NO SHOWS FOR MEDICAL APPOINTMENTS

CS 699 DATA MINING PROJECT
BY:
ASHI CHOUDHARY
SHRUTIKA SINGHODIA

TABLE OF CONTENTS

INTRODUCTION

- **DATA MINING GOAL**
- **DESCRIPTION OF DATASET**
- **DATA ATTRIBUTES**

DATA PREPROCESSING

- **DATA REDUCTION**
- **DATA CLEANING - MISSING VALUES**
- **REMOVING DUPLICATIONS**

DATA MINING

- **ATTRIBUTE SELECTION**
- **DATA MINING FILES**

CONCLUSION

- **ANALYSIS AND RECOMMENDATIONS**
- **TAKEAWAYS**

Data Mining is the mining, or discovery, in terms of patterns or rules from huge amounts of data, a new information is developed. To be useful, data mining must be carried out efficiently on large files and databases.

1. Prediction: Determine how certain attributes will behave in the future. For example, how much sales volume a store will generate in a given period.

2. Identification: Identify patterns in data. For example, newly wed couples tend to spend more money buying furniture.

3. Classification: Partition data into classes. For example, customers can be classified into different categories with different behavior in shopping.

4. Optimization: Optimize the use of limited resources such as time, space, money or materials. For example, how to best use advertising to maximize profits (sales).

DATA MINING GOAL

Our main goal in this project is to show that , after making an appointment with a doctor, receiving all the instructions from the doctor, the person does not show up for the appointment. Why? Who to blame it for? In this project, our data mining goal is to predict whether the patient will show up or not show up for the appointment after setting it up.

DESCRIPTION OF DATASET

In this modern era, people are engaged with their crucial stuff. Which is the reason why they miss their medical appointment. Missed appointments or no-shows are defined as “patients who neither kept nor canceled their scheduled appointments”.

This data set collected data for appointments created during the months of April, May and June in the city of Vitória, in Brazil. It collected 110527 records with unique appointment IDs. In which the most crucial is the patient show up or not show up to the appointment.Surveys or Studies conducted previously in primary care settings found that the rates of missed appointments in the United States vary from 5% to 55%. Missed appointments or no-shows are defined as “patients who neither kept nor canceled scheduled appointments”. Missed appointments cost the United States healthcare system more than $150 billion a year.

It causes disruption in the continuity of the provision of healthcare services, adds to the dissatisfaction of patients due to delays in getting new appointments and hinders the detection and treatment of diseases. The rates of missed appointments vary between countries and healthcare systems. Studies conducted previously in primary care settings found that the rates of missed appointments were (5%-55%, in different series) in the United States, (29.5%) in Saudi Arabia, (36%) in Israel, and (6.5%-7.7%) in the United Kingdom. Only a few studies have been conducted in Latin American populations.

Data Attributes

Patient_Id: Patient Id is the unique Id, which is the identification of the patient.

Appointment_ID: Appointment Id is the unique attribute which is assign to each patient who wants an appointment schedule.

Gender: Gender is an attribute of the person which identifies the person.

Appointment_Day: This is the day of the actual appointment which is given to each and every individual person who makes the medical appointment.

Age: This is the attribute of the appointment holder which tells how old the person is.

Neighbourhood: It is the place where appointment of the person takes place.

SMS_received: This is the attribute which shows the confirmation of the appointment of the person.

No_Show: This is the no show appointment of the person, which makes an appointment but did not appear.

Scheduled_day: This is the day where all the appointments are scheduled for the person who made a medical appointment.

Hypertension: This is one of the attributes of the patient. By which it got affected.

Alcoholism: This is the second attribute of the patient which leads to medical appointment.

Diabetes: This is the third attribute of the patient which leads to medical appointment.

Handicap: This is the fourth attribute of the patient which leads to the medical appointment.

CLASS ATTRIBUTE

The class attribute in the above dataset is No-show attribute which is either 1 which is Yes/True or 0 which is No/False.

DATA PREPROCESSING

In our dataset, the problem of missing values is not so serious as some of the attributes miss most of the values and some other attributes just miss a few values. In the dataset, we also have checked for doublications.
While cleaning the data, we removed all the ages that were more than 100 and less than 0.
Because of the special situation, we can finish the data missing in an easy way. We could remove the missing values by deleting the attributes and objects that have too many missing values and use the average value to fill the other missing values. On the other hand, if we don't delete the attributes and objects having so many missing values, they will affect the whole data set badly because they will become the outliers with other preprocessing methods. Using jupyter notebook, we then checked for missing values and cleaned the age attribute in the dataset.

1. Checking for missing values and doublications using Jupyter

```
In [21]: df.isnull().any()

Out[21]: PatientId         False
         AppointmentID     False
         Gender            False
         ScheduledDay      False
         AppointmentDay    False
         Age               False
         Neighbourhood     False
         Scholarship       False
         Hipertension      False
         Diabetes          False
         Alcoholism        False
         Handcap           False
         SMS_received      False
         No-show           False
         dtype: bool
```

```
In [22]: df.duplicated().any()

Out[22]: False
```

```
In [29]: # Cleaning age in the dataset
         df = df[(df['Age'] > 0) & (df['Age'] < 95)]
         df
```

Out[29]:

	PatientId	AppointmentID	Gender	ScheduledDay	AppointmentDay	Age	Neighbourhood	Scholarship	Hipertension	Diabetes	Alcoholism	Handca
0	2.990000e+13	5642903	F	2016-04-29T18:38:08Z	2016-04-29T00:00:00Z	62	JARDIM DA PENHA	0	1	0	0	
1	5.590000e+14	5642503	M	2016-04-29T16:08:27Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	0	0	0	
2	4.260000e+12	5642549	F	2016-04-29T16:19:04Z	2016-04-29T00:00:00Z	62	MATA DA PRAIA	0	0	0	0	
3	8.680000e+11	5642828	F	2016-04-29T17:29:31Z	2016-04-29T00:00:00Z	8	PONTAL DE CAMBURI	0	0	0	0	
4	8.840000e+12	5642494	F	2016-04-29T16:07:23Z	2016-04-29T00:00:00Z	56	JARDIM DA PENHA	0	1	1	0	
...	...	...	...	...	...	...	...	...	...	...	...	
110522	2.570000e+12	5651768	F	2016-05-03T09:15:35Z	2016-06-07T00:00:00Z	56	MARIA ORTIZ	0	0	0	0	
110523	3.600000e+12	5650093	F	2016-05-03T07:27:33Z	2016-06-07T00:00:00Z	51	MARIA ORTIZ	0	0	0	0	
110524	1.560000e+13	5630692	F	2016-04-27T16:03:52Z	2016-06-07T00:00:00Z	21	MARIA ORTIZ	0	0	0	0	
110525	9.210000e+13	5630323	F	2016-04-27T15:09:23Z	2016-06-07T00:00:00Z	38	MARIA ORTIZ	0	0	0	0	

ATTRIBUTE SELECTION

Each attribute in the dataset is considered to determine the class attribute and determine whether the person will show up for the appointment or not. The more important the attribute is, the greater impact it has on the model we build. In order to make sure that our models can handle the preprocessed datasets and that we have removed all the less important attributes, we choose five types of attribute selection methods to reduce the preprocessed dataset.

The five types of attribute selection are described below:

1. **GainratioAttributeEval**: This evaluator evaluates the worth of the attribute by measuring the gain ratio with respect to the class.

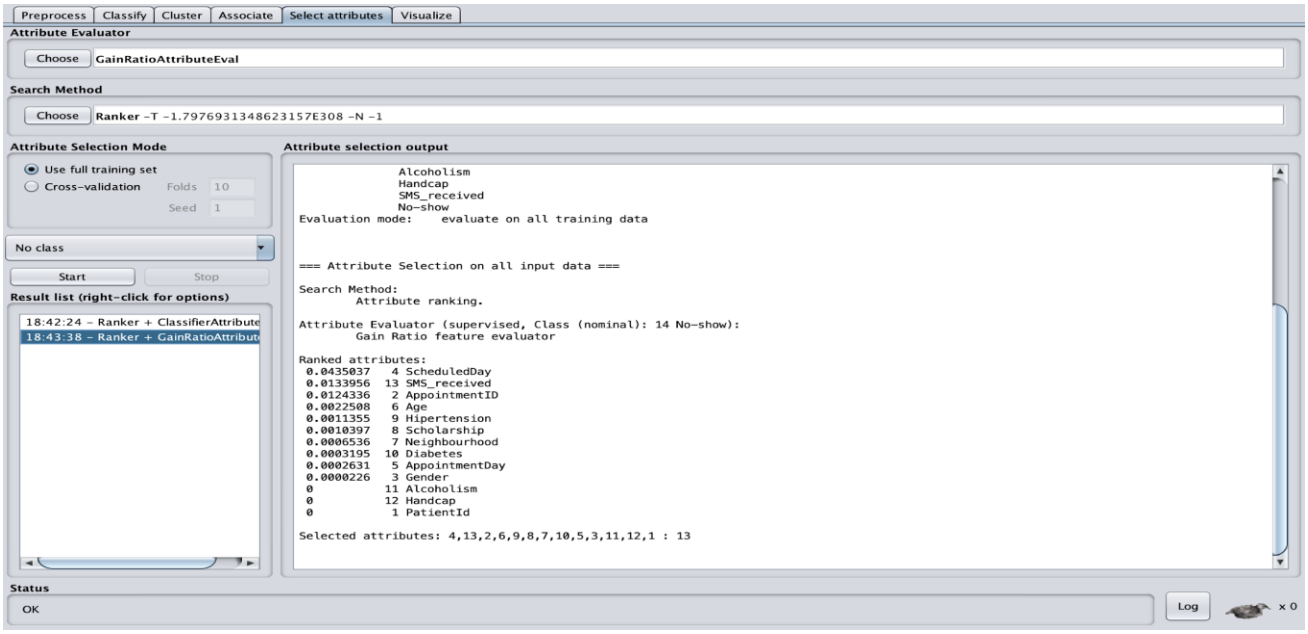
2. **InfoGainAttributeEval:** This evaluator would select attributes that contribute more information (closer to 1, higher information gain value) and remove attributes that can't contribute as much information (closer to 0, lower information gain value).

3. **OneRAttributeEval:** It generates one rule for each attribute and then evaluates the attribute according to the error rate.

4. **ClassifierAttributeEval:** This attribute selection evaluates the worth of an attribute by using a user-specified classifier.

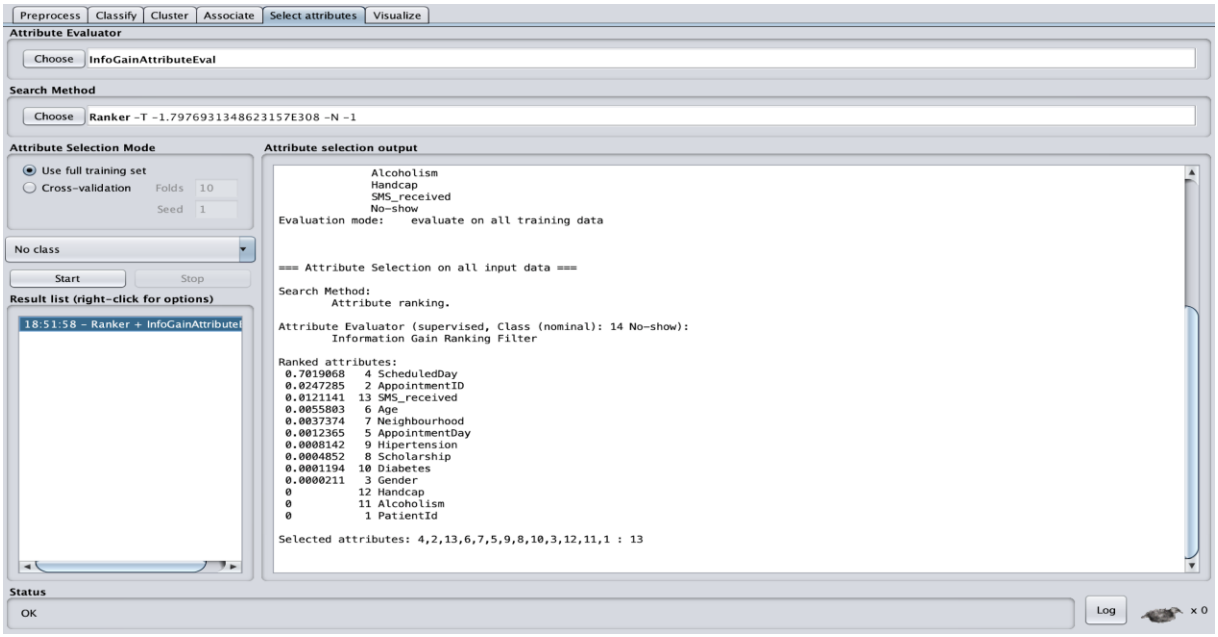
5. **Personal Selection:** We preferred selecting the attribute which helped us determine No-Show.

1. **GainratioAttributeEval(Ranker)**



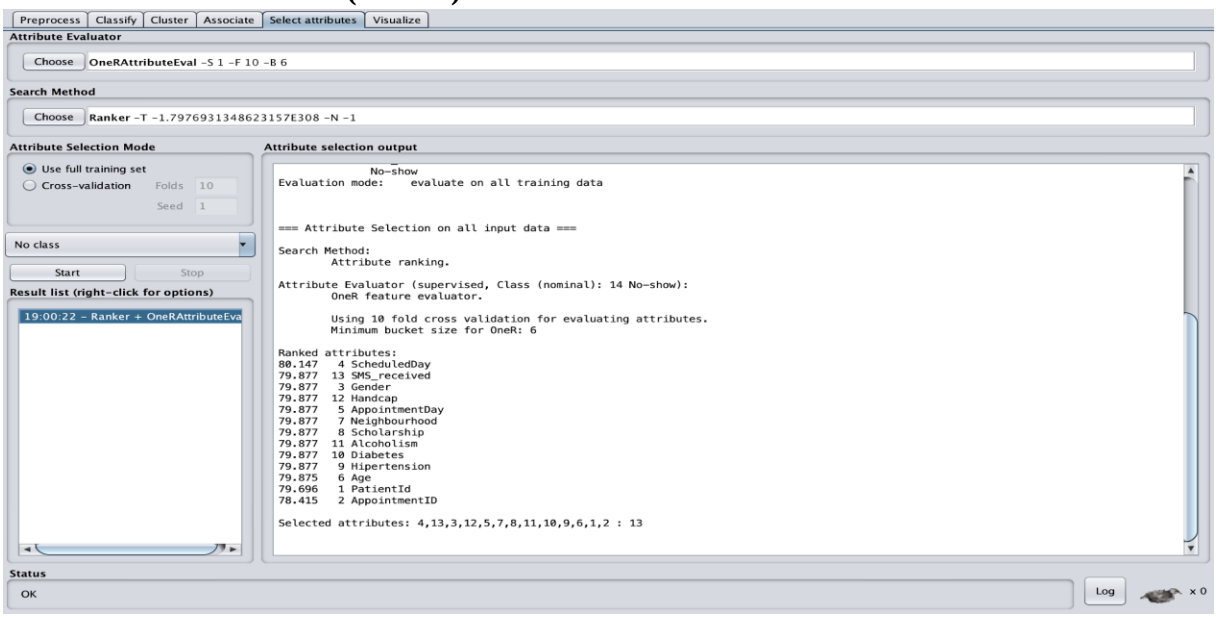
In this attribute selection method, we decided to choose top 7 attributes with the highest ranking. So, we have 4, 13, 2, 6, 9, 8, 7 in our models.

2. **InfoGainAttributeEval(Ranker)**



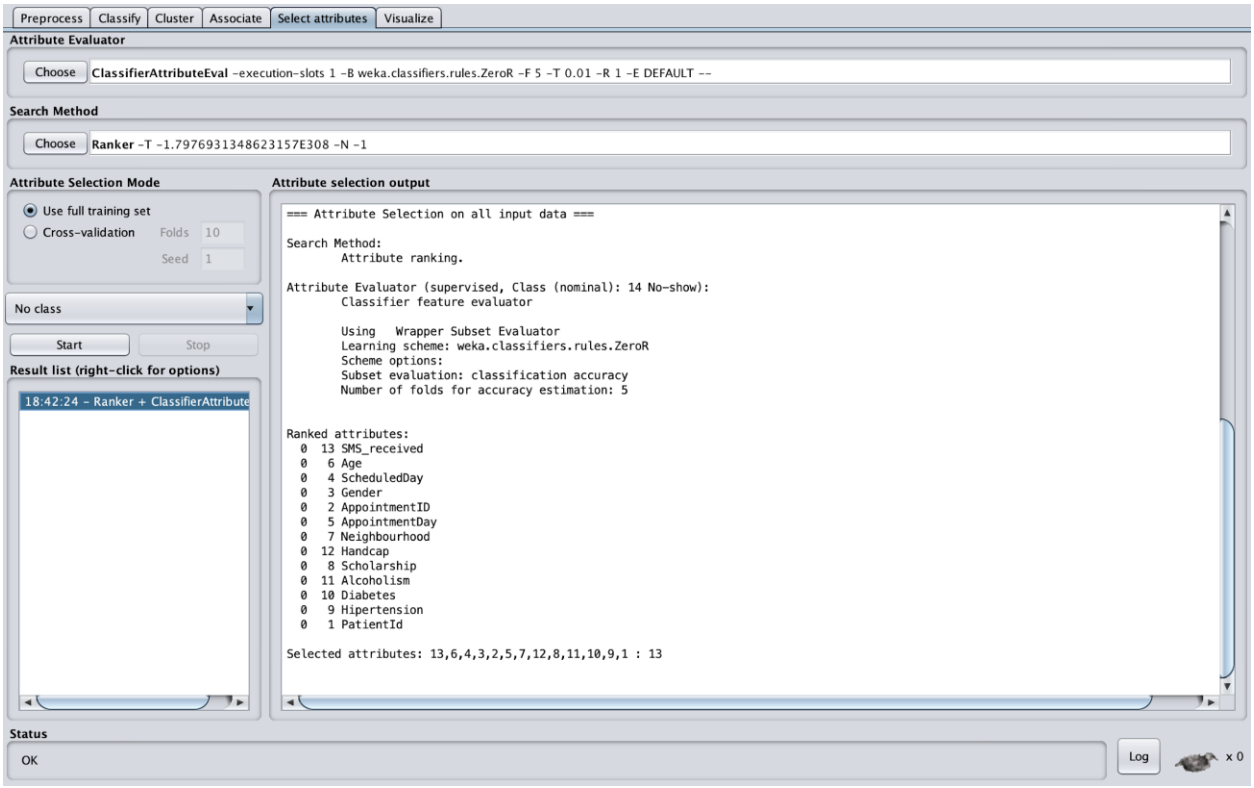
In this attribute selection method, we decided to choose top 7 attributes with the highest ranking. So, we have 4, 2, 13, 6, 7, 5, 9 for our models.

3. OneRAttributeEval(Ranker)



In this attribute selection method, we decided to choose top 7 attributes with the highest ranking. So, we have 4, 13, 3, 12, 5, 7, 8 for our models.

4. ClassifierAttributeEval(Ranker)



In this attribute selection method, we decided to choose top 5 attributes with the highest ranking. So, we have 13, 6, 4, 3, 2 for our models.

5. **Personal Selection:**

In our personal attribute selection method, we decide to use attributes 13, 6, 4, 7, 9, 3, 12 in our models.

DESCRIPTION OF CLASSIFICATION ALGORITHMS:

1. **Naive Bayes**

Naive Bayes is a simple technique for constructing classifiers that models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

2. **J48**

The decision tree method first forms a decision tree based on the training set data. If the tree cannot give correct classification to all objects, then some exceptions are added to the training set data, and the process is repeated until the correct decision set is formed. The decision tree represents the tree structure of the decision set. A decision tree consists of decision nodes, branches, and leaves. The top node in the decision tree is the root node, each branch is a new decision node, or the leaf of the tree. Each decision node represents a question or decision, usually corresponding to the attributes of the object to be classified. Each leaf node represents a possible classification result. During the traversal of the decision tree from top to bottom, a test is encountered at each node, and different test outputs for each node on the node result in different branches, and finally a leaf node is reached. It is the process of using the decision tree
to classify, using several variables to determine the category to which it belongs.

3. **Filtered Classifier**

Class for running an arbitrary classifier on data that has been passed through an arbitrary filter. Like the classifier, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure.

4. **IBk**

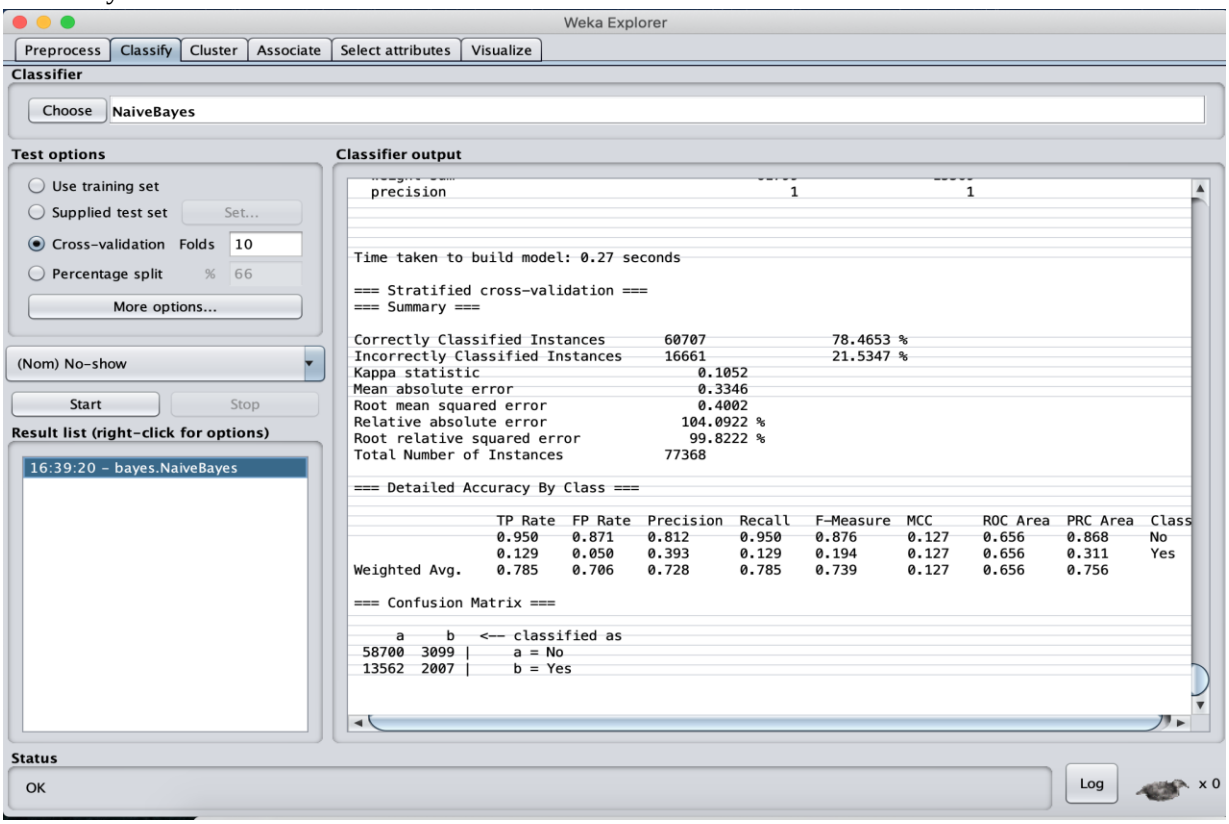
k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. Both for classification and regression, a useful technique can be to assign weights to the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones.

★ By using Weka, we applied a 70-30 partition such that 70% of the data is training dataset and 30% is the test dataset which has been randomly selected from the original dataset. We then built our models based on the labeled records in the training dataset.

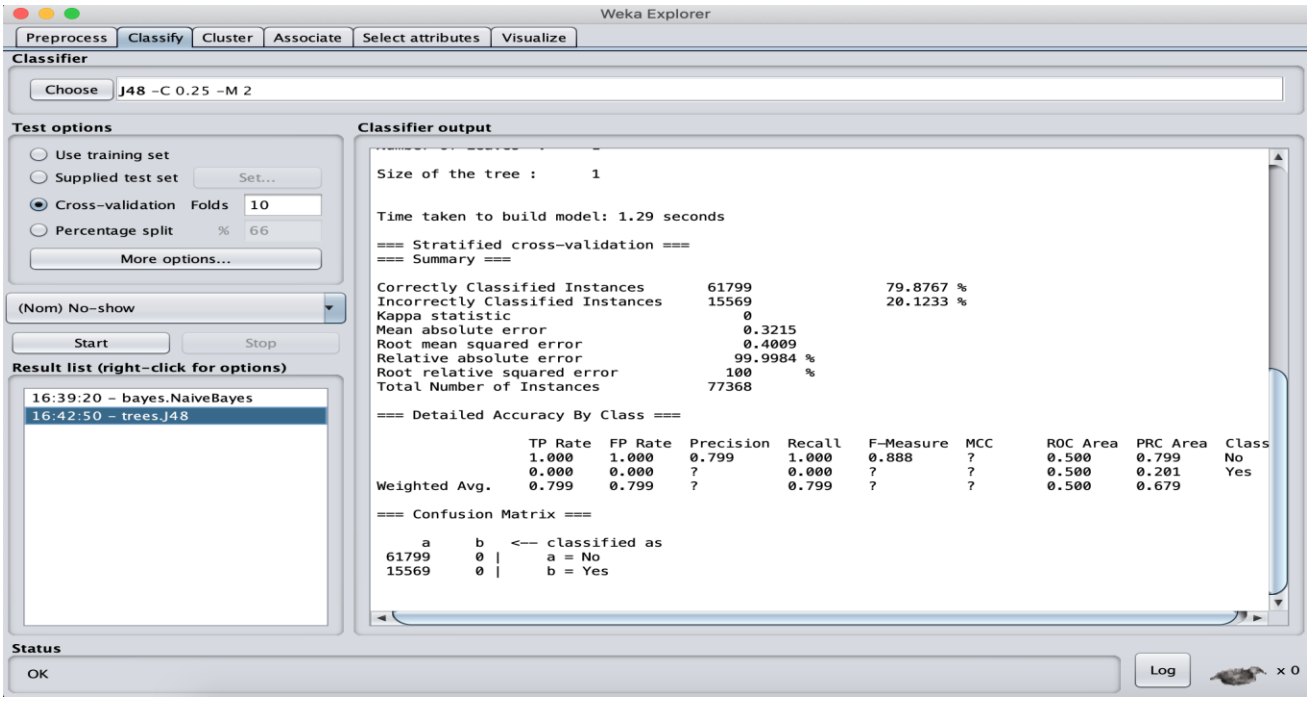
CLASSIFICATION - TRAINING DATASET

1. **GainratioAttributeEval**

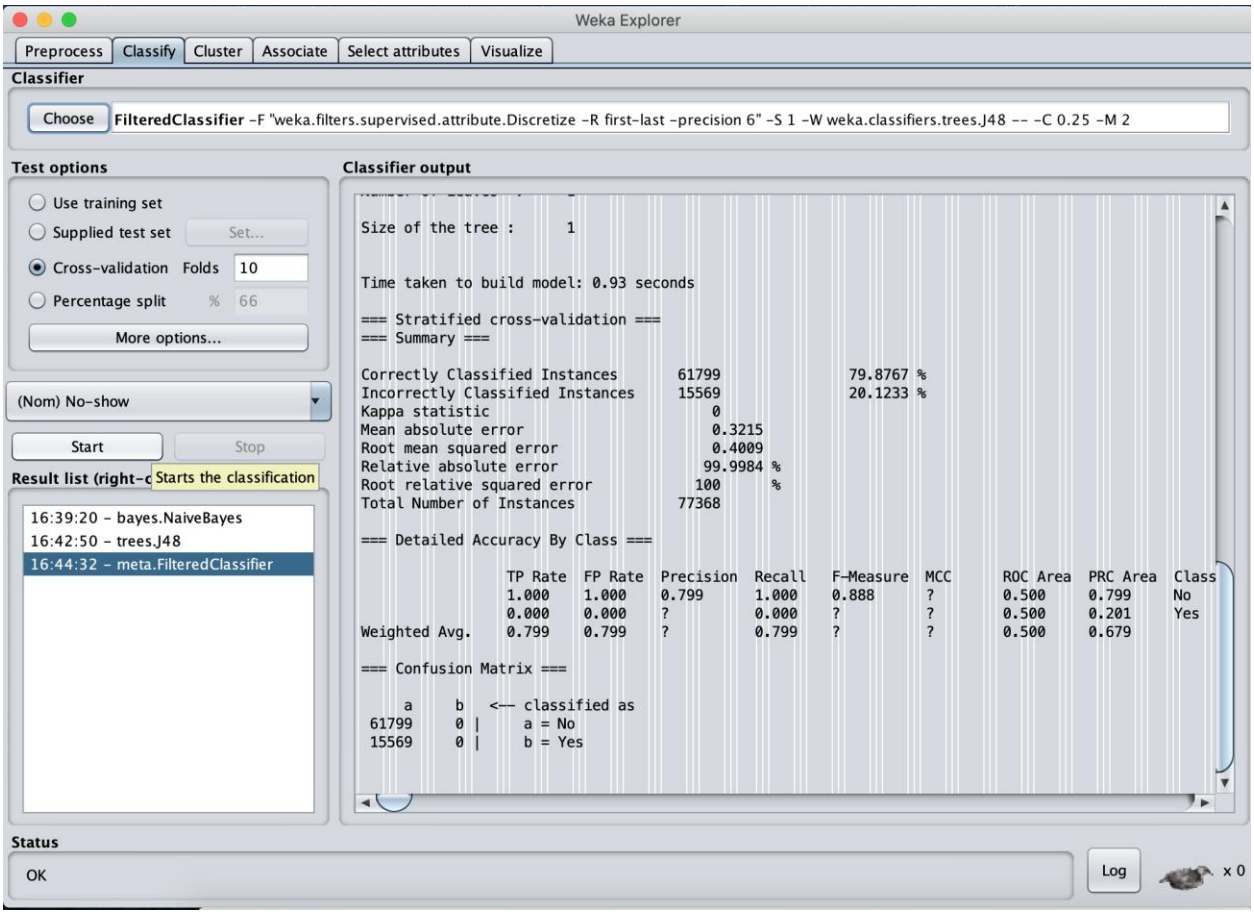
NaiveBayes:



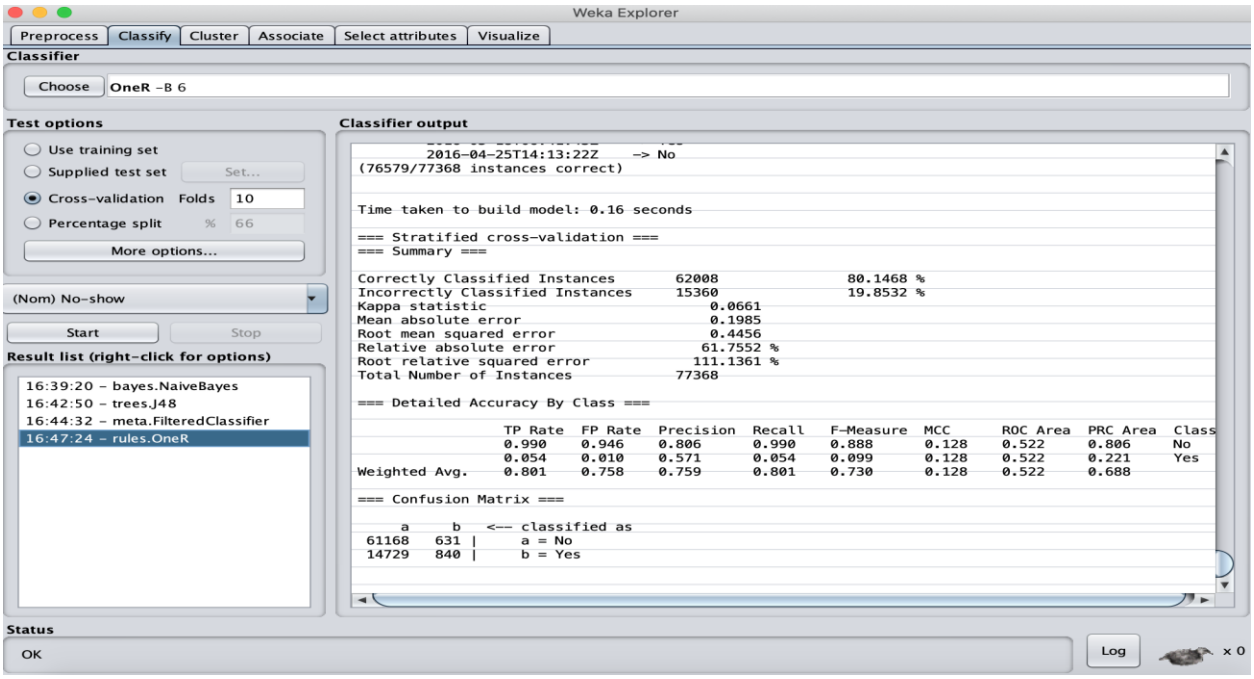
J48



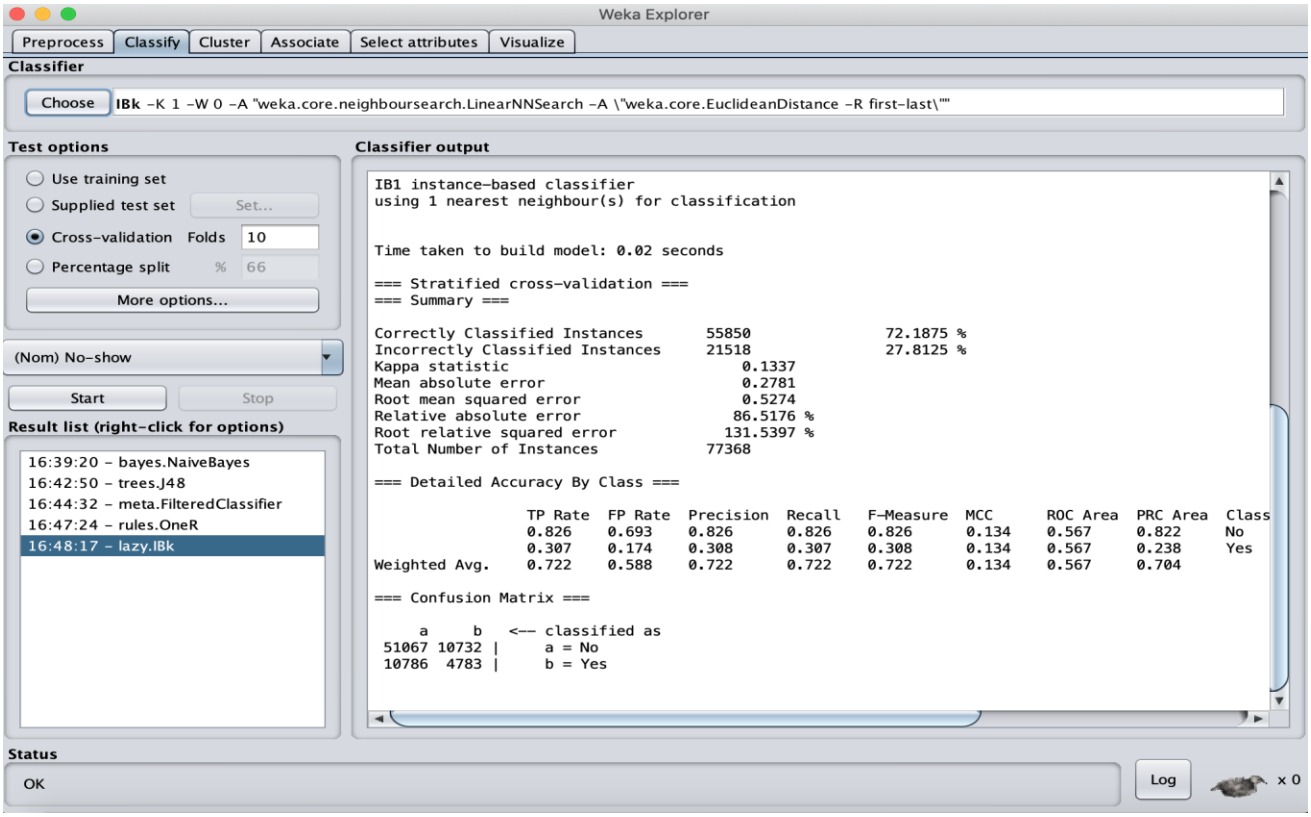
Filtered Classifier:



OneR:

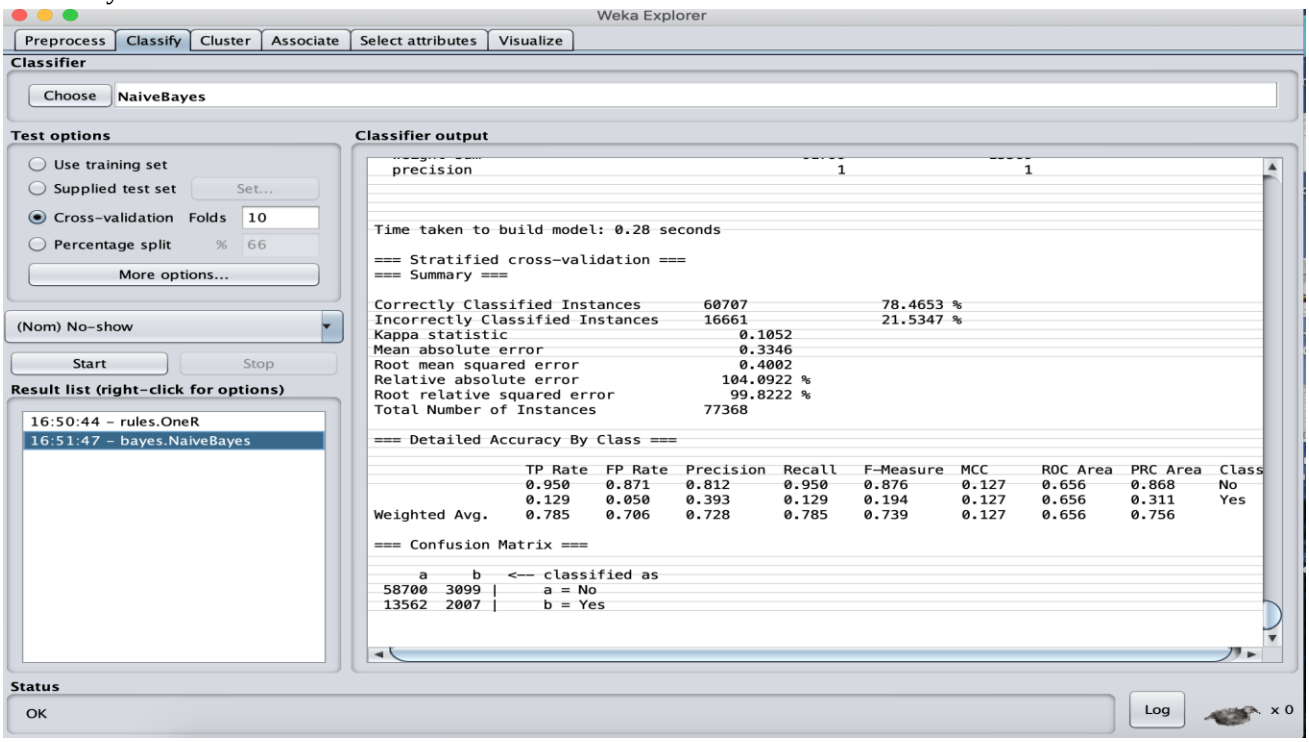

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
16:39:20 - bayes.NaiveBayes
16:42:50 - trees.J48
16:44:32 - meta.FilteredClassifier
16:47:24 - rules.OneR
Classifier output
2016-04-25T14:13:22Z -> No
(76579/77368 instances correct)
Time taken to build model: 0.16 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 62008 80.1468 %
Incorrectly Classified Instances 15360 19.8532 %
Kappa statistic 0.0661
Mean absolute error 0.1985
Root mean squared error 0.4456
Relative absolute error 61.7552 %
Root relative squared error 111.1361 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.990 0.946 0.806 0.990 0.888 0.128 0.522 0.806 No
0.054 0.010 0.571 0.054 0.099 0.128 0.522 0.221 Yes
Weighted Avg. 0.801 0.758 0.759 0.801 0.730 0.128 0.522 0.688
=== Confusion Matrix ===
a b <-- classified as
61168 631 | a = No
14729 840 | b = Yes
Status
OK
Log
x 0

IBk:

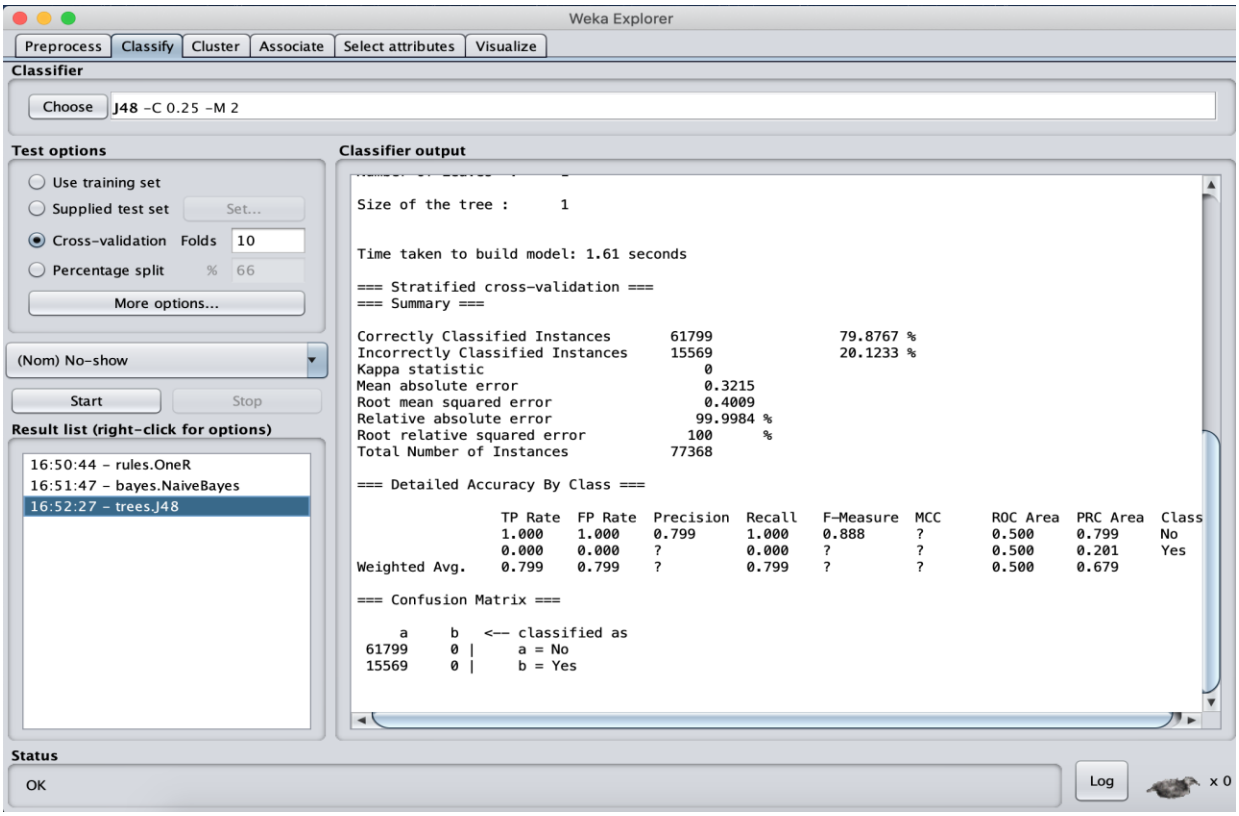

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
16:39:20 - bayes.NaiveBayes
16:42:50 - trees.J48
16:44:32 - meta.FilteredClassifier
16:47:24 - rules.OneR
16:48:17 - lazy.IBk
Classifier output
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.02 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 55850 72.1875 %
Incorrectly Classified Instances 21518 27.8125 %
Kappa statistic 0.1337
Mean absolute error 0.2781
Root mean squared error 0.5274
Relative absolute error 86.5176 %
Root relative squared error 131.5397 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.826 0.693 0.826 0.826 0.826 0.134 0.567 0.822 No
0.307 0.174 0.308 0.307 0.308 0.134 0.567 0.238 Yes
Weighted Avg. 0.722 0.588 0.722 0.722 0.722 0.134 0.567 0.704
=== Confusion Matrix ===
a b <-- classified as
51067 10732 | a = No
10786 4783 | b = Yes
Status
OK
Log
x 0

2. InfoGainAttributeEval(Ranker)

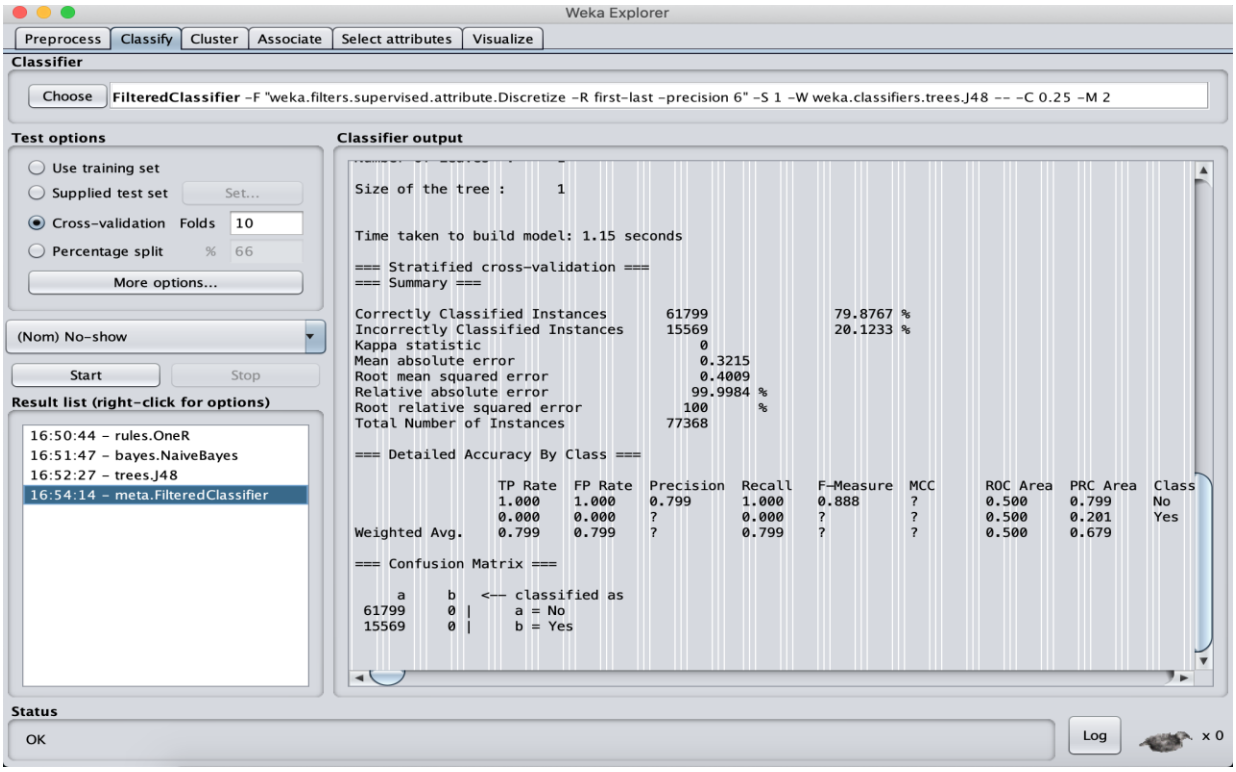
NaiveBayes:



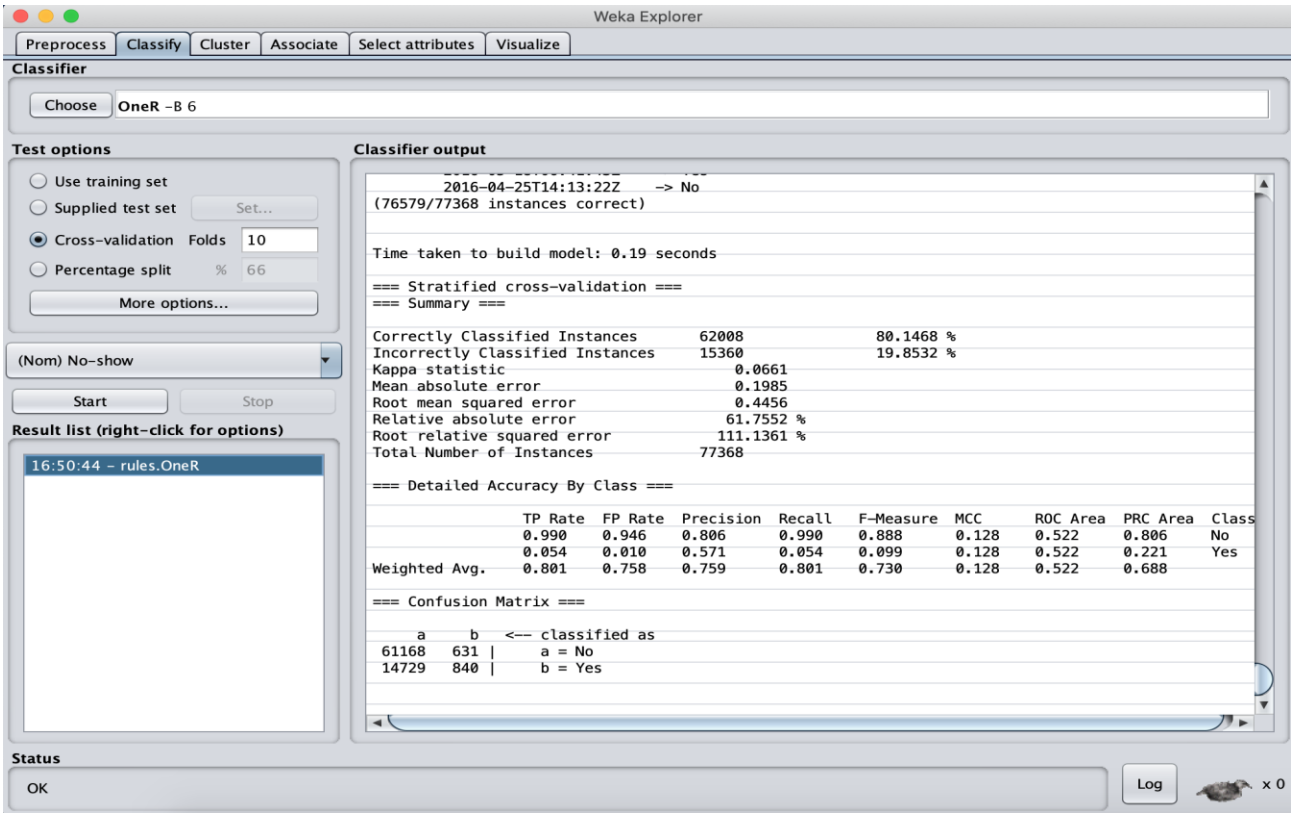
J48:



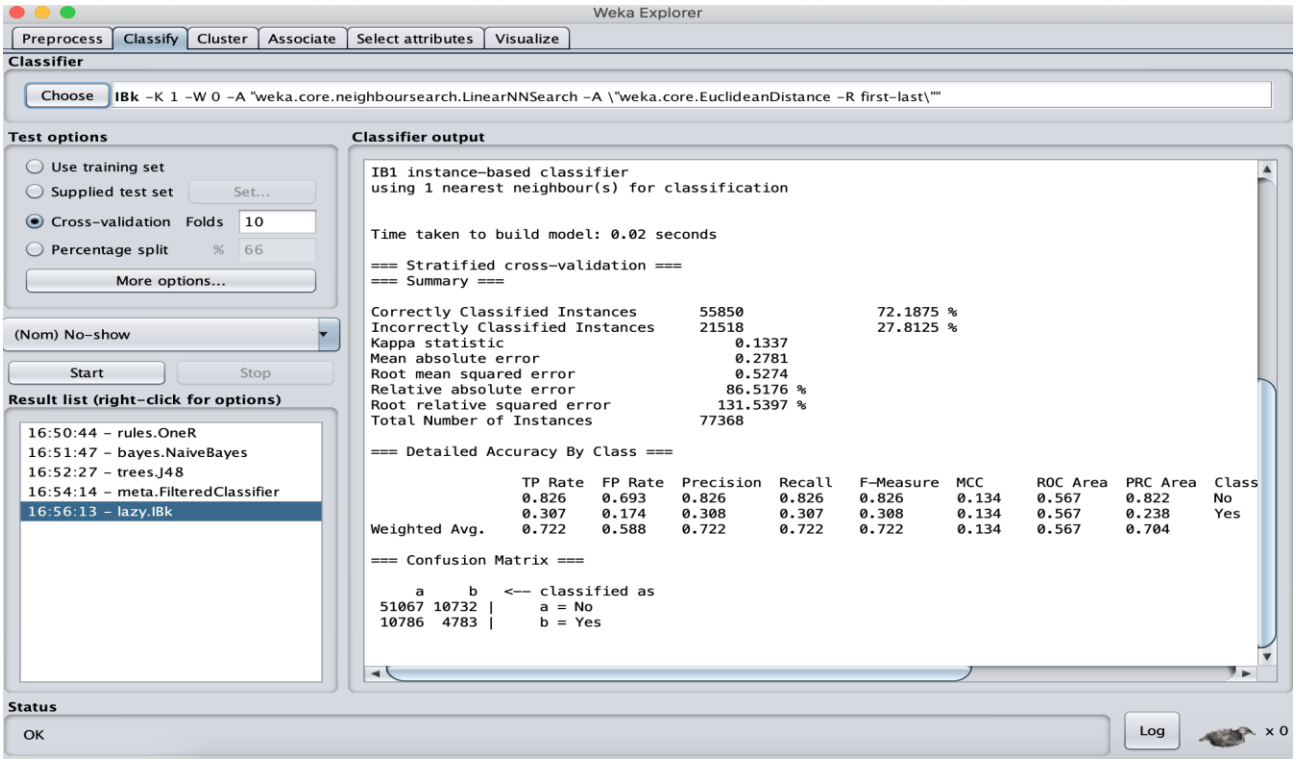
Filtered Classifier:



OneR:

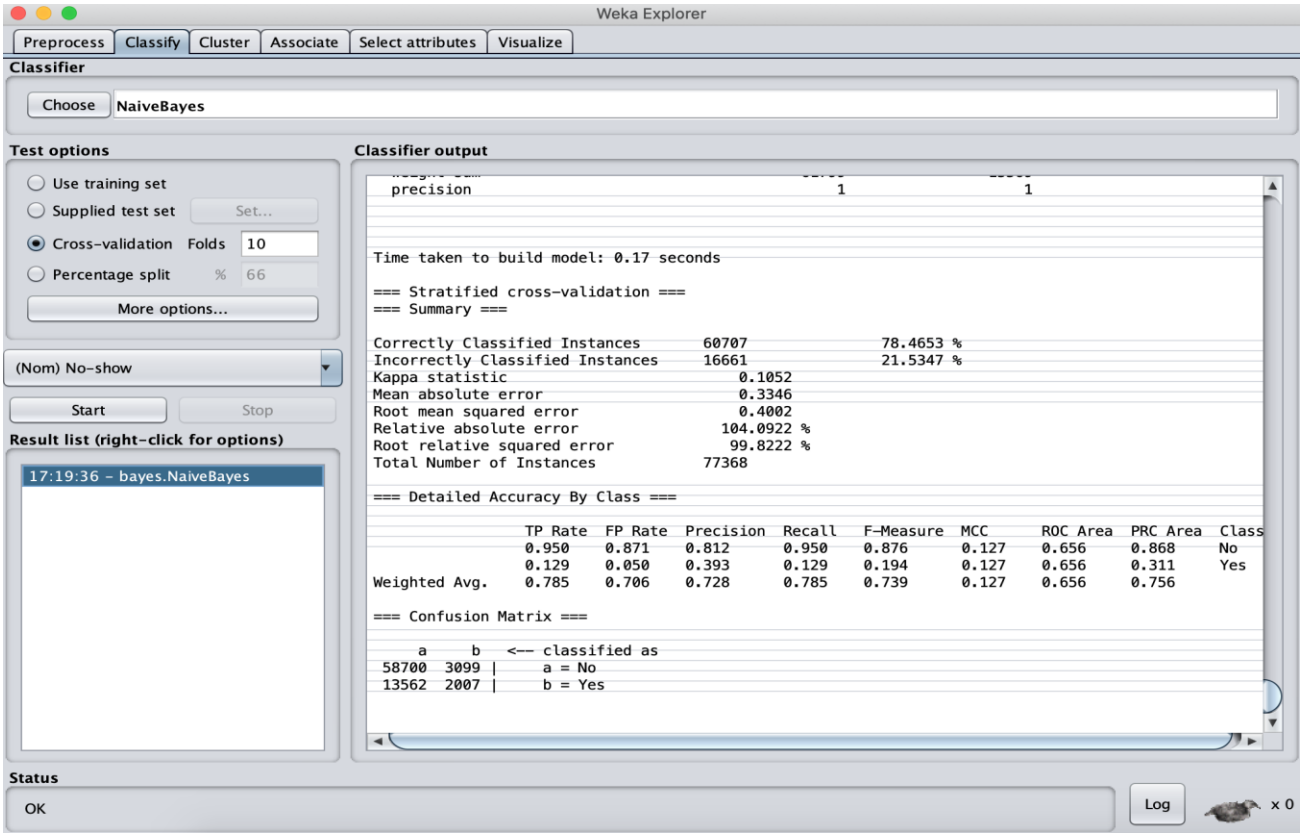


IBk:

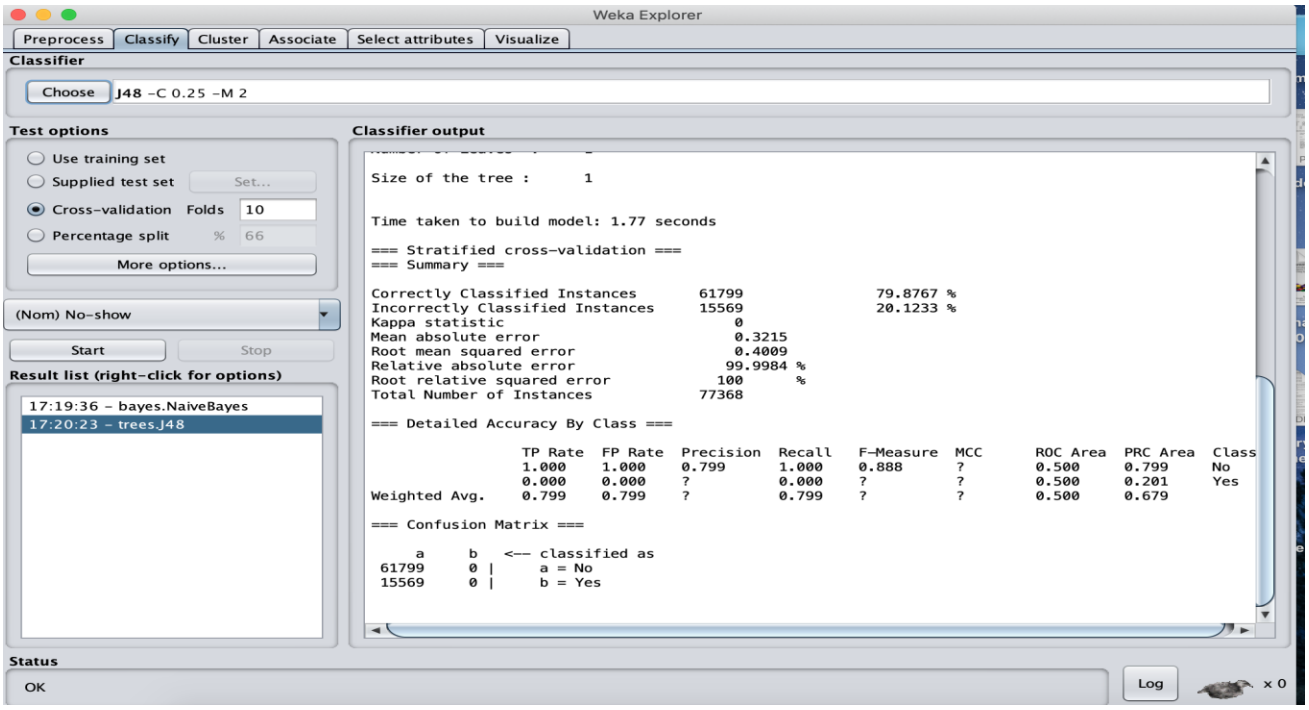


3. OneRAttributeEval:

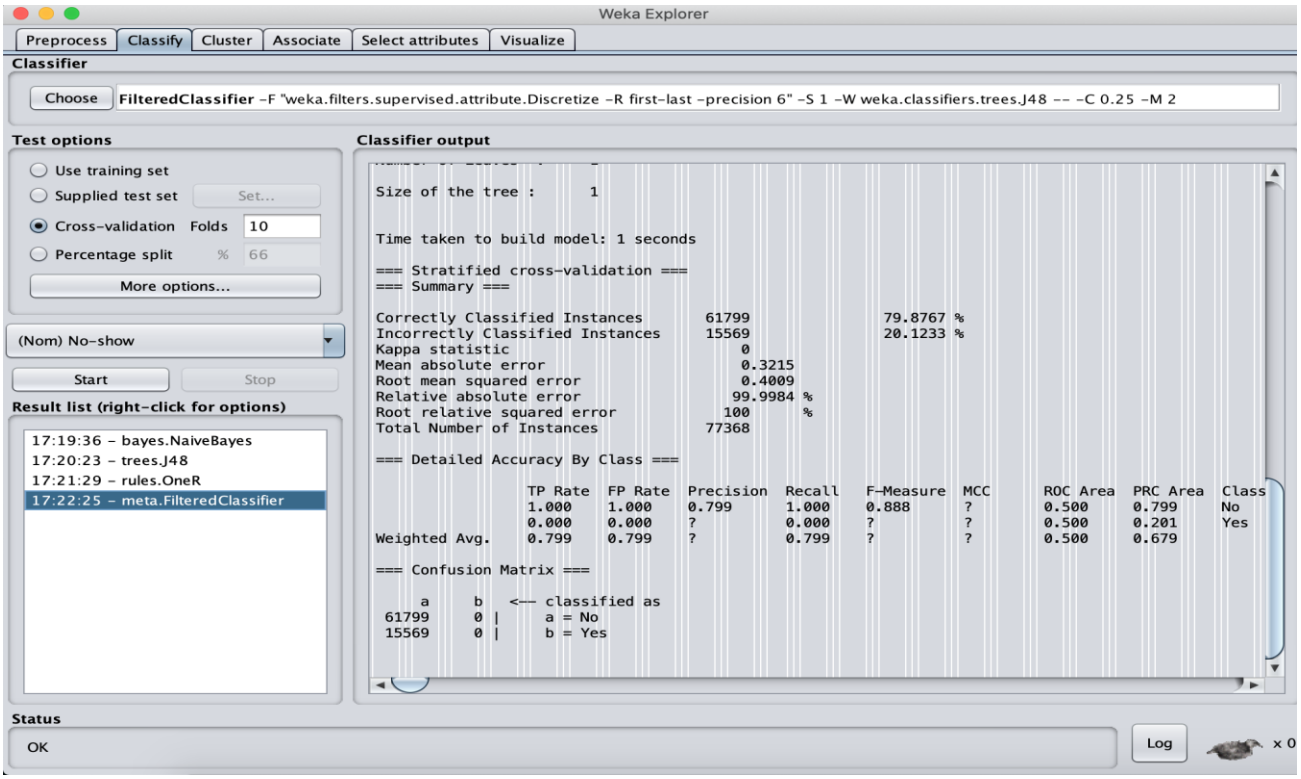
NaiveBayes:



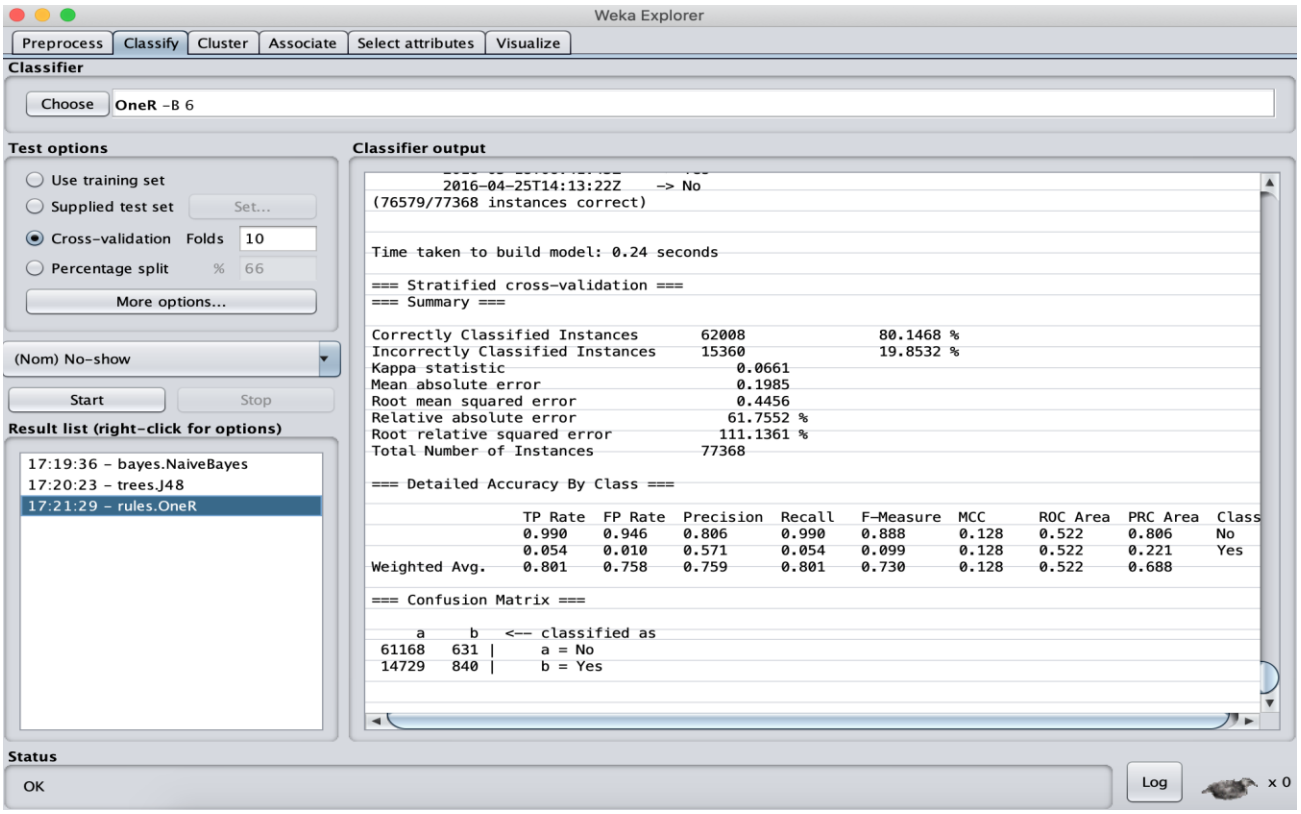
J48:



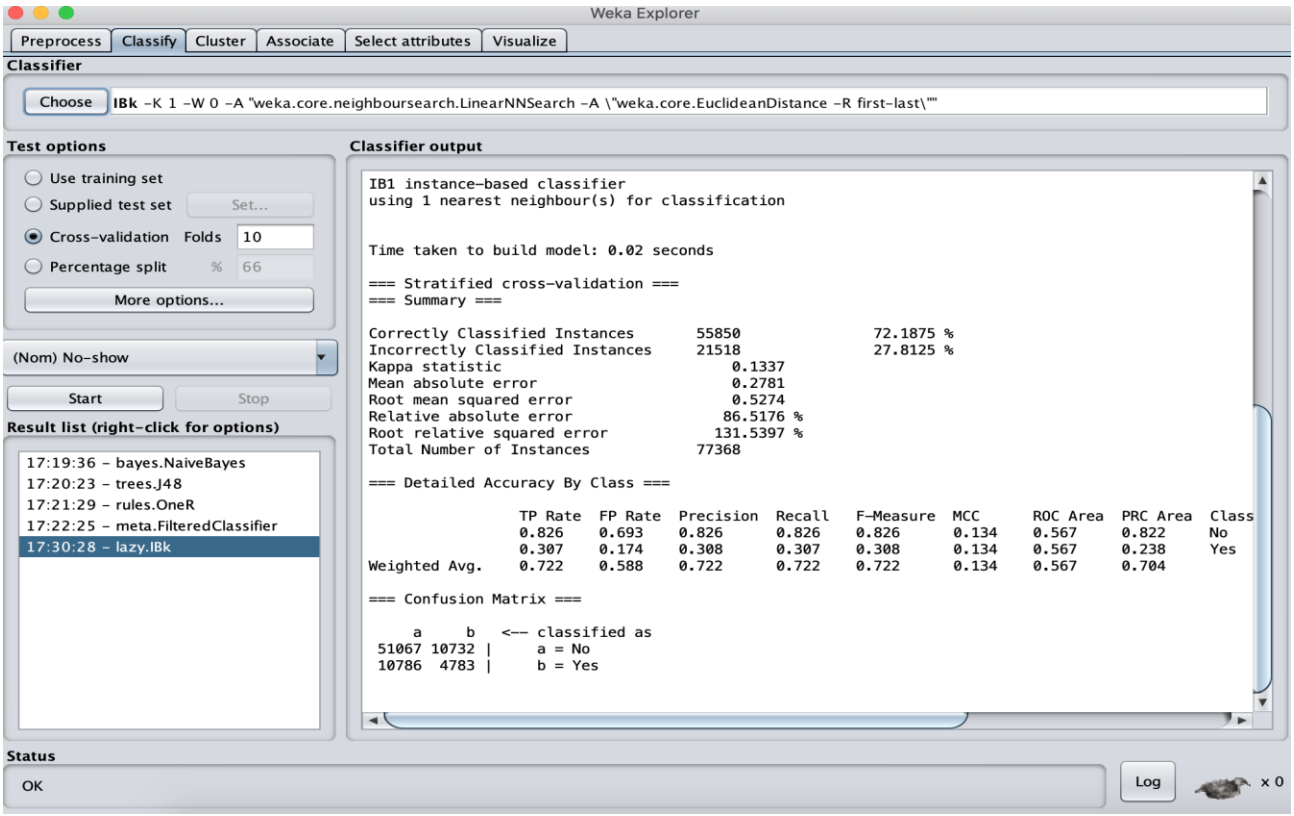
Filtered Classifier:



OneR:

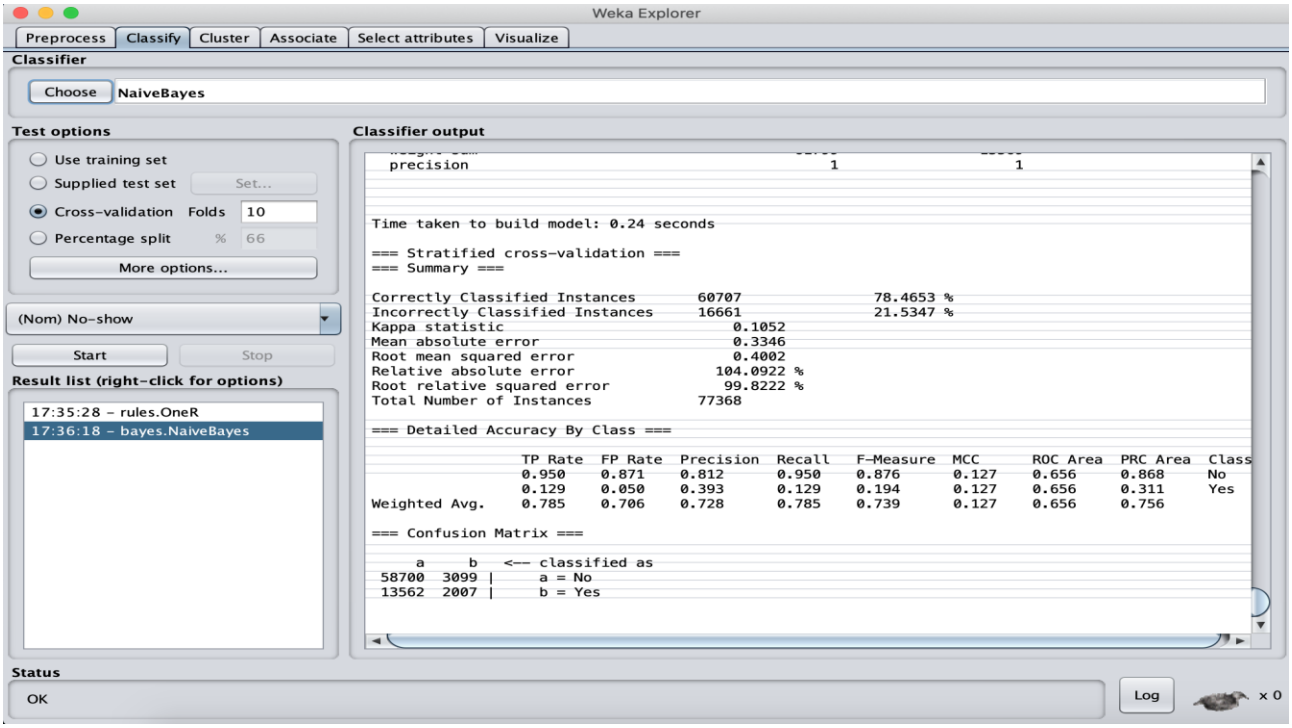

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
17:19:36 - bayes.NaiveBayes
17:20:23 - trees.J48
17:21:29 - rules.OneR
Classifier output
2016-04-25T14:13:22Z -> No
(76579/77368 instances correct)
Time taken to build model: 0.24 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 62008 80.1468 %
Incorrectly Classified Instances 15360 19.8532 %
Kappa statistic 0.0661
Mean absolute error 0.1985
Root mean squared error 0.4456
Relative absolute error 61.7552 %
Root relative squared error 111.1361 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.990 0.946 0.806 0.990 0.888 0.128 0.522 0.806 No
0.054 0.010 0.571 0.054 0.099 0.128 0.522 0.221 Yes
Weighted Avg. 0.801 0.758 0.759 0.801 0.730 0.128 0.522 0.688
=== Confusion Matrix ===
a b <-- classified as
61168 631 | a = No
14729 840 | b = Yes
Status
OK
Log
x 0

IBk:

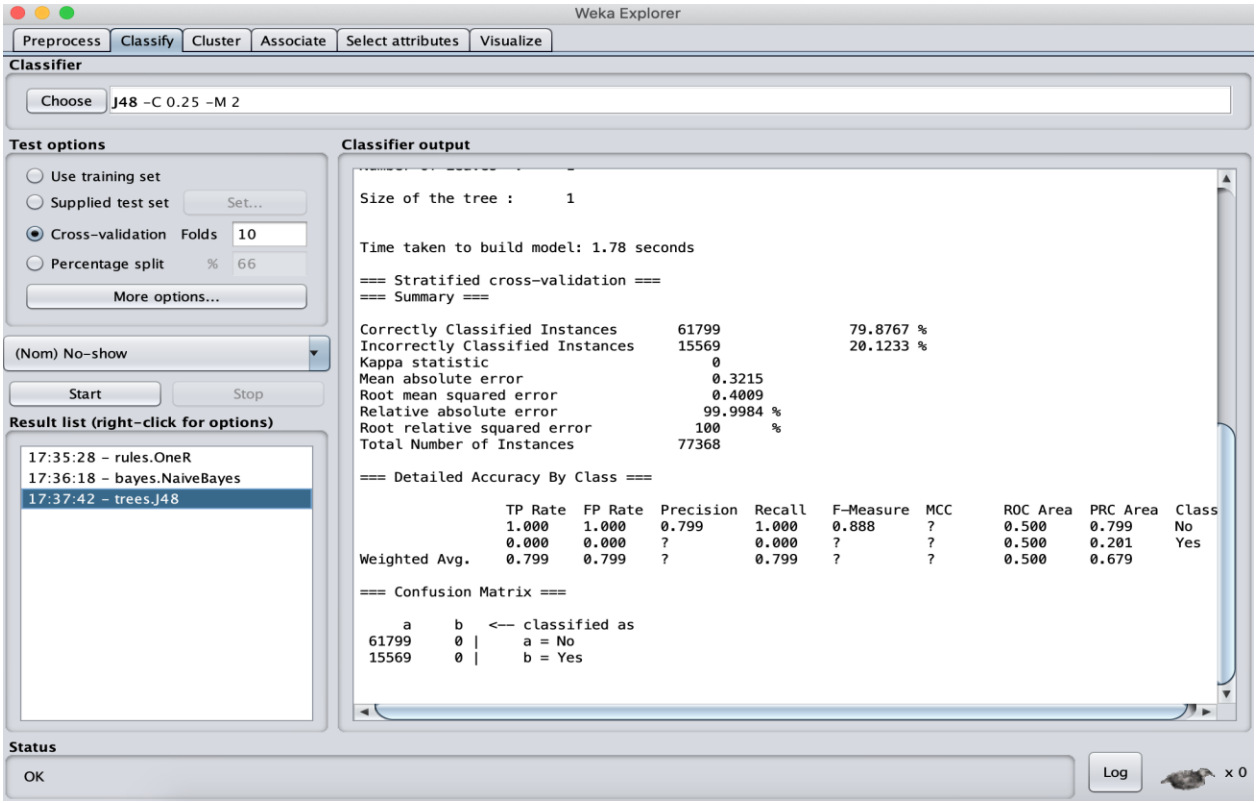

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
17:19:36 - bayes.NaiveBayes
17:20:23 - trees.J48
17:21:29 - rules.OneR
17:22:25 - meta.FilteredClassifier
17:30:28 - lazy.IBk
Classifier output
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.02 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 55850 72.1875 %
Incorrectly Classified Instances 21518 27.8125 %
Kappa statistic 0.1337
Mean absolute error 0.2781
Root mean squared error 0.5274
Relative absolute error 86.5176 %
Root relative squared error 131.5397 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.826 0.693 0.826 0.826 0.826 0.134 0.567 0.822 No
0.307 0.174 0.308 0.307 0.308 0.134 0.567 0.238 Yes
Weighted Avg. 0.722 0.588 0.722 0.722 0.722 0.134 0.567 0.704
=== Confusion Matrix ===
a b <-- classified as
51067 10732 | a = No
10786 4783 | b = Yes
Status
OK
Log
x 0

4. ClassifierAttributeEval

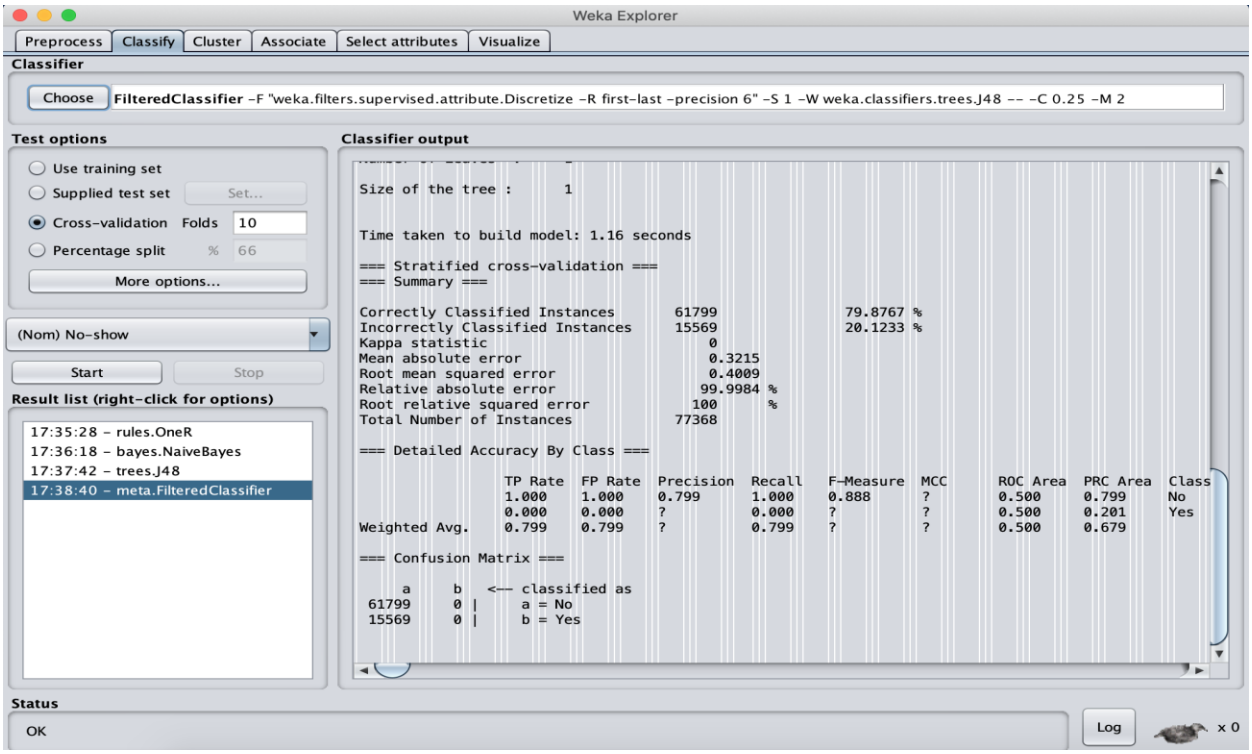
NaiveBayes:



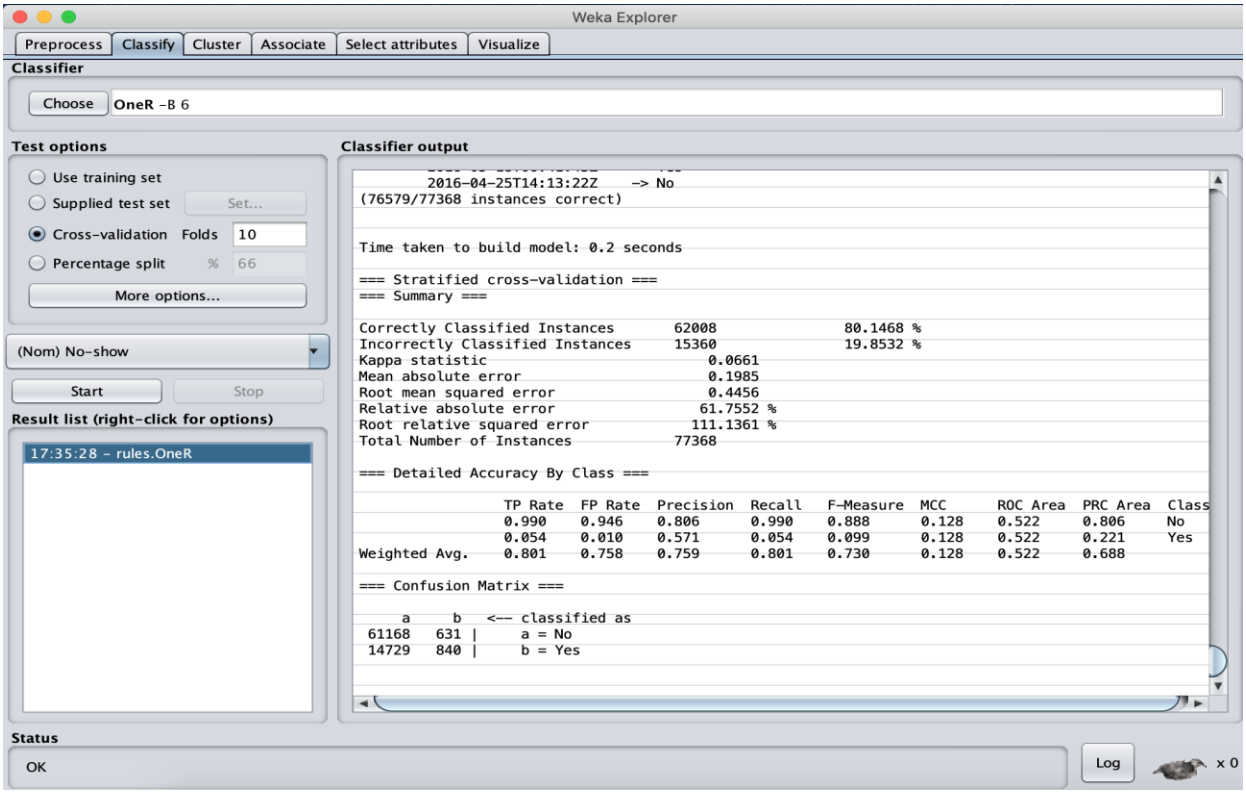
J48:



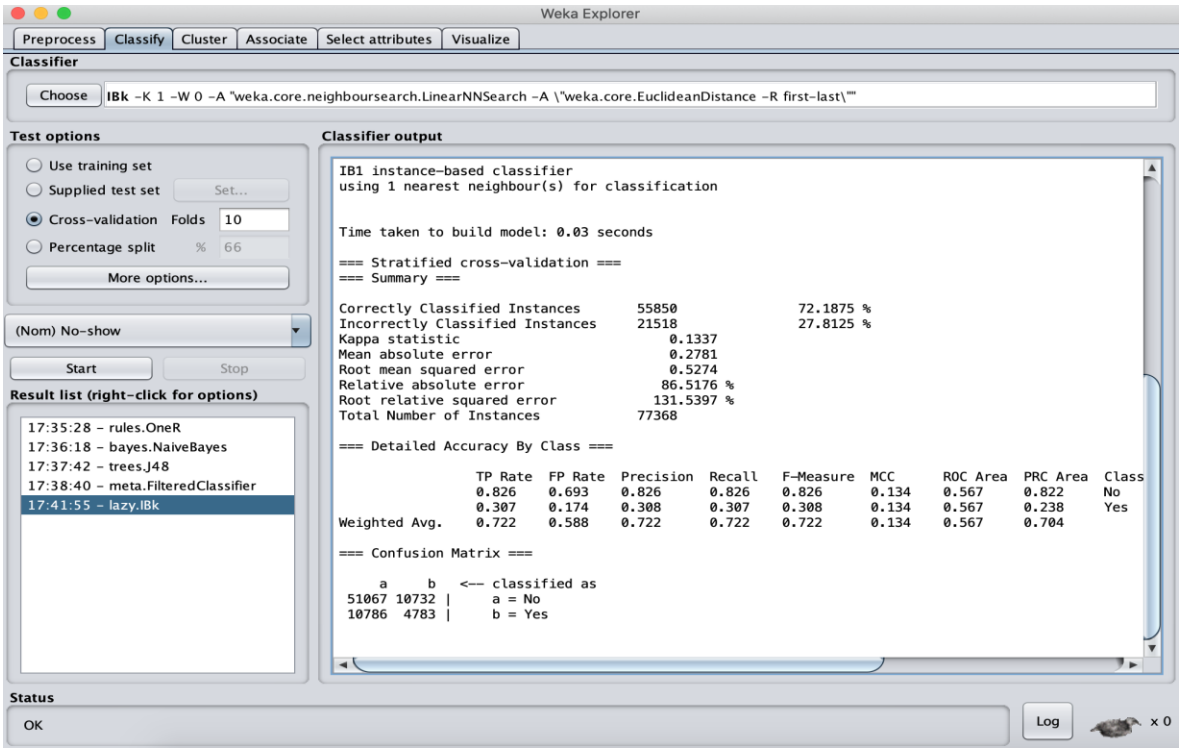
Filtered Classifier:


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
FilteredClassifier -F "weka.filters.supervised.attribute.Discretize -R first-last -precision 6" -S 1 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
17:35:28 - rules.OneR
17:36:18 - bayes.NaiveBayes
17:37:42 - trees.J48
17:38:40 - meta.FilteredClassifier
Classifier output
Size of the tree : 1
Time taken to build model: 1.16 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 61799 79.8767 %
Incorrectly Classified Instances 15569 20.1233 %
Kappa statistic 0
Mean absolute error 0.3215
Root mean squared error 0.4009
Relative absolute error 99.9984 %
Root relative squared error 100 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
1.000 1.000 0.799 1.000 0.888 ? 0.500 0.799 No
0.000 0.000 ? 0.000 ? ? 0.500 0.201 Yes
Weighted Avg. 0.799 0.799 ? 0.799 ? ? 0.500 0.679
=== Confusion Matrix ===
a b <-- classified as
61799 0 | a = No
15569 0 | b = Yes
Status
OK
Log
x 0

OneR:

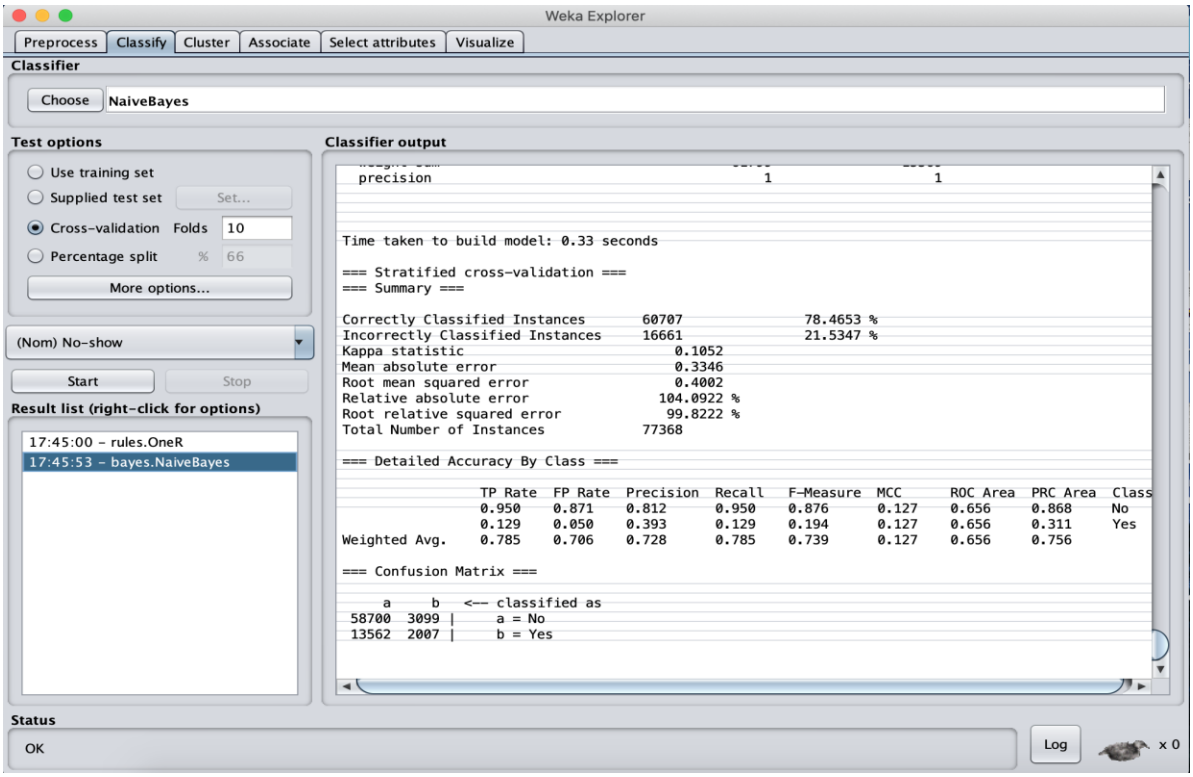

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
17:35:28 - rules.OneR
Classifier output
2016-04-25T14:13:22Z -> No
(76579/77368 instances correct)
Time taken to build model: 0.2 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 62008 80.1468 %
Incorrectly Classified Instances 15360 19.8532 %
Kappa statistic 0.0661
Mean absolute error 0.1985
Root mean squared error 0.4456
Relative absolute error 61.7552 %
Root relative squared error 111.1361 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.990 0.946 0.806 0.990 0.888 0.128 0.522 0.806 No
0.054 0.010 0.571 0.054 0.099 0.128 0.522 0.221 Yes
Weighted Avg. 0.801 0.758 0.759 0.801 0.730 0.128 0.522 0.688
=== Confusion Matrix ===
a b <-- classified as
61168 631 | a = No
14729 840 | b = Yes
Status
OK
Log
x 0

IBk:

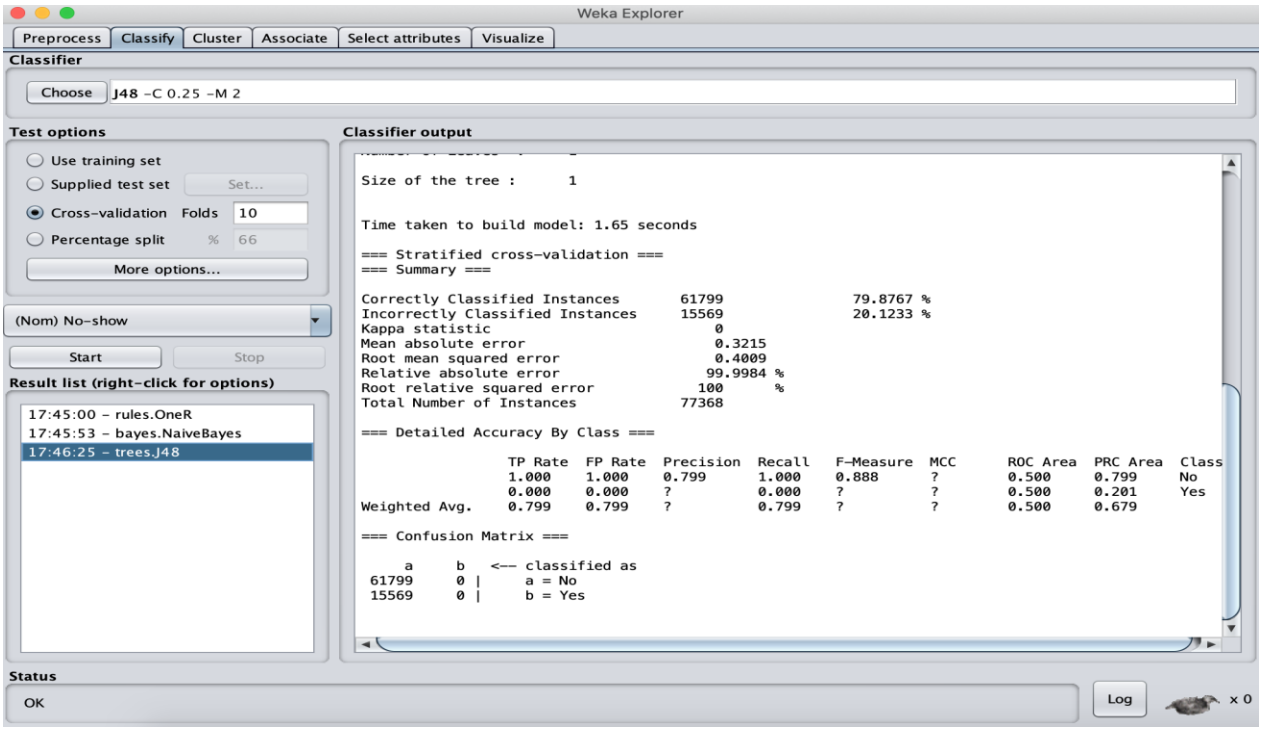


5. Personal Selection:

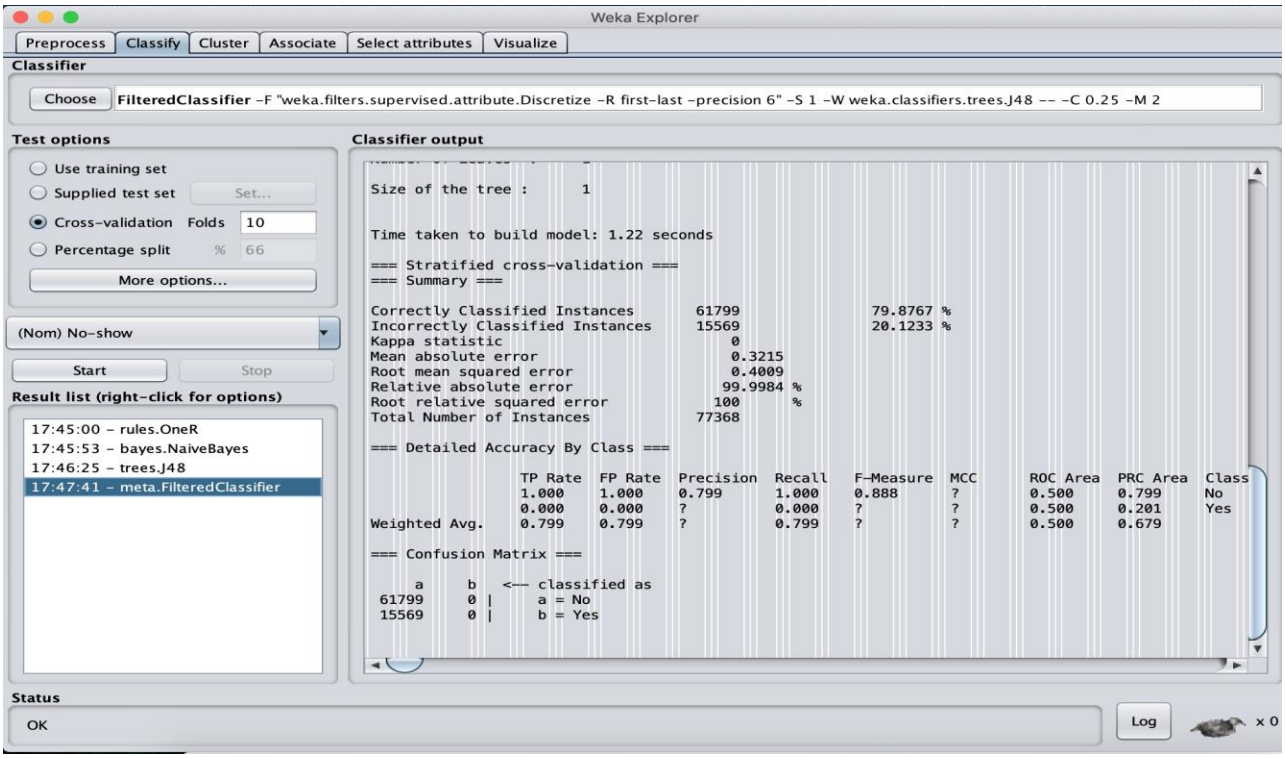
NaiveBayes:



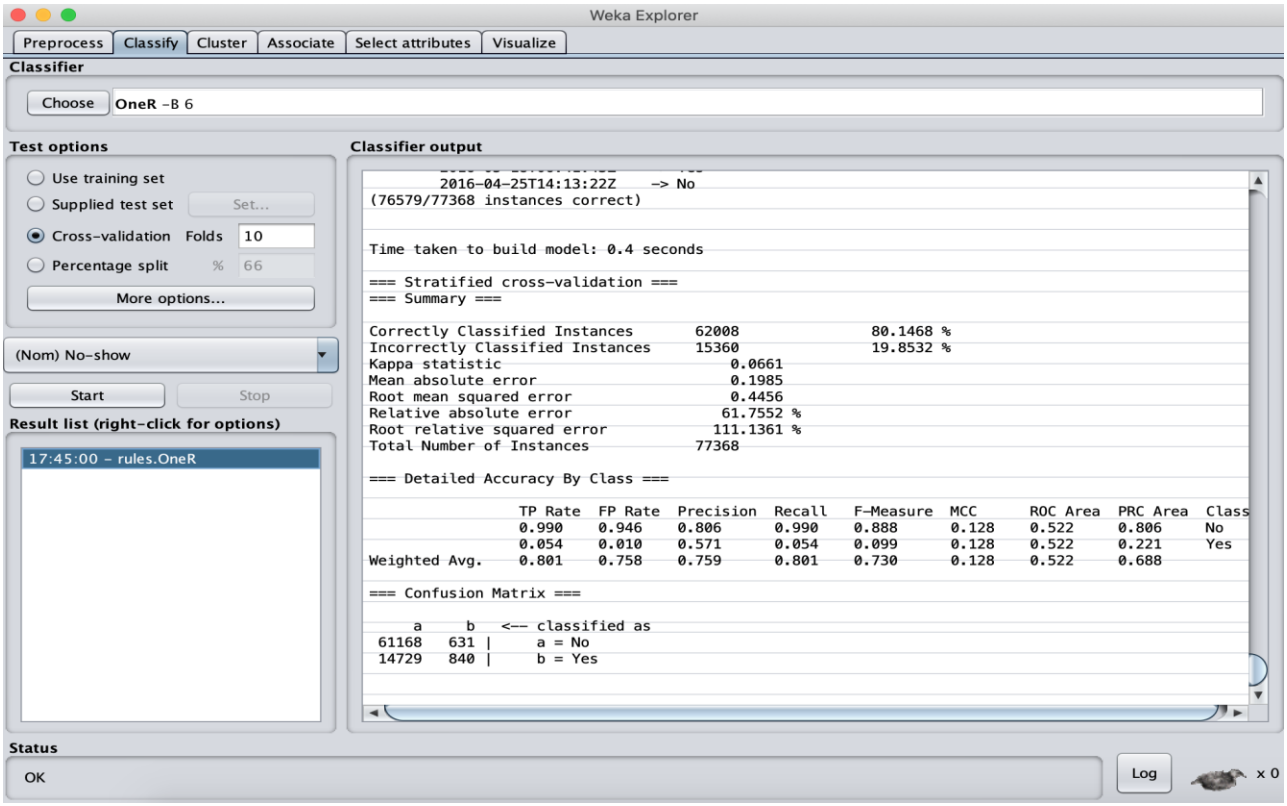
J48:



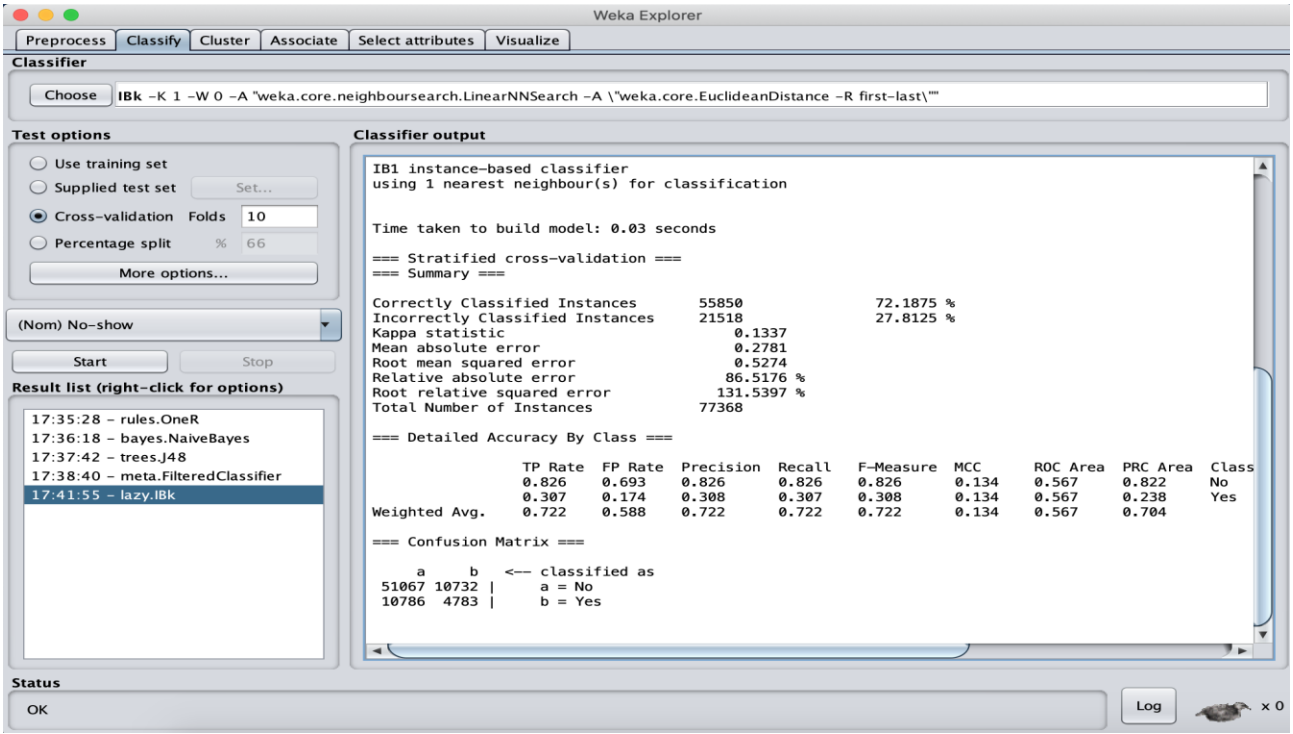
Filtered Classifier:



OneR:


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
17:45:00 - rules.OneR
Classifier output
2016-04-25T14:13:22Z -> No
(76579/77368 instances correct)
Time taken to build model: 0.4 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 62008 80.1468 %
Incorrectly Classified Instances 15360 19.8532 %
Kappa statistic 0.0661
Mean absolute error 0.1985
Root mean squared error 0.4456
Relative absolute error 61.7552 %
Root relative squared error 111.1361 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.990 0.946 0.806 0.990 0.888 0.128 0.522 0.806 No
0.054 0.010 0.571 0.054 0.099 0.128 0.522 0.221 Yes
Weighted Avg. 0.801 0.758 0.759 0.801 0.730 0.128 0.522 0.688
=== Confusion Matrix ===
a b <-- classified as
61168 631 | a = No
14729 840 | b = Yes
Status
OK
Log
x 0

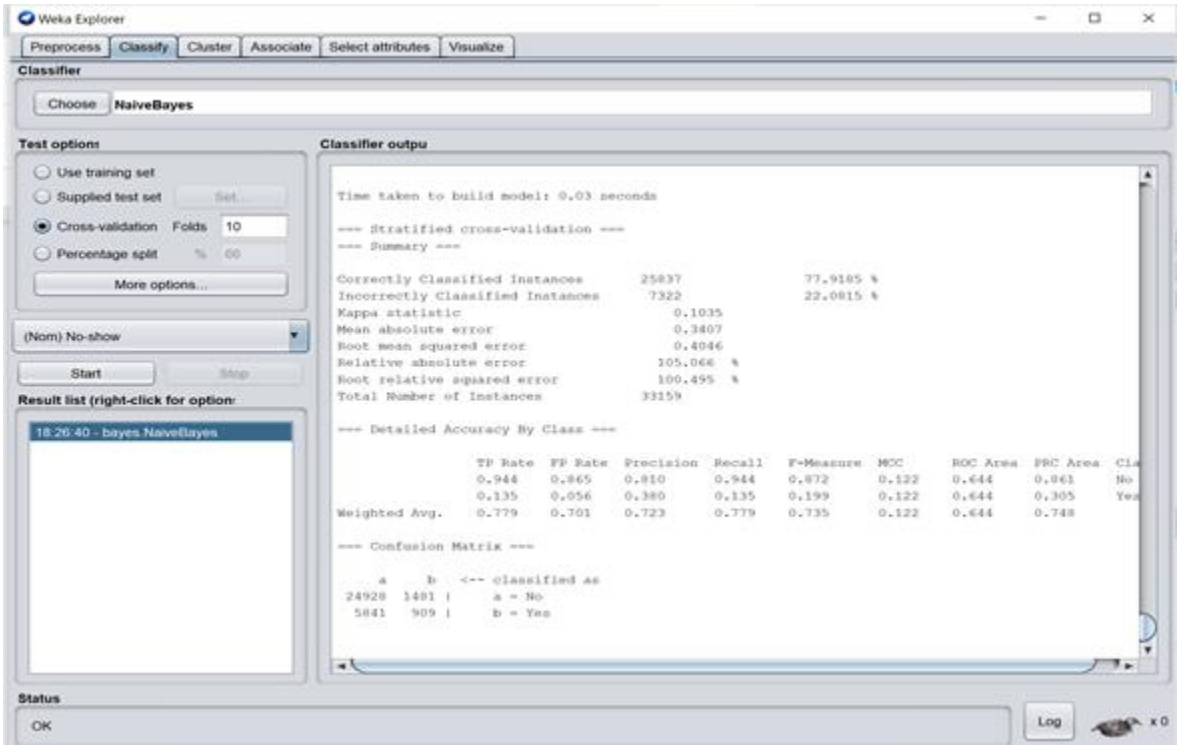
IBk:


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) No-show
Start
Stop
Result list (right-click for options)
17:35:28 - rules.OneR
17:36:18 - bayes.NaiveBayes
17:37:42 - trees.J48
17:38:40 - meta.FilteredClassifier
17:41:55 - lazy.IBk
Classifier output
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification
Time taken to build model: 0.03 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 55850 72.1875 %
Incorrectly Classified Instances 21518 27.8125 %
Kappa statistic 0.1337
Mean absolute error 0.2781
Root mean squared error 0.5274
Relative absolute error 86.5176 %
Root relative squared error 131.5397 %
Total Number of Instances 77368
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.826 0.693 0.826 0.826 0.826 0.134 0.567 0.822 No
0.307 0.174 0.308 0.307 0.308 0.134 0.567 0.238 Yes
Weighted Avg. 0.722 0.588 0.722 0.722 0.722 0.134 0.567 0.704
=== Confusion Matrix ===
a b <-- classified as
51067 10732 | a = No
10786 4783 | b = Yes
Status
OK
Log
x 0

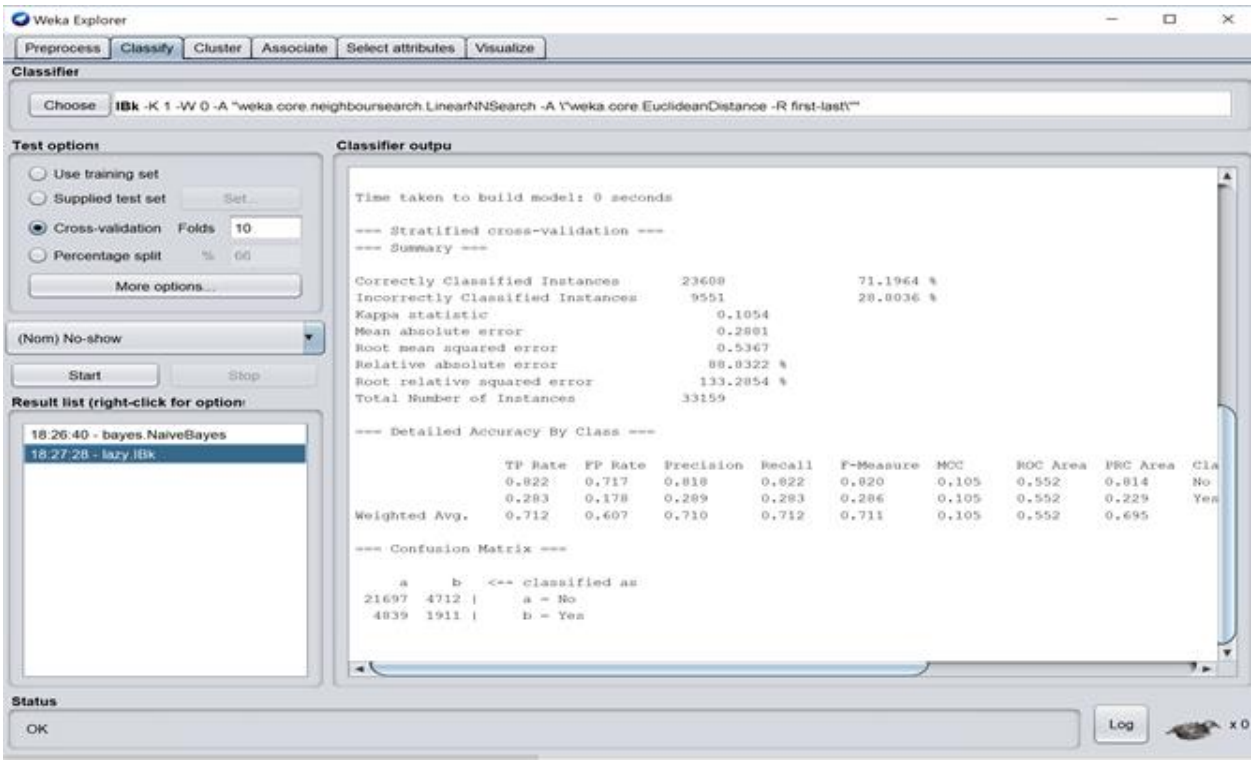
TESTING DATASET

1. **GainratioAttributeEval(Ranker)**

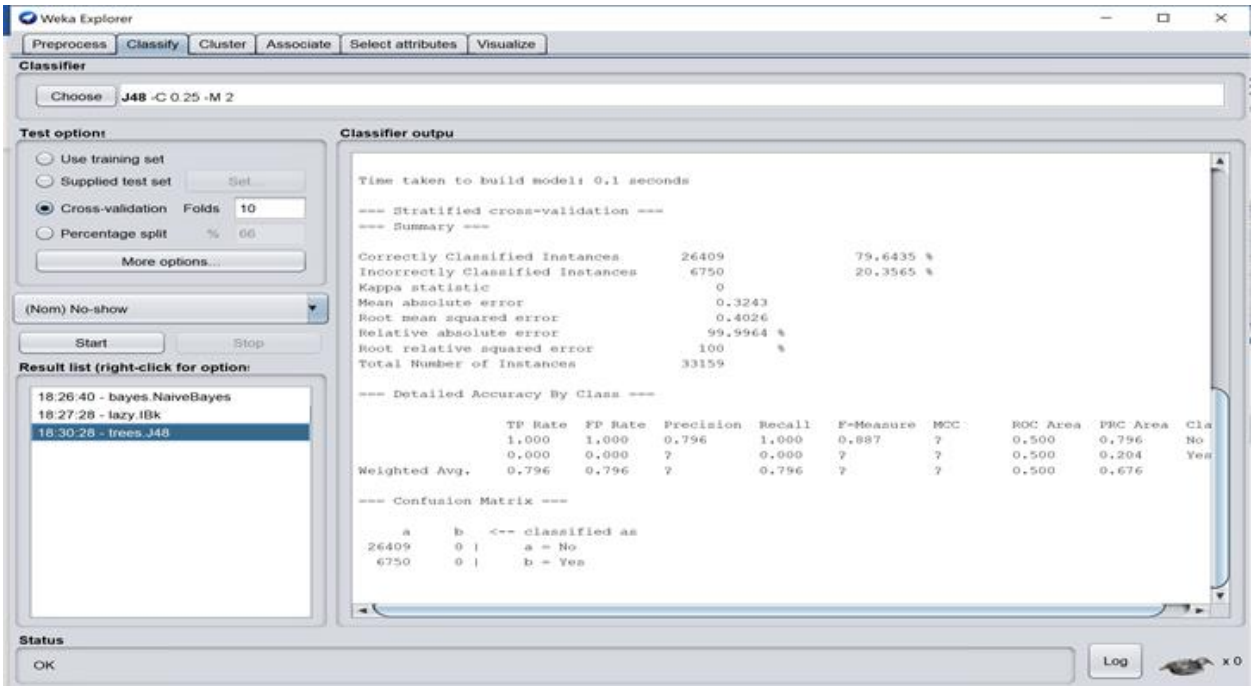
NaiveBayes:



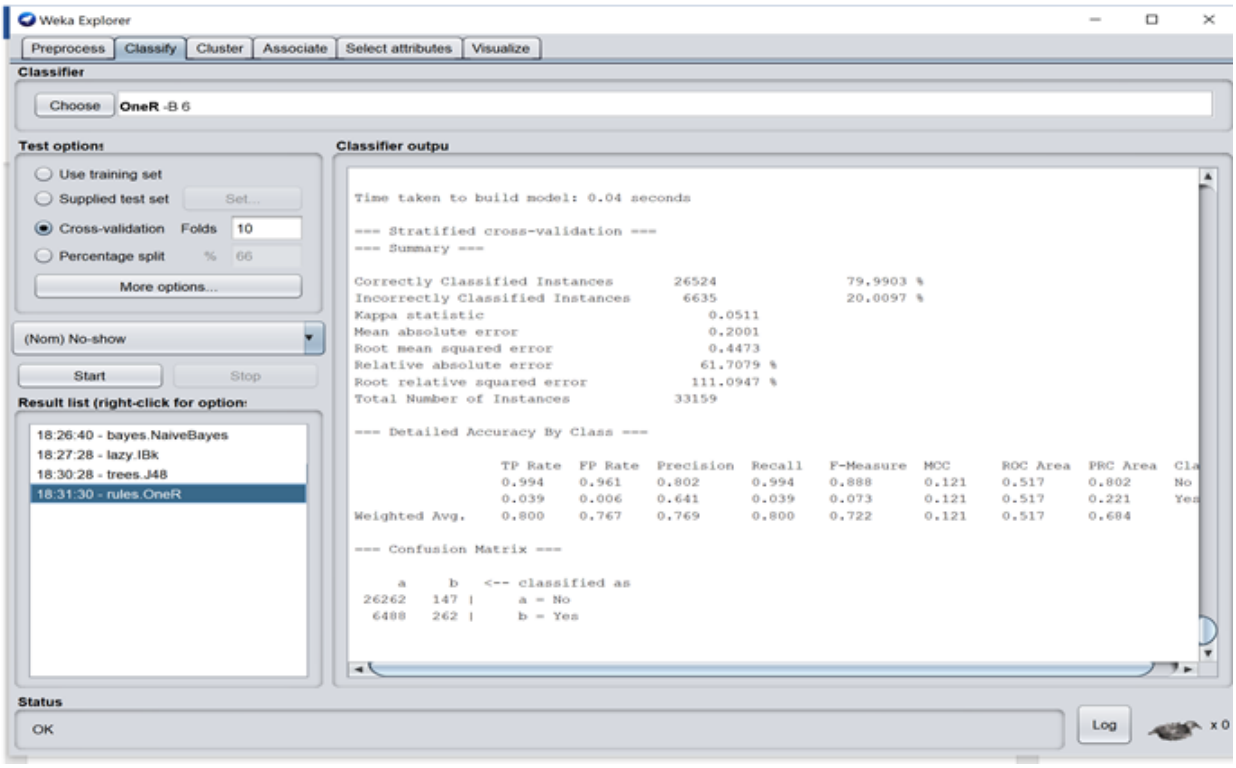
IBK



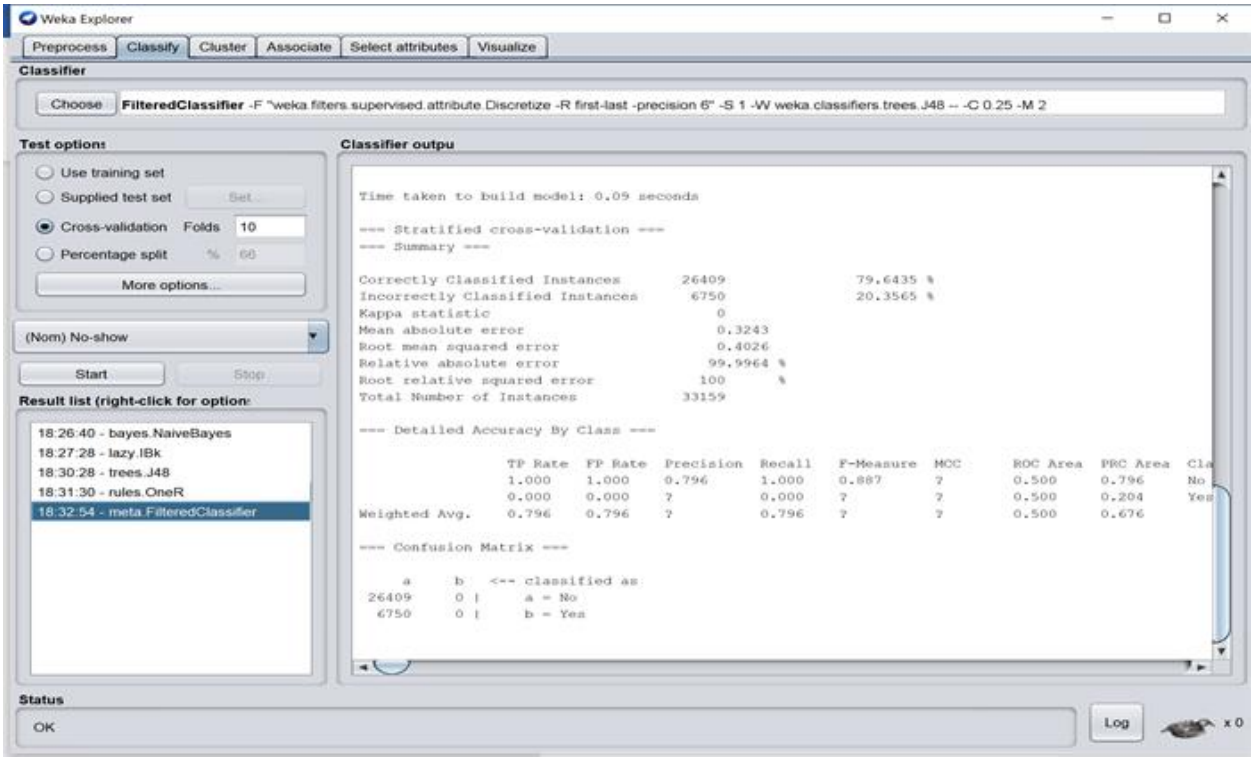
J48


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test option:
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
(Nom) No-show
Start
Stop
Result list (right-click for option:
18:26:40 - bayes.NaiveBayes
18:27:28 - lazy.IBk
18:30:28 - trees.J48
Time taken to build model: 0.1 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 26409 79.6435 %
Incorrectly Classified Instances 6750 20.3565 %
Kappa statistic 0
Mean absolute error 0.3243
Root mean squared error 0.4026
Relative absolute error 99.9964 %
Root relative squared error 100 %
Total Number of Instances 33159
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
1,000 1,000 0.796 1,000 0.887 ? 0.500 0.796 No
0.000 0.000 ? 0.000 ? ? 0.500 0.204 Yes
Weighted Avg. 0.796 0.796 ? 0.796 ? ? 0.500 0.676
=== Confusion Matrix ===
a b <-- classified as
26409 0 | a = No
6750 0 | b = Yes
Status
OK
Log
x 0

OneR

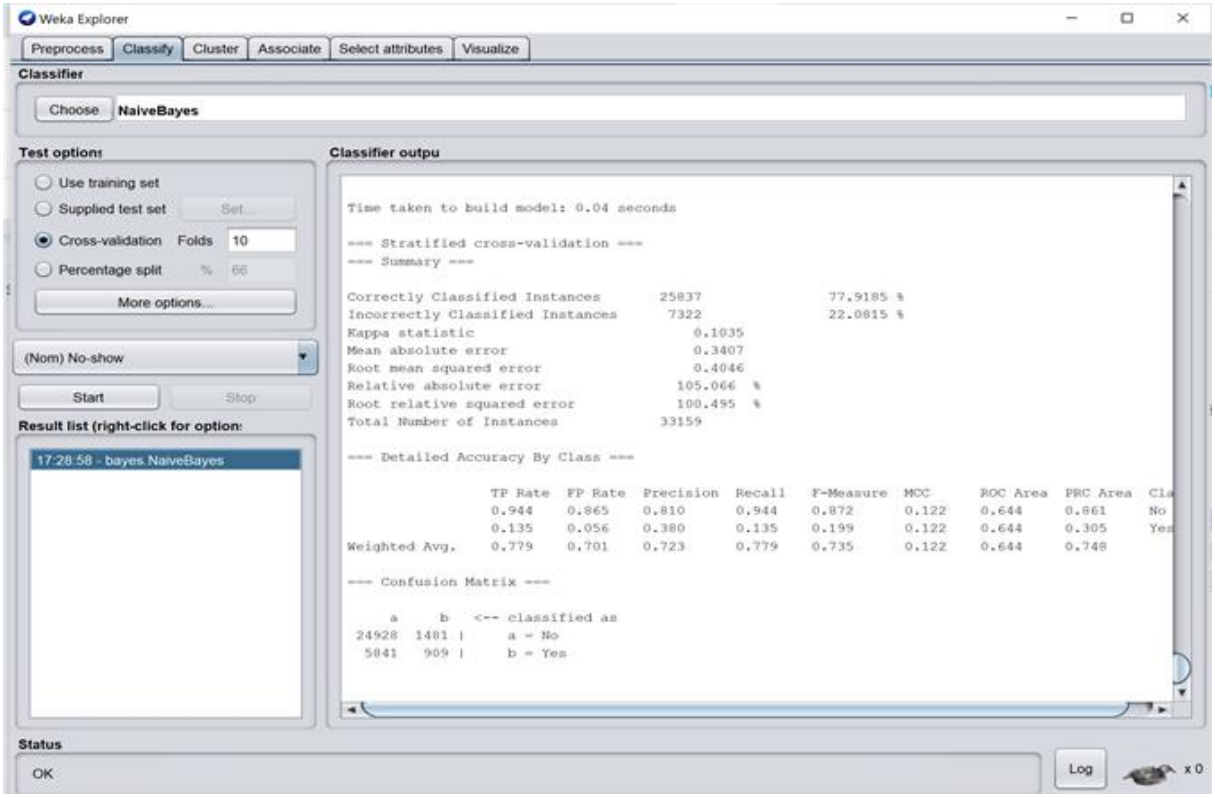

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test option:
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
(Nom) No-show
Start
Stop
Result list (right-click for option:
18:26:40 - bayes.NaiveBayes
18:27:28 - lazy.IBk
18:30:28 - trees.J48
18:31:30 - rules.OneR
Time taken to build model: 0.04 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 26524 79.9903 %
Incorrectly Classified Instances 6635 20.0097 %
Kappa statistic 0.0511
Mean absolute error 0.2001
Root mean squared error 0.4473
Relative absolute error 61.7079 %
Root relative squared error 111.0947 %
Total Number of Instances 33159
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.994 0.961 0.802 0.994 0.888 0.121 0.517 0.802 No
0.039 0.006 0.641 0.039 0.073 0.121 0.517 0.221 Yes
Weighted Avg. 0.800 0.767 0.769 0.800 0.722 0.121 0.517 0.684
=== Confusion Matrix ===
a b <-- classified as
26262 147 | a = No
6488 262 | b = Yes
Status
OK
Log
x 0

FilteredClassifier

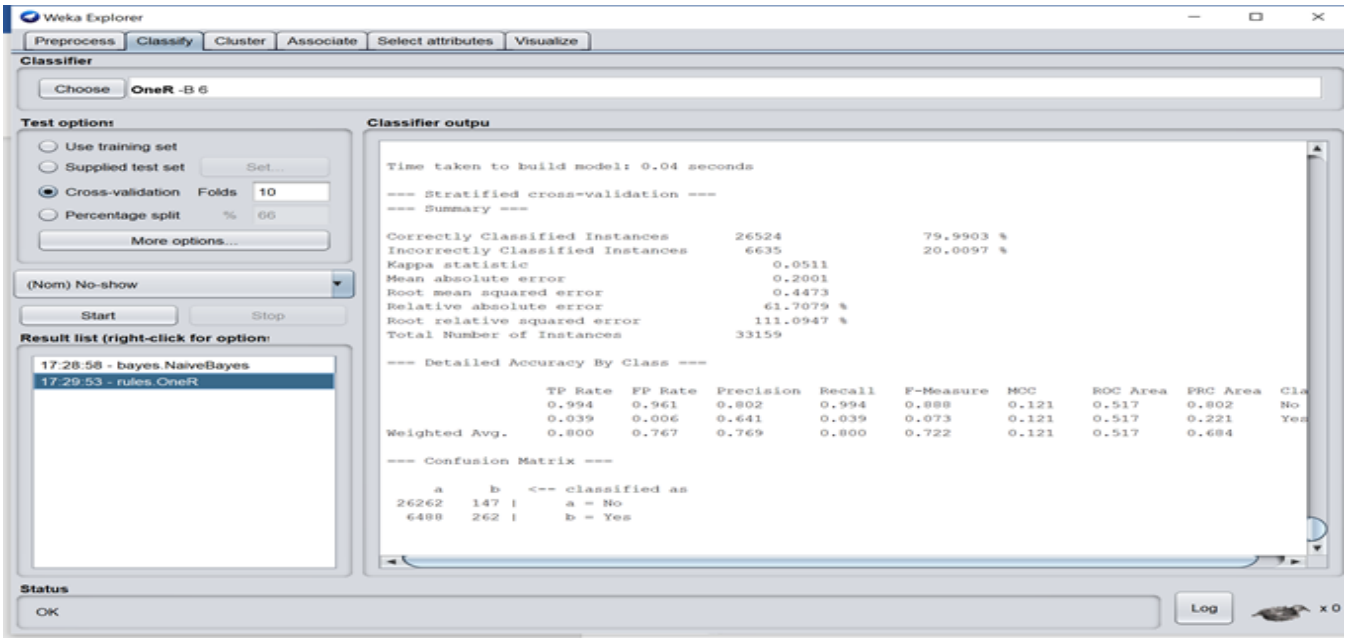


2. InfoGainAttributeEval(Ranker)

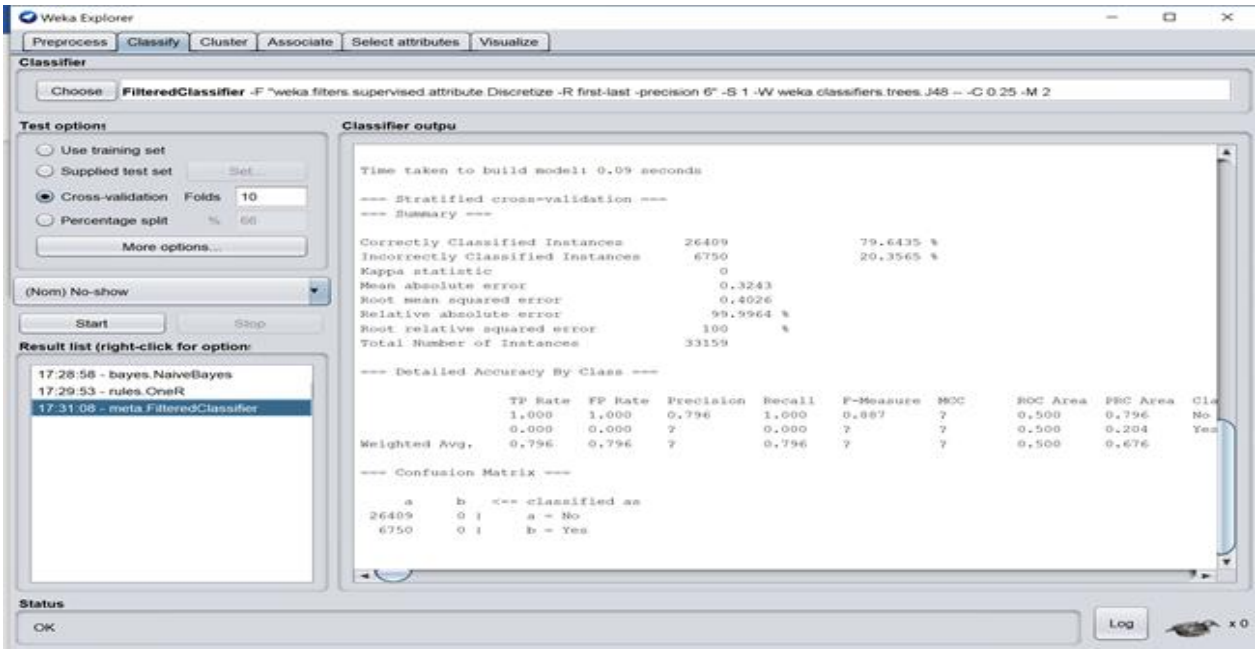
NaiveBayes



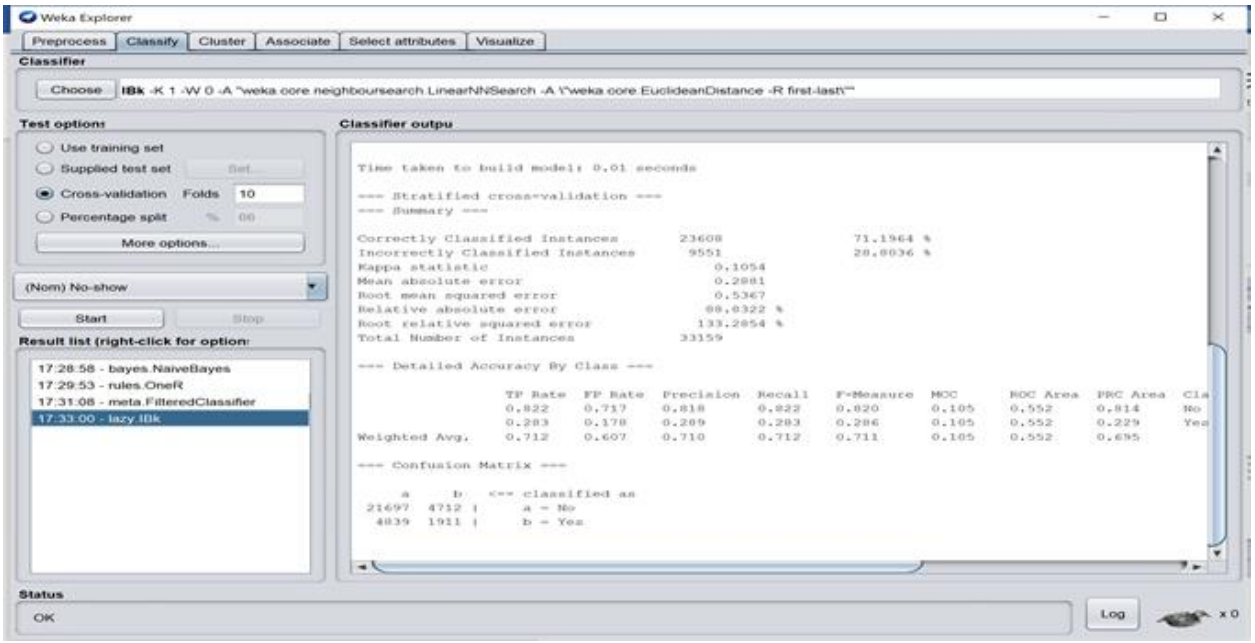
OneR



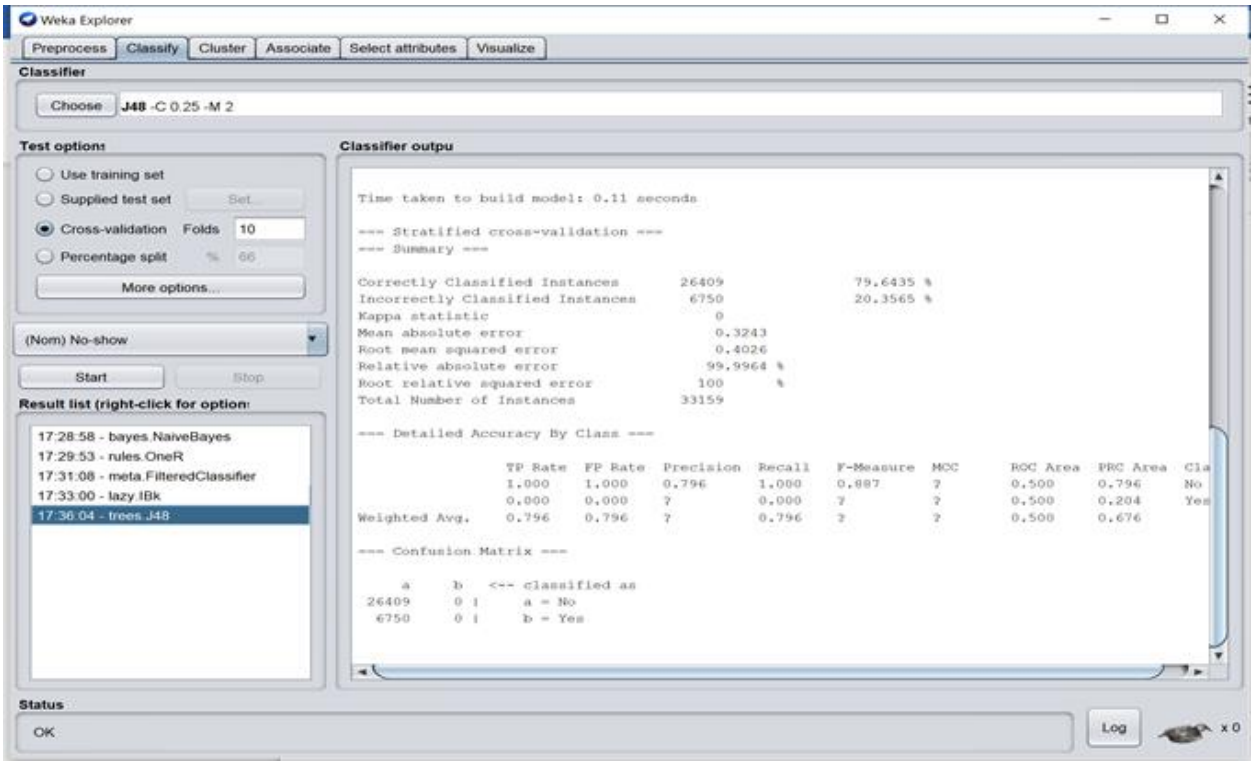
FilteredClassifier



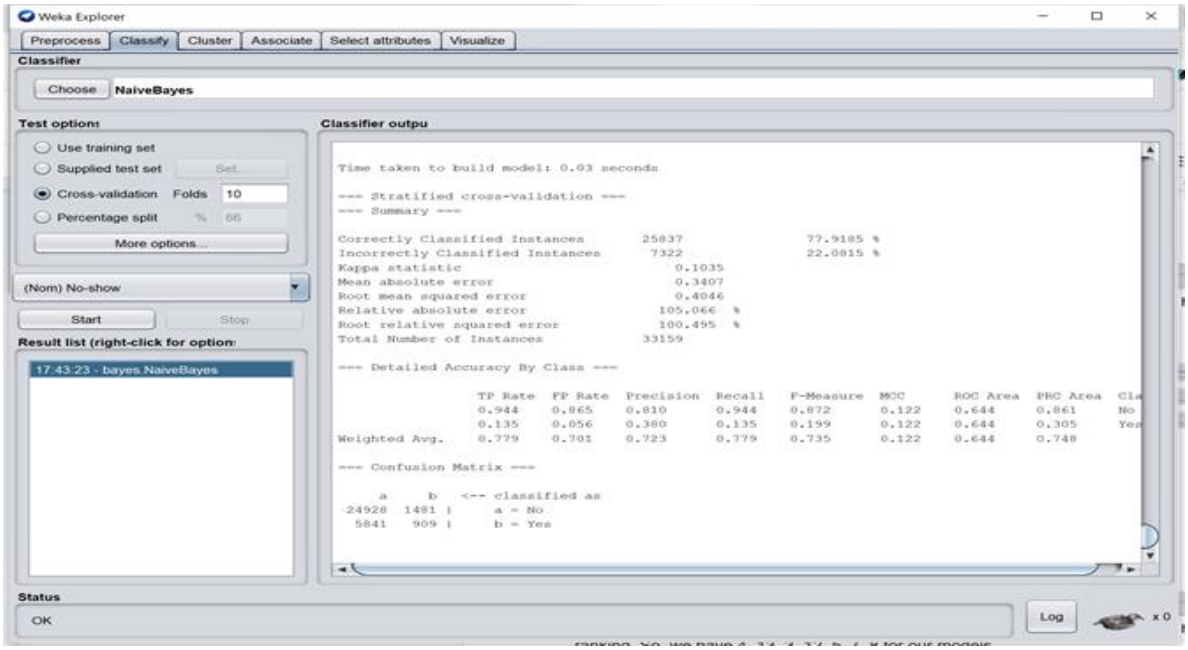
IBK



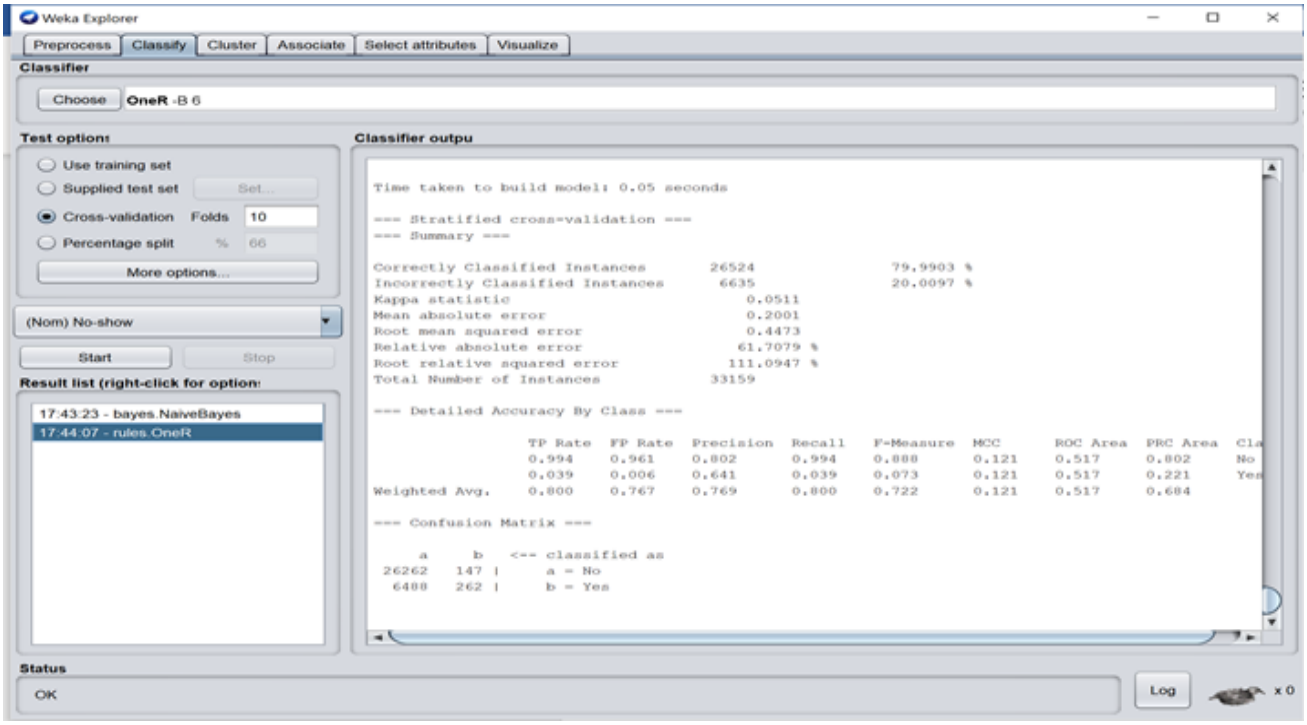
J48



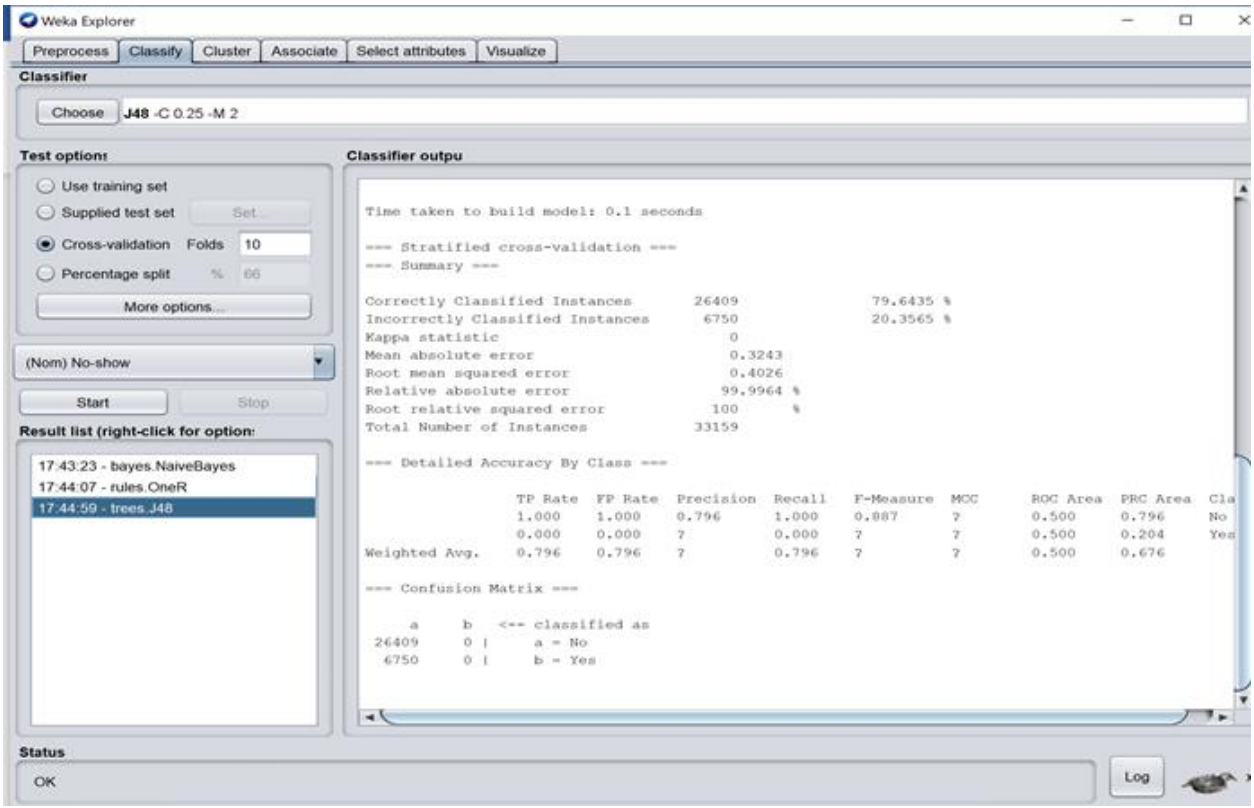
3. **OneRAttributeEval(Ranker):**
   NaiveBayes



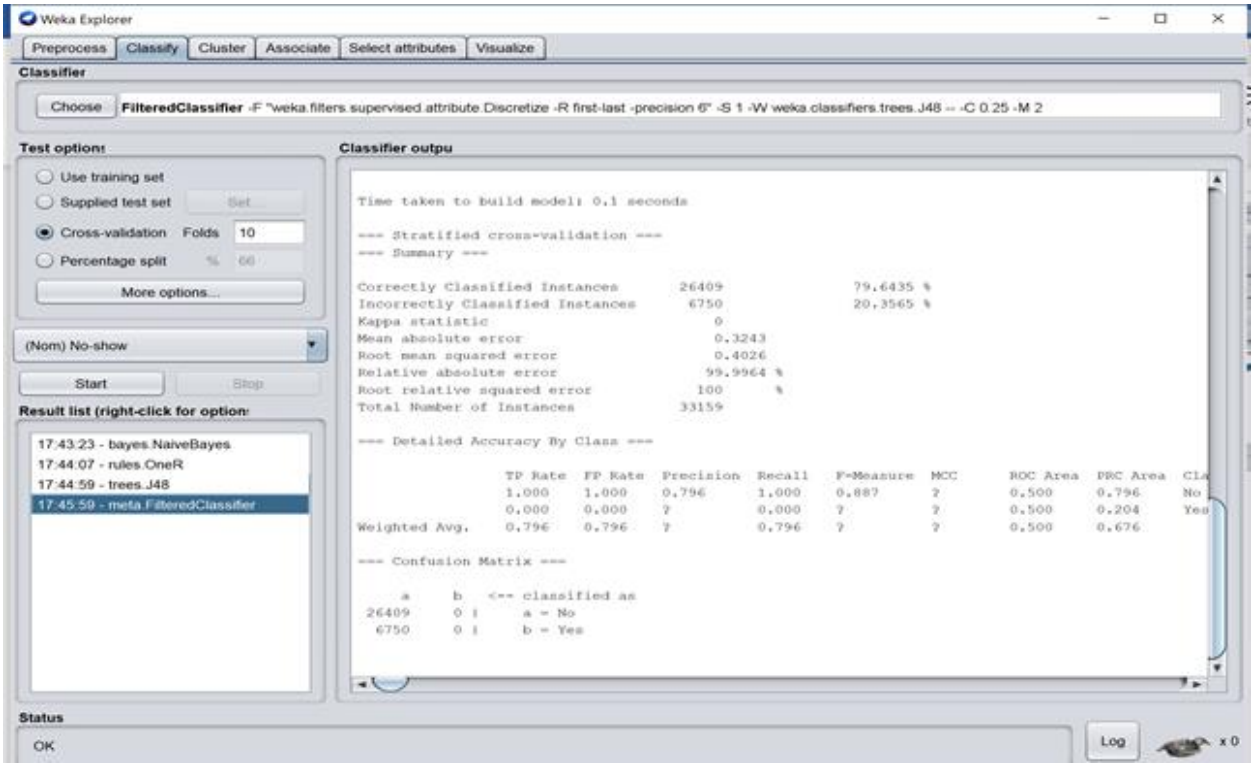
OneR



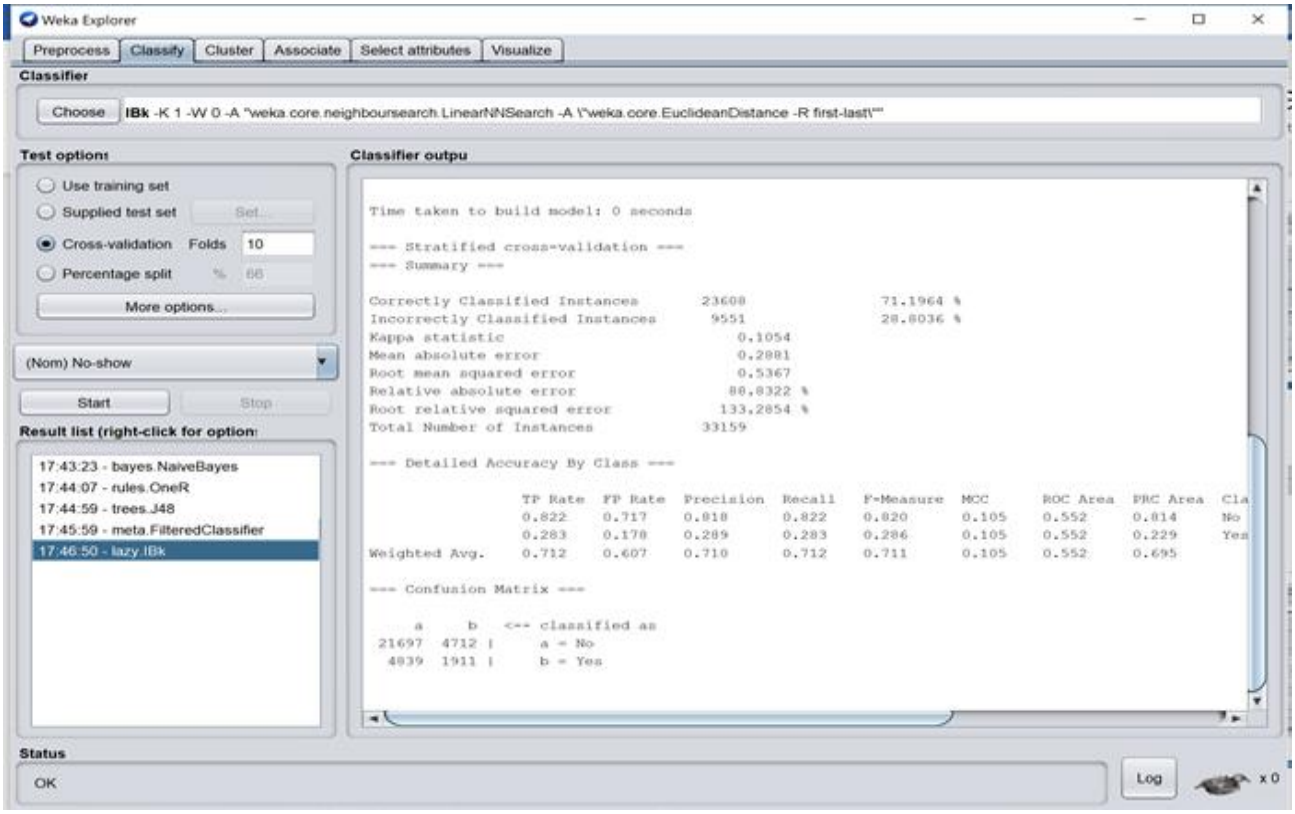
J48



FilteredClassifier

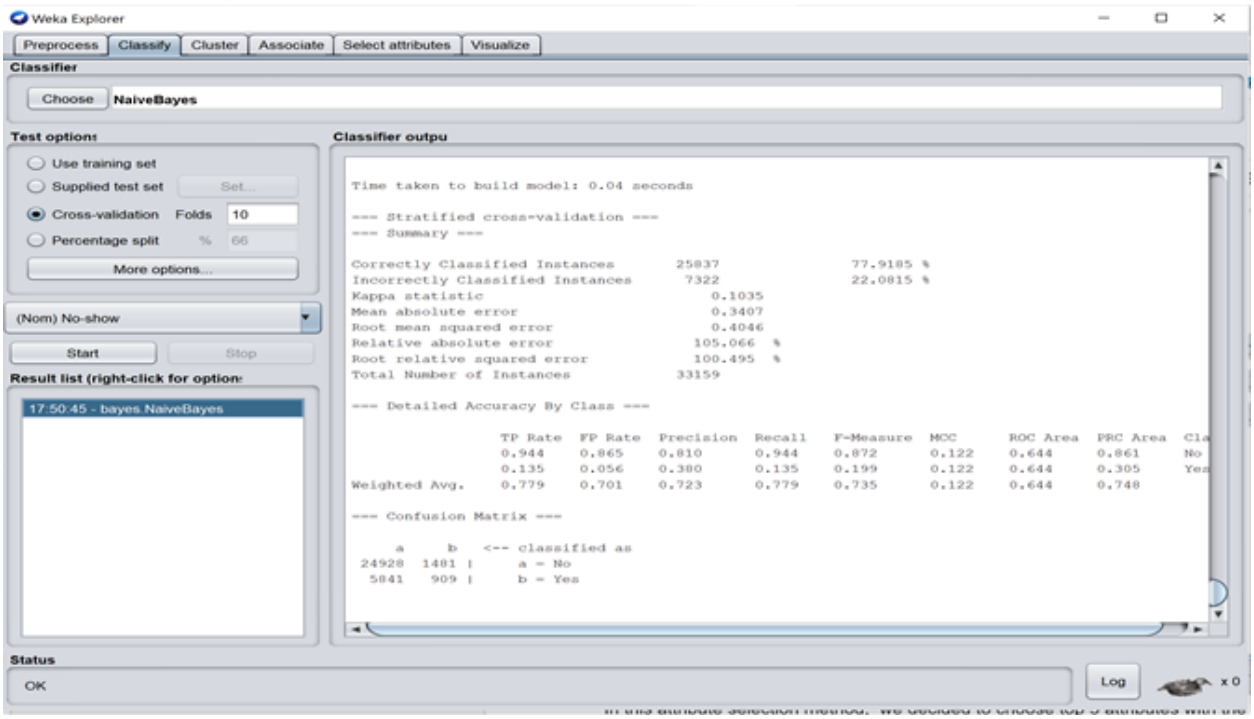

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
FilteredClassifier -F "weka.filters.supervised.attribute.Discretize -R first-last -precision 6" -S 1 -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Test option:
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
Time taken to build model: 0.1 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 26409 79.6435 %
Incorrectly Classified Instances 6750 20.3565 %
Kappa statistic 0
Mean absolute error 0.3243
Root mean squared error 0.4026
Relative absolute error 99.9964 %
Root relative squared error 100 %
Total Number of Instances 33159
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
1.000 1.000 0.796 1.000 0.887 ? 0.500 0.796 No
0.000 0.000 ? 0.000 ? ? 0.500 0.204 Yes
Weighted Avg. 0.796 0.796 ? 0.796 ? ? 0.500 0.676
=== Confusion Matrix ===
a b <-- classified as
26409 0 | a = No
6750 0 | b = Yes
(Nom) No-show
Start
Stop
Result list (right-click for option:
17:43:23 - bayes.NaiveBayes
17:44:07 - rules.OneR
17:44:59 - trees.J48
17:45:59 - meta.FilteredClassifier
Status
OK
Log
x 0

IBK

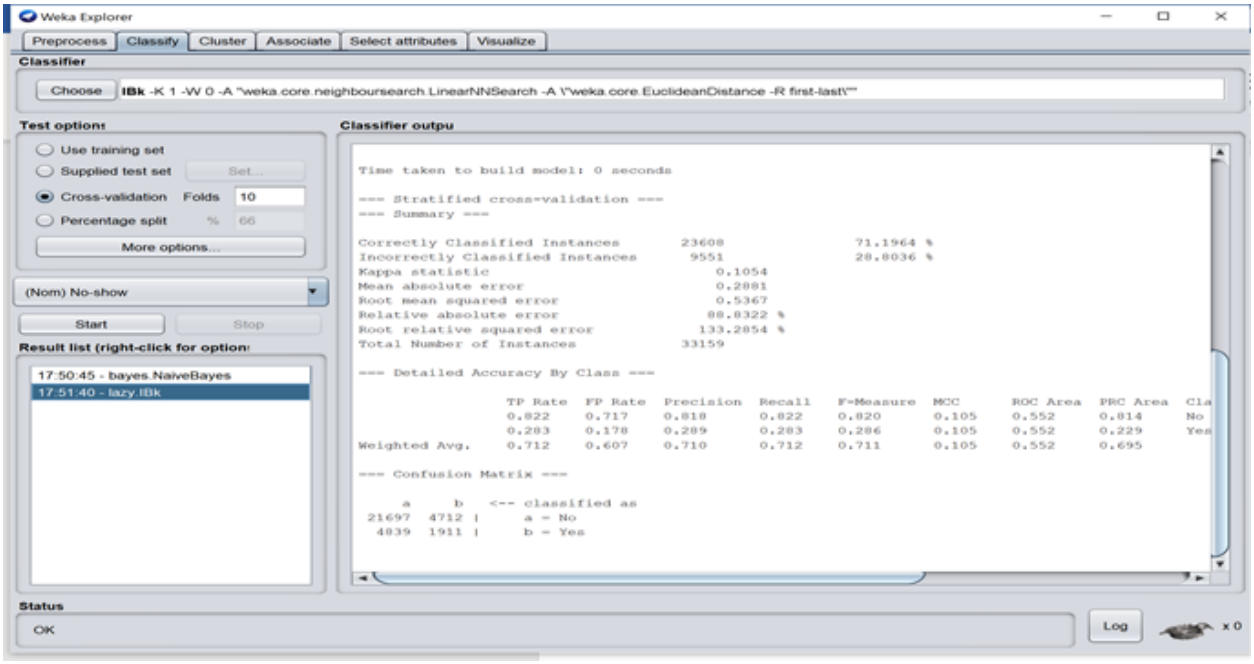

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Test option:
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
Time taken to build model: 0 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 23608 71.1964 %
Incorrectly Classified Instances 9551 28.8036 %
Kappa statistic 0.1054
Mean absolute error 0.2881
Root mean squared error 0.5367
Relative absolute error 88.8322 %
Root relative squared error 133.2854 %
Total Number of Instances 33159
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.822 0.717 0.818 0.822 0.820 0.105 0.552 0.814 No
0.283 0.178 0.289 0.283 0.286 0.105 0.552 0.229 Yes
Weighted Avg. 0.712 0.607 0.710 0.712 0.711 0.105 0.552 0.695
=== Confusion Matrix ===
a b <-- classified as
21697 4712 | a = No
4839 1911 | b = Yes
(Nom) No-show
Start
Stop
Result list (right-click for option:
17:43:23 - bayes.NaiveBayes
17:44:07 - rules.OneR
17:44:59 - trees.J48
17:45:59 - meta.FilteredClassifier
17:46:50 - lazy.IBk
Status
OK
Log
x 0

4. **ClassifierAttributeEval(Ranker)**

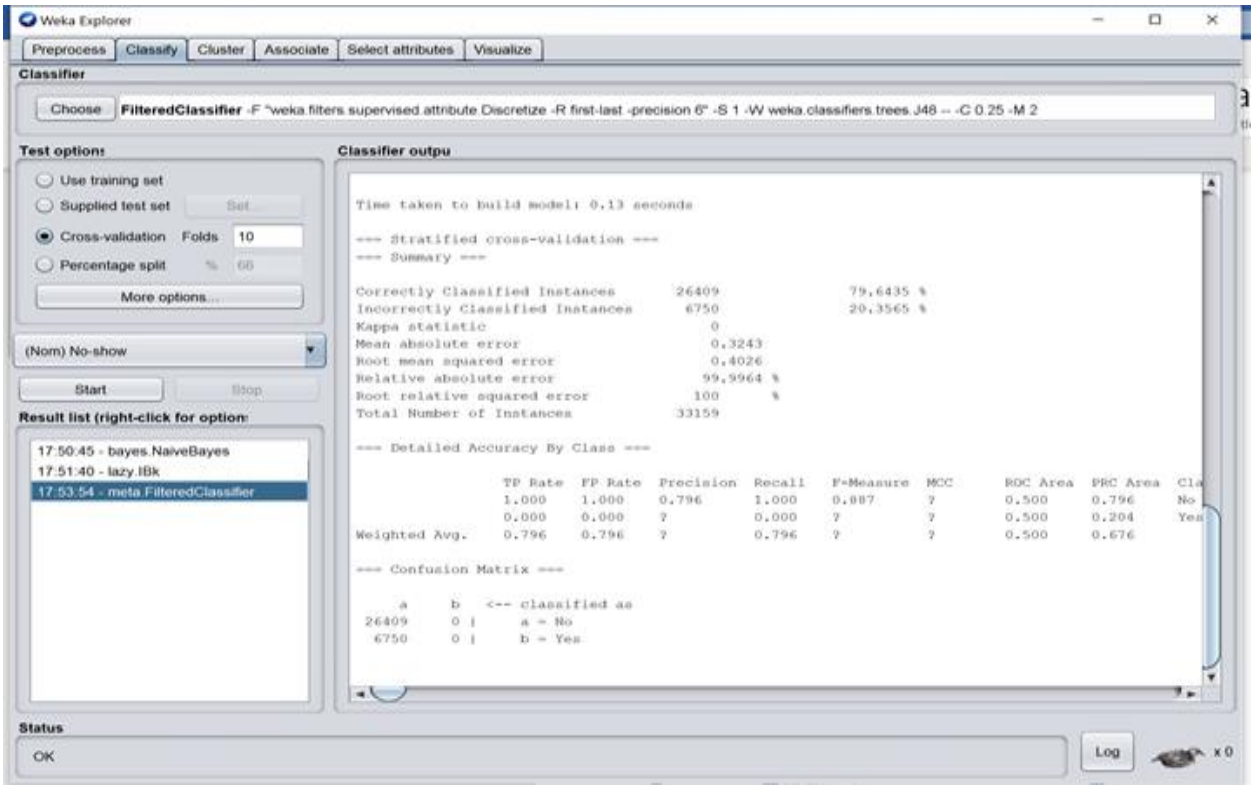
NaiveBayes



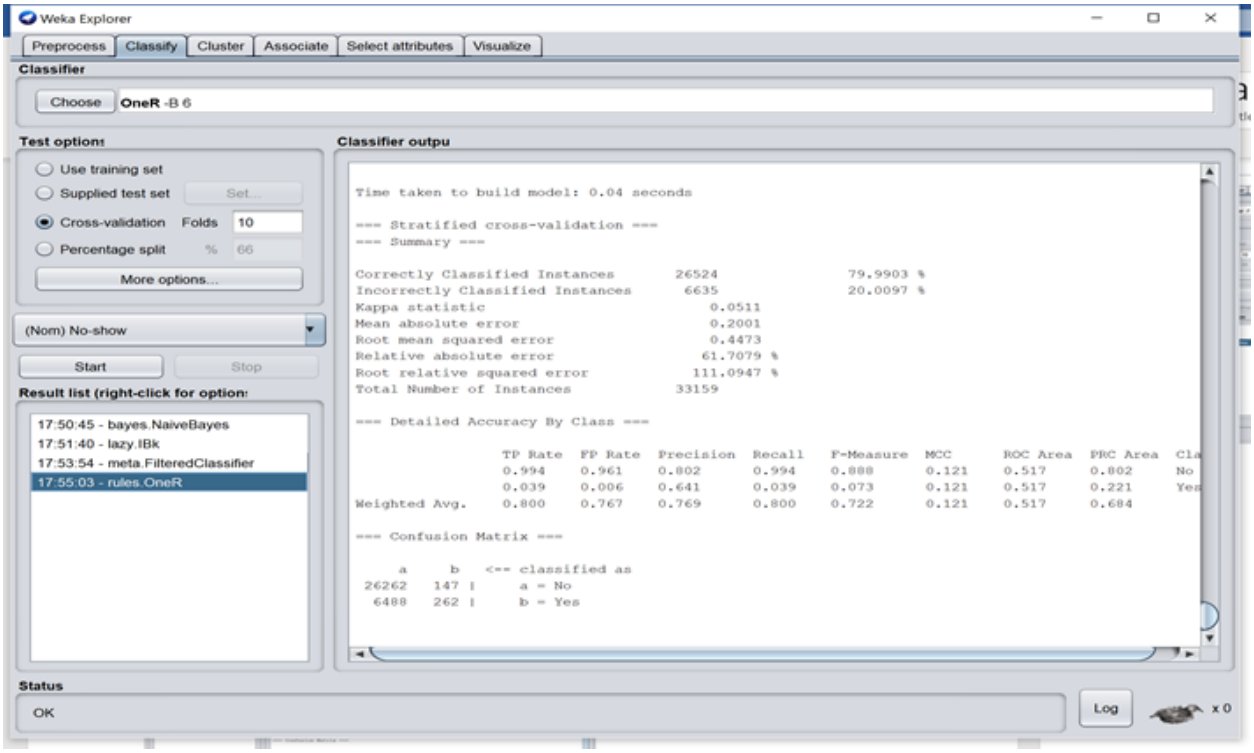
IBK



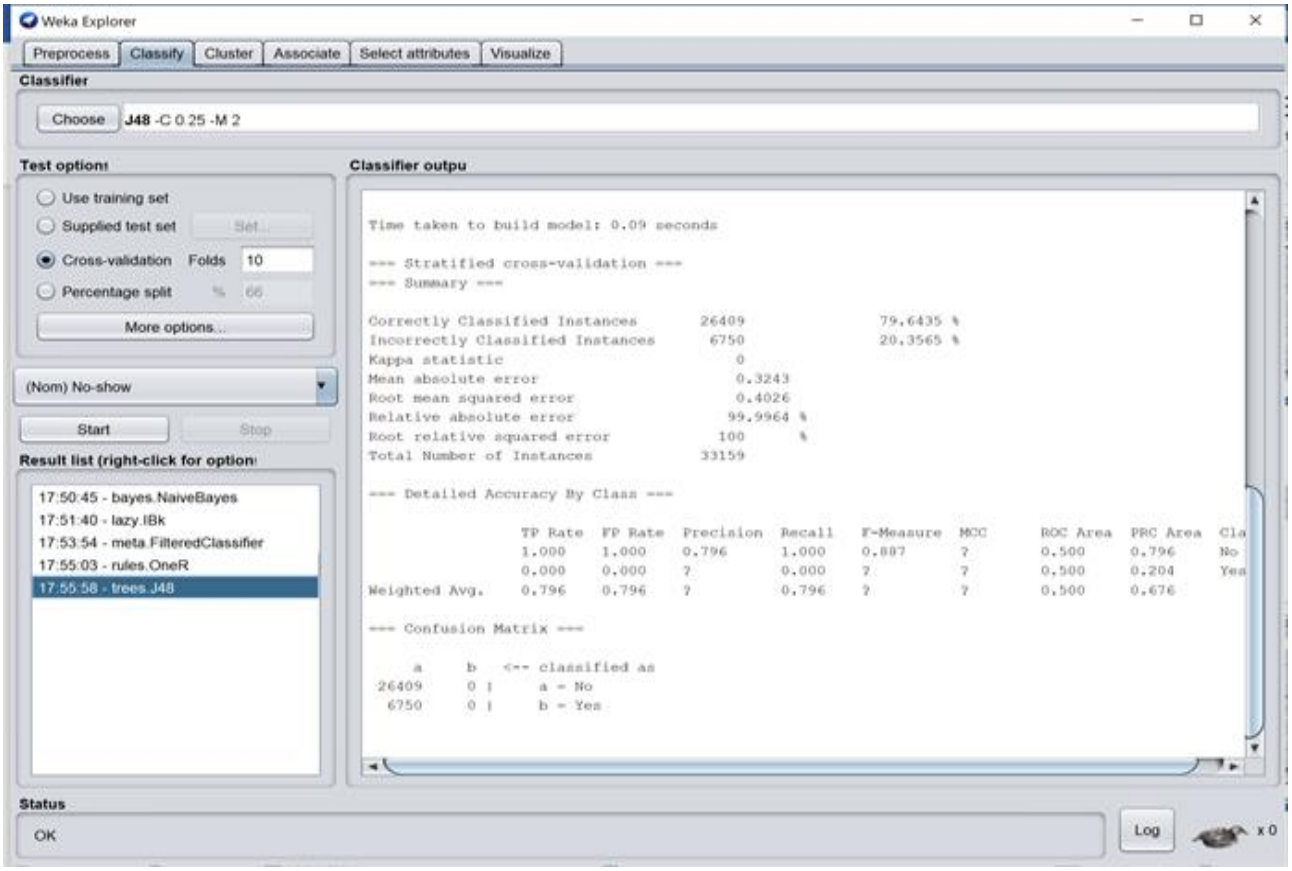
FilteredClassifier



OneR

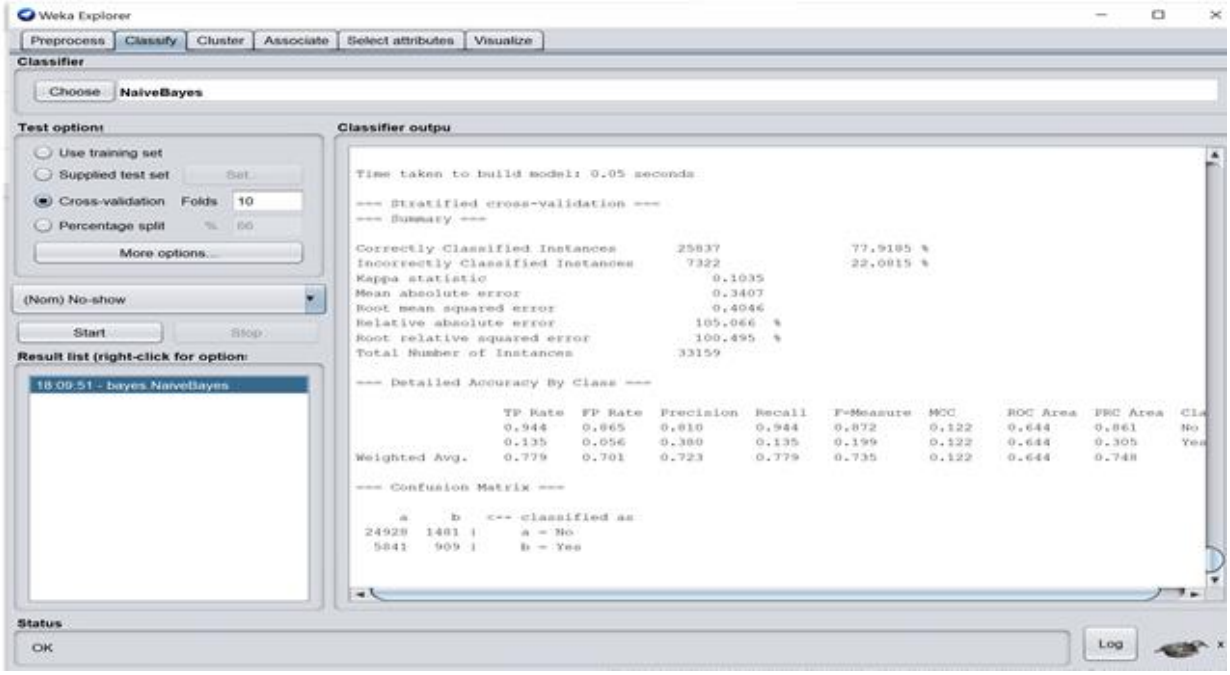


J48

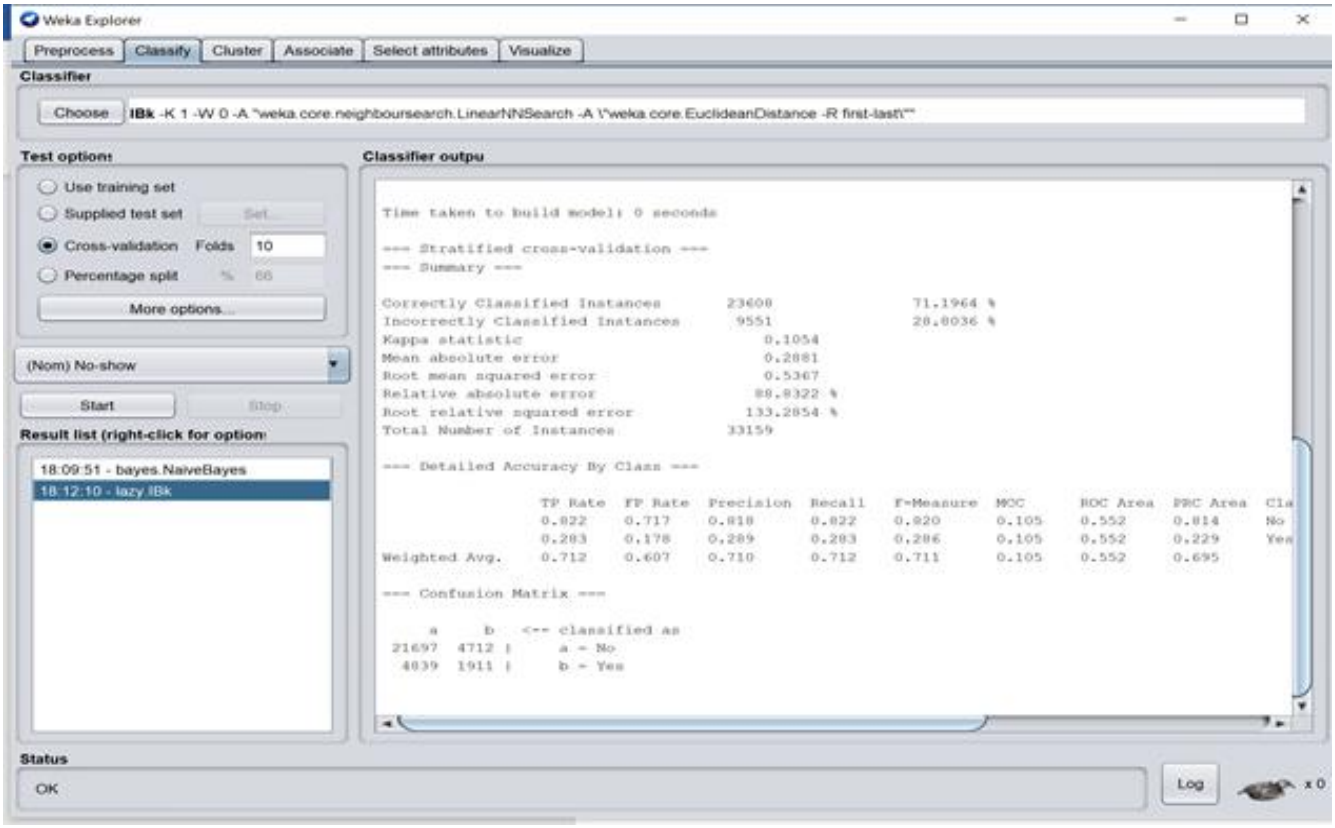


5. Personal Selection

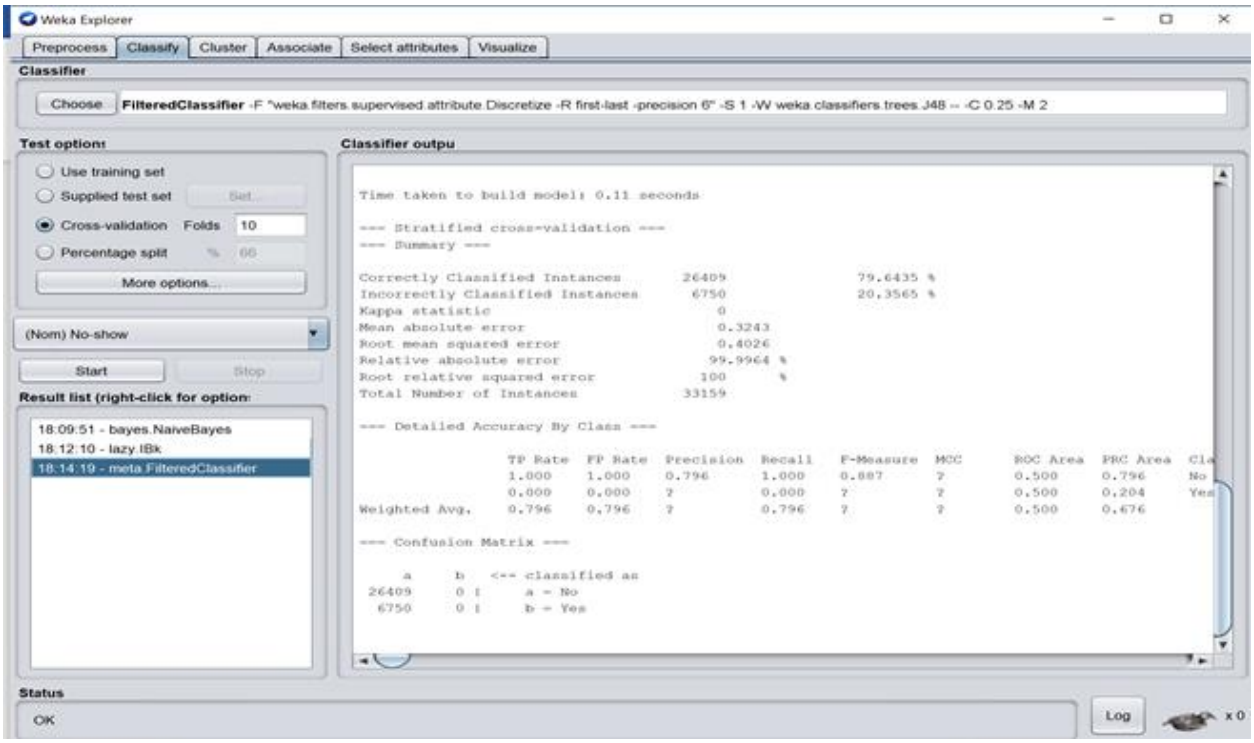
NaiveBayes



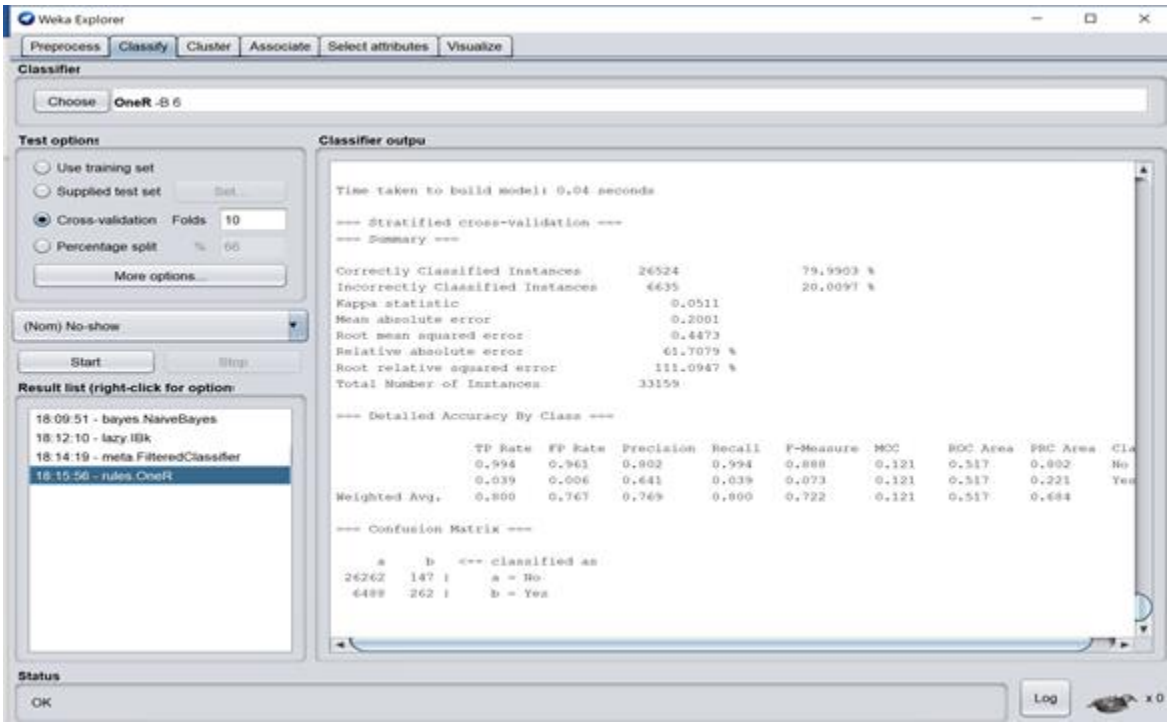
IBK



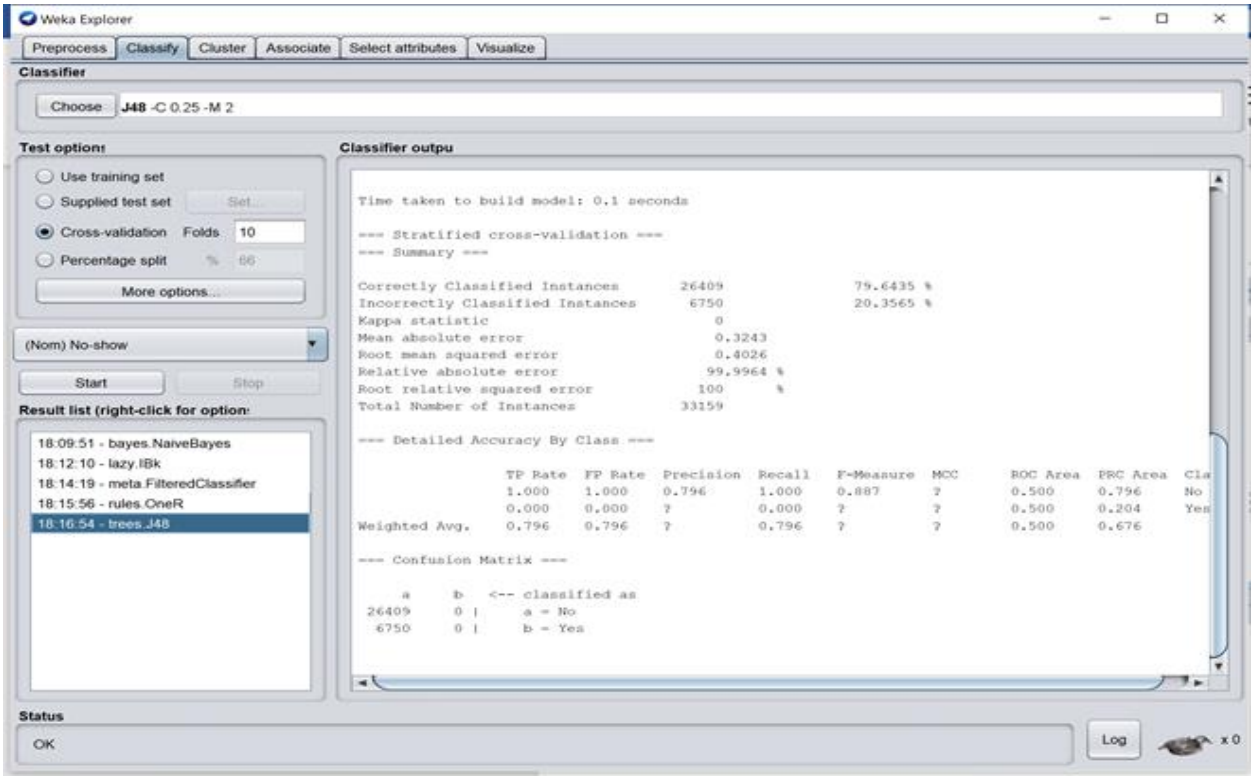
FilteredClassifier



OneR


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
OneR -B 6
Test option:
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
(Nom) No-show
Start
Stop
Result list (right-click for option:
18:09:51 - bayes.NaiveBayes
18:12:10 - lazy.IBk
18:14:19 - meta.FilteredClassifier
18:15:56 - rules.OneR
Time taken to build model: 0.04 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 26524 79.9903 %
Incorrectly Classified Instances 4635 20.0097 %
Kappa statistic 0.0511
Mean absolute error 0.2001
Root mean squared error 0.4473
Relative absolute error 61.7079 %
Root relative squared error 111.0947 %
Total Number of Instances 33159
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
0.994 0.961 0.802 0.994 0.888 0.121 0.517 0.802 No
0.039 0.006 0.641 0.039 0.073 0.121 0.517 0.221 Yes
Weighted Avg. 0.800 0.767 0.769 0.800 0.722 0.121 0.517 0.684
=== Confusion Matrix ===
a b <-- classified as
26262 147 | a = No
4488 262 | b = Yes
Status
OK
Log
x 0

J48


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test option:
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
(Nom) No-show
Start
Stop
Result list (right-click for option:
18:09:51 - bayes.NaiveBayes
18:12:10 - lazy.IBk
18:14:19 - meta.FilteredClassifier
18:15:56 - rules.OneR
18:16:54 - trees.J48
Time taken to build model: 0.1 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 26409 79.6435 %
Incorrectly Classified Instances 6750 20.3565 %
Kappa statistic 0
Mean absolute error 0.3243
Root mean squared error 0.4026
Relative absolute error 99.9964 %
Root relative squared error 100 %
Total Number of Instances 33159
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Cla
1.000 1.000 0.796 1.000 0.887 ? 0.500 0.796 No
0.000 0.000 ? 0.000 ? ? 0.500 0.204 Yes
Weighted Avg. 0.796 0.796 ? 0.796 ? ? 0.500 0.676
=== Confusion Matrix ===
a b <-- classified as
26409 0 | a = No
6750 0 | b = Yes
Status
OK
Log
x 0

CONCLUSION

Training Dataset

	NaiveBays	J48	Filtered Classifier	OneR	IBk
GainRatioAttributeEval	78.4653%	79.8767%	79.8767%	80.1468%	72.1875%
InfoGainAttributeEval	78.4653%	79.8767%	79.8767%	80.1468%	72.1875%
OneRAttributeEval	78.4653%	79.8767%	79.8767%	80.1468%	72.1875%
ClassifierAttributeEval	78.4653%	79.8767%	79.8767%	80.1468%	72.1875%
Personal Selection	78.4653%	79.8767%	79.8767%	80.1468%	72.1875%

Testing Dataset

	NaiveBayes	J48	Filtered Classifier	OneR	IBk
GainRatioAttributeEval	77.9185%	79.6435%	79.6435%	79.9903%	71.1964%
InfoGainAttributeEval	77.9185%	79.6435%	79.6435%	79.9903%	71.1964%
OneRAttributeEval	77.9185%	79.6435%	79.6435%	79.9903%	71.1964%
ClassifierAttributeEval	77.9185%	79.6435%	79.6435%	79.9903%	71.1964%
Personal Selection	77.9185%	79.6435%	79.6435%	79.9903%	71.1964%

ANALYSIS/ RECOMMENDATIONS

Our aim is to reduce or accurately predict appointment no shows since optimized scheduling can greatly improve allocation of valuable doctor time. Particularly in a developing country like Brazil, where there is a shortage of qualified professionals, optimizing medical scheduling can help lower healthcare costs, provide broader healthcare access to more people, and even save lives.
From the above output,we can conclude that all the datasets show nice performance with different classification methods. We select OneR as our final model as it is better as compared to other models. To sum up, accuracy could have been greater if there had been more relevant features such as more disease records or some measure of how busy the person's schedule is etc.
It's extremely rare that a patient will just keep making appointments without ever showing up for any of them. In order to predictively classify if they will be a No-Show, we need more data about their past, and other information about competing priorities at the time of their appointment.

The attributes we selected for our final model are:

Appointment_Day

Appointment_ID

SMS_received

Age

Neighbourhood

Appointment_Day

Hypertension

During this project, the main gains for us are:

- How to select suitable dataset for classification
- How to select valuable attributes relevant to our models
- How to use Weka in efficient manner

Shrutika Singodia - data collection, attribute description, attribute selection, data visualization, interpreting results, determining proposal,report writing

Ashi Choudhary - data collection, data preprocessing, attribute selection, data classification, data visualization, analyzing results , report writing