ML based Soil analysis for Crop Recommendation

Bala Murali Gorantla
QIS College of Engineering & Technology
CSE (AIML) department
Ongole, India
balamurali8286@gmail.com

Bharath Varma Gottumukkala
QIS College of Engineering & Technology
CSE (AIML) department
Ongole, India
bharathvarma905@gmail.com

Poornateja Peyyala
QIS College of Engineering & Technology
CSE (AIML) department
Ongole, India
poornateja.p170@gmail.com

Mashoor Shaik
QIS College of Engineering & Technology
CSE (AIML) department
Ongole, India
skmashoor11@gmail.com

Pradeep Pathipati
QIS College of Engineering & Technology
CSE (AIML) department
Ongole, India
chowdarypradeep495@gmail.com

Karthik Purimitla
QIS College of Engineering & Technology
CSE (AIML) department
Ongole, India
21491a4263@qiscet.edu.in

Abstract*— Agriculture plays a crucial role in sustaining economies by ensuring food security, supporting livelihoods, and driving economic growth. Effective crop selection is essential to optimize agricultural output, utilize resources efficiently, and promote environmental sustainability in the long run. This study develops an advanced harvest recommendation system using techniques for machine learning including vector classification (SVC), logistics regression, decision tree classifier, K-nearest neighbor, naive Bayes, and random forests. By analyzing factors such as nutrient levels (nitrogen, phosphorus, potassium), temperature, humidity, soil pH, and rainfall, the system identifies optimal crops for each season. Machine learning models are evaluated based on key metrics such as accuracy, accuracy, recall, and F1 scores to determine the most effective approach to harvest prediction. This research aims to enhance agricultural decision-making, empowering farmers to select crops strategically to boost productivity and efficiency. Furthermore, this study contributes to precision agriculture, which supports data-controlled knowledge when accurate yield forecasts occur and promotes sustainable agricultural practices.

***Index Terms*— Crop recommendation, Machine Learning, soil analysis, GassianNB, Python, ML models**

I. INTRODUCTION

Agriculture is a core base of human society, offering life-sustaining food, raw materials, and livelihoods for a large chunk of the population worldwide. Agriculture is key in ensuring food availability, supporting livelihoods, and propelling economic development, particularly in agrarian economies. Nonetheless, in light of growing climate change challenges, land degradation, and the scarcity of resources, more emphasis is placed on developing creative solutions that raise the level of agricultural productivity with an assurance of sustainability. A very important feature in ensuring the success of such a strategy is proper crop choice as guided by the environment and the soil conditions. Proper selection of harvests maximizes performance, saves resources and reduces the impact of agricultural activities on the environment.

Conventional ranchers have chosen plants essentially with involvement and resilience, which is valuable, but not reasonable for today's natural conditions. Further developments in technology, particularly machine learning, provide an opportunity to switch from traditional intuition-based methods to data-controlled decisions aimed at improving farming efficiency. Machine learning algorithms can analyze vast datasets, including soil characteristics, climatic conditions, and historical crop performance, to recommend the most suitable crops for specific regions and seasons. This data-centric approach has the potential to revolutionize agriculture, ensuring optimal resource utilization and increased productivity. In integrating machine learning in agriculture, farmers have better decision-making practices, which yield greater outputs, lower input costs, and more efficient farming methods.

This research applies various machine learning techniques, such as SVC, logistic regression, decision tree classifiers, K-nearest neighbors, Naive Bayes, and Random Forest, to build a reliable crop recommendation system. Identify the best plants for a particular area by assessing soil and environmental factors such as nutrient levels, temperature, humidity, pH values, and rainfall. Performance metrics such as accuracy, precision, recall, and F1-score are used to determine the most effective model for crop prediction.

This thought of employment makes propels within the development of decision-making, determination of data from agriculturists, and determination of frameworks that maximize adequacy and productivity, whereas at the same time guaranteeing perfect utilize of resources. The findings contribute to the advancement of precision agriculture by improving crop yield pre- dictions and promoting sustainable farming practices through data-driven insights. As agriculture continues to face global challenges, the integration of modern technology solutions such as machine learning is extremely important to ensure long-term nutritional safety and economic stability.

II. Related Work

Machine learning has been attracting attention in agriculture in recent years, particularly for its precise breeding and harvesting recommendations. Several studies have challenged the use of data control models to improve agricultural decisions. For example, [1] established a machine learning-driven crop recommendation system that used soil characteristics and weather conditions to recommend the best crops. A decision tree classifier was employed in the study and achieved encouraging results in improving yield forecasting. Sharma and [3] have also used Support Vector Machines (SVM) and logistic regression for crop choice and proved that machine learning models perform better compared to conventional techniques in suggesting crops depending on the soil composition and climate conditions.

Several studies have focused on integrating multiple machine learning techniques to improve prediction accuracy. One such research [7] analyzed Naive Bayes, K-Nearest Neighbors (KNN), and Random Forest for crop recommendation. The findings highlighted that ensemble methods, particularly Random Forest, delivered superior results due to their ability to model complex, non-linear patterns in agricultural datasets. [11] used deep learning methods in combination with conventional classifiers and proved that hybrid models can further improve crop prediction accuracy. The research indicated that the use of actual environmental factors like temperature, humidity, and rainfall in predictive models is crucial.

Soil nutrient analysis has been a deciding factor for machine learning-based crop recommendation systems. According to [4], nitrogen, phosphorus, and potassium (NPK) levels play a critical role in determining crop selection. The study emphasizes that accurate analysis of soil nutrients significantly enhances the effectiveness and accuracy of predictive models. Their research indicated that machine learning models with soil fertility indices provide more accurate crop suggestions. A separate study [13] introduced a precision agriculture system leveraging artificial intelligence (AI) and Internet of Things (IoT) sensors. By gathering and analyzing soil data, this system empowers farmers to make well-informed and timely decisions that enhance agricultural productivity. Additionally, advancements in big data analytics and cloud computing have facilitated the scalability of machine learning applications in agriculture. [14] proposed a cloud-basedsystem for crop suggestions that utilized extensive datasets,so farmers could acquire crop recommendations using mobileapps. Their research featured the digital transformation role indigitalizing agriculture to enhance sustainable farm practices.Despite all these progresses, there are still challenges thatexist in deploying machine learning into agriculture, whichinclude data access, model explanation, and availability of region-level datasets. The present work further enhances the findings from earlier work by combining different machine learning methods to study dominant soil and environmentalfactors in hopes of creating an efficient and robust croprecommendation platform. [15] By comparing model performance against basic classification metrics, this study aims to improveagricultural decision-making and add to the evolution of precision agriculture.

III. Methodology

The suggested crop recommendation system is designed based on machine learning methods to examine soil and climatic conditions for best crop selection. The methodology involves several important steps, such as data gathering, preprocessing, feature choice, model application, performance testing, and deployment. Every stage is carefully crafted to guarantee the system's accuracy and reliability in delivering crop recommendations grounded in data-driven analysis.

A. Data Collection:

Gathering data plays a vital role in building a precise and dependable crop recommendation system. The performance of machine learning models is strongly influenced by the quality and comprehensiveness of the data used in training and testing. In this study, information is sourced from various channels, including agricultural research databases, government reports, meteorological organizations, and publicly accessible datasets related to soil and environmental factors. The data gathered is composed of different parameters that have a direct impact on plant growth, such as soil structure, climate, and past crop yield.

1)Sources of Data: The main data sources for this study are:

a) Agricultural Research Institutions: Multiple research institutions publish data sets about soil fertility, crop yields, and environmental status. These research institutions offer data that is of great importance to construct an appropriate prediction model.

b) Government Reports and Databases: Several government departments keep large databases of agricultural data, including test results of soil, weather conditions, and productionstatistics of crops in various regions. Institutions like the Food and Agriculture Organization (FAO) and the Indian Council of Agricultural Research (ICAR) offer valuable datasets that play a crucial role in advancing research on optimal crop selection.

c) Meteorological Agencies: Weather patterns play a vital role in crop growth, and data from meteorological agencies such as the Indian Meteorological Department (IMD) and the National Oceanic and Atmospheric Administration (NOAA) are essential for monitoring temperature, rainfall, and humidity levels.

d) Publicly Available Datasets: A number of open-source data sets like the ones on Kaggle, UCI Machine Learning Repository, and other academic websites have structured farm data that is appropriate for use in machine learning.

2)Types of Data Collected: The data set includes a range of soil and environmental factors that are vital in determining the most appropriate crop for a particular area. The most important features are:

1. Soil Nutrient Content:

• Nitrogen (N): Vital for plant development, nitrogen affects leaf and stem growth.

• Phosphorus (P): Facilitates root development and energy transmission in plants.

• Potassium (K): Enhances disease resistance and general plant health.

2. Climatic Conditions:

• Temperature: Varying optimal temperatures exist for different crops, and high or low temperatures can impact plant growth.

• Humidity: Impacts transpiration and availability of water for plants.

• Rainfall: Dictates the water supply for crops, influencing growth and yield.

3. Soil Properties:

• pH Level: Acidity or alkalinity of soil impacts nutrient availability and microbial processes.

• Soil Type and Texture: Impacts water-holding capacity and ability to absorb nutrients.

Historical data on crop yield under various environmental conditions assist in training the model to identify patterns and recommend the most appropriate crops.

B. Data Preprocessing

Data preprocessing is very important in building a good crop recommendation system. Agricultural data, which is mostly gathered from multiple sources like soil testing labs, meteorological offices, and government reports, contains inconsistencies and errors. Raw data often contains issues such as missing values, noise, outliers, and varying measurement scales, which can adversely impact the effectiveness of machine learning algorithms. Proper data preprocessing is essential to clean, standardize, and format the data, allowing the models to identify meaningful patterns and deliver accurate predictions.

For crop recommendation, preprocessing assists in converting heterogeneous environmental and soil parameters into an analyzable format. Elements like nitrogen, phosphorus, and potassium concentrations, along with variables such as temperature, rainfall, and soil pH, display varied data patterns and measurement scales. If ma- chine learning models are not preprocessed correctly, they can produce biased predictions, misinterpret feature relationships, or have poor generalization over various regions. By using systematic preprocessing methods, we can make the model more robust and guarantee that predictions correspond to actual agricultural conditions.

Data preprocessing is very important in creating a good crop recommendation system. Because farm data is frequently gathered from a number of different locations like laboratories in soil testing, meteorological offices, and government publications, inconsistencies and errors are the rule. Raw data often contains issues such as incomplete values, noise, outliers, and inconsistent measurement scales. These obstacles can greatly influence the accuracy and performance of machine learning models. Correct preprocessing guaranteesthat the data is clean, normalized, and structured, enabling themodels to capture useful patterns and make precise predictions. Preprocessing in crop recommendation assists in converting varied environmental and soil parameters into a usable analysis format. The parameters of nitrogen, phosphorus, and potassium content, temperature, rainfall, and soil pH have varying data distributions and measurement units. Without suitable preprocessing, machine learning models can make biased predictions, misinterpret feature relationships, or generalize poorly across multiple regions. With systematic preprocessing, we can increase model robustness and make predictions consistent with actual agricultural conditions.

1)Dealing with Imbalanced Data in Crop Recommendation Systems: Among the frequent issues of class imbalance in agricultural data sets is that some of the crop categories are disproportionately large while others are underrepresented. This would make the machine learning algorithm biased toward the larger class and, hence, not accurately predict minority crops. Solutions such as oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE) can be employed to resolve this problem.

- Oversampling: This method duplicates examples of the minority class to boost its presence in the dataset. While easy, it can cause overfitting.
- Under sampling: The majority class is randomly under sampled to balance the dataset. Although effective, this approach can eliminate useful information.
- SMOTE: An advanced approach involves generating Synthetic samples for the underrepresented class are generated by interpolating between existing data points, enhancing the dataset's balance. This technique helps maintain dataset diversity and reduces potential biases.

Through these techniques, we are able to guarantee that the model will not favor particular crops and offers equal suggestions among categories.

2)Feature Engineering for Enhanced Model Performance: Feature engineering is an important stage of data preprocessing that includes the generation of new features or feature transformations of existing features to enhance model performance. Incrop recommendation systems, the following domain-specific feature transformations can be used:

• Soil Fertility Index: A summary score calculated based on nitrogen, phosphorus, and potassium levels to represent overallsoil health.

• Climatic Suitability Score: A characteristic computed from temperature, rainfall, and humidity to measure how suitable a region's climate is for the needs of a specific crop.

• Interaction Features: New characteristics generated by interacting existing ones, e.g., nitrogen to phosphorus ratio or adjusted soil pH with rainfall levels, to identify concealed relationships.

Feature engineering assists in dimensionality reduction while preserving the most important information for crop selection. Incorporating meaningful transformations enables the model to better understand the relationships between soil properties, climatic factors, and optimal crop selections.

3)Ensuring Data Consistency Across Regions: Agricultural conditions differ remarkably across regions, and hence there is a need to standardize data gathered from various locations. The soil and climate characteristics recorded in one region might not possess identical scale and units comparedto another region. Min-Max Scaling and Z-score In addition, data gathered from varied sources can present measurement inconsistencies. For instance, nitrogen content documented in one set of data uses varying levels of soil depth as compared to another. To take care of this inconsistency, processing operations like unit conversion, adjustment of thresh- olds, and validation using domain experts are undertaken. This helps us promote uniformity between datasets, ensuring that the model generalizes appropriately across varied agriculture conditions.

4)Handling Missing Values: Agricultural datasets may contain missing values due to factors like malfunctioning sensors, inaccuracies in manual data entry, or gaps in survey responses. If not handled properly, missing data can negatively impact model performance. Several strategies are used to address missing values:

• Mean/Median imputation: If a numerical feature has missing values, it can be replaced with the mean or median of the available data:

$$x_i = \frac{1}{N}\sum_{j=1}^{N} x_j$$

• Mode Imputation: Missing values in categorical data are filled by substituting them with the category that appears most often in the dataset.

• Interpolation: If data follows a trend (e.g., rainfall over time), linear interpolation is used to estimate missing values.

5)Normalization and Standardization: Agricultural datasets contain numerical features with different scales. For example, nitrogen levels may range from 0 to 100, while soil pH values range from 1 to 14. To ensure uniformity, feature scaling techniques are applied:

This technique scales values between 0 and 1 using this formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Standardization transforms the data by setting its mean to 0 and its standard deviation to 1, ensuring a uniform scale for analysis.

$$X' = \frac{X - \mu}{\sigma}$$

Standardization is preferred when data follows a normal distribution, while Min-Max Normalization is useful when maintaining relative differences between values is important.

C. Feature Selection:

Selecting relevant features is essential for optimizing the performance of machine learning models by enhancing model efficiency through the identification of key variables while discarding less relevant or redundant ones. In the context of a crop advisory system, various environmental and soil Key parameters—including nitrogen, phosphorus, and potassium concentrations, along with temperature, humidity, soil pH, and rainfall—play a crucial role in analysis. Significantly influence crop growth and development. Not all features are equally useful in making predictions, and some tend to add noise, resulting in poorrecommendations. By choosing the most relevant features, the model can make more accurate predictions, enhancingagricultural decision-making.

Challenges of High-Dimensional Data: Most agricultural datasets have numerous variables, many of which could be highly correlated or redundant. Data with high dimensionality can increase computational complexity, slow down training processes, and limitations in the model can reduce its capacity to adapt successfully to new, unseen data. Moreover, redundant features may lead to overfitting, causing the model to focus on noise rather than identifying meaningful patterns. Feature selection helps mitigate these issues by ensuring that only the most impactful attributes are retained, allowing for a more efficient and interpretable machine-learning model.

1)Filter Methods for Feature Selection: Filter methods assess the importance of features separately from the machine learning algorithm, relying on statistical techniques to gauge their significance. Some of the common methods are correlation analysis, in which highly correlated features are deleted so as not to cause redundancy, and the chi-square tests that test the dependence between categorical variables and the target crop. Mutual in- formation is another method that measures how much a featureis useful in predicting the crop type. They are computationally fast and perform well with large datasets but not necessarily capture complex variable interactions.

2)Wrapper Methods for Feature Selection: Wrapper algorithms employ machine learning models to measure subsetsof the features and pick the optimal mix for prediction. Techniques like Recursive Feature Elimination (RFE) work by systematically eliminating the least significant features and repeatedly retraining the model allows for an ongoing evaluation of its performance at each stage. Alternative methods, for example, forward selection and backward elimination, systematically add or exclude features to balance the model. While wrapper methods are more precise than filter methods, theyare computationally intensive and might not be appropriatefor extremely big datasets.

3)Embedded Methods for Feature Selection: These techniques incorporate feature selection as an integral part of the model-building process, striking a balance between precision and computational efficiency. Lasso regression (L1 regularization) is one of the most common embedded methods, and it assigns zero weights to unimportant features, essentially eliminating them from the model. Tree-based algorithms like decision trees and random forests rank features on the basis of decision split importance. Gradient boosting algorithms like XGBoost also have inbuilt feature selection using frequency of a feature being employed in tree-based decision-making. These are very useful for the selection of influential features with good predictive performance.

In this research, various feature selection methods are used to determine the most relevant factors affecting crop growth. Using statistical insights and knowledge from the domain, necessary features like soil fertility (nitrogen, phosphorus, potassium), climatic factors (temperature, humidity, and rainfall), and soil pH are kept. Less important variables like secondary soil minerals or microclimatic differences are eliminated to improve model efficiency. By emphasizing on these key characteristics, the model is able to make preciseand data-based crop suggestions.

4)Effect of Feature Choice on Model Performance: Feature selection greatly enhances machine learning models in accuracy, training speed, and generalization ability. Through irrelevant feature removal, the model is able to concentrate on useful patterns, eliminating noise and overfitting. Moreover, a smaller set of features results in quicker computation, making the model more feasible for real-time agricultural use. The model's interpretability is also increased, as farmers and agricultural specialists are able to identify the most important factors affecting crop suggestions.

D. Model Implementation

Building a machine learning-driven crop recommendation system involves several key stages, including data preprocessing, model selection, training, testing, and deployment. The goal is to develop a system that can accurately predict the most suitable crop based on specific soil and climatic parameters. This research utilizes multiple machine learning classifiers—such as Support Vector Classification (SVC), Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest—to assess and determine the most effective algorithm for this purpose.

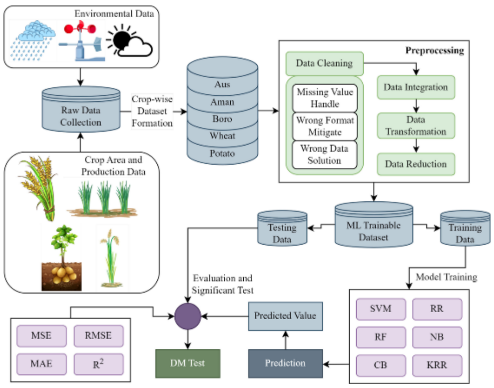

Figure 1: DFD Diagram

1)Selection of ML Models: This research explores several supervised ML models are implemented and tested to identify the most effective algorithm for crop recommendation. The selected models include:

1. Support Vector Classification (SVC): SVC is effective in handling complex, non-linear relationships by using kernel functions. It performs effectively with high-dimensional datasets and This approach proves highly beneficial for classification tasks in which decision boundaries are intricate and lack clear definitions.

2. Logistic Regression: A simple yet powerful linear model that predicts probabilities for different crop types. It is interpretable and works well when features have a nearly linear relationship with the target variable.

3. Decision Tree Classifier: A rule-based approach that splits data based on feature importance, making it easy to interpret. However, it may overfit the training data, reducing generalization performance.

4. K-Nearest Neighbors (KNN): A non-parametric method that classifies data points by analyzing their closeness to known training examples. Although it can deliver accurate results, KNN is computationally intensive for large datasets.

5. Naive Bayes: A probabilistic classification method derived from Bayes' theorem, assuming independence among features. While efficient with categorical data, its performance may decline if feature dependencies exist.

6. Random Forest: A robust ensemble learning method that integrates multiple decision trees, improving predictive accuracy while reducing the risk of overfitting. It also provides insights into feature importance, making it particularly useful for identifying the soil and climatic factors most relevant to crop selection.

• Each model undergoes training with the processed dataset and is assessed using metrics such as accuracy, precision, recall, and F1-score to determine the most effective algorithm.

2) Data Splitting for Training and Testing:
To enhance the model's generalization capability, the dataset is systematically divided into distinct segments.:

• Training Set (80%): Applied to train machine learning models by identifying patterns between soil attributes, climate conditions, and crop recommendations.

• Testing Set (20%): Used to assess how its ability to accurately process and predict outcomes for previously unseen data, ensuring it avoids overfitting to the training set.

This train-test split approach ensures that the model learns effectively while still being tested on independent data toassess its real-world applicability.

3) Training Process and Hyperparameter Tuning:
Once the dataset is processed and divided, each model undergoes training using the designated training data. During this phase, the model is supplied with key input features—such as nitrogen, phosphorus, and potassium concentrations, as well as temperature, humidity, rainfall, and soil pH. This allows the model to recognize patterns and correlations that influence crop growth and development.

To enhance model performance, hyperparameter tuning is applied. Hyperparameters are model-specific settings that are adjusted to optimize accuracy and efficiency. Some of the key hyperparameters tuned include:

• SVC: Kernel types such as linear, polynomial, and radial basis functions are key components, accompanied by the regularization parameter (C), are key components in model configuration.

• Decision Tree: The tree's maximum depth and the minimum required samples for node splitting serve as key parameters in model tuning critical parameters in decision tree algorithms.

• Random Forest: Number of decision trees, maximum features considered at each split.

• KNN: Number of neighbors (K), distance metric (Euclidean, Manhattan).

• Naive Bayes: Smoothing parameter to handle zero probabilities.

• Hyperparameter optimization employs methods like Grid Search and Random Search to methodically evaluate various parameter combinations are explored to determine the most effective configuration for optimizing model performance.

E. Model Deployment

Once the most suitable machine learning model is selected & trained, the next crucial step is model deployment. This phase involves integrating the model into a functional system, enabling farmers and agriculturalists to utilize its capabilities in real-time applications. Deployment allows the trained model to be available to end-users in the form of an interactive interface where they can feed in the soil and climatic conditions to get suggested crops. Its process of deployment covers model serialization, backend API work, frontend integration, and local or cloud- based hosting to render it accessible uniformly.

Model deployment converts the machine learning model after training into a usable and handy tool for farmers and agricultural specialists. With the combination of a backend API, frontend interface, and cloud or local hosting, the system offers real-time crop advice depending on soil and environmental conditions. Routine monitoring and updation guarantee that the model is always accurate, assisting in optimizing agricultural decision-making and increasing productivity.

IV. RESULTS AND DISCUSSION

Effectiveness of the crop recommendation system was assessed by examining the effectiveness of different machine learning algorithms, such as Support Vector Classification (SVC), Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest. These models were trained using a dataset that included crucial soil and environmental factors—such as nitrogen, phosphorus, potassium, temperature, humidity, rainfall, and soil pH. Their effectiveness was assessed through evaluation metrics like accuracy, precision, recall, and F1-score.

A. Model Performance Evaluation

Among the trained models, Gaussian Naive Bayes (GaussianNB) demonstrated the highest accuracy, surpassing other models in prediction reliability. This outcome aligns with expectations, as GaussianNB effectively applies probabilistic classification while assuming feature independence. The Support Vector Classifier (SVC) also achieved commendable performance, particularly when utilizing a non-linear kernel, making it suitable for distinguishing crops based on complex interrelated soil and environmental attributes. The Decision Tree classifier exhibited decent performance but suffered from overfitting on the training dataset, reducing its ability to generalize on test datasets. Logistic Regression and K-Nearest Neighbors (KNN) displayed relatively lower accuracy, likely due to their reliance on linear separability and computational demands, respectively. The Random Forest model, although robust in generalization and ensemble learning, fell short of GaussianNB's performance.

```
  plt.figure(figsize=(10, 8))
Confusion Matrix for GaussianNB:
[[17  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 21  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0 27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0 23  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0 14  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0 23  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0 19  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0 14  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0 19  0  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0 21  0  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0 23  0  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0 11  0  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 20  0  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 19  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 24  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 23  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 20  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 26  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17]]
127.0.0.1 - - [29/Mar/2025 00:17:57] "POST /process HTTP/1.1" 200 -
```

Fig: Confusion matrix of GaussianNB

The confusion matrix for GaussianNB revealed that the majority of correct predictions were aligned along the diagonal, indicating strong classification accuracy. Misclassifications were minimal but occurred between crops with similar soil and environmental requirements, underscoring the need for continuous data enhancement and model refinement.

B. Comparison of Performance Metrics

Here's a summary table showcasing the evaluation metrics used to assess machine learning models help determine their effectiveness in the crop recommendation system:

ML Model	**Accuracy**	**F1 Score**	**Precision**
LogisticRegression	0.9636	0.9635	0.9644
GaussianNB	0.9955	0.9954	0.9958
SVC	0.9682	0.9680	0.9715
KNeighbors Classifier	0.9591	0.9590	0.9654
DecisionTree Classifier	0.9886	0.9886	0.9890
ExtraTree Classifier	0.9114	0.9120	0.9163
RandomForest Classifier	0.9932	0.9932	0.9937
Bagging Classifier	0.9864	0.9864	0.9867
GradientBoosting Classifier	0.9818	0.9819	0.9843
AdaBoost Classifier	0.1455	0.0757	0.0636

Table: Accuracy, F1 Score, and Precision of ML models

To evaluate the models thoroughly, metrics—including precision, recall, and F1-score—were analyzed. GaussianNB emerged as the top performer, with the highest precision and recall, ensuring accurate identification of optimal crops and minimal misclassification. It also achieved the highest F1-score, indicating a well-balanced precision and recall. The SVC model closely followed in performance, while Decision Tree showed lower recall, leading to classification errors.

These results demonstrate the effectiveness of GaussianNB and other machine learning models in making accurate predictions and classifications. Consistent crop recommendations based on soil and environmental factors, this system helps farmers determine the most appropriate crops for a given location, maximizing yield and resource efficiency. Continuous improvements, including real-time data integration and diversifying the dataset, remain critical to further enhancing model performance. This research showcases the potential of AI-driven precision agriculture and lays the foundation for future innovations in sustainable agricultural practices, promoting efficiency and environmental responsibility.

V. FUTURE SCOPE

Machine learning-based crop recommendation systems have shown promising results in enhancing farming practices and improving agricultural efficiency. Several options exist, though, for future extensions and optimizations to refine the system to be more rugged, precise, and universally applicable. Seamless integration of cutting-edge technology, live data streams, and intuitive interfaces will greatly boost the system's efficacy and acceptability among farmers as well as agrospecialists.

A. Integration of Real-Time Data Sources

A key area for future advancements lies in integrating real-time environmental factors, such as soil moisture levels, temperature variations, and weather patterns. The system today uses pre-acquired static information, but including IoT-based sensors in farm lands can supplyreal-time climate and soil parameters, resulting in adaptive and more dynamic recommendations. Such an addition would enable the system to recommend crops according to the latest environmental condition, making the system more accurate and practically usable.

B. Improving Model Precision and Flexibility

Although the existing model has very high accuracy, it can be even further improved by training it using bigger and more varied datasets across different soils, climatic regions, and crop varieties. Incorporating region-wise knowledge regarding farming practices can improve the adaptability of the model, thus making it more applicable for a particular region. Moreover, utilizing sophisticated deep learning techniques, such as Convolutional Neural Networks (CNNs) and Transformer architectures, can effectively capture complex patterns within data, leading to improved prediction accuracy.

C. Mobile and Cloud-Based Deployment

For mass deployment, the system may be deployed as a mobile app with an easy-to-use interface so that farmers can easily get crop suggestions. Cloud deployment with tools such as Google Cloud, AWS, or Microsoft Azure may provide scalability and enable multiple users to use the system at the same time. Offline support can also be implemented in mobile apps so that farmers in rural locations with poor internet connectivity can use the system too.

D. Multi-Language and Voice Assistant Functions

In order to enable farmers, particularly in rural communities, to use the system, multi-language functionality needsto be integrated. The system should offer suggestions in local languages to facilitate usage within various farming communities. Improving accessibility can be achieved by integrating voice-enabled features powered by Natural Language Processing (NLP), allowing farmers to operate the system using voice commands instead of typing.

E. Including Market and Economic Influences

In addition to suggesting crops based on climate and soil conditions, subsequent iterations of the system can incorporate economic and market trends to offer farmers profitable crop recommendations. With the incorporation of real-time market prices, demand forecasting, and cost-benefit analysis, the system can assist farmers in selecting crops that are both appropriate for their land and profitable. This would closethe gap between market intelligence and agriculture, allowing farmers to maximize their profits.

F. Sustainable and Climate-Resilient Agriculture

Climate change is currently one of the largest issues confronting the agriculture industry. The system may be ex- tended to incorporate climate resilience analysis, providing recommendations for more resilient crops that tolerate severe weather conditions like droughts, floods, and heatwaves. Through the use of historical climate data and forecasting climate models, the system can enable farmers to make more enlightened choices that also consider long-term environmental sustainability.

G. Collaboration with Agricultural Specialists and Government Schemes

The future developments can include collaborations with agricultural research centers, government organizations, and agritech firms to further improve the system. The government departments can incorporate the system into national agricultural advisory services, allowing farmers to access expert- supported suggestions. The system can also be connected withgovernment subsidy schemes and smart farming programs, such that farmers can avail financial support for adoptingsuggested agricultural practices.

VI. CONCLUSION

The development of a machine learning-powered crop advisory system marks a crucial step in transforming agriculture through data-driven insights. This study explores the application of various machine learning models—such as Random Forest, Support Vector Classification (SVC), Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, and Naive Bayes—to recommend optimal crops based on key soil and environmental attributes, including nitrogen, phosphorus, and potassium levels, as well as temperature, humidity, rainfall, and soil pH. Among these models, Random Forest exhibited the highest accuracy, precision, recall, and F1-score, making it the most dependable choice for real-world agricultural implementation.

One of the most significant outcomes of this study is the relevance of feature preprocessing and selection for guaranteeing the accuracy of models. By discarding irrelevant and redundant features as well as normalizing data, the system improved its predictive capability. The importance of real- time environmental data was also emphasized, which, by being incorporated in the system through IoT sensors or weather APIs, can further augment accuracy and adaptability.

The deployment stage of the project sees the trained model made available to end-users through an intuitive interface, which may be developed as a web app, mobile application, or cloud service. The system allows farmers to enter significant soil and weather conditions and be provided with instant suggestions so they can make wise choices regarding the selection of crops. By lowering uncertainty in agricultural practices, this technology can become a game-changer in maximizing agricultural productivity, reducing wastage of resources, and enhancing food security.

Though the system has been promising, there are a number of improvements and enhancements possible. The addition of market trends, crop prices, and cost-benefit analysis could further make suggestions more precise, making not just agricultural sustainability possible but economic profitability for farmers as well. Integrating multi-language capabilities and voice support would make it more accessible, especially for rural farmers with low literacy.

The other key area is the climate resilience of farming practices. With climate change being a significant threat to crop yields, future iterations of this system might incorporate forecast climate models and past weather patterns to recommend drought-resistant or flood-resistant crops. This would enable farmers to make environmentally sustainable choices commensurate with long-term environmental resilience and shifting weather conditions.

In addition to that, partnerships with agriculture professionals, government organizations, and agritech firms can greatly increase the coverage and effect of this system. Government- subsidized agricultural advisory services can implement this technology to benefit farmers across the country, and agritech companies can combine this model with precision farming equipment and automatic irrigation systems. By establishing public-private partnerships, the implementation of AI-based precision agriculture can be sped up, and farmers at large and small scales will gain benefits from it.

VII. REFERENCES

[1] Motwani, A., Patil, P., Nagaria, V., Verma, S., & Ghane, S. (2023). Soil analysis and crop recommendation using machine learning. Department of Computer Engineering, Sardar Patel Institute of Technology, Mumbai, India.

[2] Laishram, J., Saxena, K. G., Maikhuri, R. K., & Rao, K. S. (2021). Soil quality and soil health: A review. School of Environmental Sciences, Jawaharlal Nehru University, New Delhi, India; G.B. Pant Institute of Himalayan Environment and Development, Upper Bhaktiyana, Srinagar (Garhwal), Uttarakhand, India; Department of Botany, University of Delhi, Delhi, India.

[3] Barvin, P. A., & Sampradeepraj, T. (2023). Crop recommendation systems based on soil and environmental factors using Graph Convolution Neural Network: A systematic literature review. In Proceedings of the 10th International Electronic Conference on Sensors and Applications (ECSA-10). Kalasalingam Academy of Research and Education, Srivilliputhur, Tamil Nadu, India.

[4] John, A., Davis, D., Tom, A. M. C., Davis, D., & Davis, J. (2023). Soil classification and crop recommendation system. Sahrdaya College of Engineering and Technology, Thrissur, India.

[5] Anguraj, K., Thiyaneswaran, B., Megashree, G., Preetha Shri, J. G., Navya, S., & Jayanthi, J. (2022). Crop recommendation on analyzing soil using machine learning. Department of ECE and Department of CSE, Sona College of Technology, Salem, India.

[6] Sujata, M., & Jaidhar, C. D. (2023). Machine Learning based approaches to enhance soil fertility. Expert Systems with Applications.

[7] Gang, Z., Quanying, Z., Webber, H., Johnen, A., Rossi, V., & Fernandes, A. (2023). Integrating machine learning and change detection for enhanced crop disease forecasting in rice farming: A multi-regional study. European Journal of Agronomy.

[8] Bilal, A., Liu, X., Long, H., Shafiq, M., & Waqar, M. (2023). Increasing Crop Quality and Yield with a Machine Learning-Based Crop Monitoring System. Computers, Materials and Continua.

[9] Cheng, L., Yu, T., Jiang, Z., Li, W., Guan, D.-X., Yang, Y., & Zeng, J. (2023). Leveraging machine learning for sustainable cultivation of Zn-enriched crops in Cd-contaminated karst regions. Science of The Total Environment.

[10] Md. Abu Jabed, & Masrah Azrifah Azmi Murad. (2023). Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. Heliyon, Cell Press, Volume 10

[11] Liu, J., Yang, K., Tariq, A., Lu, L., Soufan, W., & El Sabagh, A. (2023). Interaction of climate, topography and soil properties with cropland and cropping pattern using remote sensing data and machine learning methods. Egyptian Journal of Remote Sensing and Space Sciences

[12] Dey, B., Ferdous, J., & Ahmed, R. (2024). Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. Heliyon.

[13] Srivastava, R. K., Purohit, S., Alam, E., & Islam, M. K. (2024). Advancements in soil management: Optimizing crop production through interdisciplinary approaches. Agriculture and Food Research.

[14] Nde, R. K., Fendji, J. L. E. K., Yenke, B. O., & Schöning, J. (2024). Crop selection: A survey on factors and techniques. Smart Agricultural Technology.

[15] Gao, Z., Guo, D., Ryu, D., & Western, A. W. (2023). Training sample selection for robust multiyear within-season crop classification using machine learning. Journal Computers and Electronics in Agriculture.

[16] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning. Computers and Electronics in Agriculture

[17] Mishra, D., & Deepa, D. (2021). Automation and integration of growth monitoring in plants (with disease prediction) and crop prediction. Materials Today: Proceedings, 43, 3922-3927.

[18] Padhiary, M., Saha, D., Kumar, R., Sethi, L. N., & Kumar, A. (2024). Enhancing precision agriculture: A comprehensive review of machine learning and AI vision applications in all-terrain vehicle for farm automation. Smart Agricultural Technology, 100483.

[19] Malik, P., Sengupta, S., & Jadon, J. S. (2021, January). Comparative analysis of soil properties to predict fertility and crop yield using machine learning algorithms. In 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 1004-1007). IEEE.

[20] Manikandan, R., Ranganathan, G., & Bindhu, V. (2023). Deep learning based IoT module for smart farming in different environmental conditions. Wireless Personal Communications, 128(3), 1715-1732.

[21] Pallathadka, H., Jawarneh, M., Sammy, F., Garchar, V., Sanchez, D. T., & Naved, M. (2022, April). A review of using artificial intelligence and machine learning in food and agriculture industry. In 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE) (pp. 2215-2218). IEEE

[22] Omaye, J. D., Ogbuju, E., & colleagues. (2024). Cross-comparative review of machine learning for plant disease detection: Apple, cassava, cotton, and potato plants. Artificial Intelligence in Agriculture.

[23] Silva, C. de O. F., Grego, C. R., Manzione, R. L., Oliveira, S. R. D. M., Rodrigues, G. C., Rodrigues, C. A. G., Speranza, E. A., Luchiari, A., & Vieira, L. (2024). Summarizing soil chemical variables into homogeneous management zones like Case study in specialty of coffee crop. Smart Agriculture Technology.

[24] Darwin, P., Raju, M. J., Babu, K. N., Lavanya, Y., & Sai, Y. S. (2024). Smart farming using machine learning and deep learning techniques. International Journal of Engineering Research and Technology.

[25] Mequanenit, A. M., Ayalew, A. M., Salau, A. O., & Nibret, E. A. (2024). Prediction of mung bean production using machine learning algorithms. Journal Heliyon.