Comparing the Performance of Classification Algorithms for Predicting Grain Yield Prediction

Gökay Çetinakdoğan 202111050

Hasan Emre Usta 202111301

[1] Çankaya University, Ankara, Türkiye
c2111050@student.cankaya.edu.tr, c2111301@student.cankaya.edu.tr

Abstract: This study aims to compare the performance of various classification algorithms for predicting grain yield using the "Data_processed.xlsx" dataset. The evaluated algorithms include Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Artificial Neural Networks (ANN), Random Forest, Gradient Boosting, AdaBoost, Support Vector Machines (SVM), and XGBoost. Data preprocessing involved handling missing values by imputing column means and converting categorical variables into numerical representations. Additionally, SelectKBest was employed for feature selection, identifying 'VarietyClass_SDV' and 'DaysFromZerotoSowing' as the most significant predictors. The performance of the models was assessed using metrics such as accuracy, AUC, F1-score, precision, recall, and Matthews Correlation Coefficient (MCC).

Keywords: Grain Yield Prediction, Classification Algorithms, Machine Learning, ROC Curve, Confusion Matrix

1 Introduction

Yield prediction in agricultural production is crucial for efficient resource management and sustainable farming practices. Grain yield is influenced by numerous factors, including soil properties, weather conditions, and planting practices. This study utilizes machine learning algorithms to predict grain yield and compares the performance of different algorithms. The objective is to identify the algorithm that provides the most accurate predictions and contributes to increased agricultural productivity.

2 Methodology

The following steps were taken for grain yield prediction in this study:

- **2.1 Data Preprocessing:**
  - Missing values were filled with the mean for numerical columns and zero for non-numerical columns.
  - Non-numeric characters in object-type columns were removed, and all columns were converted to numeric data.
  - The target variable GrainYield was encoded into three classes (A → 0, B → 1, C → 2).
  - Features were standardized using the StandardScaler to ensure uniformity across different scales.
  - Categorical features were converted to numerical values. (Applying one-hot encoding is recommended)

- **2.2 Feature Selection:**
  - The SelectKBest method with the f_classif scoring function was used to select the top two most significant features for grain yield prediction.

- **2.3 Modeling:**
  - Multiple machine learning algorithms were applied, including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Artificial Neural Network (ANN), Random Forest, Gradient Boosting, AdaBoost, Support Vector Machine (SVM), and XGBoost.

- **2.4 Evaluation:**
  - The performance of each model was assessed using metrics such as Accuracy, Precision, Recall, F1 Score, Matthews Correlation Coefficient (MCC), and ROC-AUC Score.
  - ROC curves were plotted for probabilistic classifiers to visualize their performance.
  - A confusion matrix was generated for the best-performing model to provide further insight into its classification performance.

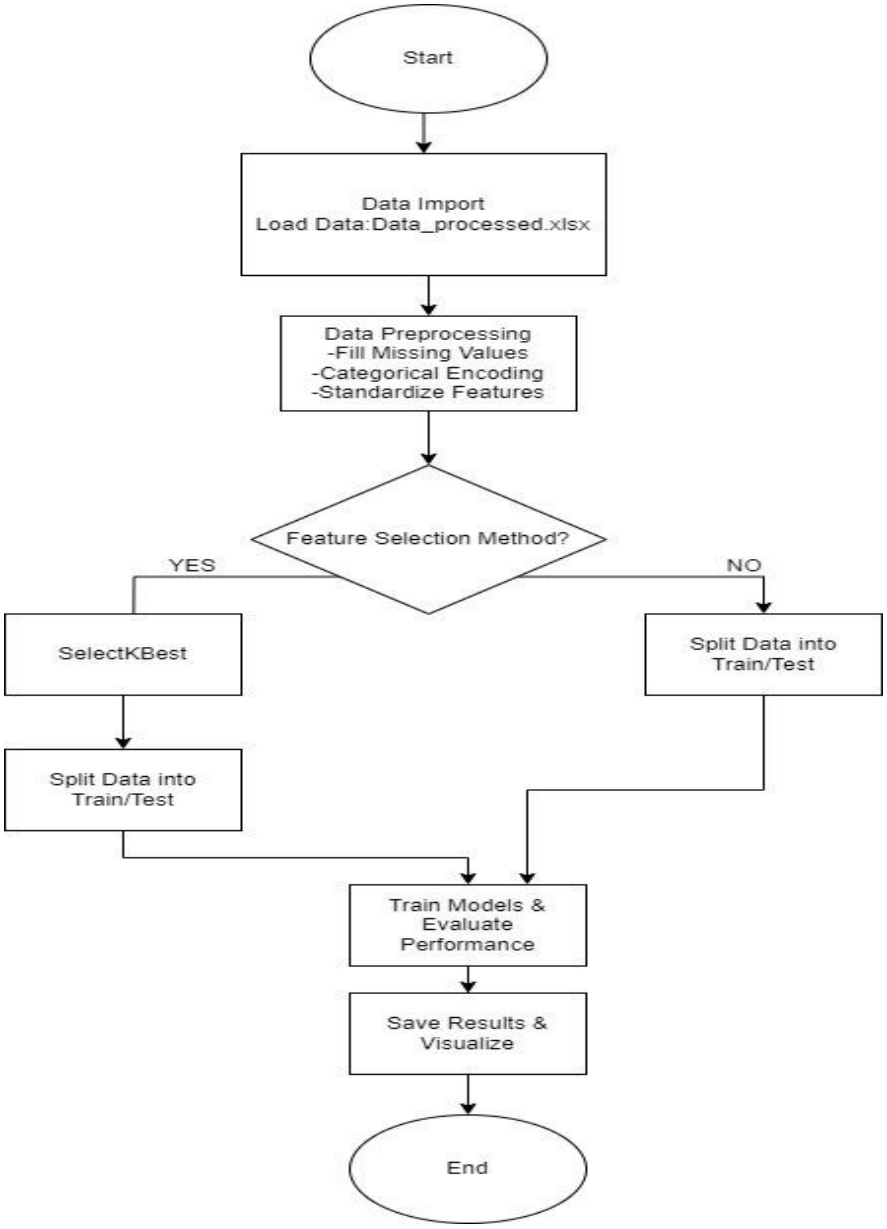

Fig. 1. flowchart.

The workflow illustrated in Figure 1 summarizes the steps followed during the project, starting from data import and preprocessing to model training and evaluation. Below is a detailed explanation of each step:

Data Import: The dataset (Data_processed.xlsx) is loaded into the system for analysis.
Data Preprocessing:
Missing values are filled to ensure no gaps in the data.
Categorical features are encoded to numerical formats suitable for machine learning algorithms.
Standardization is applied to scale features for better model performance.
Feature Selection Method:
If a feature selection method is chosen (e.g., SelectKBest), the dataset is reduced to the most important features before splitting into train/test sets.
If no feature selection method is applied, the data is directly split into train and test sets.
Train/Test Splitting: The dataset is divided into training and testing subsets for evaluation.
Model Training and Evaluation: Models are trained on the training set and their performance is evaluated using metrics like accuracy, precision, and recall on the test set.
Save Results and Visualize: Results are saved and visualizations, such as performance graphs, are generated for analysis.

3 Data

The "Data_processed.xls" dataset used in this study contains information on various soil properties. The "GrainYield" feature in the dataset represents grain yield and takes categorical values ("A", "B", "C"). The dataset has undergone preprocessing steps, including handling missing values and encoding categorical features.

Table 1 contains the important attributes in the dataset used in the study. These attributes represent specific data fields that stand out in the data processing and feature selection stages and are used in the analysis. The two most significant attributes were determined in the study using the SelectKBest and RFE methods:

These attributes were selected as variables that contribute the most to the performance of the prediction models. However, a decrease in model performance was observed in the analysis results when these two attributes were used alone. This shows that other attributes also make significant contributions to the prediction performance.**Table 1.** Data fields in the dataset.

Attribute Name	Description	Data Type
VarietyClass	Seed variety class	Numerical
Variety_HD_2824	Seed variety class	
DaysFromZerotoSowing	Days from zero to sowing	Numerical
SowingYear	Year of sowing	Numerical
GrainYield	Grain yield categories (A, B, C)	**Categorical Class Label (Target)**

Table 2. Classification performance analysis, sorted with respect to decreasing values of CA.

Algorithm	CA (Classification Accuracy)	AUC (Area Under Curve)
Logistic Regression	68.00%	78%
XGBoost	*76.70%*	86%
KNN	*73.23%*	81%
Decision Tree	*70.01%*	72%
Random Forest	*78.44%*	86%
Gradient Boosting	*75.76%*	84%
SVM	*74.02%*	85%
Naïve Bayes	*33.60%*	67%
AdaBoost	*63.18%*	72%
ANN	*72.95%*	81%

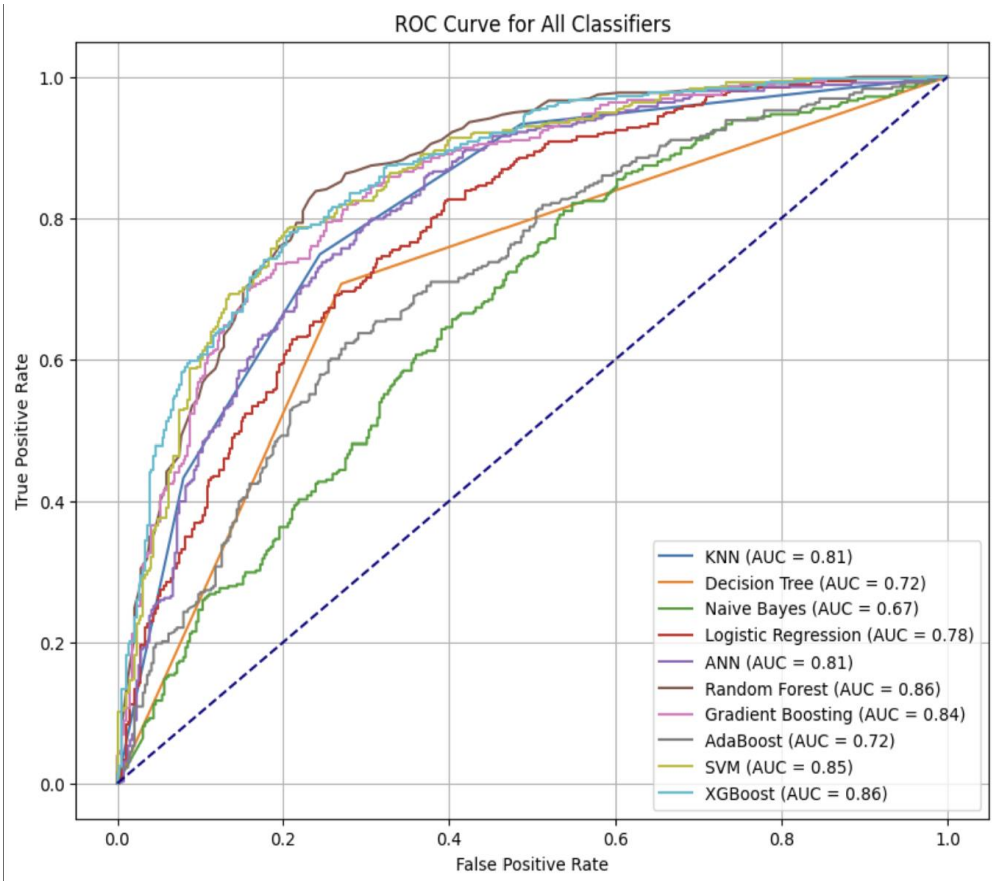

Fig. 2. ROC curve for the tested classification algorithms.

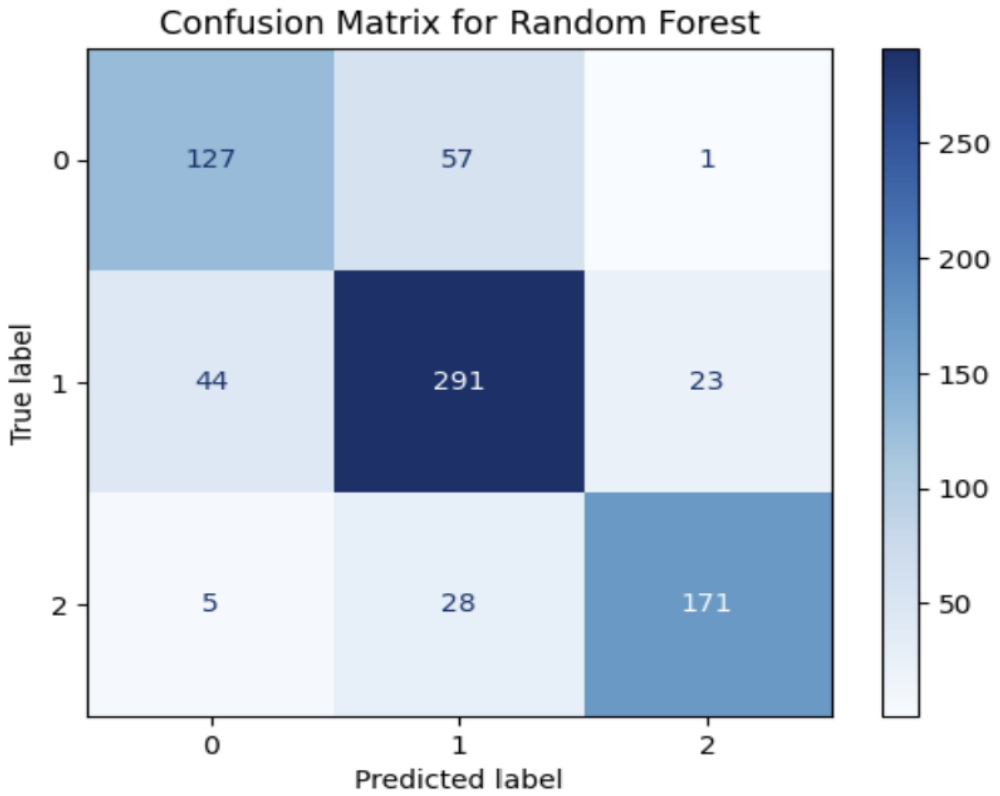

Fig. 3. Confusion matrix for the selected Random Forest algorithm, which has the highest CA (Classification Accuracy) value.

4 Analysis and Results

The results of experiments using different classification algorithms are shown in Table 2. The **Random Forest** algorithm performed best when using all features, achieving an **accuracy of 78.45%** and an **AUC value of 0.898**. Similarly, the **XGBoost** and **Gradient Boosting** algorithms also demonstrated strong performances, with accuracy values of **76.71%** and **75.77%**, respectively. The confusion matrix created for the **Random Forest** algorithm is shown in **Figure 2**.

The superior performance of the **Random Forest** algorithm can be attributed to its ability to **handle complex relationships** within the dataset, its **resilience to overfitting**, and its **ensemble learning approach**, which combines multiple decision trees for robust predictions.

On the other hand, when feature selection was applied, the overall performance of all models decreased significantly. The **Gradient Boosting** algorithm achieved the highest accuracy (**59.44%**) among models using only the top two selected features (VarietyClass_SDV and DaysFromZerotoSowing). This decline suggests that while these two features are significant, additional features contribute meaningfully to model performance.

The performance of other algorithms varied depending on the dataset's characteristics and the algorithms' hyperparameter configurations. These findings highlight the importance of both feature selection and the choice of machine learning models for achieving optimal predictive accuracy in grain yield estimation.

Classifier	Accuracy	Precision	Recall	F1 Score	MCC	AUC
KNN	0,732262	0,737917	0,732262	0,734245	0,577925	0,85032
Decision Tr	0,700134	0,700624	0,700134	0,70037	0,527384	0,757797
Naive Baye	0,336011	0,53933	0,336011	0,242639	0,139299	0,661627
Logistic Reg	0,680054	0,680737	0,680054	0,6779	0,487344	0,8209
ANN	0,729585	0,731402	0,729585	0,72999	0,57805	0,858422
Random Fo	0,784471	0,785696	0,784471	0,784399	0,657522	0,8985
Gradient Bo	0,757697	0,760139	0,757697	0,756158	0,612844	0,880734
AdaBoost	0,631861	0,639572	0,631861	0,623206	0,401171	0,777282
SVM	0,740295	0,742784	0,740295	0,735831	0,583865	0,882601
XGBoost	0,767068	0,767986	0,767068	0,766453	0,62912	0,897379

All Features | Selected Features | Best Model

Fig. 4. All Features Results

Selected Features
VarietyClass_SDV
DaysFromZerotoSowing

Classifier	Accuracy	Precision	Recall	F1 Score	MCC	AUC
KNN	0,488621	0,493098	0,488621	0,490494	0,201446	0,644001
Decision Tr	0,583668	0,578158	0,583668	0,572191	0,321687	0,713278
Naive Baye	0,544846	0,543881	0,544846	0,544112	0,277295	0,704252
Logistic Reg	0,548862	0,521658	0,548862	0,469857	0,245067	0,714644
ANN	0,578313	0,687471	0,578313	0,497875	0,309959	0,72186
Random Fo	0,585007	0,578647	0,585007	0,573142	0,324526	0,712658
Gradient Bo	0,594378	0,587912	0,594378	0,578448	0,3383	0,722071
AdaBoost	0,551539	0,542294	0,551539	0,533169	0,261707	0,702966
SVM	0,576975	0,442376	0,576975	0,493491	0,306107	0,695902
XGBoost	0,589023	0,582149	0,589023	0,572688	0,329051	0,713947

All Features | Selected Features | Best Model

Fig. 5. Selected Features Results

5 Conclusions

In this study, various machine learning classification algorithms were evaluated for their effectiveness in predicting grain yield. The dataset underwent preprocessing, including imputation of missing values, encoding of categorical variables, and standardization. Feature selection using SelectKBest identified 'VarietyClass_SDV' and 'DaysFromZerotoSowing' as the two most significant predictors.

Among the tested algorithms, Random Forest demonstrated the highest performance with an accuracy of 78.45%.
The robustness of Random Forest can be attributed to its ensemble nature, which effectively handles complex relationships within the dataset and reduces the risk of overfitting.

However, when feature selection was applied, the performance of all algorithms declined notably, indicating that additional features contribute significantly to accurate predictions. Gradient Boosting emerged as the best-performing model with selected features, achieving an accuracy of 59.44%.

These findings emphasize the importance of comprehensive feature engineering and the careful selection of machine learning models to maximize predictive accuracy.

References

https://www.geeksforgeeks.org/top-6-machine-learning-algorithms-for-classification/
https://dzone.com/refcardz/data-mining-discovering-and
https://www.activestate.com/resources/quick-reads/how-to-classify-data-in-python/