International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 12 Issue: 12 | Dec 2025 www.irjet.net p-ISSN: 2395-0072

Crop Recommendation System Using Machine Learning

Shivam Kumar[1], Suraj Thakur[2], Utsav Shukla[3], Vedant Agrawal[4], Prof. Rajeev Raghuwanshi[5]

[1]Student, CSE-AIML Department, Oriental Institute of Science & Technology, Madhya Pradesh, India
[2]Student, CSE-AIML Department, Oriental Institute of Science & Technology, Madhya Pradesh, India
[3]Student, CSE-AIML Department, Oriental Institute of Science & Technology, Madhya Pradesh, India
[4]Student, CSE-AIML Department, Oriental Institute of Science & Technology, Madhya Pradesh, India
[5]Professor, CSE-AIML Department, Oriental Institute of Science & Technology, Madhya Pradesh, India

---

Abstract - *In India, agriculture plays a significant role in the growth of the nation's economy and provides employment to a large population. Farmers often face challenges in selecting suitable crops due to limited knowledge of soil nutrients and changing environmental conditions, which adversely affects crop productivity. To address this issue, this paper presents a machine learning–based system that assists farmers in recommending appropriate crops and estimating crop yield based on various agricultural parameters. The proposed system uses features such as soil nutrients, temperature, humidity, rainfall, and soil pH to generate predictions. Supervised machine learning techniques are applied to analyze agricultural data and provide accurate recommendations. The system is implemented as a web-based application using a machine learning backend, enabling farmers to adopt a scientific and data-driven approach to farming. This approach can help improve agricultural productivity and support precision farming practices.*

***Key Words*: Agriculture, Crop Recommendation, Crop Yield Prediction, Machine Learning, Random Forest**

1.INTRODUCTION

Many developed countries have adopted modern scientific and technological techniques in agriculture to improve productivity and efficiency. These countries extensively use data-driven and automated systems to optimize farming practices. In contrast, agriculture in India still largely depends on traditional methods, despite being one of the major contributors to the nation's economy. Agriculture plays a significant role in employment generation and contributes substantially to the Gross Domestic Product. With rapid population growth and globalization, the demand for food production has increased considerably.

To meet this growing demand, farmers often rely on excessive use of chemical fertilizers to increase crop yield, which may lead to long-term environmental degradation and soil fertility issues. However, if farmers are provided with accurate information regarding suitable crops based on soil nutrients and environmental conditions, crop losses can be reduced and agricultural productivity can be improved. The availability of data related to soil composition, climatic conditions, and rainfall enables a better understanding of crop growth patterns influenced by geographical and environmental factors.

In this work, a machine learning–based predictive system is proposed to assist farmers in selecting appropriate crops and estimating expected yield. The system analyzes parameters such as soil nutrients, temperature, humidity, rainfall, and soil pH to generate recommendations. By identifying nutrient deficiencies and unsuitable crop choices, the proposed system helps in minimizing production inefficiencies. The adoption of such a scientific and data driven approach can significantly enhance farming practices and support sustainable agricultural development.

2. Literary Survey

Several research works have been carried out in the domain of crop recommendation and yield prediction using machine learning techniques. Padmakar et al. proposed a crop recommendation system for precision agriculture using soil nutrient data, soil type, and yield information. The system employed ensemble-based machine learning techniques to recommend suitable crops. Algorithms such as Random Tree, CHAID, and Support Vector Machine were used to improve prediction accuracy by combining the strengths of multiple models.

Solanki et al. presented a crop cultivation prediction system with the primary objective of reducing the risk associated with incorrect crop selection. The study evaluated multiple machine learning algorithms and implemented a k-fold cross validation approach, where the dataset was divided into five subsets to ensure reliable performance evaluation. The authors concluded that Random Forest achieved superior prediction performance compared to other models, followed by Support Vector Regression using the radial basis function kernel.

Kumar et al. introduced a supervised machine learning approach for crop yield prediction based on historical agricultural data. Their system analysed previous farming records to estimate future crop yields. The proposed model supported both qualitative and quantitative prediction of

harvest output and demonstrated effective performance in Fig. -2: Project Framework assisting agricultural planning and decision-making processes.

Cauvery et al. proposed a crop recommendation system using ensemble learning techniques to enhance crop productivity. Their approach combined predictions from multiple machine learning models to generate accurate crop recommendations. The system focused on knowledge-based learning and demonstrated improved precision by integrating results from various classifiers.

Patel et al. presented a machine learning–based crop recommendation and yield analysis system that focused on improving decision-making in agriculture through data-driven techniques. The study utilized soil parameters and climatic factors to train supervised learning models for predicting suitable crops and expected yield. Various classification and regression algorithms were evaluated to analyze their effectiveness on agricultural datasets. The results highlighted that ensemble-based models performed consistently better due to their ability to handle complex and non-linear relationships among agricultural features, making them suitable for practical farming applications.

3. System Design

To develop an efficient software application, it is essential to understand the concept of the Software Development Life Cycle (SDLC). SDLC provides a structured framework that helps in planning, designing, developing, and maintaining a software system in an organized manner. It ensures that the system is developed systematically while meeting user requirements and quality standards.

In the proposed project, the Waterfall model is adopted as the SDLC approach. The Waterfall model is a simple and straightforward life cycle model that follows a linear and sequential flow of development. In this model, each phase must be completed before the next phase begins, and there is no overlap between the stages. The output generated from one phase acts as the input for the subsequent phase, ensuring a clear flow of information throughout the development process.

The Waterfall model consists of several stages, including requirement analysis, system design, implementation, testing, deployment, and maintenance, as illustrated in Fig. 1. Each stage plays a crucial role in the successful development of the system. Following this model helps in systematic development and reduces complexity, making it suitable for projects with well-defined requirements.

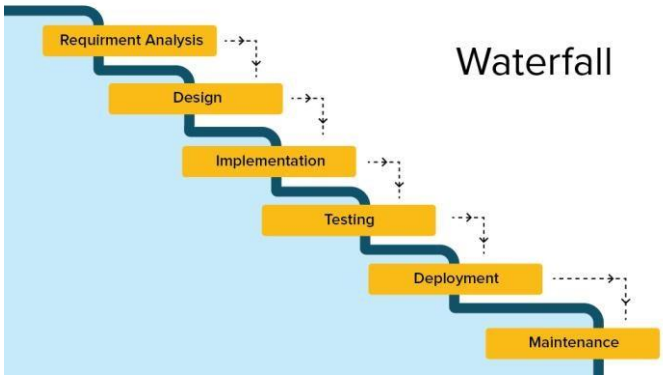

Fig. -1: Waterfall Model

4. Model Framework Overview

The most important steps involved in building a machine learning based predictive model are shown in the following " Fig. 2".

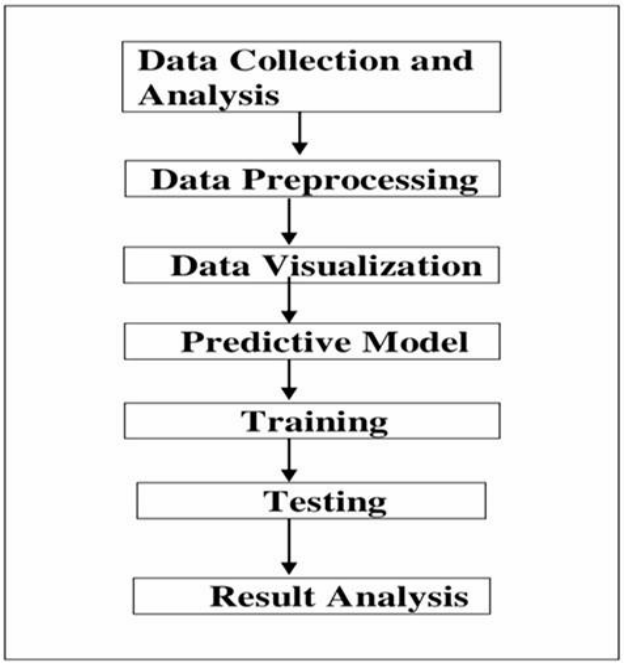

Fig. -2: Project Framework

5. System Architecture

System architecture refers to a conceptual and structured model that defines the organization, components, and operational behaviour of a software system. It provides a clear representation of how different modules interact with one another and explains the relationship and dependency between various processes involved in the system. The architecture of a system presents a high-level overview of its internal processes, methods, and stages. It allows developers and users to visualize the working of the system and understand how data flows from input to output. In the

proposed crop recommendation and yield prediction system, the architecture is designed to integrate data input, machine learning processing, and result generation efficiently. The modular design ensures smooth interaction between data collection, model prediction, and output display, thereby improving system reliability and scalability.

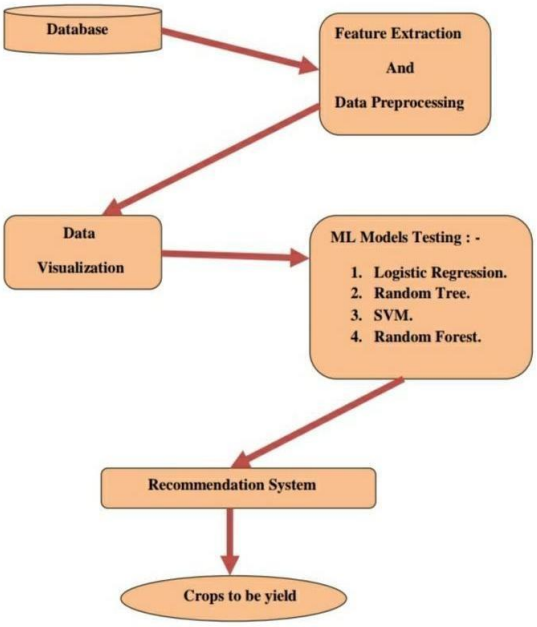

Fig. -3: System Architecture

6. Methodology and Implementation

This section describes the sequential steps followed for the implementation of the proposed predictive system. The overall process is carried out in a sequential manner to ensure proper data handling, model training, prediction, and result generation.

7. Dataset Collection

Dataset collection is the initial step in the development of any machine learning project. The data obtained from online sources is often raw and may contain inconsistencies or missing values. In this work, the datasets are collected from an open-source platform, Kaggle. The proposed system utilizes two primary datasets: a soil nutrient dataset containing information on nitrogen (N), phosphorus (P), potassium (K), and soil pH, and a climatic dataset consisting of parameters such as rainfall, humidity, and temperature. These datasets are combined to form the final dataset used for model training and prediction. The consolidated dataset contains approximately 2200 records with multiple agricultural attributes.

8. Data Analysis

Data analysis is performed prior to the preprocessing stage to gain a clear understanding of the dataset. In this step, the data is carefully examined to identify patterns, relationships, and important features that influence crop recommendation and yield prediction. Understanding the dataset at this stage is essential, as the quality of feature selection directly affects the accuracy of the prediction results.

The analysis focuses on studying the relationships between soil nutrients and environmental factors such as temperature, humidity, rainfall, and soil pH. These parameters are evaluated to determine their influence on crop suitability. Visual techniques such as heatmaps are used to represent correlations among different features, providing insights into how various factors interact with each other. The heatmap illustrating this analysis is shown in Fig. 5.

Based on the analysis, relevant features are selected for model training. Feature importance provided by ensemble learning methods helps in identifying influential parameters. This analytical process ensures that only meaningful and impactful features are considered, thereby improving the reliability and performance of the predictive system. Basic statistical analysis is performed to identify outliers and variations in the data. This helps in ensuring data consistency and improves the reliability of the predictive model.

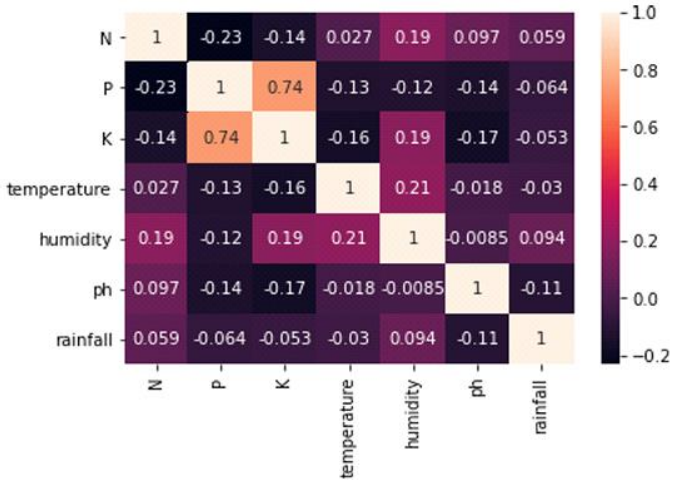

Fig. -4: Data analysis as heatmap

9. Data Visualization

Data visualization is used to represent agricultural data in a graphical form to better understand the effect of various factors on crop production. In this work, the production quantity of different crops is plotted against influencing parameters such as soil nutrients and climatic conditions.

These visual representations help in identifying trends and relationships that affect crop growth and overall productivity. Bar graphs are used to compare crop production with respect to different environmental factors, as shown in Fig. 5.

Visualization techniques also assist in understanding key parameters such as nitrogen, phosphorus, potassium, soil pH, temperature, humidity, and rainfall. Correlation plots and distribution graphs help in identifying dominant features that influence crop suitability. Such visual analysis supports feature selection and improves the effectiveness of the machine learning models.

Additionally, visual representation of prediction results helps users interpret the system output more clearly. By analyzing graphical outputs, users can gain better insights into recommended crops and expected yield trends. Overall, data visualization enhances system usability and supports informed, data-driven agricultural decision-making.

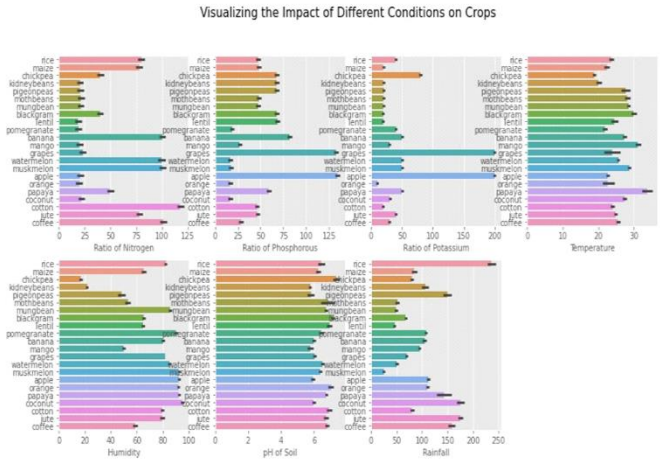

Fig. -5: Data Visualization

10. Machine Learning Algorithms

The crop recommendation component of the proposed system is formulated as a classification problem. Multiple supervised machine learning algorithms are trained and evaluated to determine the most suitable model for accurate crop prediction. The performance of each model is assessed based on prediction accuracy, and the algorithm with the best results is selected. Priority is given to achieving higher accuracy to ensure reliable crop recommendations and effective decision support for farmers.

- **Algorithm Selection**

  Various machine learning algorithms have been explored in existing literature for crop recommendation systems, including Decision Trees, Support Vector Machines, and Logistic Regression. Decision Tree models are simple and interpretable but often suffer from overfitting when applied to complex agricultural datasets. Logistic Regression is effective for basic classification problems but may fail to capture non-linear relationships among soil and climatic parameters. Support Vector Machines provide good accuracy in some cases; however, they require careful parameter tuning and are computationally expensive for large datasets. Based on these considerations and prior research findings, Random Forest was selected as the primary algorithm for the proposed system due to its robustness, ability to handle non-linear data, and superior generalization performance.

• Random Forest

Random Forest is an ensemble-based supervised machine learning algorithm that is widely used for both classification and regression tasks. It works by constructing multiple decision trees using randomly selected subsets of the training data and features. The final prediction is obtained by combining the outputs of all individual trees, using majority voting for classification problems and averaging for regression problems. This ensemble mechanism helps in reducing model variance and improves overall prediction accuracy.

One of the key advantages of Random Forest is its ability to handle non-linear relationships and complex interactions between input features, which are commonly observed in agricultural datasets. Unlike single decision tree models, Random Forest reduces the risk of overfitting by aggregating predictions from multiple trees. It is also robust to noise and missing values, making it suitable for real-world agricultural data.

In the proposed crop recommendation system, Random Forest is used to classify the most suitable crop based on soil nutrient parameters such as nitrogen, phosphorus, potassium, soil pH, along with climatic factors including temperature, humidity, and rainfall. Additionally, a Random Forest regressor is employed for crop yield prediction using historical agricultural data and encoded categorical features such as crop type and geographical region.

The trained Random Forest models are stored using the joblib library and integrated into a Flask-based backend. This allows the system to provide real-time predictions through a web interface, enabling farmers to make informed, data-driven decisions. The efficiency, scalability, and reliability

of the Random Forest algorithm make it well-suited for precision agriculture applications.

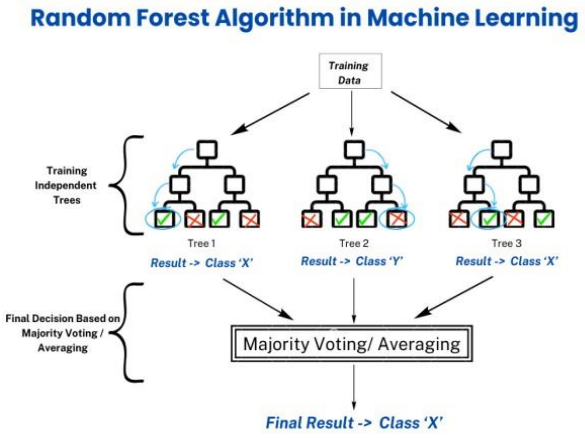

Fig. -6: Random Forest

11. Testing

Testing is an essential phase in evaluating the performance and reliability of a machine learning system. In the proposed crop recommendation and yield prediction system, the dataset is divided into training and testing subsets to analyze how well the trained models generalize to unseen data. A train– test split approach is used, where a portion of the data is used for training the model and the remaining data is used for testing its performance.

The system is developed by integrating multiple functional modules, as described in the system architecture. Each module is tested individually as well as in combination to ensure proper functionality. Different machine learning models were trained and evaluated during the development phase to identify the most suitable algorithm for crop prediction.

Model performance is evaluated using standard evaluation metrics. For crop recommendation, metrics such as accuracy, precision, and classification report are used to assess classification performance.

For crop yield prediction, regression-based evaluation measures are applied to analyze prediction accuracy. These testing procedures help ensure that the selected model performs reliably and provides meaningful predictions for real-world agricultural applications.

12. Accuracy

Accuracy is one of the most important metrics used to evaluate the performance of a machine learning classification model. It represents the proportion of correctly predicted instances out of the total number of predictions made. Accuracy helps in determining how effectively the model classifies the input data and whether it is suitable for real-world applications.

Mathematically, accuracy is defined as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

It can also be expressed using confusion matrix parameters as:

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

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

13. Precision

Precision is a performance metric that measures the correctness of positive predictions made by the model. It indicates how many of the instances predicted as positive are actually positive. Precision is particularly important when false positive predictions need to be minimized.

Precision is calculated using the confusion matrix as:

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

14. Recall

Recall is a performance evaluation metric that measures the ability of a machine learning model to correctly identify all relevant positive instances. It indicates how many of the actual positive samples are successfully predicted by the model. Recall is particularly important in applications where missing a positive instance can lead to significant consequences.

Recall is calculated using the confusion matrix as:

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

15. Result and Performance Analysis

The performance of the proposed crop recommendation and yield prediction system is evaluated based on prediction outputs obtained through the web-based application. The trained machine learning models are tested using user provided input values, and the corresponding results are generated in real time. This evaluation helps in assessing the practical applicability, consistency, and reliability of the developed system.

For crop recommendation, the system accepts soil nutrient parameters such as nitrogen, phosphorus, potassium, and soil pH, along with climatic factors including temperature, humidity, and rainfall. These parameters are processed by the Random Forest classification model to identify the most suitable crop under the given conditions. The recommendation results demonstrate that the system is able to effectively map environmental and soil conditions to appropriate crop choices. Fig. 7 illustrates the crop recommendation webpage, showing the input fields and the predicted crop output generated by the system.

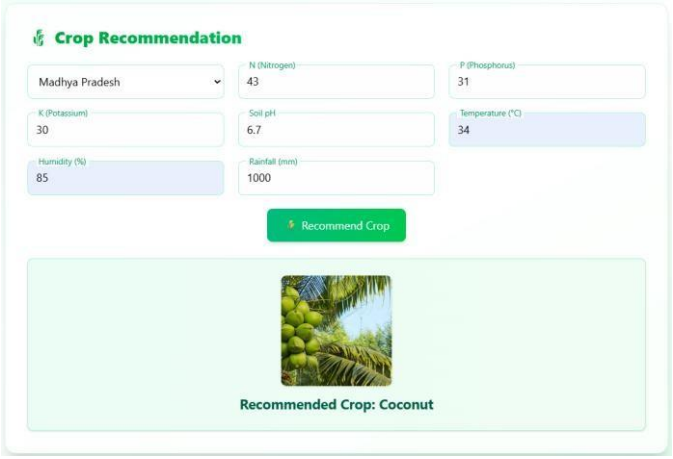

Fig. -7: Result of Crop Recommendation

The recommendation module is designed to support informed decision-making by reducing dependency on traditional trial-and-error farming practices. By analyzing multiple parameters simultaneously, the system provides a data-driven approach to crop selection, which can help farmers minimize crop mismatch and improve productivity. The consistent prediction behaviour observed during testing indicates that the model generalizes well to different input combinations provided through the interface.

In addition to crop recommendation, the system also includes a crop yield prediction module. This module estimates the expected yield based on agricultural and environmental inputs using a Random Forest regression model. Fig. 8 represents the yield prediction webpage, displaying the input parameters and the corresponding predicted yield value. This feature allows users to gain an approximate understanding of potential crop productivity before cultivation.

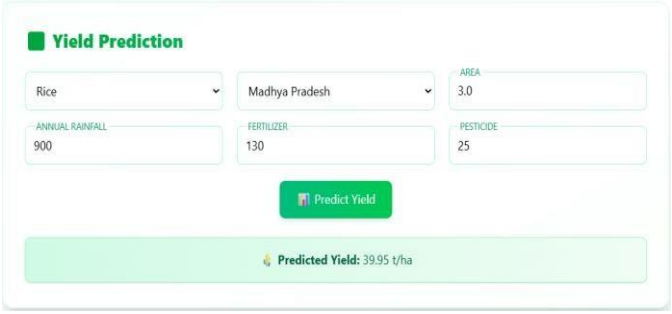

Fig. -8: Result of Yield Prediction

Overall, the results show that the integrated system successfully delivers meaningful predictions for both crop selection and yield estimation. The combination of machine learning models with a Flask-based backend ensures efficient processing and quick response time. The simple and user-friendly interface further enhances usability, making the system suitable as a practical decision-support tool for precision agriculture.

16. Conclusion

In this paper, a machine learning–based crop recommendation and yield prediction system has been proposed to support data-driven decision making in agriculture. The system utilizes soil nutrient parameters and climatic conditions to recommend suitable crops and estimate expected yield using supervised learning techniques. Random Forest algorithms were employed for both classification and regression tasks due to their robustness and ability to handle complex agricultural data.

The developed system was successfully implemented as a web-based application using a Flask backend, allowing users to interact with the models through a simple interface. Experimental evaluation demonstrated that the system is capable of generating consistent and meaningful predictions based on user-provided inputs. By reducing dependency on traditional trial-and-error farming practices, the proposed approach can assist farmers in improving crop selection and planning.

Overall, the system highlights the potential of machine learning in precision agriculture and provides a practical decision-support tool that can contribute to enhanced agricultural productivity. The proposed solution can serve as a foundation for further advancements in smart and sustainable farming technologies.

17. Future Scope

The proposed crop recommendation and yield prediction system can be further enhanced in several ways. Integration of real-time weather data and soil sensor information can improve prediction accuracy and adaptability to changing environmental conditions. The

system can also be extended to support region-specific models by incorporating larger and more diverse datasets.

In the future, advanced machine learning and deep learning techniques can be explored to capture complex temporal patterns in agricultural data. Deployment as a mobile application can improve accessibility for farmers in remote areas. Additionally, features such as fertilizer recommendation, irrigation planning, and multilingual support can be integrated to make the system more comprehensive and farmer-friendly.

18. References

- Nilesh Dumber, Omkar Chikane, Gitesh Moore, "System for Agriculture Recommendation using Data Mining" , E- ISSN : 2454- 9916 | Volume : 1 | Issue : 4 | NOV 2014.
- Government of India, "Soil Health Card Scheme," Ministry of Agriculture and Farmers Welfare, Available: https://soilhealth.dac.gov.in/home
- R. K. Solanki, D. Bein, J. A. Vasko, and N. Rale, "Prediction of Crop Cultivation Using Machine Learning Techniques," *International Journal of Computer Applications*, vol. 111, no. 6, pp. 1–5, 2015.
- T. M. Mitchell, *Machine Learning*, McGraw-Hill Education, 1997.
- Kaggle, "Crop Recommendation Dataset," Available: https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset
- Kaggle, "Yield Prediction Dataset," Available: https://www.kaggle.com/datasets/akshatgupta7/crop-yield-in-indian-states-dataset
- S. Pudumalar, E. Ramanujam, R. Harine Rajashree, C. Kavya, T. Kiruthika, and J. Nisha, "Crop Recommendation System for Precision Agriculture," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 6, pp. 32–36, 2017.