Mini Project - 2

Dataset - Adult.data

A. Implementation Steps:

Step 1 : Exploring data and missing values -

The data was visualized using Tableau. Correlation between different parameters and class was explored. Three attributes, all categorical, have missing values. The dataset follows a binary class distribution.

Step 2 : Data Preprocessing -

Different preprocessing techniques have been used in order to analyze the performance of classification algorithm.

a) Missing Values: Missing values in the dataset have been replaced by the most occurring values in the dataset for the respective attribute. Missing Values are in the columns work class, occupation and native country.

b) Normalization : The dataset was normalized before classification task using inbuilt library StandardScaler.

c) Feature Reduction : There are two kinds of implementation that were studied for this dataset. In the first one all the features were considered in the feature set for classification. In the second implementation, feature reduction was done on the basis of feature importance calculated by tree based classifier. The results of both the implementations were then studied for all algorithms.

According to the feature reduction following features have higher relevance than others:

Higher relevance : age, fnlwgt, education_num, race, occupation, relationship, capital gain, hours per week

Lower relevance: work class, education, marital status, race, sex, capital loss, native country

Step 3 : Algorithm Implementation & Results -

Seven different classification algorithms are implemented using 10 fold cross validation. Two kinds of cross folds were used - k fold and stratified k fold. The k-fold cross validation divide the instances into 10 equal folds without taking into consideration the class distribution. However, the stratified cross fold takes into consideration the class distribution and maintains balance between all classes while creating train and test fold. The results below are from stratified k fold which performed better than k fold.

Various evaluation metrics are then calculated including ROC-AUC curve and learning curve to understand the performance of each classifier.

3.1. Naive Bayes Algorithm :

For the analysis of results two different implementation was done and evaluated. The performance metrics is as below:

Metric	With all labels	With feature reduction features = 8
Accuracy	0.80	0.79
Balanced Accuracy	0.63	0.59
Precision	0.66	0.74
Recall	0.31	0.21
F-1 Measure	0.42	0.33
Mathews Correlation Coefficient	0.34	0.31
AUC	0.82	0.84
Confusion Matrix	[[23457 1263]	[[24137 583]
	[5401 2440]]	[6175 1666]]

3.2. Decision Tree

The performance measures are as follows:

Metric	With all labels	With feature reduction
Accuracy	0.81	0.80
Balanced Accuracy	0.74	0.73
Precision	0.60	0.57
Recall	0.61	0.59
F-1 Measure	0.61	0.58
Mathews Correlation Coefficient	0.48	0.45
AUC	0.74	0.73
Confusion Matrix	[[21570 3150] [2957 4884]]	[[21287 3433] [3245 4596]]

3.3. K-Nearest Neighbors

The algorithm gave best results with number of neighbors 15. The performance measures are as follows:

Metric	With all labels	With feature reduction
Accuracy	0.80	0.84
Balanced Accuracy	0.60	0.75
Precision	0.76	0.71
Recall	0.23	0.58
F-1 Measure	0.36	0.64
Mathews Correlation Coefficient	0.34	0.54
AUC	0.66	0.88
Confusion Matrix	[[24103 617] [6027 1814]]	[[22829 1891] [3256 4585]]

3.4. Random Forest

The performance measure as below with number of trees 100:

Metric	With all labels	With feature reduction
Accuracy	0.86	0.84
Balanced Accuracy	0.78	0.76
Precision	0.74	0.70
Recall	0.63	0.60
F-1 Measure	0.68	0.64
Mathews Correlation Coefficient	0.59	0.54
AUC	0.91	0.89
Confusion matrix	[[23012 1708]	[[22704 2016]
	[2903 4938]]	[3161 4680]]

3.5. Logistic Regression

Metric	With all labels	With feature reduction
Accuracy	0.80	0.82
Balanced Accuracy	0.62	0.68
Precision	0.69	0.72
Recall	0.28	0.41
F-1 Measure	0.40	0.53
Mathews Correlation Coefficient	0.35	0.45
AUC	0.71	0.84
Confusion matrix	[[23669 1051]	[[23439 1281]
	[5553 2288]]	[4587 3254]]

3.6. ADA Boost

Metric	With all labels	With feature reduction
Accuracy	0.86	0.85

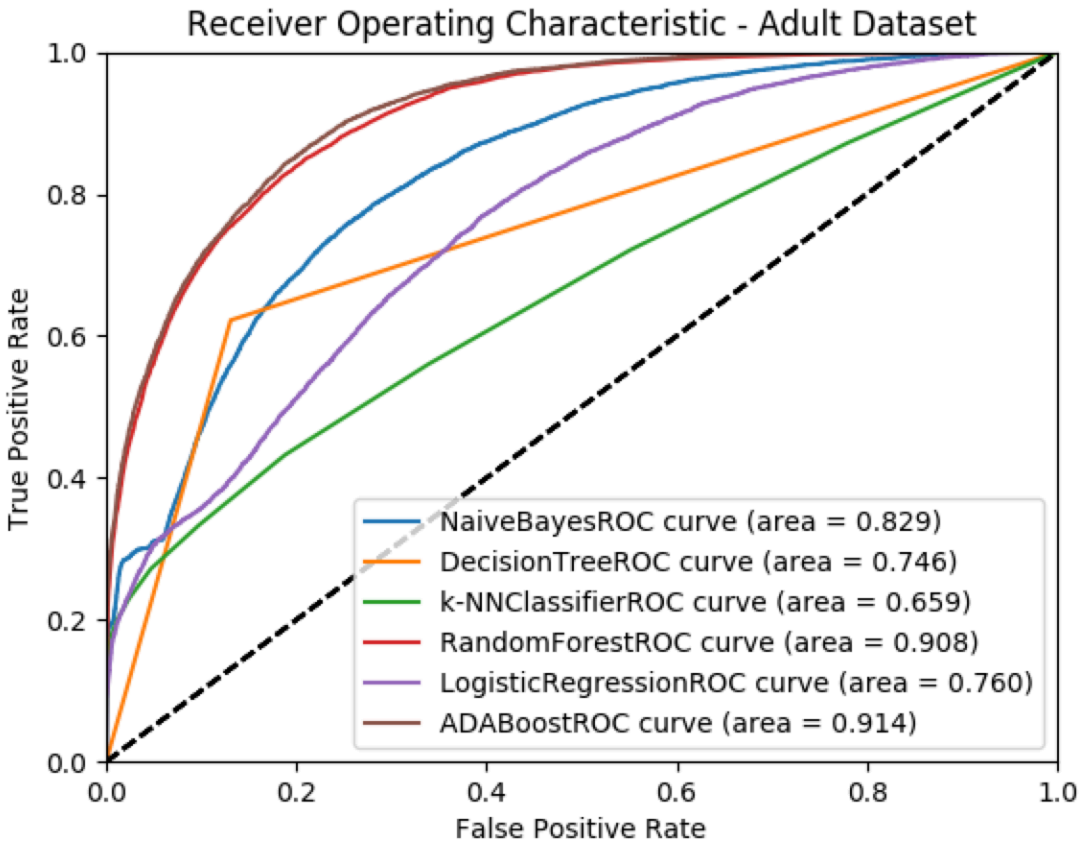
Metric	With all labels	With feature reduction
Balanced Accuracy	0.77	0.76
Precision	0.77	0.75
Recall	0.60	0.57
F-1 Measure	0.67	0.65
Mathews Correlation Coefficient	0.59	0.56
AUC	0.91	0.90
Confusion matrix	[[23346 1374] [3196 4645]]	[[23116 1604] [3283 4558]]

3.7. Support Vector Machine

Metric	With feature reduction
Accuracy	0.7977949080187950
Balanced Accuracy	0.5944369877199300
Precision	0.8285415577626760
Recall	0.20214258385410000
F-1 Measure	0.32499487389788800
Mathews Correlation Coefficient	0.34341837361572400
AUC	0.8179448001522160
Confusion Matrix	[24392, 328], [6256, 1585]]

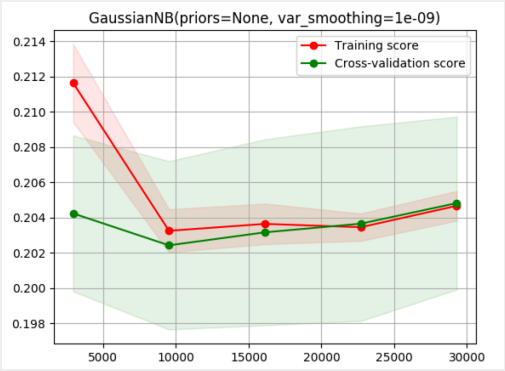
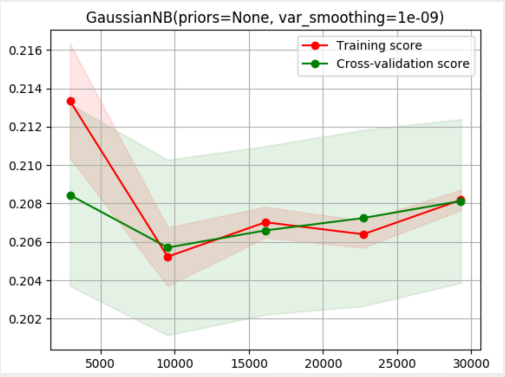
3.8. ROC Curve:

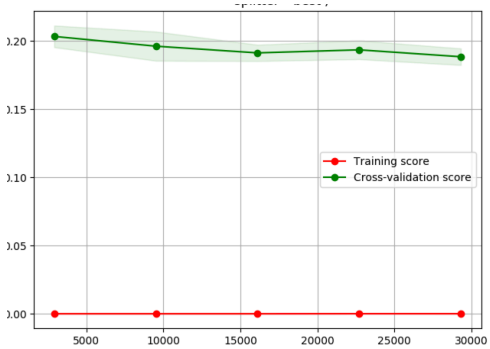
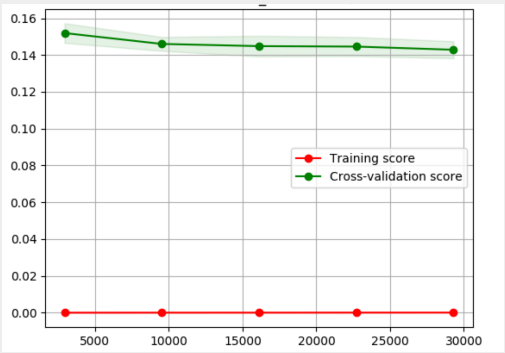
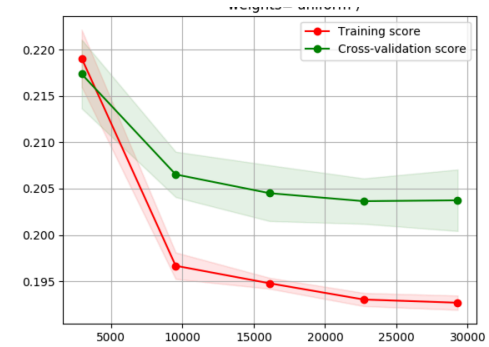
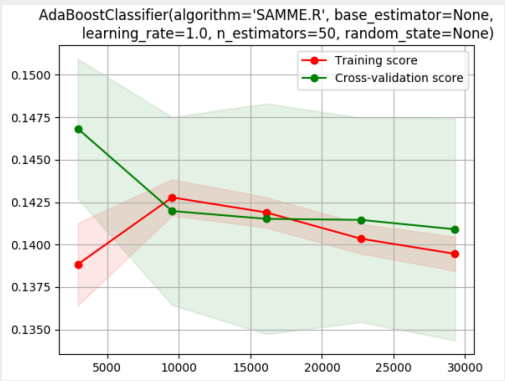
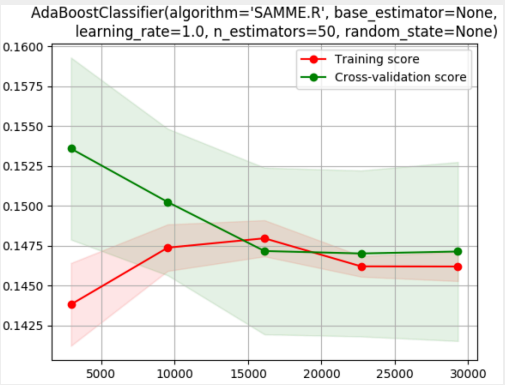
The ROC Curve for all the classifiers(except SVM) are shown below. As we can see that Random Forest and ADA boost have the maximum area under cure.

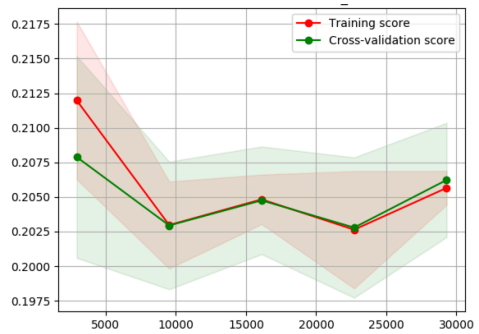
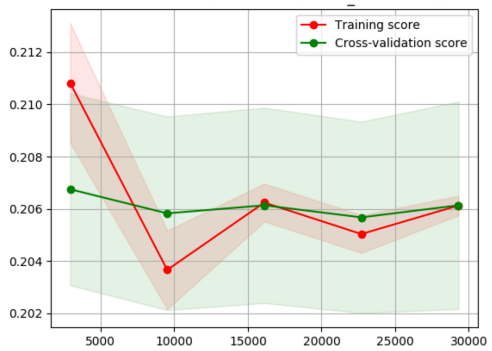


3.9. Learning Curves

Learning curves have been plotted for each algorithm. On x axis we have number of samples and on y axis we have mean squared error.

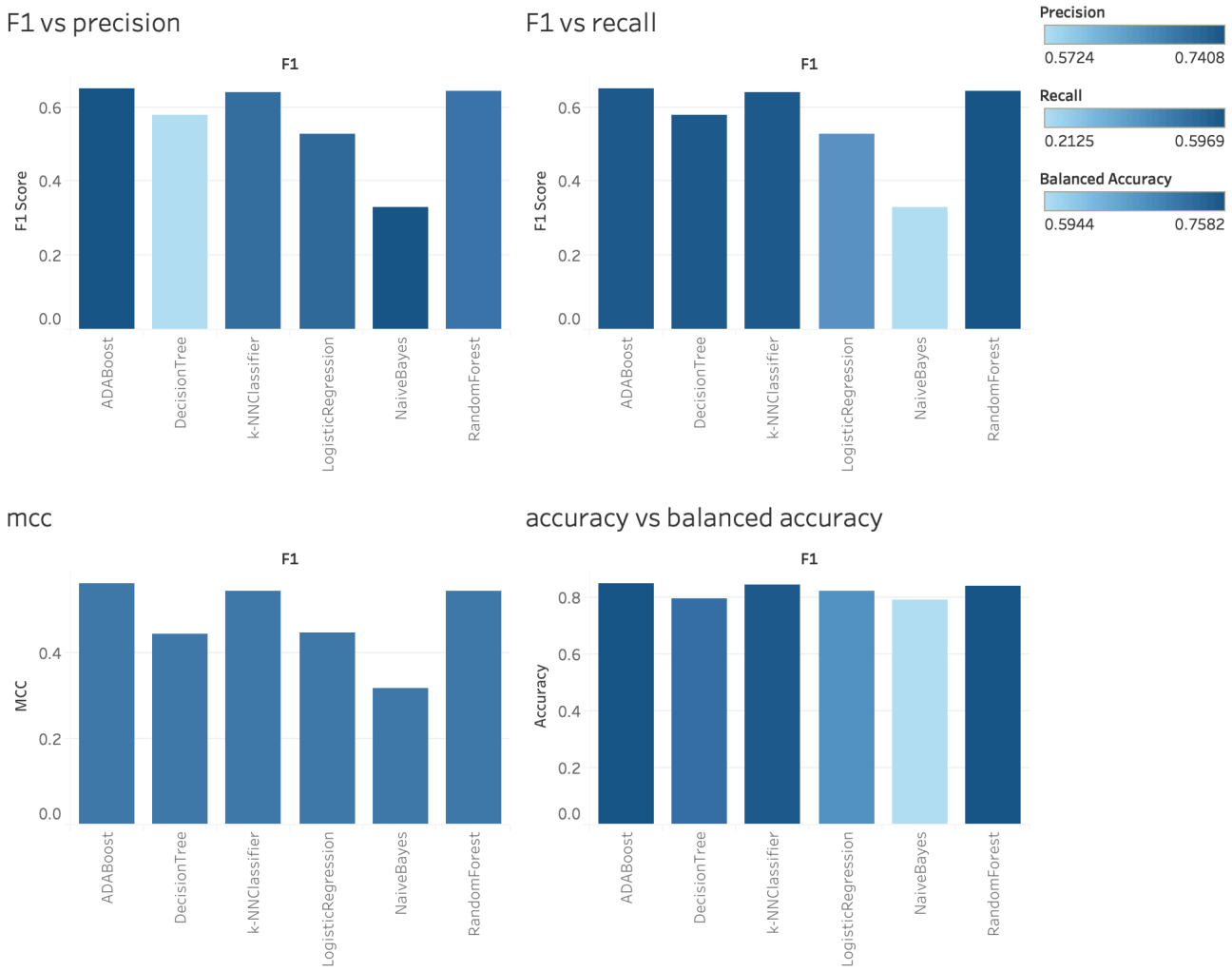
ALGO	With all features	With feature reduction
Gaussian Naive Bayes		

Decision Tree		
Random Forest		
KNN		
Ada Boost		

Logistic Regression		

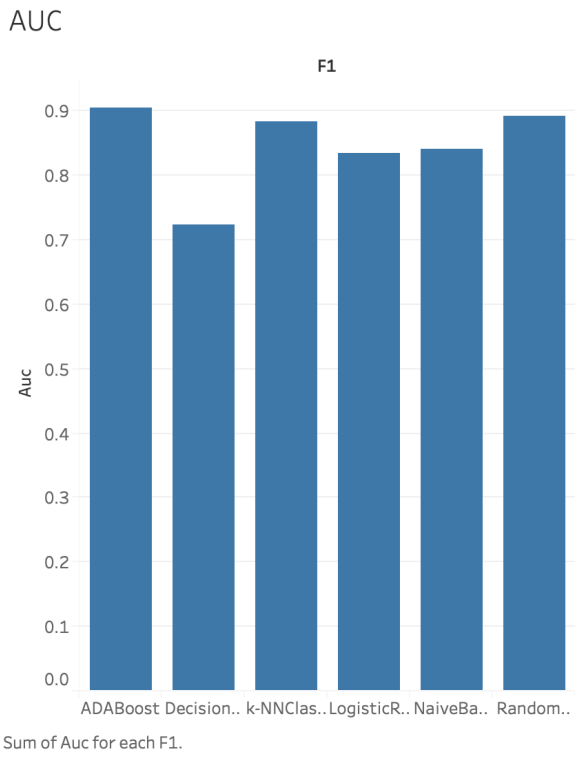
3.10. Tableau Representation

A. The dashboard below gives a brief summary of F1, precision, recall and accuracy for algorithms with feature importance.

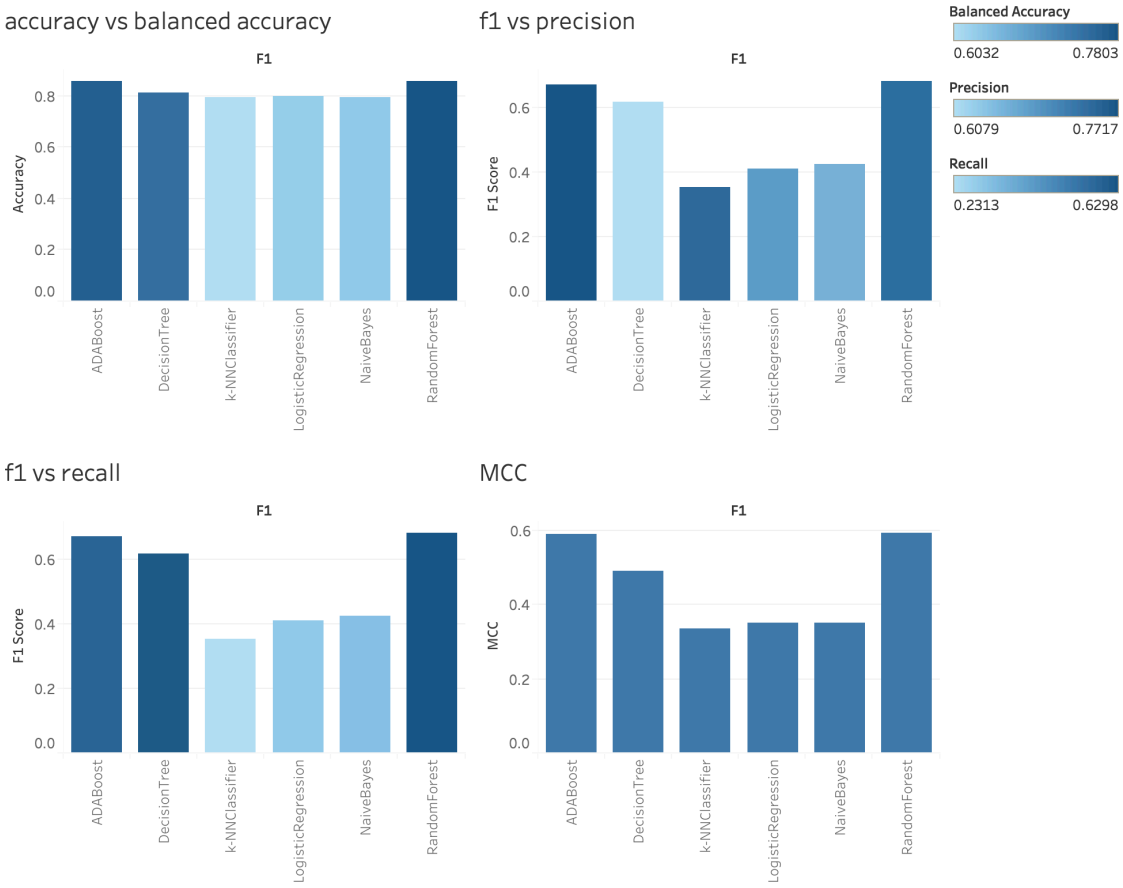


B. Area Under the curve:

The plot shows area under the curve for different algorithms implemented as per feature relevance.

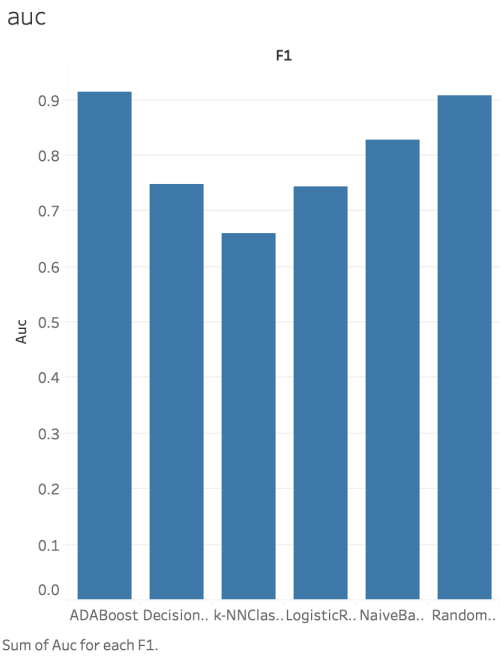


C. The dashboard below gives a brief summary of F1, precision, recall and accuracy for algorithms without considering any feature importance I.e. with all features.



D. Area Under the curve:

The plot shows area under the curve for different algorithms implemented without taking into consideration the feature relevance.



Step 4 : Conclusion

The ROC- AUC curve depicts how well the test can separate the group being tested into the two class labels. The more the area, the better it is. As per the above plotted graph, the two classification algorithms that have the highest area under curve viz. ADA Boost and Random Forest.

If we look at the evaluation parameters for all algorithm for both approaches I.e. with feature reduction and without feature reduction, ADA Boost and Random Forest seem to be well balanced in both cases however, the other algorithms improve on all evaluation parameters when feature set has been reduced as per the feature importance. Which implies that ADA Boost and Random Forest can do classification better with the whole set of features and scale the performance on its own.

The Matthews correlation coefficient measure the quality of classification where the classes are imbalanced. In this dataset the ratio of class is approx. 75:25. The MCC value for ADA Boost and Random Forest is better than the rest of algorithm.

Precision is the ability of classifier not to label a negative sample as positive. Recall is the ability of classifier to find all positive samples. A high value of precision and recall indicates that the classifier has been able to correctly classify both positive and negative classes. F1 Score is the harmonic mean of precision and recall. Higher the value of F1 better the classification is. ADA Boost and Random Forest have a better precision, recall and F1 values. Although the precision recall and F1 values are higher in some other algorithms as well eg. K-NN, but for an overall assessment, we shall consider accuracy, AUC and the performance of algorithm with different feature set.

Learning curves are a measure of error. They are plotted between the Mean square Error and sample size. Ideally, as the sample size increase, the validation and training error should decrease and in the end converge with each other. Looking at the learning curves of all classifiers above, it can be inference that the ADA Boost algorithm performs better than the other classifiers. Unlike other classifiers, ADA Boost does not have high training error or no training error. Naive Bayes, KNN and Logistic

regression has high training error which indicates high bias. Considering other metrics as well, random forest can be considered to be the next better classifier.

Hence, at a holistic view, ADA Boost and Random Forest performs better than the rest of algorithm.

Dataset - Hand Gesture Dataset

A. Implementation Steps:

Step 1 : Exploring data and missing values - The data was visualized using Tableau. Correlation between different parameters and class was explored. All attributes are continuous. Multiple attributes have missing values. The dataset follows a multi class distribution.

Step 2: Different preprocessing techniques have been used in order to analyze the performance of classification algorithm.

a) Missing Values: Missing values in the dataset have been replaced with the mean of that attribute in the dataset for the respective attribute. There are some columns where more than 50% attributes have missing value. Feature reduction was used for this dataset.

b) Normalization : The dataset was normalized before classification task using inbuilt library StandardScaler.

c) Feature Reduction : For this dataset, feature reduction was done on the basis of feature importance calculated by tree based classifier. Feature reduction was used because there are too many features and most of them have a lot of missing values. The results were then studied for all algorithms.

Step 3 : Algorithm Implementation & Results

Since the dataset has a lot of missing values, for some features more than 50% missing values. Hence, for the implementation, first feature importance was calculated for each feature and then the higher relevance features were considered.

3.1. Naive Bayes Algorithm :

The feature The performance metrics is as below:

Metric	With feature reduction features = 17
Accuracy	0.518432678148409

Metric	With feature reduction features = 17
Precision	0.4481071642757860
Recall	0.518432678148409
F-1 Measure	0.407096371846232
Mathews Correlation Coefficient	0.44214233919581300
AUC	Class 1: 0.8593701938205164
	Class 2: 0.8903712992659208
	Class 3: 0.5
	Class 4: 0.5569380911610569
	Class 5: 0.8708735020529855
Confusion matrix	[16138, 114, 0, 9, 4], [196, 13146, 0, 24, 1612], [12882, 999, 0, 377, 2086], [7351, 3304, 0, 825, 3295], [772, 4557, 0, 26, 10378]

3.2. Decision Tree

The performance measures are as follows:

Metric	With feature reduction features = 17
Accuracy	0.7173954798642680
Precision	0.7767674339784340
Recall	0.7173954798642680
F-1 Measure	0.7077470747749060
Mathews Correlation Coefficient	0.6623732913648800
AUC	Class 1: 0.7466357110288072
	Class 2: 0.9318216982036128
	Class 3: 0.7483812172475831
	Class 4: 0.8122042150471431

Metric	With feature reduction features = 17
	Class 5: 0.8921791886242659
Confusion matrix	[8042, 1516, 519, 2753, 3435], [24, 14247, 26, 133, 548], [22, 2179, 8352, 2910, 2881], [9, 1210, 279, 10649, 2628], [17, 621, 56, 304, 14735]

3.3. K-Nearest Neighbors

The algorithm gave best results with number of neighbors 15.

Metric	With feature reduction features = 17
Accuracy	0.941455919072924
Precision	0.9412659125727760
Recall	0.941455919072924
F-1 Measure	0.9412582536296040
Mathews Correlation Coefficient	0.9268306407099160
AUC	Class 1: 0.9982944573140388
	Class 2: 0.9914506414891446
	Class 3: 0.993610604224046
	Class 4: 0.9912818440002865
	Class 5: 0.9937956059798951
Confusion matrix	[16195, 47, 14, 8, 1], [60, 13824, 97, 107, 890], [344, 1, 15415, 568, 16], [194, 38, 812, 13672, 59], [71, 883, 88, 274, 14417]

3.4. Random Forest

Metric	With feature reduction features = 17
Accuracy	0.9796529867469110
Precision	0.9797446499095480
Recall	0.9796529867469110
F-1 Measure	0.9796470800586130
Mathews Correlation Coefficient	0.9745837852336250
AUC	Class 1: 0.9997686282096239
	Class 2: 0.9983342966740815
	Class 3: 0.9990804658633567
	Class 4: 0.998527973294939
	Class 5: 0.9996735169694908
	[16168, 48, 37, 10, 2], [10, 14657, 47, 70, 194], [30, 88, 15740, 359, 127], [9, 267, 104, 14331, 64], [0, 84, 13, 26, 15610]

3.5. Logistic Regression

Metric	With feature reduction features = 17
Accuracy	0.7189320699148470
Precision	0.7171474562159860
Recall	0.7189320699148470
F-1 Measure	0.7174296381356870
Mathews Correlation Coefficient	0.6487526387921610
AUC	Class 1: 0.9479015540911514
	Class 2: 0.9437552799235966
	Class 3: 0.899902895427536

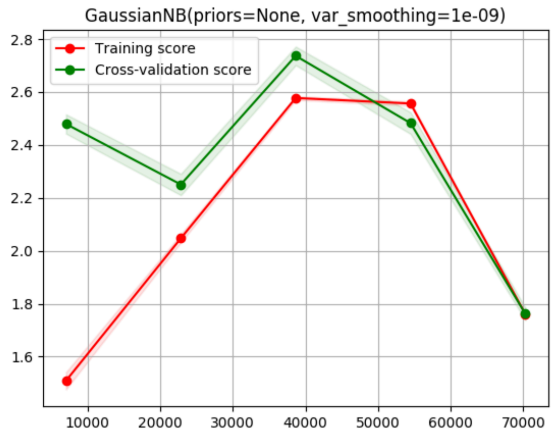
Metric	With feature reduction features = 17
	Class 4: 0.8672418729884892
	Class 5: 0.9160556097897845
Confusion Matrix	[13581, 37, 1981, 653, 13], [99, 11454, 241, 421, 2763], [2979, 94, 10632, 2392, 247], [916, 1013, 2316, 9387, 1143], [219, 2954, 323, 1146, 11091]

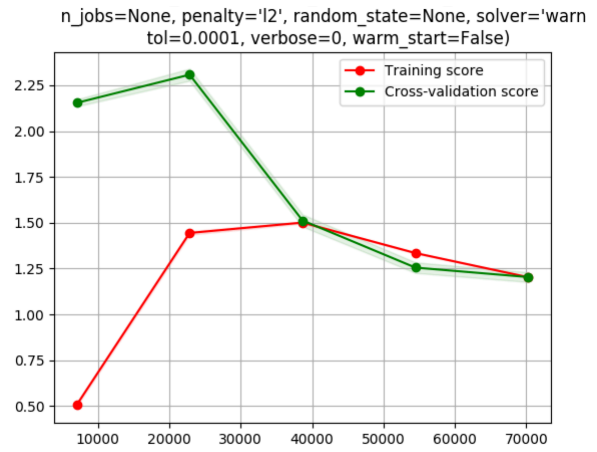
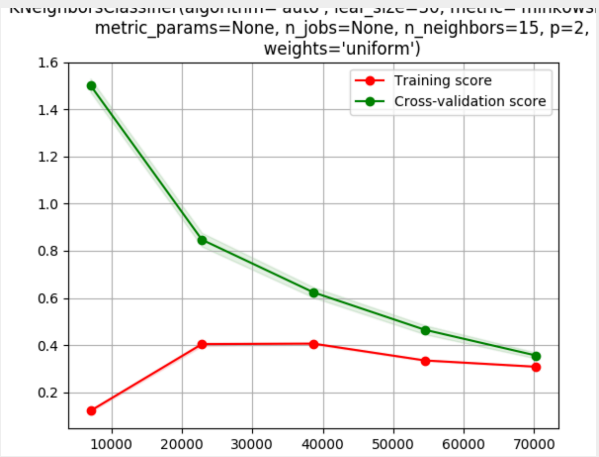
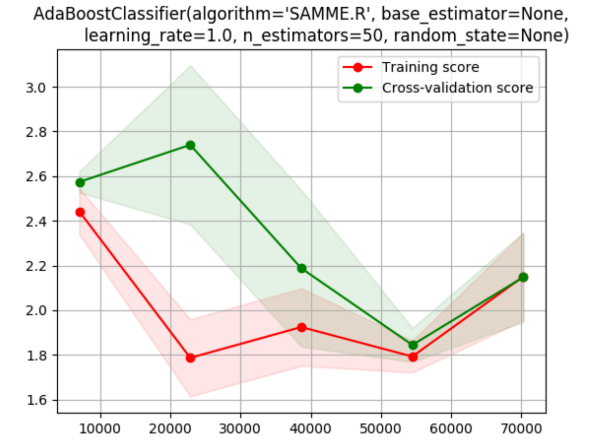
3.6. ADA Boost

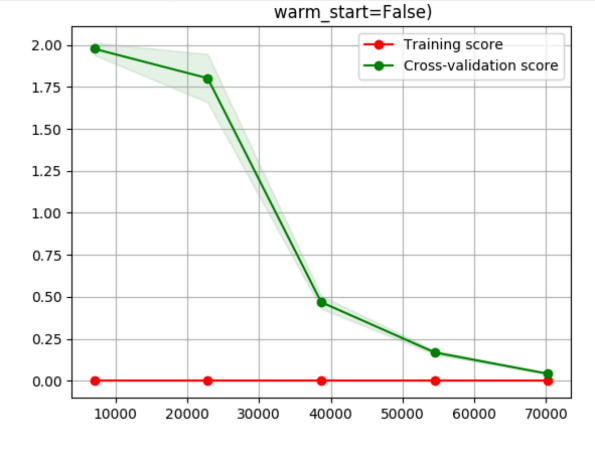
Metric	With feature reduction features = 17
Accuracy	0.8311287534413210
Precision	0.83552609234583
Recall	0.8311287534413210
F-1 Measure	0.8266541980120600
Mathews Correlation Coefficient	0.791807400866536
AUC	Class 1: 0.8901618247707649
	Class 2: 0.9639244063867358
	Class 3: 0.8405606414287916
	Class 4: 0.8749663781741921
	Class 5: 0.8674972464161144

Metric	With feature reduction features = 17
Confusion Matrix	[16042, 66, 125, 28, 4], [62, 13550, 90, 319, 957], [1304, 447, 9797, 4032, 764], [125, 922, 1008, 11660, 1060], [225, 1033, 128, 489, 13858]

3.8. Learning Curves

Gaussian Naive Bayes	
Decision Tree	

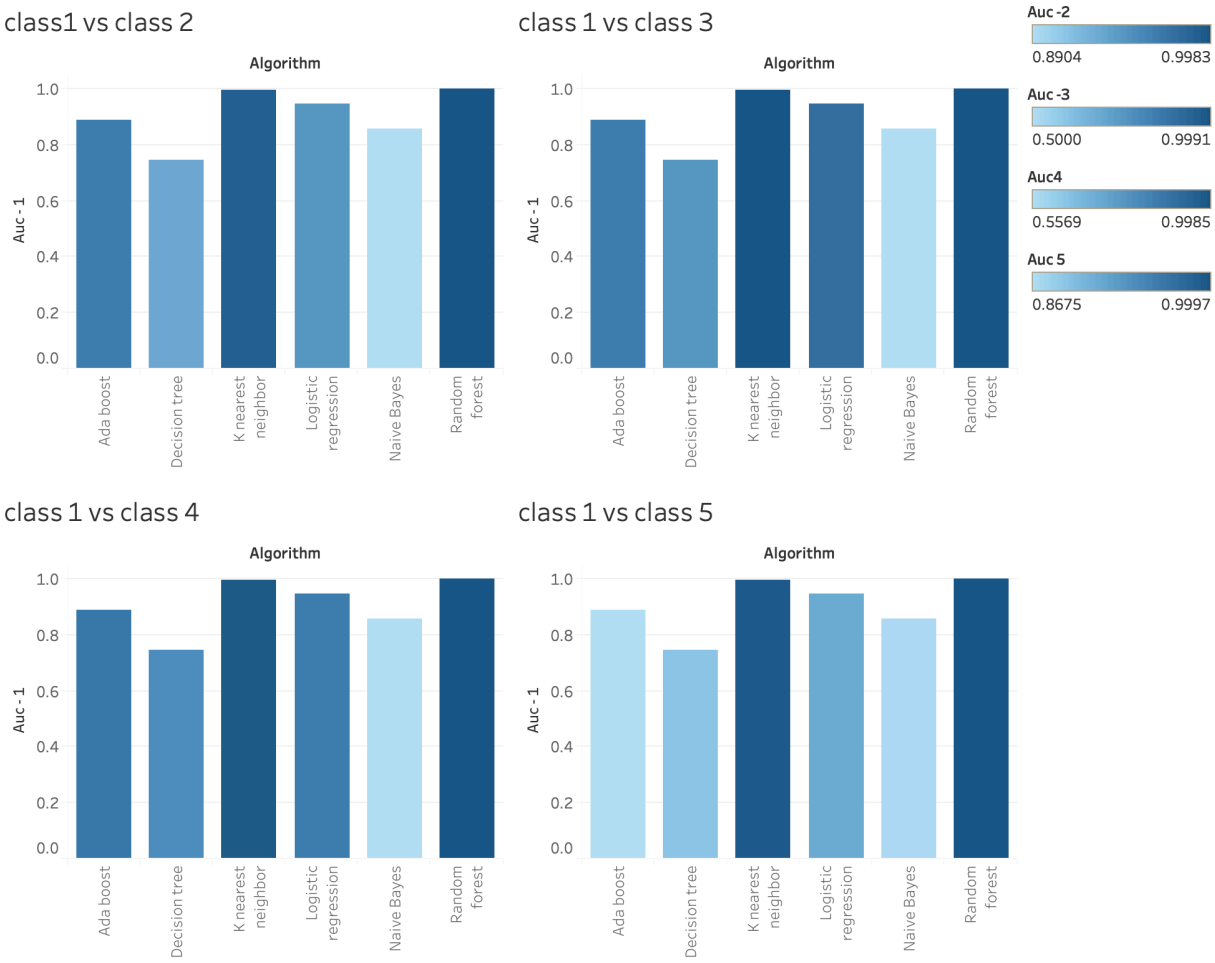
Random Forest	
KNN	
Ada Boost	

Logistic Regression	

3.9. Tableau:

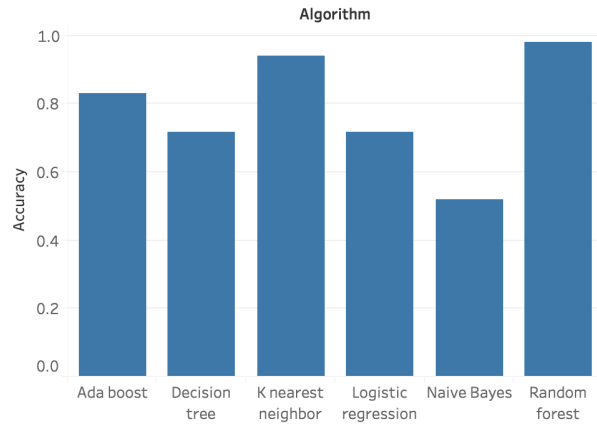
The evaluation metrics are interpreted with the help of Tableau. Various representations were made. Following are the dashboards corresponding to those representation.

A. One class AUC vs rest class AUC

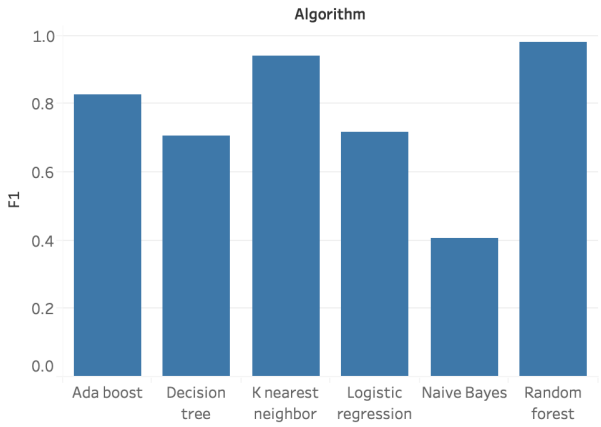


B. Accuracy, precision, recall F1 :

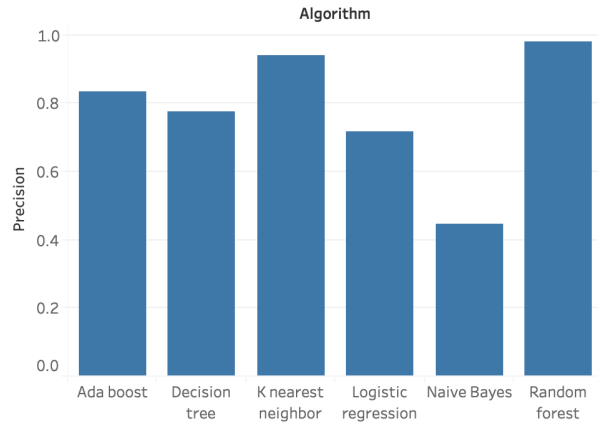
accuracy



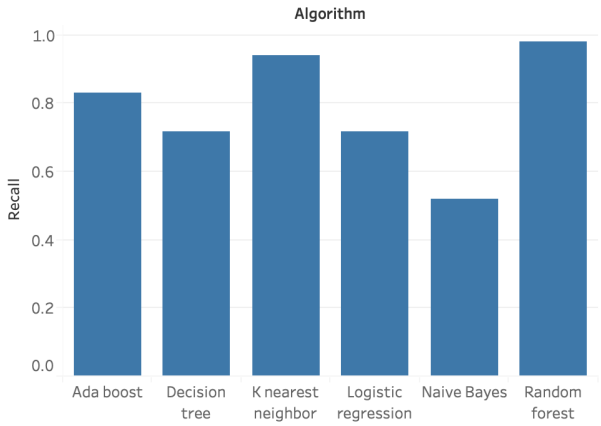
f1



precision

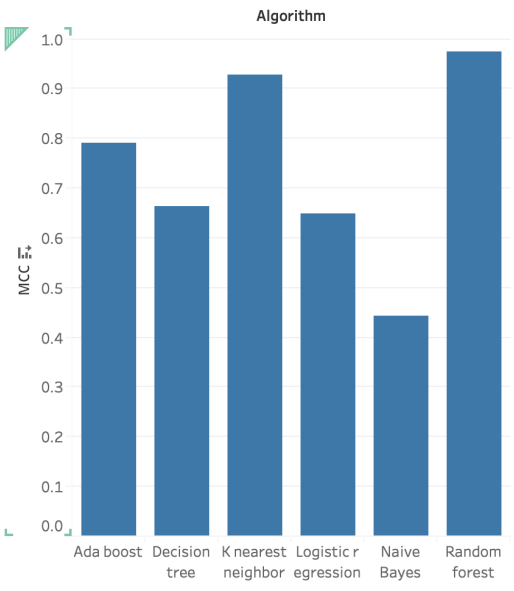


recall



C. Matthew's Correlation coefficient:

MCC



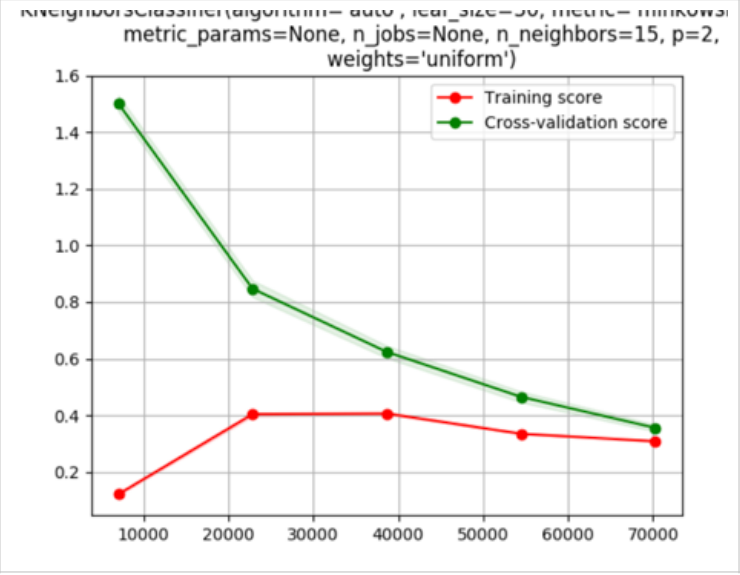
Step 4 : Conclusion

The ROC and AUC for all the classes taking into consideration one vs one is plotted using tableau. It shows that the AUC from KNN and Random Forest give the maximum area under the curve. The higher the area under curve, classifier can distinguish the classes better. AUC for all the classes is better in KNN and Random Forest.

Precision is the ability of classifier not to label a negative sample as positive. Recall is the ability of classifier to find all positive samples. A high value of precision and recall indicates that the classifier has been able to correctly classify both positive and negative classes. F1 Score is the harmonic mean of precision and recall. Higher the value of F1 better the classification is. The precision recall and F1 of Random Forest and KNN looks promising. They have been plotted in the tableau presentation above.

Matthews correlation coefficient measure the quality of classification where the classes are imbalanced. The Matthew's Correlation coefficient for the KNN and Random forest is more than 90% while for all other classifiers its less than 80%.

Learning curves are a measure of error. They are plotted between the Mean square Error and sample size. Ideally, as the sample size increase, the validation and training error should decrease and in the end converge with each other. In our algorithm implementation, for KNN, with a low sample size, the cross validation error is high and with increasing sample size, the cross validation error decreases and in the end is very close to the training set error.



Naive Bayes has very high training and cross-validation error. Logistic Regression and decision tree has no training error and a high cross validation error. In comparison to other algorithms, random forest although suffers from a little overfitting problem but with increasing sample size, performs better.

Hence, taking a holistic view, KNN Classification is the most preferred and followed by random forest algorithm.