df.drop("label", axis=1)
y = df["label"]


label_encoder = LabelEncoder()
y_encoded = label_encoder.fit_transform(y)
```

Fig9 Label Encoding for crop recommendation

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, June 2026
Impact Factor: 8.2

```python
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report

# Create model
model2 = RandomForestClassifier(random_state=42)

# Train model
model2.fit(X_train, y_train)

# Predict
y_pred = model2.predict(X_test)

# Check accuracy
print("Accuracy:", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

Fig10 Implementation of Random Forest Classifier

DOI: 10.48175/IJARSCT-36884