Comparative Analysis of Machine Learning Models for Crop Type and Yield Prediction

Bagusetty Gayatri, 2024057286, [gbagusetty@student.gitam.edu](mailto:gbagusetty@student.gitam.edu), GITAM University, VSKP

|  |  |  |
| --- | --- | --- |
| Keywords |  | Abstract |
| Crop Type Classification  Yield Prediction  Machine Learning  Model Evaluation  Smart Agriculture  Supervised Learning  Comparative Analysis |  | Agriculture plays a vital role in food security and economic growth, yet farmers often face uncertainty in selecting suitable crops and estimating yields due to changing climatic conditions, soil variability, and resource constraints. This study applies machine learning techniques to enhance agricultural decision-making by addressing two key tasks: crop type classification and crop yield prediction. A dataset of nearly 20,000 records, containing attributes such as crop year, state, season, area, production, rainfall, fertilizer, and pesticide use, was pre-processed, encoded, and analyzed to uncover patterns. Multiple models were trained and evaluated, with Random Forest and XGBoost achieving the highest classification accuracies of 67% and 66%, respectively. For yield prediction, Gradient Boosting provided the best results with the lowest mean squared error (43,055), followed by Random Forest. These findings highlight the superiority of ensemble methods in handling complex agricultural data compared to linear and single-tree models. The outcomes of this work demonstrate the potential of machine learning to support smarter agricultural planning, resource optimization, and sustainable farming practices. |

**1. Introduction**

Agriculture is the backbone of food security and economic growth, but it faces challenges due to changing climate, varying soil conditions, and resource limitations. Farmers often struggle to decide which crops to grow and how much yield to expect, making accurate predictions highly valuable. Traditional methods rely on experience and manual analysis, which can be uncertain and time-consuming. This project applies machine learning to improve decision-making in agriculture by predicting crop types and estimating yields. Using historical data that includes rainfall, fertilizers, pesticides, and production details, multiple ML models are trained and evaluated to deliver more reliable and efficient predictions.

**2. Methodology**

**2.1 Data Collection and Preprocessing**

**2.1.1 Data Sources:** The dataset used in this study was obtained from the Omdena repository[[1]](https://datasets.omdena.com/dataset/crop-yield-prediction). It contains 19,689 records with features such as crop type, season, state, area, fertilizer, pesticide, production, and yield. The dataset covers 55 different crops across 30 Indian states, making it diverse enough for classification and regression tasks.

**2.1.2 Preprocessing:**  The dataset was first inspected for missing values and duplicates, but none were found. Categorical attributes, including *Crop, State,* and *Season*, were transformed into numerical form using LabelEncoding. Numerical features such as *Area, Fertilizer,* and *Pesticide* were standardized using StandardScaler to ensure balanced feature scales. Outliers were detected through boxplots and histograms, while correlations among variables were visualized using a heatmap. Finally, the dataset was divided into training and testing subsets using the *train\_test\_split()* function to prepare it for model training.

**2.2 System Architecture**

The system works step by step to predict crop type and yield. First, the dataset is collected, which includes details like crop, season, state, rainfall, fertilizer, pesticide, and yield. Next, the data is cleaned and converted into a form that machine learning models can understand. This includes handling categories, scaling numbers, and checking for outliers.

After that, exploratory data analysis (EDA) is done to see patterns and relationships in the data through graphs and heatmaps. The prepared data is then used in two ways: classification models are trained to predict which crop is grown, and regression models are trained to estimate how much yield can be expected.

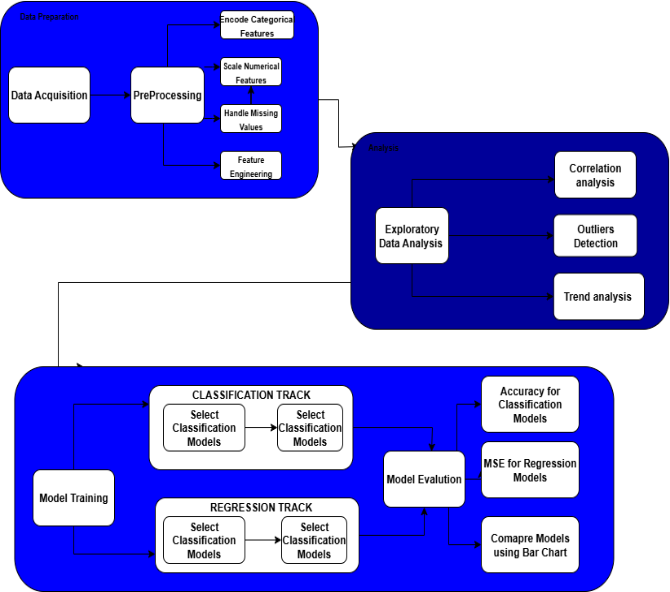


Figure 2.2 System Architecture

Finally, the models are tested and compared using accuracy for crop prediction and error rates for yield prediction. The results are shown with bar charts and graphs, making it easy to see which models perform best.

**2.3 Model Selection**

For this project, Two machine learning tasks were addressed:

* + 1. **Classification (Crop Type Prediction):**

SVC, Random Forest, Decision Tree, Logistic Regression, and XGBoost classifiers were trained to predict the type of crop grown.

* + 1. **Regression (Yield Prediction):**

SVR, Random Forest Regressor, Gradient Boosting, Linear Regression, and XGBoost regressors were employed to estimate crop yield.

**2.4 Model Training**

The dataset was split into training (80%) and testing (20%) sets to ensure fair model evaluation. Multiple machine learning models were trained for both classification and regression tasks. For crop type classification, models such as Support Vector Classifier (SVC), Random Forest, Decision Tree, Logistic Regression, and XGBoost were employed. For yield prediction (regression task), algorithms including Support Vector Regressor (SVR), Random Forest Regressor, Gradient Boosting, Linear Regression, and XGBoost were implemented. Hyperparameters such as learning rate, maximum depth, and number of estimators were tuned for better performance. Training was conducted using supervised learning techniques, where labelled crop and yield data guided the learning process.

**3. Results**

Random Forest and XGBoost achieved the best accuracy for crop classification, while Gradient Boosting and Random Forest Regressor gave the lowest errors in yield prediction, outperforming other models.

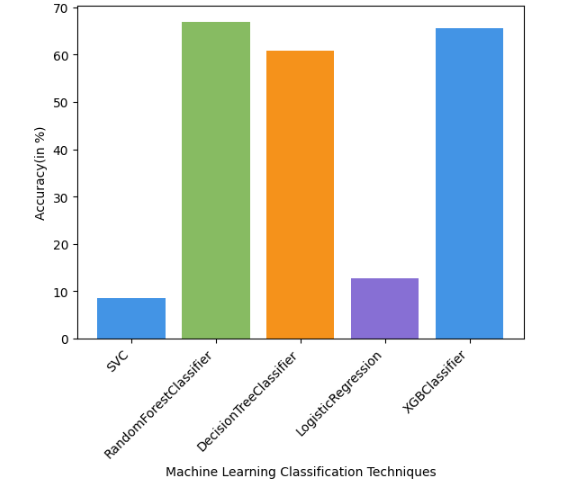


Figure 3.1 - Performance Evaluation of ML Classifiers

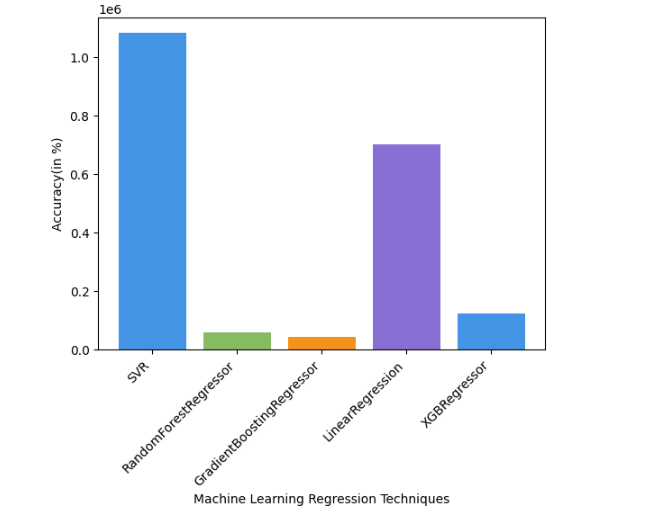


Figure 3.2 - Performance Comparison of Classifiers (Normalized Accuracy)

**3.1 Crop Type Classification**

Random Forest and XGBoost achieved the highest accuracies, while Logistic Regression and SVC performed poorly.

|  |  |
| --- | --- |
| Model | Accuracy (%) |
| Support Vector Classifier (SVC) | 8.53 |
| Random Forest Classifier | 66.89 |
| Decision Tree Classifier | 60.87 |
| Logistic Regression | 12.82 |
| XGBoost Classifier | 65.62 |

Table 3.1 - Crop Type Classification (Accuracy %)

**3.2 Yield Prediction**

Gradient Boosting Regressor outperformed other models with the lowest MSE, followed by Random Forest.

|  |  |
| --- | --- |
| Model | MSE |
| Support Vector Regressor (SVR) | 1,080,964 |
| Random Forest Regressor | 58,955 |
| Gradient Boosting Regressor | 43,055 |
| Linear Regression | 700,167 |
| XGBoost Regressor | 125,167 |

Table 3.2 - Yield Prediction (Mean Squared Error)

**4. Conclusion**

This study highlights the effectiveness of machine learning in crop type classification and yield prediction. Ensemble models such as Random Forest, Gradient Boosting, and XGBoost outperformed traditional methods. The findings demonstrate strong potential for data-driven agricultural planning. Future work will incorporate soil, climatic, and remote sensing data to improve prediction accuracy, scalability, and practical applications in smart farming.

**5. References**

[1] Omdena, *Crop Yield Prediction Dataset*, 2023. [Online]. Available: https://www.omdena.com/

[2] A. Badshah, B. Y. Alkazemi, F. Din, K. Z. Zamli, and M. Haris, “Crop Classification and Yield Prediction using Robust Machine Learning Models for Agricultural Sustainability,” *IEEE Access*, 2024.

[3] P. Mahesh and R. Soundrapandiyan, “Yield prediction for crops by gradient-based algorithms,” *PLoS ONE*, vol. 19, no. 8, Aug. 2024, Art. no. e0291928.

[4] M. Abdel-salam, N. Kumar, and S. Mahajan, “A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning,” *Neural Computing and Applications*, 2024.