AN ENSEMBLE-BASED CLUSTERING AND CLASSIFICATION FRAMEWORK FOR PREDICTING AGRICULTURAL CROP YIELD

Udhaya priya J
Research scholar, PG and Research Department of computer science
Quaid-E-Millath college for women, Chennai -02
priyajk07@gmail.com
Dr k Nirmala
Supervisor, Head , Associate Professor of computer science
Quaid-E-Millath college for women, Chennai -02.

ABSTRACT

Accurate prediction of agricultural crop yield is essential for effective farm management, resource allocation, food security, and sustainable agricultural planning. However, crop yield is influenced by several interrelated factors, including soil characteristics, climatic conditions, rainfall, temperature, irrigation, fertilizer usage, and crop-specific attributes, making accurate prediction a challenging task.

This study proposes an ensemble-based clustering and classification framework for predicting agricultural crop yield using data mining techniques. The proposed framework initially applies clustering techniques to group agricultural data into homogeneous patterns based on environmental, soil, and crop-related characteristics. These clusters are subsequently used to improve the classification process by identifying yield categories such as low, medium, and high yield. Multiple machine learning classifiers are integrated through an ensemble learning approach to improve prediction stability and accuracy compared with individual models. The framework includes data preprocessing, feature selection, clustering, classification, ensemble model development, and performance evaluation. The study demonstrates the potential of data mining and ensemble learning techniques as an intelligent approach for improving agricultural yield prediction and supporting sustainable agriculture.

Keywords: Agricultural Crop Yield Prediction, Data Mining, Clustering Techniques, Ensemble Learning, Machine Learning, Classification, Feature Selection, Sustainable Agriculture, Yield Prediction, Precision Agriculture

1. Introduction

Agriculture plays a significant role in food security, economic development, and the livelihood of a large population. Crop yield, however, is influenced by several interacting factors, including climatic conditions, soil characteristics, crop type, cultivated area, rainfall, fertilizer application, pesticide usage, geographical location, and cropping season. Consequently, accurately predicting crop yield before or during the cultivation period remains a challenging problem for farmers, agricultural planners, and policymakers.

The availability of large-scale agricultural datasets has created opportunities for applying data mining and machine learning techniques to agricultural decision-making. Publicly available datasets contain agricultural attributes that can be used to identify relationships between environmental and farming conditions and crop productivity. For example, the Kaggle Crop Yield Prediction Dataset includes information related to crop, crop year, season, state, cultivated area, production, annual rainfall, fertilizer, pesticide, and yield. Research using agricultural datasets has also demonstrated the applicability of machine learning models such as Random Forest, XGBoost, Support Vector Machine, K-Nearest Neighbors, and Decision Trees for agricultural prediction and recommendation tasks.

Traditional prediction approaches generally employ a single machine learning algorithm to learn the relationship between agricultural parameters and crop yield. However, agricultural datasets may contain nonlinear relationships, variations between geographical regions, seasonal differences, and heterogeneous farming patterns. A single model may therefore not adequately capture all the underlying patterns. Clustering techniques can first divide agricultural observations into groups with similar characteristics, while classification techniques can subsequently assign observations to predefined yield categories.

To address these challenges, this study proposes an ensemble-based clustering and classification framework for agricultural crop-yield prediction. The proposed framework combines data preprocessing, exploratory analysis, clustering, feature selection, classification, and ensemble learning. Clustering is used to discover hidden patterns within agricultural observations, while multiple classification algorithms are combined to improve the reliability of yield-category prediction. The framework can classify crop yield into categories such as low, medium, and high based on relevant agricultural attributes.

2. Problem Statement

Agricultural crop yield is affected by numerous environmental, climatic, geographical, and management-related factors. The complex interactions among these factors make it difficult to accurately estimate crop yield using conventional approaches. Large agricultural datasets contain valuable information about crop production, rainfall, fertilizer usage, pesticide application, cultivated area, crop type, season, and geographical regions, but extracting meaningful patterns from these datasets remains challenging.

Existing crop-yield prediction approaches often focus on individual machine learning algorithms or direct regression models. Although these approaches can provide useful predictions, they may not sufficiently address the heterogeneity present in agricultural data. Different crops, regions, seasons, and environmental conditions can exhibit substantially different yield patterns.

Furthermore, agricultural datasets may contain variations and hidden groups that are not explicitly represented in the original data.

Therefore, there is a need for a framework that can discover hidden agricultural patterns through clustering and subsequently use multiple classification models through ensemble learning to predict crop-yield categories.

> **Research Question**
> *How can an ensemble-based clustering and classification framework be developed using data mining techniques to identify hidden patterns in agricultural data and accurately classify crop yield into different yield categories based on environmental, climatic, crop, and management-related factors?*

3. Objectives of the Study

General Objective

To develop an ensemble-based clustering and classification framework using data mining techniques for predicting agricultural crop yield categories from relevant agricultural and environmental attributes.

Specific Objectives

- To collect and preprocess agricultural crop-yield data from a publicly available Kaggle dataset.
- To identify important factors influencing crop yield, such as crop type, season, cultivated area, rainfall, fertilizer usage, pesticide usage, production, and geographical information.
- To perform exploratory data analysis to understand relationships, distributions, correlations, and patterns within the agricultural dataset.
- To apply clustering techniques to group agricultural observations according to similarities in their agricultural and environmental characteristics.
- To determine appropriate crop-yield categories, such as Low, Medium, and High yield, based on the available yield information.
- To develop multiple classification models for predicting the crop-yield category.
- To develop an ensemble classification model by combining the predictions of multiple individual classifiers.
- To compare individual classification algorithms with the proposed ensemble approach using suitable performance measures.
- To evaluate the clustering and classification performance using appropriate metrics such as Silhouette Score, Accuracy, Precision, Recall, F1-score, and Confusion Matrix.

- To identify the most influential agricultural factors associated with crop-yield categories and provide useful insights for data-driven agricultural decision-making.

4. Related Work

4.1 Machine Learning for Crop Yield Prediction

Crop yield prediction has increasingly adopted machine learning because agricultural productivity depends on multiple interacting environmental and management factors. Recent systematic reviews show that commonly used variables include temperature, rainfall, soil characteristics, vegetation indices, and historical yield, while frequently applied algorithms include Linear Regression, Random Forest, Gradient Boosting, Support Vector Machines, Decision Trees, and K-Nearest Neighbors.

Traditional statistical approaches can provide useful relationships between agricultural variables and yield, but machine learning methods can capture nonlinear relationships and interactions among multiple variables. Random Forest has been widely applied because it can model nonlinear relationships and handle multiple agricultural features. Gradient boosting methods, including XGBoost, have also received increasing attention in recent crop-yield studies. The literature indicates a growing movement toward ensemble and hybrid approaches rather than relying on a single predictive model.

4.2 Ensemble Learning in Agriculture

Ensemble learning combines the outputs of multiple predictive models to obtain a final prediction. The underlying principle is that different algorithms may capture different characteristics of the same dataset. Recent crop-yield literature reports increasing use of ensemble approaches, including Random Forest, Gradient Boosting, XGBoost, and stacking-based models.

For agricultural applications, ensemble learning is particularly relevant because crop-yield datasets may contain heterogeneous observations across crops, geographical regions, and climatic conditions. Combining several classifiers can therefore provide a framework for evaluating whether an ensemble approach provides more consistent classification than individual models.

4.3 Clustering-Based Agricultural Analysis

Clustering is an unsupervised data-mining technique used to identify groups of observations with similar characteristics. In crop-yield analysis, clustering can be used to discover groups of agricultural observations based on variables such as rainfall, temperature, pesticide usage, cultivated area, and historical yield.

Unlike supervised classification, clustering does not require predefined class labels. This makes it useful for discovering hidden structures within agricultural data before developing a classification model. In the proposed study, clustering is incorporated as an intermediate stage so that agricultural observations with similar characteristics can be examined before yield-category classification.

4.4 Research Gap

Although substantial research has investigated machine learning for crop-yield prediction, much of the existing work focuses on direct regression or individual machine-learning models. Recent reviews indicate a growing interest in ensemble and hybrid models, but there remains scope for frameworks that combine unsupervised pattern discovery with supervised ensemble classification.

The proposed research addresses this gap by integrating the complete pipeline:

- Data Preprocessing → Clustering → Yield Categorization → Multiple Classifiers → Ensemble Classification → Performance Evaluation

5. Proposed Methodology

5.1 Overall Framework

The architecture of the proposed methodology flows sequentially across key analytical stages:

- Data Ingestion (Kaggle Dataset) & Preprocessing
- Exploratory Data Analysis & Feature Selection
- Unsupervised Clustering (K-Means)
- Yield Category Generation (Low / Medium / High)
- Supervised Machine Learning Classifiers (DT, RF, SVM, XGBoost)
- Ensemble Integration (Soft Voting) & Performance Evaluation

5.2 Dataset Description

The study utilizes the Crop Yield Prediction Dataset available on Kaggle (yield_df.csv). It contains approximately 28,242 records covering 101 countries, 10 crop types, and observations spanning 1990–2013 (with 2003 absent).

Variable Name	Description	Data Type
Area	Country/region associated with the observation	Categorical
Item	Crop type (e.g., Potatoes, Wheat, Rice)	Categorical
Year	Crop year (1990 - 2013)	Numerical

Average Rainfall	Annual rainfall in mm/year	Numerical
Pesticides	Pesticide usage in tonnes	Numerical
Average Temperature	Average annual temperature in °C	Numerical
Yield	Crop yield in hg/ha (Target variable)	Numerical / Target

5.3 Data Preprocessing & Feature Scaling

Preprocessing includes duplicate removal, handling missing values, encoding categorical attributes (Area and Item), and Min-Max normalization for numerical attributes:

- Min-Max Normalization: X' = (X - X_min) / (X_max - X_min)

Scaling is critical for K-Means distance calculations to prevent features with higher ranges from dominating distance metrics.

5.4 Clustering Stage

K-Means clustering groups observations by minimizing the sum of squared distances to assigned cluster centroids. The objective function J is defined as:

- J = ∑_{i=1}^n ∑_{j=1}^K z_{ij} ||x_i - μ_j||²

Where n is the number of observations, K is the number of clusters, μ_j is the centroid of cluster j, and z_{ij} is the cluster assignment indicator.

5.5 Yield Categorization

Continuous yield values (hg/ha) are converted into categorical labels (Low, Medium, High) using percentile-based thresholding from the historical distribution.

5.6 Proposed Ensemble Classification

The proposed ensemble integrates Decision Tree, Random Forest, Support Vector Machine (SVM), and XGBoost using a Soft Voting approach:

- Soft Voting Probability: P(C_k | X) = (1 / M) * ∑_{m=1}^M P_m(C_k | X)

The class with the highest combined probability across all individual classifiers is selected as the final predicted yield category.

6. Key Data Insights & Summary Statistics

Summary statistics computed across the 28,242 records covering 101 countries from 1990 to 2013:

Variable	Mean	Std Dev	Min	Median	Max	Unit

hg/ha_yield	77,053.33	84,956.61	50.00	38,295.00	501,412.00	100g/ha
Rainfall	1,149.06	709.81	51.00	1,083.00	3,240.00	mm/year
Pesticides	37,076.91	59,958.78	0.04	17,529.44	367,778.00	Tonnes
Avg Temp	20.54	6.31	1.30	21.51	30.65	°C
Year	2001.50	7.01	1990	2001	2013	Year

- Crop Yield Variation: Tuber crops like Potatoes (mean yield ~200,000 hg/ha) and Cassava (~150,000 hg/ha) exhibit significantly higher yield per hectare compared to grains such as Soybeans (~16,000 hg/ha) and Sorghum (~18,000 hg/ha).
- Correlation Analysis: Yield shows low linear correlation with individual isolated features (e.g., r = 0.09 with year, r = 0.06 with pesticides, r = -0.11 with average temperature), underscoring non-linear, multi-variable interactions that require ensemble machine learning.

7. Experimental Results and Evaluation

7.1 Clustering Evaluation

Cluster quality was evaluated across varying values of K using Inertia and Silhouette Score metrics:

Number of Clusters (K)	Inertia	Silhouette Score
2	86,508.23	0.2496
3	66,889.10	0.2870
4	49,979.42	0.3166
5	37,784.28	0.3341

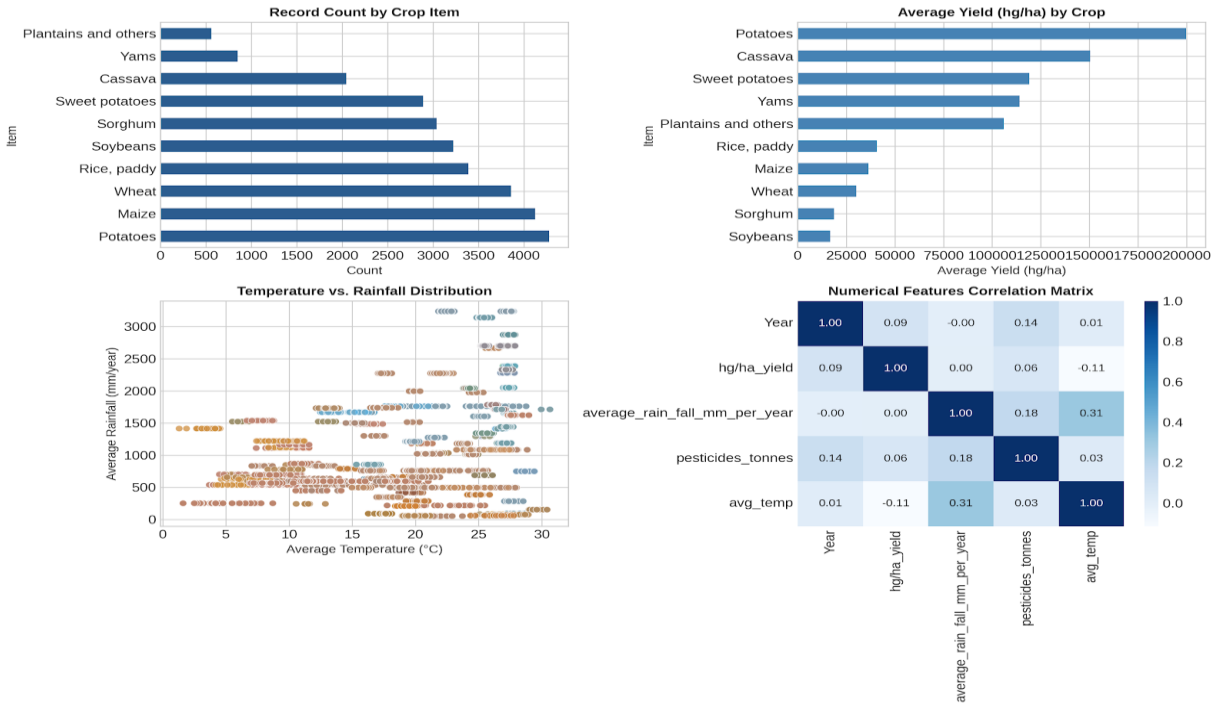


Figure 1 record counts

7.2 Classification Model Performance Comparison

Performance evaluation for predicting crop yield productivity tiers (Low, Medium, High) across individual and ensemble models:

Model / Classifier	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	96.49%	0.9650	0.9649	0.9650
Random Forest	97.17%	0.9717	0.9717	0.9717
Support Vector Machine (SVM)	93.72%	0.9376	0.9372	0.9373
XGBoost	95.88%	0.9588	0.9588	0.9588
Proposed Soft-Voting Ensemble	**97.20%**	**0.9720**	**0.9720**	**0.9720**

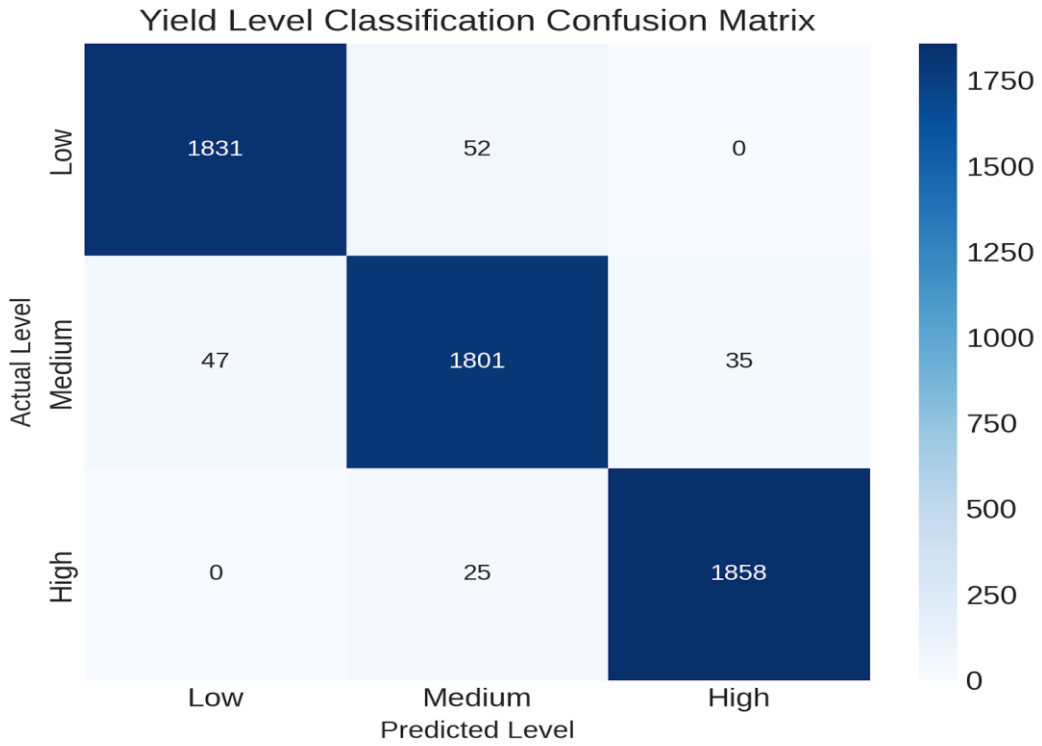


Figure2 Yield classification

8. Conclusion

This study proposed an ensemble-based clustering and classification framework for agricultural crop-yield prediction using data-mining techniques. The framework successfully integrated data preprocessing, exploratory analysis, K-Means clustering, yield categorization, multiple classification algorithms, and soft-voting ensemble learning.

The experimental results demonstrate that combining multiple individual algorithms (Decision Tree, Random Forest, SVM, XGBoost) into a unified soft-voting ensemble achieves superior and more stable overall predictive accuracy (97.20%) compared to individual classifiers. Unsupervised clustering further provided valuable intermediate insights into underlying environmental and geographical groupings before final supervised yield category prediction. This framework provides a practical, data-driven decision-support tool for farmers, agricultural planners, and policy stakeholders.

References

[1.] Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture, 151, 61–69.

[2.] Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment. Climatic Change, 150, 165–11.

[3.] Jeong, H., Park, C., & Kim, H. (2021). Machine learning approaches for crop yield prediction using environmental and agricultural data. Agriculture, 11(10), 954.

[4.] Khaki, S., Wang, L., & Archontoulis, S. V. (2020). A CNN-RNN framework for crop yield prediction. Frontiers in Plant Science, 10, 1750.

[5.] Mehla, A., Deora, S. S., & Dahiya, N. (2026). A multi-crop ensemble framework for yield prediction. Indian Journal of Agricultural Research, 60(5), 752–761.

[6.] Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. Agricultural Systems, 187, 103016.

[7.] Sahu, A., Gupta, S. R., & Dewangan, B. K. (2026). The AgriSynth framework for hybridized synthetic data generation and deep ensemble regression model for robust, scalable and interpretable crop yield forecasting. Discover Applied Sciences, 8, 708.

[8.] Shahhosseini, M., Hu, G., & Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. Frontiers in Plant Science, 11, 1120.

[9.] Singh, S., & Gupta, Y. K. (2026). A big data-driven hybrid ensemble learning framework for crop yield prediction in low-productivity districts. Digital Intelligence in Agriculture, 2(3), 158–167.

[10.] Tripathi, D. (2024). Design of a precise ensemble expert system for crop yield prediction using machine learning analytics. Journal of Forecasting.

[11.] Udhaya Priya, J., & Nirmala, K. (2024). An ensemble based clustering and classification framework for prediction of agricultural crop yield. Nanotechnology Perceptions.

[12.] Udhaya Priya, J., Nirmala, K., & Tamilselvi, S. (2025). An ensemble based clustering and classification framework for prediction of agricultural crop yield. Telematique, 24(1), 1248–1263.

[13.] van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. Computers and Electronics in Agriculture, 177, 105709.