Second Year (Semester-3) Research Assignment on

# *SMOTE-Enhanced Machine Learning Models for Crop Prediction in Imbalanced Multiclass Data*

in partial fulfilment of the requirement for the successful completion of semester 2 of MSc Big Data Analytics

Submitted By
24-PBD-045
Krishna Saverdekar
(Semester – III MSc. BDA)

Under the supervision of
*Prof. Hemal Desai*

2025-2026

Department of Computer Sciences (MSc. BDA)
St. Xavier's College (Autonomous) Ahmedabad – 380009

# DECLARATION

I, the undersigned solemnly declare that the research assignment *SMOTE-Enhanced Machine Learning Models for Crop Prediction in Imbalanced Multiclass Data* is based on my work carried out during the course of our study under the supervision of *Prof. Hemal Desai*. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

- The work contained in the report is original and has been done by me under the general supervision of my supervisor.
- The work has not been submitted to any other Institution for any other degree / diploma / certificate in this university or any other University of India or abroad.
- We have followed the guidelines provided by the department in writing the report.

Krishna Saverdekar

24-PBD-045

MSc. BDA (Big Data Analytics)

St. Xavier's College (Autonomous), Ahmedabad

# TABLE OF CONTENTS

# 1. Abstract

Precision agriculture increasingly relies on machine learning (ML) for crop recommendation, yet the practical utility of many models is limited by a critical challenge: inherent class imbalance in agricultural datasets. Standard ML algorithms trained on such data often become biased towards majority crops, leading to inaccurate and inequitable predictions for minority classes. This research addresses this gap by proposing and evaluating a robust framework centered on the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate the effects of imbalanced multiclass data. In this research, methodology involves systematic data preprocessing and transformation, followed by the application of SMOTE to balance the class distribution of a comprehensive crop dataset. A comparative analysis of high-performance supervised learning models, including Random Forest and XGBoost, on both the original imbalanced data and the SMOTE-enhanced balanced data. This study validates that integrating SMOTE is a crucial step for developing accurate, fair, and generalizable crop prediction systems, thereby offering a scalable solution that enhances the reliability of data-driven decision-making in modern agriculture.

# 2. Introduction

Agriculture, a vital pillar of the global economy, faces unprecedented challenges from climate change, land degradation, and inefficient resource utilization [4][6]. These pressures often compel farmers to make sub-optimal cultivation choices, adversely affecting productivity. In response, Precision Agriculture (PA) has emerged, leveraging Artificial Intelligence (AI) and Machine Learning (ML) to offer data-driven solutions [2][8]. By analysing complex variables such as soil properties (e.g., pH, Nitrogen, Phosphorus, Potassium) and climatic conditions (e.g., temperature, humidity, rainfall), ML models can provide powerful decision support, most notably in recommending suitable crops for specific environments [3][8].

To bridge these gaps, this research proposes and evaluates a robust framework of SMOTE-Enhanced Machine Learning Models for Crop Prediction in Imbalanced Multiclass Data. Our methodology is built on a systematic approach that begins with data preprocessing for eliminating outliers and Min-Max normalization is applied to scale features. The core of this contribution lies in addressing class imbalance through the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic data points for underrepresented crop categories [3][9]. This creates a balanced training dataset, enabling the models to learn the distinct features of all classes, thereby improving fairness and generalization. We then conduct a comparative analysis of high-performance algorithms, including XGBoost and Random Forest [8][9], to demonstrate the synergistic benefits of combining advanced models with the SMOTE balancing technique.

# 3. Review of Literature

## 3.1 Efficacy of Supervised Learning in Crop Prediction

Supervised ML algorithms are central to modern crop prediction. Studies consistently show that models can effectively translate soil and climatic data into actionable farming insights. Ensemble methods like Random Forest are frequently lauded for their high accuracy [2][5][8], while other classifiers such as Support Vector Machines (SVM), K-Nearest Neighbors (K-NN), and XGBoost have also been successfully benchmarked for this task [8][9]. This body of work confirms that variables like rainfall, temperature, humidity, and soil nutrient levels (N, P, K) are critical predictors for crop selection [4][6]. Several studies have further enhanced predictive power by utilizing multi-class ensemble approaches, demonstrating a clear consensus on the utility of supervised learning in this domain [7].

### 3.2 Limitations of Prior Models: Generalizability and Data Imbalance

Despite these successes, two significant limitations hinder the practical deployment of many existing models. First, many studies rely on localized or region-specific datasets, creating models that may not generalize well to different agro-climatic conditions [10]. Second, and more critically, agricultural datasets are often characterized by a natural class imbalance, where staple crops are heavily overrepresented compared to minority crops [3]. Standard ML algorithms trained on such skewed data develop a bias towards the majority class, leading to poor predictive performance for underrepresented crops [3][9]. This bias undermines the goal of providing equitable and diverse crop recommendations, a challenge explicitly identified by Iorzua et al. [3] as a major obstacle to building fair AI-driven agricultural systems.

### 3.3 SMOTE as a Solution for Class Imbalance

To address data imbalance, recent research has focused on data-level remediation techniques, with the Synthetic Minority Over-sampling Technique (SMOTE) emerging as a prominent solution. SMOTE rectifies class imbalances by generating synthetic data points for minority classes, creating a balanced training dataset that enables models to learn the distinct features of all crop types without bias [3][9]. Research by Sapkal and Kadam [9] demonstrates that applying SMOTE as a preprocessing step significantly improves the accuracy of various classifiers, including Random Forest and XGBoost. This finding is corroborated by Iorzua et al. [3], who successfully integrated SMOTE to build a fairer and more accurate recommendation framework. These studies establish that class balancing is not an optional tweak but a crucial step for developing robust and reliable agricultural prediction models. This research builds directly upon that conclusion by systematically evaluating the performance gains achieved by integrating SMOTE with high-performance ML models.

# 4. Objective, Data and Methodology

## 4.1. Objective

The objective of this study is to build and compare various supervised learning models such as Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Random Forest, and XGBoost, in order to identify the most effective approach for handling the combined challenges of class imbalance and multi-class classification. By incorporating appropriate data preprocessing methods, including transformations to address outliers and synthetic sampling techniques to balance underrepresented classes, the study seeks to evaluate the impact of these strategies on model performance. Ultimately, the goal is to generate reliable insights into crop suitability under diverse soil and climatic conditions, thereby contributing to more accurate, data-driven decision making in agriculture.

## 4.2. Data Source

The study uses a combined dataset of soil, climate, and crop production information to support machine learning–based crop recommendation. Soil data from the Ethiopian Agricultural Transformation Agency (ATA) includes geographic coordinates, pH, soil color, composition, electrical conductivity, and nutrients, with associated crop types primarily cereals. Climate data from NASA covers temperature, precipitation, humidity, wind, pressure, and cloud cover, while historical crop production is sourced from the Ethiopian Statistics Service (ESS).

## 4.3. Methodology

The methodological framework for this study is illustrated in Fig. 1. It follows a structured pipeline from data acquisition and preprocessing to model training and the final evaluation of classification results. Each stage was carefully designed to ensure the development of a robust and reliable crop recommendation model.

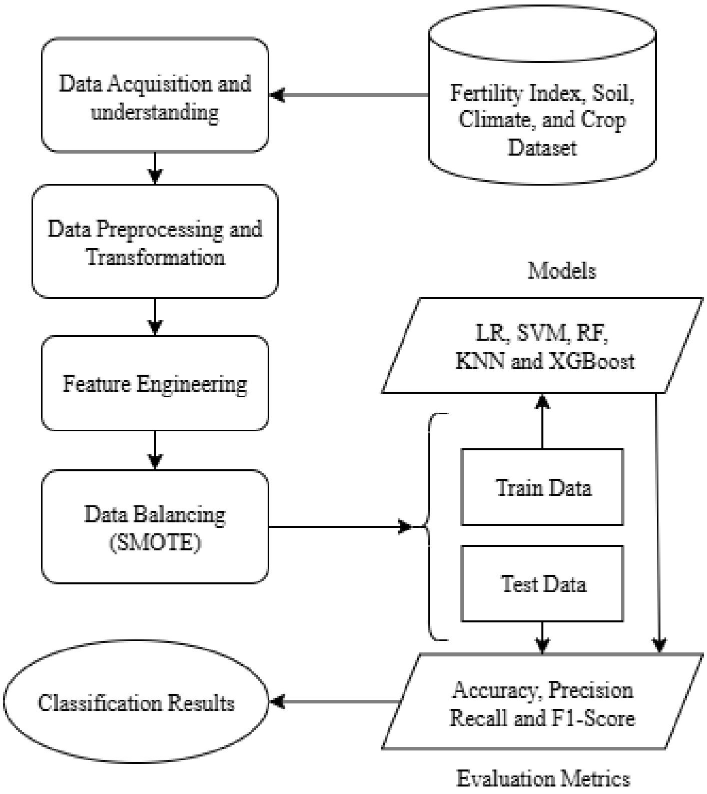
Data Acquisition and understanding
Fertility Index, Soil, Climate, and Crop Dataset
Data Preprocessing and Transformation
Models
LR, SVM, RF, KNN and XGBoost
Feature Engineering
Train Data
Data Balancing (SMOTE)
Test Data
Classification Results
Accuracy, Precision Recall and F1-Score
Evaluation Metrics

**Fig. 1** The Proposed Methodological Workflow

### 4.3.1 Data Acquisition

The research is based on the Crop Recommendation using Soil Properties and Weather Prediction dataset, a comprehensive collection of agricultural and environmental data.

### 4.3.2 Data Preprocessing and Transformation

Initial exploratory data analysis revealed that several key numerical features exhibited significant skewness and deviated from a normal distribution. To address this, a logarithmic transformation was applied, reducing the influence of outliers and aligning the data more closely with statistical assumptions. Categorical variables such as Soil color were converted into numerical format using label encoding, assigning unique integers to each category to make the data machine-readable while preserving its categorical nature. Finally, to ensure that all features contributed equally during model training, Min-Max scaling was employed to normalize the feature values within a common range, typically between 0 and 1. This step was particularly crucial for distance-based models like KNN and gradient-based algorithms, preventing features with larger numeric ranges from disproportionately influencing the learning process.

### 4.3.3 Feature Engineering

A feature engineering phase was conducted to enhance the predictive capacity of the dataset by leveraging domain knowledge in agronomy and environmental science. This process focused on constructing composite features that encapsulate meaningful relationships among variables, thereby improving model interpretability and reducing dimensionality. Several of the original 29 features were mathematically combined into more informative aggregates such as merging temperature-related variables into a "Temperature" metric resulting in a refined feature set of 15 variables. This transformation not only streamlined the dataset but also improved training efficiency and reduced the risk of overfitting, ultimately contributing to more robust and generalizable model performance.

### 4.3.4 Data Balancing (SMOTE)

A critical challenge in the dataset was the imbalance among crop classes. For the K-Nearest Neighbors (KNN) models, the was used. This technique was applied only to the training data to generate synthetic samples for underrepresented crops, creating a balanced class distribution.

### 4.3.5 Train/Test Split

For robust and unbiased model evaluation, the dataset was stratified into a training set (80%) and a testing set (20%). The training set was used solely for model development, whereas the testing set was withheld until the final evaluation phase to objectively assess generalization performance on unseen data.

### 4.3.6 Models Used

Five supervised machine learning models were selected for this comparative study to represent a diverse spectrum of classification strategies. The linear models included Logistic Regression and Support Vector Machine (SVM), both of which construct decision boundaries based on linear separability. K-Nearest Neighbors (KNN) was chosen as a distance-based, non-parametric algorithm that classifies data points based on proximity to labeled instances. To capture non-linear relationships and enhance predictive robustness, two ensemble models Random Forest and XGBoost were employed. These models leverage multiple decision trees through bagging and boosting techniques, respectively, enabling them to handle complex patterns and interactions within the dataset.

### 3.7 Evaluation and Classification

The final phase of the study involved evaluating the trained models on the held-out test set to assess their generalization capability. Given the multi-class nature of the classification task, model performance was quantified using a suite of standard evaluation metrics. Accuracy was used to measure the overall proportion of correctly classified instances. Precision assessed the model's ability to avoid false positives, while recall (or sensitivity) evaluated its effectiveness in identifying all relevant instances of each class. To provide a balanced measure that accounts for both precision and recall, the F1-score was calculated as their harmonic mean. These metrics collectively offer a robust framework for comparing model effectiveness across imbalanced and multi-class scenarios.

## 5. Model Performance Analysis

This section presents a comprehensive evaluation of the five selected machine learning models: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and XGBoost. The analysis was conducted under two distinct conditions to systematically assess the impact of class imbalance on predictive performance. First, models were trained and tested on the original, imbalanced dataset. Second, the models were trained on a balanced dataset created using the Synthetic Minority Over-sampling Technique (SMOTE) and then evaluated on the original test set.

### 5.1 Performance on the Original Imbalanced Dataset

Initially, all models were trained on the unaltered dataset, which contained a significant class imbalance. The performance, measured by classification accuracy, is shown in Fig. 2.

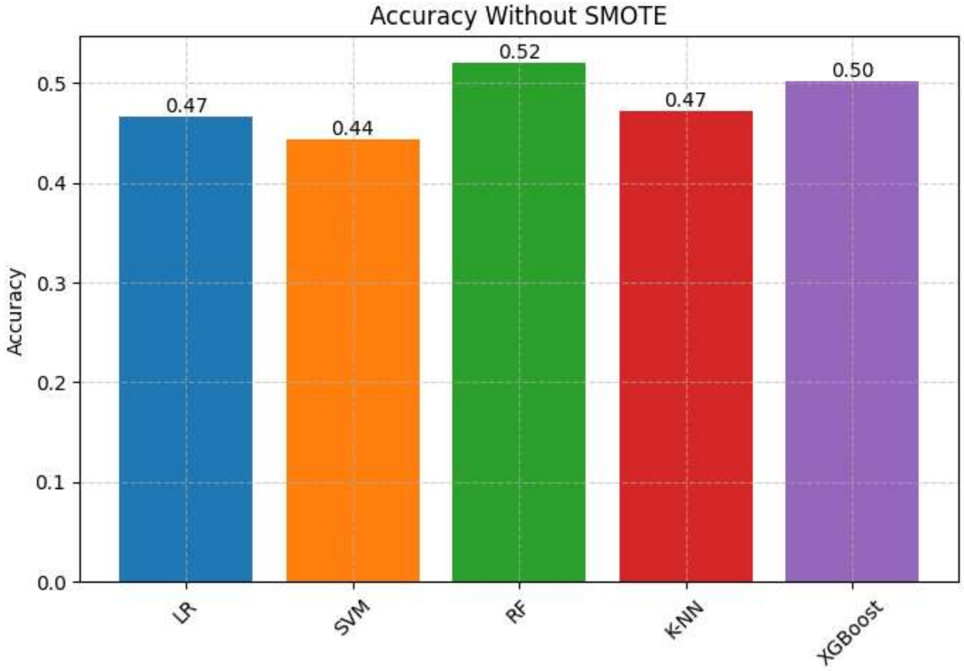

**Fig. 2** Model Accuracy on Imbalanced Data

The results indicate that the ensemble models demonstrated a performance advantage over linear and instance-based classifiers. Random Forest achieved the highest accuracy at 0.52, followed closely by XGBoost at 0.50. K-Nearest Neighbors and Logistic Regression both recorded an accuracy of 0.47, while the Support Vector Machine performed the poorest at 0.44.

While these scores provide a baseline, an accuracy of around 50% in a multi-class problem suggests that the models were struggling. This modest performance is a classic symptom of class imbalance, where models tend to become biased towards the majority classes, leading to poor predictive power for underrepresented crops. The slight superiority of the ensemble methods hints at their greater capacity to capture complex patterns, even in skewed data.

The MAE values further reinforced this trend shown in Figure 3. Random Forest recorded the lowest error (2.86), suggesting superior predictive precision. XGBoost followed with an MAE of 3.01, while KNN showed the highest error (3.13), indicating less reliable predictions. Both LR and SVM had comparable MAE scores around 3.01 and 3.03, respectively.

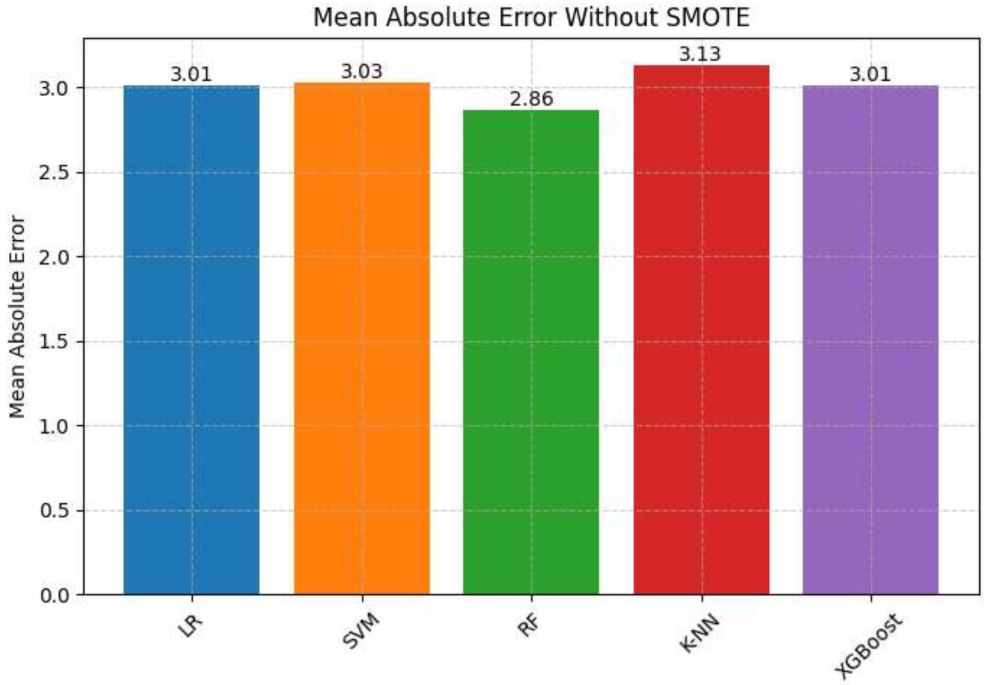

**Fig 3.** Mean Absolute Error on Imbalanced Data

### 5.2.2 The Transformative Impact of SMOTE on Model Performance

To address the issue of class imbalance, the SMOTE algorithm was applied to the training data to create a balanced class distribution. The models were then retrained and re-evaluated, with the results shown in Fig 4.

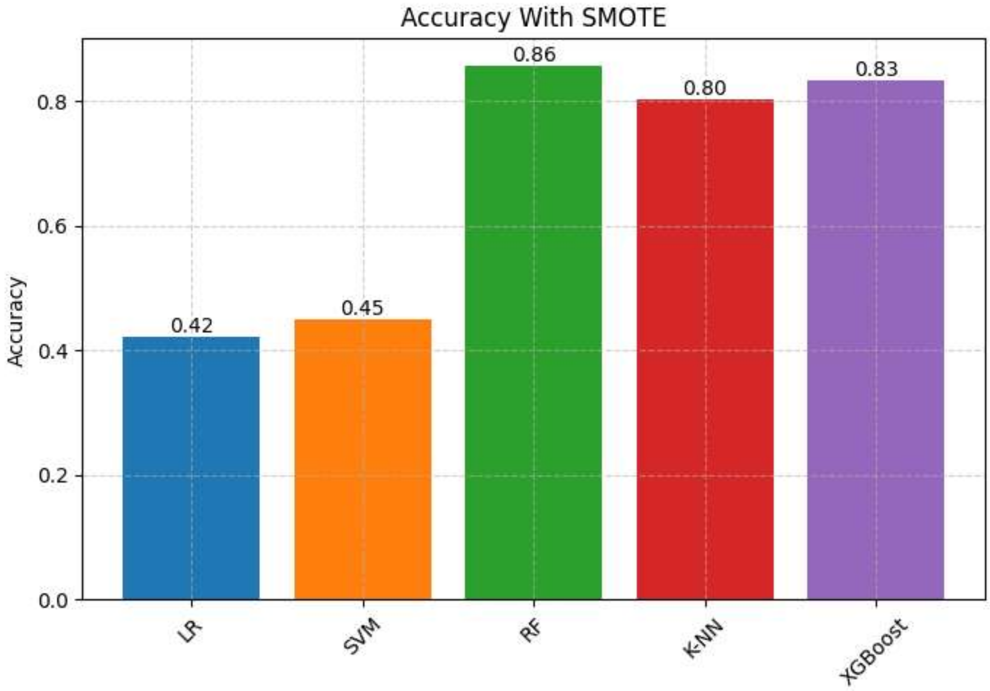

**Fig 4.** Model Accuracy on SMOTE-Balanced Data

The application of SMOTE yielded a dramatic and highly significant improvement, particularly for the non-linear and ensemble models.

Random Forest emerged as the top-performing model, with its accuracy surging from 0.52 to 0.86—a 65% improvement. XGBoost also saw a substantial gain, with its accuracy rising from 0.50 to 0.83. K-Nearest Neighbors benefited significantly as well, improving from 0.47 to 0.80.

The MAE values which are shown in Fig. 5 dropped substantially for the ensemble models. Random Forest achieved an MAE of just 0.79, the lowest among all models, followed by XGBoost (0.92) and KNN (1.07). In contrast, LR and SVM maintained higher error rates (2.43 and 2.45), indicating that SMOTE did not significantly improve their predictive accuracy.

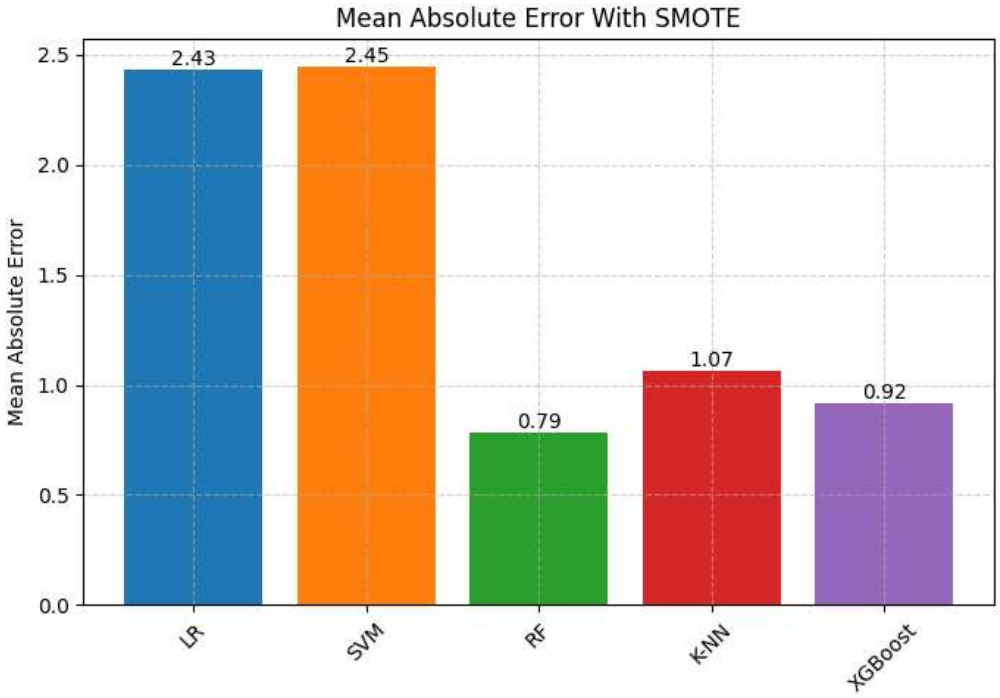

**Fig. 5** Mean Absolute Error on SMOTE-Balanced Data

This demonstrates that these models were highly effective at leveraging the balanced data to learn the distinguishing features of minority classes. By creating synthetic samples, SMOTE provided a richer, more equitable dataset, allowing the algorithms to build a more robust and generalizable decision boundary.

# 6. Findings and Conclusions

This study evaluated five supervised machine learning models for crop selection, revealing that class imbalance in the dataset is a primary barrier to predictive accuracy. The application of the Synthetic Minority Over-sampling Technique (SMOTE) was pivotal, dramatically improving performance, particularly for ensemble methods.

Random Forest emerged as the superior model, achieving an accuracy of 0.86 and a Mean Absolute Error (MAE) of 0.79 on the balanced dataset. XGBoost and KNN also saw significant gains, demonstrating their robustness for complex agricultural data once class distribution is addressed. In contrast, Logistic Regression and SVM failed to improve, indicating their unsuitability for the non-linear relationships within the data.

The central conclusion is that an ensemble approach specifically Random Forest combined with SMOTE provides the most accurate and reliable framework for this predictive crop recommendation task. This work contributes a validated methodology for handling imbalanced agricultural data, a common limitation in previous studies. Future research should focus on exploring deep learning models, implementing advanced feature selection, and integrating real-time environmental data to develop more dynamic and precise recommendation tools.

# 7. References

[1] Alemu, S. (2024). *Crop recommendation using soil properties and weather prediction dataset* [Dataset]. Mendeley Data. https://doi.org/10.17632/8V757RR4ST.1

[2] Champaneri, M., Chachpara, D., Chandvidkar, C., & Rathod, M. (2018). *Crop yield prediction using machine learning* (2319 7064). International Journal of Science and Research. https://doi.org/10.21275/SR20402185927

[3] Iorzua, J. T., Kwaghtyo, D. K., Hule, T. P., Ibrahim, A. T., & Nongu, A. D. (2025). AI-Driven Approach to Crop Recommendation: Tackling class imbalance and feature selection in precision agriculture. *Journal of Future Artificial Intelligence and Technologies*, *2*(2), 269–281. https://doi.org/10.62411/faith.3048-3719-118

[4] Kalimuthu, M., Vaishnavi, P., & Kishore, M. (2020). Crop Prediction using Machine Learning. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 926–932. https://doi.org/10.1109/icssit48917.2020.9214190

[5] Kumar, Y. J. N., Spandana, V., Vaishnavi, V., Neha, K., & Devi, V. (2020). Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector. *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, 736–741. https://doi.org/10.1109/icces48766.2020.9137868

[6] Lad, A. M., Bharathi, K. M., Saravanan, B. A., & Karthik, R. (2022). Factors affecting agriculture and estimation of crop yield using supervised learning algorithms. *Materials Today Proceedings*, *62*, 4629–4634. https://doi.org/10.1016/j.matpr.2022.03.080

[7] Meenachi, L., Ramakrishnan, S., Sivaprakash, M., Thangaraj, C., & Sethupathy, S. (2022). Multi class ensemble classification for Crop recommendation. *2022 International Conference on Inventive Computation Technologies (ICICT)*, 1319–1324. https://doi.org/10.1109/icict54344.2022.9850561

[8] Patel, K., & Patel, H. B. (2021). A Comparative analysis of Supervised Machine Learning Algorithm for agriculture crop Prediction. *2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, *349*, 1–5. https://doi.org/10.1109/icecct52121.2021.9616731

[9] Sapkal, K. G., & Kadam, A. B. (2025). Class Balancing for soil data: Predictive Modeling Approach for crop recommendation using Machine learning algorithms. *EPJ Web of Conferences*, *328*, 01026. https://doi.org/10.1051/epjconf/202532801026

[10] Shakoor, M. T., Rahman, K., Rayta, S. N., & Chakrabarty, A. (2017). *Agricultural production output prediction using supervised machine learning techniques* (No. 978-1-5386-3831-6/17; pp. 182–187). Department of Computer Science and Engineering BRAC University Dhaka, Bangladesh. https://doi.org/10.1109/nextcomp.2017.8016196

[11] Suganya, M., R, D., & R, R. (2020). Crop yield prediction using supervised learning techniques. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID3639103_code4086546.pdf?abstractid=3639103&mirid=1