FACULTY OF COMPUTING & INFORMATION TECHNOLOGY
KING ABDULAZIZ UNIVERSITY

FCIT
KAU

كلية الحاسبات وتقنية المعلومات
جامعة الملك عبدالعزيز

Department of Computer Science

Fall 2021 - 1st Semester

Project Report

CPCS-433

MACHINE LEARNING PROJECT With R Language

Artificial Intelligence Topics

Student Name	ID	Section
Omar Abdulaziz Alqurashi **(Leader)**	1742589	A
Mohammed Alzahrani	1740166	
Mohammed Saleh Alharbi	1740373	

Supervisor: Dr. Turki T. Turki

December 8, 2020

List of Contents

List of Illustrations

Figures

Tables

1 Selection

The selected features of each dataset are the following:

- **Standard Datasets:**
  - Spambase: The number of the selected features is **57**: Word_freq_make, Word_freq_address, Word_freq_all, Word_freq_3d, Word_freq_our, Word_freq_over, Word_freq_remove, Word_freq_internet, Word_freq_order, Word_freq_mail, Word_freq_receive, Word_freq_will, Word_freq_people, Word_freq_report, Word_freq_addresses, Word_freq_free, Word_freq_business, Word_freq_email, Word_freq_you, Word_freq_credit, Word_freq_your, Word_freq_font, Word_freq_000, Word_freq_money, Word_freq_hp, Word_freq_hpl, Word_freq_george, Word_freq_650, Word_freq_lab, Word_freq_labs, Word_freq_telnet, Word_freq_857, Word_freq_data, Word_freq_415, Word_freq_85, Word_freq_technology, Word_freq_1999, Word_freq_parts, Word_freq_pm, Word_freq_direct, Word_freq_meeting, Word_freq_original, Word_freq_project, Word_freq_re, Word_freq_edu, Word_freq_table, Word_freq_conference, Char_freq1, Char_freq2, Char_freq3, Char_freq4, Char_freq5, Char_freq6, Capital_run_length_average, Capital_run_length_longest, and Capital_run_length_total.

  - Twonorm: The number of the selected features is **20**: A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15, A16, A17, A18, A19, and A20.

- **Imbalanced Datasets:**
  - Glass Identification: The number of the selected features is **9**: RI, Na, Mg, Al, Si, K, Ca, Ba, and Fe.

  - Vehicle: The number of the selected features is **18**: Compactness, Circularity, Distance_circularity, Radius_ratio, Praxis_aspect_ratio, Max_length_aspect_ratio, Scatter_ratio, Elongatedness, Praxis_rectangular, Length_rectangular, Major_variance, Minor_variance, Gyration_radius, Major_skewness, Minor_skewness, Minor_kurtosis, Major_kurtosis, and Hollows_ratio.

2 Preprocessing

There are stages of preprocessing has been done on the datasets. First, the extension of each file of the four datasets has been changed from "dat" into "txt", and their headers has been deleted and the remaining is the data itself. Second, The Oversampling has been applied to the Imbalanced datasets (Glass Identification and Vehicle). (The uploaded dataset files are the edited ones)

3 Transformation

In SVM, and AdaBoost classifiers, the labels for the Glass Identification and Vehicle datasets (positive, negative) are transformed into (+1, -1). For the two other datasets (Spambase and Twonorm), the labels (1, 0) are transformed into (+1, -1).

For the XGBoost classifier, the labels for the Glass Identification and Vehicle datasets (positive, negative) are transformed into (+1, 0). For the two other datasets (Spambase and Twonorm), the labels (1, 0) are transformed into (+1, 0), although that there is no need for the last one, it is good for safety purposes.

4 Data Mining

16 classifiers are the combinations of 4 algorithms (XGBoost, SVM, Random Forests, and AdaBoost), and 4 datasets (Glass Identification, Vehicle, Spambase, and Twonorm).

For the parameters of XGBoost, the number of rounds is 5, and the objective is "binary:hinge". For the SVM, the used kernel is the default in the R language (radial basis kernel).

5 Interpretation/Evaluation

5.1 ROC Curves Plots

Each of the following four **Figures (1,2,3,4)** represents the performance (AUC) of all classifiers on the specified dataset through ROC Curves.

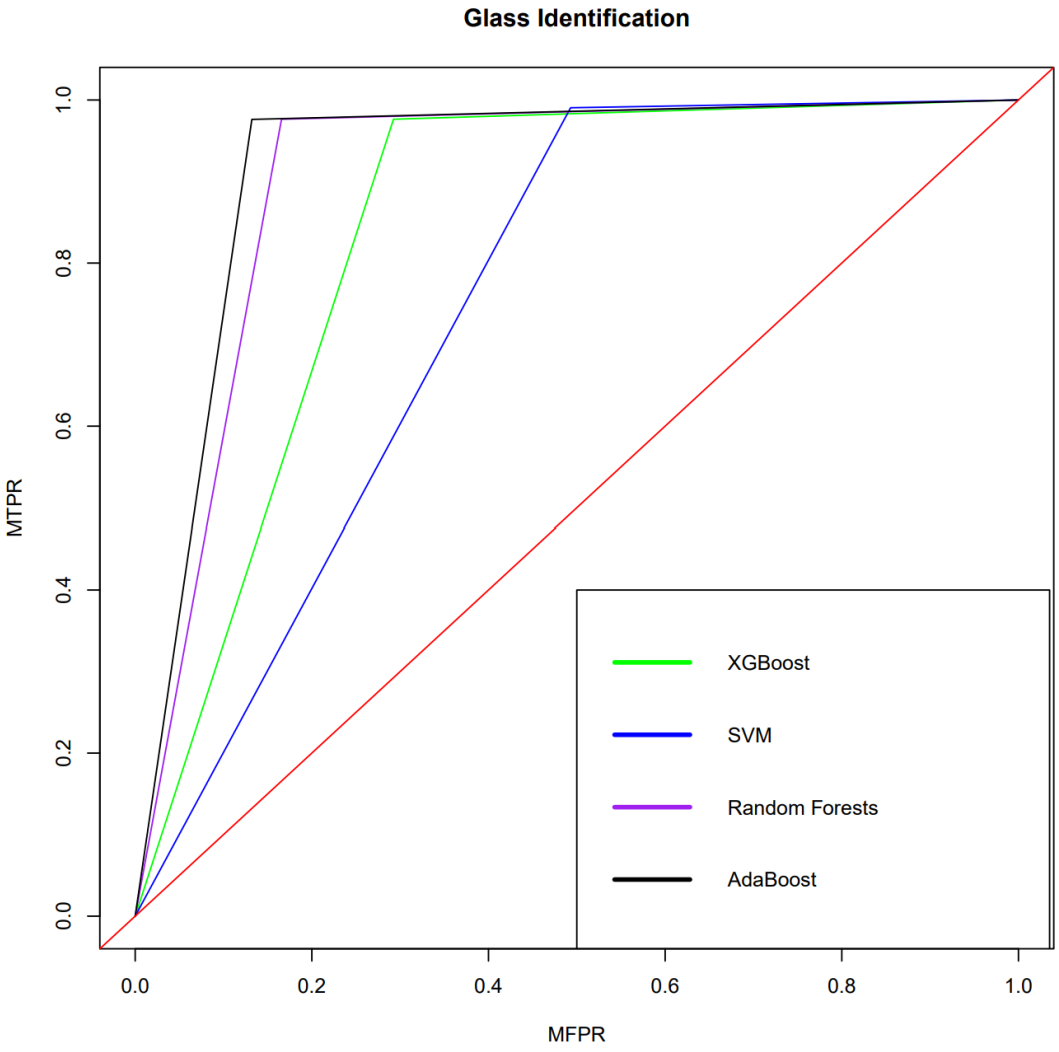

Figure 1: ROC Curves Plot for glass0 dataset

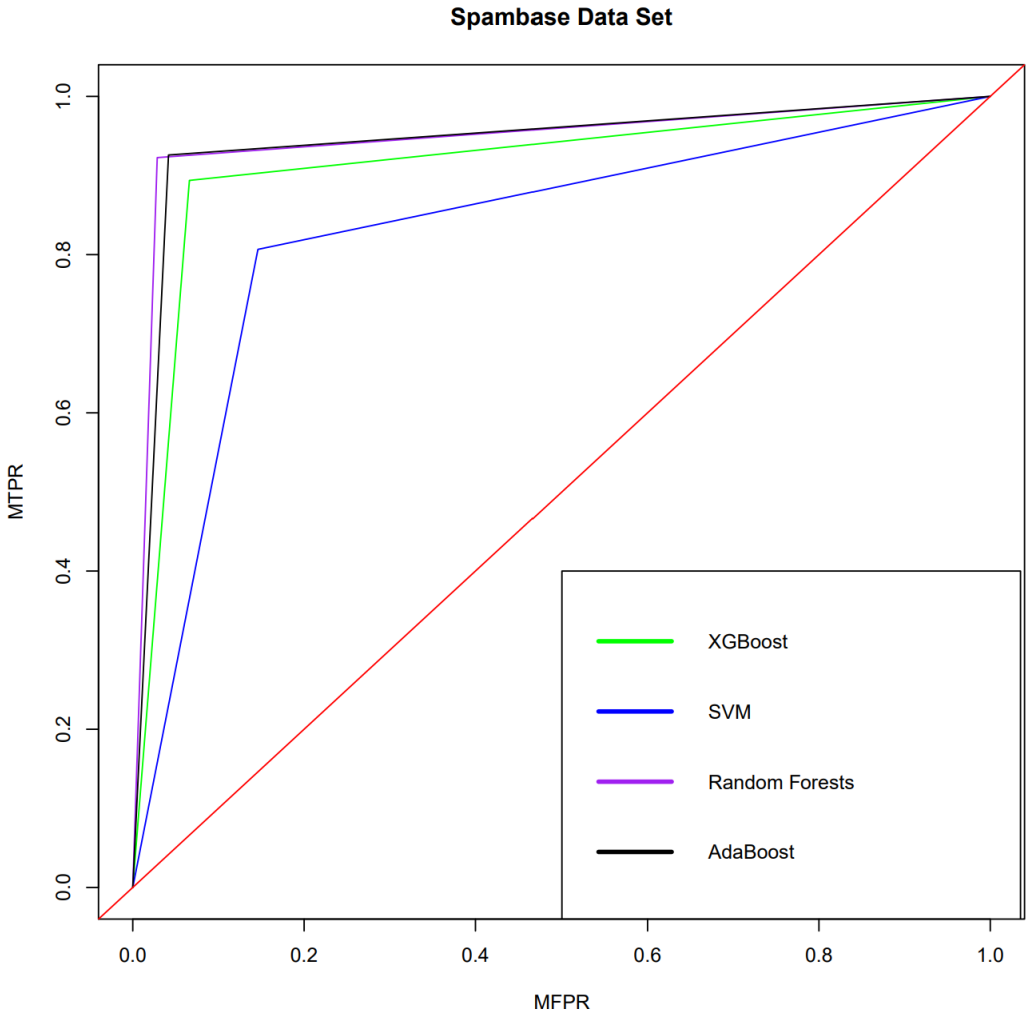

Figure 2: ROC Curves Plot for spambase dataset

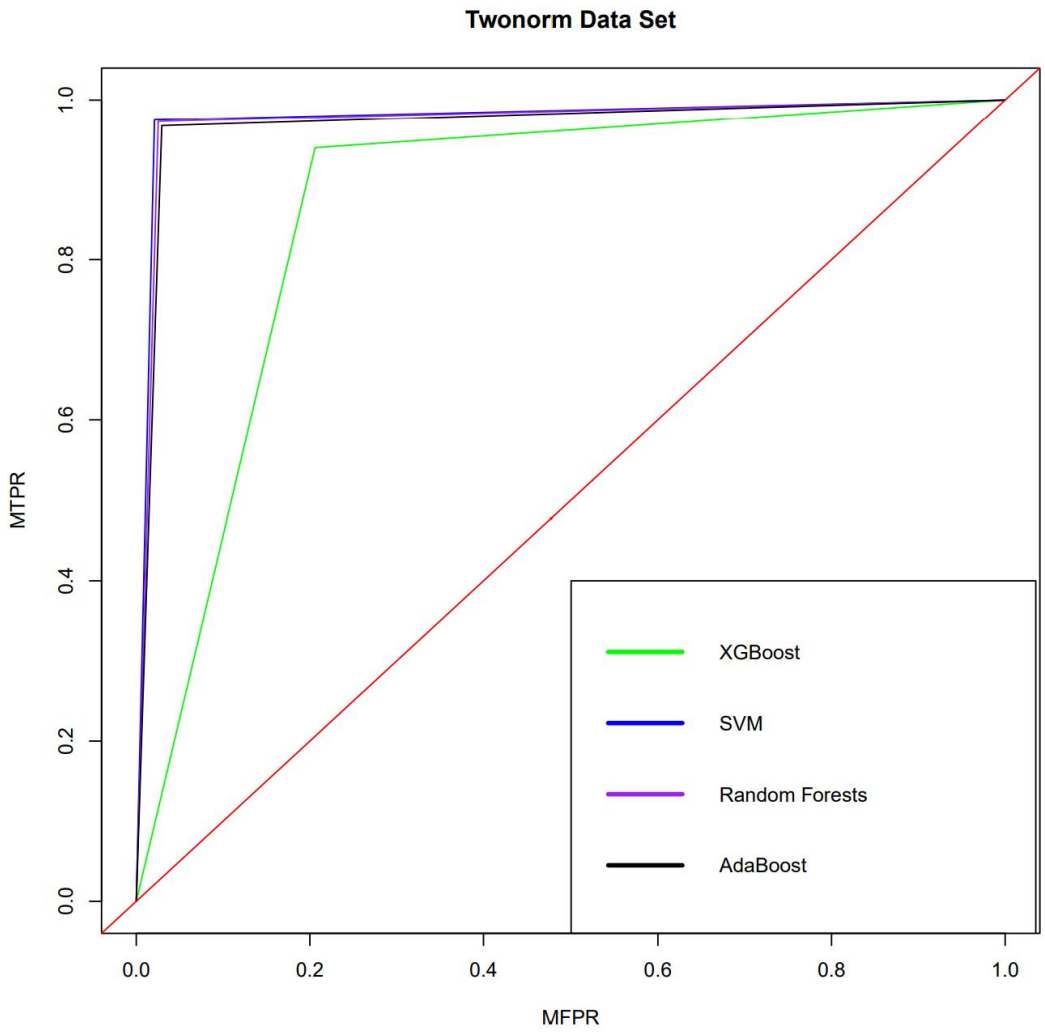

Figure 3: ROC Curves Plot for twonorm dataset

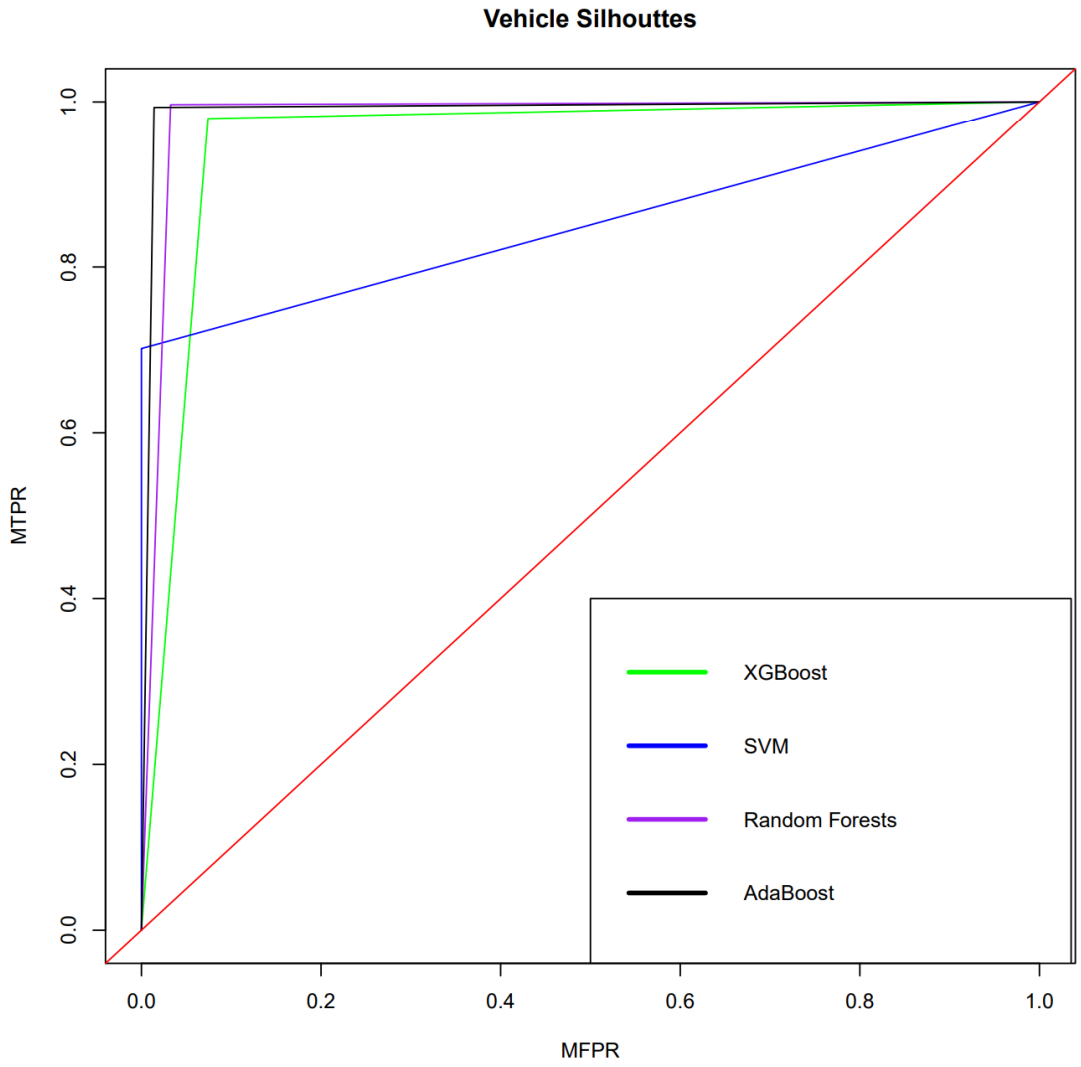

Figure 4: ROC Curves Plot for vehicle0 dataset

5.2 Performance Measures

The performance (MACC, MBAC, MF1, and Standard Deviation) of each classifier on the four datasets is illustrated through the following four tables: **Table (1,2,3,4)**

Table 1: Performance Measures and Standard Deviation for XGBoost

XGBoost		
Dataset\Measure Types	**Performance Measures**	**Standard Deviation**
glass0	MACC = 0.867510	0.037816
	MBAC = 0.841975	0.043177
	MF1 = 0.897450	0.027898
spambase	MACC = 0.917986	0.009063
	MBAC = 0.913859	0.010678
	MF1 = 0.895769	0.010239
twonorm	MACC = 0.866892	0.012559
	MBAC = 0.866909	0.013746
	MF1 = 0.875943	0.009554
vehicle0	MACC = 0.951772	0.012034
	MBAC = 0.952784	0.011698
	MF1 = 0.950967	0.013170

Table 2: Performance Measures and Standard Deviation for SVM

SVM		
Dataset\Measure Types	**Performance Measures**	**Standard Deviation**
glass0	MACC = 0.793985	0.025447
	MBAC = 0.748751	0.029246
	MF1 = 0.850736	0.018178
spambase	MACC = 0.835097	0.017753
	MBAC = 0.830275	0.016858
	MF1 = 0.794183	0.019549
twonorm	MACC = 0.976892	0.002210
	MBAC = 0.976899	0.002309
	MF1 = 0.976774	0.002732
vehicle0	MACC = 0.856866	0.033181
	MBAC = 0.850798	0.034293
	MF1 = 0.823020	0.050052

Table 3: Performance Measures and Standard Deviation for Random Forests

Random Forests		
Dataset\Measure Types	**Performance Measures**	**Standard Deviation**
glass0	MACC = 0.918262	0.024626
	MBAC = 0.905366	0.025299
	MF1 = 0.933773	0.020438
spambase	MACC = 0.952355	0.010442
	MBAC = 0.947042	0.011490
	MF1 = 0.938572	0.013464
twonorm	MACC = 0.973784	0.004835
	MBAC = 0.973791	0.004873
	MF1 = 0.973652	0.005422
vehicle0	MACC = 0.981503	0.008361
	MBAC = 0.982065	0.008377
	MF1 = 0.981011	0.008505

Table 4: Performance Measures and Standard Deviation for AdaBoost

AdaBoost		
Dataset\Measure Types	**Performance Measures**	**Standard Deviation**
glass0	MACC = 0.932349	0.024870
	MBAC = 0.922052	0.024854
	MF1 = 0.944633	0.020770
spambase	MACC = 0.945616	0.010235
	MBAC = 0.942116	0.011332
	MF1 = 0.930588	0.013297
twonorm	MACC = 0.969054	0.003651
	MBAC = 0.969028	0.003775
	MF1 = 0.968890	0.004484
vehicle0	MACC = 0.989539	0.009706
	MBAC = 0.989593	0.009928
	MF1 = 0.989195	0.010036

5.3 Statistical Tests

This section shows the statistical comparison between each classifier on the four datasets using the performance measures (MACC, MBAC, and MF1); including the rank of each classifier (Using Friedman ranking test), raw p-value, and the corrected p-value (Using correction by BH). These things are illustrated through the following tables: **Table (5,6,7)**

(Note that no classifier is statistically significant from the others)

Table 5: Statistical Comparison for MACC

MACC				
Summary				
	XGBoost	**SVM**	**RandomForests**	**AdaBoost**
	0.9010398	0.8657101	0.9564761	0.9591393
raw.pval				
Classifier	**XGBoost**	**SVM**	**RandomForests**	**AdaBoost**
XGBoost	NA	1.0000000	0.1003482	0.1003482
SVM	1.0000000	NA	0.1003482	0.1003482
RandomForests	0.1003482	0.1003482	NA	1.0000000
AdaBoost	0.1003482	0.1003482	1.0000000	NA
corrected.pval				
Classifier	**XGBoost**	**SVM**	**RandomForests**	**AdaBoost**
XGBoost	NA	1.0000000	0.6020895	0.6020895
SVM	1.0000000	NA	0.6020895	0.6020895
RandomForests	0.6020895	0.6020895	NA	1.0000000
AdaBoost	0.6020895	0.6020895	1.0000000	NA

Table 6: Statistical Comparison for MBAC

MBAC				
Summary				
	XGBoost	**SVM**	**RandomForests**	**AdaBoost**
	0.8938818	0.8516807	0.9520658	0.9556972
raw.pval				
Classifier	**XGBoost**	**SVM**	**RandomForests**	**AdaBoost**
XGBoost	NA	1.0000000	0.1003482	0.1003482
SVM	1.0000000	NA	0.1003482	0.1003482
RandomForests	0.1003482	0.1003482	NA	1.0000000
AdaBoost	0.1003482	0.1003482	1.0000000	NA
corrected.pval				
Classifier	**XGBoost**	**SVM**	**RandomForests**	**AdaBoost**
XGBoost	NA	1.0000000	0.6020895	0.6020895
SVM	1.0000000	NA	0.6020895	0.6020895
RandomForests	0.6020895	0.6020895	NA	1.0000000
AdaBoost	0.6020895	0.6020895	1.0000000	NA

Table 7: Statistical Comparison for MF1

MF1				
Summary				
	XGBoost	**SVM**	**RandomForests**	**AdaBoost**
	0.9050322	0.8611782	0.9567521	0.9583262
raw.pval				
Classifier	**XGBoost**	**SVM**	**RandomForests**	**AdaBoost**
XGBoost	NA	1.0000000	0.1003482	0.1003482
SVM	1.0000000	NA	0.1003482	0.1003482
RandomForests	0.1003482	0.1003482	NA	1.0000000
AdaBoost	0.1003482	0.1003482	1.0000000	NA
corrected.pval				
Classifier	**XGBoost**	**SVM**	**RandomForests**	**AdaBoost**
XGBoost	NA	1.0000000	0.6020895	0.6020895
SVM	1.0000000	NA	0.6020895	0.6020895
RandomForests	0.6020895	0.6020895	NA	1.0000000
AdaBoost	0.6020895	0.6020895	1.0000000	NA

Appendix

```
> source("source_code.txt")
[1] "----------------= Students' Info =----------------"
[1] "Names:"
[1] "- Omar Alqurashi     , ID: 1742589"
[1] "- Mohammed Alzahrani, ID: 1740166"
[1] "- Mohammed Alharbi  , ID: 1740373"
[1] "2020-12-08 18:35:57 +03"
[1] "2020-12-08"
[1] "====================Start====================="
[1] "----------------Preparation-----------------"
[1] "Glass Dimensions:"
[1] 214  10
[1] "Spambase Dimensions:"
[1] 4597   58
[1] "Twonorm Dimensions:"
[1] 7400   21
[1] "Vehicle Dimensions:"
[1] 846  19
[1] "====================Oversampling====================="
[1] "-----------------before-----------------"
[1] "Num. Records in Glass Dataset Labeled as 'positive':"
[1] 70
[1] "Num. Records in Glass Dataset Labeled as 'negative':"
[1] 144
[1] "Num. Records in Vehicle Dataset Labeled as 'positive':"
[1] 199
[1] "Num. Records in Vehicle Dataset Labeled as 'negative':"
[1] 647
[1] "-----------------after-----------------"
[1] "Num. Records in Glass Dataset Labeled as 'positive':"
[1] 210
[1] "Num. Records in Glass Dataset Labeled as 'negative':"
[1] 144
[1] "Num. Records in Vehicle Dataset Labeled as 'positive':"
[1] 597
[1] "Num. Records in Vehicle Dataset Labeled as 'negative':"
[1] 647
[1] "====================Oversampling End====================="
[1] ""
```

```
[1] "==== Start 5-Folds Cross-Validation ===="
[1] ""
[1] "                 i = 1                 "
[1] "=================== Training... ==================="
[1]   train-error:0.397887
[2]   train-error:0.084507
[3]   train-error:0.038732
[4]   train-error:0.038732
[5]   train-error:0.038732
[1]   train-error:0.601849
[2]   train-error:0.087571
[3]   train-error:0.081044
[4]   train-error:0.064455
[5]   train-error:0.057656
[1]   train-error:0.501858
[2]   train-error:0.208277
[3]   train-error:0.103716
[4]   train-error:0.075507
[5]   train-error:0.057432
[1]   train-error:0.523618
[2]   train-error:0.052261
[3]   train-error:0.033166
[4]   train-error:0.033166
[5]   train-error:0.017085
"----------- XGBoost Trained -----------"
[1] "----------- SVM Trained -----------"
[1] "------ Random Forests Trained -------"
[1] "---------- AdaBoost Trained ----------"
[1] "=================== Testing... ==================="
[1] "----------- XGBoost Tested -----------"
[1] "----------- SVM Tested -----------"
[1] "------ Random Forests Tested -------"
[1] "---------- AdaBoost Tested ----------"
[1] ""
```

```
[1] "_________________ i = 2 _________________"
[1] "============= Training... ============"
[1]   train-error:0.407801
[2]   train-error:0.078014
[3]   train-error:0.056738
[4]   train-error:0.056738
[5]   train-error:0.056738
[1]   train-error:0.607017
[2]   train-error:0.164264
[3]   train-error:0.069350
[4]   train-error:0.048409
[5]   train-error:0.046777
[1]   train-error:0.494595
[2]   train-error:0.226689
[3]   train-error:0.110135
[4]   train-error:0.081926
[5]   train-error:0.055405
[1]   train-error:0.520080
[2]   train-error:0.040161
[3]   train-error:0.017068
[4]   train-error:0.017068
[5]   train-error:0.016064
[1] "---------- XGBoost Trained -----------"
[1] "------------ SVM Trained -------------"
[1] "------ Random Forests Trained --------"
[1] "--------- AdaBoost Trained -----------"
[1] "============= Testing... ============="
[1] "---------- XGBoost Tested ------------"
[1] "------------- SVM Tested -------------"
[1] "----- Random Forests Tested ----------"
[1] "--------- AdaBoost Tested ------------"
[1] ""
```

```
[1] "_________________ i = 3 _________________"
[1] "============= Training... ============"
[1]   train-error:0.415493
[2]   train-error:0.080986
[3]   train-error:0.042254
[4]   train-error:0.042254
[5]   train-error:0.035211
[1]   train-error:0.610114
[2]   train-error:0.107667
[3]   train-error:0.065797
[4]   train-error:0.053834
[5]   train-error:0.052202
[1]   train-error:0.504054
[2]   train-error:0.226858
[3]   train-error:0.120101
[4]   train-error:0.084459
[5]   train-error:0.056588
[1]   train-error:0.526633
[2]   train-error:0.040201
[3]   train-error:0.032161
[4]   train-error:0.026131
[5]   train-error:0.026131
[1] "---------- XGBoost Trained -----------"
[1] "------------ SVM Trained -------------"
[1] "------ Random Forests Trained --------"
[1] "--------- AdaBoost Trained -----------"
[1] "============= Testing... ============="
[1] "---------- XGBoost Tested ------------"
[1] "------------- SVM Tested -------------"
[1] "----- Random Forests Tested ----------"
[1] "--------- AdaBoost Tested ------------"
[1] ""
```

```
[1] "_________________ i = 4 _________________"
[1] "============= Training... ============"
[1]   train-error:0.406360
[2]   train-error:0.084806
[3]   train-error:0.028269
[4]   train-error:0.028269
[5]   train-error:0.028269
[1]   train-error:0.606745
[2]   train-error:0.122926
[3]   train-error:0.061463
[4]   train-error:0.062823
[5]   train-error:0.049497
[1]   train-error:0.501351
[2]   train-error:0.181081
[3]   train-error:0.109797
[4]   train-error:0.076351
[5]   train-error:0.051182
[1]   train-error:0.511558
[2]   train-error:0.038191
[3]   train-error:0.029146
[4]   train-error:0.027136
[5]   train-error:0.024121
[1] "---------- XGBoost Trained -----------"
[1] "------------ SVM Trained -------------"
[1] "------ Random Forests Trained --------"
[1] "--------- AdaBoost Trained -----------"
[1] "============= Testing... ============="
[1] "---------- XGBoost Tested ------------"
[1] "------------- SVM Tested -------------"
[1] "----- Random Forests Tested ----------"
[1] "--------- AdaBoost Tested ------------"
[1] ""
```

```
[1] "_________________ i = 5 _________________"
[1] "============= Training... ============"
[1]   train-error:0.406360
[2]   train-error:0.063604
[3]   train-error:0.045936
[4]   train-error:0.045936
[5]   train-error:0.045936
[1]   train-error:0.603425
[2]   train-error:0.101386
[3]   train-error:0.063876
[4]   train-error:0.060342
[5]   train-error:0.050829
[1]   train-error:0.500169
[2]   train-error:0.196115
[3]   train-error:0.106250
[4]   train-error:0.081419
[5]   train-error:0.055574
[1]   train-error:0.518593
[2]   train-error:0.039196
[3]   train-error:0.028141
[4]   train-error:0.028141
[5]   train-error:0.025126
[1] "---------- XGBoost Trained -----------"
[1] "------------ SVM Trained -------------"
[1] "------ Random Forests Trained --------"
[1] "--------- AdaBoost Trained -----------"
[1] "============= Testing... ============="
[1] "---------- XGBoost Tested ------------"
[1] "------------- SVM Tested -------------"
[1] "----- Random Forests Tested ----------"
[1] "--------- AdaBoost Tested ------------"
[1] ""
```

```
[1] "---------Calculations of Performance Measures----------"
[1] "------------------------XGBoost-------------------------"
[1] "glass0:"
[1] "MACC = 0.867510, SD = 0.037816"
[1] "MBAC = 0.841975, SD = 0.043177"
[1] "MF1  = 0.897450, SD = 0.027898"
[1] "spambase:"
[1] "MACC = 0.917986, SD = 0.009063"
[1] "MBAC = 0.913859, SD = 0.010678"
[1] "MF1  = 0.895769, SD = 0.010239"
[1] "twonorm:"
[1] "MACC = 0.866892, SD = 0.012559"
[1] "MBAC = 0.866909, SD = 0.013746"
[1] "MF1  = 0.875943, SD = 0.009554"
[1] "vehicle0:"
[1] "MACC = 0.951772, SD = 0.012034"
[1] "MBAC = 0.952784, SD = 0.011698"
[1] "MF1  = 0.950967, SD = 0.013170"
[1] "--------------------------SVM--------------------------"
[1] "glass0:"
[1] "MACC = 0.793985, SD = 0.025447"
[1] "MBAC = 0.748751, SD = 0.029246"
[1] "MF1  = 0.850736, SD = 0.018178"
[1] "spambase:"
[1] "MACC = 0.835097, SD = 0.017753"
[1] "MBAC = 0.830275, SD = 0.016858"
[1] "MF1  = 0.794183, SD = 0.019549"
[1] "twonorm:"
[1] "MACC = 0.976892, SD = 0.002210"
[1] "MBAC = 0.976899, SD = 0.002309"
[1] "MF1  = 0.976774, SD = 0.002732"
[1] "vehicle0:"
[1] "MACC = 0.856866, SD = 0.033181"
[1] "MBAC = 0.850798, SD = 0.034293"
[1] "MF1  = 0.823020, SD = 0.050052"
```

```
[1] "------------------Random Forests----------------------"
[1] "glass0:"
[1] "MACC = 0.918262, SD = 0.024626"
[1] "MBAC = 0.905366, SD = 0.025299"
[1] "MF1  = 0.933773, SD = 0.020438"
[1] "spambase:"
[1] "MACC = 0.952355, SD = 0.010442"
[1] "MBAC = 0.947042, SD = 0.011490"
[1] "MF1  = 0.938572, SD = 0.013464"
[1] "twonorm:"
[1] "MACC = 0.973784, SD = 0.004835"
[1] "MBAC = 0.973791, SD = 0.004873"
[1] "MF1  = 0.973652, SD = 0.005422"
[1] "vehicle0:"
[1] "MACC = 0.981503, SD = 0.008361"
[1] "MBAC = 0.982065, SD = 0.008377"
[1] "MF1  = 0.981011, SD = 0.008505"
[1] "----------------------AdaBoost-----------------------"
[1] "glass0:"
[1] "MACC = 0.932349, SD = 0.024870"
[1] "MBAC = 0.922052, SD = 0.024854"
[1] "MF1  = 0.944633, SD = 0.020770"
[1] "spambase:"
[1] "MACC = 0.945616, SD = 0.010235"
[1] "MBAC = 0.942116, SD = 0.011332"
[1] "MF1  = 0.930588, SD = 0.013297"
[1] "twonorm:"
[1] "MACC = 0.969054, SD = 0.003651"
[1] "MBAC = 0.969028, SD = 0.003775"
[1] "MF1  = 0.968890, SD = 0.004484"
[1] "vehicle0:"
[1] "MACC = 0.989539, SD = 0.009706"
[1] "MBAC = 0.989593, SD = 0.009928"
[1] "MF1  = 0.989195, SD = 0.010036"
[1] "============== Cross-Validation Finished ==============="
[1] ""
```

```
[1] "================== Statistical Tests ==================="
[1] "------------------------MACC---------------------------"
$summary
       XGBoost       SVM RandomForests  AdaBoost
[1,] 0.9010398 0.8657101     0.9564761 0.9591393

$raw.pval
              XGBoost       SVM RandomForests  AdaBoost
XGBoost            NA 1.0000000     0.1003482 0.1003482
SVM         1.0000000        NA     0.1003482 0.1003482
RandomForests 0.1003482 0.1003482            NA 1.0000000
AdaBoost      0.1003482 0.1003482     1.0000000        NA

$corrected.pval
              XGBoost       SVM RandomForests  AdaBoost
XGBoost            NA 1.0000000     0.6020895 0.6020895
SVM         1.0000000        NA     0.6020895 0.6020895
RandomForests 0.6020895 0.6020895            NA 1.0000000
AdaBoost      0.6020895 0.6020895     1.0000000        NA

[1] "------------------------MBAC---------------------------"
$summary
       XGBoost       SVM RandomForests  AdaBoost
[1,] 0.8938818 0.8516807     0.9520658 0.9556972

$raw.pval
              XGBoost       SVM RandomForests  AdaBoost
XGBoost            NA 1.0000000     0.1003482 0.1003482
SVM         1.0000000        NA     0.1003482 0.1003482
RandomForests 0.1003482 0.1003482            NA 1.0000000
AdaBoost      0.1003482 0.1003482     1.0000000        NA

$corrected.pval
              XGBoost       SVM RandomForests  AdaBoost
XGBoost            NA 1.0000000     0.6020895 0.6020895
SVM         1.0000000        NA     0.6020895 0.6020895
RandomForests 0.6020895 0.6020895            NA 1.0000000
AdaBoost      0.6020895 0.6020895     1.0000000        NA
```

```
[1] "------------------------MF1----------------------------"
$summary
       XGBoost       SVM RandomForests  AdaBoost
[1,] 0.9050322 0.8611782     0.9567521 0.9583262

$raw.pval
              XGBoost       SVM RandomForests  AdaBoost
XGBoost            NA 1.0000000     0.1003482 0.1003482
SVM         1.0000000        NA     0.1003482 0.1003482
RandomForests 0.1003482 0.1003482            NA 1.0000000
AdaBoost      0.1003482 0.1003482     1.0000000        NA

$corrected.pval
              XGBoost       SVM RandomForests  AdaBoost
XGBoost            NA 1.0000000     0.6020895 0.6020895
SVM         1.0000000        NA     0.6020895 0.6020895
RandomForests 0.6020895 0.6020895            NA 1.0000000
AdaBoost      0.6020895 0.6020895     1.0000000        NA

[1] "========================END============================"
[1] "-----------------= Students' Info =-------------------"
[1] "Names:"
[1] "- Omar Alqurashi    , ID: 1742589"
[1] "- Mohammed Alzahrani, ID: 1740166"
[1] "- Mohammed Alharbi  , ID: 1740373"
[1] "2020-12-08 18:38:14 +03"
[1] "2020-12-08"
>
```