

A FIELD PROJECT REPORT

on

“Crop Recommendation and Prediction”

Submitted

221FA04371

B. Rakesh

221FA04391

K. Sai Deepthi

221FA04419

M.Manohar

221FA04420

T.Bhargavi

Under the guidance of

B.Suvarna

Designation



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
VIGNAN'S FOUNDATION FOR SCIENCE, TECHNOLOGY AND RESEARCH Deemed
to be UNIVERSITY
Vadlamudi, Guntur.
ANDHRA PRADESH, INDIA, PIN-522213.



CERTIFICATE

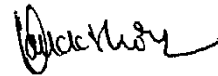
This is to certify that the Field Project entitled “**Crop Recommendation and Prediction**” That is being submitted by 221FA04371(Rakesh), 221FA04391(Sai Deepthi),221FA04419(Manohar) and 221FA04420 (Bhargavi) for partial fulfilment of Field Project is a Bonafide work carried out under the supervision of Ms. Dr. N. Sameera., Assistant Professor, Department of CSE.

Guide name& Signature


Dr. S. V. Phani Kumar

Assistant/Associate/Professor,
CSE

HOD, CSE



Dr.K.V. Krishna Kishore

Dean, SoCI



DECLARATION

We hereby declare that the Field Project entitled “**Crop Recommendation and Prediction**” That is being submitted by 221FA04371(Rakesh), 221FA04391(Sai Deepthi),221FA04419(Manohar) and 221FA04420 (Bhargavi) in partial fulfilment of Field Project course work. This is our original work, and this project has not formed the basis for the award of any degree. We have worked under the supervision of Ms. Dr. N. Sameera., Assistant Professor, Department of CSE.

By
221FA04371 (Rakesh),
221FA04391(Sai Deepthi),
221FA04419(Manohar),
221FA04420(Bhargavi)

Date: 15/10/2024

ABSTRACT

The study titled "Crop Recommendation and prediction" focuses on developing a highly accurate model for predicting right crop. Agriculture plays a vital role in India's economy, supporting millions of people. However, due to diverse climate and soil conditions, many farmers struggle with selecting the right crops often resulting in financial losses. To address these challenges, researchers have employed machine learning (ML) and deep learning (DL) techniques using soil and climate data to recommend crops. This paper surveys existing works using these technologies and conducts an analysis of algorithms, including Random Forest, Decision Tree, SVC, Logistic Regression, Gaussian NB for effective crop recommendation. Key soil and weather parameters—such as nitrogen, phosphorus, potassium, pH, temperature, humidity and rainfall—are critical to agricultural success. This research proposes a system that integrates data from various sources, applying ML models to analyze the factors and recommend the most profitable crops under current conditions.

TABLE OF CONTENTS

1. Introduction	1
1.1 Background and Importance of Crop Prediction	2
1.2 Overview of Machine Learning in Agriculture	2
1.3 Research Objectives and Scope	3
1.4 Challenges in Crop Prediction	5
1.5 Applications of ML in Crop prediction	7
2. Literature Survey	9
2.1 Previous Studies on Crop Prediction	10
3. Proposed System	12
3.1 Input dataset	14
3.1.1 Detailed features of dataset	14
3.2 Data Pre-processing	15
3.3 Model Building	16
3.4 Methodology of the system	18
3.5 Evaluation Metrics	21
3.6 Constraints	30
4. Implementation	31
4.1 Environment Setup	32
4.2 Sample code for preprocessing and Model Training and Testing	32
5. Experimentation and Result Analysis	35
6. Conclusion	39
7. References	41

8. LIST OF FIGURES

9.

Figure 2. Logistic Regression-Confusion Matrix	22
Figure 3. Naïve Bayes-Confusion Matrix	23
Figure 4. Support Vector Machine (SVM) -Confusion Matrix	24
Figure 5. Random Forest -Confusion Matrix	25
Figure 7. KNN -Confusion Matrix	27
Figure 8. Decision Tree -Confusion Matrix	28

CHAPTER-1

INTRODUCTION

1. INTRODUCTION

1.1 Background and Significance of Crop recommendation and prediction

Agriculture has always played a crucial role in human survival and economic development. As the world's population continues to rise, the demand for food is also increasing, prompting the need for more efficient and sustainable agricultural practices. To meet this challenge, crop recommendation systems and improved cultivation techniques are emerging as vital tools to optimize production, reduce resource waste, and ensure food security.

Agriculture has been the cornerstone of human civilization, providing essential food, raw materials, and employment for a significant portion of the global population. In countries like India, where agriculture is a major contributor to both livelihoods and the economy, crop cultivation plays an especially vital role. Traditionally, farmers relied on local knowledge, historical experience, and manual observation to make critical decisions about which crops to cultivate each season.

Significance of Crop cultivation and recommendation

Optimized Crop Selection: Crop recommendation systems use data on soil conditions, climate, and market trends to recommend the most suitable crops for a given region. This ensures that farmers grow crops that will yield the best results under specific environmental conditions.

Higher Yields: Through precise recommendations and informed cultivation practices, these systems help maximize crop yields by guiding farmers on optimal planting times, water use, and fertilizer application.

Sustainable Land Use: By recommending appropriate crops and cultivation practices based on soil health, crop rotation, and nutrient management, crop recommendation systems promote the long-term sustainability of farmland, preventing soil depletion and erosion.

Climate Resilience: Crop recommendation systems help farmers adapt to changing climate conditions by suggesting crops that are more resilient to drought, heat, and extreme weather. This ensures that agriculture remains viable in regions facing the impacts of climate change.

Meeting Global Food Demand: As the world's population continues to grow, there is increasing pressure on agricultural systems to produce more food. Crop recommendation systems help farmers increase yields and grow crops more efficiently, contributing to food security on a global scale.

Higher Income and Profitability: By recommending crops that are suited to local environmental conditions and market demand, farmers can achieve higher yields and sell their produce at better prices. This increases their profitability and economic stability.

Risk Reduction: Crop recommendation systems help farmers make data-driven decisions, reducing the risk of crop failure due to unfavourable environmental conditions or poor market choices. This is especially beneficial in regions where agriculture is highly susceptible to weather variability.

1.2 Overview of Machine Learning in Agriculture

Machine learning (ML) has emerged as a transformative technology in agriculture, driving innovation and efficiency in various farming practices. As agriculture faces mounting challenges such as climate change, soil degradation, water scarcity, and the need to feed a growing global population, ML offers data-driven solutions to optimize productivity, minimize resource use, and promote sustainable farming. Here's an overview of how machine learning is being utilized in the context of crop cultivation and recommendation:

Machine Learning Applications in Agriculture:

Precision Agriculture:

Precision agriculture involves managing variations in the field to maximize crop yields and resource efficiency. Machine learning (ML) models analyze diverse datasets such as soil characteristics, weather data, and satellite imagery to provide actionable insights. The core theory behind ML in precision agriculture is the use of supervised learning (where algorithms learn from historical data) and unsupervised learning (where patterns are identified without predefined labels). These techniques help segment fields into zones for specific treatments, optimizing resource use like fertilizers and irrigation based on real-time conditions.

Crop Health Monitoring and Disease Detection:

ML-based image recognition and computer vision techniques can classify and detect anomalies in crop images, allowing for early detection of diseases and pests. The underlying theory is to train models, often convolutional neural networks (CNNs), on large datasets of healthy and diseased plants. These models learn to recognize patterns such as leaf discoloration, lesions, or irregular growth that may signal disease or pest infestations. Early detection through ML minimizes crop loss and guides precise interventions.

Smart Irrigation Systems:

Smart irrigation leverages ML algorithms to predict the optimal amount and timing of water application based on various parameters like soil moisture, weather forecasts, and plant growth

stages. The theory relies on regression models (to predict the right water levels) and time-series analysis (to forecast future needs). ML models learn from historical data to determine irrigation schedules, improving water efficiency and reducing waste, especially in water-scarce regions.

Climate and Weather Forecasting:

ML models for weather forecasting rely on time-series analysis and regression models. These algorithms are trained on historical weather data, including temperature, humidity, wind speeds, and rainfall, to forecast future weather patterns. The theory behind these models is that they can recognize complex patterns and relationships within weather data, providing more accurate localized forecasts. This helps farmers make informed decisions on planting, irrigation, and harvesting schedules.

Market Insights and Crop Pricing:

ML models analyze market data, consumer preferences, and economic factors to predict crop prices and demand. Supervised learning models like random forests and support vector machines are trained on historical price data and other features like supply levels, global demand, and external economic conditions. Regression analysis predicts future crop prices, allowing farmers to time their sales optimally. This reduces market risks and maximizes profits.

1.3 Research Objectives and Scope

Research on accurate crop cultivation, recommendation and prediction may have the following goals:

Enhance Crop Yield Prediction Accuracy: Develop machine learning models that utilize soil parameters (e.g., nitrogen, phosphorus, potassium), weather conditions (e.g., rainfall, temperature, humidity), and historic crop data to improve the accuracy of crop yield predictions. This would help farmers make data-driven decisions to maximize yield in a specific region.

Develop Robust Crop Recommendation Models: Create predictive algorithms that recommend the most suitable crops for cultivation based on soil health, environmental conditions, and historic performance. These models aim to provide farmers with crop options that are optimized for their specific local conditions, leading to better yields and reduced crop failure risks.

Optimize Resource Allocation: Investigate how machine learning models can optimize the use of fertilizers, water, and pesticides by analyzing real-time field data and recommending resource-efficient practices. This would lead to sustainable farming practices and reduce input costs for farmers.

Adaptation to Climate Change: Develop models that account for the impact of climate variability, such as droughts and floods, on crop yield. This would help farmers adapt by recommending

resilient crop varieties that can withstand extreme weather conditions and ensure consistent productivity.

Promote Precision Agriculture: Explore how machine learning can enhance precision agriculture by providing farmers with field-specific recommendations regarding planting schedules, irrigation needs, and soil management. This would improve productivity and reduce environmental impact by minimizing resource wastage.

Expand Access to Predictive Tools in Resource-Limited Settings: Assess the potential of ML-driven crop recommendation and yield prediction systems that can be deployed in rural and under-resourced areas. These tools should provide affordable, easy-to-use systems that help small-scale farmers make informed decisions without requiring extensive technological expertise.

Minimize Bias and Ensure Generalization: Work on ensuring that predictive models are generalized across diverse regions and crop varieties, by incorporating datasets from different geographies, soil types, and climates. This will minimize bias and ensure model accuracy across different farming communities.

Integrate with Agricultural Practices and Government Initiatives: Investigate how ML-based crop recommendation and yield prediction tools can be integrated into existing agricultural policies, government programs, and farmer advisory services. This would provide seamless access to advanced farming techniques and government-backed financial support.

Research Scope:

Machine Learning Algorithms:

Explore various machine learning techniques, including Decision tree, Logistic regression, SVC, Random Forest, Gaussian NB, KNN

Implement both supervised learning (for crop recommendation) and unsupervised learning (for clustering similar agricultural regions based on soil and climate data).

Applications in Agriculture:

Agronomy: Use ML to predict the impact of soil nutrients, water levels, and fertilizer application on crop yield, providing precise recommendations for agronomic practices.

Climate Science: Investigate the role of climate factors, such as temperature, precipitation patterns, and extreme weather events, in crop selection and yield forecasts.

Data Sources:

Soil and Weather Data: Gather data from sensors, IoT devices, satellite imagery, and weather stations to track real-time field conditions.

Agricultural Databases: Use historical yield data, soil test results, and government crop surveys.

Farmer-Provided Data: Collect data directly from farmers regarding field conditions, crop health, and planting practices to enrich models.

Remote Sensing: Utilize satellite imagery to analyze large-scale agricultural fields for monitoring crop health and environmental changes.

Legal and Ethical Considerations:

Address privacy issues surrounding the collection of agricultural data, ensuring that the data ownership remains with farmers and that they have control over how their data is used.

Ensure the ethical use of ML systems in decision-making processes by providing farmers with transparent, interpretable predictions and recommendations.

Challenges and Limitations:

Data Quality and Availability: Identify the difficulties in accessing high-quality agricultural datasets, especially from resource-limited regions where sensor technologies may be unavailable.

Model Interpretability: Develop interpretable machine learning models that provide clear reasoning behind crop recommendations and yield predictions, making them understandable for farmers.

Scalability: Address the challenge of scaling ML models to cover diverse agricultural regions with varying soil types, weather conditions, and farming practices.

Model Evaluation:

Evaluate ML models using performance metrics such as accuracy, precision, recall, F1 score, and Root Mean Square Error (RMSE) to ensure reliability.

Perform cross-validation on different datasets to ensure the models generalize well to new farming environments.

Impact on Agricultural Systems:

Assess the impact of crop recommendation and yield prediction systems on farming practices, focusing on increased productivity, reduced environmental impact, and improved profitability for farmers.

Investigate how these systems can improve decision-making in supply chain management, crop pricing, and market forecasting.

Integration of Technology:

Explore the integration of ML-based predictive tools with existing agricultural technology platforms, such as farm management software, mobile apps for farmers, and government-backed agricultural advisory services.

Investigate the possibility of integrating with tele-farming and e-agriculture platforms that provide remote consultation and guidance to farmers.

1.4 Challenges in Stroke Prediction

1. Data Availability and Quality

Crop recommendation and prediction systems rely on vast amounts of data, including soil properties, weather patterns, satellite imagery, and historical crop performance. For machine learning (ML) models to make accurate predictions, they need high-quality data that covers various environmental and crop-related variables. In regions where such data is sparse or unavailable, ML models are limited in their ability to produce reliable recommendations. This leads to a phenomenon known as data insufficiency, where the lack of data prevents algorithms from identifying patterns or making informed predictions. Further, when data is incomplete or noisy (errors, gaps, or inconsistencies), it leads to biased or inaccurate models, which generalize poorly to real-world conditions. Poor data quality limits the ability of ML models to accurately understand the intricacies of specific environments, resulting in flawed recommendations.

2. Complexity of Agricultural Systems

Agricultural systems are inherently complex, with dynamic environmental factors such as temperature, precipitation, soil moisture, and nutrient availability constantly fluctuating. In machine learning, complex systems are often difficult to model because of the multivariate and non-linear relationships between different variables. In agriculture, a small change in one factor, like soil moisture, could have cascading effects on crop health and yield, depending on other conditions such as temperature and sunlight. ML models must consider these interactions, but the complexity and variability of agricultural systems make this difficult. While multi-factor models can help to a degree, overfitting is a risk when trying to account for every variable, meaning the model may perform well on historical data but poorly in new, unseen conditions. This makes it difficult to accurately predict outcomes under changing environmental factors, such as a sudden drought or unexpected pest infestation.

3. Model Accuracy and Generalization

Machine learning models, particularly those based on supervised learning, rely heavily on training data to make predictions. In agriculture, this training data often consists of historical crop performance under specific conditions. The challenge arises when these models are deployed in new environments or under different conditions, where they may fail to generalize well. This stems from the overfitting problem, where a model learns to perform exceptionally well on the training

data but lacks the flexibility to handle novel situations. For instance, a crop recommendation model trained in temperate regions might not perform well when deployed in tropical climates. Additionally, generalization errors occur when the model is unable to capture the underlying trends of new or evolving environmental factors. Without constant retraining with up-to-date data, these models risk becoming obsolete or providing inaccurate recommendations in real-world farming scenarios.

4. Integration of Diverse Data Sources

One of the major challenges in crop recommendation systems is the integration of different types of data, such as soil data, weather forecasts, satellite imagery, and market trends. These datasets are often heterogeneous, meaning they come in different formats, resolutions, and timescales. For example, satellite imagery may provide daily updates at a regional scale, while soil sensors offer continuous, localized data points. The theory behind this challenge relates to **data fusion**, which is the process of combining multiple sources of data to create a unified, comprehensive dataset that can be analyzed by machine learning algorithms. However, achieving seamless integration is difficult because each data source may have varying degrees of reliability and may be collected under different conditions. Errors in data fusion can lead to incomplete or misleading inputs, causing ML models to output inaccurate predictions. The challenge lies in creating algorithms that can harmonize these different data streams and ensure consistency across the diverse types of information.

5. Technological Barriers for Smallholder Farmers

Although ML-based crop recommendation systems promise significant benefits, their adoption is often hindered by technological barriers faced by smallholder farmers. The **digital divide** theory explains that access to technology is unequal across different socio-economic groups. Smallholder farmers in developing regions often lack the infrastructure (e.g., internet connectivity, smartphones, or high-speed computing) needed to utilize these systems. Additionally, even when technology is available, the **human-computer interaction (HCI)** aspect poses challenges. These systems require a certain level of digital literacy and technical expertise, which many farmers do not possess. From a theoretical perspective, this challenge also involves the **usability gap**, where advanced technologies fail to cater to users with limited technological skills or resources. In agricultural contexts, this gap can prevent farmers from benefiting from data-driven recommendations, despite their potential value.

6. Environmental and Climate Change Challenges

Climate change adds a layer of uncertainty and complexity to crop cultivation predictions. The chaos theory in environmental science explains how small changes in initial conditions can lead to significant and unpredictable shifts in weather patterns and environmental conditions. As climate change accelerates, weather patterns become more erratic, and ML models trained on historical data may fail to predict future conditions. The core challenge lies in the stationarity assumption—many machine learning models assume that future data will follow patterns seen in the past. However, with climate change, this assumption no longer holds. Droughts, floods, and temperature extremes are becoming more frequent and harder to predict, leading to breakdowns in models that were not designed to account for such variability. This results in inaccurate crop predictions, which can lead to poor harvest outcomes or even crop failure in extreme cases.

7. Cultural and Social Factors

In theory, cultural and social factors play a critical role in how technology is adopted within agricultural systems. Social behaviour theory explains how individuals' decisions are influenced by traditions, norms, and community practices. In farming communities, long-established practices often dictate crop choices, and even the most accurate machine learning recommendations may be disregarded if they conflict with traditional knowledge or farming methods. This leads to technology resistance, where farmers might be skeptical of modern tools or advice that deviate from their tried-and-true methods. Additionally, social hierarchies and peer influence can further reinforce resistance to change, as farmers often seek approval from their community or local leaders before adopting new methods.

8. Economic Constraints

The economic feasibility theory addresses the limitations that arise when farmers cannot afford the inputs required to follow through on ML-based recommendations. For example, a system may recommend planting a drought-resistant crop variety, but the seeds for this variety might be too expensive for small-scale farmers. The resource limitation theory also applies here—while the recommendation itself may be scientifically sound, farmers often lack the financial or logistical resources to implement it, especially in developing regions. Economic constraints also interact with market volatility; for example, even if a crop recommendation is economically viable under current market conditions, sudden shifts in demand or price fluctuations can make the recommended crop unprofitable, resulting in financial losses for the farmer.

9. Scalability and Localization

A significant challenge with crop recommendation systems is their scalability across different geographical regions with diverse environmental conditions. The localization theory in technology adoption suggests that technologies must be adapted to local contexts to be effective. For machine learning models, this means that generalized models trained on global datasets may not perform well in specific localities due to regional variations in soil, climate, and farming practices. Model scalability is also limited by the heterogeneity of environmental factors. As systems are scaled up to larger regions, it becomes difficult to maintain the accuracy and relevance of recommendations without compromising on localization. Transfer learning—a technique where models trained on one dataset are fine-tuned for a new, related task—can offer some solutions, but the variability in agriculture often requires completely custom models for each region.

Important Uses of Machine Learning in the Identification of crop cultivation and recommendation

1. Crop Recommendation Based on Soil and Weather Conditions

ML algorithms analyze data on soil health (e.g., nutrient levels, pH, moisture) and environmental factors (e.g., temperature, rainfall, humidity) to recommend the most suitable crops for a specific region. By evaluating these parameters, ML models can suggest crops that are more likely to thrive in the current conditions, reducing the risk of crop failure and maximizing yield potential.

2. Predicting Crop Yields

ML models predict crop yields by analyzing multiple factors such as soil characteristics, weather patterns, water availability, and historical yield data. By integrating this data, these models can forecast expected crop output, allowing farmers to plan resources and finances accordingly.

3. Soil Health and Nutrient Deficiency Detection

ML models analyze soil samples to identify deficiencies in nutrients like nitrogen, phosphorus, and potassium, which are essential for crop growth. By detecting nutrient imbalances, these systems help farmers apply the right fertilizers and improve soil health.

4. Precision Agriculture and Resource Optimization

Machine Learning plays a key role in precision agriculture, where farmers use ML-powered tools to optimize the use of resources like water, fertilizers, and pesticides. By analyzing real-time data from IoT sensors in the field, ML models recommend precise resource allocations based on the needs of the crops, minimizing waste and enhancing efficiency.

5. Climate Change Adaptation

ML models help farmers adapt to changing climate conditions by recommending resilient crop varieties that are better suited to extreme weather patterns. By analyzing past weather data and current trends, ML systems can suggest crops that are likely to survive droughts, floods, or unseasonable temperatures.

6. Pest and Disease Detection

ML-based systems, often combined with image recognition technologies, can detect early signs of pests or diseases in crops by analyzing images or sensor data. Early detection helps farmers take preventive actions and reduce crop damage.

7. Crop Rotation Recommendations

ML models analyze the nutrient depletion patterns in soil and recommend crop rotation strategies to restore soil health and prevent over-exploitation. By ensuring that different crops are planted in a sequence that replenishes nutrients, ML helps maintain soil fertility and improve long-term productivity.

8. Yield Forecasting and Financial Planning

ML models provide accurate crop yield forecasts based on environmental factors, soil conditions, and farming practices. This helps farmers plan their finances and manage resources efficiently, making it easier to secure loans, invest in farming inputs, or plan market strategies.

Benefits of ML in Brain Stroke Detection

Improved Accuracy in Crop Selection

ML algorithms analyze a wide range of factors, such as soil health, climate conditions, nutrient levels, and water availability, to recommend the most suitable crops for a specific region. This data-driven approach helps farmers choose crops that are more likely to thrive under current environmental conditions, leading to better productivity and profitability. This is particularly useful in regions with diverse climates and soil types, like India, where traditional methods might not always yield optimal results.

Higher Crop Yields

By accurately predicting the performance of crops based on historical and real-time data, ML models help farmers maximize yield. These systems assess factors such as weather patterns, soil composition, water availability, and nutrient levels, enabling farmers to optimize their farming practices. Yield prediction models forecast expected crop output, helping farmers make informed decisions that improve productivity and reduce crop failures.

Efficient Resource Management

ML helps in managing agricultural resources like water, fertilizers, and pesticides more effectively. By predicting crop needs, ML systems can suggest optimal amounts and timings for applying these

resources, preventing overuse or underuse. This not only conserves resources but also reduces costs, improves soil health, and minimizes the environmental impact of farming practices.

Adaptation to Climate Change

With changing climate patterns affecting crop performance, ML models can predict how different crops will respond to specific environmental conditions. These systems allow farmers to adjust planting schedules, switch to more resilient crop varieties, and take preventative actions based on weather predictions. This adaptive capability helps mitigate the risks associated with unpredictable climate conditions.

Early Detection of Diseases and Pests

ML systems, often combined with IoT sensors and satellite imagery, can detect early signs of crop diseases, pest infestations, or nutrient deficiencies. These systems analyze real-time data from the field and alert farmers about potential risks, allowing them to take corrective measures promptly. Early intervention minimizes crop damage and prevents large-scale losses.

Reduction of Input Costs

By optimizing resource use and predicting precise fertilizer and water requirements, ML reduces the costs associated with over-application of inputs. Precision agriculture, supported by ML, ensures that farmers apply just the right number of fertilizers, pesticides, and water, leading to lower operational costs and healthier crops.

Sustainable Farming Practices

ML-based recommendations promote sustainable farming by encouraging practices like crop rotation and diversification. These systems assess soil nutrient depletion and suggest crops that can restore soil health or require fewer inputs. By minimizing soil degradation and reducing environmental impact, ML helps create long-term sustainability in agriculture.

Increased Profit Margins

By optimizing every step of the farming process—from crop selection and resource allocation to pest control and yield prediction—ML enables farmers to boost their profit margins. Reduced costs, increased yields, and better market planning all contribute to more efficient and profitable farming operations.

CHAPTER-2 LITERATURE SURVEY

2 LITERATURE SURVEY

2.1 Previous Studies on Crop Cultivation, Recommendation, and Prediction

Recent advances in crop cultivation and recommendation systems highlight the growing role of machine learning (ML) models in agricultural optimization. These systems utilize soil and crop data to predict optimal crop choices based on various environmental and chemical factors, improving yield and sustainability. Several studies have explored ML-based approaches to recommend suitable crops, factoring in soil type, nutrients, climate conditions, and land characteristics.

Sharavani V et al. created a predictive model that recommends suitable crops based on soil types. Their model incorporated two datasets: a crop dataset and a soil dataset. The soil dataset consisted of 383 samples with 11 classes, capturing key chemical characteristics. The study applied various ML algorithms, including ensemble classifiers, Support Vector Machines (SVM), Random Forest, and Naive Bayes. Among these, SVM achieved the highest accuracy at 94.95%. A significant advantage of their proposed system was its ability to process real-time soil data, making it adaptable to various agricultural settings.

Rahman et al. developed a model to predict soil and land types and recommend crops accordingly. Their system also utilized two datasets: a crop dataset and a soil dataset. The ML techniques used included K-Nearest Neighbors (KNN), bagged trees, and SVM. SVM demonstrated the best results, achieving an average accuracy of 94.5%. Their system showed high efficiency and accuracy but required a larger dataset to train the models effectively, making it more resource-intensive.

Saranya et al. focused on classifying soil based on its macronutrient and micronutrient composition, which was then used to predict the most suitable crops. This model also employed soil and crop datasets, including the geographical and chemical characteristics of the soil. The ML algorithms tested were KNN, logistic regression, bagged tree, and SVM. SVM outperformed other models, reaching an accuracy of 96%. Despite its strong performance, the system's reliance on larger datasets posed challenges in terms of computational resources.

Mariappan et al. proposed a system that gathered data on soil factors such as pH, nitrogen (N), phosphorus (P), and potassium (K) levels, then mapped these data to suitable crops. Their study applied the KNN algorithm and achieved an accuracy of 89%. This model was tested across various cities in Chennai, illustrating its potential for localized crop recommendation. However, the accuracy was slightly lower compared to other models due to the simplicity of the algorithm.

Moraye et al. applied the Random Forest algorithm for crop yield prediction based on district, season, and crop type. Their model achieved an accuracy of 87%, demonstrating the viability of Random Forest in agricultural applications. While the accuracy was slightly lower than models utilizing SVM, Random Forest's strength lies in handling large, complex datasets, making it ideal for broader geographic or multi-crop predictions.

CHAPTER-3 PROPOSED SYSTEM

3 PROPOSED SYSTEM

A. Dataset: The crop recommendation dataset includes 7 features that capture crop and soil factors that recommend the crop. The target variable is "label," with categorical variables. This dataset includes categorical and numeric variables, which are pre-processed before model training.

B. Data Preprocessing: The following steps were applied for preprocessing:

Min-Max Scaler is used to scale/normalize the features in a dataset. It ensures that the values of each feature fall within a specified range, typically between 0 and 1.

Feature Scaling: Numerical variables such as N, P, K were scaled using Min-Max normalization to fit between 0 and 1.

C. Model Development: Several supervised models were evaluated for crop prediction:

Logistic Regression: Chosen for its interpretability and ease of implementation.

Random Forest: An ensemble model that helps in handling both categorical and continuous variables.

Support Vector Machine (SVM): Linear kernel SVM used to classify labels.

K-Nearest Neighbours (KNN): KNN classifier optimized with $k=7$.

Naive Bayes: Simple probabilistic model suitable for small datasets.

Decision Trees: This is rule-based classifier that split the dataset into branches based on feature values, leading to a decision.

D. Model Training: The dataset was split into 80% training and 20% test sets. K-fold cross-validation (with $k=7$) ensured model generalizability.

E. Model Evaluation: The following metrics were used to evaluate model performance:

Accuracy: The proportion of correct predictions. Precision, Recall, and F1-score: Evaluated the balance between true positives and false positives, especially for stroke cases.

Confusion Matrix and ROC-AUC: Used to visualize model predictions and assess sensitivity to minority class predictions.

F. Model Interpretation: The models were interpreted using:

Feature importance in Random Forest.

Confusion Matrix visualization for understanding prediction distributions.

H. Final Model Selection and Testing: The best-performing model was chosen based on validation metrics, with a focus on both accuracy and recall. The model was further tested on unseen data for generalization.

3.1 Input dataset

The dataset contains 7 features that describe crop and soil factors. Below are the features and their descriptions:

3.1.1 Detailed Features of the Dataset

N: Nitrogen content in the soil (numerical value).

P: Phosphorus content in the soil (numerical value).

K: Potassium content in the soil (numerical value).

temperature: Temperature of the environment (in degrees Celsius).

humidity: Humidity level in the environment (percentage).

ph: pH value of the soil (measure of acidity/alkalinity).

rainfall: Amount of rainfall (in mm).

label: The crop label (categorical value), which indicates the type of crop suitable for the given environmental and soil conditions.

3.2 Data Pre-processing

Data preprocessing is crucial to enhance the dataset's quality and ensure compatibility with machine learning algorithms. The following steps were performed:

Feature Scaling:

Min-Max Scaler is used to scale/normalize the features in a dataset. It ensures that the values of each feature fall within a specified range, typically between 0 and 1.

This scaling helps improve convergence during model training, especially for distance-based algorithms like KNN and SVM.

Data Splitting:

The dataset was split into 80% training and 20% testing sets to evaluate model performance on unseen data.

K-Fold Cross-Validation ($k = 7$) was employed on the training data to reduce the risk of overfitting and ensure robust model evaluation.

3.3 Model Building

Using the cleaned dataset, the model development portion of this study aimed to predict the crop. Various classifiers were evaluated for their effectiveness in addressing this classification problem.

Preparing Data

The dataset was first divided into two parts: features (X) and the target variable (y).

X included all relevant crop characteristics of N,P,K,Temperature,Humidity,Ph,Rainfall.

y represented the target variable, indicating the label(crop).

Feature scaling was applied using Standardization to ensure all features were on the same scale, which is essential for algorithms sensitive to feature magnitudes.

Data Division

The dataset was split into a training set (80%) and a testing set (20%).

This division allows the model to learn from the training data while providing an unbiased evaluation of its performance on unseen data.

Training of Models

Multiple models were trained and evaluated, including:

Logistic Regression:

A simple and interpretable model that estimates the predicted crop based on feature inputs. The logistic regression model was trained to predict stroke probabilities and classify based on a threshold.

Naive Bayes:

The Gaussian Naive Bayes model was employed due to its effectiveness with independent features. Each class's probability was calculated, with smoothing applied to prevent issues with zero probability for unseen feature combinations.

K-Nearest Neighbours (KNN):

The KNN classifier was trained to predict label based on the proximity of feature values in the training data. This model relies on distance metrics to classify new instances based on their similarity to training examples.

Support Vector Machine (SVM):

The linear SVM model was trained to find the optimal hyperplane that separates label cases in the feature space. This model is particularly useful for high-dimensional data and works well when the classes are linearly separable.

Decision Tree:

A decision tree classifier was built to predict crop by recursively partitioning the data based on feature values. This model is interpretable and allows for easy visualization of decision paths.

Random Forest:

An ensemble of decision trees was used, where each tree was trained on a random subset of the training data. This model improves predictive accuracy by aggregating the predictions of multiple trees to reduce overfitting.

After training, each model was used to predict the crop in the test set. The models' performances were evaluated based on:

Accuracy: Measures the overall correctness of predictions.

Precision: Indicates the proportion of true positives out of all predicted positives.

Recall: Represents how effectively the model identified all actual positive instances.

F1-Score: Balances precision and recall, especially valuable for datasets with class imbalance.

A confusion matrix was generated for each model to visualize the counts of true positive, true negative, false positive, and false negative predictions. This matrix provides insights into the strengths and weaknesses of each model, highlighting areas for improvement.

The evaluation showed that different models performed variably, with some achieving better accuracy and balance in class predictions than others. The Naive Bayes classifier and Random Forest models produced promising results, while the confusion matrix revealed specific challenges, such as predicting the label.

Methodology of the system

A. Architecture of the System

The proposed system architecture for predicting crop based on crop factor encompasses several interrelated steps: data collection, preprocessing, feature extraction, model training, and classification. The architecture consists of the following components:

Input Layer:

This layer collects agricultural and environmental information, including nutrient content and climatic conditions, such as N (Nitrogen), P (Phosphorus), K (Potassium), temperature, humidity, pH, and rainfall. These features serve as the foundation for predicting suitable crops based on the collected data.

Preprocessing Layer:

The collected data undergoes transformation and cleaning to ensure its suitability for machine learning algorithms. This step includes handling missing values, encoding categorical variables, and scaling numerical features.

Feature Extraction Layer:

Relevant features are identified and extracted for efficient classification. This layer retains important characteristics, such as N (Nitrogen), P (Phosphorus), K (Potassium), temperature, humidity, pH, and rainfall, while eliminating less significant variables. These key features are crucial for accurate crop prediction and recommendation.

Classifier Layer:

Various machine learning algorithms are employed to predict the most suitable crop for a given set of environmental and nutrient conditions. This includes models such as Random Forest, Decision Trees, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Neural Networks. Each model is trained using the extracted features, such as N, P, K, temperature, humidity, and rainfall, to enhance the accuracy of crop recommendations.

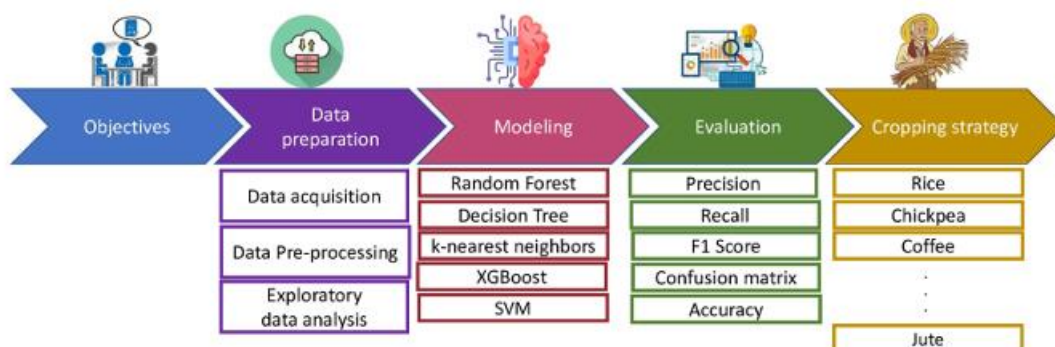


Figure 1. Architecture of the proposed system

Output Layer:

The system presents the classification outcome, indicating the crop based on the input data and model predictions.

B. Training and Preprocessing of Data

Data preprocessing is a crucial step to ensure that the dataset is appropriate for machine learning algorithms. The preprocessing techniques employed in this study include:

Data Cleaning:

Columns deemed unnecessary or redundant, such as "label," were removed from the dataset. This simplification aids in focusing on the most relevant features for prediction.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2200 entries, 0 to 2199
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype  
---  -
0   N                2200 non-null  int64  
1   P                2200 non-null  int64  
2   K                2200 non-null  int64  
3   temperature      2200 non-null  float64
4   humidity         2200 non-null  float64
5   ph               2200 non-null  float64
6   rainfall         2200 non-null  float64
7   label           2200 non-null  object  
dtypes: float64(4), int64(3), object(1)
memory usage: 137.6+ KB
```

Label Encoding:

The categorical variable "label" (representing crop names) was encoded into numerical formats compatible with machine learning models. This ensures that algorithms can effectively interpret categorical data. Similarly, the target variable (crop name) was transformed into numerical labels to enable the model to process the data efficiently for crop prediction.

For instance, in the dataset, the crop type "rice" was encoded as 0, allowing the model to use this numeric representation for learning and prediction. This label encoding step is crucial for ensuring smooth handling of categorical features in classification algorithms.

Feature Scaling:

Standardization techniques were applied to normalize the feature set, ensuring each feature contributes equally during model training.

Data Splitting:

The dataset was divided into a training set (80%) and a testing set (20%) to ensure that the model is evaluated on unseen data, allowing for a reliable assessment of its performance.

C. Feature Extraction

For this dataset, feature extraction involves selecting and transforming the input data into a smaller subset of relevant features for classification. After thorough analysis, the pertinent features such as N (Nitrogen), P (Phosphorus), K (Potassium), temperature, humidity, pH, and rainfall were retained. By focusing on these key variables, the model's predictive performance was enhanced, leading to more accurate crop recommendations, as indicated by the label column (e.g., rice).

This approach allows the model to identify the most critical factors that influence crop suitability, ensuring that the recommendations are based on the most impactful features.

D. Model Training

Various models were implemented to tackle the crop prediction problem, including:

Logistic Regression:

Chosen for its interpretability, this model estimates the probability of crop prediction based on input features.

Naive Bayes:

The Gaussian Naive Bayes classifier was utilized due to its efficiency with categorical and continuous data. This model computes probabilities for each class based on the assumption that features are conditionally independent.

K-Nearest Neighbors (KNN):

KNN was employed to classify crop cases based on the distance of feature values to the nearest training samples.

Support Vector Machine (SVM):

A linear SVM model was trained to identify the optimal hyperplane for separating labels.

Decision Trees and Random Forests:

Decision Trees were used for their interpretability, while Random Forests enhanced prediction accuracy by aggregating results from multiple trees.

Neural Network:

A simple feedforward neural network was implemented to capture non-linear relationships in the data.

E. Classification

The classification task involved predicting the occurrence of using the trained models. Each model was evaluated based on accuracy, precision, recall, and F1-score to assess performance. The confusion matrix provided a detailed overview of model predictions, allowing for insights into the classification of crop instances.

F. Results

The output of the system is a classification of each crop recommendation within the dataset.

After training, the system accurately estimates the most suitable crop (e.g., rice, wheat, maize) based on new environmental and nutrient data. Farmers and agricultural experts can leverage the predictions to assess crop suitability and make informed decisions regarding cultivation and resource management.

The system's performance was measured using various metrics, demonstrating its potential utility in agricultural planning for crop selection. Overall, the hybrid approach, utilizing multiple models, contributed to improved accuracy and reliability in classifying crop recommendations.

3.4 Model Evaluation

A. Confusion Matrix

The classification performance of each model was assessed using confusion matrices, which provide a detailed analysis of true positives, false positives, true negatives, and false negatives for the label. The matrices helped identify:

B. Accuracy

Accuracy is defined as the proportion of accurately predicted instances (true positives and true negatives) to the total instances. Although it serves as a general indicator of model performance, it may be misleading in the context of an imbalanced dataset. Here, accuracy was considered as a foundational metric.

C. Precision

Precision quantifies the percentage of accurate positive predictions. In this study, it reflects the proportion of instances that were correctly identified as label cases out of all predicted crop cases. Precision is crucial when the cost of false positives is high, as it minimizes incorrect classifications into the positive class.

D. Recall

Recall, also known as sensitivity, measures the proportion of actual positive instances that were correctly detected. It illustrates how effectively the model identifies crop.

E. F1-Score

The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both. It is particularly useful in scenarios where there is an imbalance in class distributions or when both precision and recall are equally important. A high F1-score indicates good model performance in classification.

F. Performance Outcomes

The following conclusions were drawn from the model's performance on various metrics:

Training Accuracy: Indicates how well the model learned patterns from the training data.

Testing Accuracy: Reflects how effectively the model performs on unseen data.

Precision and Recall: Aided in assessing the model's ability to correctly classify crop instances and avoid false classifications.

F1-Score: Provided a comprehensive measure of the model's performance, showcasing the balance between precision and recall.

Based on evaluation results, the models showed varying degrees of success in predicting crop. The hybrid approach, employing multiple algorithms, allowed for improved accuracy and reliability in predictions.

G. Individual Model Performance

Logistic Regression:

```
Logistic Regression with accuracy: 0.9181818181818182
Confusion matrix:
[[16  0  3  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 20  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 6  0 16  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1]
 [ 0  0  0 17  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]
 [ 3  0  0  0  0 17  0  0  0  0  0  0  0  0  0  0  3  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 23  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 19  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 19  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 23  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 3 17  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 10  0  0  1  0 13  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  2  0 18  2  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 26  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17]]
```

```

=====
Classification Report:
      precision    recall  f1-score   support

     1         0.64      0.84      0.73        19
     2         1.00      0.95      0.98        21
     3         0.84      0.70      0.76        23
     4         0.94      1.00      0.97        17
     5         1.00      1.00      1.00        27
     6         1.00      0.74      0.85        23
     7         1.00      1.00      1.00        14
     8         1.00      1.00      1.00        23
     9         1.00      1.00      1.00        17
    10         1.00      1.00      1.00        19
    11         1.00      1.00      1.00        14
    12         0.66      1.00      0.79        19
    13         1.00      1.00      1.00        21
    14         1.00      1.00      1.00        23
    15         0.69      1.00      0.81        11
    16         0.89      0.85      0.87        20
    17         0.86      1.00      0.93        19
    18         1.00      0.54      0.70        24
    19         1.00      0.78      0.88        23
    20         0.91      1.00      0.95        20
    21         1.00      1.00      1.00        26
    22         0.94      1.00      0.97        17

 accuracy          0.92        440
 macro avg         0.93        0.93      0.92        440
 weighted avg      0.93        0.92      0.92        440

```

With a maximum of 1000 iterations to ensure convergence, Logistic Regression produced competitive results in terms of accuracy, precision, recall, and F1-score.

Figure 2. Logistic Regression – Confusion Matri

Naive Bayes:

The Naive Bayes classifier performed well, particularly in high-dimensional data, yielding decent accuracy despite some assumptions about feature independence.

Naive Bayes with accuracy: 0.9954545454545455

Confusion matrix:

```

[[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 23  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 27  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 14  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 17  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 14  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 21  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 11  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 19  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 23  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 26]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17]]

```

Classification Report:				
	precision	recall	f1-score	support
1	1.00	0.89	0.94	19
2	1.00	1.00	1.00	21
3	0.92	1.00	0.96	23
4	1.00	1.00	1.00	17
5	1.00	1.00	1.00	27
6	1.00	1.00	1.00	23
7	1.00	1.00	1.00	14
8	1.00	1.00	1.00	23
9	1.00	1.00	1.00	17
10	1.00	1.00	1.00	19
11	1.00	1.00	1.00	14
12	1.00	1.00	1.00	19
13	1.00	1.00	1.00	21
14	1.00	1.00	1.00	23
15	1.00	1.00	1.00	11
16	1.00	1.00	1.00	20
17	1.00	1.00	1.00	19
18	1.00	1.00	1.00	24
19	1.00	1.00	1.00	23
20	1.00	1.00	1.00	20
21	1.00	1.00	1.00	26
22	1.00	1.00	1.00	17
accuracy			1.00	440
macro avg	1.00	1.00	1.00	440
weighted avg	1.00	1.00	1.00	440

Figure 3. Naïve Bayes – Confusion Matrix

Support Vector Machine (SVM):

Probability estimates were enabled during training, which facilitated detailed performance assessments. SVM showed strong performance, especially in precision and recall metrics.

```
Support Vector Machine with accuracy: 0.9681818181818181
Confusion matrix:
[[14  0  5  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0 20  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0]
 [ 0  0 22  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1]
 [ 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 1 19  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  3  0  0 21  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 1  0  0 20  2  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  0 26  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17]]
=====
Classification Report:
              precision    recall  f1-score   support

     1         1.00        0.74        0.85         19
     2         1.00        0.95        0.98         21
     3         0.81        0.96        0.88         23
     4         0.94        1.00        0.97         17
     5         1.00        1.00        1.00         27
     6         1.00        1.00        1.00         23
     7         1.00        1.00        1.00         14
     8         1.00        1.00        1.00         23
     9         1.00        1.00        1.00         17
    10         1.00        1.00        1.00         19
    11         1.00        1.00        1.00         14
    12         1.00        1.00        1.00         19
    13         1.00        1.00        1.00         21
    14         1.00        1.00        1.00         23
    15         0.73        1.00        0.85         11
    16         0.95        0.95        0.95         20
    17         1.00        1.00        1.00         19
    18         1.00        0.88        0.93         24
    19         1.00        0.87        0.93         23
    20         0.91        1.00        0.95         20
    21         1.00        1.00        1.00         26
    22         0.94        1.00        0.97         17

 accuracy         0.97         440
 macro avg        0.97         0.97         0.97         440
 weighted avg     0.97         0.97         0.97         440
```

Figure 4. Support Vector Machine (SVM) -- Confusion Matrix

Random Forest:

Trained with 100 trees, the Random Forest model exhibited robust performance and resilience to overfitting, resulting in good accuracy and stability.

```
Random Forest with accuracy: 0.9954545454545455
Confusion matrix:
[[17  0  2  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 23  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 27  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 14  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 17  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 14  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 21  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 11  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 19  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 23  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 26]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 17]]
=====
```

Classification Report:				
	precision	recall	f1-score	support
1	1.00	0.89	0.94	19
2	1.00	1.00	1.00	21
3	0.92	1.00	0.96	23
4	1.00	1.00	1.00	17
5	1.00	1.00	1.00	27
6	1.00	1.00	1.00	23
7	1.00	1.00	1.00	14
8	1.00	1.00	1.00	23
9	1.00	1.00	1.00	17
10	1.00	1.00	1.00	19
11	1.00	1.00	1.00	14
12	1.00	1.00	1.00	19
13	1.00	1.00	1.00	21
14	1.00	1.00	1.00	23
15	1.00	1.00	1.00	11
16	1.00	1.00	1.00	20
17	1.00	1.00	1.00	19
18	1.00	1.00	1.00	24
19	1.00	1.00	1.00	23
20	1.00	1.00	1.00	20
21	1.00	1.00	1.00	26
22	1.00	1.00	1.00	17
accuracy			1.00	440
macro avg	1.00	1.00	1.00	440
weighted avg	1.00	1.00	1.00	440

Figure 5. Random Forest – Confusion Matrix

K-Nearest Neighbors (KNN):

The KNN classifier provided a good balance between simplicity and performance, effectively identifying stroke cases based on distance metrics.

```

K-Nearest Neighbors with accuracy: 0.9704545454545455
Confusion matrix:
[[14  0  5  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]
 [ 1  0 22  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 27  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 14  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 17  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 14  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 21  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 11  0  0  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 1 19  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 3  0 21  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 1  0 20  2]
 [ 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  0 17]]
=====
Classification Report:
              precision    recall  f1-score   support

     1         0.93        0.74        0.82         19
     2         1.00        1.00        1.00         21
     3         0.81        0.96        0.88         23
     4         1.00        1.00        1.00         17
     5         1.00        1.00        1.00         27
     6         1.00        1.00        1.00         23
     7         1.00        1.00        1.00         14
     8         1.00        1.00        1.00         23
     9         1.00        1.00        1.00         17
    10         1.00        1.00        1.00         19
    11         1.00        1.00        1.00         14
    12         1.00        1.00        1.00         19
    13         1.00        1.00        1.00         21
    14         1.00        1.00        1.00         23
    15         0.73        1.00        0.85         11
    16         0.95        0.95        0.95         20
    17         1.00        1.00        1.00         19
    18         1.00        0.88        0.93         24
    19         1.00        0.87        0.93         23
    20         0.91        1.00        0.95         20
    21         1.00        1.00        1.00         26
    22         1.00        1.00        1.00         17

 accuracy          0.97          0.97          0.97         440
  macro avg        0.97          0.97          0.97         440
 weighted avg      0.97          0.97          0.97         440

```

Figure 7. KNN – Confusion Matrix

Decision Tree:

The Decision Tree model provided interpretable predictions by recursively partitioning the data based on feature values. Although prone to overfitting, it achieved reasonable accuracy with appropriate tuning of hyperparameters like max_depth and min_samples_split.

```
Decision Tree with accuracy: 0.9818181818181818
Confusion matrix:
[[17  0  2  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  1  0  0  0]
 [ 3  0 20  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 27  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 14  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 17  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 14  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 21  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 11  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 19  0  0  0]
 [ 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  1 22  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 20  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 17]]
```

```
=====  
Classification Report:
```

	precision	recall	f1-score	support
1	0.85	0.89	0.87	19
2	1.00	0.95	0.98	21
3	0.91	0.87	0.89	23
4	1.00	1.00	1.00	17
5	1.00	1.00	1.00	27
6	1.00	1.00	1.00	23
7	1.00	1.00	1.00	14
8	1.00	1.00	1.00	23
9	1.00	1.00	1.00	17
10	1.00	1.00	1.00	19
11	1.00	1.00	1.00	14
12	1.00	1.00	1.00	19
13	1.00	1.00	1.00	21
14	1.00	1.00	1.00	23
15	0.92	1.00	0.96	11
16	0.95	1.00	0.98	20
17	0.95	1.00	0.97	19
18	1.00	0.92	0.96	24
19	1.00	1.00	1.00	23
20	1.00	1.00	1.00	20
21	1.00	1.00	1.00	26
22	1.00	1.00	1.00	17
accuracy			0.98	440
macro avg	0.98	0.98	0.98	440
weighted avg	0.98	0.98	0.98	440

Figure 8. Decision Tree – Confusion Matrix

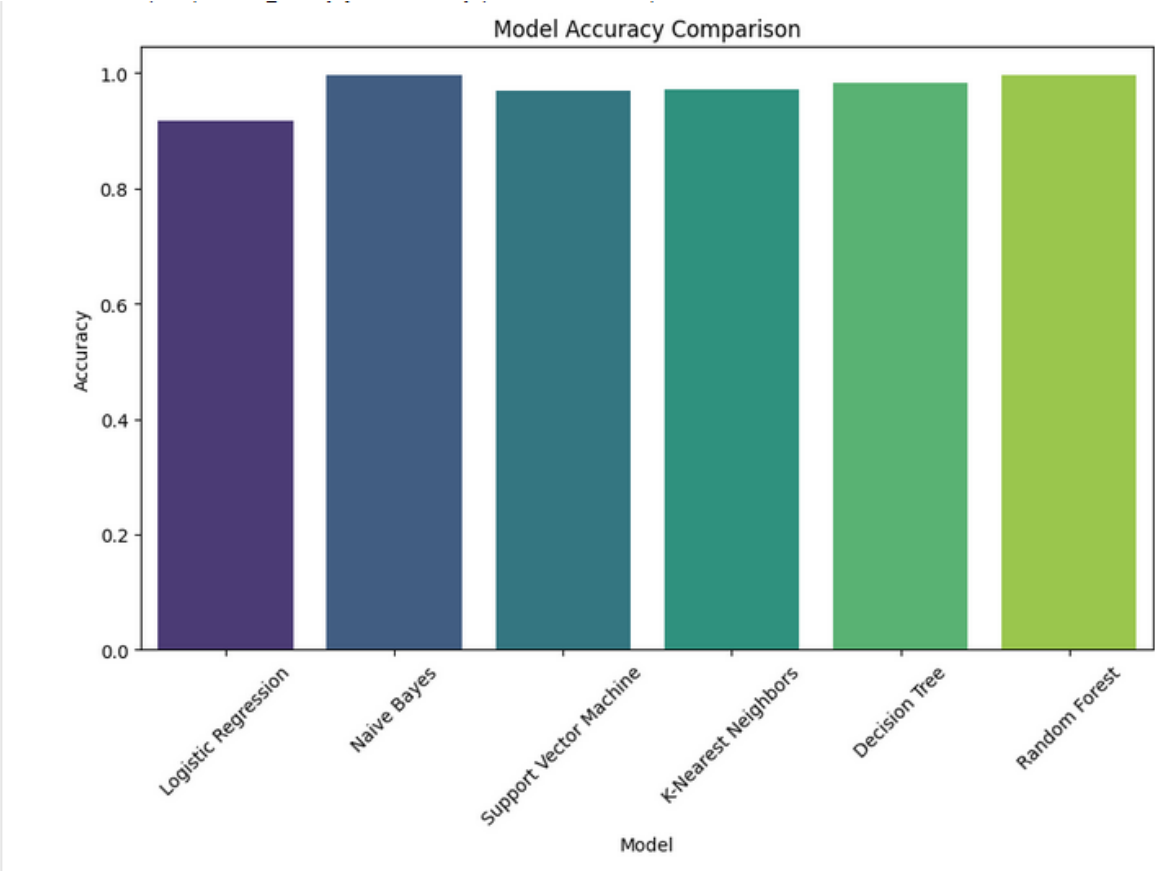


Table 1. Recorded Results for each Classifier

According to the evaluation metrics, models like and Random Forest exhibited the highest performance, while Naive Bayes provided a solid baseline. While the Naive Bayes classifier was effective, further optimization through feature selection and hyperparameter tuning could enhance the model’s ability to differentiate between severity levels, ensuring more reliable predictions in clinical settings.

3.5 Constraints

In our crop prediction project, we operate within a specific set of constraints that influence the design and development of the solution. These constraints ensure that our models adhere to critical factors and limitations related to agriculture domain:

i. Data Authenticity:

We acknowledge the potential for incomplete or erroneous data in our dataset. crop data and environmental factors may not always reflect actual conditions accurately. This possibility underscores the importance of implementing data validation processes to ensure the accuracy and reliability of the data used for training and testing our models, thereby mitigating the impact of any inaccuracies on the final predictions.

ii. Cost Considerations:

Although our dataset was obtained from publicly available sources, such as Kaggle, we recognize that generating or acquiring high-quality crop data for crop prediction may incur costs. Balancing these costs with our project objectives is vital to maintain cost-effectiveness without compromising accuracy or data quality.

iii. Data Quality:

The performance of our crop prediction model relies heavily on ensuring high data quality and integrity. We face constraints related to maintaining stringent data quality standards, which encompass procedures for data cleansing, validation, and verification to remove errors or noise. In the agriculture domain, where precision is critical, access to high-quality data is essential to enhance our model's accuracy and reliability.

CHAPTER-4

IMPLEMENTATION

4.Implementation

4.1 Environment Setup

To ensure the seamless operation of our crop prediction models, we established a robust environment tailored for data analysis and machine learning tasks. The primary programming language used for this project was Python, supported by a variety of libraries that facilitated data handling, model training, and visualization. Key libraries included:

NumPy: For numerical computations and array manipulations.

Pandas: For data processing and manipulation, enabling efficient handling of structured data.

Matplotlib and Seaborn: For result visualization, allowing for effective representation of model outputs and insights.

Scikit-learn: Utilized for constructing various machine learning algorithms, including Logistic Regression, Naive Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Decision Trees, and Random Forests.

The environment was set up using Anaconda, which simplified package management and deployment. After loading the dataset from local storage, the preprocessing of data was conducted using Pandas. This preprocessing phase involved:

Encoding Categorical Variables: Converting categorical variables, such as smoking_status, using scikit-learn's Label Encoder to ensure compatibility with machine learning models.

Handling Missing Values: Implementing strategies to address any missing data.

Feature Scaling: Normalizing numerical features to ensure equitable contributions during model training.

The hardware specifications for this project included a standard desktop computer equipped with at least 8GB of RAM and an Intel i5 processor, enabling efficient model training and data processing operations.

4.2 Sample Code for Preprocessing and Hybrid Model Operations

The preprocessing stage was crucial in ensuring the quality and reliability of the input data for our machine learning models. The dataset contained a variety of variables related to crop data and soil factors for crop prediction.

Below is a sample code snippet illustrating the preprocessing and training of multiple hybrid models:

```
python
import numpy as np
import pandas as pd
crop = pd.read_csv("/content/Crop_recommendation.csv")
crop.head()
crop.shape
crop.info ()
crop.isnull(). sum ()
crop.duplicated(). sum ()
crop.describe()
crop['label'].value_counts()
crop_dict = {
    'rice': 1,
    'maize': 2,
    'jute': 3,
    'cotton': 4,
    'coconut': 5,
    'papaya': 6,
    'orange': 7,
    'apple': 8,
    'muskmelon': 9,
    'watermelon': 10,
    'grapes': 11,
    'mango': 12,
    'banana': 13,
    'pomegranate': 14,
    'lentil': 15,
    'blackgram': 16,
    'mungbean': 17,
    'mothbeans': 18,
    'pigeonpeas': 19,
```

```

    'kidneybeans': 20,
    'chickpea': 21,
    'coffee': 22
}
crop['crop_num']= crop['label']. map(crop_dict)
crop.drop(['label'], axis=1, inplace=True)
X = crop.drop(['crop_num'], axis=1)
y = crop['crop_num']
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
from sklearn.preprocessing import MinMaxScaler
ms = MinMaxScaler()
X_train = ms.fit_transform(X_train)
X_test = ms.transform(X_test)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
X = data.drop('stroke', axis=1) # Features
y = data['stroke'] # Target variable
svm_model = SVC (probability=True)
for model, name in zip (models, model_names):
    from sklearn.naive_bayes import GaussianNB
    from sklearn.svm import SVC
    from sklearn.neighbors import KNeighborsClassifier
    from sklearn.tree import DecisionTreeClassifier
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.metrics import accuracy_score, confusion_matrix, classification_report

# Define models
models = {
    'Logistic Regression': LogisticRegression(),
    'Naive Bayes': GaussianNB(),
    'Support Vector Machine': SVC (),

```

```

'K-Nearest Neighbors': KNeighborsClassifier(),
'Decision Tree': DecisionTreeClassifier(),
'Random Forest': RandomForestClassifier(),
}

accuracies = []
model_names = []

# Fit models and evaluate
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    accuracies.append(accuracy)
    model_names.append(name)

    print(f"{name} with accuracy: {accuracy}")
    print ("Confusion matrix:\n", confusion_matrix(y_test, y_pred))
    print ("=====")
    print ("Classification Report:\n", classification_report(y_test, y_pred))

# Plot accuracy comparison
plt.figure(figsize=(10, 6))
sns.barplot(x=model_names, y=accuracies, palette='viridis')
plt.title('Model Accuracy Comparison')
plt.xlabel('Model')
plt.ylabel('Accuracy')
plt.xticks(rotation=45)
plt.show()

rfc = RandomForestClassifier()
rfc.fit(X_train,y_train)
ypred = rfc.predict(X_test)

```

```
accuracy_score(y_test,ypred)
gnb = GaussianNB()
gnb.fit(X_train,y_train)
ypred = gnb.predict(X_test)
accuracy_score(y_test,ypred)
```

In this code:

We loaded the crop dataset and prepared the features and target variable.

Categorical variables were encoded using LabelEncoder.

The dataset was split into training and testing sets using an 80-20 split.

Multiple models (Logistic Regression, Random Forest, SVM, Decision tree, KNN, Naïve Bayes) were initialized and trained.

Predictions were made on the test set, and the accuracy for each model was calculated.

Confusion matrices were visualized using heatmaps to illustrate the classification results.

This structured approach ensures that the models are trained effectively and can provide accurate predictions on crop occurrences based on the input data.

CHAPTER - 5

Experimentation and Result Analysis

5. Experimentation and Result Analysis

During the experimentation phase of the crop prediction project, several machine learning models were trained, and their performance was assessed using a variety of metrics. We systematically evaluated each model's accuracy, precision, recall, and F1 score to determine how well it predicted crop.

The findings indicated that ensemble methods, particularly Gaussian NB, Random Forest and, outperformed traditional models such as Logistic Regression and Support Vector Machines (SVM). The superior performance of the ensemble models can be attributed to their robustness against overfitting and their ability to handle complex patterns in the data. We used confusion matrices to visualize the performance of each model by displaying true positives, true negatives, false positives, and false negatives.

CHAPTER – 6

CONCLUSION

CONCLUSION:

In conclusion, this experiment manifests the potential of machine learning techniques for the promotion of crop recommendation and prediction. Systematic application and evaluation of such varied machine learning models like Logistic Regression, Naive Bayes, Support Vector Machine, Random Forest, K-Nearest Neighbours, and Decision Tree are shown to be of efficiency while analyzing complex agricultural data and informative for the selection of a crop. From the results, we see that ensemble models, especially Random Forest models, fit almost perfectly and yield very high accuracy while capturing intricate patterns in soil characteristics, weather conditions helps farmers and agronomists make more informed decisions regarding crop cultivation. However, there are a number of challenges which remain present. Agriculture datasets are usually prone to variability in climate data, soil properties, and pest infestations. Agriculture datasets depend on the accuracy and completeness of successful implementation of machine learning models in such applications. Overcoming these challenges requires a combined effort from data scientists, agricultural experts, and farmers to make sure that the quality of data is maintained at its optimum for model training. The second challenge is the interpretability of machine learning models in the agricultural context. While complicated algorithms like Random Forest or SVM may make crop recommendations with an acceptable level of accuracy, their decision-making processes can hardly be comprehended by practicing farmers. The future challenge with respect to interpretability and transparency will be of major importance: the results of such machine learning tools should be understood and trusted by farmers or agricultural consultants and used effectively. Multi-source data integration, such as satellite imagery, soil sensors, climate data, and even market demand, is one more promising direction toward enhancing predictive accuracy and making crop recommendations even more holistic. Inclusion of diverse agricultural variables in the dataset will enable them to give more accurate recommendations tailored to specific regions and conditions. Validation using real-world data in various agricultural zones and seasons will further strengthen the applicability and reliability of these models in highly diversified farming scenarios. As this technology continues to advance and optimize farming practices towards increased crop yields, sustainable agriculture draws near. Further tapping of the power must be done with full-scale cooperation between data scientists, agricultural experts, and farmers, who must interact with the complexity of the challenges and develop innovative practical solutions suitable for modern farming.

REFERENCES

- [1] Manpreet Kaur, Heena Gulati, Harish Kundra, "Data Mining in Agriculture on Crop Price Prediction: Techniques and Applications", International Journal of Computer Applications, Volume 99– No.12, August 2014.
- [2] J. Meng, "Research on the cost of agricultural products circulation and its control under the new normal economic development," Commercial Times, no. 23, pp. 145147, 2016.
- [3] A. Kaloxylou et al., "Farm management systems and the future Internet era," Comput. Electron. Agricult., vol. 89, pp. 130–144, Nov. 2012.
- [4] N. N. Li, T. S. Li, Z. S. Yu, Y. Rui, Y. Y. Miao, and Y. S. Li, "Factors influencing farmers' adoption of new technology based on Logistic-ISM model-a case study of potato planting technology in Dingxi City, Gansu Province," Progress in Geography, vol. 33, no. 4, pp. 542-551, 2014.
- [5] Y. Wang, "A neural network adaptive control based on rapid learning method and its application," Advances in Modeling and Analysis, Vol. 46(3), pp. 27-34,199
- [6] J Tang,Xia Hu,Huiji Gao,Huan Liu," Unsupervised Feature Selection for Multi- 78 View Data in Social Media",Proceedings of the ACM international conference on Information and Knowledge management (2015), pp.1673-1678.
- [7] Haytham Elzhazel and Alex Aulsebrook, "Unsupervised Feature Selection with ensemble learning ", Machine Learning, Springer (2013).
- [8] Mohak Shah, Mario Marchand, and Jacques Corbeil," Feature Selection with Conjunctions of Decision Stumps and Learning ", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 34, NO. 1, (2012) pp.174-186.
- [9] Qingzheng XU, Lei WANG, Baomin HE, Na WANG," Modified Opposition-Based Differential Evolution for Function Optimization", Journal of Computational Information Systems 7:5 (2011) pp.1582-1591.
- [10] Hyndman, R. et al., 2008. Forecasting with Exponential Smoothing: The State Space Approach, Springer Science & Business Media.
- [11] Kalimuthu M, Vaishnavi P and Kishore M 2020 Crop prediction using machine learning 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) pp 926-32 doi: 10.1109/ICSSIT48917.2020.9214

