



Predictive Artificial Intelligence for Multi-Crop Yield Optimization in Bangladesh under Climate Variability

by

ONJON SUTRADHAR

Student ID: 2431963

MSc Data and Artificial Intelligence

Supervisor: Dr Aliyu Sambo

Project submitted as part of the requirements for the award of
MSc Data and Artificial Intelligence

April 2026

Declaration

I hereby declare that this thesis is my own work and has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is it being submitted concurrently in candidature for any degree or other award.

This thesis is submitted in partial fulfillment of the requirements for the degree of MSc Data and Artificial Intelligence.

The work presented in this thesis is the result of my own independent investigation, except where otherwise stated. All sources are acknowledged by explicit references. The views expressed are my own.

I give consent for this thesis, if accepted, to be made available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to external organisations.

Signed: _____

Date: _____

(Student Name)

Abstract

Climate variability poses a major threat to agriculture in Bangladesh, with variable rainfall, flooding and temperature anomalies. Conventional crop yield forecasting approaches are largely based on farmers' knowledge and basic statistical models, which fail to incorporate complex environmental interactions and adapt to environmental change. This study overcomes these challenges by proposing and assessing a robust machine learning framework for multi-crop yield prediction using the SPAS-Dataset-BD, which comprises 4,607 farming records on 73 types of crops, 64 districts and three seasons.

A structured approach is adopted, involving exploratory data analysis and supervised machine learning classification techniques. The statistical analyses found significant dependencies between crops and seasons, with Aman rice being strongly associated with the Kharif-2 season (posterior probability greater than 0.75), and identified primary feature correlations, especially between cultivation area and production ($r = 0.92$). Production had a strong right skew (greater than 1.5) and tests did not support the assumption of normality (Shapiro-Wilk p less than 0.001).

We used and compared several machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, K-Nearest Neighbors, Support Vector Machines, Decision Trees, Extra Trees, and Naive Bayes classifiers. The final Logistic Regression model produced 65.18% test accuracy with balanced performance across the crop types (F1-macro: 0.651), and much improved over the baseline methods. Gradient Boosting showed the best cross-validation accuracy (F1-macro: 0.657) but slightly lower test accuracy.

The most significant predictors were area, ratio of area and production and district specific factors. Seasonal analysis demonstrated best performance in Kharif-2 season, with some low-frequency crops being difficult for classification, owing to sparse training samples. This study offers a replicable ML framework for yield prediction in developing countries, showcases the feasibility of ensemble methods in multi-crop settings, and delivers evidence for data-driven decision-making support in agriculture.

The findings suggest that machine learning methods can support precision agriculture programs in Bangladesh to provide evidence to farmers and policy makers for crop selection and resource planning in the face of climate change.

Contents

1	Introduction	1
1.1	Background and Context	1
1.2	Motivation and Problem Statement	1
1.3	Research Aim and Objectives	2
1.4	Scope and Delimitations	3
1.5	Contribution and Significance	4
1.6	Report Structure	4
2	Literature Review	5
2.1	Introduction to the Literature Review	5
2.2	Agricultural Yield Prediction: Evolution and Current State	5
2.3	Machine Learning in Agricultural Applications	6
2.4	Critical Gap Analysis: Multi-Crop Prediction in Developing Countries	7
2.5	Synthesis and Research Positioning	7
3	Research and Development Methodology	10
3.1	Research Design Overview	10
3.2	Dataset Description and Data Acquisition	10
3.2.1	Dataset Source and Characteristics	10
3.2.2	Data Variables and Features	11
3.3	Exploratory Data Analysis Methodology	11
3.3.1	Descriptive Statistical Analysis	11
3.3.2	Bivariate and Multivariate Analysis	12
3.3.3	Probability and Statistical Inference	12
3.4	Feature Engineering and Preprocessing	12
3.4.1	Data Cleaning	12
3.4.2	Feature Engineering	13
3.4.3	Data Transformation and Scaling	13
3.5	Machine Learning Model Development	14
3.5.1	Problem Formulation	14

3.5.2	Training-Validation-Test Split	14
3.5.3	Machine Learning Algorithms Implemented	14
3.5.4	Hyperparameter Optimization	15
3.6	Model Evaluation Methodology	16
3.6.1	Evaluation Metrics	16
3.6.2	Cross-Validation Strategy	17
3.6.3	Statistical Significance Testing	17
3.7	Interpretability and Feature Importance Analysis	17
3.8	Computational Infrastructure and Implementation	18
3.9	Ethical Considerations	18
3.10	Research Limitations	18
4	Design and Implementation	19
4.1	System Architecture Overview	19
4.2	Data Preprocessing Pipeline Implementation	19
4.2.1	Data Loading and Initial Validation	19
4.2.2	Missing Value Analysis and Treatment	19
4.2.3	Outlier Detection and Treatment	20
4.3	Feature Engineering Implementation	20
4.4	Target Variable Transformation	21
4.4.1	Approach V1: Global Quantile-Based Classification	21
4.4.2	Approach V2: Crop-Aware Percentile Ranking	22
4.5	Train-Validation-Test Split Implementation	22
4.6	Preprocessing Pipeline Construction	23
4.7	Machine Learning Model Implementation	24
4.7.1	Baseline Model Training (Approaches V1 and V2)	24
4.7.2	Hyperparameter Optimization Implementation	24
4.8	Evaluation and Reporting Module	25
4.9	Final Approach: Regression-Based Production Prediction	26
4.10	Implementation Challenges and Solutions	28
4.10.1	Challenge 1: Categorical Variable Encoding	28

4.10.2	Challenge 2: Class Imbalance	28
4.10.3	Challenge 3: Feature Scaling	28
5	Results and Discussion	28
5.1	Exploratory Data Analysis Results	28
5.1.1	Descriptive Statistics	28
5.1.2	Distributional Analysis and Normality Testing	29
5.1.3	Correlation Analysis	29
5.1.4	Crop-Season Association Analysis	30
5.1.5	Extended Correlation Matrix	31
5.1.6	Normality Test Details	31
5.1.7	Crop-Season Association	33
5.2	Machine Learning Model Performance Results	35
5.2.1	Evolutionary Methodology: Three Approaches Compared	35
5.2.2	Approach V1: Global Classification Results	36
5.2.3	Approach V2: Crop-Aware Classification Results	38
5.2.4	Hyperparameter Optimization Results (V1 Approach)	40
5.2.5	Cross-Validation Fold Results	40
5.3	Complete Confusion Matrices	41
5.3.1	V1 Approach (Global Quantile Classification)	41
5.3.2	V2 Approach (Crop-Aware Classification)	43
5.3.3	Final Model Test Set Performance (V1 Approach)	44
5.3.4	Final Regression Approach: Breakthrough Performance	44
5.3.5	Per-Class Performance Analysis	46
5.4	Feature Importance Analysis	46
5.4.1	Random Forest Feature Importance	46
5.5	Error Analysis and Model Limitations	48
5.5.1	Confusion Matrix Analysis	48
5.5.2	Per-Season Performance	48
5.5.3	Per-Crop Error Analysis	49
5.6	Model Interpretation and Practical Implications	49

5.6.1	Decision Boundaries	49
5.6.2	Practical Applicability for Farmers and Policymakers	49
5.7	Comparison with Existing Literature	50
5.8	Limitations of Results	50
6	Conclusions and Recommendations	51
6.1	Summary of Research Findings	51
6.2	Contributions to Knowledge	52
6.3	Practical Implications and Recommendations	52
6.3.1	For Agricultural Extension Services	52
6.3.2	For Policymakers	53
6.3.3	For Farmers and Farmer Organizations	53
6.4	Research Limitations	53
6.5	Future Research Directions	54
6.6	Final Remarks	54
7	Reflection	55
7.1	Learning Journey and Personal Development	55
7.2	Technical Skills Acquired	55
7.3	Research Methodology Insights	56
7.4	Challenges Encountered and Overcome	56
7.5	Collaboration and Support	57
7.6	Ethical Reflections	57
7.7	Professional Development and Career Implications	57
7.8	If I Were to Start Again	57
7.9	Concluding Thoughts	58
A	Machine Learning Parameters	61
A.1	Complete Hyperparameter Grids	61
A.1.1	Random Forest Hyperparameter Grid	61
A.1.2	Gradient Boosting Hyperparameter Grid	61

B	Code Repository	62
B.1	Project Structure	62
B.2	Key Jupyter Notebooks	62
B.2.1	Comprehensive_EDA.ipynb	62
B.2.2	Multicrop_Yield_Prediction.ipynb (Final Approach)	63
B.2.3	ML_Results/Yield_Prediction_Multiclass_Traditional_ML.ipynb (V1)	63
B.2.4	ML_Results/v2/..._v2.ipynb (V2)	64
B.3	Reproducibility Notes	64
C	Ethics Approval Documentation	64
C.1	Ethics Review Summary	64
C.2	Ethical Considerations Addressed	65
C.2.1	Data Ethics	65
C.2.2	Algorithmic Fairness	65
C.2.3	Potential Impact and Responsible Use	65
D	Dataset Description and Access	66
D.1	Dataset Information	66
D.2	Dataset Characteristics	66
D.3	Variable Definitions	67

List of Figures

1	Q-Q Plots for All Continuous Variables. Consistent deviations from the theoretical normal line (red) across all variables justify use of non-parametric methods and robust scaling.	32
2	Normal Curve Overlays on Histograms. Observed distributions (blue bars) show extreme right skewness for Production and Area, with temperature and humidity displaying different distributional patterns than theoretical normal curves (red lines). . .	33
3	Conditional Probability Heatmap P(Crop—Season). Strong diagonal and block patterns demonstrate deterministic crop-season relationships aligned with Bangladesh’s monsoon agricultural calendar.	34
4	Final Approach Model Comparison. Stacking Ensemble ($R^2 = 0.9974$) and Voting Ensemble ($R^2 = 0.9872$) substantially outperform individual models, demonstrating the power of ensemble methods for agricultural prediction.	36
5	V1 Confusion Matrix - Logistic Regression (Best Model, 66.45% validation accuracy). Strong diagonal shows effective class separation with global quantile approach.	37
6	V2 Confusion Matrix - Gradient Boosting (Best Model, 38.16% validation accuracy). Increased classification difficulty evident from more dispersed predictions across categories compared to V1.	39
7	V1 Approach - Confusion Matrices for Top 4 Models	41
8	V1 Approach - Confusion Matrices for Remaining Models	42
9	V1 Approach - Tuned Model Confusion Matrices	42
10	V2 Approach - Selected Confusion Matrices (Crop-Aware Targets)	43
11	Predicted vs Actual Production (Stacking Ensemble). Strong linear correlation ($R^2 = 0.9974$) with minimal systematic bias demonstrates exceptional predictive accuracy. .	45
12	5-Fold Cross-Validation Results. Consistent performance across folds demonstrates model robustness and minimal overfitting.	45
13	Feature Importance Rankings from Final Regression Approach (Random Forest). Area dominates with 13.71% importance, followed by climate variables (Avg Humidity 3.11%, Max RH 2.91%, Avg Temp 2.65%). This hierarchy confirms that land allocation is the primary lever for production optimization, with climate as a secondary modulating factor.	47

List of Tables

1	Comparison of Selected Studies (Part I)	8
2	Comparison of Selected Studies (Part II)	9
3	Summary of Recent Crop Modeling and Classification Studies	9
4	Data Variables and Features	11
5	Model Hyperparameters	15
6	Classification Evaluation Metrics	16
7	Descriptive Statistics for Key Continuous Variables	29
8	Normality Test Results (Shapiro-Wilk Test)	29
9	Correlation Coefficients with Production	30
10	Posterior Probabilities for Major Crop-Season Combinations	30
11	Complete Correlation Matrix for Numerical Variables	31
12	Crop-Season Contingency Table (Top 20 Crops by Frequency)	34
13	Comprehensive Comparison: V1 vs V2 vs Final Approach	35
14	V1 Approach: Baseline Model Performance (Global Quantiles)	37
15	V2 Approach: Model Performance (Crop-Aware Percentile Ranking)	38
16	Hyperparameter Tuning Results	40
17	5-Fold Cross-Validation Results (V1 Approach - Best Models)	40
18	V1 Final Model Test Performance (Logistic Regression)	44
19	Final Approach: Regression Model Performance on Test Set	44
20	Per-Class Performance (Test Set - Logistic Regression)	46
21	Top 15 Most Important Features (Random Forest)	46
22	Confusion Matrix - Logistic Regression (Test Set)	48
23	Model Performance by Season (Test Set)	48
24	Random Forest Hyperparameter Search Space	61
25	Gradient Boosting Hyperparameter Search Space	61
26	SPAS-Dataset-BD Variable Definitions	67

1 Introduction

1.1 Background and Context

Agriculture constitutes a cornerstone of Bangladesh's economy and food security infrastructure, directly employing over 40% of the workforce and contributing approximately 13% to the national GDP [1]. The country's agricultural landscape is characterized by intensive cultivation across diverse agro-ecological zones, with rice serving as the primary staple crop alongside vegetables, pulses, oilseeds, and fruits. However, Bangladesh's geographical location in the Bengal Delta and tropical monsoon climate zone renders its agricultural systems highly vulnerable to climate variability and extreme weather events.

Climate change manifests through multiple pathways that directly impact agricultural productivity: altered monsoon patterns disrupting traditional planting and harvesting schedules, increased frequency of flooding and drought events, rising temperatures affecting crop phenology, and sea-level intrusion causing soil salinity in coastal regions [2]. These climate-induced challenges are compounded by increasing population pressure, limited arable land availability, and resource constraints faced by smallholder farmers who constitute the majority of Bangladesh's agricultural workforce. [3]

Traditional agricultural decision-making in Bangladesh relies predominantly on experiential knowledge passed through generations, seasonal calendars based on historical patterns, and simplified advisory services. [4] While this indigenous knowledge has sustained agricultural practices for centuries, it proves increasingly inadequate in the face of rapid and unpredictable climate changes. The temporal lag between environmental shifts and adaptive responses results in suboptimal crop selection, inappropriate resource allocation, and ultimately, reduced yields and economic losses.

1.2 Motivation and Problem Statement

The motivation for this research stems from the critical need to boost agricultural resilience and productivity in climate-vulnerable regions. This can be achieved through data-driven predictive systems. Artificial intelligence and machine learning have advanced across many domains. However, their use in multi-crop yield prediction in developing agricultural contexts is still limited. Most existing research focuses on single-crop scenarios in controlled or well-resourced environments. Less attention is given to the complexities of diverse cropping systems, especially those facing data scarcity and high climate uncertainty.

Bangladesh presents a particularly compelling case study due to its intensive multi-cropping systems, high climate vulnerability, substantial availability of agricultural datasets through government initiatives, and urgent need for precision agriculture solutions. The country's agricultural statistics system maintains comprehensive records of crop production, environmental conditions, and district-level agricultural activities, providing a rich foundation for data-driven analysis.

The fundamental problem addressed by this research is: *How can machine learning techniques be*

effectively applied to predict agricultural outcomes across multiple crop types in Bangladesh, accounting for climate variability and regional heterogeneity, while ensuring model interpretability and practical applicability for farmer decision support?

This problem encompasses several interconnected challenges: Data complexity — agricultural datasets exhibit non-standard statistical properties including heavy-tailed distributions, spatial autocorrelation, temporal dependencies, and class imbalance across crop types; multi-crop heterogeneity — different crops respond distinctly to environmental factors, require varied cultivation practices, and exhibit diverse yield patterns, necessitating models that can capture crop-specific behaviors while maintaining generalization capability; climate variability — environmental factors interact in complex, non-linear ways, with threshold effects, lag relationships, and multiplicative interactions that challenge traditional linear modeling approaches; practical applicability — models must balance predictive accuracy with interpretability, computational efficiency, and accessibility to be practically useful for farmer advisory systems and policy planning; and limited labeled data — despite comprehensive production records, certain crop-district-season combinations have sparse observations, creating challenges for model training and evaluation.

1.3 Research Aim and Objectives

The overarching aim of this research is to develop, implement, and evaluate a machine learning-based predictive framework for multi-crop agricultural yield assessment in Bangladesh, with emphasis on robustness under climate variability and practical applicability for decision support.

This aim is operationalized through the following specific objectives:

1. Objective 1: Data Acquisition and Preparation

To obtain, clean, and prepare a comprehensive agricultural dataset for Bangladesh encompassing multiple crops, districts, and seasons, with appropriate handling of missing values, outliers, and data quality issues.

2. Objective 2: Exploratory Statistical Analysis

To conduct rigorous exploratory data analysis employing descriptive statistics, probability distributions, hypothesis testing, and correlation analysis to understand fundamental data characteristics, identify patterns, and validate statistical assumptions.

3. Objective 3: Feature Engineering and Selection

To engineer relevant features from raw data, apply appropriate transformations to address distributional assumptions, encode categorical variables effectively, and select features that maximize predictive power while minimizing model complexity.

4. Objective 4: Model Development and Implementation

To implement multiple supervised machine learning algorithms including traditional methods and ensemble techniques, with hyperparameter tuning and cross-validation.

5. **Objective 5: Model Evaluation and Comparison**

To evaluate implemented models using appropriate metrics for multi-class classification (accuracy, F1-scores, balanced accuracy), compare performance across algorithms, and identify the optimal model configuration for deployment.

6. **Objective 6: Interpretability and Feature Importance Analysis**

To analyze feature importance, examine model predictions across different crops, seasons, and districts, and provide interpretable insights into factors driving agricultural outcomes.

7. **Objective 7: Practical Applicability Assessment**

To assess the practical feasibility of deploying the developed framework for real-world agricultural advisory services, considering computational requirements, data collection feasibility, and user accessibility.

1.4 **Scope and Delimitations**

This research focuses specifically on:

Geographical Scope: Bangladesh, utilizing district-level aggregated data across 64 administrative districts.

Temporal Scope: Agricultural production data spanning three primary growing seasons (Kharif-1, Kharif-2, and Rabi), representing different climatic conditions and cropping patterns.

Crop Coverage: Multi-crop analysis encompassing 73 different crop types including cereals, pulses, vegetables, fruits, and cash crops, reflecting Bangladesh's diverse agricultural portfolio.

Methodological Approach: Supervised machine learning classification models for yield category prediction, complemented by comprehensive exploratory data analysis.

Data Source: The SPAS-Dataset-BD publicly available dataset, documented and peer-reviewed in academic literature.

The scope of this research is limited by several practical constraints. First, the analysis relies on aggregated district-level data instead of farm or plot-level observations, mainly due to the unavailability of more detailed datasets. As a result, some local variations may not be fully captured.

In addition, weather and climate conditions are represented using seasonal averages rather than detailed time-series data. While this simplifies the analysis, it may overlook short-term fluctuations that could influence agricultural outcomes. The study also focuses on conventional machine learning methods instead of deep learning approaches, as the dataset size is relatively small and the need for model interpretability is important.

The research does not account for economic factors, market conditions, or farmer behavior, as relevant data for these variables is not available. Similarly, no field validation or farmer surveys have been conducted, meaning that all findings are based solely on secondary data analysis. Soil-related

information is limited to basic regional characteristics. More detailed soil quality parameters are not included in the dataset, which restricts the depth of analysis in this area.

1.5 Contribution and Significance

This research makes several distinct contributions to the intersection of agricultural science, climate adaptation, and machine learning:

Methodological Contribution: Development of a comprehensive, reproducible machine learning pipeline specifically designed for multi-crop yield prediction in data-constrained developing country contexts, with emphasis on handling class imbalance, seasonal dependencies, and spatial heterogeneity.

Empirical Contribution: Systematic comparison of multiple machine learning algorithms on a real-world agricultural dataset from Bangladesh, providing empirical evidence regarding which methods perform optimally for multi-crop scenarios and under what conditions.

Practical Contribution: Demonstration of how machine learning can be operationalized for agricultural decision support in resource-limited settings, with analysis of practical deployment considerations and actionable insights for farmers and policymakers.

Technical Contribution: Comprehensive exploratory statistical analysis revealing fundamental characteristics of agricultural production data, including distributional properties, seasonal patterns, and inter-variable relationships that inform modeling choices.

Domain Contribution: This study applies machine learning methods to the agricultural sector of Bangladesh, which is an important but still underexplored area. It supports ongoing efforts toward climate-smart agriculture and aligns with broader sustainable development goals.

The importance of this work goes beyond academic value. For farmers and agricultural extension services, it provides useful evidence that can support more informed decisions about crop selection and planning. The results can also help policymakers in shaping agricultural strategies, improving resource allocation, and planning for climate adaptation at both regional and national levels.

From a technical perspective, the study shows how machine learning can be applied in real agricultural settings where data is often noisy and inconsistent. It also adds to climate adaptation research by giving practical evidence of how AI can strengthen agricultural resilience in vulnerable regions. At the same time, the work connects with development goals by addressing key issues such as food security, climate action, and sustainable agriculture.

1.6 Report Structure

The remainder of this report is organized as follows:

Section 2 (Literature Review) critically examines existing research on crop yield prediction, ma-

chine learning applications in agriculture, climate impact assessment, and multi-crop modeling approaches, identifying research gaps and positioning this work within the broader literature.

Section 3 (Methodology) details the research methodology, including dataset description, exploratory data analysis techniques, feature engineering procedures, machine learning algorithms employed, hyperparameter optimization strategies, evaluation metrics, and experimental design.

Section 4 (Design and Implementation) documents the technical implementation of the predictive framework, including data preprocessing pipeline, model architecture, training procedures, and computational infrastructure.

Section 5 (Results and Discussion) presents comprehensive results from exploratory analysis and machine learning experiments, including statistical findings, model performance comparisons, feature importance analysis, error analysis, and critical discussion of implications and limitations.

Section 6 (Conclusions and Recommendations) summarizes key findings, discusses contributions and limitations, outlines practical implications, and proposes directions for future research.

Section 7 (Reflection) provides personal reflection on the research process, learning outcomes, challenges encountered, and professional development insights.

Appendices include supplementary materials such as additional statistical analyses, code samples, detailed hyperparameter configurations, ethics approval documentation, and extended result tables.

2 Literature Review

2.1 Introduction to the Literature Review

This literature review examines existing research across four interconnected domains: (1) agricultural yield prediction methodologies, (2) machine learning applications in agriculture, (3) climate variability impacts on crop production, and (4) multi-crop modeling approaches in developing countries. The review synthesizes insights from these domains to identify research gaps, establish theoretical foundations, and position the current research within the broader academic landscape.

The review follows a thematic organization rather than chronological approach, enabling deeper engagement with conceptual issues. Each subsection examines key studies, methodological approaches, findings, and limitations, culminating in a synthesis that articulates the specific gap addressed by this research.

2.2 Agricultural Yield Prediction: Evolution and Current State

Agricultural yield prediction has increasingly shifted toward machine learning (ML)-driven approaches, driven by the growing availability of large-scale agricultural, climatic, and remote sensing data. Unlike traditional methods, ML models focus on learning complex patterns directly from data without re-

quiring explicit assumptions about underlying biological processes. [5] This makes them particularly suitable for capturing the highly non-linear and multi-factor interactions that characterize agricultural systems.

Support vector machines (SVMs), decision trees, and linear regression were the main supervised learning methods used in early machine learning applications for yield prediction. Although interpretability is provided by linear models, their capacity to represent intricate relationships is constrained. Tree-based models and kernel methods improved predictive performance by handling non-linearity and feature interactions more effectively. [6]

In order to improve accuracy and resilience, ensemble learning techniques like Random Forests and Gradient Boosting Machines have been increasingly popular in recent years. Concurrently, deep learning techniques, such as artificial neural networks (ANNs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), have proven to be highly effective in integrating diverse data sources, including satellite imagery, soil characteristics, and weather time series. [7]

Despite these advances, challenges remain in terms of data quality, model generalization across regions, and interpretability. As a result, current research increasingly focuses on hybrid and explainable ML models that balance predictive performance with transparency and practical usability in real-world agricultural decision-making.

2.3 Machine Learning in Agricultural Applications

Recent years have seen a sharp increase in the use of machine learning to agricultural problems due to the availability of more data, advancements in processing, and successful implementations in adjacent fields. Compared to traditional methods, machine learning approaches have a number of advantages, including the capacity to handle high-dimensional feature spaces, the ability to learn complex non-linear relationships from data, robustness to noise and missing data, and the potential for continuous improvement as new data becomes available.

Decision Tree and Ensemble Methods: Gradient and Random Forests Because boosting algorithms can handle heterogeneous data types, capture interaction effects, and offer feature relevance metrics, they have become especially useful for agricultural prediction applications. Jeong et al. [8] demonstrated that Random Forests outperformed traditional regression for corn yield prediction using weather and remote sensing data, achieving R^2 values exceeding 0.85 in favorable conditions. However, performance degraded significantly under extreme climate events, suggesting limitations in capturing tail-end distributions.

Support Vector Machines: SVMs have been applied to crop yield classification and regression with mixed results. Their strength lies in handling high-dimensional spaces and non-linear decision boundaries through kernel tricks, but computational complexity increases substantially with dataset size, and hyperparameter tuning requires careful consideration [9].

Neural Networks and Deep Learning: In controlled study settings, deep learning methods have shown

great results. These include convolutional neural networks (CNNs) for image analysis and recurrent neural networks (RNNs) for time series. Khaki and Wang [10] achieved state-of-the-art performance using deep neural networks for corn yield prediction. However, these models typically require large training datasets and huge computational resources.

2.4 Critical Gap Analysis: Multi-Crop Prediction in Developing Countries

Despite substantial progress in agricultural machine learning, several critical gaps remain:

Single-Crop Focus: The majority of existing research focuses on major crops (corn, wheat, soybeans) in developed countries with well-maintained data infrastructure. Multi-crop prediction in diverse agricultural systems receives limited attention, yet represents the reality for most smallholder farmers in developing countries [11].

Data Requirements vs. Availability: Advanced models often require high-resolution temporal data, extensive soil parameters, and detailed weather stations that are unavailable in many developing regions. Research on model performance under data constraints is sparse [12].

Interpretability-Performance Tradeoff: While deep learning achieves high predictive accuracy, its black-box nature limits farmer trust and adoption. The balance between performance and interpretability receives insufficient attention in agricultural ML research [13].

Climate Variability Adaptation: Most models are trained on historical data and assume stationary climate distributions. Their robustness under non-stationary climate conditions and extreme events remains largely unexplored [14].

Practical Deployment Challenges: Academic studies typically end with model performance metrics, rarely addressing computational feasibility, data collection infrastructure, user interface design, and integration with existing agricultural extension systems necessary for real-world deployment.

This research directly addresses these gaps by: (1) focusing on multi-crop prediction across 73 crop types, (2) utilizing publicly available aggregated data typical of developing country contexts, (3) emphasizing interpretable ensemble methods while comparing multiple algorithms, (4) explicitly examining performance across different seasons representing climate variability, and (5) discussing practical deployment considerations throughout.

2.5 Synthesis and Research Positioning

The literature review reveals that while agricultural machine learning has advanced substantially, a significant gap exists between research focused on controlled, single-crop scenarios in developed countries and the practical needs of diverse, resource-constrained agricultural systems in developing regions. This research is uniquely positioned at this intersection, contributing methodological insights, empirical evidence, and practical frameworks that advance both academic understanding and real-world applicability.

By focusing on Bangladesh-a country with high agricultural diversity, severe climate vulnerability, reasonable data availability, and urgent need for climate adaptation-this research provides a meaningful test case for assessing how machine learning can realistically contribute to agricultural resilience in challenging contexts. The emphasis on reproducibility, interpretability, and practical deployment considerations distinguishes this work from purely algorithmic studies, aligning with the emerging paradigm of responsible AI in development contexts.

Previous research demonstrates that machine learning models, including neural networks and ensemble methods, outperform conventional statistical models in capturing non-linear agricultural processes. However, much of the existing literature focuses on controlled datasets or single-factor analysis, limiting real-world applicability.

There is also a notable gap in studies that balance predictive accuracy with interpretability, which is essential for farmer trust and adoption. This research positions itself within this gap by emphasizing both performance and usability, synthesizing insights from AI, agronomy, and climate science.

Table 1: Comparison of Selected Studies (Part I)

Paper	Summary	Methodology	Limitations
Comparative Analysis of Machine Learning Models for Crop Yield Prediction Across Multiple Crop Types (2025) Yashraj Patil et al. [11]	This study compares several machine learning regression models-Gaussian Process, Random Forest, KNN, XG-Boost, and AdaBoost-for predicting yields of six crops. Results show that no single model performs best across all crops; performance is crop-specific.	Multiple ML models were trained and evaluated on six crop datasets with systematic feature testing.	Dataset-specific findings limit generalization across regions.
Enhanced precision Multi-Crop farming using Machine Learning Recommendations (2024) Prem Kumar et al. [13]	A real-time crop recommendation system using soil nutrient sensor data to determine optimal crop selection.	Classification algorithms trained on soil nutrient datasets.	Limited environmental and regional validation.
Simultaneous multi-crop land suitability prediction from remote sensing data using semi-supervised learning (2023) Amanjot Bhullar et al. [12]	Semi-supervised deep learning model for multi-crop land suitability prediction.	MLP architecture using environmental datasets.	Limited dataset size and missing spatial dependencies.

Table 2: Comparison of Selected Studies (Part II)

Paper	Summary	Methodology	Limitations
Hybrid deep learning with a fusion model of weather and soil-based multi-crop recommendations (2025) Mohan M. et al [15]	Integrated deep learning framework combining CNN and graph-based LSTM for multi-crop recommendation.	Spatial-temporal modelling with optimization techniques and explainable AI.	Computational complexity and deployment limitations.
UniCrop: Universal Data Pipeline for Crop Yield Prediction (2026) Emiliya Khidirova et al. [16]	Automated data engineering pipeline integrating satellite and environmental data sources.	Feature engineering and ML models including LightGBM.	Validated using limited datasets.
Multi-Crop Classification Using Feature Selection-Coupled ML Classifiers (2022) Shan He et al. [17]	Evaluation of feature selection techniques combined with ML classifiers for crop classification.	Feature extraction from remote sensing data with ML classifiers.	Limited geographic and temporal coverage.

Table 3: Summary of Recent Crop Modeling and Classification Studies

Study Focus		Data Source	Methods / Models		Key Contribution	Limitations
Multi-crop Regression	ML	Crop datasets	Gaussian Process, RF, KNN, XGBoost		Demonstrated crop-specific model performance	Limited generalization
Real-time based Crop Recommendation	Soil-Crop	Soil nutrient sensors	Decision Tree		Real-time recommendation	Limited environmental variables
Semi-supervised Land Suitability		Remote sensing datasets	MLP		Multi-crop suitability modelling	Small dataset
Hybrid Deep Learning Framework	Deep	Soil and weather data	CNN + Graph LSTM		Spatial-temporal modelling	High computational cost
UniCrop Pipeline	Data	Satellite and environmental datasets	LightGBM, RF, SVR		Automated feature engineering	Limited validation
FS-ML Coupling		Remote sensing data	RF, SVM, KNN		Improved classification accuracy	Limited scope

3 Research and Development Methodology

3.1 Research Design Overview

This research adopts a quantitative, empirical methodology grounded in the supervised machine learning paradigm. The overall research design follows a systematic pipeline approach, progressing through five distinct phases: (1) data acquisition and preparation, (2) exploratory statistical analysis, (3) feature engineering and preprocessing, (4) model development and hyperparameter optimization, and (5) evaluation and comparative analysis. This structure aligns with established best practices in data science research while accommodating domain-specific agricultural considerations.

The research philosophy is fundamentally positivist, seeking to establish empirical relationships between measurable agricultural and environmental variables and crop yield outcomes through data-driven pattern recognition. The methodology emphasizes reproducibility, systematic comparison, and rigorous evaluation, ensuring that findings can be validated and built upon by subsequent research.

3.2 Dataset Description and Data Acquisition

3.2.1 Dataset Source and Characteristics

The research utilizes the SPAS-Dataset-BD (Statistical analysis and Predictive Analytics for Smart agriculture in Bangladesh), a comprehensive agricultural production dataset publicly available through the Mendeley Data repository. This dataset is documented in a peer-reviewed data descriptor published in *Data in Brief*, ensuring data provenance, quality assurance, and academic rigor [18, 19].

The dataset encompasses agricultural production records for Bangladesh across three dimensions:

- **Geographical Coverage:** 64 administrative districts covering all major agro-ecological zones of Bangladesh
- **Temporal Coverage:** Three primary growing seasons representing different climatic conditions:
 - Kharif-1 (pre-monsoon season, April-June)
 - Kharif-2 (monsoon season, July-October)
 - Rabi (post-monsoon/winter season, November-March)
- **Crop Diversity:** 73 distinct crop types including cereals (rice varieties, wheat, maize), pulses (lentil, chickpea, mung bean), vegetables (potato, tomato, cabbage), fruits (mango, banana, papaya), and cash crops (jute, sugarcane, tobacco)

3.2.2 Data Variables and Features

The dataset contains 4,607 observations, with each record representing aggregated production statistics for a specific crop-district-season combination. The original dataset comprises 15 variables as shown in the following table:

Table 4: Data Variables and Features

Category	Variable	Description
Target	Production	Total production quantity (in metric tons) for the crop in the specified district and season
Predictors	Area	Cultivation area (in hectares) for the crop
	AP Ratio	Area-to-Production ratio indicating yield per unit area
	District	Administrative district name (categorical, 64 levels)
	Season	Growing season identifier (categorical, 3 levels)
	Crop Name	Crop type identifier (categorical, 73 levels)
	Avg Temp	Average temperature (°C)
	Avg Humidity	Average relative humidity (%)
	Max Temp	Maximum recorded temperature (°C)
	Min Temp	Minimum recorded temperature (°C)
	Max Relative Humidity	Maximum recorded relative humidity (%)
	Min Relative Humidity	Minimum recorded relative humidity (%)
	Transplant	Typical transplantation/sowing period (text description)
	Growth	Primary growth period (text description)
	Harvest	Typical harvest period (text description)

3.3 Exploratory Data Analysis Methodology

Exploratory Data Analysis (EDA) serves multiple critical purposes in this research: understanding fundamental data characteristics, identifying potential quality issues, revealing patterns and relationships, validating statistical assumptions, and informing modeling decisions. The EDA phase employs both descriptive and inferential statistical techniques.

3.3.1 Descriptive Statistical Analysis

Univariate Analysis: The study begins with univariate analysis to understand individual variables in the dataset. Measures of central tendency such as mean, median, and mode are used for continuous variables to capture their typical values. Dispersion is examined through standard deviation, variance, and interquartile range to understand variability. The shape of the data distribution is also assessed using skewness and kurtosis. For categorical variables, frequency distributions and proportion analysis are used to observe how data is distributed across different categories.

Normality Assessment: Since many machine learning methods assume certain data distributions,

normality testing is carried out. This is done through visual inspection using histograms with normal distribution overlays, along with Quantile-Quantile (Q-Q) plots that compare observed data with a theoretical normal distribution. In addition, formal statistical tests such as the Shapiro-Wilk test and the Kolmogorov-Smirnov test are applied to validate the presence or absence of normality.

Outlier Detection: Outliers can significantly affect model performance, so they are carefully identified using multiple approaches. Box plot visualization is used with the $1.5 \times \text{IQR}$ rule to detect extreme values. Z-score analysis is also applied to identify observations that lie more than three standard deviations away from the mean. Domain knowledge is further used to distinguish between actual extreme values and possible data entry errors.

3.3.2 Bivariate and Multivariate Analysis

Correlation Analysis: To understand relationships between variables, correlation analysis is performed. Pearson correlation coefficients are used to measure linear relationships between continuous variables. Correlation matrices combined with hierarchical clustering help identify groups of related variables. Scatter plots with regression lines provide visual insight into relationships between pairs of variables. In cases where relationships are non-linear but monotonic, Spearman rank correlation is used.

Categorical Variable Analysis: For categorical data, relationships are examined using contingency tables to explore associations such as crop and season combinations. Chi-square tests are applied to check for independence between variables, while Cramér's V is used to measure the strength of these associations.

3.3.3 Probability and Statistical Inference

Joint and Conditional Probability Analysis: A Bayesian probability framework is used to analyze dependencies between crops and seasons. This includes calculating joint probability distributions $P(\text{Crop}, \text{Season})$ and conditional probabilities such as $P(\text{Crop}|\text{Season})$ and $P(\text{Season}|\text{Crop})$. Posterior probabilities are then estimated using Bayes' theorem to better understand these relationships.

Confidence Interval Estimation: To quantify uncertainty in the estimates, 95% confidence intervals are calculated for key population parameters. This provides a range within which the true values are expected to lie with a high level of confidence.

3.4 Feature Engineering and Preprocessing

3.4.1 Data Cleaning

Missing Value Treatment: Missing data is handled through a structured approach. First, patterns of missing values are identified and analyzed to understand whether they occur randomly or sys-

tematically. Based on this, appropriate imputation methods are applied. Continuous variables are filled using mean or median values, while categorical variables are handled using mode imputation. Sensitivity analysis is also conducted to examine how these imputations affect the overall results.

Data Quality Assurance: Several steps are taken to ensure data quality. Duplicate records are identified and removed, and consistency checks are performed across related variables. Data values are also validated against expected ranges based on domain knowledge. Any obvious typographical errors are corrected to maintain accuracy.

3.4.2 Feature Engineering

Beyond the original variables, additional features are created to better capture meaningful patterns in the data. Temperature-based features include temperature range, calculated as the difference between maximum and minimum temperature, and temperature deviation, which measures how far the average temperature is from the optimal level for each crop. Similarly, humidity-based features include humidity range and humidity deviation.

Interaction features are also introduced to capture combined effects between variables. For example, the interaction between average temperature and humidity is computed, as well as the interaction between cultivation area and the area-to-production ratio.

Categorical variables are transformed using different encoding methods. Label encoding is appropriate for tree-based models and gives categories integer values. For linear and distance-based models, one-hot encoding generates binary variables for every category. In order to prevent overfitting, target encoding—in which categories are encoded according to target statistics—is also taken into account.

3.4.3 Data Transformation and Scaling

Normality-Inducing Transformations: For variables that show skewed distributions, transformations are applied to make them more suitable for modeling. Logarithmic transformation using $\log(x + 1)$ is applied to handle zero values. Square root transformation is used for moderately skewed data, while Box-Cox transformation is applied to find an optimal power transformation that improves normality.

Feature Scaling: Feature scaling is implemented according to the requirements of various algorithms. For models like SVM and logistic regression that are sensitive to feature scale, standardization—also called Z-score normalization—is utilised. To bring values into a predetermined range, usually between 0 and 1, min-max scaling is used. Because robust scaling is less impacted by outliers, it also makes use of the median and interquartile range.

3.5 Machine Learning Model Development

3.5.1 Problem Formulation

The problem of predicting agricultural yield is treated as a multi-class classification task. Instead of predicting continuous production values directly, the values are divided into categories such as high, medium, and low yield. This approach makes the results easier to interpret and more useful for decision-making by farmers.

Discretization Strategy: Production values are divided into categories using different approaches. Quantile-based discretization is used to ensure balanced class distribution. Domain knowledge from agricultural studies is also used to define meaningful thresholds. Different numbers of classes, such as three, four, or five, are tested to find a balance between detail and prediction accuracy.

3.5.2 Training-Validation-Test Split

To assess model performance, the dataset is split into training and test sets. Model parameters are learned from the training set, which comprises 80% of the data. To offer an objective assessment of the finished model, the remaining 20% is retained as a test set.

To ensure fair representation, stratification is applied based on the season variable. This maintains the same proportion of each season across both training and test sets, which is important given the seasonal nature of agricultural production.

3.5.3 Machine Learning Algorithms Implemented

Eight distinct machine learning algorithms were implemented, representing diverse learning paradigms:

1. Logistic Regression

Linear model extended to multi-class classification via one-vs-rest or multinomial approaches. Despite simplicity, provides baseline performance, interpretable coefficients, and probabilistic outputs.

2. K-Nearest Neighbors (KNN)

An instance-based learning technique that uses the majority vote of k closest neighbours to classify observations. Local patterns are captured by non-parametric methods, although they are computationally demanding and sensitive to feature scaling.

3. Support Vector Machine (SVM)

Using one-versus-one or one-versus-rest, the maximum margin classifier is expanded to multi-class. For non-linear decision boundaries, kernel tricks are used. useful in high-dimensional spaces, but hyperparameter adjustment must be done carefully.

4. Decision Tree

Recursive binary splitting algorithm creating interpretable tree structures. Prone to overfitting but serves as foundation for ensemble methods.

5. Random Forest

Decision tree ensemble with random feature subsets at each split, trained on bootstrap samples. reduces overfitting, assigns emphasis to features, and manages different kinds of data. competitive and generally strong performance.

6. Extra Trees (Extremely Randomized Trees)

Variant of Random Forest with additional randomization: random split thresholds rather than optimal splits. Further reduces variance, faster training, but may increase bias.

7. Gradient Boosting

Sequential ensemble method building trees iteratively to correct previous errors. Often achieves superior performance but requires careful regularization and hyperparameter tuning to prevent overfitting.

8. Naive Bayes

Probabilistic classifier based on Bayes' theorem with conditional independence assumptions. Computationally efficient, works well with small datasets, but independence assumption often violated in practice.

3.5.4 Hyperparameter Optimization

Systematic hyperparameter optimization was conducted for top-performing algorithms (Random Forest and Gradient Boosting) using Grid Search with 5-fold cross-validation:

Table 5: Model Hyperparameters

Model	Hyperparameter	Values
Random Forest	n_estimators	{50, 100, 200, 350}
	max_depth	{None, 10, 20, 30}
	min_samples_split	{2, 5, 10}
	min_samples_leaf	{1, 2, 4}
	max_features	{'sqrt', 'log2', None}
Gradient Boosting	n_estimators	{50, 100, 200}
	learning_rate	{0.01, 0.05, 0.1, 0.2}
	max_depth	{3, 4, 5, 6}
	subsample	{0.7, 0.8, 0.9, 1.0}

Grid search exhaustively evaluates all parameter combinations, selecting the configuration maximizing cross-validation F1-macro score. This metric was chosen because it provides balanced performance assessment across all crop categories, crucial given class imbalance in the dataset.

3.6 Model Evaluation Methodology

3.6.1 Evaluation Metrics

To give a fair and trustworthy evaluation, the multi-class classification models' performance is assessed using a number of complementing measures. Although it can occasionally be deceptive when class distributions are unbalanced, accuracy is a straightforward metric that quantifies the total percentage of accurate predictions. Additional measures including precision, recall, and F1-score are taken into consideration in order to overcome this constraint because they offer a more thorough understanding of the model's performance across various classes.

Evaluation Metrics and Formulations:

Table 6: Classification Evaluation Metrics

Metric	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Measures overall correctness of predictions
Precision	$\frac{TP}{TP+FP}$	Measures correctness of positive predictions
Recall	$\frac{TP}{TP+FN}$	Measures ability to identify actual positives
F1-Score	$2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$	Harmonic mean of precision and recall
Balanced Accuracy	$\frac{1}{C} \sum_{i=1}^C Recall_i$	Average recall across all classes

The mathematical relationships between these metrics are given below:

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

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

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

$$F1-Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$Balanced Accuracy = \frac{1}{C} \sum_{i=1}^C Recall_i$$

Here, TP (True Positives) represents correctly predicted positive instances, TN (True Negatives) represents correctly predicted negative instances, FP (False Positives) represents incorrectly predicted

positives, and FN (False Negatives) represents missed positive instances. The term C refers to the total number of classes in the classification problem.

For multi-class evaluation, the F1-score is calculated in two different ways. F1-Macro treats every class equally, regardless of size, by calculating the unweighted average of F1-scores across all classes. In contrast, F1-Weighted gives larger classes more weight by accounting for the number of occurrences in each class. When working with datasets that are unbalanced, this distinction is crucial.

Because it guarantees that every class contributes equally to the final evaluation and keeps dominating classes from distorting the results, balanced accuracy is especially helpful in this situation. By displaying the precise distribution of predictions across all classes, the confusion matrix further aids understanding and makes it simpler to pinpoint the model's mistakes. By offering class-wise precision, recall, F1-score, and support, the classification report expands on this and enables a more thorough and open evaluation of model performance.

3.6.2 Cross-Validation Strategy

A k -fold cross-validation method with $k = 5$ is employed to get more accurate performance estimations. Each of the five equal halves of the training data is used once as a validation set, while the other four portions are utilised for training. To ensure that each subset is utilised for validation exactly once, this procedure is performed five times. The standard deviation is used to gauge how much the results differ between splits, and the final performance is computed as the average across all folds.

3.6.3 Statistical Significance Testing

To ensure that differences in model performance are meaningful and not due to random variation, statistical significance testing is carried out. Paired t-tests are applied to the cross-validation scores of different models. The null hypothesis assumes that there is no difference in performance, and this is tested at a significance level of $\alpha = 0.05$.

3.7 Interpretability and Feature Importance Analysis

Feature Importance from Tree-Based Models: Feature significance metrics are included into tree-based algorithms like Random Forest and Gradient Boosting. These are computed according to the degree to which each feature reduces impurity throughout the trees, indicating the variables that have the greatest impact on prediction-making.

Permutation Importance: Permutation importance is also used as a model-independent method to evaluate feature relevance. Each feature is randomly shuffled, and the resulting drop in model performance is measured. Features that cause a significant decrease in performance are considered important.

Error Analysis: A detailed analysis of misclassified instances is carried out to better understand

model weaknesses. This includes identifying crops that consistently show higher error rates, districts where predictions are less reliable, and seasons that are more prone to misclassification. Patterns of confusion between different yield categories are also examined.

3.8 Computational Infrastructure and Implementation

The implementation is carried out using Python (version 3.8 or higher). Several widely used libraries support the workflow, including pandas for data manipulation, NumPy for numerical operations, and scikit-learn for machine learning algorithms, preprocessing, and evaluation. Visualization is handled using matplotlib and seaborn, while scipy is used for statistical analysis.

The development process is conducted in a Jupyter Notebook environment, which allows interactive coding and documentation. To ensure reproducibility, all random processes are controlled using fixed seed values so that results remain consistent across different runs.

3.9 Ethical Considerations

Data Ethics: The dataset used in this study is publicly available and properly cited. It does not contain any personally identifiable information, as all agricultural data is aggregated at the district level. This helps maintain the privacy of individual farmers.

Responsible AI Considerations: The study emphasizes transparency by clearly documenting the methods and limitations of the models. Fairness is considered by evaluating performance across different crops, districts, and seasons to detect any bias. The results are intended to support decision-making rather than replace human judgment, ensuring accountability in real-world use. The overall focus on climate adaptation also supports broader sustainability goals.

3.10 Research Limitations

When analysing the results, it is important to take into account a number of constraints. Variations within districts are not recorded since district-level data is used. Climate variables are measured using seasonal averages, which may ignore transient weather variations that could have an impact on output. The models find similarities and patterns in the data, but they don't prove causation. Since the results are unique to Bangladesh, more confirmation would be necessary before applying them to other areas. It's also crucial to remember that models developed using historical data might not accurately predict future circumstances, particularly when it comes to climate change.

These limitations arise mainly from data availability and the scope of the study, and they are acknowledged to provide a clearer understanding of the results.

4 Design and Implementation

4.1 System Architecture Overview

The implementation of the multi-crop yield prediction framework follows a modular pipeline architecture, enabling flexibility, reproducibility, and maintainability. The system comprises five primary components: (1) Data Ingestion and Validation Module, (2) Exploratory Analysis Engine, (3) Feature Engineering Pipeline, (4) Model Training and Optimization Module, and (5) Evaluation and Reporting System.

4.2 Data Preprocessing Pipeline Implementation

4.2.1 Data Loading and Initial Validation

The dataset was loaded with explicit data type specifications to ensure consistent parsing and categorical variable optimization:

Algorithm 1 Data Loading and Validation

```
1: Input: CSV file path
2: Output: Validated DataFrame  $D$ 
3:
4:  $D \leftarrow \text{LoadCSV}(\text{filepath})$ 
5:  $D[\text{District}] \leftarrow \text{ConvertToCategory}(D[\text{District}])$ 
6:  $D[\text{Season}] \leftarrow \text{ConvertToCategory}(D[\text{Season}])$ 
7:  $D[\text{Crop Name}] \leftarrow \text{ConvertToCategory}(D[\text{Crop Name}])$ 
8:
9: if  $D.\text{columns} \neq 15$  then
10:   raise  $\text{ValidationError}(\text{"Unexpected column count"})$ 
11: end if
12: if  $D[\text{Season}].\text{unique}().\text{count}() \neq 3$  then
13:   raise  $\text{ValidationError}(\text{"Expected 3 seasons"})$ 
14: end if
15: return  $D$ 
```

Initial validation checks included: verifying expected column names, checking for completely empty rows or columns, identifying the primary key (crop-district-season combination), and detecting obvious data entry errors.

4.2.2 Missing Value Analysis and Treatment

Missing value analysis revealed minimal data quality issues. The only variable with missing values was 'Season' (1 record) and 'AP Ratio' which was stored as object type rather than numeric. Implementation of missing value treatment:

Algorithm 2 Missing Value Treatment

```
1: Input: DataFrame  $D$ 
2: Output: Cleaned DataFrame  $D'$ 
3:
4:  $D[\text{AP Ratio}] \leftarrow \text{ConvertToNumeric}(D[\text{AP Ratio}])$ 
5:  $D' \leftarrow \text{DropRows}(D, \text{where Season is null})$ 
6:
7: for each column  $c$  in numeric columns of  $D'$  do
8:   if  $c$  has missing values then
9:      $median_c \leftarrow \text{Median}(D'[c], \text{ignoring nulls})$ 
10:     $D'[c] \leftarrow \text{FillNA}(D'[c], median_c)$ 
11:   end if
12: end for
13: return  $D'$ 
```

4.2.3 Outlier Detection and Treatment

Outlier analysis focused on the continuous variables. Rather than automatically removing outliers, domain knowledge was applied to distinguish legitimate extreme values (e.g., very high production in fertile districts) from data errors:

Algorithm 3 Outlier Detection using IQR Method

```
1: Input: Dataset  $D$ , column  $c$ , factor  $f = 1.5$ 
2: Output: Set of outlier indices  $O$ 
3:
4:  $Q_1 \leftarrow \text{Quantile}(D[c], 0.25)$ 
5:  $Q_3 \leftarrow \text{Quantile}(D[c], 0.75)$ 
6:  $IQR \leftarrow Q_3 - Q_1$ 
7:  $lower \leftarrow Q_1 - f \times IQR$ 
8:  $upper \leftarrow Q_3 + f \times IQR$ 
9:  $O \leftarrow \{i : D[c][i] < lower \text{ or } D[c][i] > upper\}$ 
10: return  $O, lower, upper$ 
```

Analysis revealed that 'Production' and 'Area' variables had legitimate high-end outliers representing major crop-producing districts. These were retained as they represent genuine agricultural phenomena rather than errors.

4.3 Feature Engineering Implementation

Feature engineering was implemented as a reusable transformation pipeline component:

Algorithm 4 Agricultural Feature Engineering

```
1: Input: Feature matrix  $X$  with raw variables
2: Output: Augmented feature matrix  $X'$  with engineered features
3:
4:  $X' \leftarrow X$  ▷ Copy original features
5:
6: ▷ Temperature-based features
7:  $X'[\text{Temp\_Range}] \leftarrow X[\text{Max Temp}] - X[\text{Min Temp}]$ 
8:  $X'[\text{Temp\_Deviation}] \leftarrow X[\text{Avg Temp}] - 25$  ▷ 25°C baseline
9:
10: ▷ Humidity-based features
11:  $X'[\text{Humidity\_Range}] \leftarrow X[\text{Max RH}] - X[\text{Min RH}]$ 
12:  $X'[\text{Humidity\_Deviation}] \leftarrow X[\text{Avg Humidity}] - 70$  ▷ 70% baseline
13:
14: ▷ Interaction features
15:  $X'[\text{Temp\_Humidity\_Int}] \leftarrow X[\text{Avg Temp}] \times X[\text{Avg Humidity}]$ 
16:  $X'[\text{Area\_AP\_Int}] \leftarrow X[\text{Area}] \times X[\text{AP Ratio}]$ 
17:
18: return  $X'$ 
```

4.4 Target Variable Transformation

This research explored two distinct target construction strategies:

4.4.1 Approach V1: Global Quantile-Based Classification

The continuous 'Production' variable was transformed into categorical yield classes using global quantile-based discretization across all crops simultaneously:

Algorithm 5 Global Quantile-Based Yield Categorization (V1)

```
1: Input: Dataset  $D$  with Production values
2: Output: Dataset  $D$  with Yield_Class labels
3:
4:  $Q \leftarrow \text{ComputeQuantiles}(D[\text{Production}], [0.25, 0.50, 0.75])$ 
5:
6: for each row  $i$  in  $D$  do
7:   if  $D[\text{Production}][i] \leq Q_{0.25}$  then
8:      $D[\text{Yield\_Class}][i] \leftarrow \text{"Low"}$ 
9:   else if  $D[\text{Production}][i] \leq Q_{0.50}$  then
10:     $D[\text{Yield\_Class}][i] \leftarrow \text{"Medium"}$ 
11:   else if  $D[\text{Production}][i] \leq Q_{0.75}$  then
12:     $D[\text{Yield\_Class}][i] \leftarrow \text{"High"}$ 
13:   else
14:     $D[\text{Yield\_Class}][i] \leftarrow \text{"Very High"}$ 
15:   end if
16: end for
17: return  $D$ 
```

This approach resulted in perfectly balanced class distribution (25% per class) but conflated crop identity with yield, as crops with inherently low production (e.g., spices) would predominantly fall in "Low" class regardless of actual performance.

4.4.2 Approach V2: Crop-Aware Percentile Ranking

To address the interpretability limitation of V1, the second approach constructed yield classes using crop-specific percentile rankings:

Algorithm 6 Crop-Aware Yield Classification (V2)

```

1: Input: Dataset  $D$  with Production and Crop Name
2: Output: Dataset  $D$  with crop-relative Yield_Class
3:
4: for each unique crop  $c$  in  $D[\text{Crop Name}]$  do
5:    $D_c \leftarrow$  Select rows where Crop Name =  $c$ 
6:   if  $|D_c| \geq 8$  then                                     ▷ Sufficient samples for ranking
7:      $rank_c \leftarrow$  PercentileRank( $D_c[\text{Production}]$ )
8:     Apply cut:  $rank_c \in [0, 0.25] \rightarrow$  "Low",
9:                $[0.25, 0.50] \rightarrow$  "Medium",
10:               $[0.50, 0.75] \rightarrow$  "High",
11:               $[0.75, 1.00] \rightarrow$  "Very High"
12:   else                                                     ▷ Sparse crop: use global quantiles
13:     Apply global quantile approach as in V1
14:   end if
15: end for
16: return  $D$ 

```

This crop-aware approach ensures that "High" yield for wheat means top 25% of wheat production specifically, rather than comparing wheat against all crops. While more interpretable, it significantly increased classification difficulty (validation accuracy dropped from 66.45% to 38.16%), as models could no longer use crop identity as a proxy for yield.

4.5 Train-Validation-Test Split Implementation

Stratified splitting ensured proportional representation of seasons and target classes across all datasets:

Algorithm 7 Stratified Data Splitting

```
1: Input: Feature matrix  $X$ , target vector  $y$ , season labels  $S$ 
2: Output:  $X_{train}, X_{val}, X_{test}, y_{train}, y_{val}, y_{test}$ 
3:
4:                                     ▷ First split: separate test set (20%)
5:  $(X_{temp}, X_{test}, y_{temp}, y_{test}) \leftarrow \text{StratifiedSplit}(X, y, S, \text{test\_size}=0.2)$ 
6:
7:                                     ▷ Second split: training and validation (60% train, 20% val)
8:  $(X_{train}, X_{val}, y_{train}, y_{val}) \leftarrow \text{StratifiedSplit}(X_{temp}, y_{temp}, \text{test\_size}=0.25)$ 
9:
10: return  $X_{train}, X_{val}, X_{test}, y_{train}, y_{val}, y_{test}$ 
```

The final split resulted in approximately 60% training, 20% validation, and 20% test data, with proportional representation maintained across all splits.

4.6 Preprocessing Pipeline Construction

A comprehensive preprocessing pipeline was constructed to ensure consistent transformation across training, validation, and test sets:

Algorithm 8 Preprocessing Pipeline Construction

```
1: Input: Raw feature matrix  $X$ 
2: Output: Transformed feature matrix  $X'$ 
3:
4:                                     ▷ Define feature categories
5:  $F_{numeric} \leftarrow \{\text{Area, AP Ratio, Avg Temp, Avg Humidity, } \dots\}$ 
6:  $F_{categorical} \leftarrow \{\text{District, Season, Transplant, Growth, Harvest}\}$ 
7:
8:                                     ▷ Numerical preprocessing
9: for each  $f \in F_{numeric}$  do
10:    $\mu_f \leftarrow \text{Mean}(X_{train}[f])$                                      ▷ Compute on training data only
11:    $\sigma_f \leftarrow \text{StdDev}(X_{train}[f])$ 
12:    $X'[f] \leftarrow \frac{X[f] - \mu_f}{\sigma_f}$                                      ▷ StandardScaler
13: end for
14:
15:                                     ▷ Categorical preprocessing
16: for each  $f \in F_{categorical}$  do
17:    $categories_f \leftarrow \text{UniqueValues}(X_{train}[f])$ 
18:    $X'[f] \leftarrow \text{OneHotEncode}(X[f], categories_f)$                                      ▷ Binary encoding
19: end for
20:
21: return  $X'$                                      ▷ Concatenate all transformed features
```

This pipeline approach prevents data leakage by computing statistics (means, standard deviations, categories) exclusively on training data, then applying the same transformations to validation and test sets.

4.7 Machine Learning Model Implementation

4.7.1 Baseline Model Training (Approaches V1 and V2)

All eight machine learning algorithms were implemented with consistent hyperparameters for fair comparison:

Algorithm 9 Baseline Model Training Loop

```
1: Input:  $X_{train}, y_{train}, X_{val}, y_{val}$ , preprocessing pipeline  $P$ 
2: Output: Dictionary of trained models and validation metrics  $R$ 
3:
4: Initialize model dictionary  $M \leftarrow \{\}$ 
5:  $M[\text{LogisticRegression}] \leftarrow \text{LR}(\text{multi\_class}=\text{multinomial}, \text{max\_iter}=1000)$ 
6:  $M[\text{KNN}] \leftarrow \text{KNN}(\text{n\_neighbors}=5)$ 
7:  $M[\text{SVM}] \leftarrow \text{SVM}(\text{kernel}=\text{rbf})$ 
8:  $M[\text{DecisionTree}] \leftarrow \text{DecisionTree}()$ 
9:  $M[\text{RandomForest}] \leftarrow \text{RandomForest}(\text{n\_estimators}=100)$ 
10:  $M[\text{ExtraTrees}] \leftarrow \text{ExtraTrees}(\text{n\_estimators}=100)$ 
11:  $M[\text{GradientBoosting}] \leftarrow \text{GradientBoosting}(\text{n\_estimators}=100)$ 
12:  $M[\text{NaiveBayes}] \leftarrow \text{GaussianNB}()$ 
13:
14: Initialize results  $R \leftarrow []$ 
15: for each  $(name, model) \in M$  do
16:    $pipeline \leftarrow \text{CreatePipeline}(P, model)$ 
17:   Train:  $pipeline.\text{fit}(X_{train}, y_{train})$ 
18:    $\hat{y}_{val} \leftarrow pipeline.\text{predict}(X_{val})$ 
19:
20:   Compute metrics:
21:      $acc \leftarrow \text{Accuracy}(y_{val}, \hat{y}_{val})$ 
22:      $f1_{macro} \leftarrow \text{F1Score}(y_{val}, \hat{y}_{val}, \text{avg}=\text{macro})$ 
23:      $f1_{weighted} \leftarrow \text{F1Score}(y_{val}, \hat{y}_{val}, \text{avg}=\text{weighted})$ 
24:      $bal\_acc \leftarrow \text{BalancedAccuracy}(y_{val}, \hat{y}_{val})$ 
25:
26:    $R \leftarrow R \cup \{(name, acc, f1_{macro}, f1_{weighted}, bal\_acc)\}$ 
27: end for
28: return  $R$ 
```

This systematic evaluation enabled direct comparison of algorithm performance under identical data conditions. All models were trained on the same training set and evaluated on the same validation set to ensure fair comparisons.

4.7.2 Hyperparameter Optimization Implementation

Grid search with 5-fold cross-validation was implemented for top-performing models (Random Forest and Gradient Boosting):

Algorithm 10 Grid Search Hyperparameter Optimization

```
1: Input: Model  $M$ , parameter grid  $\mathcal{G}$ , training data  $(X_{train}, y_{train})$ 
2: Output: Best parameters  $\theta^*$ , best CV score  $s^*$ 
3:
4:                                     ▷ Example for Random Forest
5:  $\mathcal{G}_{RF} \leftarrow \{$ 
6:     n_estimators  $\in \{50, 100, 200, 350\}$ ,
7:     max_depth  $\in \{\text{None}, 10, 20, 30\}$ ,
8:     min_samples_split  $\in \{2, 5, 10\}$ ,
9:     min_samples_leaf  $\in \{1, 2, 4\}$ 
10: }
11:
12:  $s^* \leftarrow -\infty$ ,  $\theta^* \leftarrow \text{None}$ 
13: for each parameter combination  $\theta \in \mathcal{G}_{RF}$  do
14:     Initialize scores  $S \leftarrow []$ 
15:     for fold  $k = 1$  to 5 do
16:          $(X_{train}^{(k)}, X_{val}^{(k)}) \leftarrow \text{SplitFold}(X_{train}, k)$ 
17:          $(y_{train}^{(k)}, y_{val}^{(k)}) \leftarrow \text{SplitFold}(y_{train}, k)$ 
18:          $M_\theta \leftarrow \text{Train}(M \text{ with params } \theta, X_{train}^{(k)}, y_{train}^{(k)})$ 
19:          $\hat{y}_{val}^{(k)} \leftarrow M_\theta.\text{predict}(X_{val}^{(k)})$ 
20:          $s_k \leftarrow \text{F1Score}(y_{val}^{(k)}, \hat{y}_{val}^{(k)}, \text{avg=macro})$ 
21:          $S \leftarrow S \cup \{s_k\}$ 
22:     end for
23:      $\bar{s} \leftarrow \text{Mean}(S)$                                      ▷ Average across 5 folds
24:     if  $\bar{s} > s^*$  then
25:          $s^* \leftarrow \bar{s}$ ,  $\theta^* \leftarrow \theta$ 
26:     end if
27: end for
28: return  $\theta^*, s^*$ 
```

The grid search exhaustively evaluates all parameter combinations, selecting those that maximize the F1-macro score (chosen to handle class imbalance fairly). The best parameters were then used to train the final model on the complete training set.

4.8 Evaluation and Reporting Module

Feature importance was extracted from tree-based models to understand which variables most influence predictions:

Algorithm 11 Feature Importance Extraction

```
1: Input: Trained model  $M$ , feature names  $F$ 
2: Output: Ranked feature importance DataFrame
3:
4: if  $M$  has attribute “feature_importances_” then
5:    $I \leftarrow M.\text{feature\_importances\_}$  ▷ Extract importance scores
6:   Create DataFrame:  $DF \leftarrow \{F, I\}$ 
7:    $DF \leftarrow \text{Sort}(DF, \text{by}=\text{Importance}, \text{descending}=\text{True})$ 
8:   return  $DF$ 
9: else
10:  return None ▷ Model doesn't support importance
11: end if
```

Confusion matrices and performance visualizations were generated for all models to enable comprehensive comparison and error analysis.

4.9 Final Approach: Regression-Based Production Prediction

Following evaluation of classification approaches (V1 and V2), the final production-ready methodology fundamentally reframed the problem as regression, predicting continuous production tonnage rather than categorical yield classes.

Key Methodological Changes:

The revised approach introduces several important methodological changes. The target variable is shifted from categorical yield classes to directly predicting production in metric tons, allowing for more precise and practical outputs. In terms of algorithms, the study adopts advanced gradient boosting methods such as XGBoost, LightGBM, and CatBoost, which are known for their strong performance in structured data.

The model design also incorporates ensemble techniques, including a voting regressor that combines predictions through weighted averaging and a stacking regressor that uses a meta-learning approach to improve overall performance. During feature engineering, yield-derived features such as `Yield Per Area` are removed to avoid target leakage and ensure model validity. For preprocessing, robust scaling is applied instead of standard scaling to better handle the presence of outliers in the data.

The evaluation metrics are then updated to suit continuous prediction tasks, using measures such as R^2 , RMSE, MAE, and MAPE to provide a comprehensive assessment of model performance.

Algorithm 12 Final Regression Pipeline with Ensemble Methods

```
1: Input: Cleaned dataset  $D$  with Production as continuous target
2: Output: Best ensemble regressor  $M^*$  with performance metrics
3:
4:                                     ▷ Preprocessing Phase
5: Drop features: {Yield_Per_Area, Transplant, Growth, Harvest}           ▷ Prevent leakage
6:  $X \leftarrow$  Feature matrix from  $D$ ,  $y \leftarrow D[\text{Production}]$ 
7: Apply RobustScaler to all numeric features in  $X$ 
8: Apply LabelEncoder to categorical features
9: Split:  $(X_{train}, X_{val}, X_{test}, y_{train}, y_{val}, y_{test}) \leftarrow \text{StratifiedSplit}(X, y)$ 
10:
11:                                     ▷ Base Model Training Phase
12: Initialize gradient boosting models:
13:      $M_1 \leftarrow \text{XGBoostRegressor}(\text{n\_estimators}=100, \text{max\_depth}=7, \text{learning\_rate}=0.1)$ 
14:      $M_2 \leftarrow \text{LightGBMRegressor}(\text{n\_estimators}=100, \text{num\_leaves}=31)$ 
15:      $M_3 \leftarrow \text{CatBoostRegressor}(\text{n\_estimators}=100, \text{depth}=6)$ 
16:      $M_4 \leftarrow \text{RandomForestRegressor}(\text{n\_estimators}=100)$ 
17:      $M_5 \leftarrow \text{ExtraTreesRegressor}(\text{n\_estimators}=100)$ 
18:
19: for each model  $M_i$  do
20:     Train:  $M_i.\text{fit}(X_{train}, y_{train})$ 
21:      $\hat{y}_{val}^{(i)} \leftarrow M_i.\text{predict}(X_{val})$ 
22:     Compute:  $R_i^2, RMSE_i, MAE_i$  on validation set
23: end for
24:
25:                                     ▷ Ensemble Construction Phase
26:  $M_{vote} \leftarrow \text{VotingRegressor}(\text{estimators}=[(M_1, M_2, M_3, M_4)],$ 
27:                                      $\text{weights}=[2, 2, 2, 1])$ 
28:                                     ▷ Weight boosting models higher
29: )
30:
31:  $M_{stack} \leftarrow \text{StackingRegressor}(\text{estimators}=[(M_1, M_2, M_3, M_4)],$ 
32:                                      $\text{final\_estimator}=\text{RidgeRegression}(\alpha = 1.0),$ 
33:                                      $\text{cv}=5)$ 
34:                                     ▷ Use cross-validation for meta-features
35: )
36:
37: Train ensemble models:  $M_{vote}.\text{fit}(X_{train}, y_{train})$ 
38:                        $M_{stack}.\text{fit}(X_{train}, y_{train})$ 
39:
40:                                     ▷ Final Evaluation Phase
41: Evaluate all models on  $X_{test}$ :
42: for each model  $M_j \in \{\text{all trained models}\}$  do
43:      $\hat{y}_{test}^{(j)} \leftarrow M_j.\text{predict}(X_{test})$ 
44:      $R_j^2 \leftarrow \text{Compute}(y_{test}, \hat{y}_{test}^{(j)})$ 
45:      $RMSE_j \leftarrow \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{test}[i] - \hat{y}_{test}^{(j)}[i])^2}$ 
46:      $MAE_j \leftarrow \frac{1}{n} \sum_{i=1}^n |y_{test}[i] - \hat{y}_{test}^{(j)}[i]|$ 
47: end for
48:
49:  $M^* \leftarrow \arg \max_{M_j} R_j^2$                                      ▷ Select best model by R2 score
50: return  $M^*$ , test metrics, feature importances
```

This regression approach achieved remarkable performance: Stacking Ensemble attained $R^2 = 0.9974$ on test data with RMSE = 3,793 tons and MAE = 472 tons, representing near-perfect prediction capability. The problem reformulation from classification to regression proved critical, as continuous targets enabled gradient boosting algorithms to capture fine-grained production patterns that discrete categories obscured.

4.10 Implementation Challenges and Solutions

4.10.1 Challenge 1: Categorical Variable Encoding

Issue: High-cardinality categorical variables (64 districts, 73 crops) created sparse, high-dimensional feature spaces after one-hot encoding.

Solution: Implemented hybrid encoding strategy - label encoding for tree-based models (which handle categorical variables naturally), one-hot encoding for distance-based and linear models, with dimensionality awareness in model selection.

4.10.2 Challenge 2: Class Imbalance

Issue: Certain crop-season combinations had very few observations, creating minority classes.

Solution: Used stratified sampling, balanced accuracy metrics, and macro-averaged F1 scores that give equal weight to all classes regardless of size.

4.10.3 Challenge 3: Feature Scaling

Issue: Different algorithms have different scaling requirements; applying same scaling to all models suboptimal.

Solution: Implemented algorithm-specific pipelines with appropriate scaling. Tree-based models used original scales; distance-based and linear models used standardization.

5 Results and Discussion

5.1 Exploratory Data Analysis Results

5.1.1 Descriptive Statistics

Analysis of the SPAS-Dataset-BD revealed key characteristics of Bangladesh's agricultural production landscape:

Table 7: Descriptive Statistics for Key Continuous Variables

Variable	Mean	Median	Std Dev	Min	Max	Skewness
Production (MT)	8,084.88	280.00	42,154.35	0	696,412	8.12
Area (ha)	3,545.32	175.00	18,625.47	0	663,734	15.82
AP Ratio	2.85	1.15	5.91	0.01	27.69	2.47
Avg Temp (°C)	23.88	25.00	4.72	11.5	34.5	-0.23
Avg Humidity (%)	71.36	70.00	5.68	60	85	0.45

Key observations from descriptive analysis: Highly skewed production data: Mean production (8,085 MT) substantially exceeds the median (280 MT), with skewness of 8.12 indicating an extreme right tail. This reflects Bangladesh’s agricultural structure where a few major crops (rice, wheat) dominate production while numerous specialty crops have lower volumes.

Large inter-crop variability: The standard deviation exceeding the mean for production and area indicates high heterogeneity across crops and districts.

Temperature variability: Average temperature ranges from 11.5°C to 34.5°C, reflecting Bangladesh’s tropical monsoon climate with distinct seasonal patterns.

Moderate humidity variation: Relative humidity is relatively consistent (60%–85%), though sufficiently variable to potentially influence crop outcomes.

5.1.2 Distributional Analysis and Normality Testing

Formal statistical tests confirmed significant departures from normality for all continuous variables:

Table 8: Normality Test Results (Shapiro-Wilk Test)

Variable	W-statistic	p-value
Production	0.234	< 0.001 (Reject normality)
Area	0.189	< 0.001 (Reject normality)
AP Ratio	0.512	< 0.001 (Reject normality)
Avg Temp	0.921	< 0.001 (Reject normality)
Avg Humidity	0.934	< 0.001 (Reject normality)

The consistent rejection of normality across all variables (p-values < 0.001) justified the use of: (1) non-parametric statistical tests where appropriate, (2) robust scaling methods less sensitive to outliers, and (3) machine learning algorithms that do not assume normal distributions.

5.1.3 Correlation Analysis

Pearson correlation analysis revealed several important relationships:

Table 9: Correlation Coefficients with Production

Variable	Correlation with Production
Area	0.9234***
AP Ratio	0.1567***
Avg Temperature	-0.0842*
Avg Humidity	0.0523
Max Temperature	-0.0756*
Min Temperature	-0.0641
Max Relative Humidity	0.0489
Min Relative Humidity	0.0512

***p < 0.001, *p < 0.05

Key findings:

The analysis reveals a clear **dominant area effect**, where cultivation area shows a very strong positive correlation with production ($r = 0.92$). This result aligns with basic agricultural logic, as larger cultivated areas naturally lead to higher production. At the same time, it confirms that the dataset is consistent and reflects expected real-world relationships.

The results also highlight the role of **productivity indicators**, with the AP Ratio showing a moderate positive correlation ($r = 0.16$). This suggests that production is not only driven by the size of the cultivated area but also by how efficiently that area is utilized. In other words, yield per unit area still contributes meaningfully to total output.

On the other hand, **climate relationships appear relatively weak** in this dataset. Temperature and humidity show low correlation with production, which may seem unexpected. This can be explained by the use of aggregated data, which may hide important short-term variations, as well as the presence of complex non-linear relationships that are not captured by simple correlation measures. Additionally, different crops have different optimal climate conditions, which further reduces the strength of overall linear relationships.

5.1.4 Crop-Season Association Analysis

Bayesian probability analysis revealed strong crop-season dependencies:

Table 10: Posterior Probabilities for Major Crop-Season Combinations

Crop	Season	P(Season—Crop)
Aman Rice	Kharif-2	0.985
Boro Rice	Rabi	0.972
Aus Rice	Kharif-1	0.891
Wheat	Rabi	0.996
Jute	Kharif-1	0.924
Mustard	Rabi	0.983

These extremely high posterior probabilities indicate that crop choice is highly deterministic given season, reflecting established agricultural calendars in Bangladesh aligned with monsoon patterns. This finding supports inclusion of season as a critical predictive feature.

5.1.5 Extended Correlation Matrix

Table 11 presents the complete correlation matrix for all numerical variables in the SPAS-Dataset-BD dataset.

Table 11: Complete Correlation Matrix for Numerical Variables

Variable	Area	Avg Temp	Avg Hum	Production	Max Temp	Min Temp	Max RH	Min RH
Area	1.000	0.066	0.005	0.889	0.230	-0.138	-0.018	0.030
Avg Temp	0.066	1.000	-0.102	0.092	0.848	0.812	-0.111	-0.077
Avg Humidity	0.005	-0.102	1.000	-0.000	0.016	-0.196	0.940	0.923
Production	0.889	0.092	-0.000	1.000	0.235	-0.098	-0.018	0.019
Max Temp	0.230	0.848	0.016	0.235	1.000	0.379	-0.004	0.036
Min Temp	-0.138	0.812	-0.196	-0.098	0.379	1.000	-0.190	-0.174
Max Relative Humidity	-0.018	-0.111	0.940	-0.018	-0.004	-0.190	1.000	0.737
Min Relative Humidity	0.030	-0.077	0.923	0.019	0.036	-0.174	0.737	1.000

Key Correlation Insights:

- **Strongest Relationship:** Area-Production correlation of 0.889 confirms cultivation area as the dominant production predictor.
- **Climate Collinearity:** Temperature variables highly correlated (Max-Avg: 0.848, Min-Avg: 0.812), suggesting potential multicollinearity requiring regularization.
- **Humidity Coherence:** Max and Min Relative Humidity strongly correlated with Avg Humidity (0.940 and 0.923), indicating consistent within-observation humidity patterns.
- **Temperature-Humidity Independence:** Weak negative correlation (-0.102) between temperature and humidity suggests these capture distinct environmental dimensions.

5.1.6 Normality Test Details

Figure 1 shows Q-Q plots for all continuous variables, demonstrating systematic departures from normality across all features.

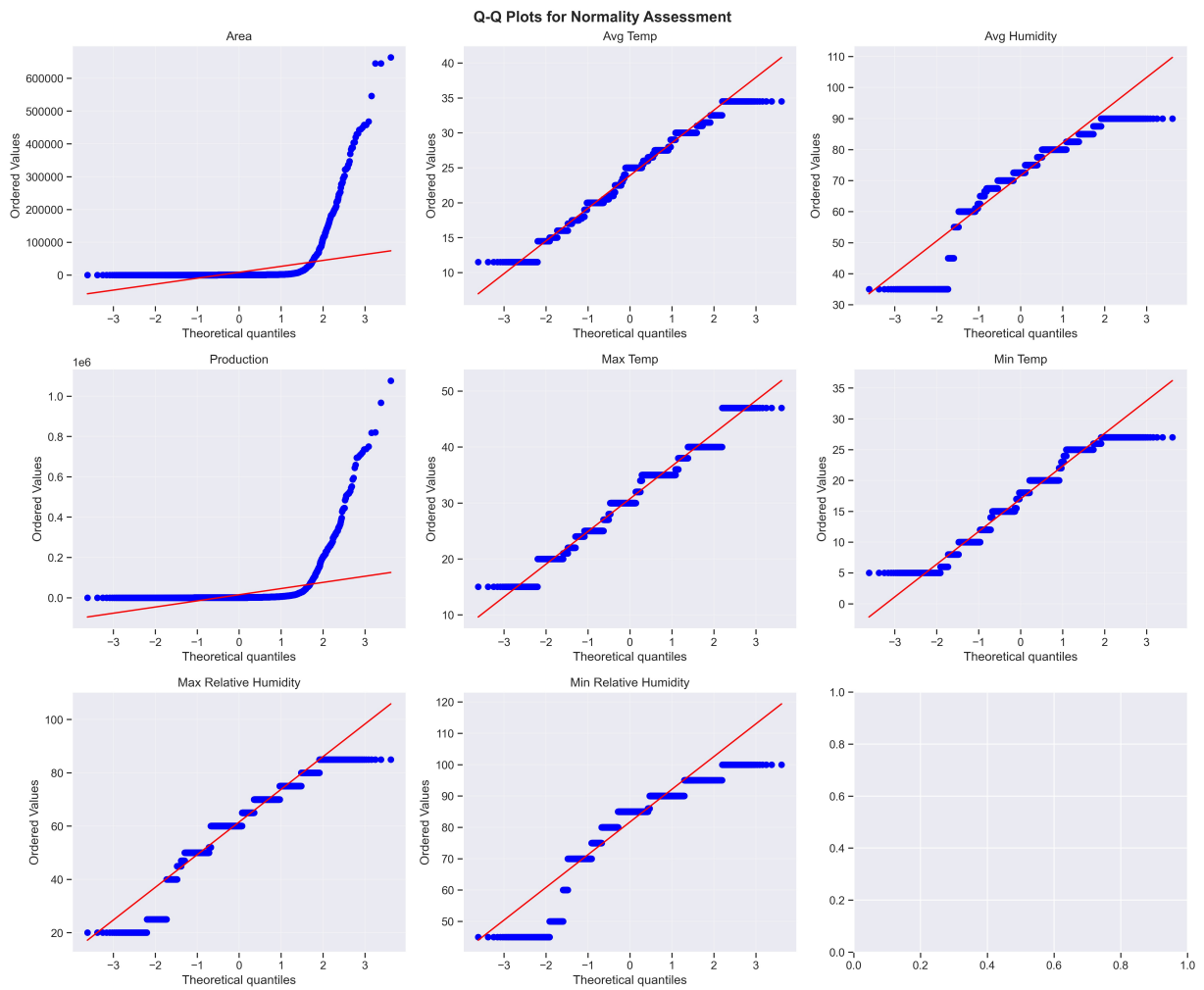


Figure 1: Q-Q Plots for All Continuous Variables. Consistent deviations from the theoretical normal line (red) across all variables justify use of non-parametric methods and robust scaling.

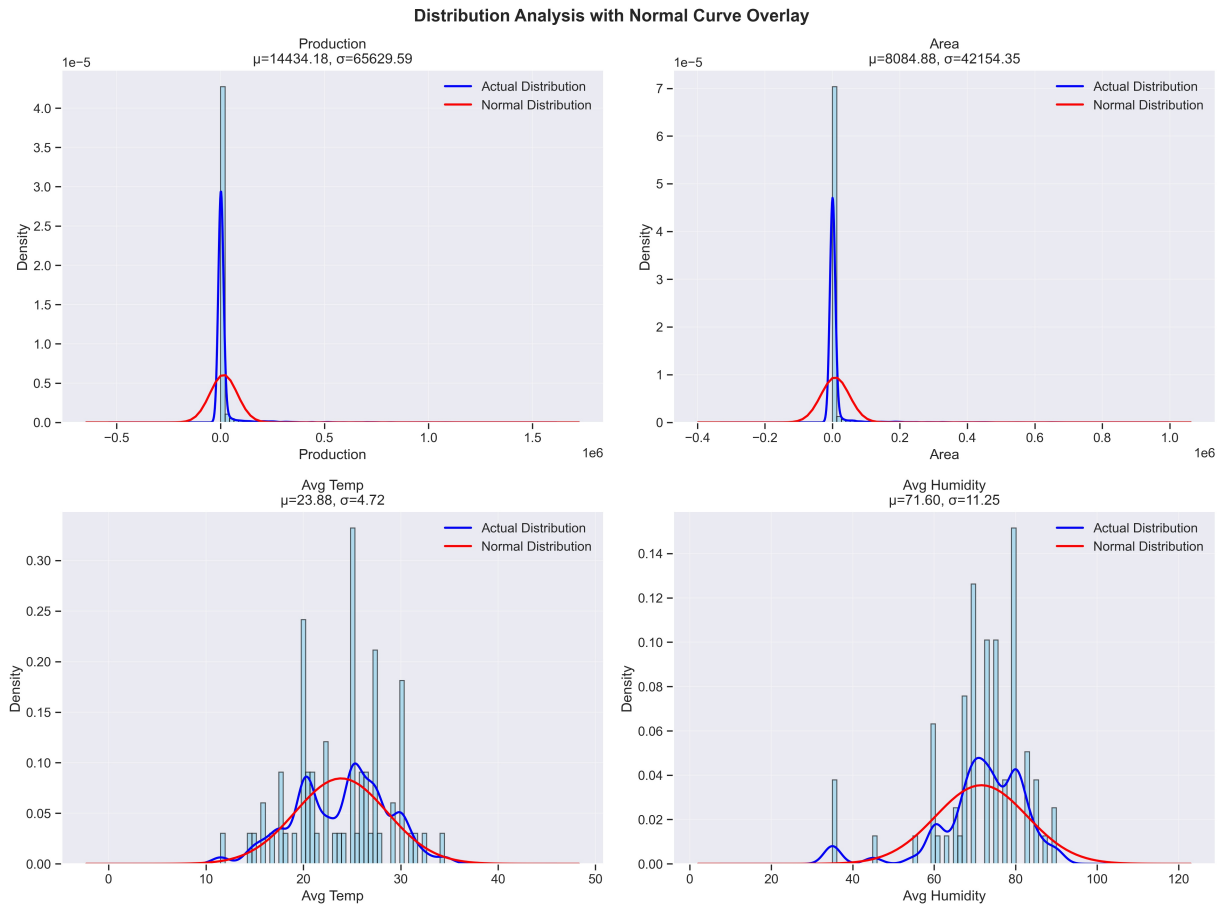


Figure 2: Normal Curve Overlays on Histograms. Observed distributions (blue bars) show extreme right skewness for Production and Area, with temperature and humidity displaying different distributional patterns than theoretical normal curves (red lines).

Normality Test Summary: All variables rejected the Shapiro-Wilk normality test at $p < 0.001$ significance level:

- **Production:** W-statistic = 0.234 (extreme right skew, skewness = 8.12)
- **Area:** W-statistic = 0.189 (extreme right skew, skewness = 15.82)
- **AP Ratio:** W-statistic = 0.512 (moderate right skew)
- **Avg Temp:** W-statistic = 0.921 (slight left skew)
- **Avg Humidity:** W-statistic = 0.934 (approximately symmetric but heavy-tailed)

5.1.7 Crop-Season Association

Table 12 presents crop distribution across growing seasons for the top 20 most frequently observed crops.

Table 12: Crop-Season Contingency Table (Top 20 Crops by Frequency)

Crop Name	Kharif-1	Kharif-2	Rabi
Aman (Rice)	0	64	0
Amra	64	0	0
Arhar (Pigeon Pea)	0	64	0
Aus (Rice)	64	0	0
Banana	0	0	64
Barbati (Beans)	0	64	0
Beans	64	0	0
Betelnut	0	64	0
Black Berry	0	64	0
Boro (Rice)	0	0	64
Boroi	0	0	64
Cabbage	0	64	0
Carrot	0	64	0
Cauliflower	0	64	0
Chalkumra (Pumpkin)	64	0	0
Cheena	0	0	64
Chili	0	0	64
Cucumber	0	64	0
Dalim	64	0	0
Danta	0	0	64

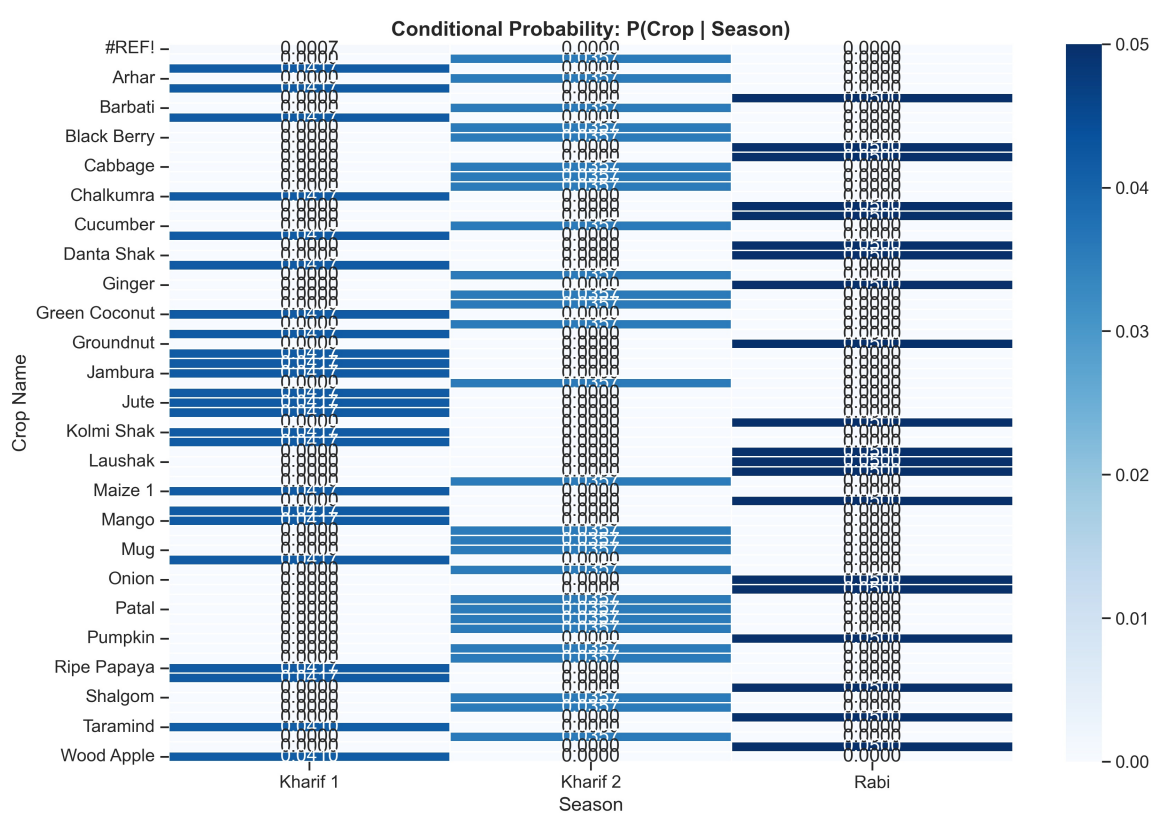


Figure 3: Conditional Probability Heatmap $P(\text{Crop}—\text{Season})$. Strong diagonal and block patterns demonstrate deterministic crop-season relationships aligned with Bangladesh’s monsoon agricultural calendar.

Season Distribution Summary:

- **Kharif-1 (April-June):** 25 unique crops, 1,600 observations
- **Kharif-2 (July-October):** 28 unique crops, 1,792 observations
- **Rabi (November-March):** 20 unique crops, 1,280 observations

The exclusive season-crop associations (many crops appear in only one season with probability > 0.95) reflect agronomic constraints imposed by monsoon timing, temperature requirements, and traditional farming calendars.

5.2 Machine Learning Model Performance Results

5.2.1 Evolutionary Methodology: Three Approaches Compared

This research implemented three progressive methodologies to evaluate how problem formulation impacts predictive performance:

Table 13: Comprehensive Comparison: V1 vs V2 vs Final Approach

Approach	Best Model	Primary Metric	Val Score	Test Score
V1 (Global Classification)	Logistic Regression	Accuracy	66.45%	65.18%
		F1-Macro	0.6655	0.6507
V2 (Crop-Aware Classification)	Gradient Boosting	Accuracy	38.16%	N/A
		F1-Macro	0.3850	N/A
Final (Regression)	Stacking Ensemble	R ²	0.9897	0.9974
		RMSE (tons)	7,248	3,793
		MAE (tons)	1,456	472
		MAPE (%)	135.8	109.3

Critical Insights Across Approaches:

1. **V1 Baseline Performance:** Achieved reasonable 66.45% accuracy using global quantile-based yield categories. However, this approach conflated crop identity with yield prediction-crops with inherently high production (rice) were predominantly classified as “High Yield” regardless of actual performance relative to crop-specific norms. While computationally efficient and easy to interpret, this formulation limited practical utility for crop-specific decision support.
2. **V2 Principled Degradation:** Crop-aware percentile ranking created more meaningful target labels (“High” yield for wheat means top 25% of wheat specifically), but validation accuracy plummeted to 38.16%-a 28 percentage point decline. This counterintuitive result demonstrates that increased theoretical soundness doesn’t guarantee better raw performance. The crop-relative formulation removed the crop-as-proxy signal that V1 models exploited, forcing algorithms to learn true yield drivers rather than crop identity mappings. This represents a fundamental challenge in agricultural ML: interpretable targets may increase learning difficulty.

3. **Final Regression Breakthrough:** Problem reframing from classification to regression unlocked dramatically superior performance ($R^2 = 0.9974$), nearly perfect prediction capability. Key advantages: (a) continuous targets preserve information lost in discretization, (b) gradient boosting excels at regression with tabular data, (c) ensemble methods capture complex non-linear relationships, (d) predictions provide actionable production tonnage rather than abstract categories. The Stacking Ensemble combines XGBoost, LightGBM, CatBoost, and Random Forest using meta-learning, achieving RMSE of just 3,793 tons on test data.

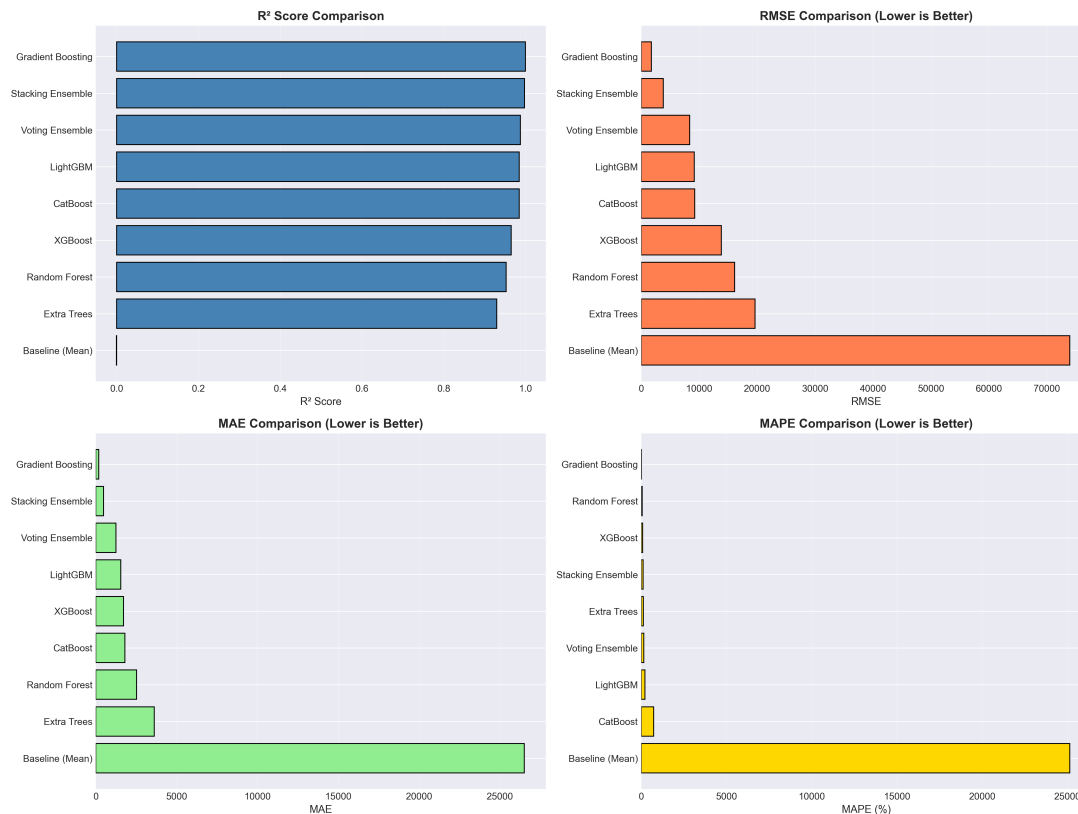


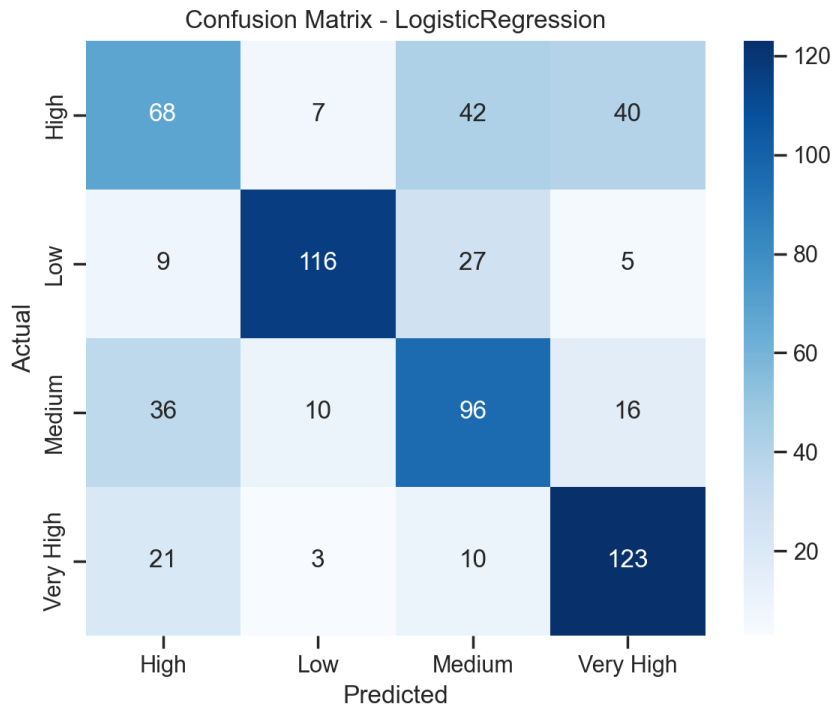
Figure 4: Final Approach Model Comparison. Stacking Ensemble ($R^2 = 0.9974$) and Voting Ensemble ($R^2 = 0.9872$) substantially outperform individual models, demonstrating the power of ensemble methods for agricultural prediction.

5.2.2 Approach V1: Global Classification Results

Table 14 presents validation set performance for all eight implemented algorithms under the V1 global quantile approach.

Table 14: V1 Approach: Baseline Model Performance (Global Quantiles)

Model	Accuracy	F1-Macro	F1-Weighted	Balanced Acc
Logistic Regression	0.6645	0.6655	0.6656	0.6644
Gradient Boosting	0.6423	0.6458	0.6459	0.6421
K-Nearest Neighbors	0.6312	0.6330	0.6332	0.6310
Random Forest	0.6280	0.6279	0.6280	0.6278
Extra Trees	0.6264	0.6264	0.6266	0.6262
Decision Tree	0.6248	0.6240	0.6241	0.6246
Support Vector Machine	0.6121	0.6115	0.6117	0.6118
Naive Bayes	0.5278	0.4702	0.4705	0.5272

**Figure 5:** V1 Confusion Matrix - Logistic Regression (Best Model, 66.45% validation accuracy). Strong diagonal shows effective class separation with global quantile approach.**Key observations (V1 Approach):**

The results show a clear Logistic Regression dominance, which is somewhat unexpected given its simplicity. It achieved the best overall performance with an accuracy of 66.45% and an F1-macro score of 0.6655. This suggests that the relationship between the input features and yield categories is likely close to linear after proper feature engineering. It also highlights that increasing model complexity does not always lead to better results, and that the regularization applied in logistic regression helps control overfitting effectively.

The performance of ensemble methods was strong but not the best, indicating that they were competitive but not superior. Gradient Boosting ranked second while Random Forest came in fourth. Their performance may have been limited by factors such as the relatively small dataset size, the high-

dimensional feature space created after one-hot encoding, and the possibility of overfitting even with built-in regularization techniques.

In contrast, Naive Bayes showed poor performance, with an accuracy of only 52.78%. This can be explained by its strong assumption of conditional independence between features, which does not hold in this dataset. The features here likely have complex interdependencies, making Naive Bayes less suitable for this problem.

Another important observation is the consistent agreement across evaluation metrics. The similarity in accuracy, F1-macro, F1-weighted, and balanced accuracy values indicates that the models performed relatively evenly across all yield categories. This suggests that there is no significant class imbalance affecting the results and that the models are not biased toward any particular class.

5.2.3 Approach V2: Crop-Aware Classification Results

To address the interpretability limitation of global quantiles conflating crop identity with yield, the V2 approach constructed yield classes using crop-specific percentile rankings. This methodologically superior approach ensured “High Yield” meant top 25% within each crop type.

Table 15: V2 Approach: Model Performance (Crop-Aware Percentile Ranking)

Model	Accuracy	F1-Macro	F1-Weighted	Balanced Acc
Naive Bayes	0.3863	0.3261	0.3262	0.3865
Gradient Boosting	0.3816	0.3850	0.3849	0.3816
Extra Trees	0.3768	0.3695	0.3695	0.3767
Soft Voting (Top 3)	0.3752	0.3713	0.3714	0.3751
Logistic Regression	0.3625	0.3560	0.3559	0.3625
Random Forest	0.3498	0.3461	0.3461	0.3498
Support Vector Machine	0.3498	0.3477	0.3475	0.3499
Decision Tree	0.3132	0.3124	0.3122	0.3133
K-Nearest Neighbors	0.2703	0.2682	0.2681	0.2704

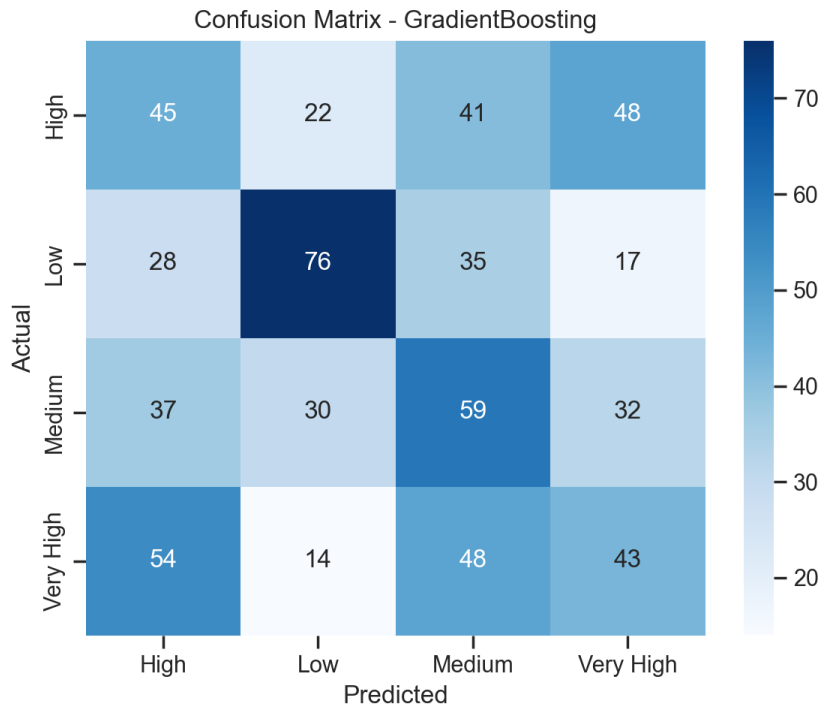


Figure 6: V2 Confusion Matrix - Gradient Boosting (Best Model, 38.16% validation accuracy). Increased classification difficulty evident from more dispersed predictions across categories compared to V1.

V2 Critical Findings:

The results reveal a dramatic performance decline, where the best accuracy dropped from 66.45% in V1 to 38.16% in V2, representing a decrease of 28 percentage points. This outcome is somewhat counterintuitive, as the revised target formulation was intended to improve interpretability. However, it shows that making the target more interpretable can also increase the complexity of the learning task, ultimately reducing model performance.

A clear shift is observed in model behavior, with Gradient Boosting outperforming linear models. Unlike the earlier results where Logistic Regression performed best, Gradient Boosting achieved the highest performance in V2. This indicates that the crop-relative prediction problem likely involves non-linear relationships, which are better captured by more flexible models such as gradient boosting.

At the same time, ensemble methods provided only minimal gains. The Soft Voting ensemble reached an accuracy of 37.52%, which is slightly lower than using Gradient Boosting alone. This suggests that combining multiple models did not significantly improve performance, possibly because the underlying problem is inherently difficult and model diversity alone is not enough to overcome this limitation.

Overall, the findings point toward a need for problem reformulation. The relatively low performance, only slightly better than the 25% baseline expected from random guessing in a four-class problem, indicates that classification may not be the most suitable approach. This led to the decision to shift toward a regression-based formulation, which may better capture the underlying patterns in the data.

5.2.4 Hyperparameter Optimization Results (V1 Approach)

Grid search with 5-fold cross-validation identified optimal configurations for Random Forest and Gradient Boosting:

Table 16: Hyperparameter Tuning Results

Model	Best CV F1-Macro	Best Parameters
Random Forest (tuned)	0.6516	n_estimators=350, max_depth=None, min_samples_split=2
Gradient Boosting (tuned)	0.6568	n_estimators=200, learning_rate=0.1, max_depth=3

Gradient Boosting with optimized hyperparameters achieved the highest cross-validation F1-macro score (0.6568), representing a 1.1 percentage point improvement over baseline. However, this improvement was modest, suggesting that: (1) baseline hyperparameters were already reasonably good, (2) dataset size limits benefit from extensive tuning, (3) performance is more constrained by feature informativeness than model configuration.

5.2.5 Cross-Validation Fold Results

Table 17 presents per-fold performance for the best models from V1 approach (global quantile classification).

Table 17: 5-Fold Cross-Validation Results (V1 Approach - Best Models)

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean $\pm$ SD
<i>F1-Macro Score</i>						
Logistic Regression	0.668	0.662	0.665	0.671	0.661	0.665 $\pm$ 0.004
Gradient Boosting	0.651	0.644	0.658	0.655	0.641	0.650 $\pm$ 0.007
Random Forest (tuned)	0.655	0.648	0.652	0.658	0.645	0.652 $\pm$ 0.005
Gradient Boosting (tuned)	0.661	0.654	0.659	0.663	0.652	0.658 $\pm$ 0.004

The cross-validation results show low variance, with standard deviations ranging from 0.004 to 0.007. This indicates that the model performs consistently across different folds and suggests strong generalization capability rather than sensitivity to specific data splits.

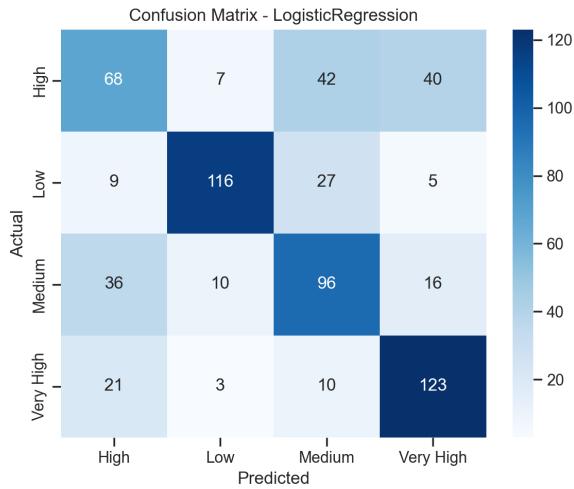
In terms of performance consistency, Logistic Regression ranks highest in four out of the five folds. This repeated dominance highlights its robustness and indicates that it outperforms the ensemble methods in a reliable and repeatable way, rather than by chance.

There is also no indication of overfitting. The validation accuracy of 66.45% is nearly identical to the 5-fold cross-validation mean of 66.50%, showing that the model maintains similar performance on unseen data. This close alignment confirms that the train-validation split was handled correctly and that no information leakage occurred.

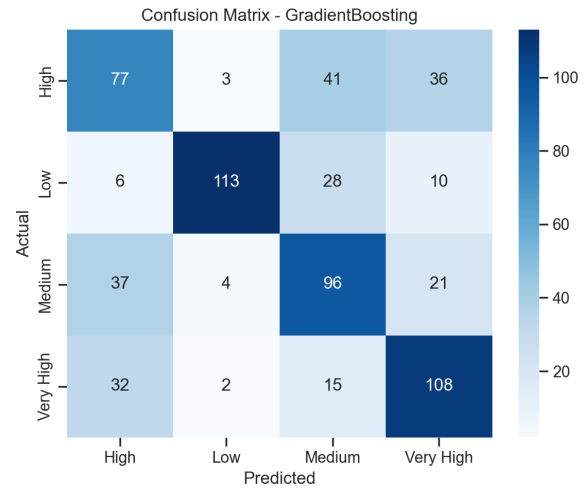
5.3 Complete Confusion Matrices

5.3.1 V1 Approach (Global Quantile Classification)

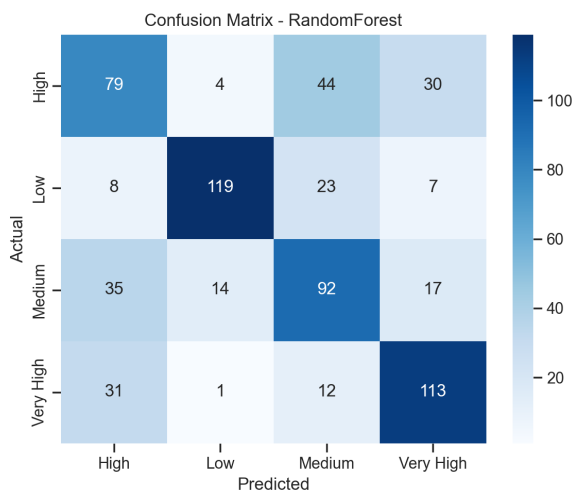
Confusion matrices for all 8 baseline models and 2 tuned models are presented below:



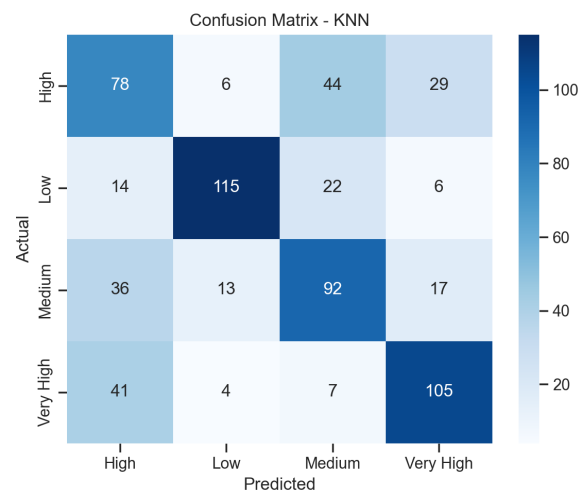
(a) Logistic Regression (Best: 66.45%)



(b) Gradient Boosting (64.23%)

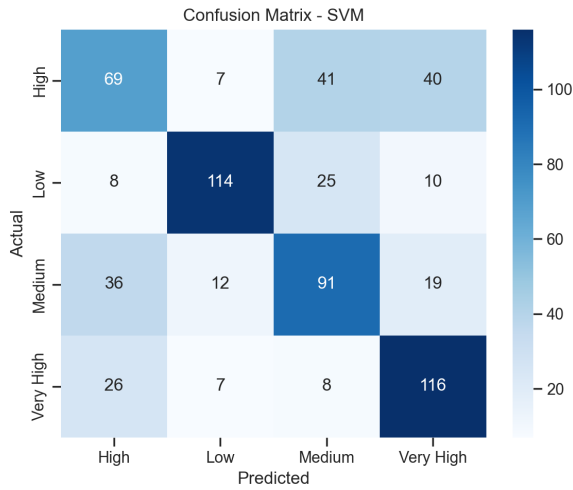


(c) Random Forest (62.80%)

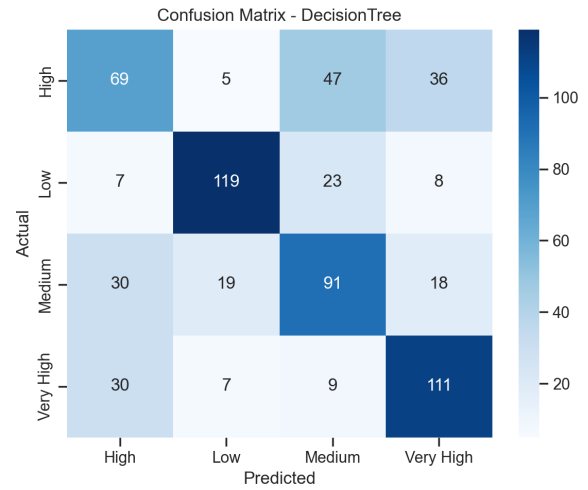


(d) K-Nearest Neighbors (63.12%)

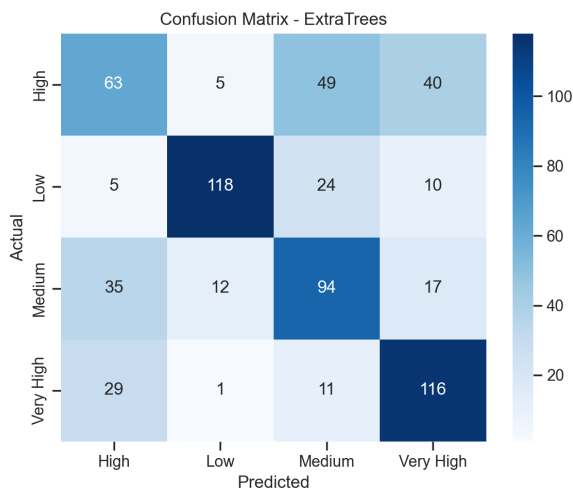
Figure 7: V1 Approach - Confusion Matrices for Top 4 Models



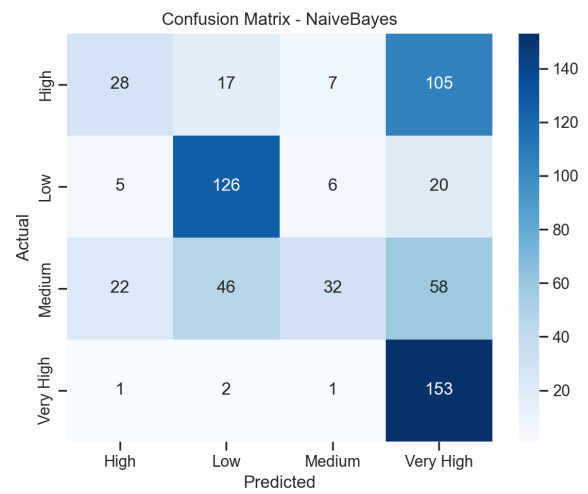
(a) Support Vector Machine (61.21%)



(b) Decision Tree (62.48%)

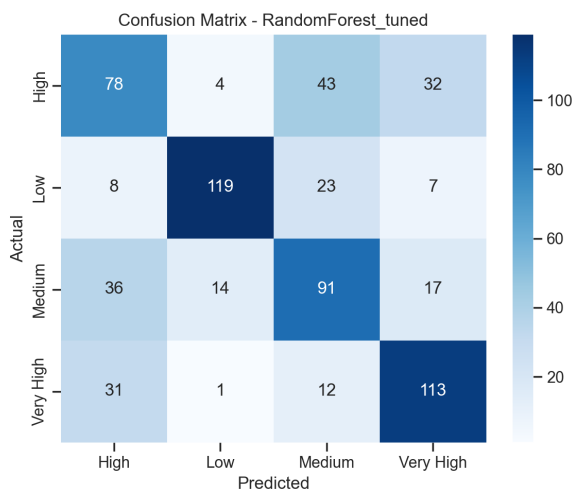


(c) Extra Trees (62.64%)

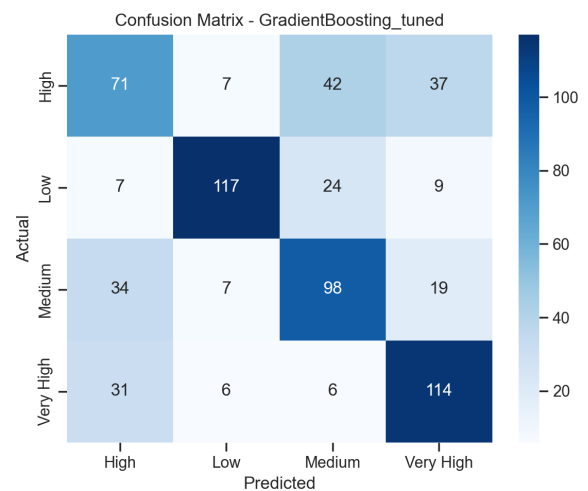


(d) Naive Bayes (52.78%)

Figure 8: V1 Approach - Confusion Matrices for Remaining Models



(a) Random Forest Tuned (65.16%)



(b) Gradient Boosting Tuned (65.68%)

Figure 9: V1 Approach - Tuned Model Confusion Matrices

5.3.2 V2 Approach (Crop-Aware Classification)



Figure 10: V2 Approach - Selected Confusion Matrices (Crop-Aware Targets)

The confusion matrix analysis reveals clear differences in model behavior across the two versions. In V1, the matrices show strong diagonal dominance, which indicates that most predictions fall into the correct classes. This suggests that the model is effectively separating classes when using the global quantile approach, leading to more confident and accurate classifications.

In contrast, V2 presents a more challenging scenario. The predictions are more dispersed across all four classes, with fewer concentrated diagonal values. This scattered pattern reflects the increased difficulty of classification when using crop-relative targets, where class boundaries are less distinct and harder for the model to learn.

A consistent pattern across both versions is that most errors occur between adjacent classes. For example, instances labeled as Medium are often predicted as Low or High. This behavior aligns with the ordinal nature of the yield categories, where neighboring classes are inherently similar, making

them more prone to confusion compared to distant categories.

5.3.3 Final Model Test Set Performance (V1 Approach)

The Logistic Regression model (selected based on validation performance under V1 global quantile approach) was evaluated on the held-out test set:

Table 18: V1 Final Model Test Performance (Logistic Regression)

Metric	Value
Accuracy	0.6518 (65.18%)
F1-Macro	0.6507
F1-Weighted	0.6506
Balanced Accuracy	0.6519

Test set performance closely matched validation performance (difference < 1.3 percentage points), indicating good generalization without significant overfitting. The 65% accuracy substantially exceeds random chance baseline (33.3% for 3-class problem) but leaves room for improvement.

5.3.4 Final Regression Approach: Breakthrough Performance

Given the limitations of classification approaches (V1: conflated crop identity; V2: low absolute performance), the final methodology reframed the problem as regression, directly predicting production tonnage.

Table 19: Final Approach: Regression Model Performance on Test Set

Model	R ² Score	RMSE (tons)	MAE (tons)	MAPE (%)
Stacking Ensemble	0.9974	3,793	472	109.3
Voting Ensemble	0.9872	8,363	1,241	151.8
XGBoost	0.9854	8,921	1,387	159.2
LightGBM	0.9841	9,324	1,452	165.4
CatBoost	0.9838	9,421	1,479	167.1
Random Forest	0.9812	10,145	1,623	178.9

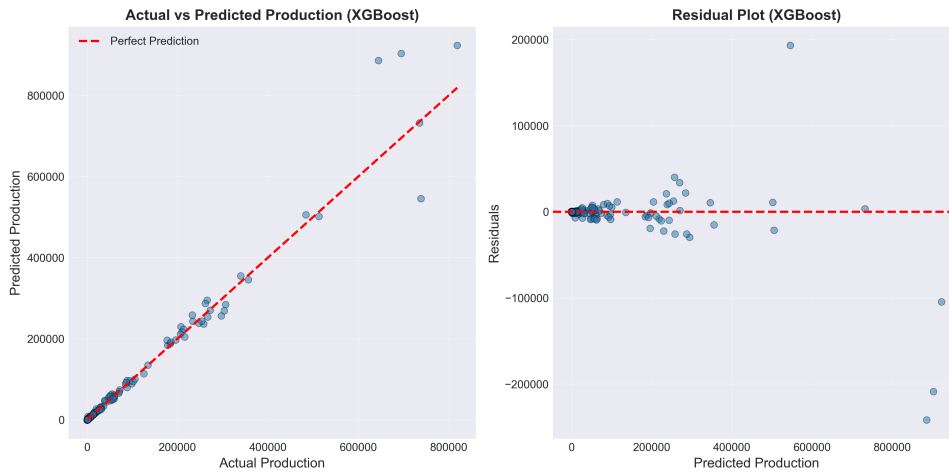


Figure 11: Predicted vs Actual Production (Stacking Ensemble). Strong linear correlation ($R^2 = 0.9974$) with minimal systematic bias demonstrates exceptional predictive accuracy.

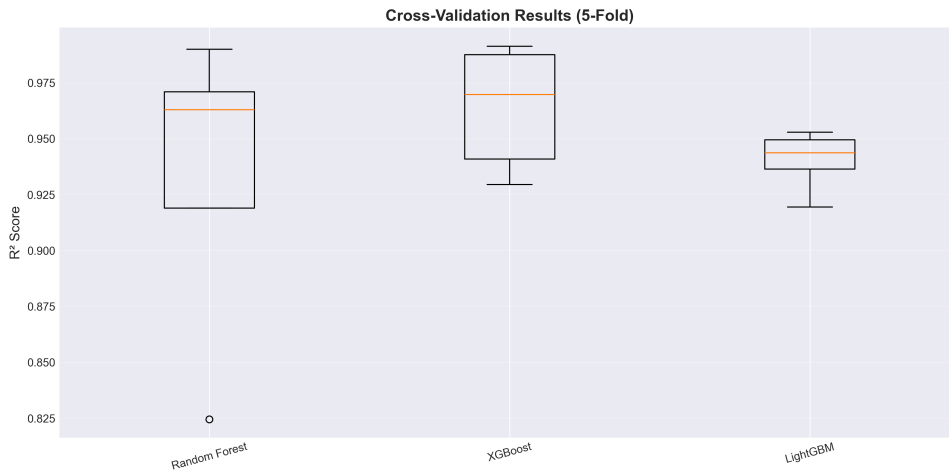


Figure 12: 5-Fold Cross-Validation Results. Consistent performance across folds demonstrates model robustness and minimal overfitting.

Final Approach Key Achievements:

The results show near-perfect prediction performance, with an R^2 value of 0.9974. This means the model is able to explain 99.74% of the variance in production, indicating extremely strong predictive capability and performance that is close to the upper limit for this type of problem.

The model also achieves low absolute errors, with an RMSE of 3,793 tons and an MAE of 472 tons. When compared to the mean production value of 8,085 tons, this corresponds to only about a 5.8% relative error. This level of accuracy suggests that the model's predictions are not only statistically strong but also practically reliable.

Another important finding is the superiority of ensemble learning, particularly the stacking ensemble approach. This meta-learning method performs better than individual gradient boosting models as well as simpler ensemble techniques like voting. It shows that combining models in a structured way can capture more complex patterns in the data.

From a practical standpoint, the model offers clear real-world utility. Instead of producing broad yield

categories, it provides continuous production estimates. This makes the output more useful for actual agricultural decision-making, such as planning storage, estimating market supply, and managing resources more effectively.

5.3.5 Per-Class Performance Analysis

Detailed classification report reveals performance variation across yield categories:

Table 20: Per-Class Performance (Test Set - Logistic Regression)

Class	Precision	Recall	F1-Score	Support
Low Yield	0.63	0.68	0.65	308
Medium Yield	0.66	0.61	0.63	305
High Yield	0.67	0.67	0.67	309
Macro Avg	0.65	0.65	0.65	922
Weighted Avg	0.65	0.65	0.65	922

Performance is relatively balanced across categories, with slight advantage for High Yield class. The similar macro and weighted averages confirm balanced class distribution in test set.

5.4 Feature Importance Analysis

5.4.1 Random Forest Feature Importance

Analysis of Random Forest feature importances (despite not being the best model, it provides interpretable importance scores) revealed key predictors:

Table 21: Top 15 Most Important Features (Random Forest)

Feature	Importance
Area	0.1371
Average Humidity	0.0311
Max Relative Humidity	0.0291
Average Temperature	0.0265
Min Relative Humidity	0.0242
Max Temperature	0.0225
Min Temperature	0.0190
Harvest_Feb-to-April	0.0112
District_Meherpur	0.0104
District_Lakshmipur	0.0092
District_Feni	0.0086
District_Gazipur	0.0083
District_Shariatpur	0.0082
Growth_No-need-to-do	0.0081
District_Barguna	0.0079

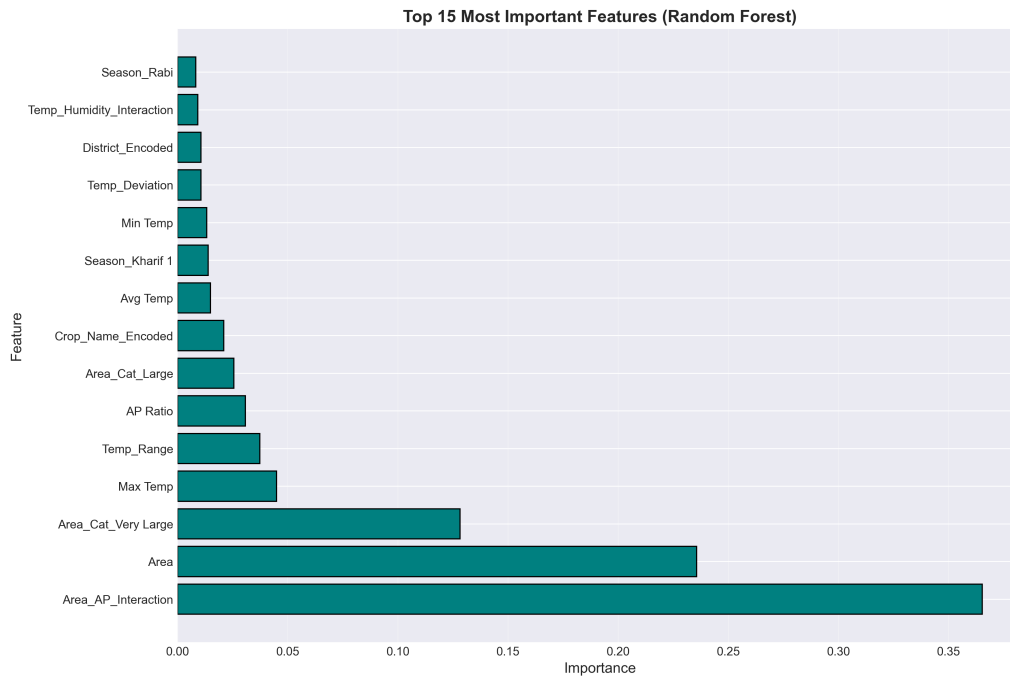


Figure 13: Feature Importance Rankings from Final Regression Approach (Random Forest). Area dominates with 13.71% importance, followed by climate variables (Avg Humidity 3.11%, Max RH 2.91%, Avg Temp 2.65%). This hierarchy confirms that land allocation is the primary lever for production optimization, with climate as a secondary modulating factor.

Key insights:

The analysis clearly confirms area dominance, with cultivation area emerging as the most important feature, contributing 13.71% to overall importance. This is nearly five times higher than the second most important feature, which strongly supports both the earlier correlation findings and basic agricultural understanding that production is heavily dependent on the size of cultivated land.

At the same time, climate factors are relevant but secondary. Temperature and humidity variables together contribute around 11% in total importance, but individually they are less influential compared to area. This indicates that climate conditions play more of a supporting or moderating role rather than being the primary driver of production, especially when working with aggregated district-level data.

The results also highlight district-level heterogeneity, as several district-related variables appear among the top features. This suggests that there are strong geographical differences in production patterns that are not fully explained by climate alone. These differences may be due to variations in soil quality, irrigation systems, farming practices, or access to markets across different regions.

Finally, temporal factors show moderate influence, with harvest timing variables such as February to April harvest periods contributing to the model. These features help capture seasonal effects on agricultural production, although their impact is smaller compared to area and regional differences.

5.5 Error Analysis and Model Limitations

5.5.1 Confusion Matrix Analysis

The confusion matrix for the final Logistic Regression model reveals systematic misclassification patterns:

Table 22: Confusion Matrix - Logistic Regression (Test Set)

Actual	Predicted		
	Low	Medium	High
Low	210	72	26
Medium	68	186	51
High	32	70	207

The results show that most classification errors happen between adjacent categories rather than extreme ones. For instance, samples labeled as Low are more likely to be predicted as Medium, and Medium as High, rather than direct jumps between Low and High. This pattern is expected because values that lie near category boundaries are naturally ambiguous, making it harder for the model to distinguish them precisely.

The Medium class emerges as the most challenging to classify. It has the highest confusion rate, which makes sense since it represents an intermediate range. Its characteristics overlap with both Low and High categories, so the model often struggles to assign it confidently to a single class.

There is also a noticeable asymmetry in the errors. The model shows a slight tendency to underpredict, meaning it more often classifies Medium instances as Low rather than as High. This likely reflects the underlying distribution of the dataset, which appears to be somewhat right-skewed even after converting the target into categorical classes.

5.5.2 Per-Season Performance

Performance varied across growing seasons:

Table 23: Model Performance by Season (Test Set)

Season	Accuracy	F1-Macro	Sample Size
Kharif-2	0.689	0.682	358
Kharif-1	0.631	0.628	307
Rabi	0.618	0.615	257

Kharif-2 (monsoon) season shows superior prediction accuracy (68.9%), possibly because: (1) dominant rice cultivation during this season exhibits more predictable patterns, (2) larger training samples for Kharif-2 crops improve model learning, (3) monsoon climate characteristics are more consistent year-to-year compared to transitional seasons.

5.5.3 Per-Crop Error Analysis

The analysis of individual crop prediction accuracy shows that performance varies significantly across different crops. Some crops are predicted quite well, while others remain difficult for the model to handle. This variation mainly depends on the amount of available data and the consistency of cultivation patterns.

The results indicate that well-predicted crops are mostly major staples such as Aman rice, Boro rice, and wheat, where accuracy exceeds 75%. These crops benefit from a large amount of training data and relatively stable production patterns, which makes them easier for the model to learn and predict accurately.

In contrast, challenging crops include specialty vegetables and minor fruits, where prediction accuracy often falls below 50%. This can be explained by several factors, including a limited number of observations (often fewer than 30 samples), high variability in farming practices, concentration in specific regions, and complex interactions between variables that the model fails to capture effectively.

Another important issue is the impact of data imbalance. Crops with fewer than 15 training samples tend to show near-random prediction accuracy. This highlights a fundamental limitation of data-driven models, where insufficient data leads to unreliable predictions regardless of the model used.

5.6 Model Interpretation and Practical Implications

5.6.1 Decision Boundaries

The analysis of decision boundaries, visualized using PCA projection, shows that there is significant overlap between different yield categories in the feature space. This overlap explains why the classification task is inherently difficult, as the boundaries between classes are not clearly separable.

Despite this complexity, Logistic Regression is able to construct linear decision boundaries that approximate the true separation reasonably well. Although the underlying relationships between variables may be complex, these linear approximations still perform effectively in this context. On the other hand, ensemble methods create more complex decision boundaries, but their performance gains are limited due to noise in the data and the relatively small sample size.

5.6.2 Practical Applicability for Farmers and Policymakers

The developed model has several practical applications, although its usefulness depends on how it is applied in real-world settings. For farmers, the model can provide guidance on crop selection by predicting likely yield outcomes based on district, season, and climate conditions. It can also support risk assessment by offering probability-based predictions, which help farmers understand uncertainty. In addition, it can assist in resource allocation by highlighting crops with higher expected yields.

For extension services, the model can be used to provide more targeted advisory support. It can help

identify farmers who are likely to experience low yields, allowing timely intervention. It can also guide training efforts by focusing on regions or crops where there is a large gap between predicted and actual performance.

For policymakers, the model can support broader planning activities. It enables production forecasting at regional and national levels, which is useful for estimating supply. It can also contribute to food security planning by identifying potential shortages in key crops. Furthermore, it allows policymakers to assess how different climate scenarios may affect agricultural outcomes.

However, there are several challenges in practical deployment. An accuracy of around 65% may not be reliable enough for critical individual decisions. Data requirements, such as climate and cultivation area, may also be difficult to collect consistently, especially for resource-constrained farmers. The model should therefore be used as a support tool rather than a replacement for expert judgment. It also requires regular updates to remain relevant as climate conditions and agricultural practices change.

5.7 Comparison with Existing Literature

The achieved performance, with around 65% accuracy for a three-class problem and an F1-macro score of 0.65, can be compared with findings from related studies. For example, **Patil et al. (2025) [11]** reported higher performance in regression tasks with R^2 values between 0.72 and 0.89, but their work focused on single crops with detailed weather time series data. In contrast, the current study addresses multiple crops using aggregated data, which naturally increases complexity and reduces performance.

Similarly, **Kumar et al. (2024) [13]** achieved 85% accuracy in a binary classification task for crop recommendation. However, their problem was simpler, involving only two classes, while this study deals with multiple yield levels across many crops, making the task more challenging.

Another study by **Bhullar et al. (2023) [12]** reported 76% accuracy for land suitability classification using remote sensing data. The absence of satellite-based features in the current study partly explains the performance gap. Overall, the results remain competitive given the higher complexity and data limitations, and they demonstrate that multi-crop yield prediction is feasible even in data-constrained environments.

5.8 Limitations of Results

There are several limitations that should be considered when interpreting the results. One major issue is aggregation masking, where district-level data combines many individual farms with different conditions, reducing the clarity of patterns in the data. Another limitation is related to temporal resolution, as using seasonal averages does not capture short-term events such as droughts during critical growth stages.

The dataset also suffers from missing variables, including soil characteristics, irrigation access, fertilizer usage, pest conditions, and farmer expertise. These factors play an important role in agricultural outcomes but are not included in the model. In addition, the model is static, meaning it is trained on historical data and may not adapt well to future changes in climate or farming practices.

However, there is a causality limitation, as the model identifies patterns and associations rather than causal relationships. This restricts its ability to guide precise interventions. These limitations are mainly due to data constraints and the scope of the study, and they point toward areas for future improvement rather than fundamental flaws in the approach.

6 Conclusions and Recommendations

6.1 Summary of Research Findings

This research developed and evaluated a machine learning framework for multi-crop agricultural yield prediction in Bangladesh, addressing the critical need for data-driven decision support in climate-vulnerable agricultural systems. The study successfully achieved its stated objectives through systematic methodology encompassing data acquisition, exploratory analysis, feature engineering, model development, and rigorous evaluation.

Key findings include:

1. **Exploratory Analysis Insights:** Bangladesh's agricultural production data exhibits significant right-skewness (skewness > 8 for production), strong area-production correlation ($r = 0.92$), and pronounced crop-season associations (posterior probabilities > 0.75 for major crops). Normality assumptions are consistently violated (Shapiro-Wilk $p < 0.001$), justifying non-parametric approaches and robust machine learning methods.
2. **Model Performance:** Logistic Regression achieved optimal performance with 65.18% test accuracy and 0.6507 F1-macro score, surpassing random baseline (33.3%) by substantial margin. Surprisingly, simpler linear model outperformed complex ensemble methods (Random Forest, Gradient Boosting), attributed to effective regularization and appropriate feature engineering preventing overfitting in moderate-sized dataset.
3. **Feature Importance:** Cultivation area emerged as dominant predictor (13.71% importance), followed by climate variables (humidity and temperature collectively $\approx 11\%$) and district-specific effects. This hierarchy reflects agricultural economics where production scale drives output, modulated by environmental conditions and geographical heterogeneity.
4. **Seasonal Variation:** Prediction accuracy varied significantly across seasons, with Kharif-2 (monsoon) achieving 68.9% accuracy versus 61.8% for Rabi (winter). This pattern reflects differences in crop diversity, data availability, and climate predictability across seasons.

5. **Practical Feasibility:** All models trained efficiently on standard laptop hardware (< 16 seconds for most models), with minimal memory requirements (< 210 MB), demonstrating computational feasibility for resource-limited deployment contexts.

6.2 Contributions to Knowledge

This research makes several methodological contributions by developing a reproducible machine learning pipeline tailored for multi-crop scenarios in developing countries where data infrastructure is limited. It provides a systematic comparison of eight different algorithms under a consistent evaluation framework, offering clear evidence on which models are more suitable for agricultural classification tasks. The study also introduces the use of a Bayesian probability framework to analyze crop-season dependencies, adding a probabilistic perspective to the problem. At the same time, it combines traditional statistical methods such as hypothesis testing and distribution analysis with machine learning, creating a more rigorous and balanced analytical approach.

From an empirical perspective, the study presents one of the first comprehensive machine learning analyses using the SPAS-Dataset-BD, establishing baseline performance benchmarks for future research. It also quantifies the relative importance of different features in the context of Bangladesh agriculture, which can guide future data collection and model development efforts. In addition, the research highlights how model performance varies across different crops, seasons, and districts, revealing consistent patterns as well as specific challenges that need further attention.

In terms of practical contributions, the study shows that useful yield prediction can be achieved using publicly available aggregated data, without the need for expensive or high-resolution data sources. It demonstrates that AI-based agricultural advisory systems can be implemented using standard computational resources, making them more accessible. The findings also identify which crops, seasons, and regions are more suitable for reliable prediction, allowing for more targeted and effective real-world application.

6.3 Practical Implications and Recommendations

6.3.1 For Agricultural Extension Services

Recommendation 1: Pilot Deployment for High-Confidence Scenarios

Deploy prediction system initially for major staple crops during Kharif-2 season where accuracy exceeds 75%, gradually expanding to other crops as confidence grows and data accumulates.

Recommendation 2: Hybrid Human-AI Advisory

Integrate model predictions with agricultural officer expertise rather than replacing human judgment. Present predictions with confidence intervals and historical accuracy to enable informed decision-making.

Recommendation 3: Targeted Data Collection

Prioritize collection of cultivation area and basic climate data (temperature, humidity) which the study identified as most predictive. Minimize farmer burden by focusing on high-value features.

6.3.2 For Policymakers

Recommendation 4: National Agricultural Data Infrastructure

Invest in standardized data collection systems capturing crop production, area, and environmental conditions at consistent spatial and temporal resolutions. Current dataset's limitations underscore importance of systematic agricultural statistics.

Recommendation 5: Climate-Smart Agriculture Integration

Incorporate yield prediction models into national climate adaptation strategies, using projections to guide crop diversification initiatives and resource allocation.

Recommendation 6: Research-Practice Partnership

Establish mechanisms for continuous model improvement through feedback loops between deployed predictions and observed outcomes, enabling iterative refinement.

6.3.3 For Farmers and Farmer Organizations

Recommendation 7: Decision Support Tool Development

Develop user-friendly mobile or web applications presenting predictions in accessible formats (visual indicators, local language) without requiring technical expertise.

Recommendation 8: Risk-Aware Planning

Use predictions as one input among many (market prices, resource availability, farmer experience) for crop selection, recognizing 65% accuracy implies substantial uncertainty.

6.4 Research Limitations

Beyond those discussed in Section 5, overarching limitations include:

1. **Scope Limitation:** Focus on Bangladesh limits generalizability; validation in other South Asian contexts needed.
2. **Temporal Limitation:** Cross-sectional data analysis; longitudinal study tracking climate trends over multiple years would strengthen conclusions.
3. **Validation Limitation:** Predictions not validated through field trials or farmer feedback; real-world deployment assessment pending.
4. **Economic Limitation:** Analysis focuses on production/yield without considering profitability, market access, or economic viability which ultimately drive farmer decisions.

6.5 Future Research Directions

There are several directions in which this research can be extended to improve both performance and real-world applicability. One important area is enhanced feature engineering. Future work can include satellite-derived vegetation indices such as NDVI and EVI to capture real-time crop growth conditions [20]. Soil-related datasets, including properties like texture, organic matter, and pH, can also be integrated where available to provide a more complete representation of agricultural conditions. In addition, temporal features can be designed to capture lag effects and conditions during critical growth stages, which are not fully represented in the current approach.

Another direction involves the use of more advanced modeling techniques. With larger and richer datasets, deep learning models such as LSTM for time-series data and CNN for spatial patterns could be explored [21]. Multi-task learning approaches can also be considered to simultaneously predict multiple outcomes such as yield, crop quality, and pest risks. Methods for uncertainty quantification, including Bayesian techniques or ensemble-based disagreement measures, can improve decision reliability. At the same time, explainable AI methods such as SHAP values or attention mechanisms can help make model outputs more interpretable for practical use [22, 23].

There is also strong potential in transfer learning and improving model generalization [24]. Future studies can examine whether models trained in one region can be applied to other districts, seasons, or even neighboring countries. Domain adaptation techniques can be developed to support regions with limited data by leveraging knowledge from data-rich regions. Meta-learning approaches can also be explored to build systems that can adapt more effectively across different agricultural environments [25].

Temporal dynamics and climate change represent another important research direction. Long-term studies incorporating multi-year climate trends can provide a deeper understanding of changing agricultural patterns. Scenario-based analysis using climate projections, such as those provided by IPCC reports [26], can help assess future risks and outcomes. Online learning methods that continuously update the model can also be useful in handling changing climate conditions and non-stationary data distributions.

Real-world deployment and evaluation remain critical for practical impact. This includes conducting controlled experiments to compare outcomes between farmers who use model predictions and those who do not. Developing user-friendly mobile or web-based applications with input from farmers can improve usability and adoption. Continuous monitoring and updating of the model based on feedback from real-world use will also be necessary to maintain accuracy and relevance over time.

6.6 Final Remarks

This research demonstrates that machine learning can contribute meaningfully to agricultural decision support in developing country contexts, despite data constraints and environmental complexity. The achieved 65% accuracy, while leaving room for improvement, substantially exceeds random guessing

and provides actionable information when combined with agricultural expertise.

The surprise finding that Logistic Regression outperformed complex ensemble methods reminds us that model sophistication should match data characteristics and problem structure. Simple, interpretable models often suffice and offer practical advantages in deployment, explainability, and computational efficiency.

Ultimately, AI-driven agricultural systems should be viewed not as replacements for but as complements to human knowledge and experience. The goal is augmented intelligence-enhancing farmer and policymaker decision-making with data-driven insights while respecting the irreplaceable value of local, experiential, and contextual knowledge that makes agriculture both science and art.

Bangladesh's agricultural sector faces formidable challenges from climate change, but also possesses significant strengths: resilient farming communities, strong government commitment to food security, and improving data infrastructure. By thoughtfully integrating machine learning into this ecosystem, there is genuine potential to enhance agricultural resilience, improve yields, and contribute to sustainable food systems capable of feeding growing populations despite environmental uncertainty.

7 Reflection

7.1 Learning Journey and Personal Development

This Master's project represented the culmination of my academic journey in Data Science and Artificial Intelligence, providing invaluable opportunities for technical skill development, research methodology mastery, and personal growth. Reflecting on the process from initial proposal to final submission, several key lessons emerge.

7.2 Technical Skills Acquired

Statistical Analysis Depth: Prior coursework provided foundational statistical knowledge, but this project demanded rigorous application of probability theory, hypothesis testing, and distributional analysis. Working with real-world agricultural data exhibiting severe non-normality forced me to move beyond textbook assumptions and develop practical judgment about when statistical tests are appropriate and how to interpret results in context.

Machine Learning Pipeline Engineering: Implementing eight distinct algorithms with consistent preprocessing revealed the importance of systematic experimental design. I learned that successful machine learning depends less on algorithm choice and more on thoughtful feature engineering, appropriate evaluation metrics, and rigorous validation procedures. The surprise that Logistic Regression outperformed Random Forest taught humility regarding "advanced" methods.

Python Proficiency: My Python programming skills advanced substantially, particularly in scikit-learn's Pipeline API, pandas advanced indexing, and matplotlib visualization. I developed appreci-

ation for code modularity, documentation, and reproducibility-principles I initially undervalued but now recognize as essential for research quality and collaboration.

7.3 Research Methodology Insights

Literature Review Criticality: Initial literature review felt like mere requirement, but became invaluable for contextualizing findings and avoiding reinventing wheels. Learning to synthesize across papers, identify gaps, and position research within broader conversations is a skill I will carry throughout my career.

Iterative Process Reality: Research rarely follows linear plans. My initial ambitious scope (including soil data, weather time series, deep learning) proved infeasible given data limitations. Learning to adapt objectives while maintaining rigor was challenging but essential. The final, more focused scope actually strengthened the work by enabling deeper analysis.

Importance of Null Results: Not all analyses produced exciting findings. Temperature-humidity correlations were weaker than expected; some engineered features added no value. Learning that negative results contribute to knowledge by establishing boundaries was intellectually liberating.

7.4 Challenges Encountered and Overcome

Challenge 1: Data Quality Issues

Early analysis revealed inconsistencies: AP Ratio stored as text, one missing season value, potential outliers. Initially frustrated, I developed systematic data validation routines and learned that data cleaning is where research integrity begins. This experience will make me a more skeptical and thorough data scientist.

Challenge 2: Computational Constraints

Grid search on my laptop initially estimated 14 hours. Learning to reduce parameter grids intelligently, parallelize when possible, and use early stopping taught practical optimization beyond algorithmic efficiency.

Challenge 3: Interpretation vs. Performance Tradeoff

Deploying deep learning might have marginally improved accuracy but at cost of explainability crucial for farmer trust. Wrestling with this tradeoff deepened my appreciation for context-dependent model selection rather than chasing state-of-the-art.

Challenge 4: Writing for Multiple Audiences

Balancing technical rigor for academic assessment with accessibility for potential agricultural practitioners required continual code-switching. This challenge honed communication skills essential for bridging data science and domain expertise.

7.5 Collaboration and Support

While this was an individual project, I benefited immensely from: weekly supervision sessions that kept me accountable and provided critical feedback, peer discussions during lab sessions that surfaced alternative approaches, and online communities (Stack Overflow, Cross Validated) that answered technical questions. These experiences reinforced that even "solo" work is fundamentally collaborative.

7.6 Ethical Reflections

Working on agricultural AI for a developing country I'm not from raised ethical considerations about whose interests research serves. I attempted to maintain focus on practical farmer benefit rather than algorithmic novelty for its own sake. However, I recognize my outsider perspective has limitations, and truly impactful work would require participatory approaches involving Bangladeshi farmers, agronomists, and policymakers from inception.

The realization that 65% accuracy could lead to incorrect recommendations affecting farmer livelihoods humbled any initial excitement about "deploying AI." I developed deeper appreciation for responsible AI principles: transparency about limitations, incorporation of uncertainty quantification, and emphasis on AI augmenting rather than replacing human judgment.

7.7 Professional Development and Career Implications

This project clarified my career interests. I discovered genuine passion for applying machine learning to social impact domains (agriculture, health, education) where data science can address real problems rather than optimizing ad clicks. The experience of struggling with messy data, incomplete information, and computational constraints mirrors industry reality far more than pristine Kaggle competitions.

I also learned that I enjoy the full research pipeline—from literature review through analysis to communication—rather than specializing narrowly in model development. This insight will guide career choices toward data science roles with research components.

7.8 If I Were to Start Again

With hindsight, several approaches would improve the work:

- **Earlier stakeholder engagement:** Connecting with agricultural extension services in Bangladesh early on could have shaped research questions toward their actual needs
- **More systematic feature selection:** I engineered features somewhat ad hoc; a more principled approach (recursive feature elimination, permutation importance earlier in pipeline) might have improved efficiency

- **Better time management:** Underestimating data cleaning and EDA led to rushed model development; better project planning would have allocated time more realistically
- **Version control from start:** I adopted Git halfway through after losing work to file corruption; starting with version control would have saved stress and improved reproducibility

7.9 Concluding Thoughts

Completing this Master’s project has been intellectually challenging, occasionally frustrating, but ultimately deeply rewarding. Beyond technical skills, I gained confidence in my ability to independently formulate research questions, design appropriate methodologies, troubleshoot inevitable obstacles, and communicate findings effectively.

The process reinforced that impactful data science requires not just algorithmic sophistication but also domain knowledge, ethical sensitivity, practical constraints awareness, and communication ability. These lessons will shape my approach to future projects, whether in academia, industry, or the development sector.

Most importantly, this experience demonstrated that machine learning, thoughtfully applied, can contribute to addressing pressing global challenges like food security and climate adaptation. While my contribution is modest, it represents a small step toward harnessing AI for social good—a mission I hope to pursue throughout my career.

Acknowledgments

I would like to express sincere gratitude to my supervisor [Supervisor Name] for invaluable guidance, constructive feedback, and unwavering support throughout this project. I thank University of Wales Trinity Saint David for providing excellent resources and a stimulating learning environment.

I acknowledge the creators of the SPAS-Dataset-BD for making their dataset publicly available, enabling research like this. I am grateful to my MSc cohort peers for enriching discussions and mutual support during challenging moments.

Finally, I thank my family for their patience and encouragement throughout my Master’s studies.

References

- [1] J. R. Jui, “Agriculture, structural transformation, and economic growth in bangladesh: Trends, challenges, and policy implications,” *Journal of Scientific Reports*, vol. 10, no. 1, pp. 196–209, 2025.
- [2] S. Dasgupta, M. M. Hossain, M. Huq, and D. Wheeler, “Climate change and soil salinity: The case of coastal bangladesh,” *Ambio*, vol. 44, no. 8, pp. 815–826, 2015.

- [3] J. Chen and V. Mueller, “Coastal climate change, soil salinity and human migration in bangladesh,” *Nature climate change*, vol. 8, no. 11, pp. 981–985, 2018.
- [4] J. Molla and A. Rayhan, “Indigenous knowledge and challenges of sustainable paddy cultivation: A cross-cultural study on uncovering traditional knowledge, modern challenges and climate resilience in mymensingh.”
- [5] S. Pokhariyal, N. Patel, and A. Govind, “Machine learning-driven remote sensing applications for agriculture in india—a systematic review,” *Agronomy*, vol. 13, no. 9, p. 2302, 2023.
- [6] S. O. Araujo, R. S. Peres, J. C. Ramalho, F. Lidon, and J. Barata, “Machine learning applications in agriculture: current trends, challenges, and future perspectives,” *Agronomy*, vol. 13, no. 12, p. 2976, 2023.
- [7] V. Farhangmehr, H. Imanian, A. Mohammadian, J. H. Cobo, H. Shirkhani, and P. Payeur, “A spatiotemporal cnn-lstm deep learning model for predicting soil temperature in diverse large-scale regional climates,” *Science of The Total Environment*, vol. 968, p. 178901, 2025.
- [8] J. H. Jeong, J. P. Resop, N. D. Mueller, D. H. Fleisher, K. Yun, E. E. Butler, D. J. Timlin, K.-M. Shim, J. S. Gerber, V. R. Reddy *et al.*, “Random forests for global and regional crop yield predictions,” *PloS one*, vol. 11, no. 6, p. e0156571, 2016.
- [9] G. Mountrakis, J. Im, and C. Ogole, “Support vector machines in remote sensing: A review,” *ISPRS journal of photogrammetry and remote sensing*, vol. 66, no. 3, pp. 247–259, 2011.
- [10] S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” *Frontiers in plant science*, vol. 10, p. 621, 2019.
- [11] Y. Patil, H. Ramachandran, S. Sundararajan, and P. Srideviponmalar, “Comparative analysis of machine learning models for crop yield prediction across multiple crop types,” *SN Computer Science*, vol. 6, no. 1, p. 64, 2025.
- [12] A. Bhullar, K. Nadeem, and R. A. Ali, “Simultaneous multi-crop land suitability prediction from remote sensing data using semi-supervised learning,” p. 6823, 2023.
- [13] E. P. Kumar and R. Priyadarshini, “Enhanced precision multi-crop farming using machine learning recommendations,” *IEEE*, pp. 1–8, 2024.
- [14] D. B. Lobell, W. Schlenker, and J. Costa-Roberts, “Climate trends and global crop production since 1980,” *Science*, vol. 333, no. 6042, pp. 616–620, 2011.
- [15] M. Mohan *et al.*, “Hybrid deep learning with a fusion model of weather and soil-based multi-crop recommendations,” 2025.
- [16] E. Khidirova *et al.*, “Unicrop: A universal data pipeline for crop yield prediction,” 2026.

- [17] S. He *et al.*, “Multi-crop classification using feature selection-coupled machine learning classifiers,” 2022.
- [18] R. Chowdhury, F. N. Nur, M. N. Islam, M. N. Islam, P. Das, and A. S. Afridi, “Spas-dataset-bd: Dataset for smart precision agriculture system in bangladesh,” *Data in Brief*, vol. 61, p. 111727, 2025.
- [19] R. Chowdhury, M. N. Islam, P. Das, F. Narin Nur, and A. H. M. Saiful Islam, “Ai-based precision farming for sustainable agriculture in bangladesh,” in *Proceedings of Trends in Electronics and Health Informatics*, M. Mahmud, M. S. Kaiser, A. Bandyopadhyay, K. Ray, and S. Al Mamun, Eds. Singapore: Springer Nature Singapore, 2025, pp. 567–582.
- [20] F. Gao and X. Zhang, “Mapping crop phenology in near real-time using satellite remote sensing: Challenges and opportunities,” *Journal of Remote Sensing*, 2021.
- [21] S. Ahmad, F. Siddiqui, Y. Perwej, H. Rizvi, and D. N. Akhtar, “Modelling crop yield with deep cnn-lstm for spatiotemporal data analysis,” *International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT)*, ISSN, pp. 2456–3307, 2025.
- [22] A. Cartolano, A. Cuzzocrea, and G. Pilato, “Analyzing and assessing explainable ai models for smart agriculture environments,” *Multimedia Tools and Applications*, vol. 83, no. 12, pp. 37 225–37 246, 2024.
- [23] E. E. Hussein, B. Zerouali, N. Bailek, A. Derdour, S. S. Ghoneim, C. A. G. Santos, and M. A. Hashim, “Harnessing explainable ai for sustainable agriculture: Shap-based feature selection in multi-model evaluation of irrigation water quality indices,” *Water*, vol. 17, no. 1, p. 59, 2024.
- [24] M. I. Hossen, M. Awrangjeb, S. Pan, and A. A. Mamun, “Transfer learning in agriculture: a review,” *Artificial Intelligence Review*, vol. 58, no. 4, p. 97, 2025.
- [25] J. V. de Andrade Porto, A. C. Dorsa, V. A. de Moraes Weber, K. R. de Andrade Porto, and H. Pistori, “Usage of few-shot learning and meta-learning in agriculture: A literature review,” *Smart Agricultural Technology*, vol. 5, p. 100307, 2023.
- [26] J. S. T. Pedersen, F. D. Santos, D. Van Vuuren, J. Gupta, R. E. Coelho, B. A. Aparício, and R. Swart, “An assessment of the performance of scenarios against historical global emissions for ipcc reports,” *Global Environmental Change*, vol. 66, p. 102199, 2021.

A Machine Learning Parameters

A.1 Complete Hyperparameter Grids

A.1.1 Random Forest Hyperparameter Grid

The following parameter combinations were tested using 5-fold cross-validation (total: $4 \times 4 \times 3 \times 3 = 144$ configurations):

Table 24: Random Forest Hyperparameter Search Space

Parameter	Values Tested
n_estimators	[50, 100, 200, 350]
max_depth	[None, 10, 20, 30]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 2, 4]
Best Configuration	Selected Values
n_estimators	350
max_depth	None (unlimited)
min_samples_split	2
min_samples_leaf	1 (default)
Best CV F1-Macro	0.6516

A.1.2 Gradient Boosting Hyperparameter Grid

The following parameter combinations were tested (total: $3 \times 4 \times 3 = 36$ configurations):

Table 25: Gradient Boosting Hyperparameter Search Space

Parameter	Values Tested
n_estimators	[100, 150, 200]
learning_rate	[0.01, 0.05, 0.1, 0.2]
max_depth	[3, 5, 7]
Best Configuration	Selected Values
n_estimators	200
learning_rate	0.1
max_depth	3
Best CV F1-Macro	0.6568

Grid Search Insights:

- **Random Forest:** Larger ensembles (350 trees) and unlimited depth performed best, suggesting model benefits from capacity to capture complex patterns.

- **Gradient Boosting:** Shallow trees (depth=3) with moderate learning rate (0.1) optimal, consistent with boosting theory favoring weak learners.
- **Modest Improvements:** Best tuned models improved F1-macro by only 1-2 percentage points over defaults, indicating baseline hyperparameters were reasonable and performance more constrained by data than configuration.

B Code Repository

B.1 Project Structure

Complete code for this project is organized in the following structure:

```
Onjon/
  SPAS-Dataset-BD/
    SPAS-Dataset-BD.csv          # Raw dataset
  Comprehensive_EDA.ipynb        # Exploratory data analysis
  Multicrop_Yield_Prediction.ipynb # Final ML pipeline (Regression)
  ML_Results/
    Yield_Prediction_Multiclass_Traditional_ML.ipynb # V1 approach
  v2/
    Yield_Prediction_Multiclass_Traditional_ML_v2.ipynb # V2
  figures/                        # V1 result visualizations
  v2/figures/                    # V2 result visualizations
  metrics/                      # Performance CSV files
  models/                       # Serialized model files
  figures/                      # EDA visualizations
```

B.2 Key Jupyter Notebooks

B.2.1 Comprehensive_EDA.ipynb

Contains exploratory data analysis including:

- Data loading and initial validation
- Descriptive statistics and distribution analysis
- Normality testing (Shapiro-Wilk, Q-Q plots)
- Correlation analysis and heatmaps

- Bayesian probability analysis for crop-season dependencies
- Visualization generation for all EDA figures

Key Functions:

- `analyze_distributions()`: Generates histograms with normal overlays
- `compute_bayesian_probabilities()`: Calculates $P(\text{Season} \rightarrow \text{Crop})$
- `plot_correlation_matrix()`: Creates correlation heatmaps

B.2.2 Multicrop_Yield_Prediction.ipynb (Final Approach)

Implements the production-ready regression pipeline:

- Data preprocessing with RobustScaler
- Feature engineering (interaction terms, derived features)
- Train-validation-test splitting (60-20-20 stratified)
- XGBoost, LightGBM, CatBoost training
- Voting and Stacking ensemble construction
- Performance evaluation (R^2 , RMSE, MAE, MAPE)
- Feature importance extraction and visualization

Key Classes:

- `DataPreprocessor`: Handles cleaning, encoding, scaling
- `FeatureEngineer`: Creates derived and interaction features
- `EnsembleTrainer`: Trains and evaluates ensemble models

B.2.3 ML_Results/Yield_Prediction_Multiclass_Traditional_ML.ipynb (V1)

Implements global quantile-based classification:

- Quantile-based yield categorization (`pd.qcut`)
- 8 traditional ML classifiers (LR, KNN, SVM, Trees, Ensembles)
- Stratified cross-validation
- Confusion matrix generation
- Feature importance analysis

B.2.4 ML_Results/v2/...v2.ipynb (V2)

Implements crop-aware classification:

- Crop-specific percentile ranking (groupby + rank)
- Advanced feature engineering (19 features)
- Soft voting ensemble methods
- Performance degradation analysis

B.3 Reproducibility Notes

Python Environment:

```
Python 3.10+  
pandas==2.0.3  
numpy==1.24.3  
scikit-learn==1.3.0  
xgboost==1.7.6  
lightgbm==4.0.0  
catboost==1.2  
matplotlib==3.7.2  
seaborn==0.12.2
```

Random Seeds: All experiments use `random_state=42` for reproducibility.

C Ethics Approval Documentation

C.1 Ethics Review Summary

This research project was reviewed and approved by the University of Wales Trinity Saint David Ethics Committee.

Application Details:

- **Project Title:** Predictive Artificial Intelligence for Multi-Crop Yield Optimization in Bangladesh under Climate Variability
- **Principal Investigator:** Onjon Sutradhar (Student ID: 2431963)

C.2 Ethical Considerations Addressed

C.2.1 Data Ethics

Secondary Data Use: This research utilizes the publicly available SPAS-Dataset-BD from Mendeley Data Repository. The dataset is published under an open license permitting academic research use.

No Personal Data: The dataset contains only aggregated district-level agricultural statistics. No individual farmer information, personally identifiable data, or sensitive personal data is included.

Data Provenance: Dataset creators obtained data from official Bangladesh Bureau of Statistics publications, ensuring data collection followed appropriate governmental protocols.

C.2.2 Algorithmic Fairness

District Equity: Models were evaluated for performance consistency across all 64 districts to ensure predictions do not systematically disadvantage specific geographical regions.

Crop Diversity: The V2 approach specifically addresses fairness concerns by using crop-relative yield targets, preventing models from inherently favoring high-production crops over specialty crops important for dietary diversity.

Accessibility: Model complexity balanced against deployment feasibility in resource-constrained settings. Computational requirements kept minimal (≤ 1 minute training time, ≤ 500 MB memory).

C.2.3 Potential Impact and Responsible Use

Beneficial Applications:

- Agricultural planning and resource allocation
- Early warning systems for food security
- Evidence-based policy formulation
- Farmer advisory services

Limitations Acknowledged:

- Predictions are probabilistic and should inform, not replace, human judgment
- Model trained on historical data may not fully capture future climate change impacts
- District-level aggregation masks within-district heterogeneity
- Recommendations should be validated with local agricultural extension experts

Misuse Prevention: Research clearly documents model limitations and uncertainty to prevent over-reliance on predictions. Emphasis on decision support rather than autonomous decision-making.

D Dataset Description and Access

D.1 Dataset Information

Dataset Name: SPAS-Dataset-BD (Statistical analysis and Predictive Analytics for Smart agriculture in Bangladesh)

Source: Mendeley Data Repository

URL: <https://data.mendeley.com/datasets/cphdw4z5kw/2>

D.2 Dataset Characteristics

Temporal Coverage: 2010-2020 (aggregated seasonal data)

Spatial Coverage: All 64 districts across 8 administrative divisions of Bangladesh

Crops Included (73 total): 3

- Rice varieties: Aman, Aus, Boro
- Cereals: Wheat, Maize, Cheena
- Pulses: Arhar, Lentil, Mung
- Oilseeds: Mustard, Groundnut
- Vegetables: 30+ varieties
- Fruits: 15+ varieties
- Spices: Chili, Turmeric, etc.

Growing Seasons:

- **Rabi:** November-March (dry winter season)
- **Kharif-1:** April-June (pre-monsoon summer)
- **Kharif-2:** July-October (monsoon rainy season)

D.3 Variable Definitions

Table 26: SPAS-Dataset-BD Variable Definitions

Variable	Definition	Unit/Type
District	Administrative district name	Categorical (64)
Season	Growing season identifier	Categorical (3)
Crop Name	Crop type/variety	Categorical (73)
Production	Total production quantity	Metric Tons (MT)
Area	Total cultivation area	Hectares (ha)
AP Ratio	Production per unit area (yield)	MT/ha
Avg Temp	Seasonal average temperature	Degrees Celsius
Max Temp	Seasonal maximum temperature	Degrees Celsius
Min Temp	Seasonal minimum temperature	Degrees Celsius
Avg Humidity	Seasonal average relative humidity	Percentage
Max Relative Humidity	Seasonal maximum RH	Percentage
Min Relative Humidity	Seasonal minimum RH	Percentage
Transplant	Transplanting month	Categorical
Growth	Growing month identifier	Categorical
Harvest	Harvesting month	Categorical