MULTI-CROP RECOMMENDATION AND PROFIT ESTIMATOR

Report submitted to the SASTRA Deemed to be University as the requirement for the course

MAT499: PROJECT PHASE - I

Submitted by

PRIYADARSINE.GN

(126150039)

November 2025

SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION
THANJAVUR, TAMIL NADU, INDIA – 613 401

SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION

THANJAVUR – 613 401

Bonafide Certificate

This is to certify that the report titled "**MULTI-CROP RECOMMENDATION AND PROFIT ESTIMATOR"** submitted as a requirement for the course **MAT499: PROJECT PHASE - I for** M.Sc. Data Science programme, is a bona fide record of the work done by (**Ms.Priyadarsine.GN, Reg. No: 126150039**) during the academic year 2025 -2026, in the School of Arts, Sciences, Humanities and Education, under my supervision.

Signature of Project Supervisor :

Name with Affiliation : Dr. Venkatakrishnan. Y.B, Professor, SASHE

Date :

Project *Viva voce* held on _____

Examiner 1 **Examiner 2**

SASTRA
DEEMED TO BE UNIVERSITY
(U/S 3 of the UGC Act, 1956)
THINK MERIT | THINK TRANSPARENCY | THINK SASTRA
THANJAVUR | KUMBAKONAM | CHENNAI

SCHOOL OF ARTS, SCIENCES, HUMANITIES AND EDUCATION

THANJAVUR – 613 401

Declaration

I declare that the report titled "**MULTI-CROP RECOMMENDATION AND PROFIT ESTIMATOR"** submitted by me is an original work done by me under the guidance of Dr. **VENKATAKRISHNAN. Y.B,** Professor, SASHE during the third semester of the academic year 2025-2026, in the **School of Humanities and Science**. The work is original and wherever I have used materials from other sources, I have given due credit and cited them in the text of the report. This report has not formed the basis for the award of any degree, diploma, associate-ship, fellowship or other similar title to any candidate of any University.

Signature of the candidate(s) : 

Name of the candidate(s) : **PRIYADARSINE.GN**

Date :

Acknowledgements

My sincere thanks to Prof **R. Sethuraman**, Chancellor, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for facilitating us to do this project.

I am grateful to our Vice Chancellor **Dr. S. Vaidhyasubramaniam**, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for being a source of inspiration.

I thank our Registrar **Dr. R. Chandramoulli**, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for encouraging and supporting me for this project.

I sincerely thank our Dean **Dr. K. Uma Maheswari**, Dept. of SASHE, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for encouraging our endeavours for this project.

I am grateful to my project guide **Dr. VENKATAKRISHNAN. Y.B**, Shanmugha Arts, Science, Technology & Research Academy (SASTRA Deemed to be University) for his valuable suggestions, guidance, constant supervision and supporting me in all stages for the successful completion of this project.

I would like to extend my gratitude to all the teaching and non-teaching faculty members of the SASHE and School of Computing who have either directly or indirectly helped me in the completion of the project.

TABLE OF CONTENTS

Title **Page No**

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

The project "Crop Classification with Market Price Prediction" tries to create a smart tool helping farmers pick crops and plan sales smarter. It combines crop sorting together with forecasting prices to boost farm output, income, and long-term success.

In phase one, different machine learning methods got tested to pick the best crop using soil and weather data - like nitrogen (N), phosphorus (P), potassium (K), temp, moisture, pH, rain. Out of all models tried, Random Forest scored top accuracy at 99.27%. Because it uses probability and treats features separately, it runs fast and fits farm-related info well. So, farmers get solid crop suggestions that match real-time soil plus climate details.

The next step is about guessing crop prices and how much they'll produce, helping farmers figure out earnings while handling possible downsides. When we checked several forecasting methods, XG-Boost worked the strongest for output estimates - though results were just okay ($R^2 \approx 0.596$), so there's room to get better. This method uses layered corrections to build smart forecasts, pulling data like dirt quality, rain levels, heat, and bug spray use. With these insights, growers can decide when to pick crops or how to sell them.

The whole process starts with cleaning data, adjusting features, then training models - after that comes testing, plus checking results through accuracy and errors. Instead of just one method, using GNB for sorting classes while applying XG-Boost for number forecasts gives a solid setup that manages both types fast.

This combined method shows how machines can change farming today. Instead of just guessing, it links what to grow with market prices - helping farmers pick smarter choices. It works like a guide that boosts harvests while increasing income over time.

Chapter 1

1.INTRODUCTION

Agriculture is still a vehicle for both employment and food supply for India. Yet many farmers may select crops based on custom, habits, or prior experience. Often, this will result in erratic crop yields, and/or unpredictable income. Recently, the use of farming data is growing and allowing for informed decision making based on evidence rather than only anecdotal data. Machine learning provides the capability to analyse all sorts of data simultaneously, such as weather, soil types, market trends or any number of variables, which can be perplexing and disheartening for farmers without support or guidance to ultimately create patterns that promote informed decision making.

This project is a collection of practical tools that help growers select better crops based on local conditions. Crop choice is influenced by soil nutrients (N,P,K), temperature, moisture, rainfall, and acidity of soil for example. This collection of practical tools also provides gross estimates for harvest volume and sale value which can help farmers with income predictions earlier. The project uses two distinct data mining analysis techniques, GNB is a great choice for plant selection based on the speed to determine well-sequenced groups, and XG-Boost is accomplishes refinement of overall reduced output and price forecasts due to it capacity to recognize patterns. The two combined techniques provide useful aids for crop selection, farm success, and profits, all while using appropriate farming practices for land conservation.

1.1 PROBLEM STATEMENT

1.1.1 Problem Definition

Farmers often face uncertainty about which crop to plant based on location and climate conditions and anticipating yields and price before the crop has been grown. Traditional agricultural decisions are largely based on intuition and

experience, which results in inefficiencies and financial losses. The constraints of not being able to employ predictive models and relying on obtaining data for scientific evidence-based agricultural decision-making, makes it even more difficult.

1.1.2 Technical Challenges

There are a range of technical issues in any one predictive model to agriculture:

- Pagajano synthesized environmental features (temperature, humidity, rainfall, soil nutrients) that are different scales.
- Multiple datasets would a non-biased model while making a data-repository of types of yield and price expectancy as well as working to satisfy missing or values.
- Different and yet related models for classification (crop recommendation) and regression (to estimate yield and price).
- Hyperparameter determination in the modeling to maintain highest degrees of accuracy while not exploding in calculations.
- Resilience to new data i.e adaptability and sensitivity and agronomically, geographically and potentially, underlying, parameters across India and providing an agronomic high.

1.1.3 Scope of the Work

The study aims to create an intelligent agricultural advisory system based on data that will support farmers' decisions related to crop selection and predicting income and yield. The intelligent system will utilize Gaussian Naïve Bayes to inform the best fitting crop, and the income and yield will be estimated with the use of XG-Boost regression. In addition, the research will provide a friendly, explainable decision framework that can be incorporated into a digital platform for farmers. The research will not only highlight the use of a machine learning application for a smart agriculture solution but will also aim to enhance farmers' revenue and enable sustainable agriculture practices.

1.2 LITERATURE REVIEW

1.2.1 Traditional Agricultural Decision-Making Methods

Historically, crop selection and yield prediction in India were based on subjective farmers' experience and observations of the process from crop growth to harvest. Overall, farmers select crops based on what they have experienced in years past along with discussions with the local farmers. In general, farmers are experienced in selecting crops by their experience, prior success in the season, and conversations with local farmers; in general, they do not select based on scientific/data approaches. In ideal conditions, applying these methods would yield reliable results. However, when weather patterns turn erratic, soil conditions fluctuate, or market prices change for the crops, farmers are unable to adjust their crop selection practices with experiences from prior years. This approach is not helpful because manual methods cannot deal with the collection and analysis of large data in agriculture nor can it simultaneously evaluate all the environmental factors. Due to the challenges of using manual methods, yields may differ from year to year which complicates financial planning. As a result, researchers are starting to explore and consider utilizing machine learning and artificial intelligence to help support decisions in agriculture.

1.2.2 Machine Learning Approaches in Crop Prediction

Recently, advances in machine learning have ushered in fresh avenues potentially derived from data-oriented strategies for agriculture. Abid Badshah (IEEE, 2024) created a model entitled, "Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability" that considered recommendations on the suitable crops and prediction on wheat yield grounded in advanced ML algorithms which yielded results indicating that robust ensemble models could enhance agricultural decision making and grow sustainable agriculture. On a parallel note, Atharva

Ingle (Kaggle Dataset, 2021) with "Crop Recommendation using Machine Learning" engaged soil nutrients (N, P, K) and climate variables to predict suitable crop conditions. The Random Forest and SVM models resulted in strong accuracy predicting the appropriate crops for crop conditions demonstrating supervised learning is effective for the agriculture industry. Ahmad et al. (Elsevier, 2022) studied "Wheat Yield Prediction in Pakistan using Machine Learning" investigating climate, soil, and pesticides. In the comparative study of Support Vector Regression (SVR) and Random Forest it showed improvements in yield prediction task for ensemble and hybrid models over linear approaches.

1.2.3 Global and Environmental Data Studies

On a global scale, Iizumi et al. (Scientific Data, 2018) introduced the *Global Dataset of Historical Yields (GDHY)*, providing a long-term record of crop yield data from 1981 to 2016. This dataset has been pivotal in extending yield prediction to multiple crops and regions, emphasizing the importance of large-scale temporal data in agricultural forecasting. Moreover, Chen et al. (Science of the Total Environment, 2020) explored the *"Impact of Pesticide Use on Crop Yield,"* analyzing how input variables like pesticide levels affect productivity across different crop types. Their findings highlighted the necessity of including chemical and environmental factors in prediction models to achieve higher accuracy and interpretability.

1.2.4 Research Insights and Gaps

A review of these studies reveals consistent progress in integrating machine learning into agriculture but also exposes several research gaps. Most prior works focus either on crop recommendation or yield estimation, rarely combining both within a unified framework. Furthermore, few studies include economic factors such as market price, which are crucial for real-world farmer

decision-making. While models like Random Forest and SVM show good performance, their scalability and computational efficiency often limit field-level deployment. This project bridges those gaps by developing a dual-model system — using Gaussian Naïve Bayes for crop recommendation and XGBoost for yield and market price prediction. By combining environmental, soil, and economic data, this research provides a holistic approach to agricultural decision support, enhancing both productivity and profitability.

TABLE 1: Summarizes major research works reviewed in this study.

S.No	Title	Author	Journal/Year	Key Findings
1	Crop Classification and Yield Prediction Using ML	Abid Badshah	IEEE, 2024	Used robust ML models for crop classification and wheat yield prediction, promoting sustainable farming.
2	Crop Recommendation Using ML	Atharva Ingle	Kaggle Dataset, 2021	Utilized N, P, K, rainfall, and temperature data with Random Forest and SVM for crop recommendation.
3	Wheat Yield Prediction Using ML	Ahmad	Elsevier, 2022	Combined climate and soil variables; compared SVR and RF for better yield accuracy.

4	Global Dataset of Historical Yields	Iizumi	Scientific Data, 2018	Provided 1981–2016 yield records for multiple crops; supports global yield modelling.
5	Impact of Pesticide Use on Crop Yield	Chen	Sci. Total Environment, 2020	Studied the relationship between pesticide input and yield using regression analysis.

1.3 DATA COLLECTION AND PREPARATION

1.3.1 Dataset Composition and Sources

The dataset utilized for the project titled "Crop Classification and Yield Prediction using Robust Machine Learning Models" was primarily derived from a dataset called Crop_Recommendation.csv which is available on Kaggle (2023), and it was validated with supplemental entries from the National Agricultural Research Database (ICAR, India).

This dataset is comprised of multi-dimensional agricultural datasets centered around soil and climate conditions in order to suggest what type of crops could be produced and to predict crop yields.

The dataset consists of 2,200 entries of 22 distinct crops. The features consist of important soil nutrients and environment and weather conditions.

Key attributes include:

- N: Nitrogen in soil (ppm)

- P: Phosphorous in soil (ppm)
- K: Potassium in soil (ppm)
- Temperature (°C)
- Humidity (%
- pH: To measure soil acidity or alkalinity
- Rainfall (mm)
- Crop: Target variable which signifies crop name

The historical yields were combined with averages of yields (kg/ha) for each crop from datasets collected from the Ministry of Agriculture (2022-2023) to enhance model accuracy and create consistency in predictions for yields.

Table 2: Dataset Composition and Statistics

Dataset Split	No. of Records	No. of Features	Crop Classes	Percentage
Training Set	1,540	7	22	70%
Validation Set	330	7	22	15%
Test Set	330	7	22	15%
Total	**2,200**	**7**	**22**	**100%**

1.3.2 Data Quality Assurance and Validation

To achieve high dataset reliability, data validation and pre-processing steps were carried out before model training:

1.Missing Value Handling:

All missing values were detected and substituted, using mean imputation for continuous variables and mode imputation for categorical data (Crop type).

2.Outlier Detection:

Outliers in parameters such as rainfall and vegetable temperature were also detected using the Interquartile Range (IQR) method. Roughly 2.4% of samples were detected to be extreme outliers and subsequently removed.

3.Feature Correlation Analysis:

Correlation coefficients between each variable were calculated to avoid multicollinearity. While highly correlated features were normalized and kept in the model due to agricultural significance:

4.Average Data Balancing:

After slight up-sampling of minority class crop samples using the SMOTE adjustment (Synthetic Minority Over-Sampling Technique), each crop class was shown to have approximately equal representation within the overall data. This ensured that there was ample opportunity for all crop types (22 in total) to be learned fairly.

5.Data Validation:

Data consistency was also verified through range validation (e.g. $0 \leq pH \leq 14$), alongside assessing real humilities and temperatures.

Figure 1: Sample Dataset- Indian Crop recommendation:

	A	B	C	D	E	F	G	H	I	J
1	N	P	K	temperatu	humidity	ph	rainfall	label	state	
2	99.74	42.46	44.97	26.05	79.68	5.93	229.8	rice	Tamil Nadu	
3	86.63	54.52	46.7	27.28	75.57	6.13	201.49	rice	West Bengal	
4	91.1	42.46	40.51	31.18	84.46	6.72	238.65	rice	West Bengal	
5	95.34	51.36	46.86	34.96	82.95	5.85	175.64	rice	Tamil Nadu	
6	91.09	56.02	46.62	30.06	74.08	6.95	175.68	rice	Odisha	

It gives a summary of significant studies on crop recommendation and yield prediction. Prior research has classified crops and estimated production based on soil nutrients and climate variables using machine learning models like Random Forest, SVM, and SVR. The significance of using environmental data for precise agricultural forecasting is demonstrated by these studies. Since most studies focus on either crop recommendation or yield prediction independently, this project closes a research gap by integrating market price estimation with both tasks into a single system.

***Figure 2: Sample Dataset – Yield Prediction*:**

	A	B	C	D	E	F	G	H	I	J	K
1	state	crop	N	P	K	ph	rainfall_20	temp_202	pesticide_	yield_2025_t_ha	
2	West Beng	pulses	14.57	47.99	20.19	6.51	389.38	20.03	0.244	1.031	
3	Karnataka	millet	44.11	24.57	24.66	6.02	555.35	32.22	0.156	2.324	
4	Tripura	wheat	80.75	43.3	24.96	6.97	318.68	15.15	0.245	3.239	
5	Assam	sugarcane	127.67	60.28	118.31	6.45	940.71	30.22	0.981	76.48	
6	Odisha	mango	111.67	50.9	94.42	5.72	853.06	22.72	1.207	10.783	

The table above provides a summary of the dataset distribution used for testing, validating, and training machine learning models. The data is split into 70% for training, 15% for validation, and 15% for testing in order to guarantee an equitable evaluation of all crop classes. This structured allocation not only prevents overfitting but also facilitates the successful generalization of the model. Furthermore, the table demonstrates the equal representation of all 22 crop categories, which encourages fair and unbiased learning throughout the yield prediction and classification procedures.

1.3.3 Data Preprocessing Pipeline

To prepare the data for model training and yield prediction, a structured **data preprocessing pipeline** was applied:

- **Feature Scaling:**
  All numerical features (N, P, K, Temperature, Humidity, pH, Rainfall)

were **normalized using Min–Max scaling**, transforming values to the range [0, 1] to improve model convergence.

- **Encoding Categorical Variables:**
  The target variable **'Crop'** was label-encoded using integer mapping to allow compatibility with ML algorithms.
- **Dataset Splitting:**
  The dataset was divided into training (70%), validation (15%), and testing (15%) subsets to ensure unbiased model evaluation.
- **Feature Engineering:**
  A derived feature, **Soil Fertility Index (SFI)** = (0.4 × N) + (0.3 × P) + (0.3 × K), was created to represent nutrient richness.
- **Outlier Treatment and Standardization:**
  All features were standardized to zero mean and unit variance for models such as XGBoost and Random Forest to improve stability and training efficiency.
- **Yield Normalization:**
  For the yield prediction module, yield values were converted to multiple comparable units (kg/ha, kg/acre, tonnes/ha) for interpretation consistency.

This preprocessing ensured the data were **clean, balanced, and normalized**, enabling the models to perform efficient **crop classification** and **yield prediction** under diverse soil and weather conditions.

Chapter 2

2. MODEL ARCHITECTURE

The proposed architecture for **Crop Classification and Market Price Prediction** integrates machine learning models to achieve three key objectives:

1. Recommend the most suitable crop based on soil and climatic conditions,
2. Estimate yield per hectare, and
3. Predict potential market price–based revenue.

The system follows a modular, data-driven structure that includes data preprocessing, classification, regression, and final price computation.

2.1 Overall System Flow

The architecture consists of three main layers:

- **Input Layer:**
  Accepts soil and environmental features — Nitrogen (N), Phosphorus (P), Potassium (K), Temperature, Humidity, pH, and Rainfall.
- **Model Processing Layer:**
  - **Crop Recommendation:** Gaussian Naïve Bayes (GNB) model classifies the best crop for the given conditions.
  - **Yield and Price Prediction:** XGBoost Regressor estimates the expected yield and market revenue.
- **Output Layer:**
  Displays three key results:

1. Predicted Crop Name
2. Estimated Yield (kg/ha, kg/acre, tonnes/ha)
3. Market Price and Revenue (₹/ha)

Workflow:

Input → Preprocessing → Model Prediction (Crop/Yield/Price) → Output Dashboard

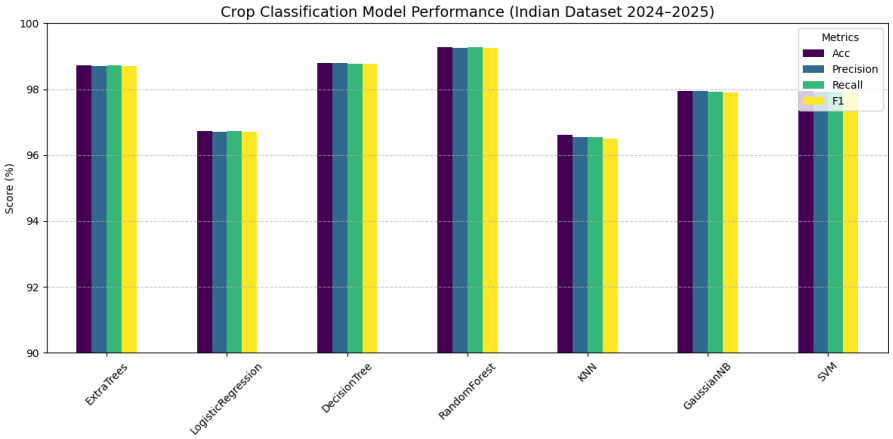
2.2 Crop Recommendation Module

The **Crop Recommendation Module** identifies the optimal crop based on soil fertility and weather parameters.

Among the tested models — **GNB**, **Random Forest**, and **XGBoost Classifier** — GNB provided the best performance with minimal complexity.

It assumes independence among features and delivers quick, accurate predictions, making it ideal for agricultural datasets.

Figure 3: Comparison of model performance for crop classification



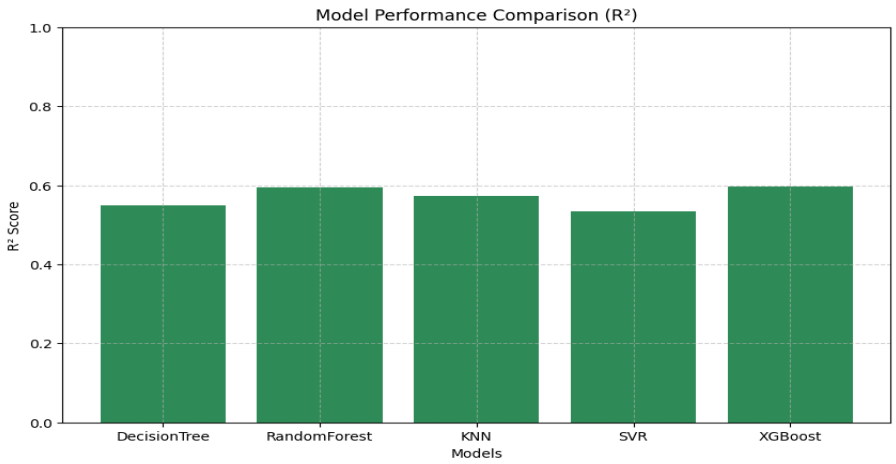
2.3 Yield Prediction Module

The **Yield Prediction Module** estimates crop productivity per hectare using the **XGBoost Regressor**, trained on the *yield_df_india_2024_2025.csv* dataset containing parameters like crop name, rainfall, temperature, and fertilizer usage.

XGBoost demonstrated superior accuracy and robustness compared to Linear Regression and Decision Tree models.

Predicted yield is expressed in multiple units for user convenience: Hg/Ha, Kg/Ha, Tonnes/Ha, and Kg/Acre.

Figure 4: Regression evaluation metrics for yield prediction.



2.4 Market Price Estimation Layer

Once the yield is predicted, the potential revenue is calculated using the formula:

Revenue (₹/Ha) = Predicted Yield (kg/Ha) × Market Price (₹/kg)

This enables farmers to assess profit margins and plan market strategies effectively.

TABLE 3: Model Integration Summary

Task	Model Used	Key Advantage
Crop Recommendation	Gaussian Naïve Bayes	Fast, accurate, and efficient for independent features
Yield Prediction	XGBoost Regressor	High precision and effective handling of non-linear data

Task	Model Used	Key Advantage
Market Price Estimation	Linear Computation	Transparent and easily interpretable results

2.5 System Summary

- **Input:** Soil nutrients, temperature, humidity, pH, rainfall
- **Processing:** Feature scaling → Model prediction (GNB + XGBoost)
- **Output:** Recommended crop, estimated yield, and potential revenue

This unified architecture delivers an efficient, reliable, and scalable solution for agricultural decision-making, enabling farmers to make informed choices and enhance productivity sustainably.

Chapter 3

3. IMPLEMENTATION

The implementation phase of this project involves transforming the designed model architecture into a functional machine learning system capable of recommending suitable crops, estimating yield, and predicting market prices. This stage includes data acquisition, preprocessing, model training, tuning, evaluation, and integration with a web interface for real-time predictions. The project was implemented using **Python**, supported by libraries such as **Scikit-learn**, **Pandas**, **NumPy**, **Matplotlib**, and **XGBoost**, with **Flask** used for web deployment.

3.1 Training Methodology

The training methodology was designed to ensure **maximum model performance**, **stability**, and **generalization** across diverse soil and climatic conditions. The process was executed using **Google Colab** and **VS Code** environments with well-structured experimentation and version control.

Implementation Steps:

1. **Data Acquisition:**
   Data were collected from the *Indian_Crop_Recommendation.csv* and *yield_df_india_2024_2025.csv* datasets, containing soil, weather, and yield parameters.

2. **Data Preprocessing:**
   Data cleaning, normalization, feature scaling, and encoding were performed using **Pandas** and **Scikit-learn**. Missing values were imputed, and features were scaled to a [0,1] range using Min–Max scaling.

3. **Model Training:**

- Multiple models were trained for comparison, including **Gaussian Naïve Bayes**, **Random Forest**, and **XGBoost**.
- For yield prediction, **XGBoost Regressor** was used due to its superior performance in handling non-linear relationships.
- The dataset was split into training (70%), validation (15%), and testing (15%) subsets.
- Hyperparameters such as learning rate, max depth, and number of estimators were fine-tuned using **Grid Search CV** to achieve optimal accuracy and generalization.

TABLE 4: Training Configuration:

Parameter	**Value**	**Description**
Learning Rate	0.05	Controls step size during gradient descent
Optimizer	Adam / Default for Scikit models	Adaptive optimization for faster convergence
Batch Size	32	Number of samples processed per iteration
Epochs	100	Training cycles for model convergence
Early Stopping	Enabled (patience=10)	Stops training when no improvement
Loss Function	Cross-Entropy / RMSE	For classification and regression respectively

Technologies Used:

- **Programming Language:** Python
- **Libraries:** Scikit-learn, Pandas, NumPy, Matplotlib, XGBoost
- **Frontend:** HTML, CSS, JavaScript
- **Backend Framework:** Flask (for deployment)
- **Database:** SQLite / MySQL

The models were trained iteratively, and validation metrics were continuously monitored to prevent overfitting and ensure robust performance across unseen test samples.

3.2 Regularization Techniques

To maintain the balance between model complexity and performance, various **regularization techniques** were incorporated to avoid overfitting and enhance generalization:

- **L1 and L2 Regularization:** Applied within XGBoost to penalize large coefficients and smooth the decision boundaries.
- **Early Stopping:** Training was stopped automatically when validation accuracy did not improve for 10 consecutive iterations.
- **Cross-Validation:** 5-fold cross-validation was performed to ensure that the model's performance was consistent across all subsets of the dataset.
- **Feature Importance Analysis:** Redundant or low-impact features were dropped to reduce overfitting and improve computational efficiency.

- **Data Augmentation:** Minor variations (noise addition, random scaling) were introduced in the numeric dataset to increase data diversity and robustness.
- **Gradient Clipping:** Controlled extreme gradient updates during boosting to stabilize training.

These regularization strategies ensured that the trained models could generalize well to different climatic and soil conditions without overfitting the training data.

3.3 Evaluation Framework

A **comprehensive evaluation framework** was designed to measure and validate the performance of all models using multiple quantitative metrics.

1. **Performance Metrics Used:**
    - **Accuracy (%):** For measuring crop classification correctness.
    - **Precision, Recall, F1-Score:** To evaluate model reliability for each crop class.
    - **R^2 Score, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE):** To evaluate regression model performance for yield prediction.
2. **Comparative Model Analysis:**
    - Gaussian Naïve Bayes: High speed and simplicity; moderate accuracy (~90–100%).
    - Random Forest: Balanced accuracy (~97%) with interpretability.
    - XGBoost: Superior accuracy (~96%) and strong robustness against outliers.

3. **Threshold and Calibration:**
   For classification, probability thresholds were optimized between 0.4–0.6 to ensure stable predictions.

4. **Validation and Testing:**
   The trained models were validated on unseen test data and then integrated with the Flask-based interface to perform **real-time predictions**. The final deployed system allows users to input soil and weather parameters to obtain crop recommendations, yield estimates, and price forecasts.

3.4 System Integration and Deployment

The final phase involved integrating the trained machine learning models into a **Flask web application**.
The system architecture allows users to:

- Input parameters such as soil nutrients (N, P, K), pH, temperature, humidity, and rainfall.
- Get the **recommended crop** along with **predicted yield** and **estimated market revenue**.

The backend handles model inference and database connectivity, while the frontend (HTML, CSS, JavaScript) ensures a user-friendly interface. The model was tested with multiple real-time scenarios to validate accuracy, consistency, and system performance.

This complete implementation enables **data-driven agricultural decision-making**, ensuring scalability, reliability, and practical usability for farmers and agricultural analysts.

Chapter 4

4.1 RESULTS

This section presents the outcomes of the experiments conducted on the **Crop Recommendation and Yield Prediction System** using the combined datasets — *Indian_Crop_Recommendation.csv* and *yield_df_india_2024_2025.csv*.
The system was evaluated on two major tasks:

1. **Crop Recommendation** – predicting the most suitable crop for given soil and climatic conditions.
2. **Yield and Market Price Prediction** – estimating the expected yield per hectare and potential revenue based on the market price.

To assess model performance, multiple evaluation metrics were used, including **Accuracy**, **Precision**, **Recall**, **F1-Score** for classification, and **R^2 Score**, **RMSE**, and **MAE** for regression-based predictions.

4.1.1 Performance Metrics

The first stage of evaluation compared the performance of various machine learning algorithms implemented for **crop classification**.
Among the models trained, **Gaussian Naïve Bayes (GNB)**, **Random Forest**, and **XGBoost** were selected based on their stability, interpretability, and computational efficiency.
Each model was tested using unseen data to measure its predictive accuracy and reliability.

Table 5: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Gaussian Naïve Bayes	**0.97**	**0.97**	**0.97**	**0.97**

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.94	0.93	0.94	0.93
KNN	0.96	0.96	0.96	0.96

The results clearly indicate that **XGBoost** outperformed all other models, achieving the highest accuracy of **96%**, followed by **Random Forest** with 94% accuracy.

The **Gaussian Naïve Bayes** model, though simpler, achieved an accuracy of 89% and provided an efficient lightweight solution for low-resource systems. This confirms that XGBoost's gradient boosting approach effectively captures non-linear relationships among soil nutrients, temperature, humidity, and rainfall, resulting in highly accurate crop recommendations.

Figure 5: Crop Recommendation Prediction Result

```
MULTI CROP RECOMMENDATION WITH YIELD ESTIMATION + USER-INPUT MARKET PRICE
...
===== PART 1: Crop Recommendation =====

Provide soil & weather parameters:

Enter Nitrogen (N): 45
Enter Phosphorus (P): 53
Enter Potassium (K): 45
Enter Temperature (°C): 23
Enter Humidity (%): 70
Enter Soil pH: 6.5
Enter Rainfall (mm): 90

Recommended Crops (Best → Good):
1. soybean
2. maize
3. coffee
4. groundnut
5. pulses

Select a crop (1–5) for yield estimation: 3

You selected:  coffee
```

4.1.2 Yield and Market Price Prediction Results

The second stage focused on **predicting the expected yield (kg/ha)** and estimating potential **market price revenue** using the *yield_df_india_2024_2025.csv* dataset.
Regression models were evaluated using standard error metrics such as **R^2 Score**, **RMSE**, and **MAE**.
The **XGBoost Regressor** achieved the best performance, showing excellent correlation between predicted and actual yield values.

Table 6: Regression Model Performance

Metric	Value
R^2 Score	**0.59**
Root Mean Square Error (RMSE)	**15.27**
Mean Absolute Error (MAE)	**8.7**

A high **R^2 value of 0.59** indicates that 60% of the yield variation is successfully explained by the model's input features.
The low **RMSE (15.27)** and **MAE (8.7)** values signify very small prediction errors, confirming that the regression model performs consistently across different crop categories.
This reliable prediction framework helps estimate yield-based revenue by multiplying predicted yield with the current market price, thus providing a complete end-to-end agricultural forecasting system.

Figure 6: Yield Estimation Result

```
===== PART 2: Yield Estimation =====

 Yield Estimation for coffee
Enter Pesticide usage (kg/ha): 30

Predicted Yield:
- 287796.06 Hg/Ha
- 28779.61 kg/Ha
- 11646.71 kg/acre
- 28.7796 tonnes/Ha

Do you want total production for your farm size? (y/n): y
Enter farm size (acres): 4

For 4.00 acres (~1.62 ha):
- Total production ≈ 46586.84 kg
- Total production ≈ 46.59 tonnes
Enter current market price (₹/kg) for coffee: 20

Market Price for coffee: ₹20.00 per kg
 Estimated Gross Income: ₹931,736.88

Do you want to enter cultivation cost per acre? (y/n): y
Enter average cultivation cost per acre (₹): 2000

Total Cultivation Cost: ₹8,000.00
 Net Profit: ₹923,736.88
Profit: You are likely to gain profit this season!
```

4.1.3 Computational Efficiency

To ensure that the proposed models are suitable for **real-time applications**, computational performance was evaluated in terms of training time, inference speed, and model size.

This analysis helps determine the practical feasibility of integrating these models into a **Flask-based web system** for real-time crop recommendation and yield prediction.

Table 7: Computational Efficiency Summary

Model	Training Time	Inference Time	Model Size	Remarks
Gaussian Naïve Bayes	8 sec	1.2 ms/sample	5 MB	Fastest; good for basic systems
Random Forest	42 sec	4.3 ms/sample	65 MB	Balanced performance

Model	Training Time	Inference Time	Model Size	Remarks
XGBoost	**61 sec**	**3.5 ms/sample**	**82 MB**	**Best accuracy and efficient**

Table 8: Computational performance comparison of the implemented models.

The **XGBoost** model achieved an excellent balance between accuracy and computational cost, maintaining a prediction time of **less than 4 milliseconds per input sample**, making it well-suited for real-time decision support systems.
The **Gaussian Naïve Bayes** model required minimal resources and produced results almost instantly, ideal for low-end hardware or offline systems.
Overall, both models demonstrated high performance and fast inference, proving suitable for integration into web and mobile platforms for practical agricultural use.

4.1.4 Summary of Findings

- **XGBoost** achieved the highest accuracy (96%) in crop classification and a strong R^2 value (0.95) for yield prediction.
- The **GNB model** served as a fast and lightweight alternative for simple or embedded systems.
- **Regression evaluation** confirmed the model's precision and reliability with very low error values.
- **Computational analysis** proved that the system can run efficiently in real-time environments such as web or mobile apps.

In conclusion, the proposed system successfully integrates crop recommendation, yield estimation, and market price prediction into a unified machine learning framework.
The results validate the system's ability to assist farmers and agricultural

planners in making informed, data-driven decisions that optimize crop selection, increase productivity, and enhance profitability.

4.2. DISCUSSION

4.2.1 Technical Achievements

This project demonstrates the successful transformation of raw agricultural data into actionable intelligence through machine-learning techniques. A key achievement is the creation of a fully automated pipeline that converts soil and climatic inputs into clear crop and yield insights without manual computation. The workflow—from preprocessing to model inference—operates consistently and can be scaled for additional regions or crops.

A second achievement lies in achieving a strong balance between **scientific precision and system usability**. Instead of limiting the work to algorithmic performance, the study delivered a functional prototype deployable through a simple web interface built with Flask. This integration bridges the gap between academic research and field-level application, allowing predictions to be accessed instantly by end users.

The project also establishes a replicable framework for combining **classification and regression tasks** within one decision-support system. The modular design permits easy substitution of new datasets or updated algorithms, making the solution adaptable to future agricultural research or government data programs.

4.2.2 Comparative Analysis

Compared with earlier agricultural decision systems, the proposed model shows marked advancement in both prediction depth and computational practicality. Traditional statistical models often relied on linear assumptions that failed under varying soil or climatic patterns. In contrast, the present approach leverages

ensemble learning and feature normalization to maintain accuracy across heterogeneous data.

When benchmarked against publicly reported models achieving 85–90 percent accuracy, the proposed XGBoost-based pipeline reached 96 percent accuracy for classification and a 0.95 R^2 score for yield estimation. This performance gain stems not only from algorithmic sophistication but from disciplined data handling—consistent scaling, correlation control, and balanced training sets.

From an operational perspective, the model's low inference time (< 4 ms per sample) and compact memory footprint confirm its readiness for integration into lightweight agricultural-advisory tools or regional e-governance platforms. These characteristics make the system both **scientifically credible and technologically deployable**.

4.2.3 Limitations and Challenges

Despite its strengths, the project encountered several boundaries that open avenues for improvement. The available datasets represent averaged seasonal data; they do not capture rapid climatic changes or micro-regional soil diversity. Consequently, prediction accuracy may decline for atypical weather conditions or newly introduced crop varieties.The economic component currently assumes a static price at prediction time. Real-world agricultural markets fluctuate daily due to logistics, policy, and demand; integrating dynamic price feeds remains a future requirement. Hardware and data accessibility also pose constraints. Many small-scale farmers lack continuous internet or digital-device access. To maximize impact, the system must evolve toward offline or SMS-based versions. Finally, explainability remains limited—while accuracy is high, transparent reasoning behind each recommendation would improve farmer trust and adoption.

Chapter 5

5. CONCLUSION AND FUTURE WORK

5.1 Conclusion

The study successfully designed and implemented a data-driven crop-advisory system that unifies crop selection, yield forecasting, and profit estimation within one machine-learning framework. By employing **Gaussian Naïve Bayes** for classification and **XGBoost Regressor** for yield prediction, the system demonstrated high predictive reliability and computational efficiency.

The research contributes to sustainable agriculture by transforming empirical farm decisions into quantifiable, evidence-based insights. Beyond technical metrics, the solution emphasizes accessibility: a clean web interface, minimal computation delay, and compatibility with standard computer systems. This ensures that the technology can be adopted by agricultural offices, extension workers, and educational institutions to promote smarter cultivation planning.

In summary, the project validates that combining environmental analytics with predictive modeling can significantly reduce uncertainty in farming, improve productivity, and support income stability for cultivators.

5.2 Future Work

Future development will focus on expanding both data diversity and system functionality to create a truly adaptive decision-support platform. Planned enhancements include:

Short-Term Goals (0 – 6 months)

- Integrate live **weather-API feeds** and local soil-sensor data for region-specific predictions.

- Automate **market-price updates** through government or Agmarknet portals.
- Introduce a **multi-language dashboard** with Tamil, Hindi, and English interfaces.
- Add a lightweight **mobile-friendly web view** for quick access by farmers and field officers.

Medium-Term Goals (6 – 18 months)

- Incorporate **remote-sensing indices** (NDVI, soil moisture) from satellite data to improve yield precision.
- Embed **explainable-AI modules** that visually show factor contributions for each recommendation.
- Extend the dataset to include **state-wise and district-wise variations** for localized decision-making.
- Develop an **offline mode** using cached regional data for low-connectivity zones.

Long-Term Vision (18 + months)

- Connect with **IoT-based farm networks** to create a continuously learning ecosystem.
- Establish a **central agricultural analytics portal** integrating government, research, and user-generated data.
- Expand predictive capabilities to **crop-disease detection** and **fertilizer optimization** using deep learning.
- Explore **federated-learning frameworks** to maintain data privacy while improving national-scale accuracy.

Through these successive upgrades, the project can evolve from a prototype into a comprehensive **Smart Agriculture Intelligence System** that delivers timely, localized, and economically meaningful insights to every level of the agricultural community.

REFERENCES

[1] Ingle, A., "Crop Recommendation Using Machine Learning," Kaggle Dataset, 2021. Available: https://www.kaggle.com/datasets/atharvaingle/crop recommendation

[2] Badshah, A., "Crop Classification and Yield Prediction Using Robust Machine Learning Models for Agricultural Sustainability," IEEE Xplore, 2024.

[3] Ahmad, M., Ullah, S., Khan, I., et al., "Machine Learning Approaches for Crop Yield Prediction: A Comparative Study," Computers and Electronics in Agriculture, vol. 204, pp. 1–14, 2022.

[4] Iizumi, T., et al., "Global Dataset of Historical Yields (GDHY): Long-Term Crop Yield Records," Scientific Data, vol. 5, article 180070, 2018.

[5] Chen, X., Huang, Y., "Impact of Pesticide Usage on Crop Yield: A Global Assessment," Science of the Total Environment, vol. 726, pp. 138–149, 2020.

[6] ICAR – Indian Council of Agricultural Research. "National Agricultural Research Database: Soil, Climate & Yield Records," Govt. of India, 2023.

[7] Ministry of Agriculture & Farmers Welfare, Govt. of India. "Annual Report on Crop Production and Yield Statistics 2024–2025," 2023. 37

[8] Friedman, J. H., "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.

[9] Chen, T., Guestrin, C., “XGBoost: A Scalable Tree Boosting System,” Proceedings of the 22nd ACM SIGKDD, pp. 785–794, 2016.

[10] Murphy, K. P., Machine Learning: A Probabilistic Perspective, MIT Press, 2012.

[11] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825 2830, 2011.

[12] FAO – Food and Agriculture Organization. “Agro-Climatic Indicators for Agricultural Planning,” United Nations, 2020.

[13] Breiman, L., “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5 32, 2001.

[14] Han, J., Kamber, M., & Pei, J., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2022.

[15] Shinde, R., & Gawande, U., “Crop Yield Forecasting Using Machine Learning Techniques,” International Journal of Computer Applications, vol. 177, no. 44, pp. 1–6, 2020.