Crop Recommendation System Using Machine Learning

Pathlavath Sudeendra[1], Prof.Shashikala Tapaswi[2]
Department of Computer Science and Engineering, ABV-IIITM Gwalior, India
[1]bobyjljsn@gmail.com,[2]stapaswi@iiitm.ac.in

***Abstract*—Agriculture plays a crucial role in India's economy, significantly contributing to the country's GDP and providing employment to a large portion of the population. Despite advancements in agricultural technology, many farmers still rely on traditional methods and make farming decisions based on weather cues, which can lead to suboptimal crop choices and reduced yields. To address this challenge, this project introduces a machine learning-based Crop Recommendation System designed to assist farmers in selecting the most suitable crop for their land. The system utilizes various machine learning classifiers, including Random Forest, Logistic Regression, and Support Vector Machine, to evaluate soil nutrients (such as nitrogen, phosphorus, and potassium) and environmental factors (including temperature, humidity, pH, and rainfall) in order to recommend the best crop for cultivation. Trained on a dataset encompassing multiple crops and their respective environmental and soil conditions, the system involves preprocessing steps like feature scaling and standardization before applying different machine learning models. Among the models tested, the Random Forest Classifier achieved the highest accuracy of 99.3%, establishing it as the primary model for crop prediction. The system is implemented to provide real-time crop suggestions based on user inputs, potentially offering a valuable tool for farmers and agricultural planners. This machine learning approach aims to improve agricultural efficiency by ensuring that crops are well-suited to their growing conditions, promoting sustainable farming, and aiding farmers in making data-driven decisions, ultimately enhancing productivity and supporting the agricultural sector in India.**

***Index Terms*—Crop recommendation,Nitrogen value, pH value,Potassium value, Phosphorous value, Humidity value, Logistic Regression,Support Vector,Random Forest.**

I. Introduction

The Crop Recommendation System represents a significant advancement in agricultural technology, utilizing machine learning (ML) algorithms to enhance crop selection and production efficiency. As agriculture remains a cornerstone of many economies, particularly in countries like India, optimizing crop yields through innovative methods is increasingly crucial. Traditional farming practices, which often rely on historical weather patterns and personal experience, can lead to suboptimal crop choices and reduced productivity. This project addresses these limitations by integrating machine learning techniques to analyze various environmental and soil parameters, offering precise crop recommendations tailored to specific conditions.

Machine learning has revolutionized the field of agriculture by enabling the development of sophisticated models that forecast crop yields with high accuracy. These models utilize diverse datasets, including soil characteristics, climate patterns, historical yield records, and agronomic practices, to provide comprehensive analyses and predictions. The convergence of machine learning with meteorological data further enhances crop selection by allowing farmers to make informed decisions based on current and forecasted weather conditions. This approach not only improves decision-making but also supports sustainable agricultural practices by optimizing resource use and increasing overall productivity.

In India, agriculture is a major economic sector and a primary source of livelihood for millions. Despite its significance, the sector often relies on traditional farming methods, which can result in inefficient crop choices and lower yields. Many farmers still depend on personal experience and environmental observations rather than data-driven insights. As global food demands and climate variability increase, there is an urgent need for modern approaches to support more effective agricultural practices.

Machine learning offers a transformative solution by analyzing key factors such as soil nutrients, temperature, humidity, pH levels, and rainfall to provide tailored crop recommendations. This shift towards data-driven agriculture helps farmers optimize their crop selection, manage resources more efficiently, and improve overall productivity. The integration of machine learning with weather forecasting and data mining strategies further enhances crop yield predictions and decision-making processes. By adopting these advanced techniques, the agricultural sector can move towards more sustainable and productive practices, bridging the gap between traditional methods and modern technology.

To achieve these goals, the project aims to analyze key agricultural factors by using machine learning models to evaluate critical environmental and soil parameters such as nitrogen, phosphorus, potassium levels, temperature, humidity, pH, and rainfall, all of which are essential in determining crop suitability. It will develop an accurate predictive model by training and comparing various machine learning algorithms to identify the most effective model for predicting the optimal crop, with particular emphasis on models like the Random Forest Classifier, known for its high accuracy. Additionally, the project will implement a user-friendly interface by developing a web-based application using Flask, which will allow users to

easily input their data and receive real-time crop recommendations. To support sustainable agriculture, the system will facilitate the adoption of data-driven practices that enhance crop yield and boost the agricultural sector's productivity. Finally, it aims to bridge the gap between traditional methods and modern technology by making advanced technological tools accessible to farmers, thus integrating modern data-driven insights into traditional farming practices and ultimately improving decision-making and agricultural outcomes.

II. Literature Survey

The literature on machine learning-based crop recommendation systems highlights significant advancements, challenges, and promising applications within the agricultural sector. Previous studies have primarily utilized basic data preprocessing techniques. For instance, one study proposed a crop recommendation system employing the Random Forest algorithm and applied fundamental preprocessing methods such as normalization and handling missing values. The experimental results demonstrated that the Random Forest model achieved an accuracy of 95%. However, the study also suggested that accuracy could be further improved through the use of more advanced preprocessing techniques and a comparison with other machine learning models.

Previous studies on crop recommendation systems have frequently employed traditional machine learning techniques with basic data preprocessing. For example, Bondre et al. [1] proposed a crop yield prediction system using the Random Forest algorithm. However, the study was constrained by the use of a limited set of crop types and rudimentary preprocessing methods. Consequently, the Random Forest model achieved an accuracy of only 86.35%. This finding underscores the potential for improved accuracy through the application of more advanced preprocessing techniques and the use of a more comprehensive dataset.

Suresh et al. [5] focused on detailing crop information and their nutritional values, including nitrogen (N), phosphorus (P), and potassium (K) levels. Their research analyzed these nutritional parameters within a limited dataset to assess their impact on crop yield. While the study offered valuable insights into crop nutrition, it was constrained by the dataset's size and scope. Additionally, although the research included the design and deployment of a system based on these findings, it did not extensively explore advanced data preprocessing techniques or compare multiple machine learning models.

Reddy et al. [3] concentrated on ensemble machine learning algorithms in their study published in the International Journal of Scientific Research in Science and Technology. Their research focused on comparing various ensemble methods, including Decision Trees, Naive Bayes, and Random Forests, to identify the most effective algorithms for optimizing crop yield. By visualizing the performance of these algorithms, the study offered valuable insights into selecting the best approach for crop recommendation, demonstrating the utility of ensemble methods in enhancing agricultural decision-making.

Garanayak et al. [2] utilized modern machine learning techniques in their research, focusing on a subset of five crops and employing Random Forest Regression, SVM Regression, and Polynomial Regression. By splitting the dataset into training and test sets, they evaluated the performance of these models and achieved an accuracy of 94.7% for rice. Their study compared the accuracy of different regression techniques for crop recommendation, offering valuable insights into the effectiveness of contemporary methods in optimizing crop selection.

III. METHODOLOGY

The methodology section outlines the systematic approach taken to develop the Crop Recommendation system. It encompasses data collection, preprocessing, model development, training, and evaluation.

A. Data Collection

Various datasets relevant to crop recommendation were gathered, each containing environmental and soil data linked to different crop types. The selected dataset includes features such as soil pH, temperature, humidity, rainfall, phosphorus (P), potassium (K), and nitrogen (N). It ensures a balanced distribution across 22 crops, including rice, maize, chickpeas, kidney beans, pigeon peas, moth beans, mung bean, black gram, lentil, pomegranate, banana, mango, grapes, watermelon, muskmelon, apple, orange, papaya, coconut, cotton, jute, and coffee. This dataset comprises 2200 records and was essential for developing models capable of accurately recommending the best crop based on specific environmental factors.

	N	P	K	temperature	humidity	ph	rainfall
0	90	42	43	20.879744	82.002744	6.502985	202.935536
1	85	58	41	21.770462	80.319644	7.038096	226.655537
2	60	55	44	23.004459	82.320763	7.840207	263.964248
3	74	35	40	26.491096	80.158363	6.980401	242.864034
4	78	42	42	20.130175	81.604873	7.628473	262.717340
...	...	...	...	...	...	...	...
2195	107	34	32	26.774637	66.413269	6.780064	177.774507
2196	99	15	27	27.417112	56.636362	6.086922	127.924610
2197	118	33	30	24.131797	67.225123	6.362608	173.322839
2198	117	32	34	26.272418	52.127394	6.758793	127.175293
2199	104	18	30	23.603016	60.396475	6.779833	140.937041

2200 rows × 7 columns

Fig. 1. Dataset

B. Data Preprocessing

1) Data Cleaning:

- The first step in the preprocessing pipeline was to clean the dataset. This involved checking for and handling missing values, a common issue in real-world datasets.

Fortunately, the dataset did not contain any missing values, as confirmed by the 'isnull().sum()' function. Following this, the dataset was scanned for duplicates using the 'duplicated().sum()' function. Any duplicate rows identified were removed to maintain the dataset's integrity, ensuring that each instance was unique and contributed to the learning process without introducing bias.

2) Data Balancing:

- In machine learning, particularly in classification tasks, having a balanced dataset is crucial for training models that can generalize well across all classes. The dataset used for the Crop Recommendation System was inherently balanced, meaning that each of the 22 crop types was equally represented. This balance was critical as it prevented the models from becoming biased towards the more frequently occurring classes, ensuring that the system could accurately recommend crops across varying environmental conditions.

3) Encoding Categorical Variables:

- The crop labels, which were originally in string format, needed to be converted into a numerical format suitable for machine learning algorithms. This was achieved through label encoding. A dictionary mapping ('crop-dict') was created to convert the crop names into corresponding numerical values, which were then added to the dataset as a new column, 'crop-num'. This transformation was essential because machine learning algorithms require numerical input, and this step ensured that the crop labels were appropriately encoded for the modeling process.

C. Feature Extraction

Feature scaling is a vital preprocessing step in the Crop Recommendation System, ensuring that all input features contribute equally to the model's learning process. In this project, two primary scaling techniques were employed: Min-Max Scaling and Standardization. These techniques were applied to the environmental variables, including soil pH, temperature, humidity, rainfall, phosphorus (P), potassium (K), and nitrogen (N).

1) Min-Max Scaling:

- The first step in scaling involved normalizing the data using Min-Max Scaling. This technique adjusts the range of the data so that all features lie between 0 and 1. Min-Max Scaling is particularly useful when dealing with features that have different units or scales, such as temperature and rainfall, which may naturally have vastly different ranges. This disparity could potentially skew the results of distance-based algorithms like K-Nearest Neighbors (KNN). By applying Min-Max Scaling, all features were brought into the same range, ensuring that no single feature dominated the model's learning process due to its magnitude.

2) Standardization:

- Following Min-Max Scaling, the data was further standardized using StandardScaler. Standardization shifts the data distribution to have a mean of 0 and a standard deviation of 1. This process is crucial for algorithms that are sensitive to the distribution of the data, such as Support Vector Machines (SVM) and Logistic Regression. Standardization ensures that each feature contributes equally to the model, especially when the model relies on assumptions about the data's distribution. By applying this technique, the models were better equipped to converge more quickly during training and achieve higher accuracy.

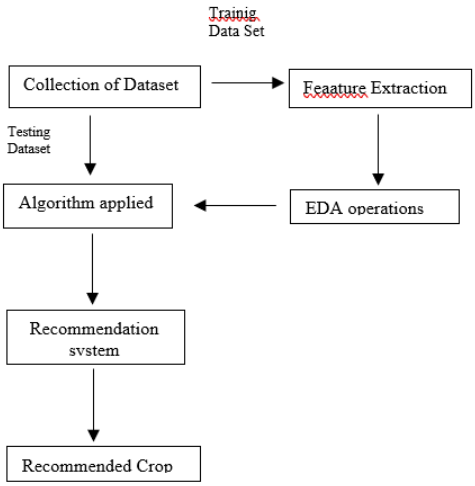

Fig. 2. Steps Involved in a model

D. Model Exploration

In the Crop Recommendation System, ensemble models like Random Forest and Gradient Boosting are preferred over simpler algorithms due to their ability to capture complex relationships between environmental factors and crop suitability. Factors such as soil pH, temperature, and nutrient levels interact in non-linear ways, which simpler models like Logistic Regression or Decision Trees may not fully capture. Ensemble methods, by combining the predictions of multiple decision trees, offer a more accurate and robust analysis. This makes them ideal for generating reliable crop recommendations. Random Forest, in particular, stands out due to its high accuracy and effectiveness in handling diverse agricultural data.

1) Logistic Regression: Logistic Regression is a linear model used to estimate the probability of recommending a certain crop based on environmental features. Its advantages include simplicity and interpretability, making it easy to understand the impact of each factor. It also serves as a baseline for comparison with more complex models. However, its main drawbacks are its assumption of linear relationships and its limited ability to handle non-linearities and complex patterns

in the data. In the Crop Recommendation System, Logistic Regression provided initial insights but was less effective in capturing the complex interactions of crop growth conditions.

2) Decision Tree: Decision Tree uses a tree-like structure to make decisions by splitting the data based on feature values, which directly maps environmental conditions to the recommended crop. Its advantages include interpretability, with a clear visual representation of decision rules, and the fact that it does not require feature scaling, simplifying preprocessing. However, it is prone to overfitting by capturing noise as true signals, and small data changes can lead to different trees, reducing reliability. In the Crop Recommendation System, Decision Trees offered clear decision paths but required careful pruning and validation to manage the risk of overfitting.

3) Random Forest: RandomForest is an ensemble method that builds multiple decision trees and combines their predictions to improve accuracy and robustness. It enhances predictive accuracy by averaging results from several trees and provides insights into feature importance, identifying key environmental factors affecting crop recommendations. Although it is less interpretable and requires more computational resources than a single decision tree, RandomForest proved to be the top-performing model in the Crop Recommendation System, delivering the highest accuracy and handling complex interactions between features effectively.

4) Bagging Classifier: Bagging Classifier is an ensemble method that builds multiple models from different subsets of the training data and combines their predictions to improve stability and accuracy. Its advantages include reduced variance, leading to more reliable recommendations, and enhanced accuracy by averaging out errors from individual models. However, it involves complex implementation and increased computational load due to managing multiple training processes. In the Crop Recommendation System, Bagging Classifier contributed to improved stability and accuracy of crop recommendations, making it a valuable addition to the ensemble methods used.

5) Gradient Boosting: Gradient Boosting builds models sequentially, with each new model correcting the errors of its predecessor, which enhances prediction accuracy by focusing on difficult cases. It excels in capturing complex patterns and is adaptable to various loss functions, making it highly flexible. However, it risks overfitting if not well-tuned and can be slow to train, especially with large datasets. In the Crop Recommendation System, Gradient Boosting improved accuracy by refining predictions through error correction, proving effective for handling complex data distributions, though it required careful tuning to avoid overfitting.

6) Support Vector Machine (SVM): SVM is a robust algorithm that identifies the optimal hyperplane to separate different crop types in a high-dimensional space, making it effective for complex datasets with many features. It handles non-linear relationships through various kernel functions but can be computationally intensive and sensitive to parameter tuning. In the Crop Recommendation System, SVM was advantageous for managing the dataset's complexity and high dimensionality, though its training required careful parameter adjustment due to its computational demands.

7) K-Nearest Neighbors (KNN): KNN is a non-parametric method that classifies crops based on the majority vote from neighboring data points, capturing complex, non-linear relationships between features and crop types. It is simple to understand and implement but can be computationally intensive and sensitive to noisy data. In the Crop Recommendation System, KNN effectively handled non-linear relationships but required careful management due to its computational demands and sensitivity to noise.

IV. Experiments and Results

In our Crop Recommendation System project, we evaluated various machine learning models to determine the best for predicting optimal crops based on environmental factors. The models tested included Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, Random Forest, Bagging Classifier, and Gradient Boosting.

The Logistic Regression model served as a baseline with an accuracy of 96.36%, precision of 96.44%, and recall of 96.36%. While effective, it was outperformed by more advanced models. The SVM model achieved an accuracy of 96.82%, precision of 97.15%, and recall of 96.82%, showing improved performance but still not surpassing ensemble methods.

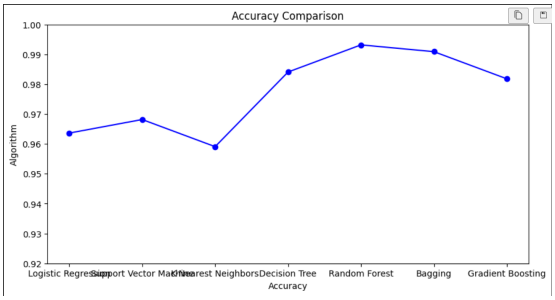

Fig. 3. Accuracy Comparison

The KNN model, which captures non-linear relationships, had an accuracy of 95.91%, precision of 96.54%, and recall of 95.91%. Its performance was dependent on parameter tuning. The Decision Tree model achieved an accuracy of 98.18%, precision of 98.24%, and recall of 98.18%. It provided clear decision paths but was prone to overfitting.

The Gradient Boosting model also performed well, with an accuracy of 98.18%, precision of 98.43%, and recall of 98.18%. Its iterative improvement of predictions showed promise, though it required more computational resources. The Bagging Classifier achieved an accuracy of 98.59%, precision of 97.16%, and recall of 98.09%, providing enhanced stability but slightly less accuracy than Random Forest.

The Random Forest model emerged as the most effective, with the highest accuracy of 99.32%, precision of 99.37%, and recall of 99.32%. Its ability to handle diverse data and reduce overfitting made it the top choice for crop recommendations.

Model	Precision	Recall	Accuracy
Logistic Regression	0.964442	0.963636	0.963401
Support Vector Machine	0.971517	0.968182	0.968182
Bagging	0.971620	0.980909	0.985909
Decision Tree	0.982402	0.981818	0.981818
Random Forest	0.993735	0.993182	0.993182
K-Nearest Neighbors	0.965390	0.959091	0.959091
Gradient Boosting	0.984271	0.981818	0.981818

Fig. 4. Comparison Table

A. Performance Comparison

The performance comparison between the Crop Recommendation System and the previous base model highlights a significant improvement in both accuracy and reliability. This enhancement is primarily attributed to the adoption of advanced techniques. By incorporating additional features and employing sophisticated data preprocessing methods, the current model has been able to better capture the complexities of the agricultural environment. These improvements have led to more precise and reliable crop predictions, demonstrating the effectiveness of the new approaches in enhancing the system's overall performance.

Model	Previous Model Accuracy	Our Model Accuracy
Support Vector Machine	0.75	0.96
K-Nearest Neighbor	0.90	0.95
Random Forest	0.95	0.99

Fig. 5. Model Comparison

Enhanced data preprocessing techniques, such as careful feature scaling and dataset balancing, played a crucial role in refining the model's inputs. These improvements helped in reducing noise and ensuring that the model was trained on high-quality data. As a result, the current model demonstrates superior performance, highlighting the effectiveness of these methodological advancements over the earlier approaches

B. Website Implementation

The Crop Recommendation System website features a user-friendly interface that allows users to input soil and environmental data, such as nitrogen, phosphorus, potassium levels, temperature, humidity, pH, and rainfall. This information is processed by trained machine learning models to recommend the most suitable crop for cultivation. The website is developed using Flask, a lightweight Python web framework, which manages HTTP requests, data processing, and model inference. The frontend, designed with HTML, CSS, and JavaScript, provides a smooth user experience, featuring clear input forms and result displays.

The system incorporates several technologies and features: Flask handles the backend operations, including data processing and model integration, with machine learning models serialized using Python's pickle module. Data preprocessing is managed through StandardScaler and MinMaxScaler to ensure proper scaling before model input. Key features of the website include a Home Page that offers an overview of the project, a Predict Page where users can input data and select a model for predictions, and a Model Selection option that lets users choose between different machine learning models. The Prediction Display provides the recommended crop or an informative message if the data does not match any known crop.

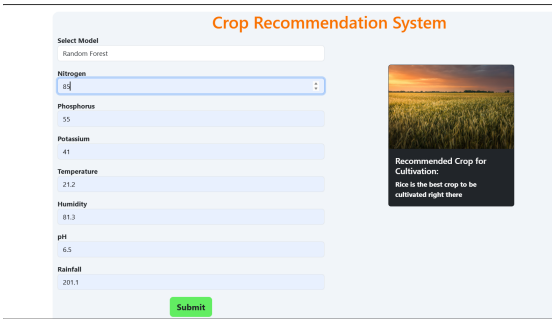

Fig. 6. Crop Recommendation Using Certain Values

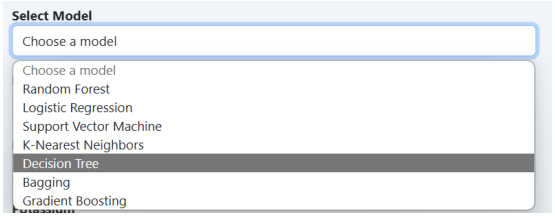

Fig. 7. Selection of Model

V. Conclusion

In conclusion, the Crop Recommendation System using Machine Learning presents a significant advancement in optimizing agricultural practices. By leveraging advanced machine learning models, the system provides accurate recommendations for the most suitable crops based on critical environmental factors such as soil pH, temperature, humidity, and nutrient levels. The project effectively integrates various machine learning techniques, including ensemble methods like Random Forest and Gradient Boosting, which excel at capturing complex relationships among these variables. The Random Forest model, in particular, has proven to be the most accurate, demonstrating its effectiveness in predicting optimal crop choices.

The comprehensive data preprocessing, involving feature scaling, encoding, and balancing, ensured that the models were well-prepared for training and performed efficiently. This system not only enhances precision agriculture but also supports farmers in making data-driven decisions, leading to better crop yields and more sustainable farming practices.

REFERENCES

[1] Devdatta A. Bondre and Santosh Mahagaonkar. Prediction of crop yield and fertilizer recommendation using machine learning algorithms. *Research Article*, 2022.

[2] Goutam Sahu SachiNandan Mohanty Garanayak, Mamata and Alok Kumar Jagadev. Agricultural recommendation system for crops using different machine learning regression methods. *IGIGlobal*, 2021.

[3] Bhagyashri Dadore Reddy, D. Anantha and Aarti Watekar. Crop recommendation system to maximize crop yield in region using machine learning. *Research Article*, 2019.

[4] Dr. Prem Kumar Ramesh Shilpa Mangesh Pande. Crop recommender system using machine learning approach. In *Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC)*, pages 1–6. IEEE, 2021.

[5] A.Senthil Kumar Suresh, G. Efficient crop yield recommendation system using machine learning for digital farming. *Research Article*, 2021.