CENG-464 Data Mining Project

Comparing the Performance of Classification Algorithms for Predicting Grain Yield

GÖKAY ÇETİNAKDOĞAN

HASAN EMRE USTA

Outline

- The aim of study
- Methodology
- Dataset
- Data Preprocessing
- All Features
- Feature Selection

The aim of this study

- To compare the performance of different algorithms.

- By comparing the performance of different machine learning algorithms, the model that provides the highest accuracy was determined.

Methodology

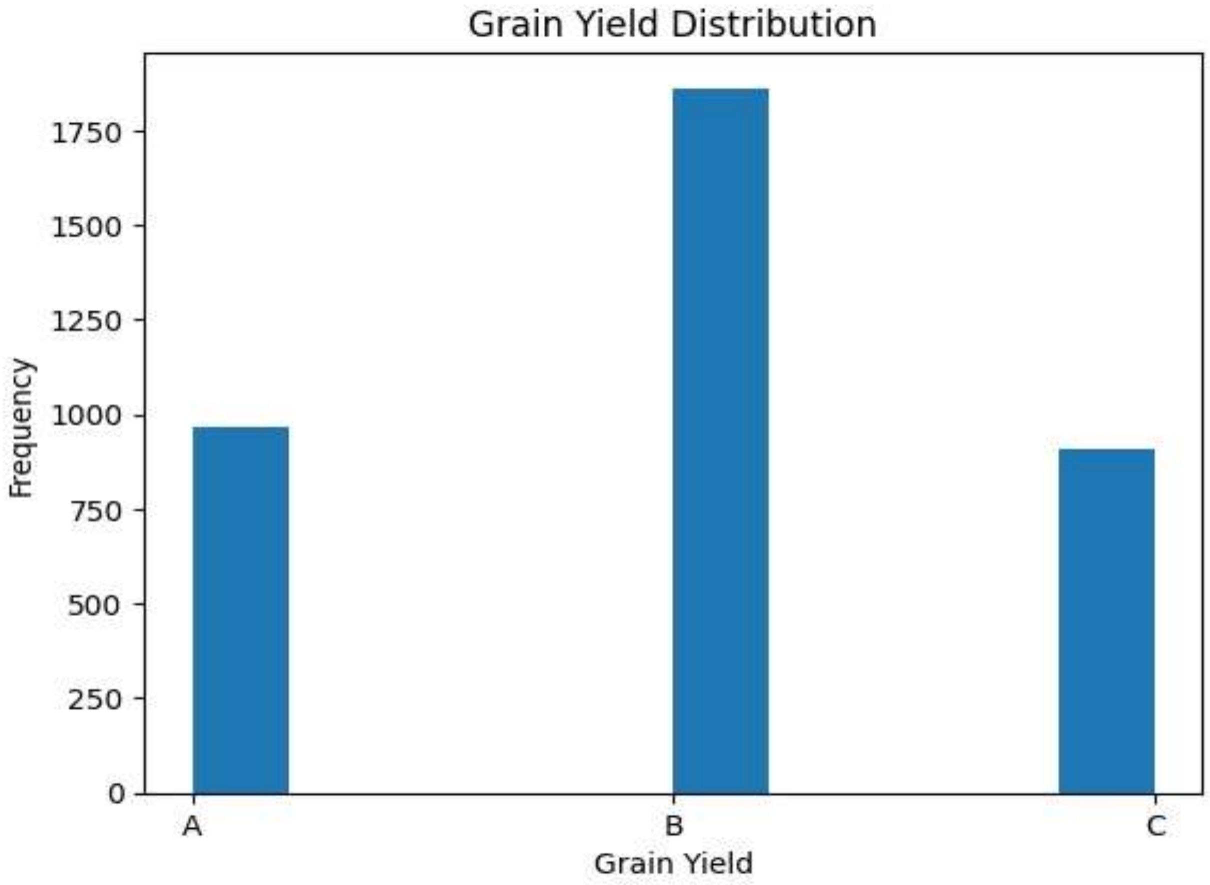
- Several classification algorithms from the Python sklearn package were trained and validated with 80-20 train-test splits.

- The sklearn package contains efficient and effective implementations of many of the most used methods,and thus was selected.

- The performance of the models was found using standard classification metrics including accuracy, F1 score and ROC-AUC

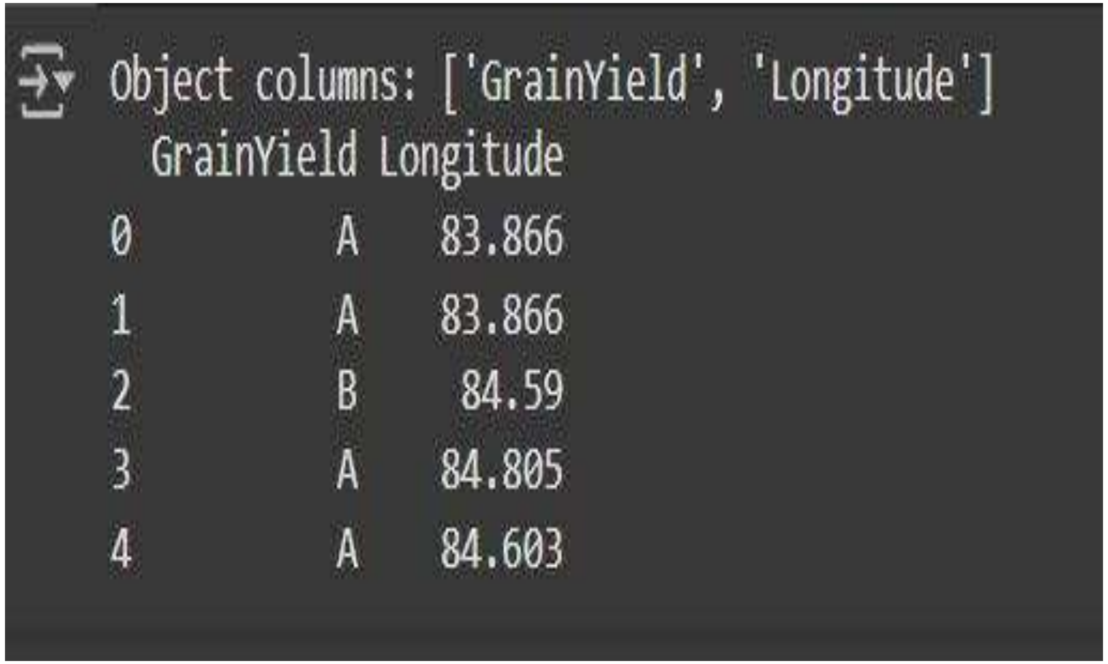
Dataset

- «Data_processed.xlsx»
- 3735 row and 120 column
- 119 numeric and 1 categoric
- Target column «Grain Yield»



Data Preprocessing

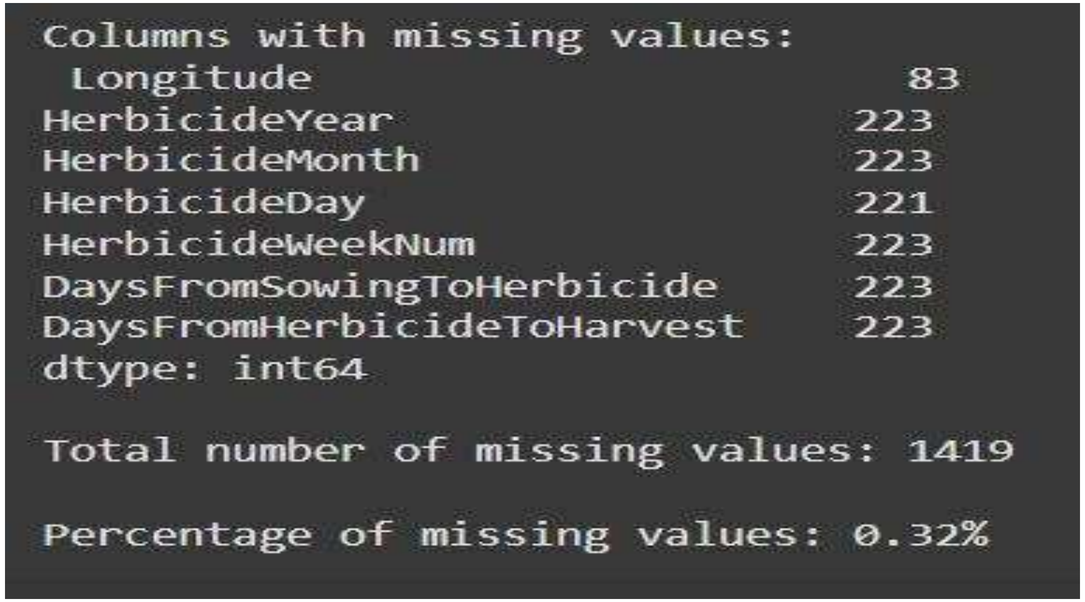
- Data cleaning



```
Object columns: ['GrainYield', 'Longitude']
  GrainYield Longitude
0          A    83.866
1          A    83.866
2          B     84.59
3          A    84.805
4          A    84.603
```

Data Preprocessing

- Missing data were filled with the column means for numeric columns and with zero for other columns.



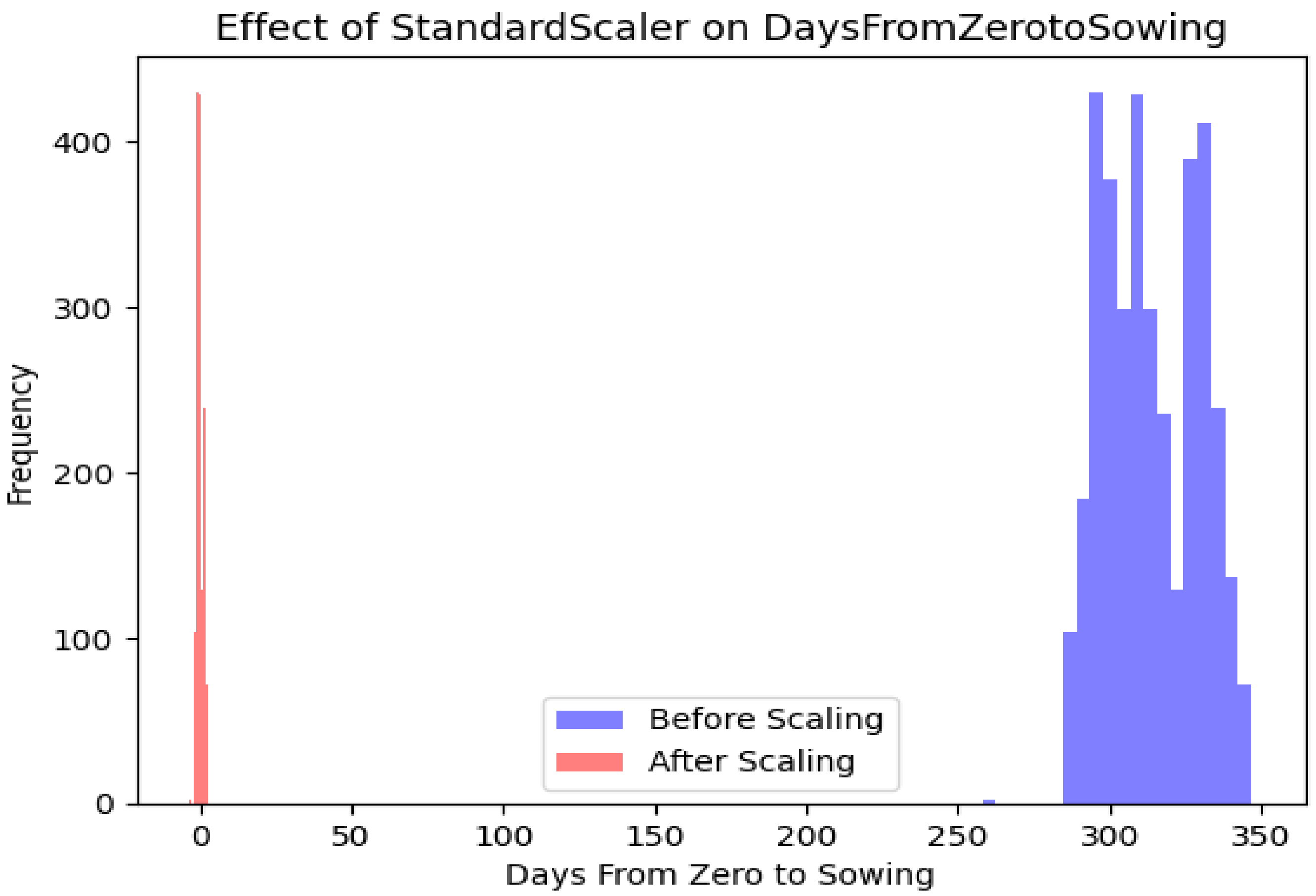
```
Columns with missing values:
 Longitude                        83
HerbicideYear                    223
HerbicideMonth                   223
HerbicideDay                     221
HerbicideWeekNum                 223
DaysFromSowingToHerbicide        223
DaysFromHerbicideToHarvest       223
dtype: int64

Total number of missing values: 1419

Percentage of missing values: 0.32%
```

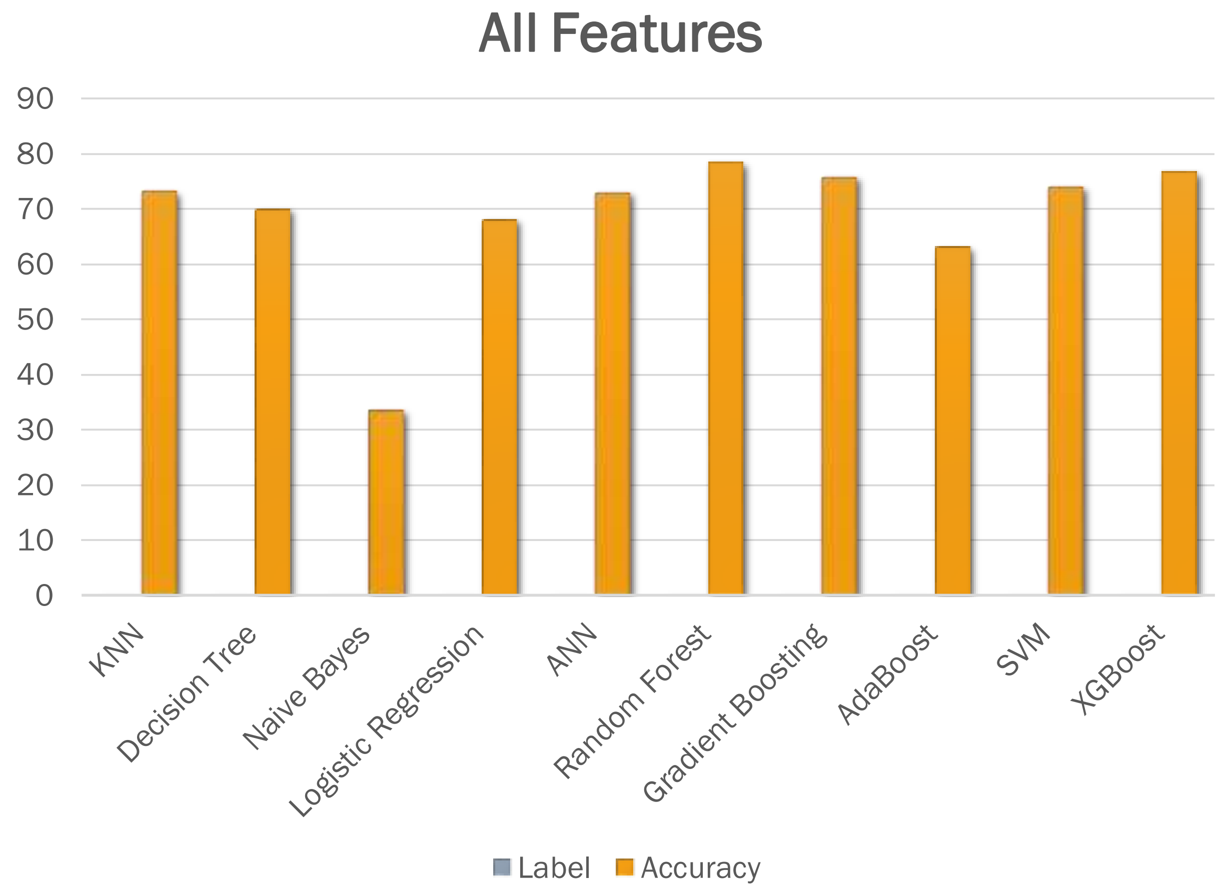
Data Preprocessing

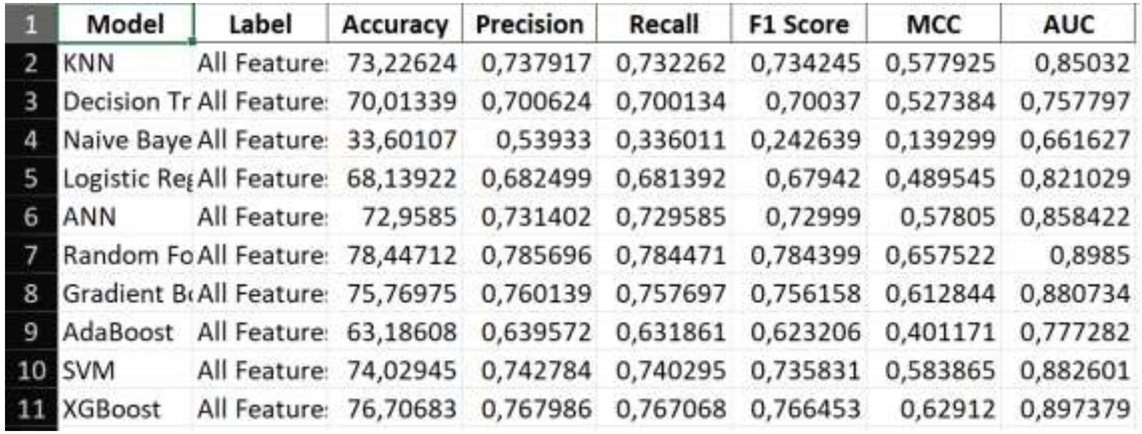
- The target column "Grain Yield" is classified as A = 0, B = 1, C = 2.
- Standard Scaler



All Features

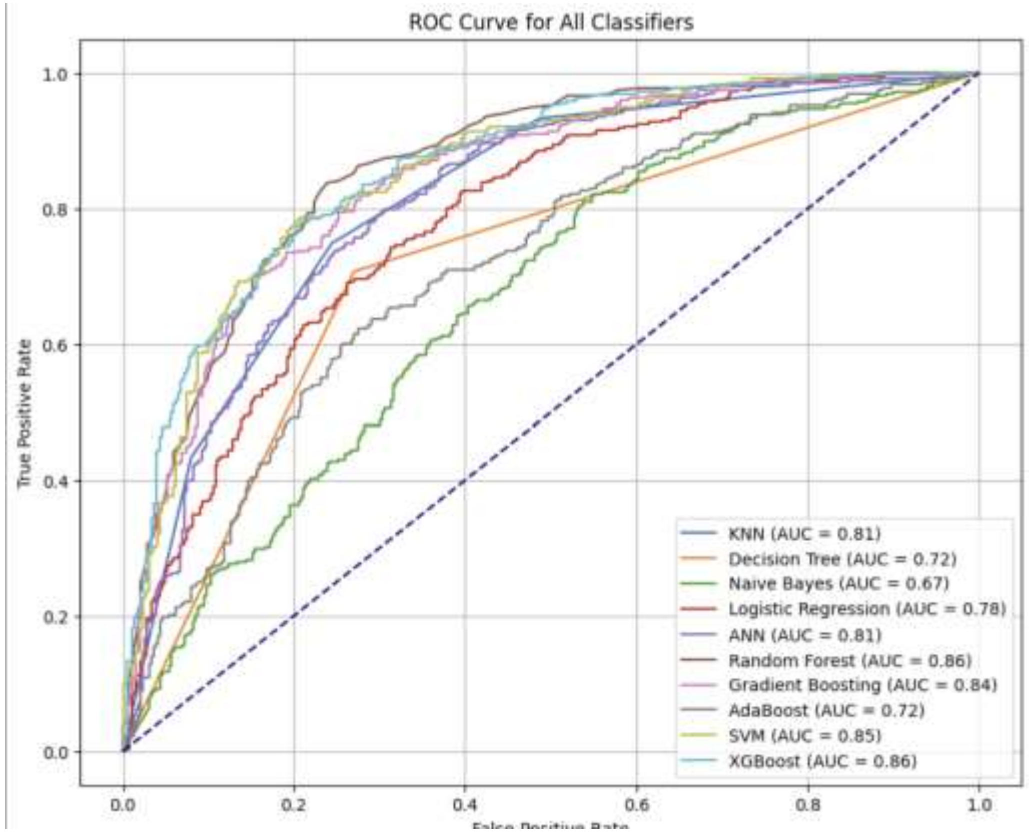
The best performance with all features is Random Forest 78.45% accuracy.

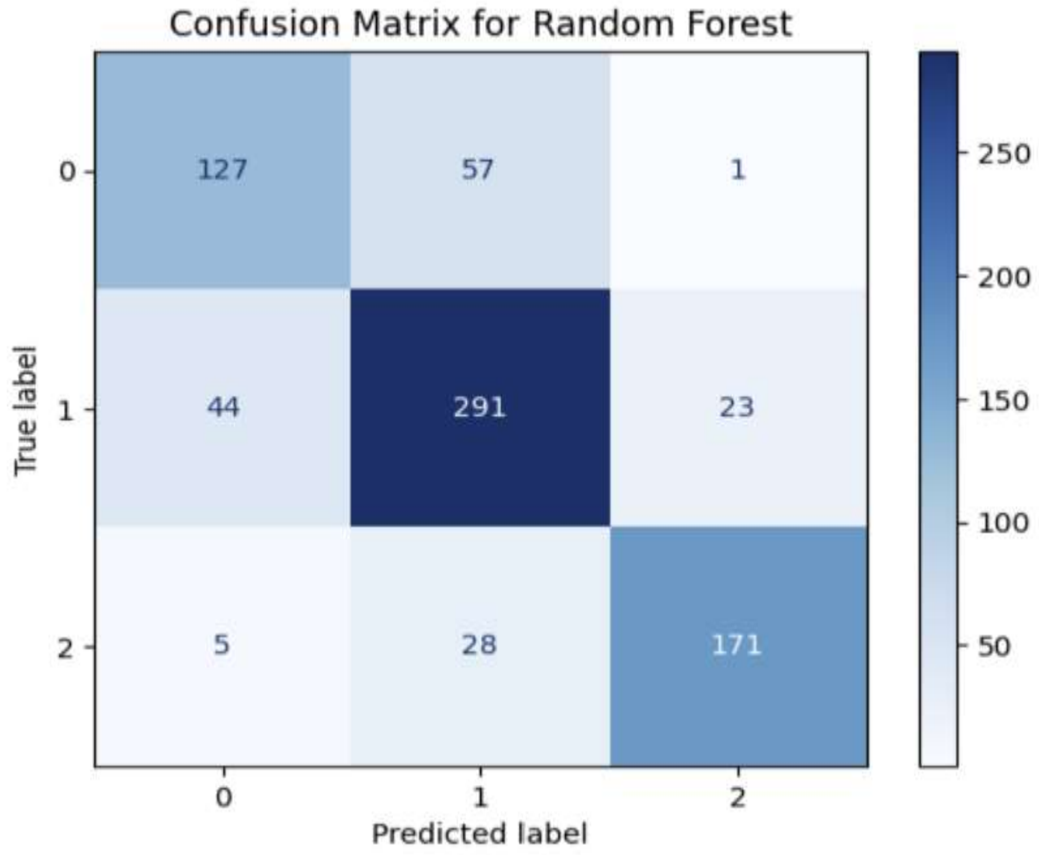




1	Model	Label	Accuracy	Precision	Recall	F1 Score	MCC	AUC
2	KNN	All Feature	73,22624	0,737917	0,732262	0,734245	0,577925	0,85032
3	Decision Tr	All Feature	70,01339	0,700624	0,700134	0,70037	0,527384	0,757797
4	Naive Baye	All Feature	33,60107	0,53933	0,336011	0,242639	0,139299	0,661627
5	Logistic Reg	All Feature	68,13922	0,682499	0,681392	0,67942	0,489545	0,821029
6	ANN	All Feature	72,9585	0,731402	0,729585	0,72999	0,57805	0,858422
7	Random Fo	All Feature	78,44712	0,785696	0,784471	0,784399	0,657522	0,8985
8	Gradient B	All Feature	75,76975	0,760139	0,757697	0,756158	0,612844	0,880734
9	AdaBoost	All Feature	63,18608	0,639572	0,631861	0,623206	0,401171	0,777282
10	SVM	All Feature	74,02945	0,742784	0,740295	0,735831	0,583865	0,882601
11	XGBoost	All Feature	76,70683	0,767986	0,767068	0,766453	0,62912	0,897379

Roc Curve and Confusion Matrix

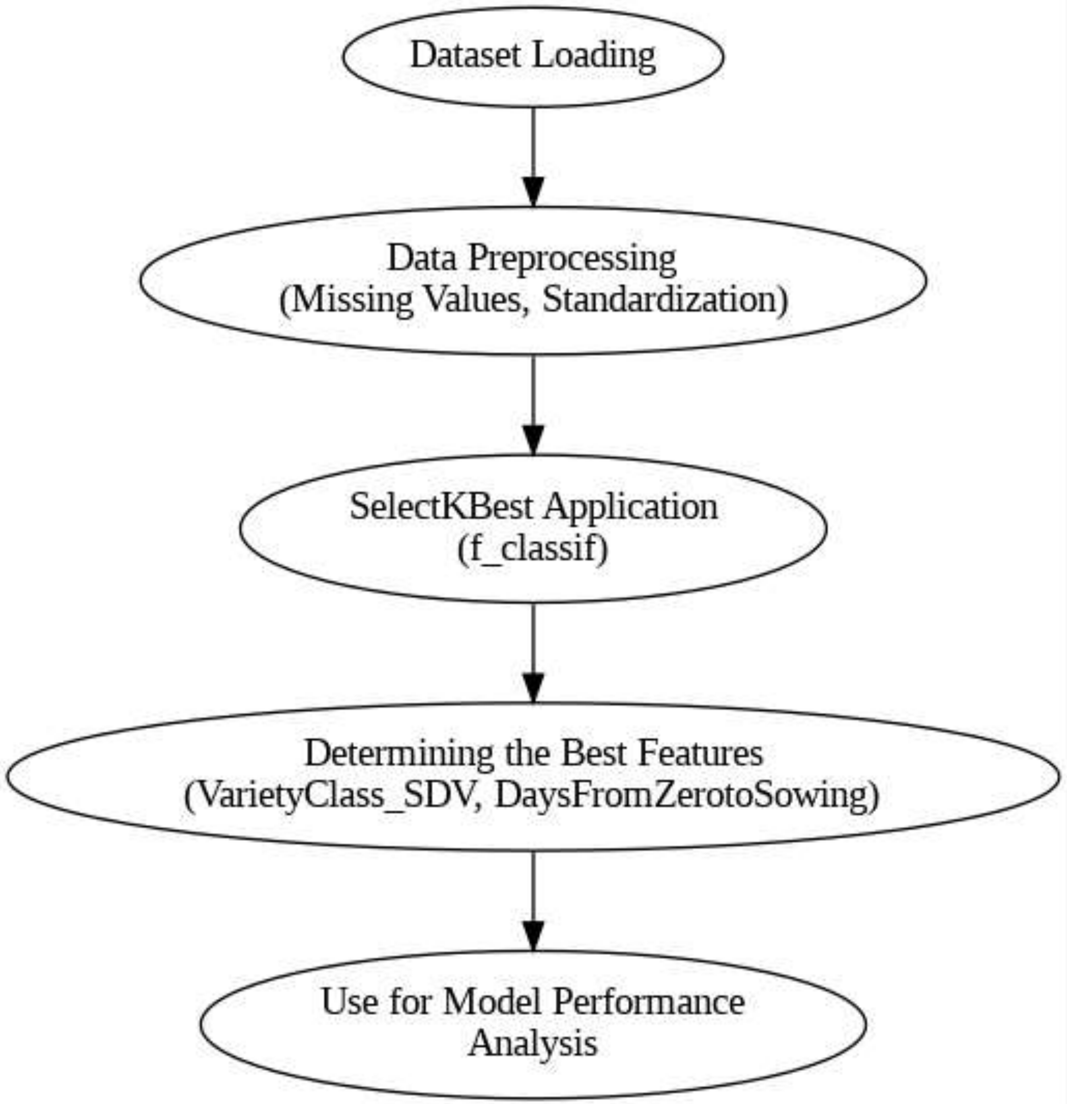




Feature Selection

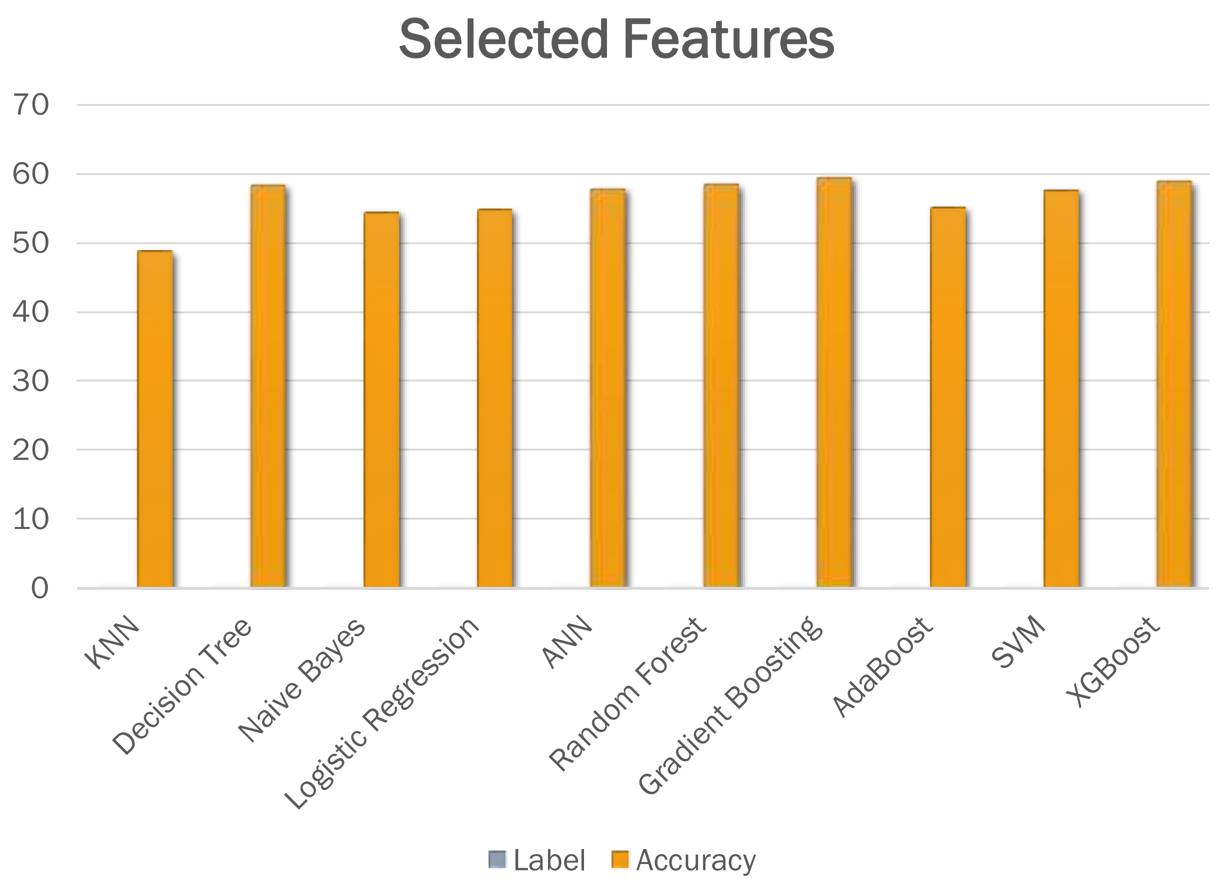
- The SelectKBest method is used to select the two best features (with the f_classif score function)

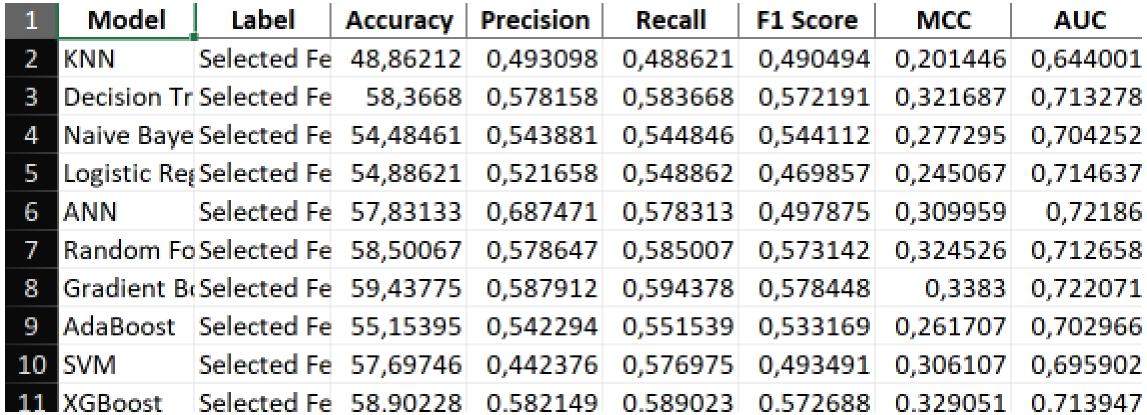
- Selected features: VarietyClass_SDV and DaysFromZerotoSowing..


Dataset Loading
Data Preprocessing (Missing Values, Standardization)
SelectKBest Application (f_classif)
Determining the Best Features (VarietyClass_SDV, DaysFromZerotoSowing)
Use for Model Performance Analysis

Selected Features

- VarietyClass_SDV and
- DaysFromZerotoSowing
- SelectKBest method
- With selected features, Gradient Boosting has the best value with 59.44% accuracy rate.

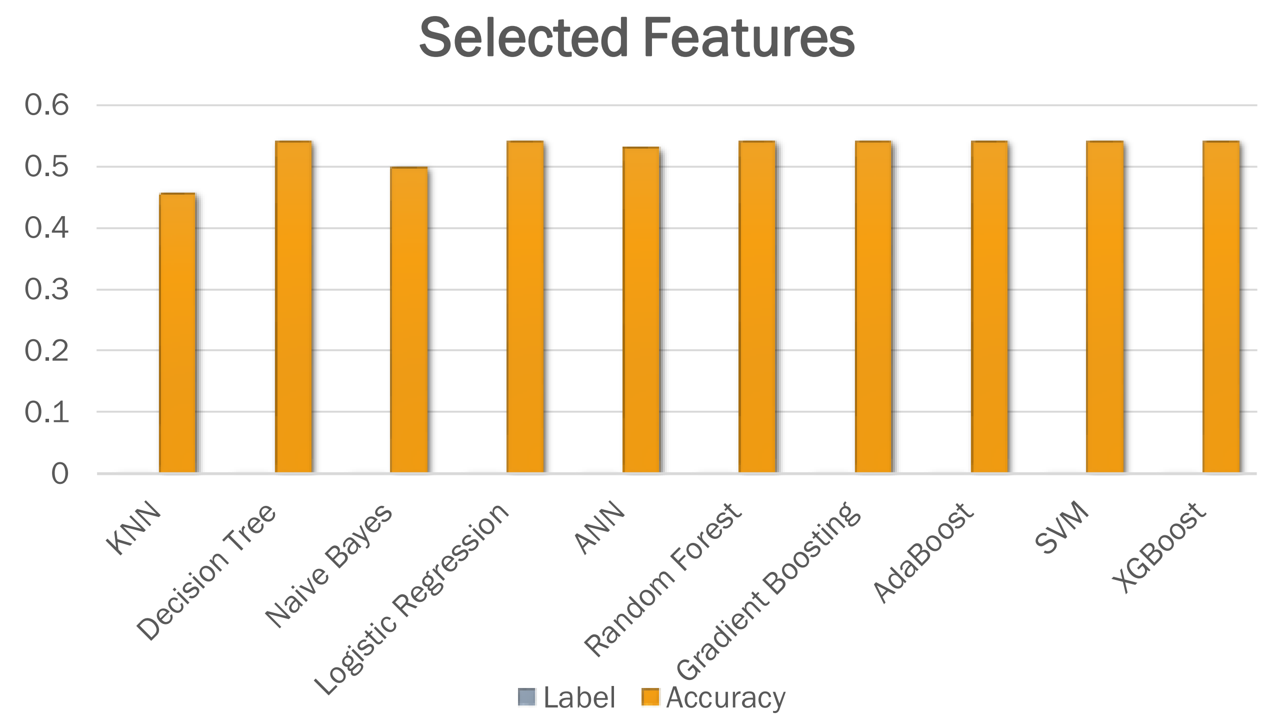


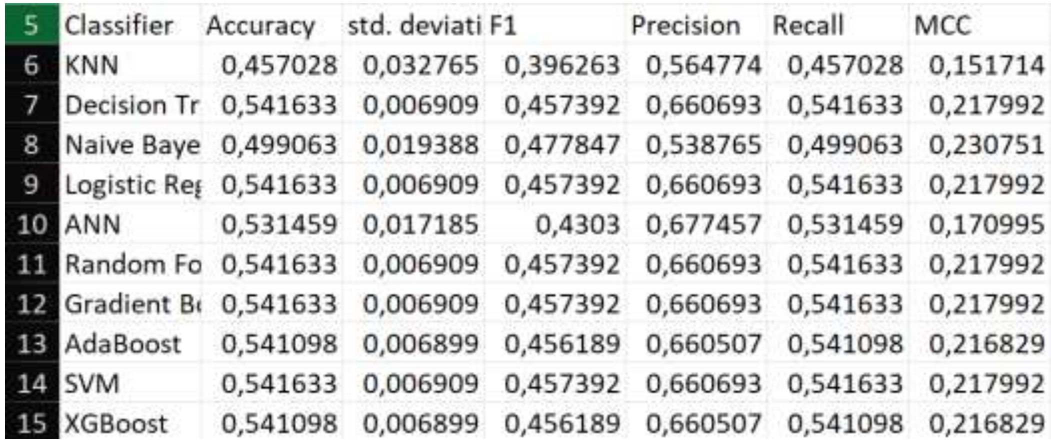


1	Model	Label	Accuracy	Precision	Recall	F1 Score	MCC	AUC
2	KNN	Selected Fe	48,86212	0,493098	0,488621	0,490494	0,201446	0,644001
3	Decision Tr	Selected Fe	58,3668	0,578158	0,583668	0,572191	0,321687	0,713278
4	Naive Baye	Selected Fe	54,48461	0,543881	0,544846	0,544112	0,277295	0,704252
5	Logistic Re	Selected Fe	54,88621	0,521658	0,548862	0,469857	0,245067	0,714637
6	ANN	Selected Fe	57,83133	0,687471	0,578313	0,497875	0,309959	0,72186
7	Random Fo	Selected Fe	58,50067	0,578647	0,585007	0,573142	0,324526	0,712658
8	Gradient B	Selected Fe	59,43775	0,587912	0,594378	0,578448	0,3383	0,722071
9	AdaBoost	Selected Fe	55,15395	0,542294	0,551539	0,533169	0,261707	0,702966
10	SVM	Selected Fe	57,69746	0,442376	0,576975	0,493491	0,306107	0,695902
11	XGBoost	Selected Fe	58.90228	0.582149	0.589023	0.572688	0.329051	0.713947

Selected Features

- Variety_HD_2824
- SowingYear
- RFE method
- With selected features, SVM has the best value with 54.16% accuracy rate.





5	Classifier	Accuracy	std. deviati	F1	Precision	Recall	MCC
6	KNN	0,457028	0,032765	0,396263	0,564774	0,457028	0,151714
7	Decision Tr	0,541633	0,006909	0,457392	0,660693	0,541633	0,217992
8	Naive Baye	0,499063	0,019388	0,477847	0,538765	0,499063	0,230751
9	Logistic Reg	0,541633	0,006909	0,457392	0,660693	0,541633	0,217992
10	ANN	0,531459	0,017185	0,4303	0,677457	0,531459	0,170995
11	Random Fo	0,541633	0,006909	0,457392	0,660693	0,541633	0,217992
12	Gradient Bo	0,541633	0,006909	0,457392	0,660693	0,541633	0,217992
13	AdaBoost	0,541098	0,006899	0,456189	0,660507	0,541098	0,216829
14	SVM	0,541633	0,006909	0,457392	0,660693	0,541633	0,217992
15	XGBoost	0,541098	0,006899	0,456189	0,660507	0,541098	0,216829