Career Longevity for NBA Rookies

Data Mining

-By VENKATA VISHNUVARDHAN ALURI

Goal of our Data Mining project

The aim of this project is to predict the career longevity for the National Basketball Association (NBA) rookies. This is achieved by performing several data mining tasks on the data set such as data preprocessing, attribute selection, data classifier modelling and performance comparison. The class attribute for prediction is the Target_5Yrs attribute details of which will be elaborated in the following section of this report.

Description of Dataset

This Dataset consists of twenty-one attributes with thirteen hundred and forty instances. Each record gives us the information about the player's basketball career such as games and minutes played, field goals made and so forth. All the attributes excluding the Name and the class attribute, are numeric attributes. The class attribute i.e. the Target_5Yrs is a nominal attribute bearing Boolean values 0 and 1 for prediction. Each attribute description is elaborated in the following table.

	Description
Name	Name
GP	Games Played
MIN	MinutesPlayed
PTS	PointsPerGame
FGM	FieldGoalsMade
FGA	FieldGoalAttempts
FG%	FieldGoalPercent
3P Made	3PointMade
3PA	3PointAttempts
3P%	3PointAttempts
FTM	FreeThrowMade
FTA	FreeThrowAttempts
FT%	FreeThrowPercent
OREB	OffensiveRebounds
DREB	DefensiveRebounds
REB	Rebounds
AST	Assists
STL	Steals
BLK	Blocks
TOV	Turnovers
TARGET_5Yrs	Outcome: 1 if career length >= 5 yrs, 0 if < 5...

Description of tools involved

Weka

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection.

R Studio

RStudio is a free and open-source integrated development environment for R, a programming language for statistical computing and graphics. RStudio was founded by JJ Allaire, creator of the programming language ColdFusion.

Microsoft Excel

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, MAC OS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications.

Data Mining Algorithms used

Information Gain

ID3 uses information gain as its attribute selection measure. This measure is based on pioneering work by Claude Shannon on information theory, which studied the value or “information content” of messages. The attribute with the highest information gain is chosen as the splitting attribute, it minimizes the information needed to classify the tuples and reflects the least randomness.

Gain Ratio

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. C4.5, a successor of ID3, uses an extension to information gain known as gain ratio, which attempts to overcome this bias. It applies a kind of normalization to information gain using a “split information” value and represents the potential information generated by splitting the training data set into v partitions, corresponding to the v outcomes of a test on the attribute.

Correlation Attribute Evaluation

Evaluates the worth of an attribute by measuring the correlation (Pearson's) between it and the class.

OneR AttributeEval

Evaluates the worth of an attribute by using the OneR classifier.

Classifiers

J48

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm, the critical distribution of the data is easily understandable. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data-mining tool, J48 is an open source Java implementation of the C4.5 algorithm.

Naïve Bayes

Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class-conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naive."

Adaboost

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire, who won the 2003 Gödel Prize for their work. It can be used in conjunction with many other types of learning algorithms to improve performance. The output of the other learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers.

Bagging

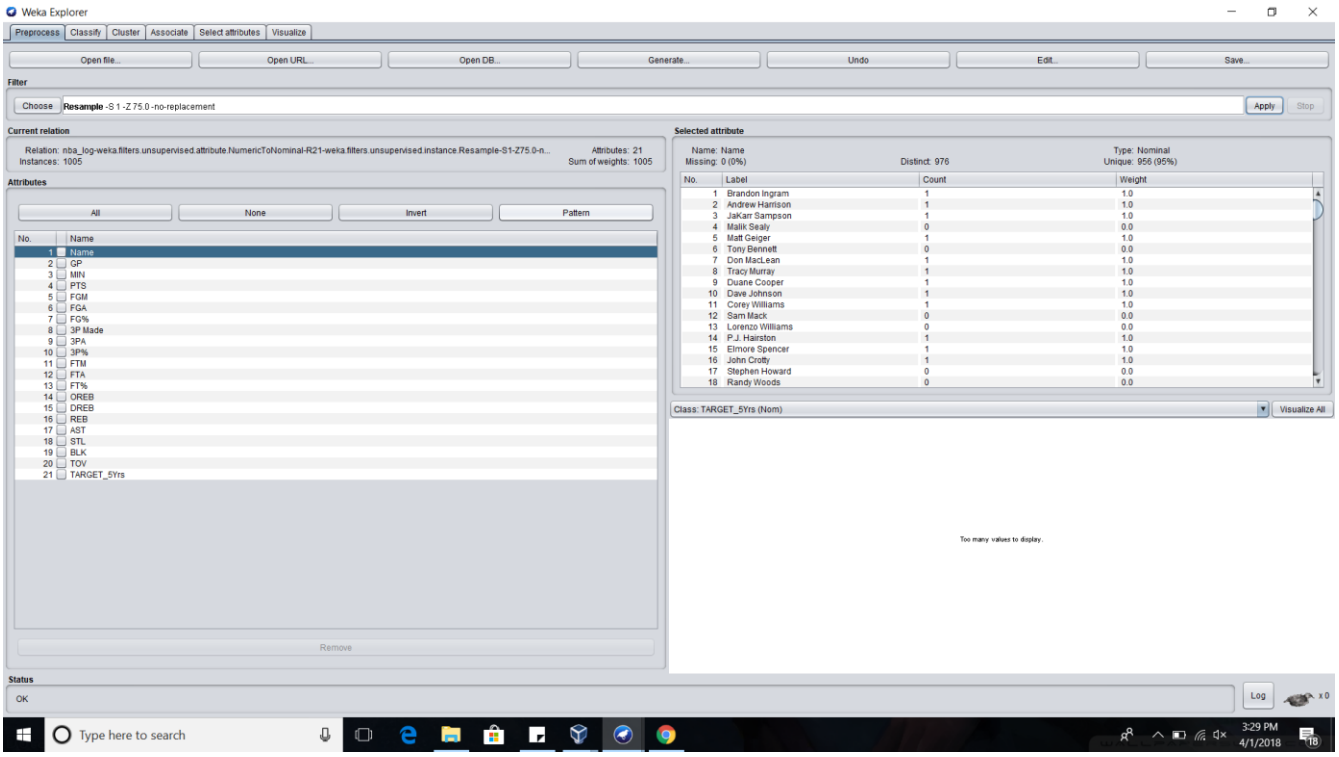
Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. Although it is usually applied to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

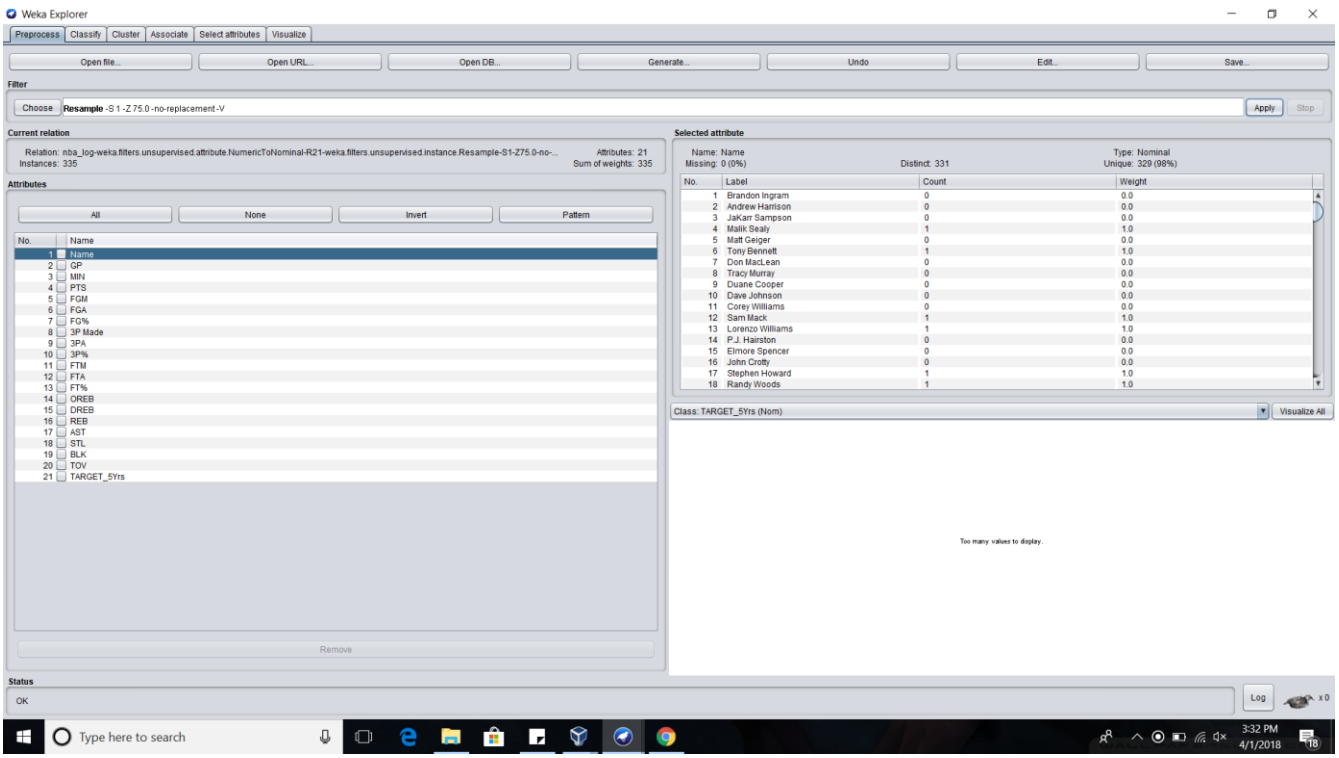
Procedure

Preprocessing

Datasets are never ideal. Our dataset has several missing values that should be addressed before classification and prediction in order to achieve accuracy. Data Smoothing helps addressing the issue where missing values are replaced with mean values of respective attributes. This is implemented using R studio where attribute values are fed in R to obtain mean.

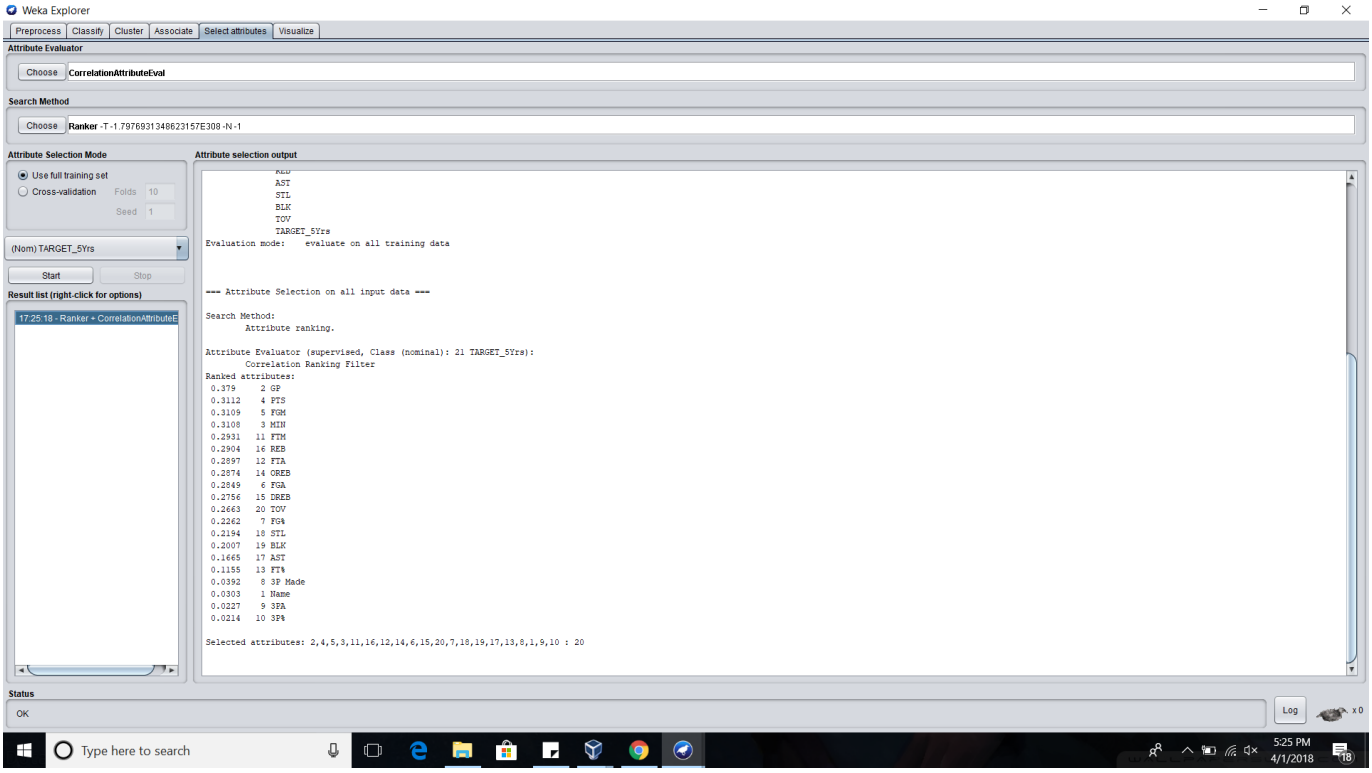
Weka implements several filters to preprocess the data set, once the dataset was loaded, it has to be split into training and test set for modelling and testing. Selecting an unsupervised filter, resampling and dividing the set into 75% training and 25 % test by applying ‘True’ to the no replacement option the training set was achieved. The test set was achieved by applying ‘True’ to the invert selection while resampling. Below are screenshots for training and test sets after split. 1340 instances are split into 1005 and 335 respectively.

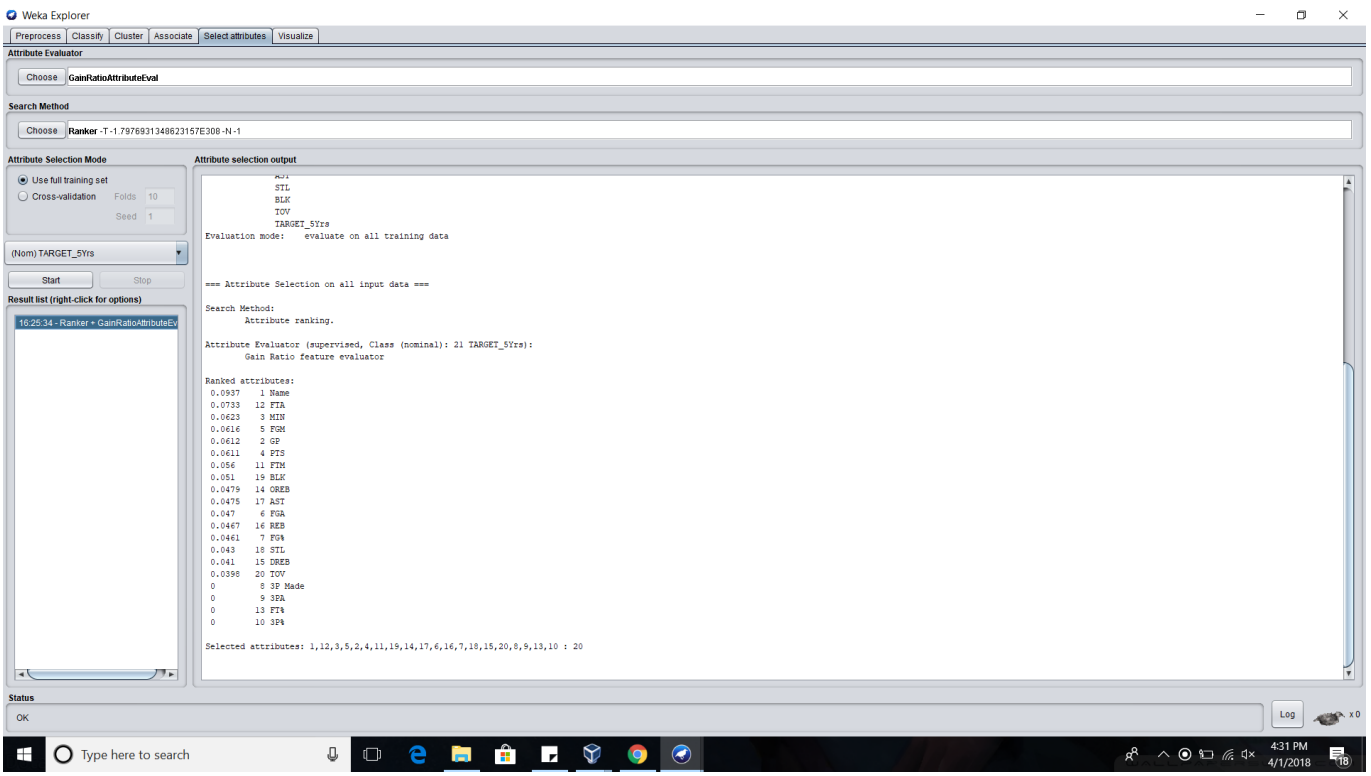


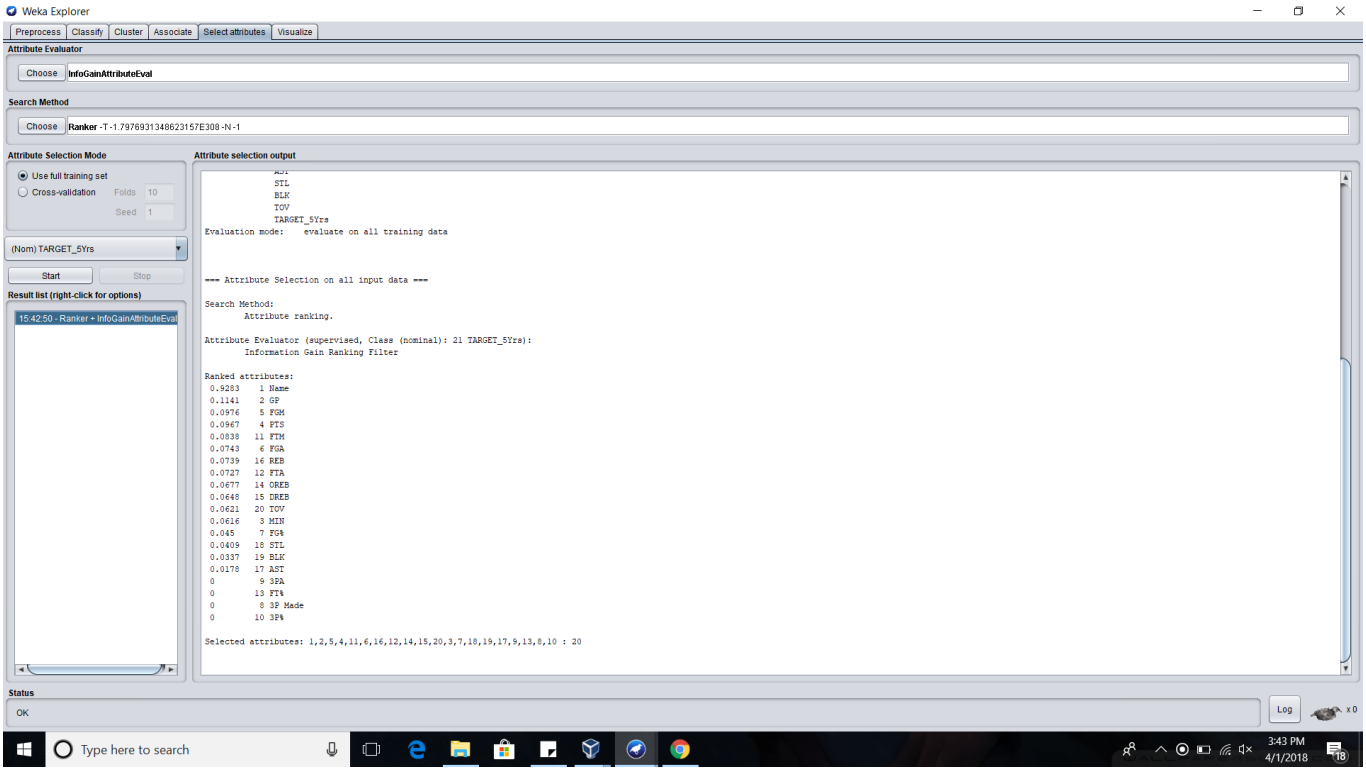


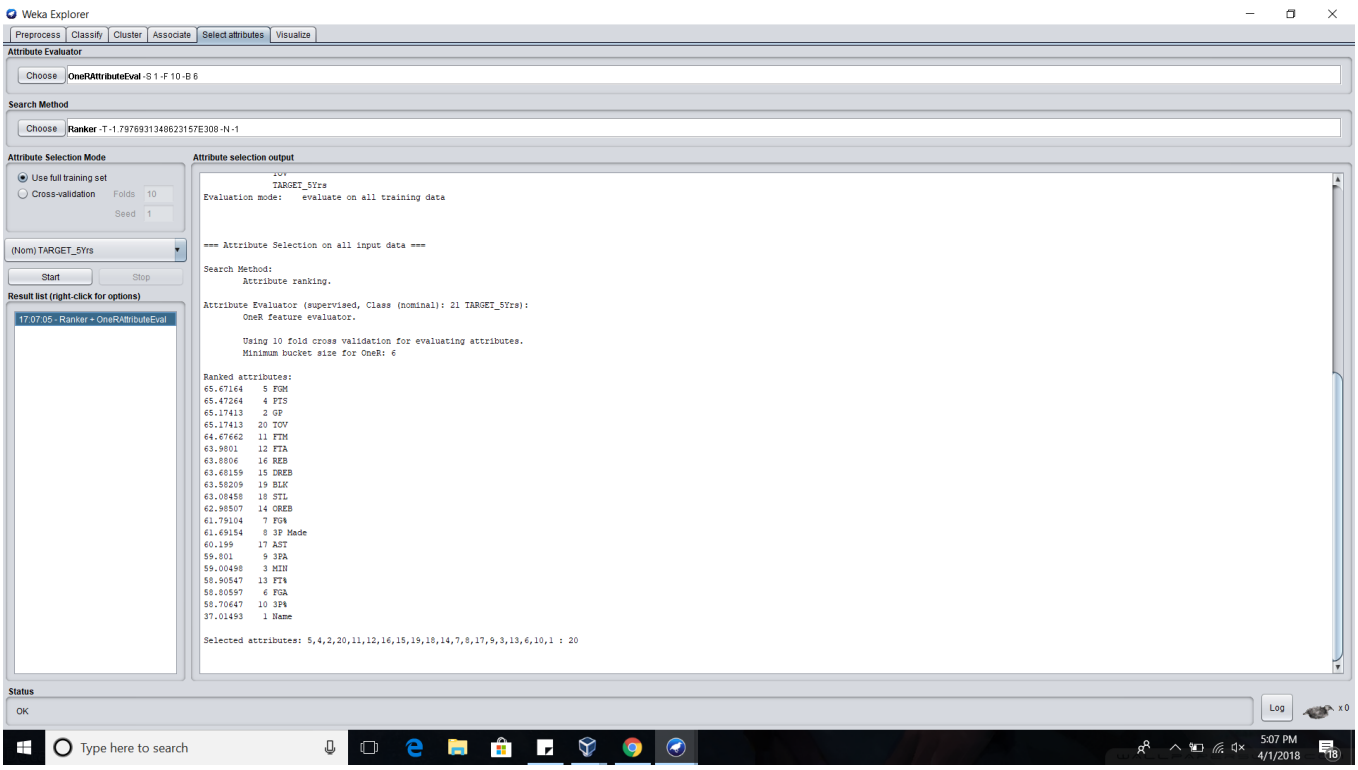
Attribute Selection

Once the dataset is divided into training and test sets, we click on Select Attributes and run each of the algorithms. For each we select the attributes that are least likely to influence the dataset and remove them from the training set. Below are the screenshots for ranks of attributes after selecting algorithms Correlation Attribute Evaluation-Ranked, Gain Ratio-Ranked, Info Gain-Ranked and OneR – Ranked in order.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
CorrelationAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) TARGET_5Yrs
Start
Stop
Result list (right-click for options)
17:25:18 - Ranker + CorrelationAttributeE
Attribute selection output
AST
STL
BLK
TOV
TARGET_5Yrs
Evaluation mode: evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 21 TARGET_5Yrs):
Correlation Ranking Filter
Ranked attributes:
0.379 2 GP
0.3112 4 PTS
0.3109 5 FGM
0.3108 3 MIN
0.2931 11 FTM
0.2904 16 REB
0.2897 12 FTA
0.2874 14 OREB
0.2849 6 FGA
0.2756 15 DREB
0.2663 20 TOV
0.2262 7 FG%
0.2194 18 STL
0.2007 19 BLK
0.1665 17 AST
0.1155 13 FT%
0.0392 8 3P Made
0.0303 1 Name
0.0227 9 3PA
0.0214 10 3P%
Selected attributes: 2,4,5,3,11,16,12,14,6,15,20,7,18,19,17,13,8,1,9,10 : 20
Status
OK
Log
x 0
Type here to search
5:25 PM
4/1/2018


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
GainRatioAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds 10
Seed 1
(Nom) TARGET_5Yrs
Start
Stop
Result list (right-click for options)
16:25:34 - Ranker + GainRatioAttributeEv
Attribute selection output
STL
BLK
TOV
TARGET_5Yrs
Evaluation mode: evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 21 TARGET_5Yrs):
Gain Ratio feature evaluator
Ranked attributes:
0.0937 1 Name
0.0733 12 FTA
0.0623 3 MIN
0.0616 5 FGM
0.0612 2 GP
0.0611 4 PTS
0.056 11 FTM
0.051 19 BLK
0.0479 14 OREB
0.0475 17 AST
0.047 6 FGA
0.0467 16 REB
0.0461 7 FG%
0.043 18 STL
0.041 15 DREB
0.0398 20 TOV
0 8 3P Made
0 9 3PA
0 13 FT%
0 10 3P%
Selected attributes: 1,12,3,5,2,4,11,19,14,17,6,16,7,18,15,20,8,9,13,10 : 20
Status
OK
Log
x 0
Type here to search
4:31 PM
4/1/2018

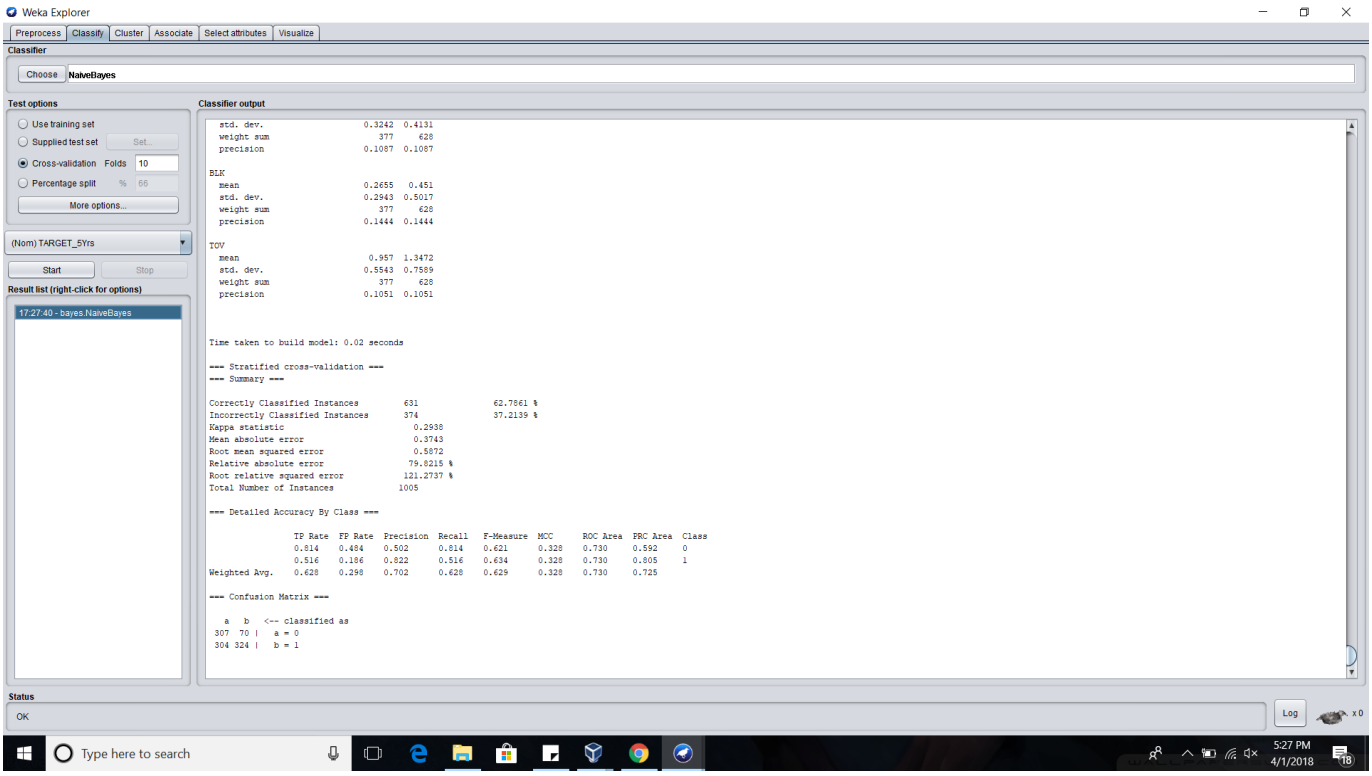

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Attribute Evaluator
Choose
InfoGainAttributeEval
Search Method
Choose
Ranker -T -1.7976931348623157E308 -N -1
Attribute Selection Mode
Use full training set
Cross-validation
Folds
10
Seed
1
(Nom) TARGET_5Yrs
Start
Stop
Result list (right-click for options)
15:42:50 - Ranker + InfoGainAttributeEval
Attribute selection output
STL
BLK
TOV
TARGET_5Yrs
Evaluation mode: evaluate on all training data
=== Attribute Selection on all input data ===
Search Method:
Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 21 TARGET_5Yrs):
Information Gain Ranking Filter
Ranked attributes:
0.9283 1 Name
0.1141 2 GP
0.0976 5 FGM
0.0967 4 PTS
0.0838 11 FTM
0.0743 6 FGA
0.0739 16 REB
0.0727 12 FTA
0.0677 14 OREB
0.0648 15 DREB
0.0621 20 TOV
0.0616 3 MIN
0.045 7 FG%
0.0409 18 STL
0.0337 19 BLK
0.0178 17 AST
0 9 3PA
0 13 FT%
0 8 3P Made
0 10 3P%
Selected attributes: 1,2,5,4,11,6,16,12,14,15,20,3,7,18,19,17,9,13,8,10 : 20
Status
OK
Log
x 0
Type here to search
3:43 PM
4/1/2018



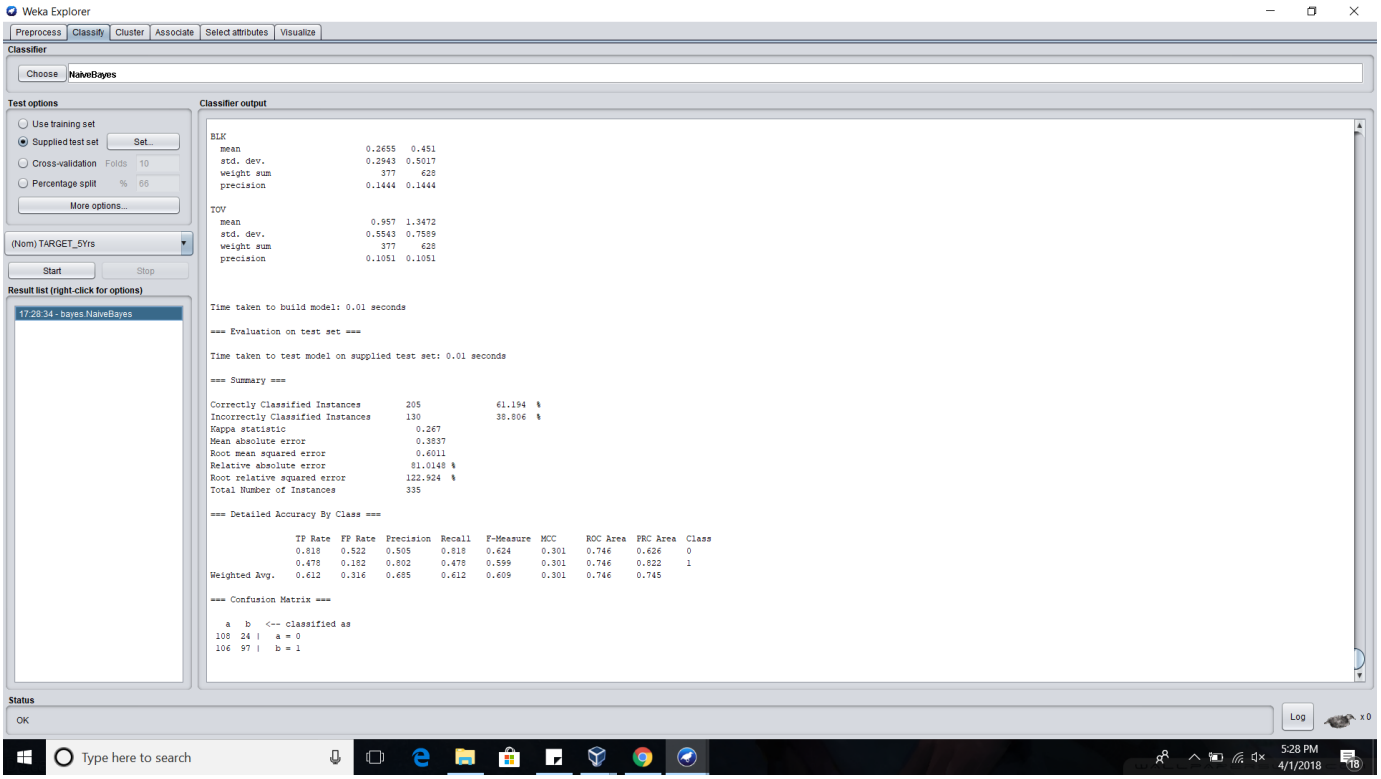
Classification

After the attributes are removed based on the result of the select attribute algorithm, we click on the classify tab of Weka. Here we use several classification algorithms as mentioned above for each select attribute algorithm and report our findings. While classifying, the test set is supplied in the supplied test slot. Screenshots for every classification with respect to each selection attributes are followed.

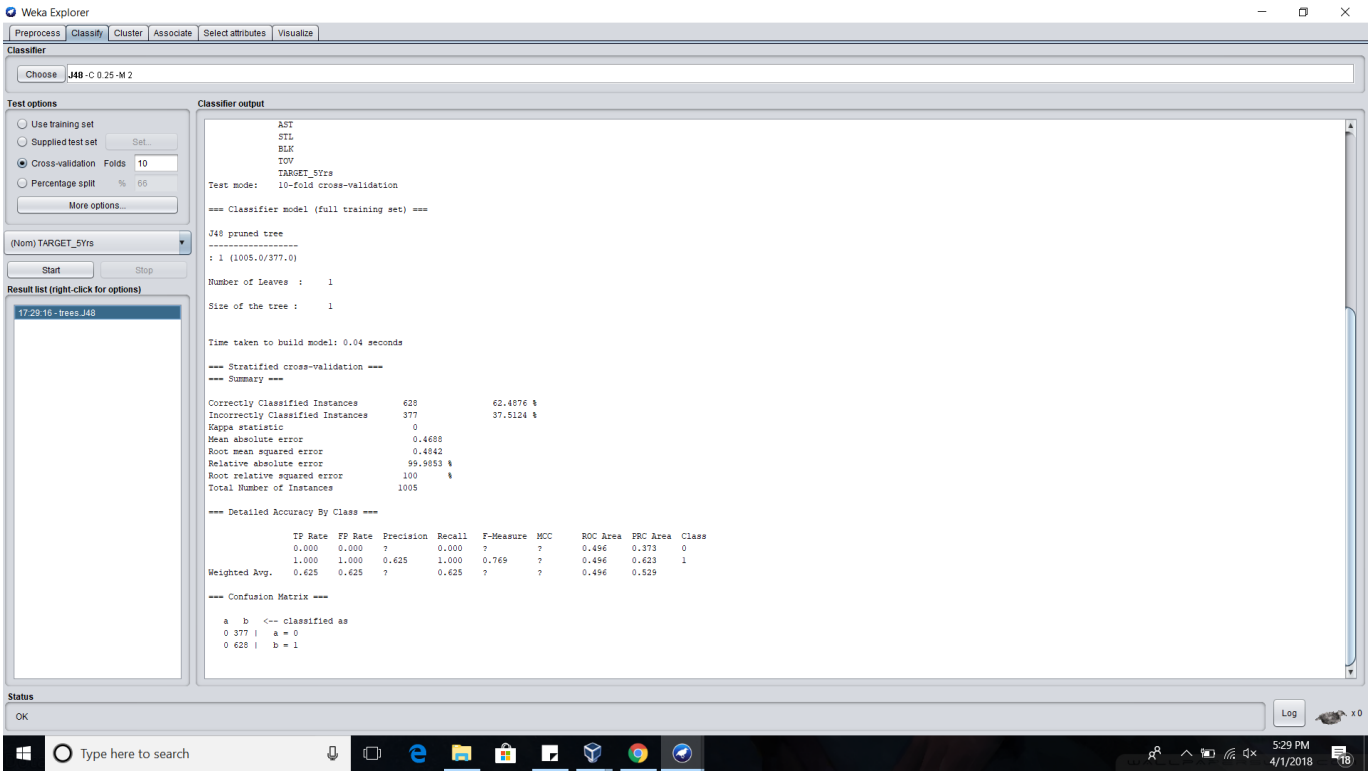
Correlation Attribute Evaluation – Naïve Bayes(Training)



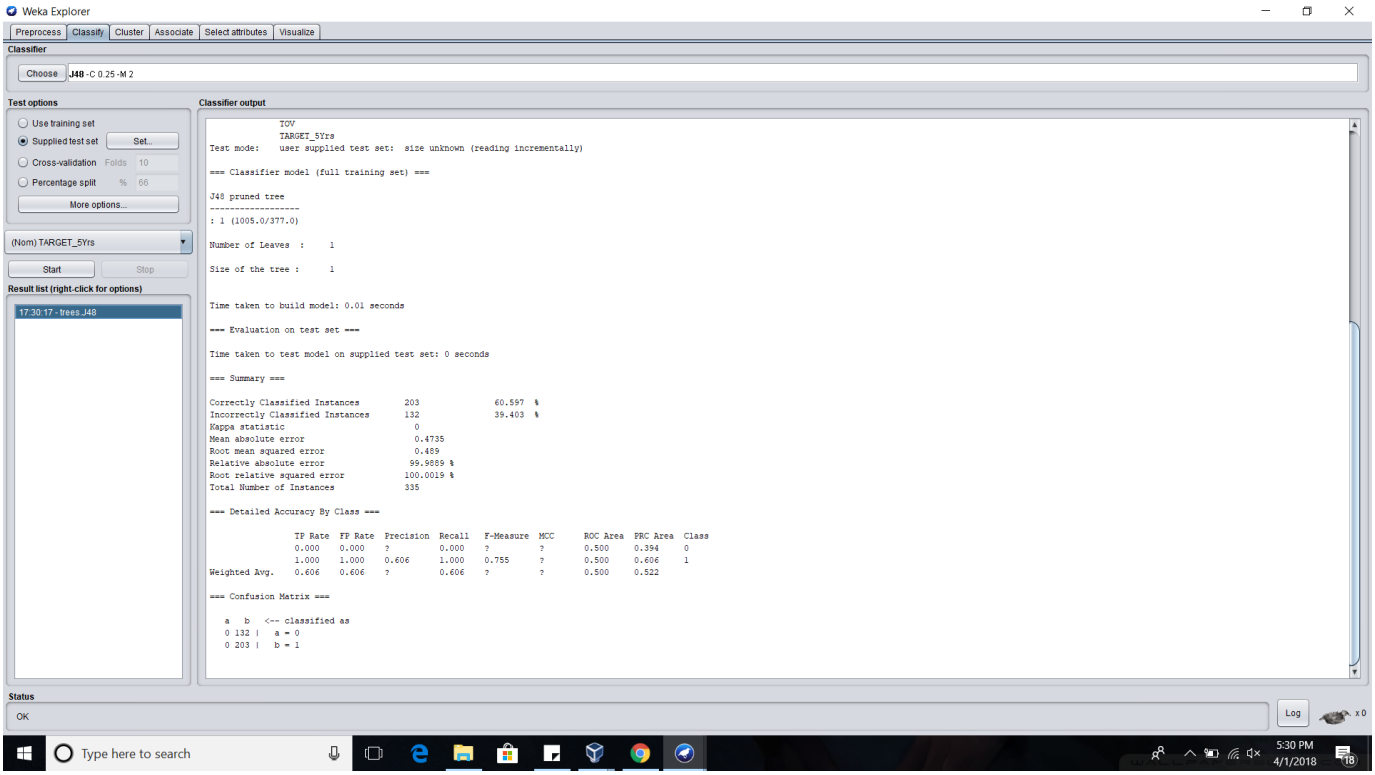
Correlation Attribute Evaluation – Naïve Bayes(Test)



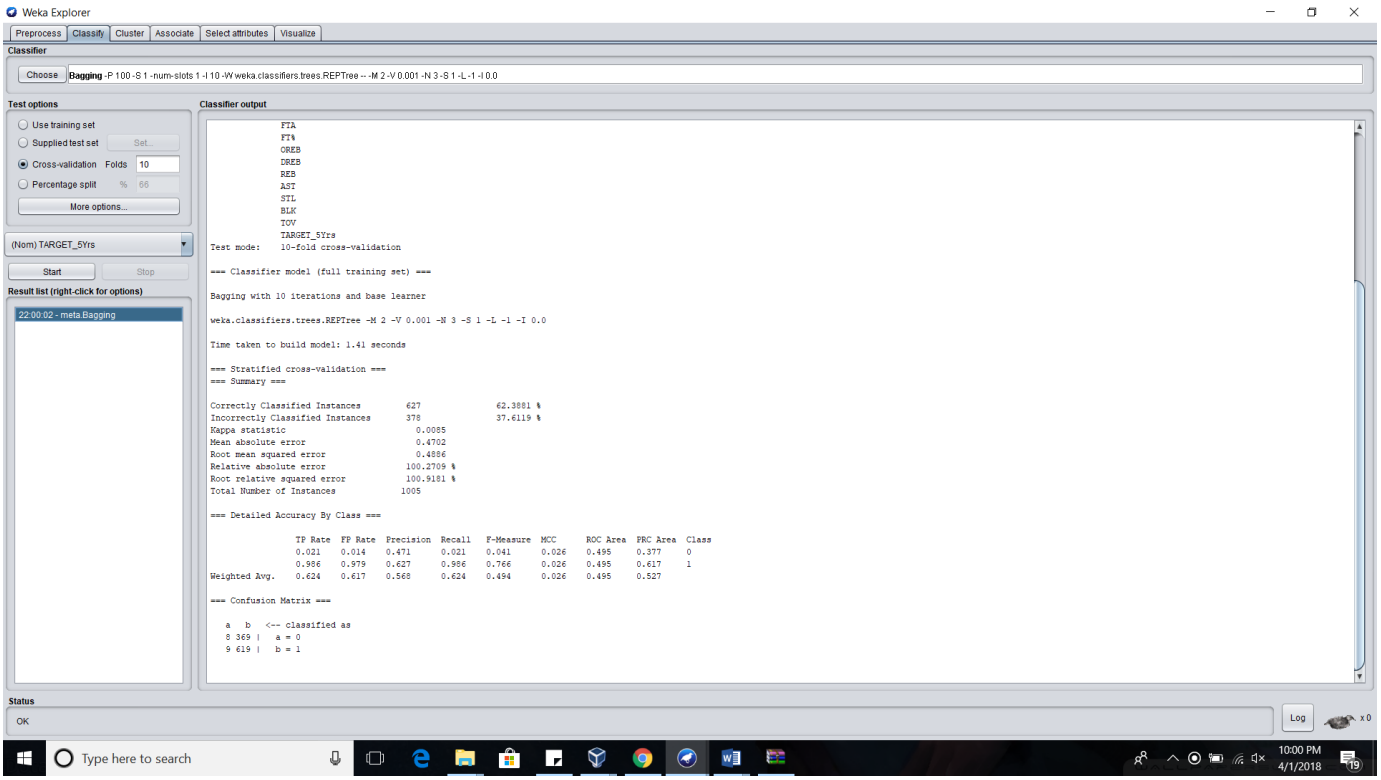
Correlation Attribute Evaluation – J48(Training)



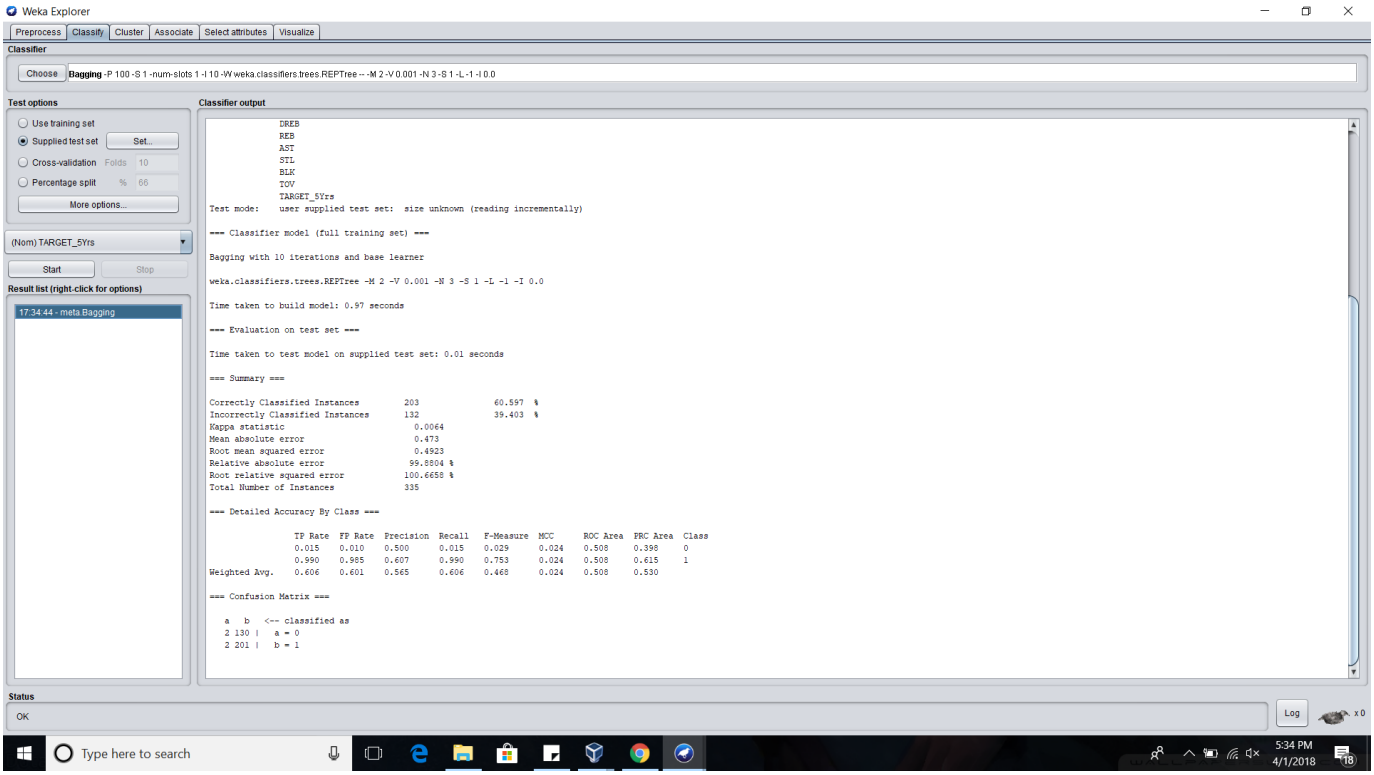
Correlation Attribute Evaluation – J48(Test)



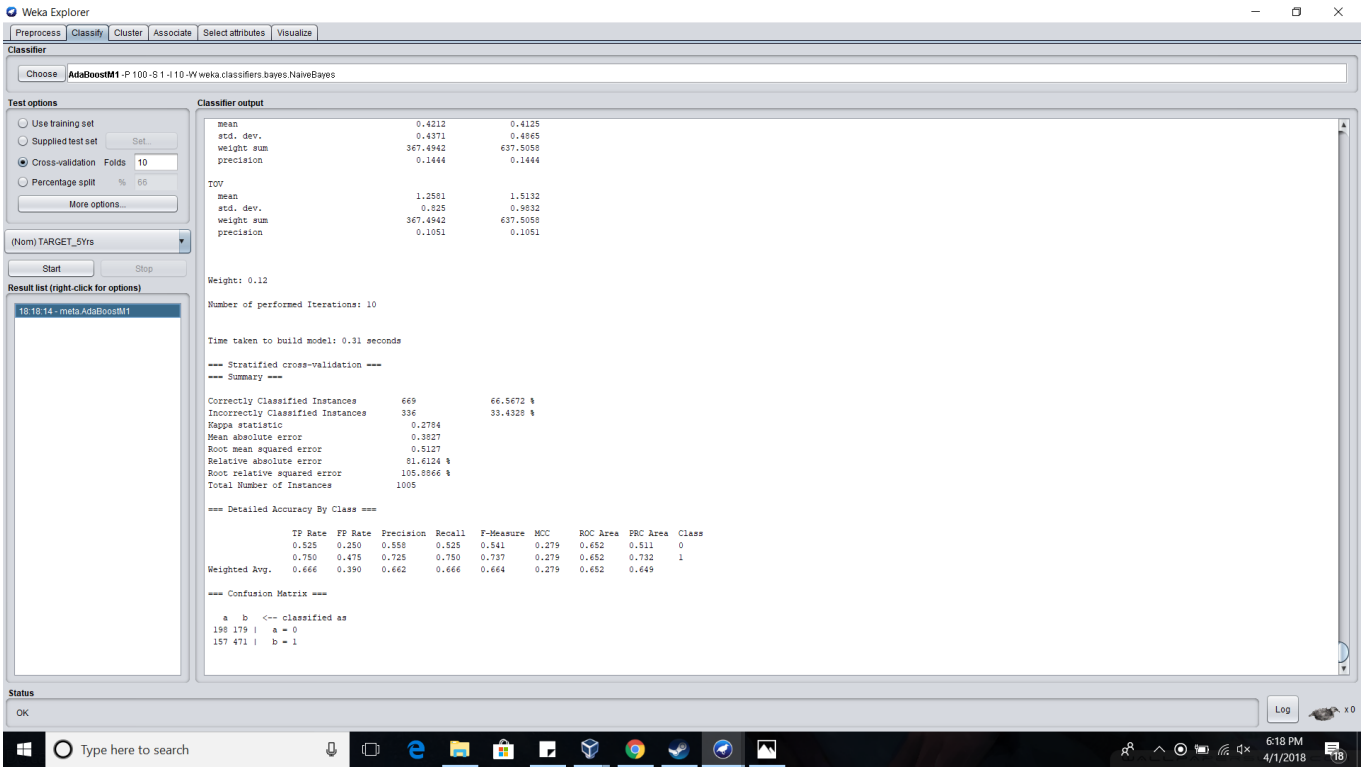
Correlation Attribute Evaluation – Bagging(Training)



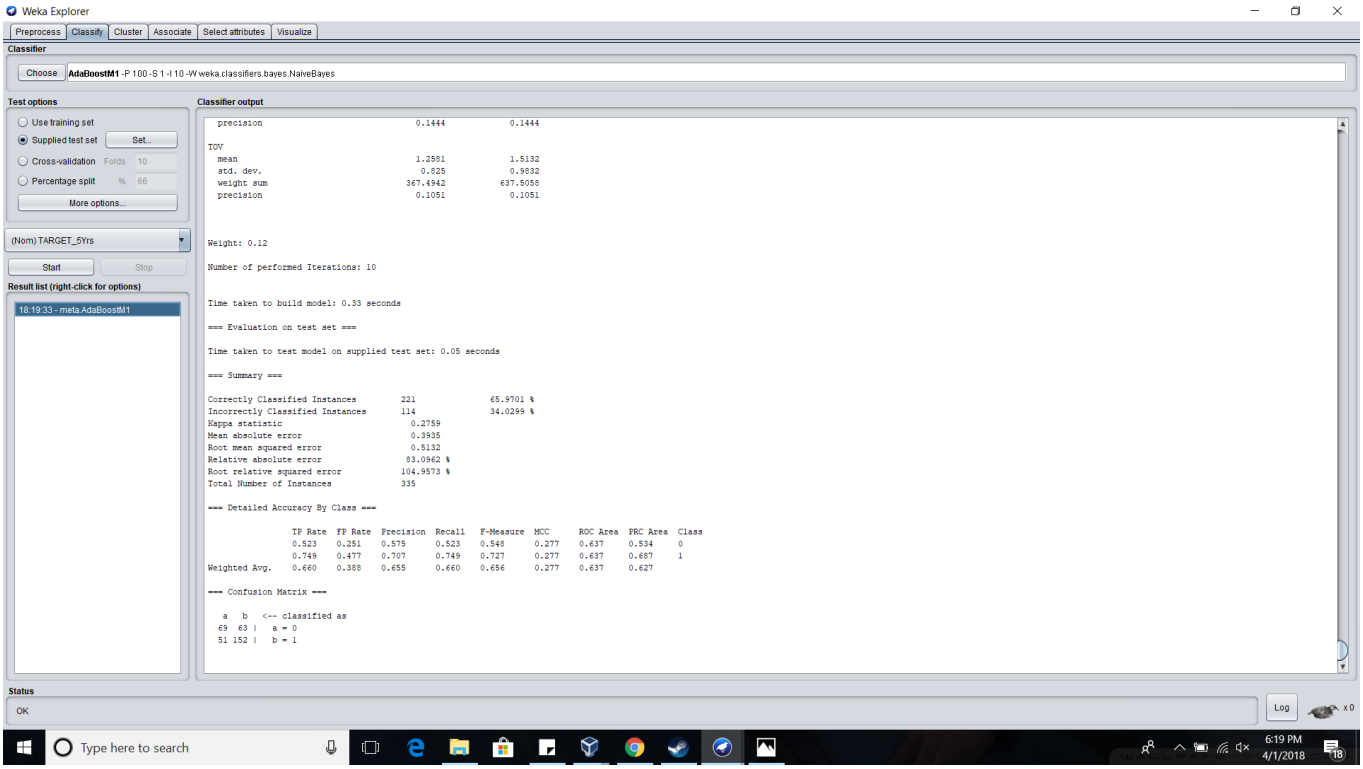
Correlation Attribute Evaluation – Bagging(Test)



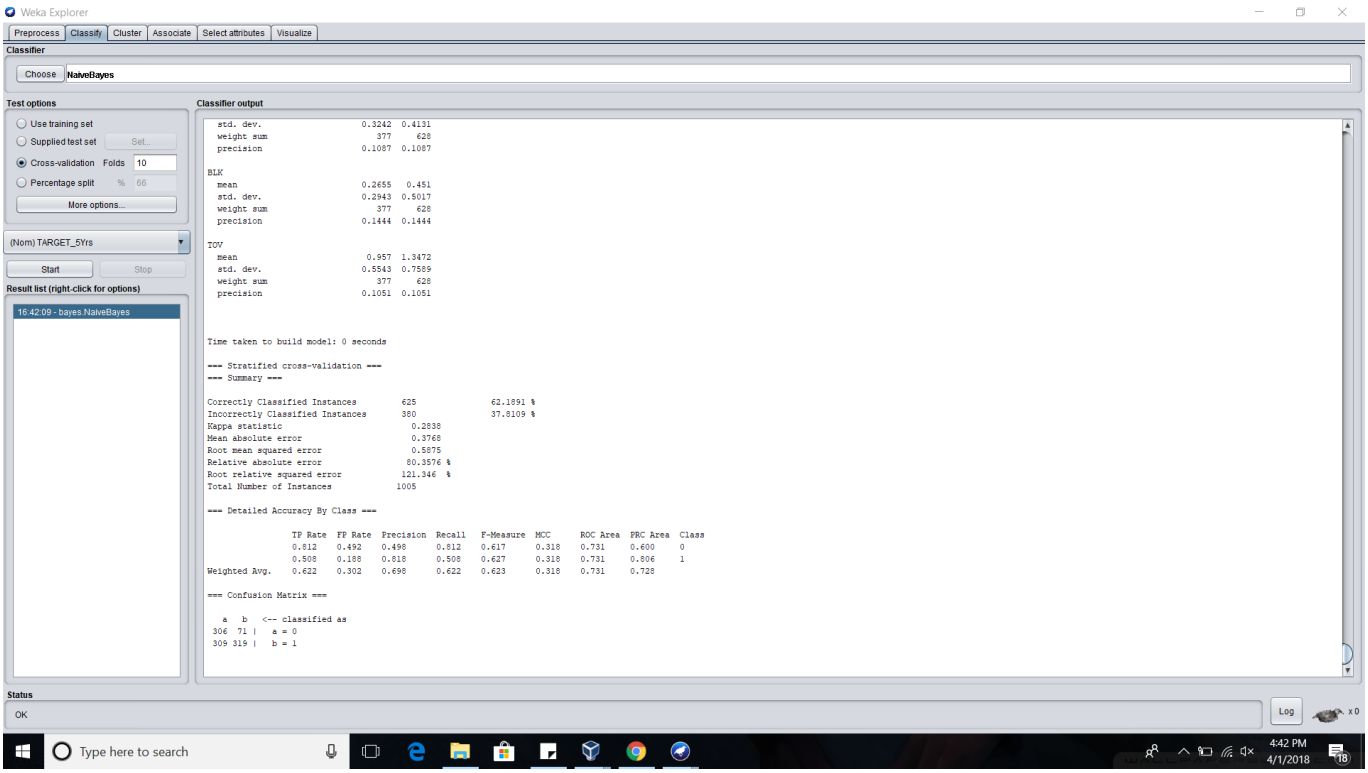
Correlation Attribute Evaluation – AdaBoost(Training)



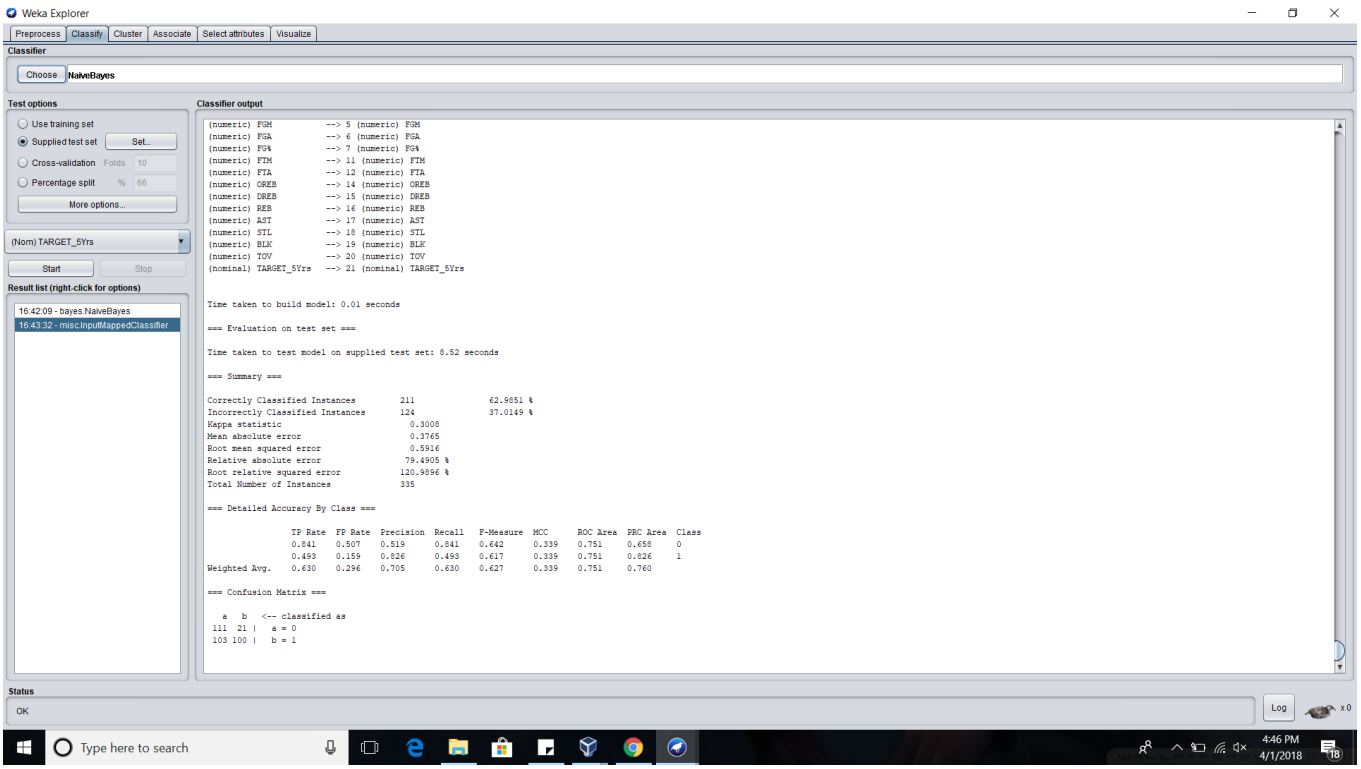
Correlation Attribute Evaluation – AdaBoost(Test)



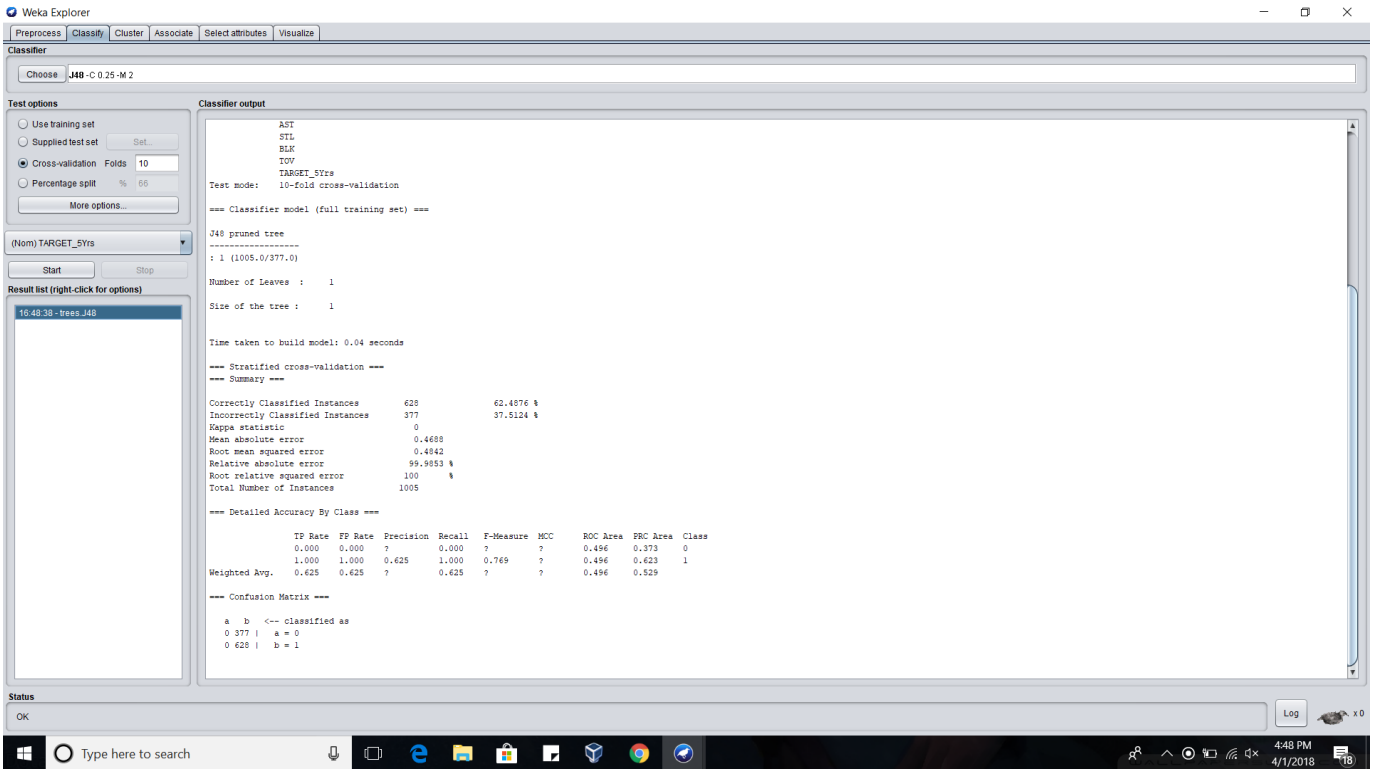
Gain Ratio – Naïve Bayes(Training)



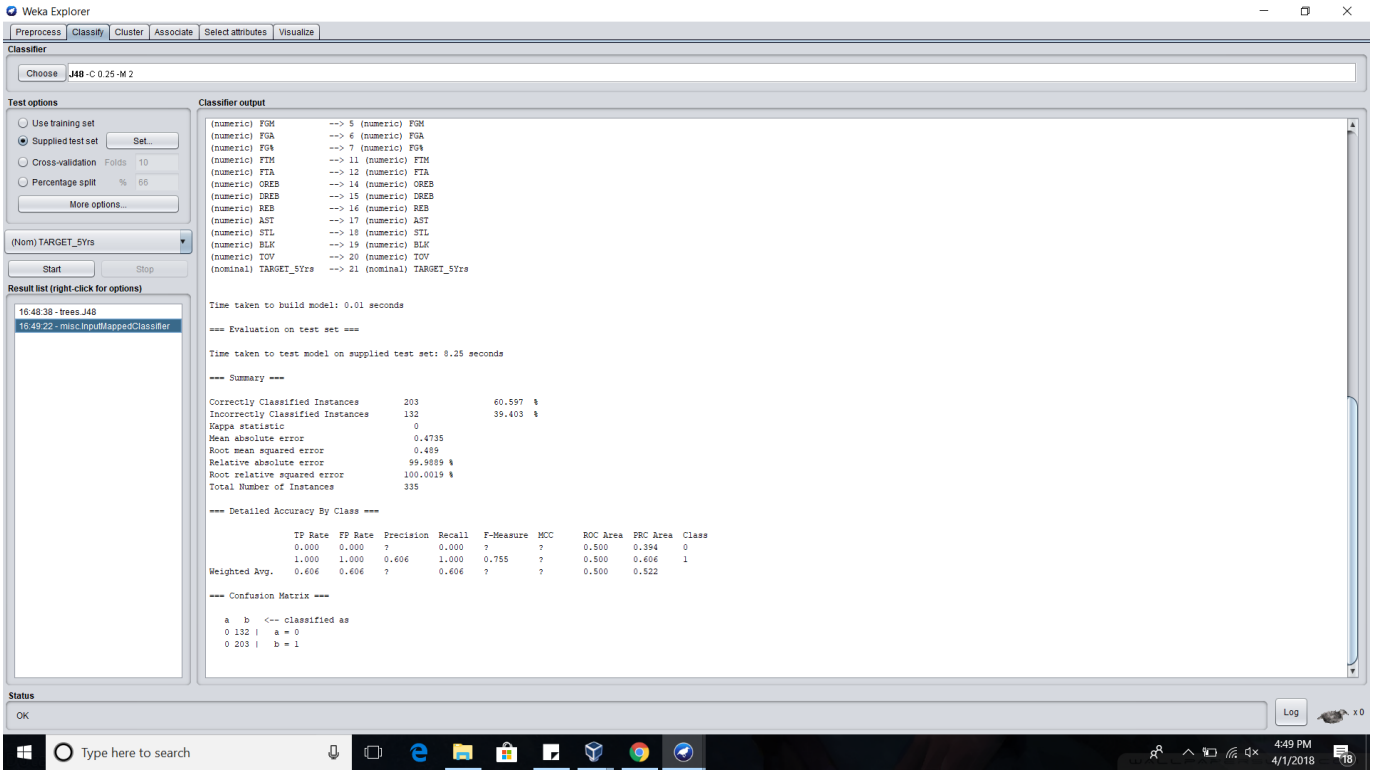
Gain Ratio – Naïve Bayes(Test)



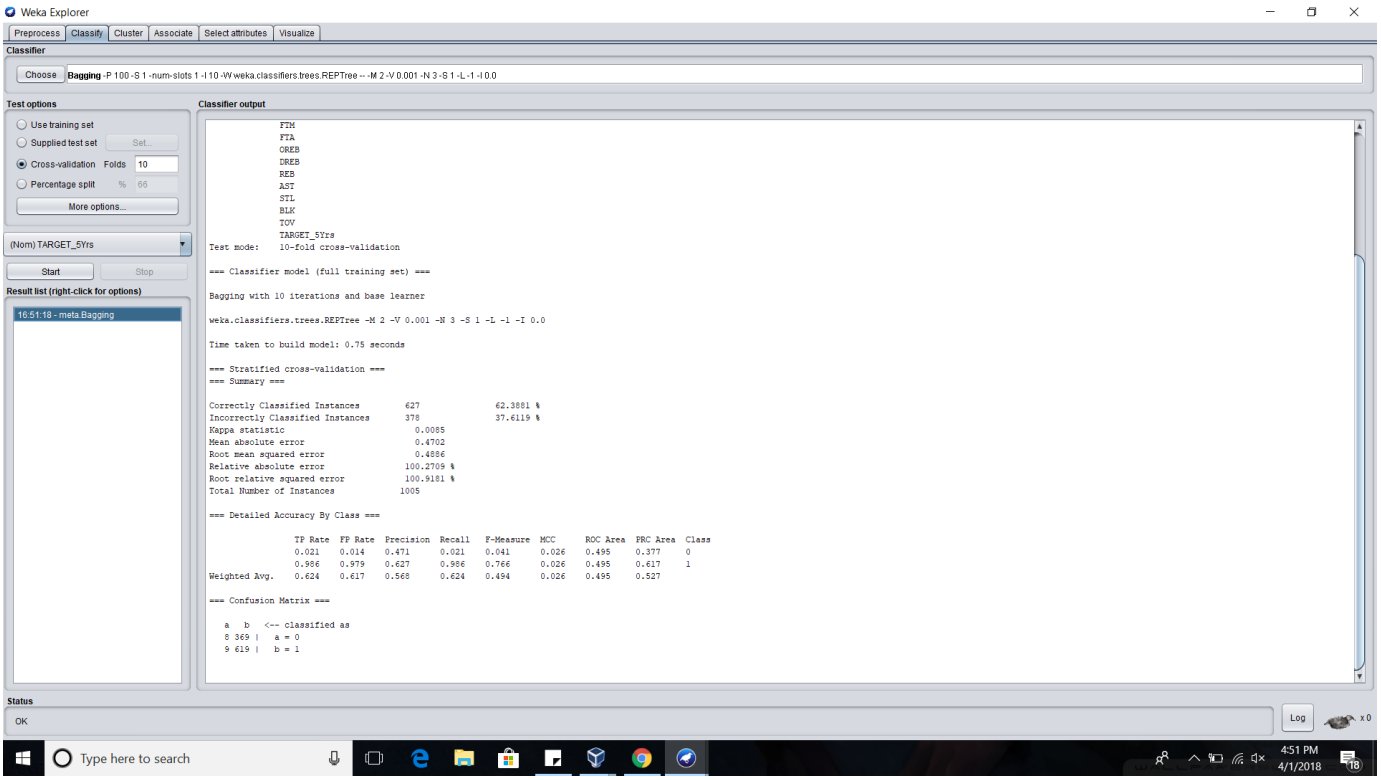
Gain Ratio – J48(Training)



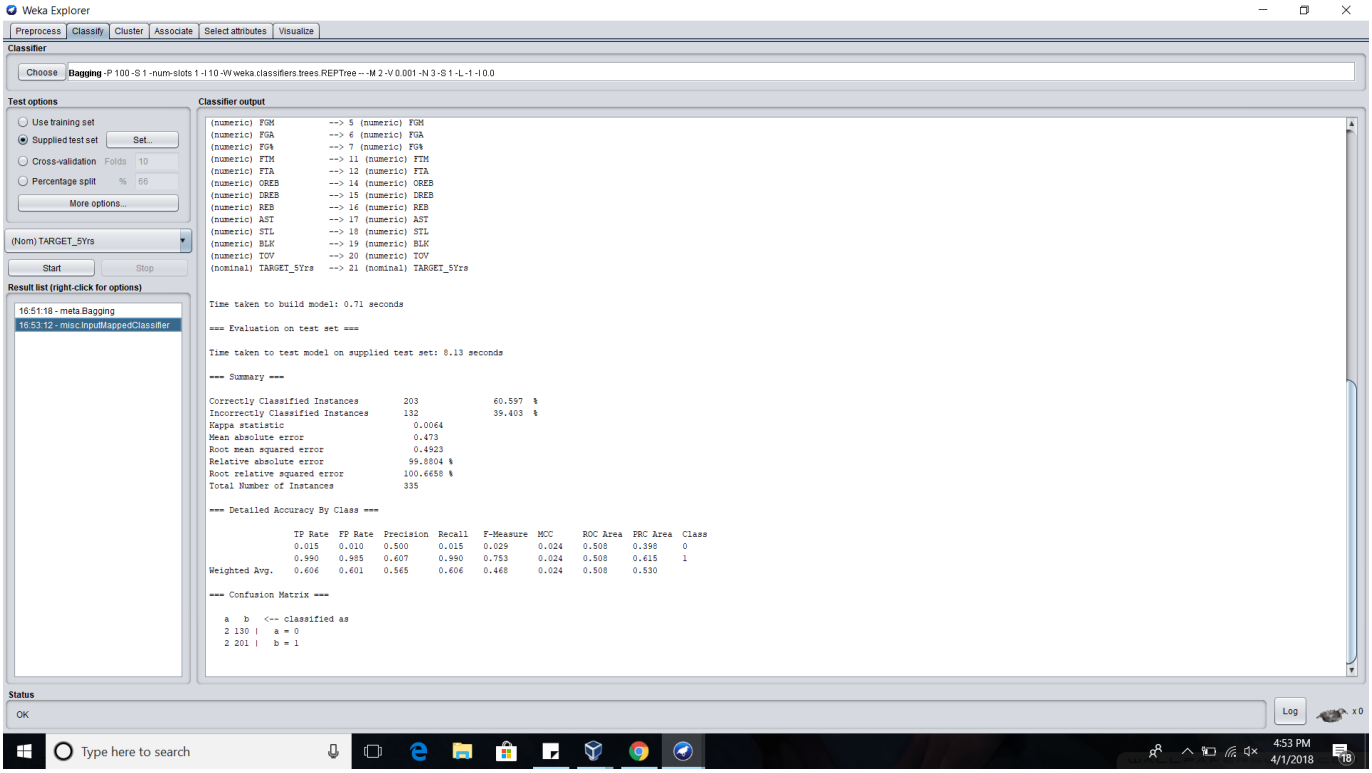
Gain Ratio – J48(Test)



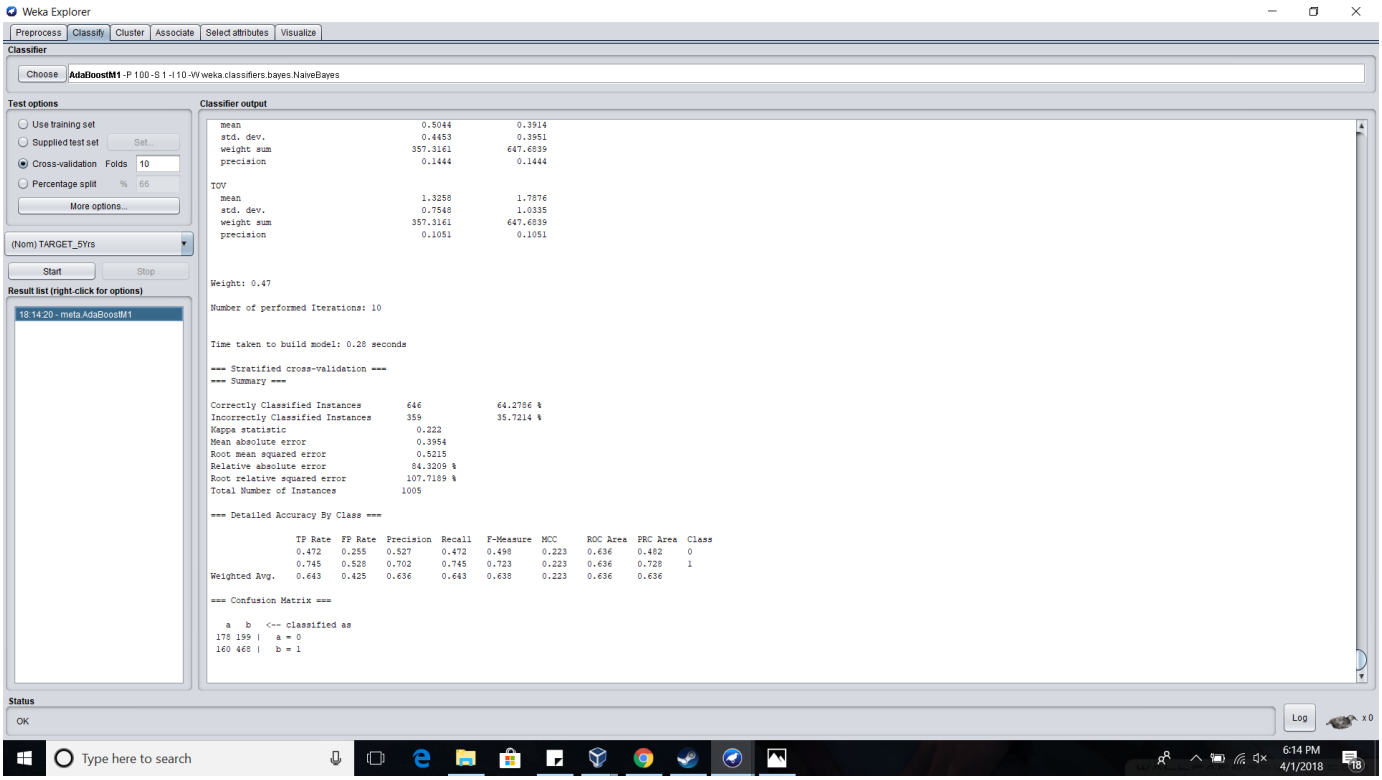
Gain Ratio – Bagging(Training)



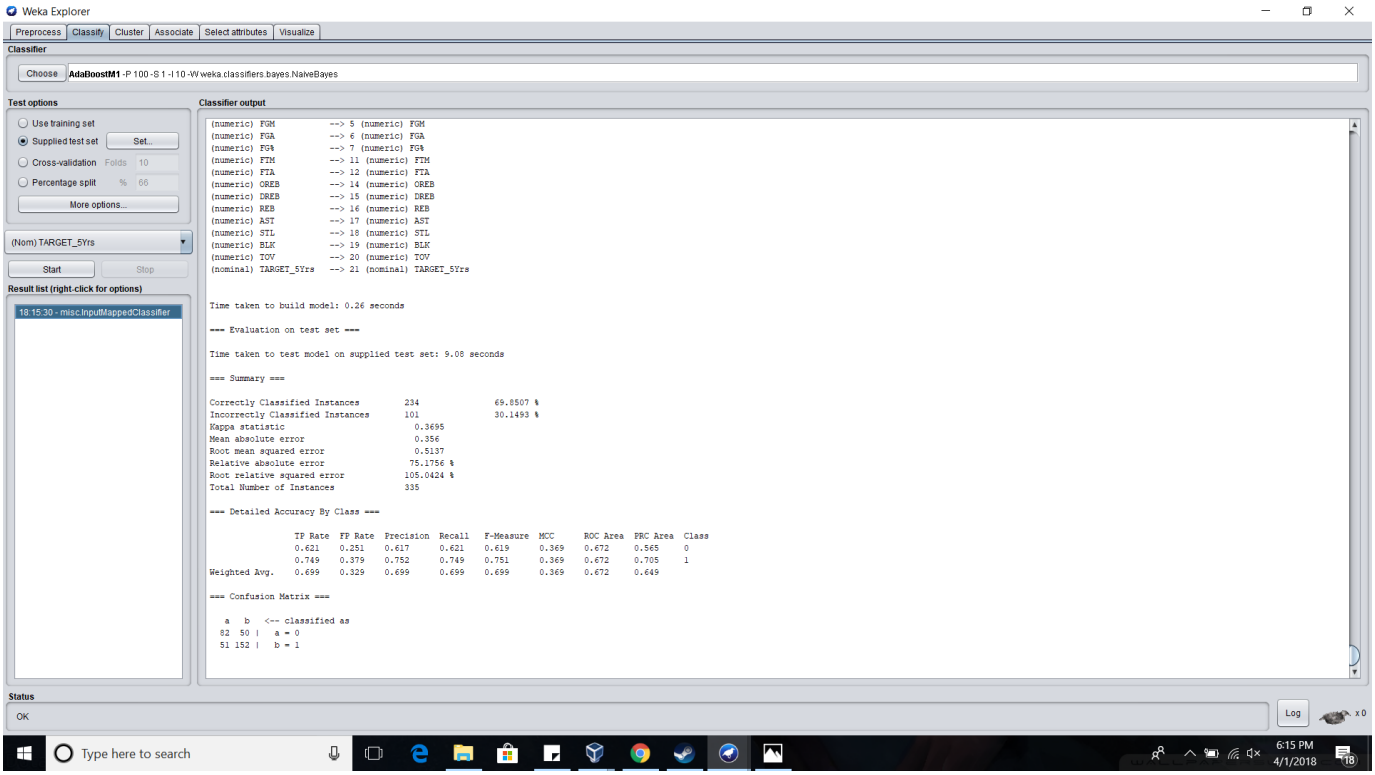
Gain Ratio – Bagging(Test)



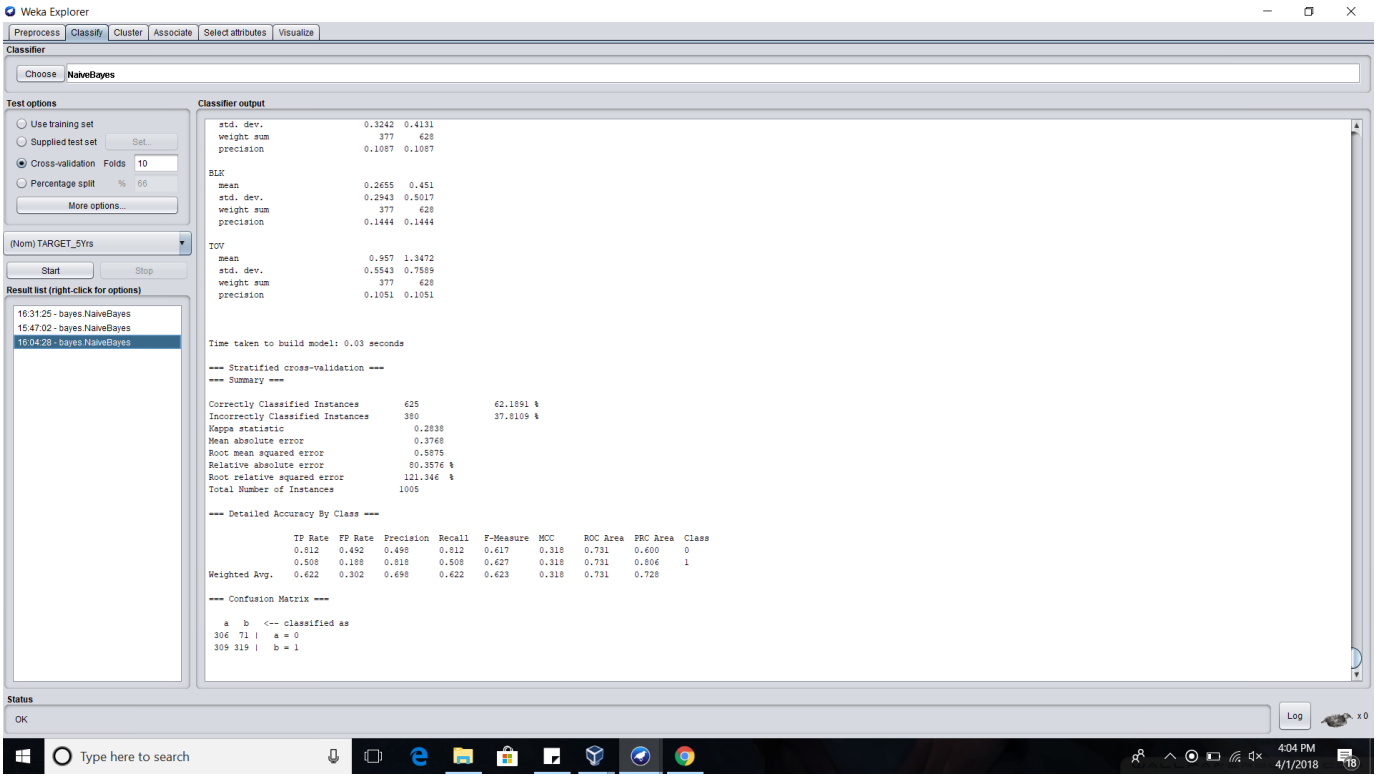
Gain Ratio – AdaBoost(Training)



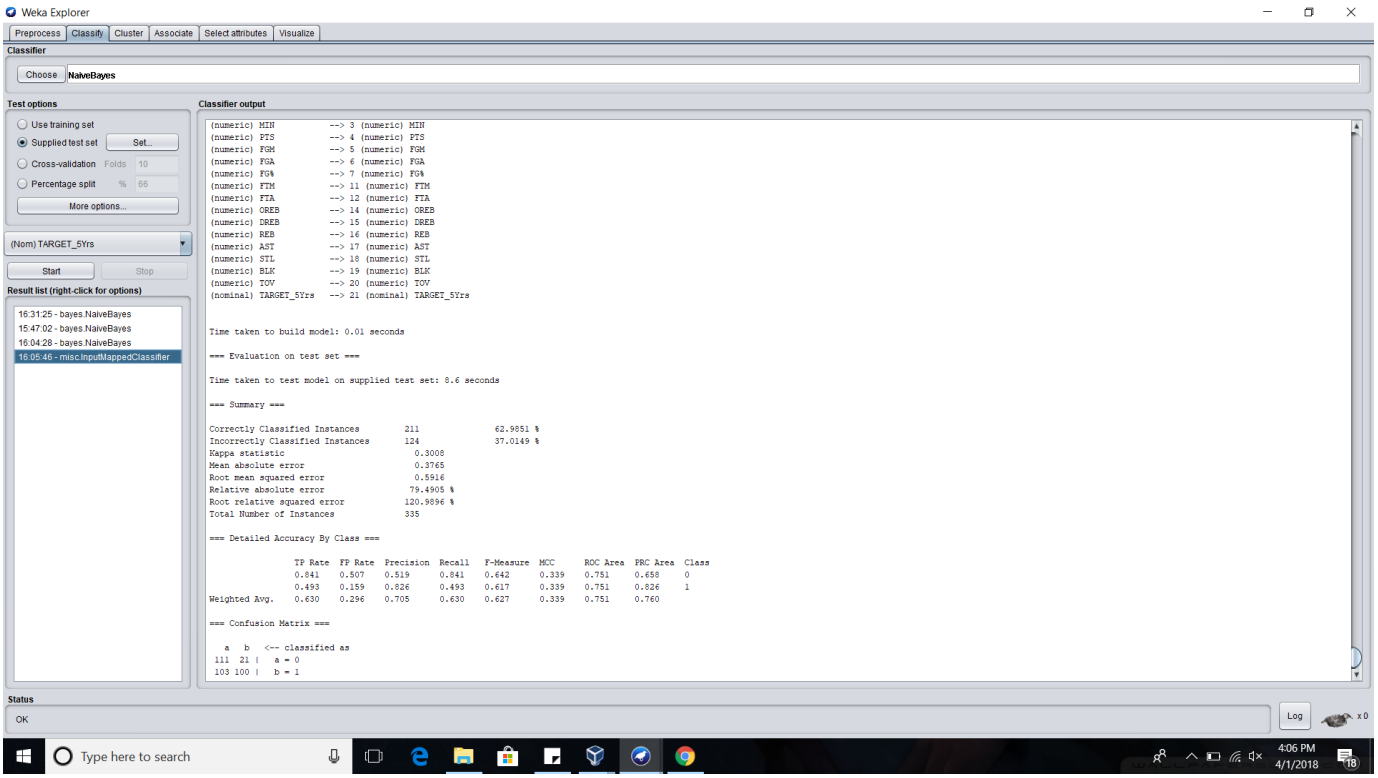
Gain Ratio – AdaBoost(Test)



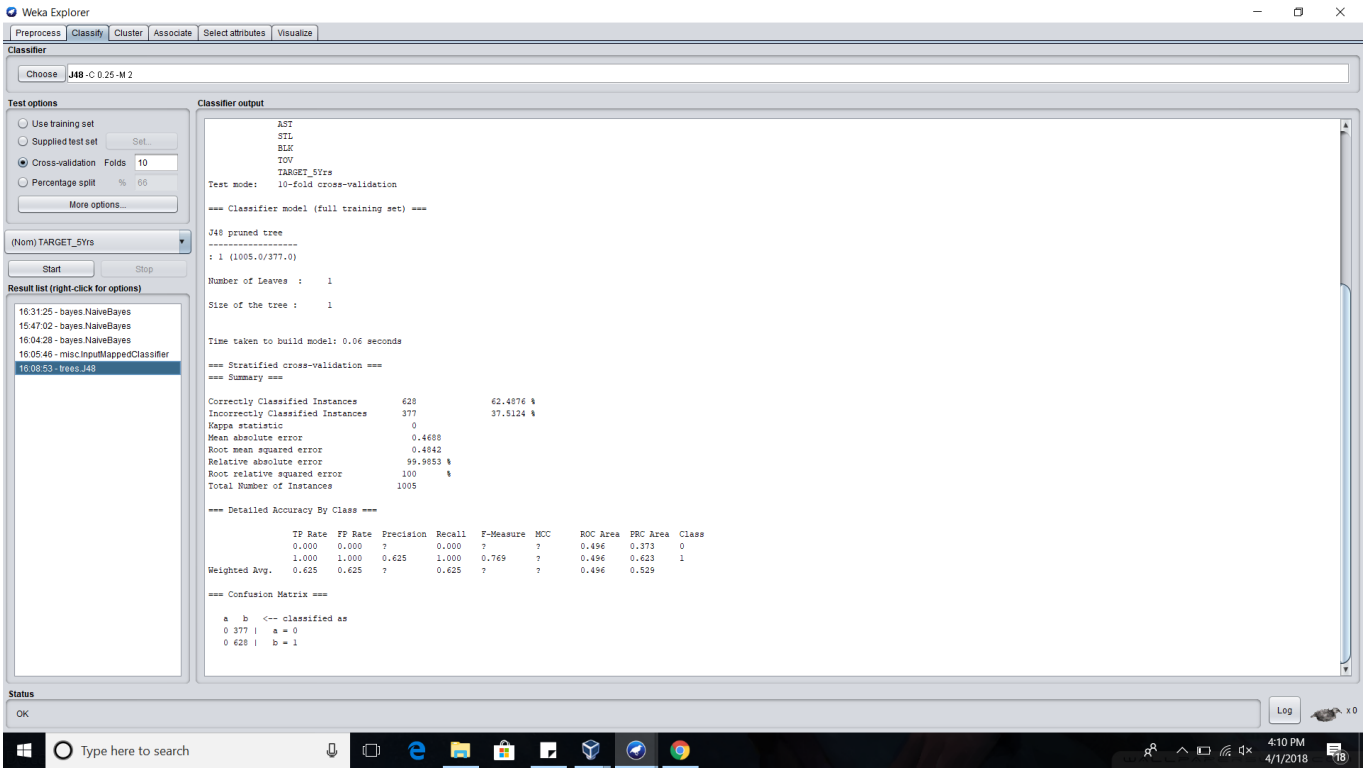
Info Gain – Naïve Bayes(Training)



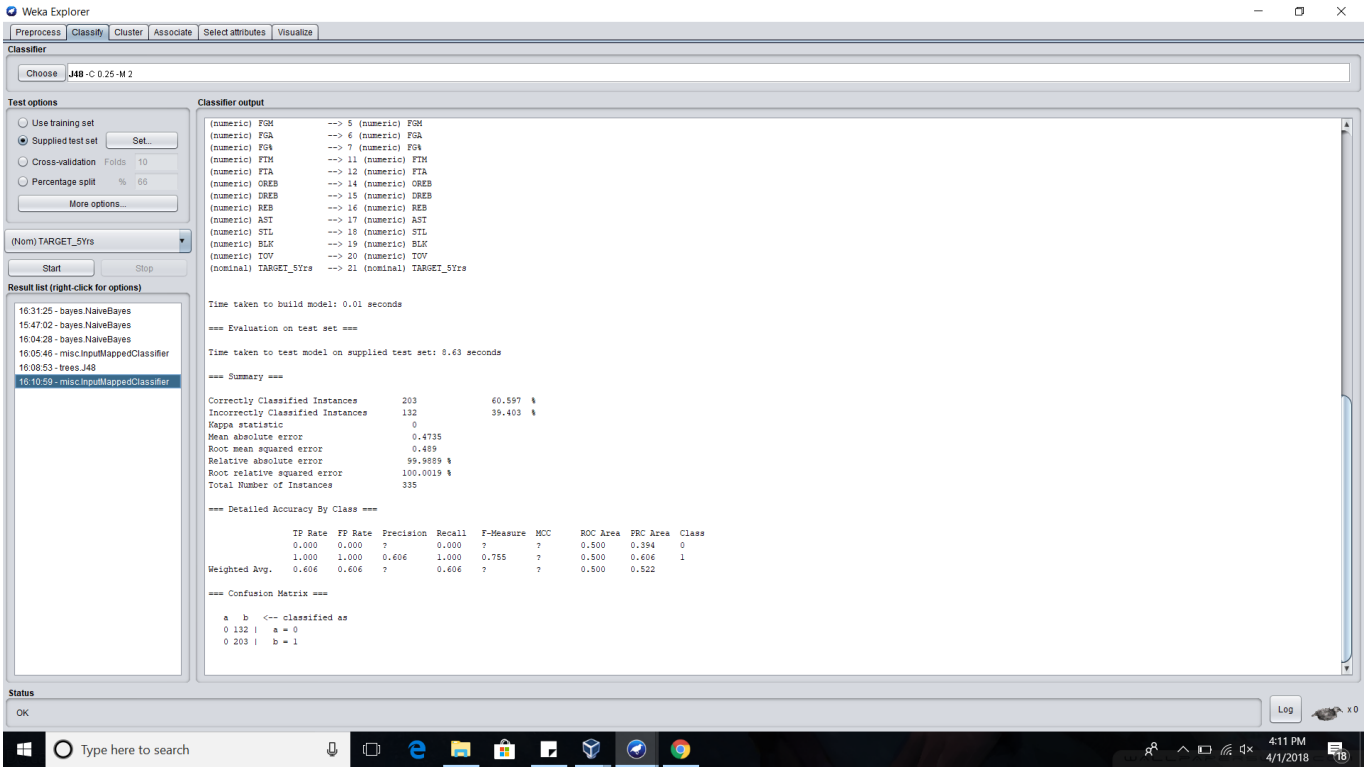
Info Gain – Naïve Bayes(Test)



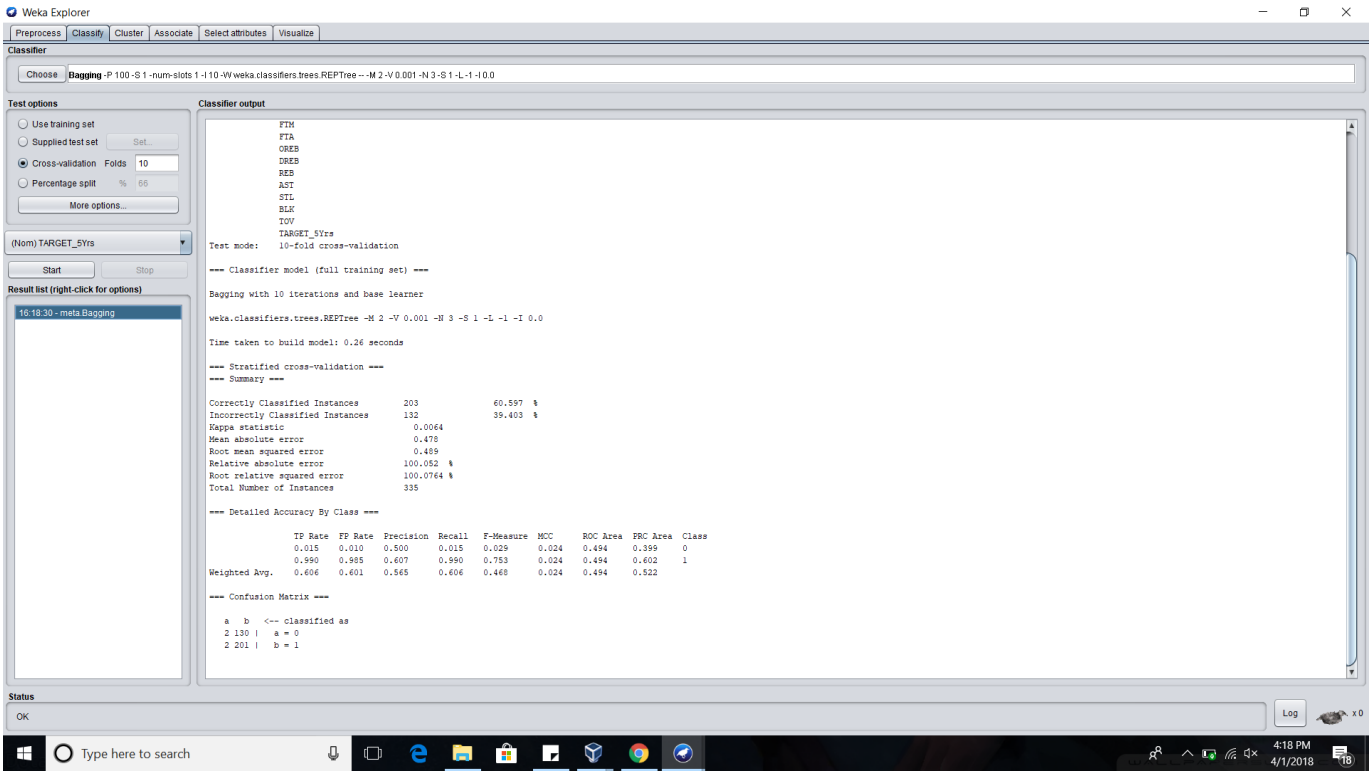
Info Gain – J48(Training)



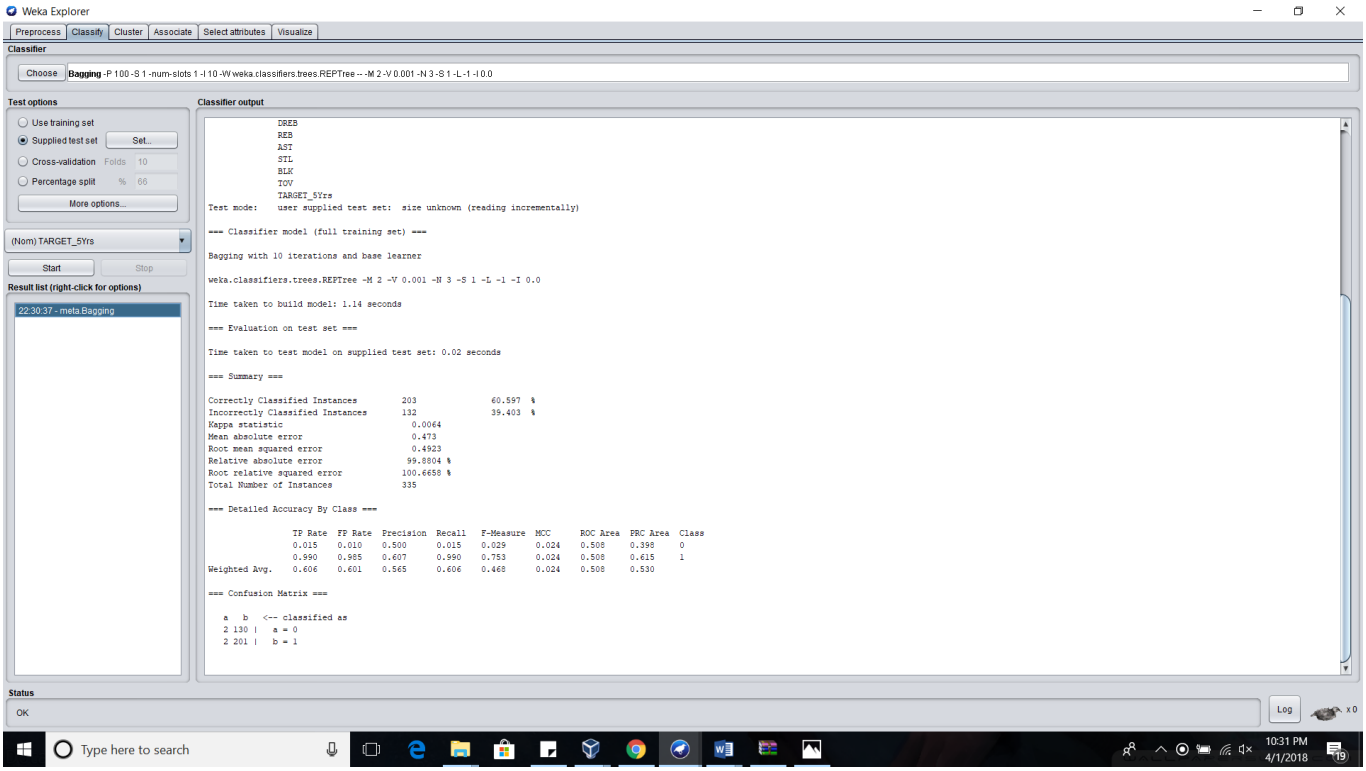
Info Gain – J48(test)



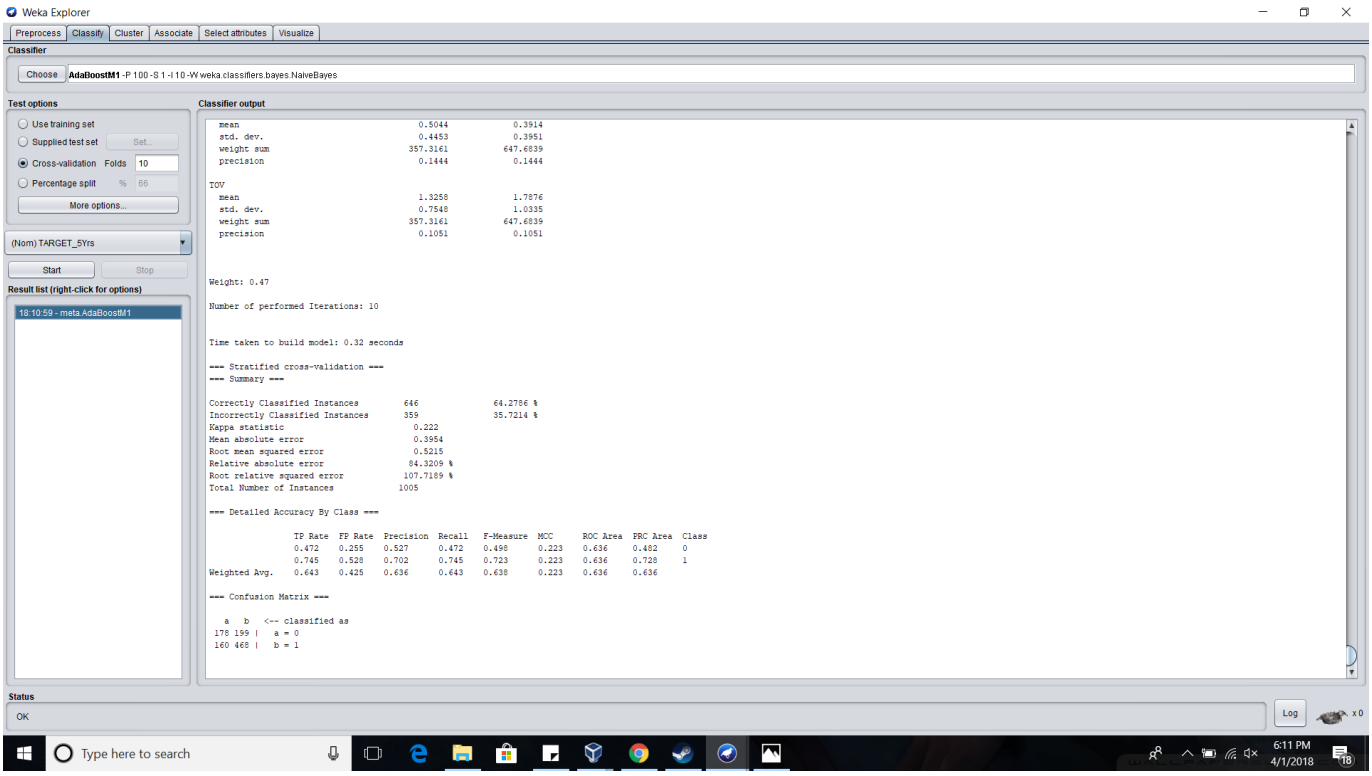
Info Gain – Bagging(Training)



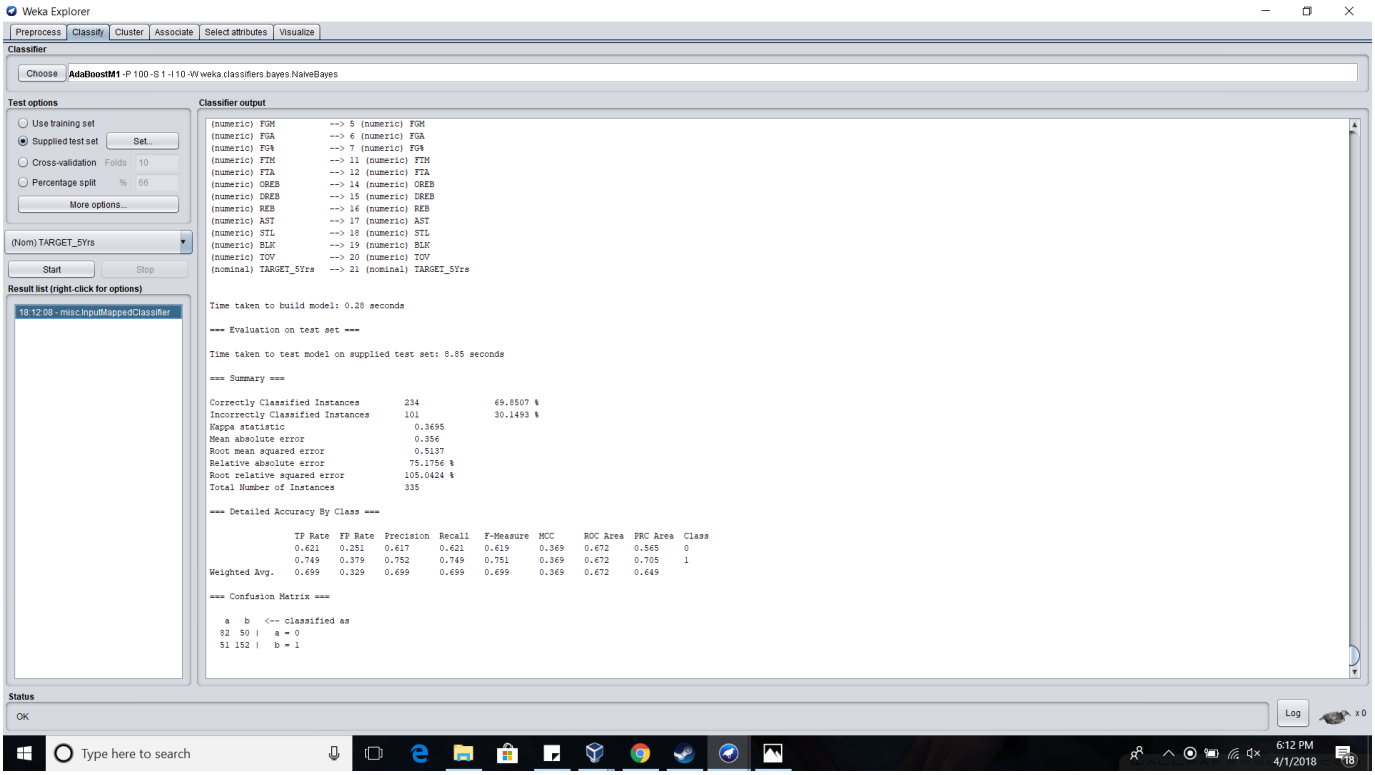
Info Gain – Bagging(test)



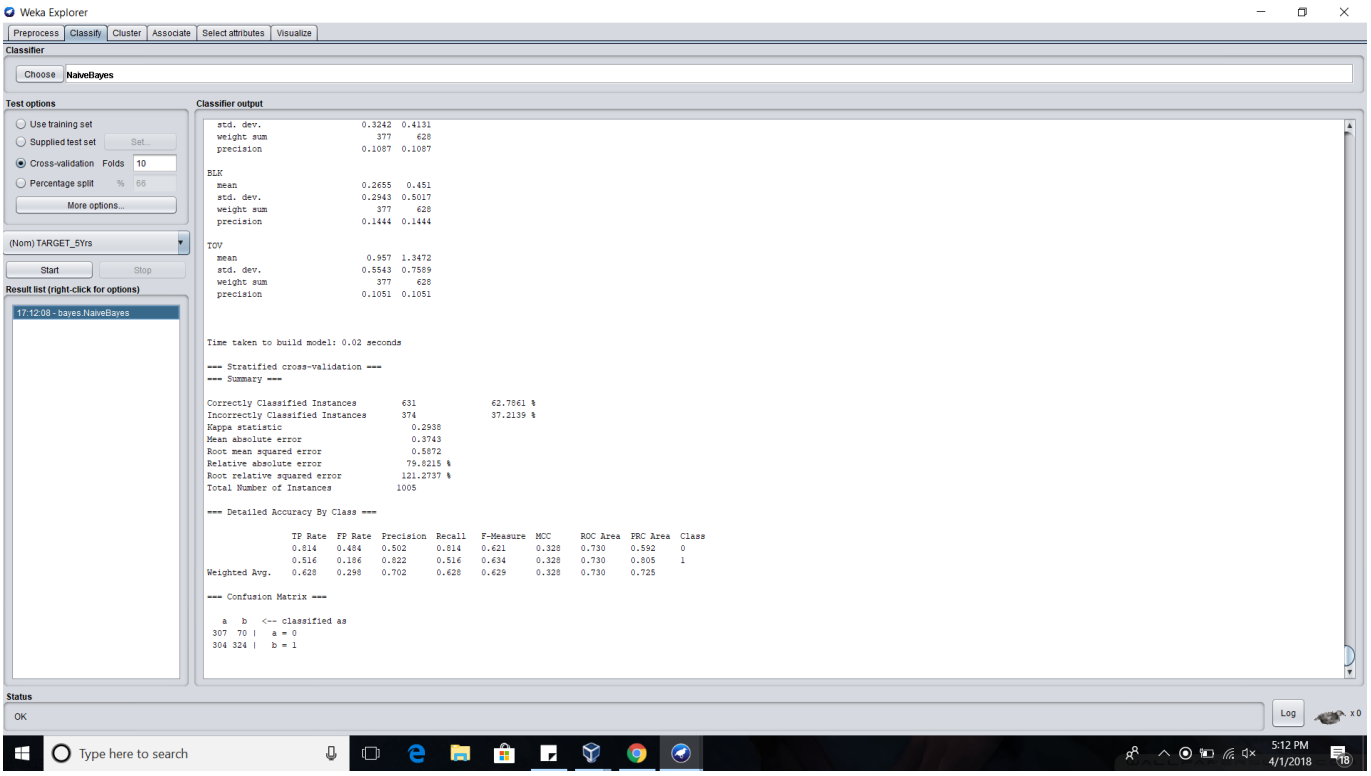
Info Gain – AdaBoost(Training)



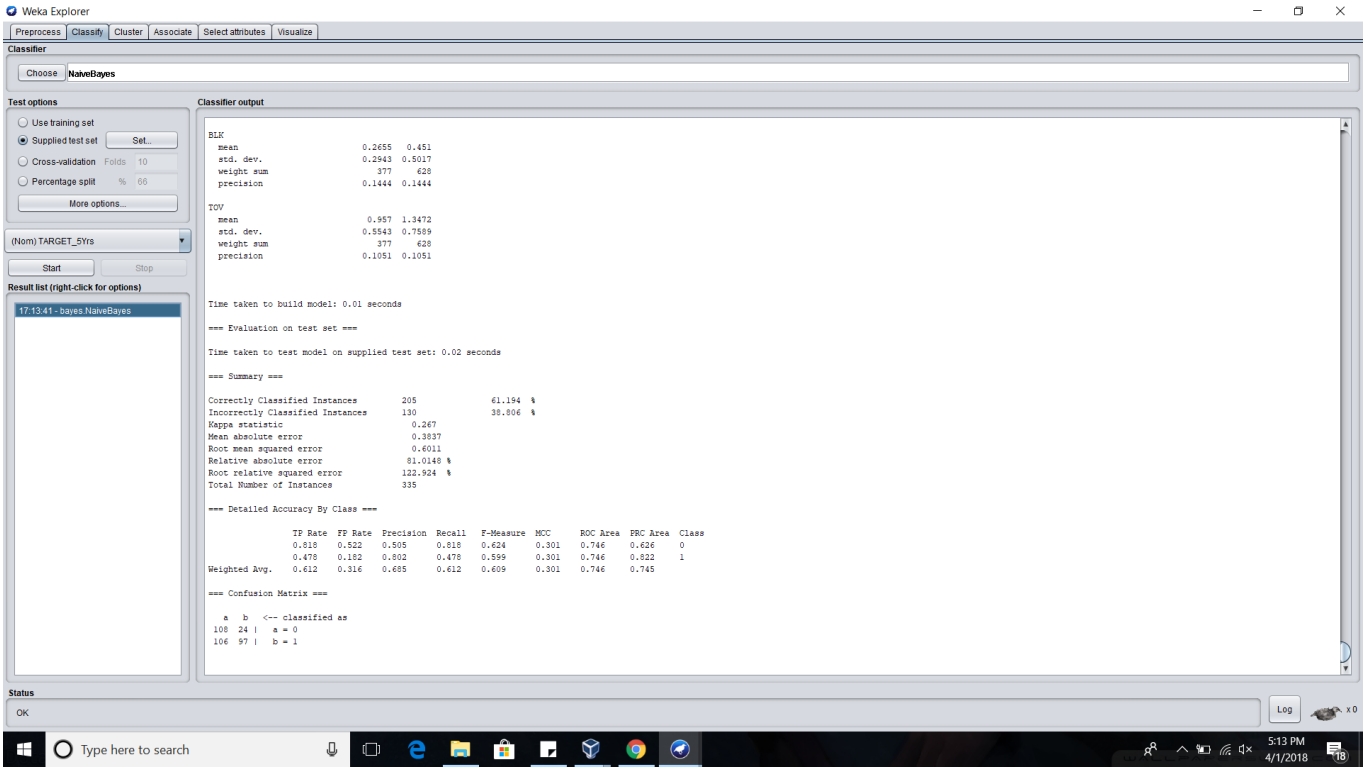
Info Gain – AdaBoost(Test)



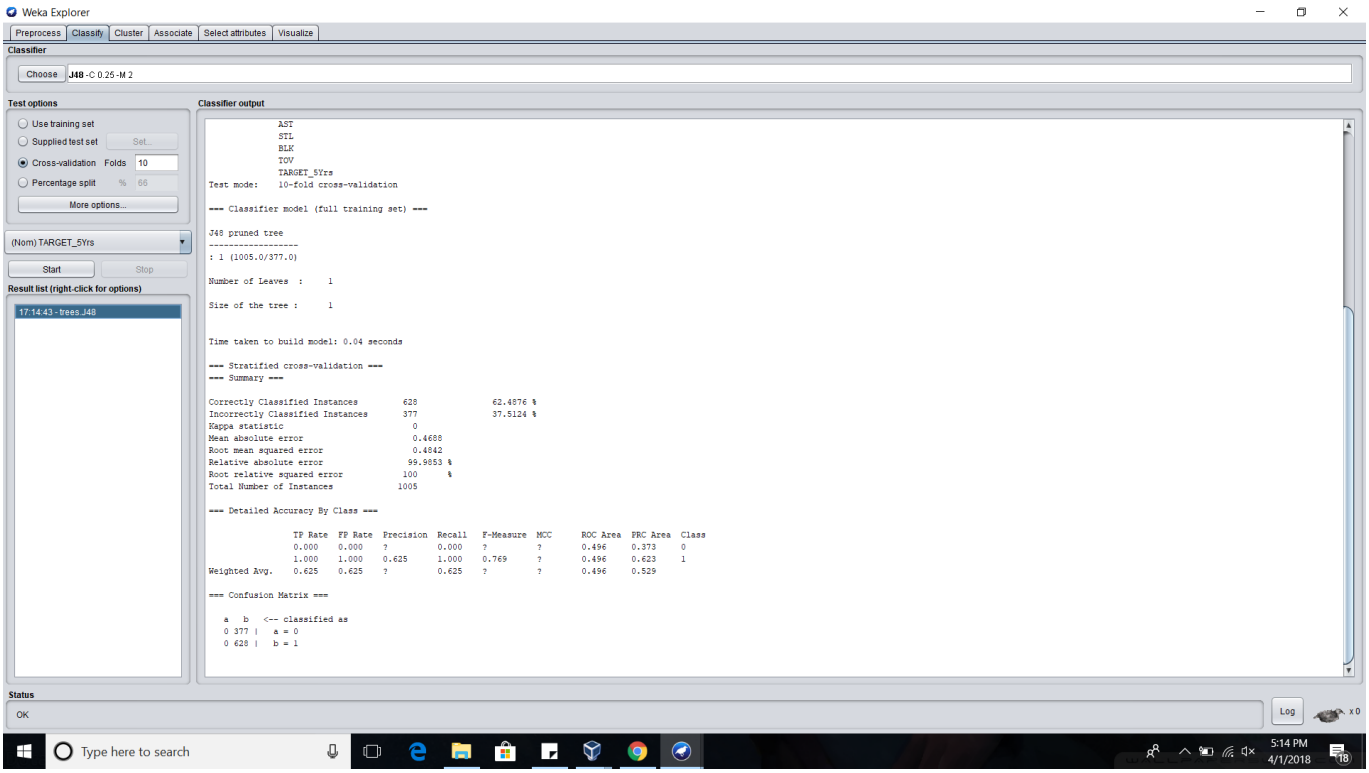
OneR – Naïve Bayes(Training)



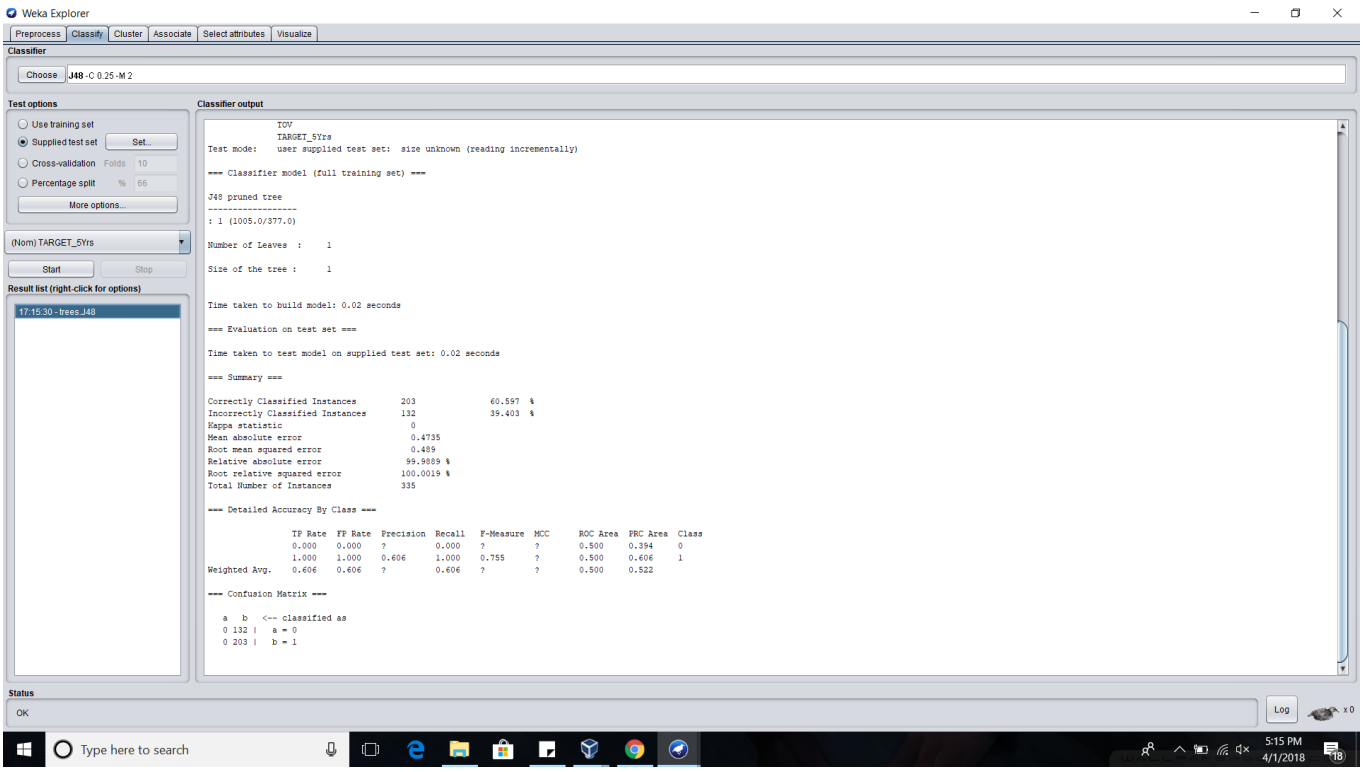
OneR – Naïve Bayes(Test)



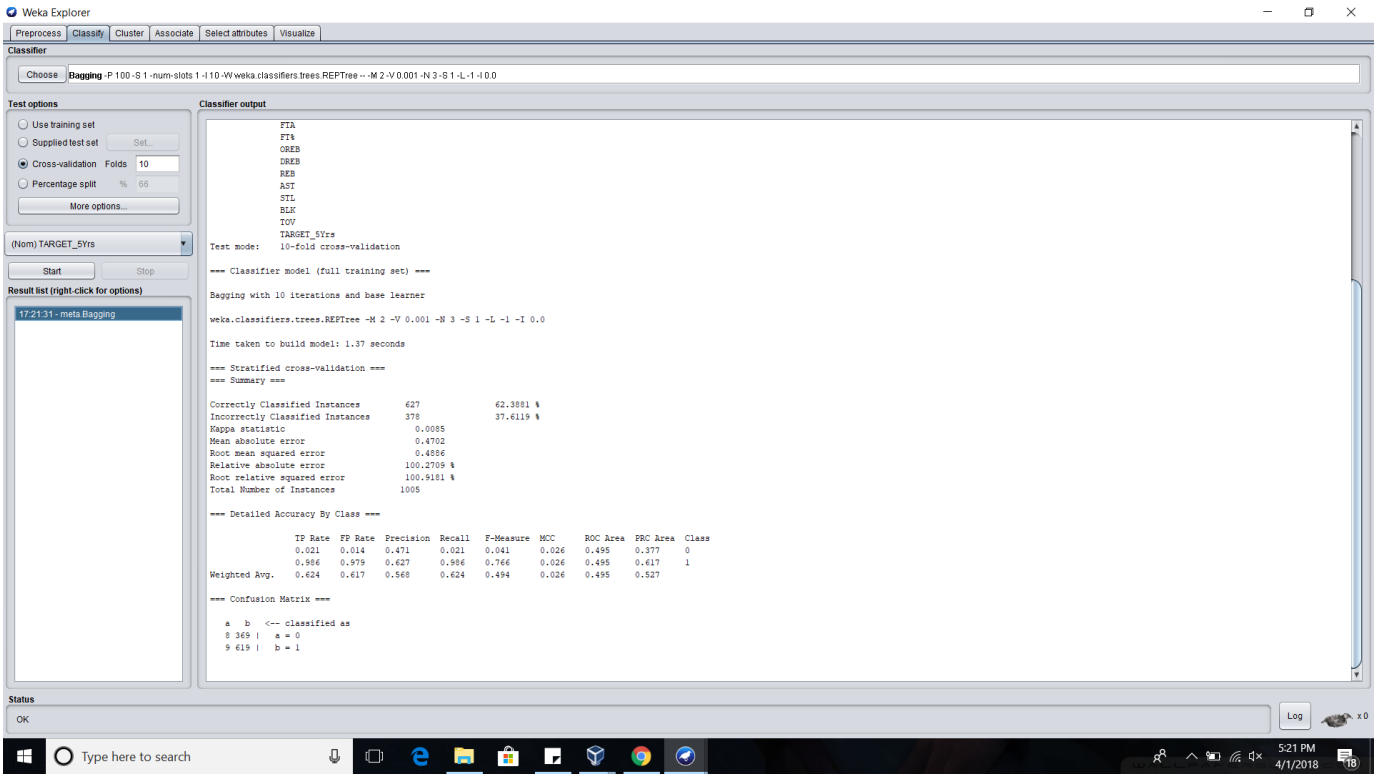
OneR – J48(Training)



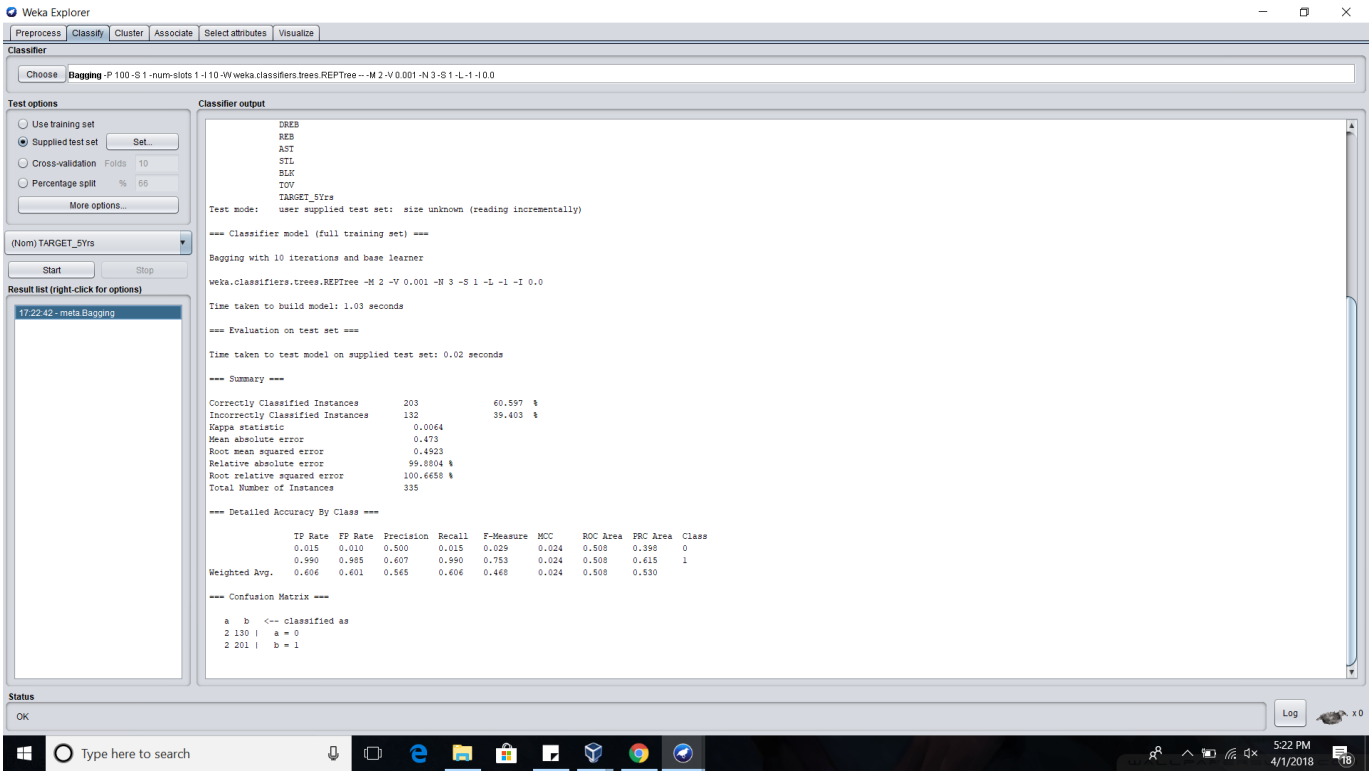
OneR – J48(Test)



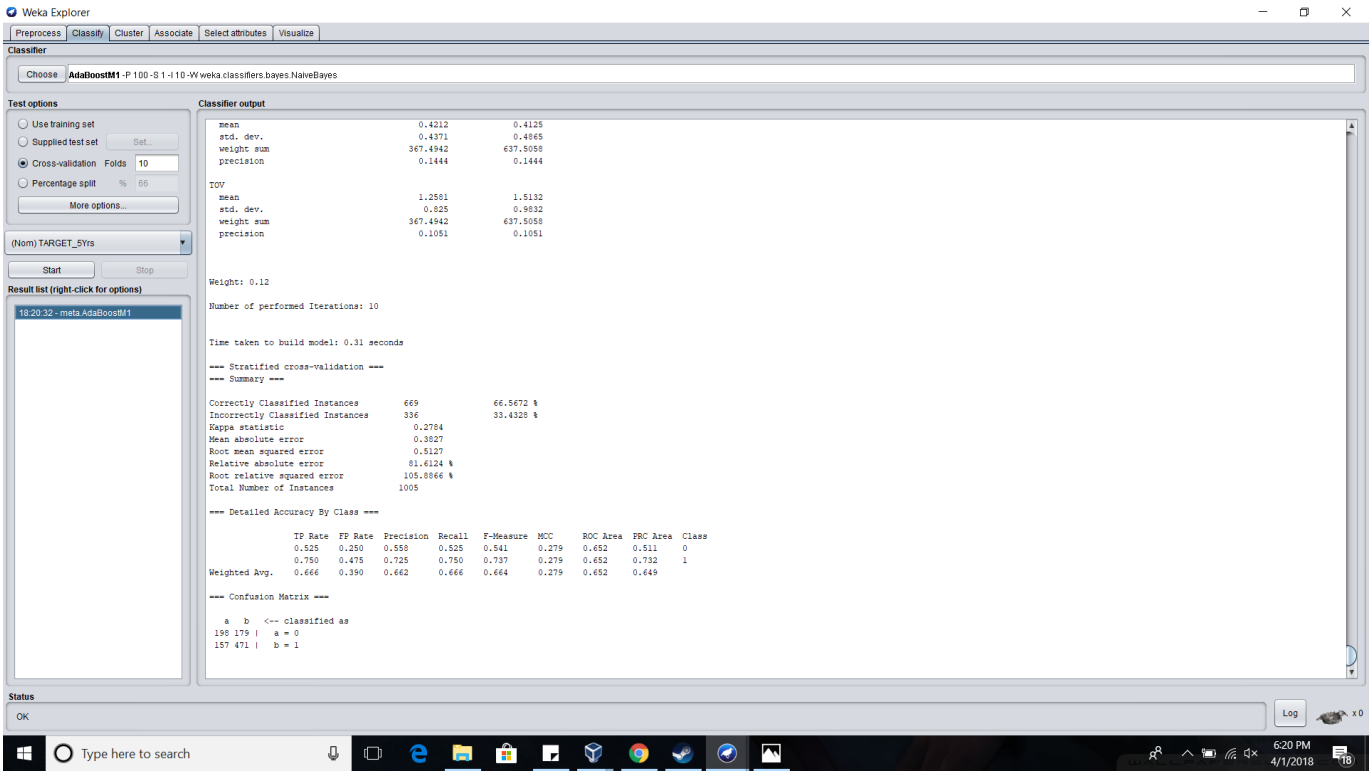
OneR – Bagging(Training)



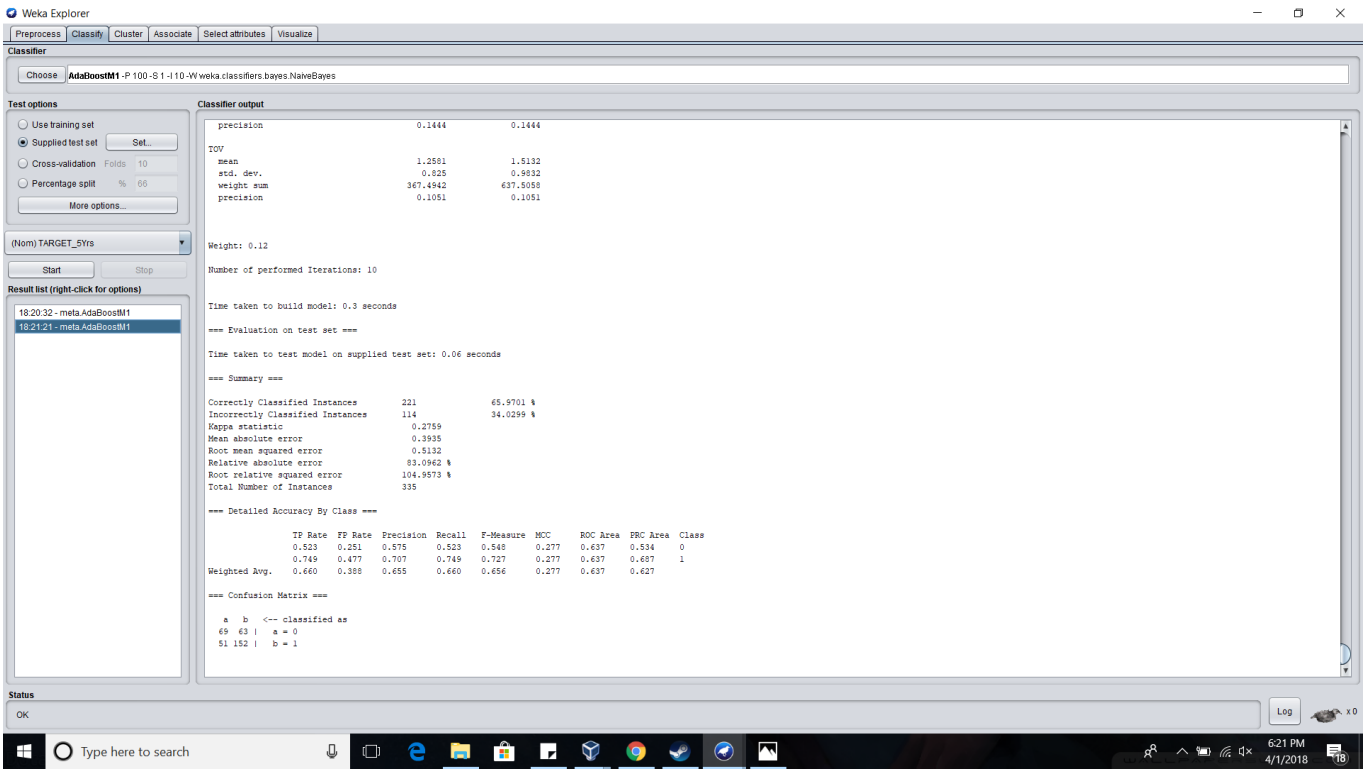
OneR – Bagging(Test)



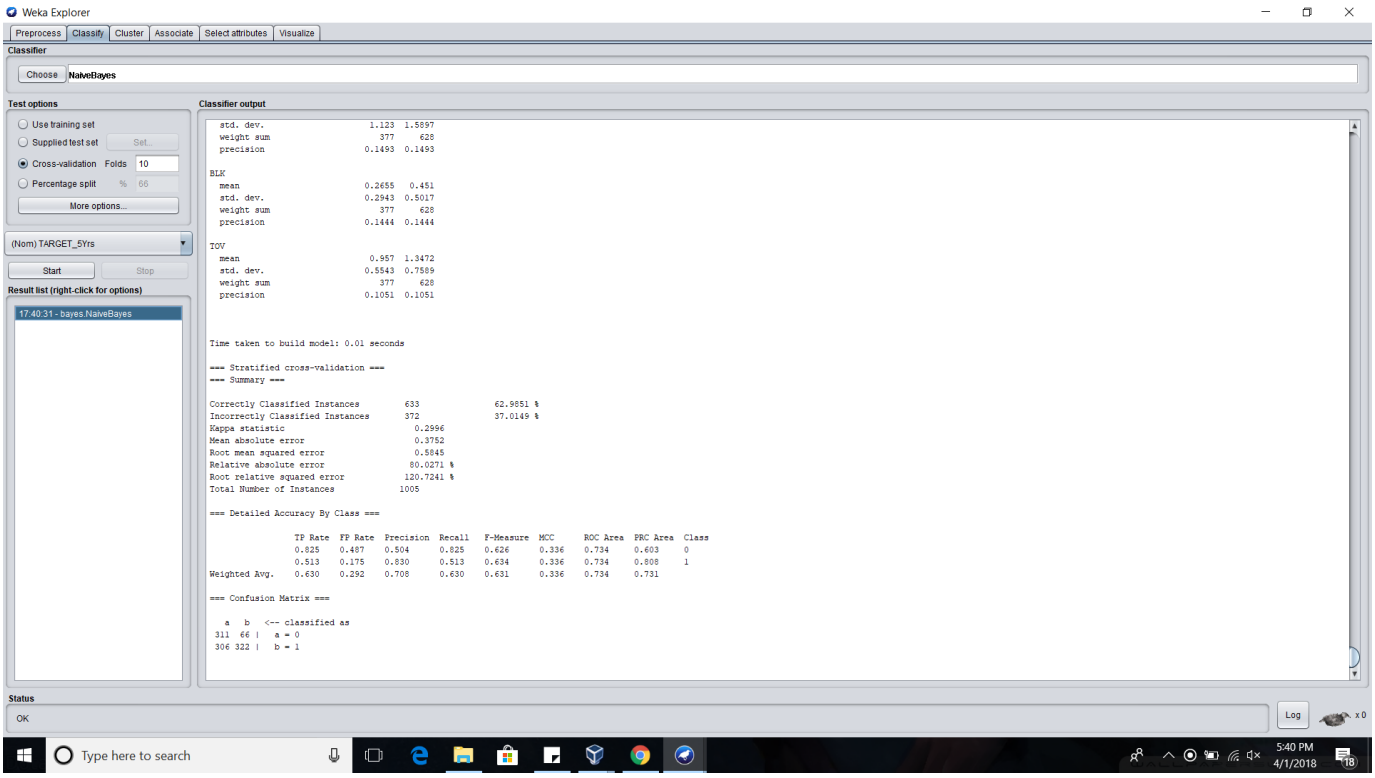
OneR – AdaBoost(Training)



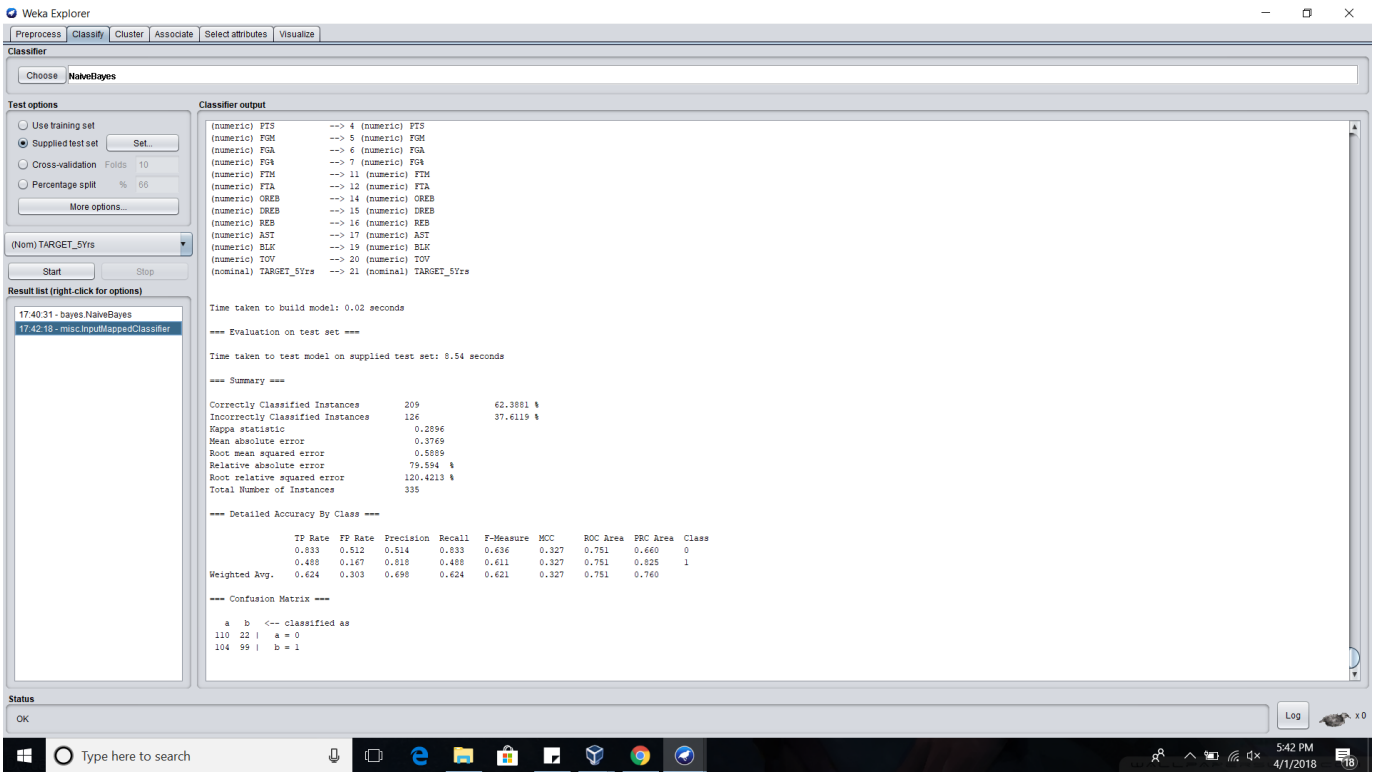
OneR – AdaBoost(Test)



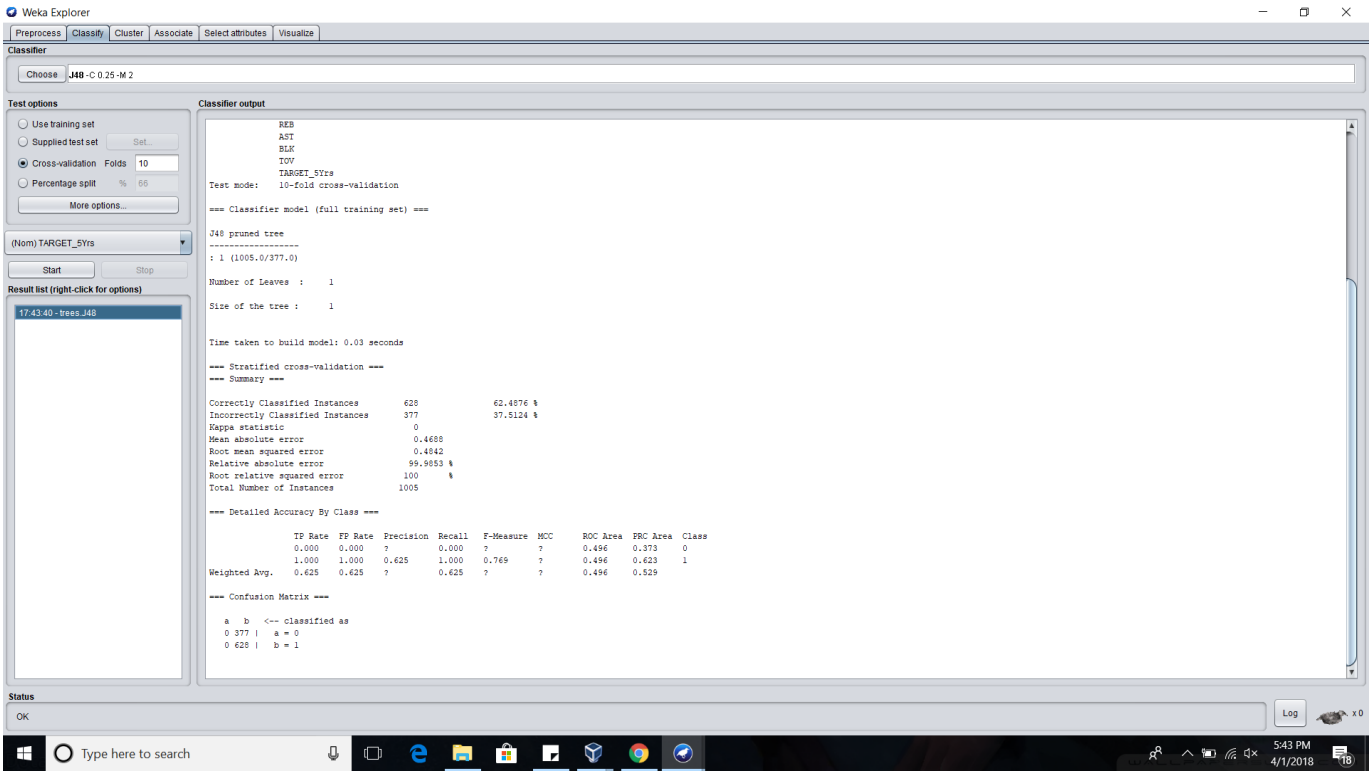
Manual – Naïve Bayes(Training)



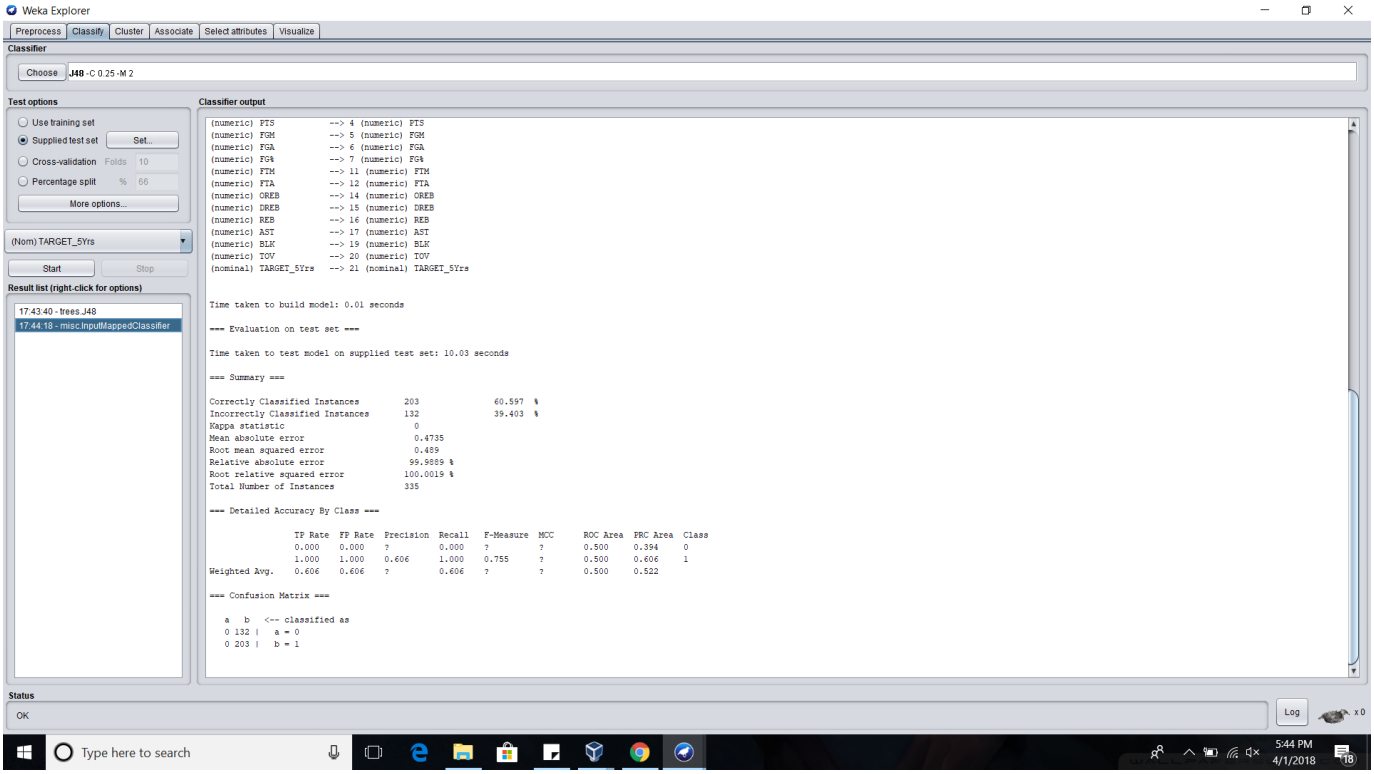
Manual – Naïve Bayes(Test)



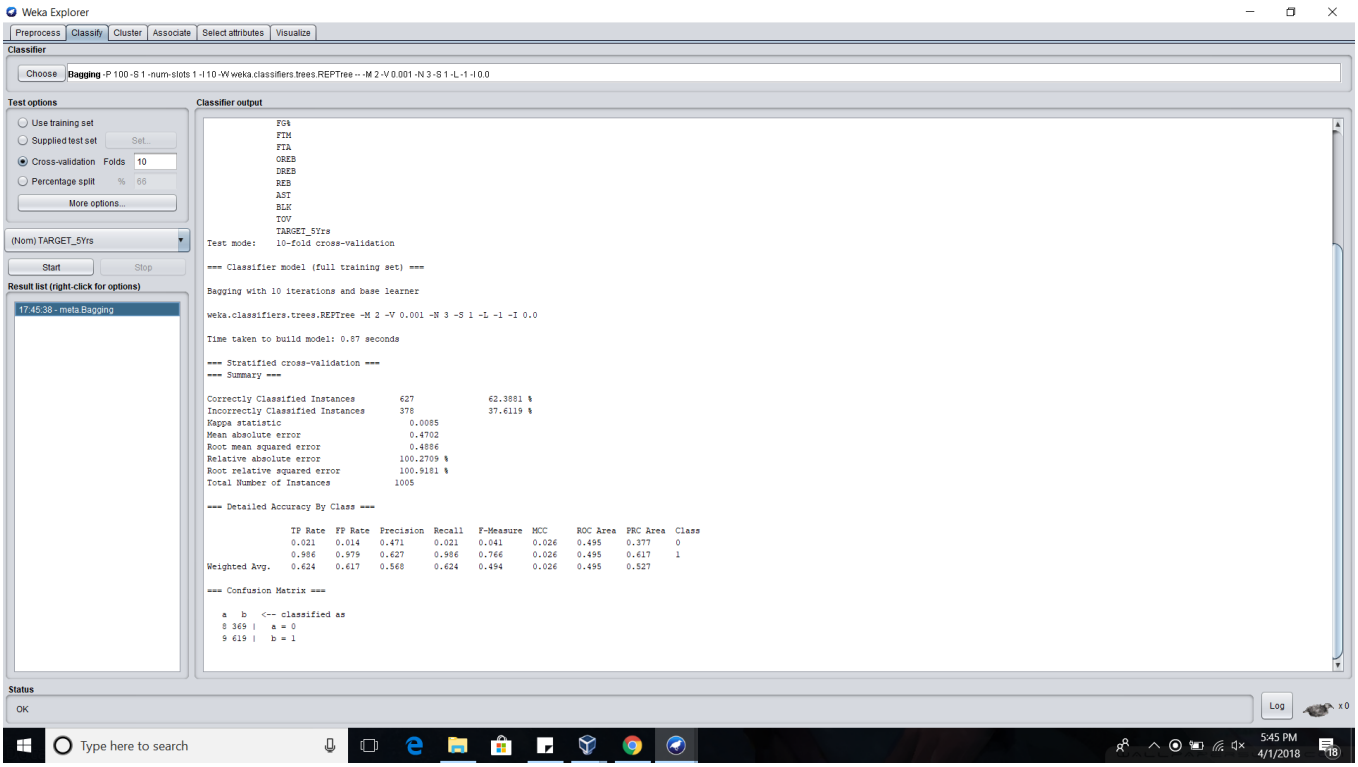
Manual – J48(Training)



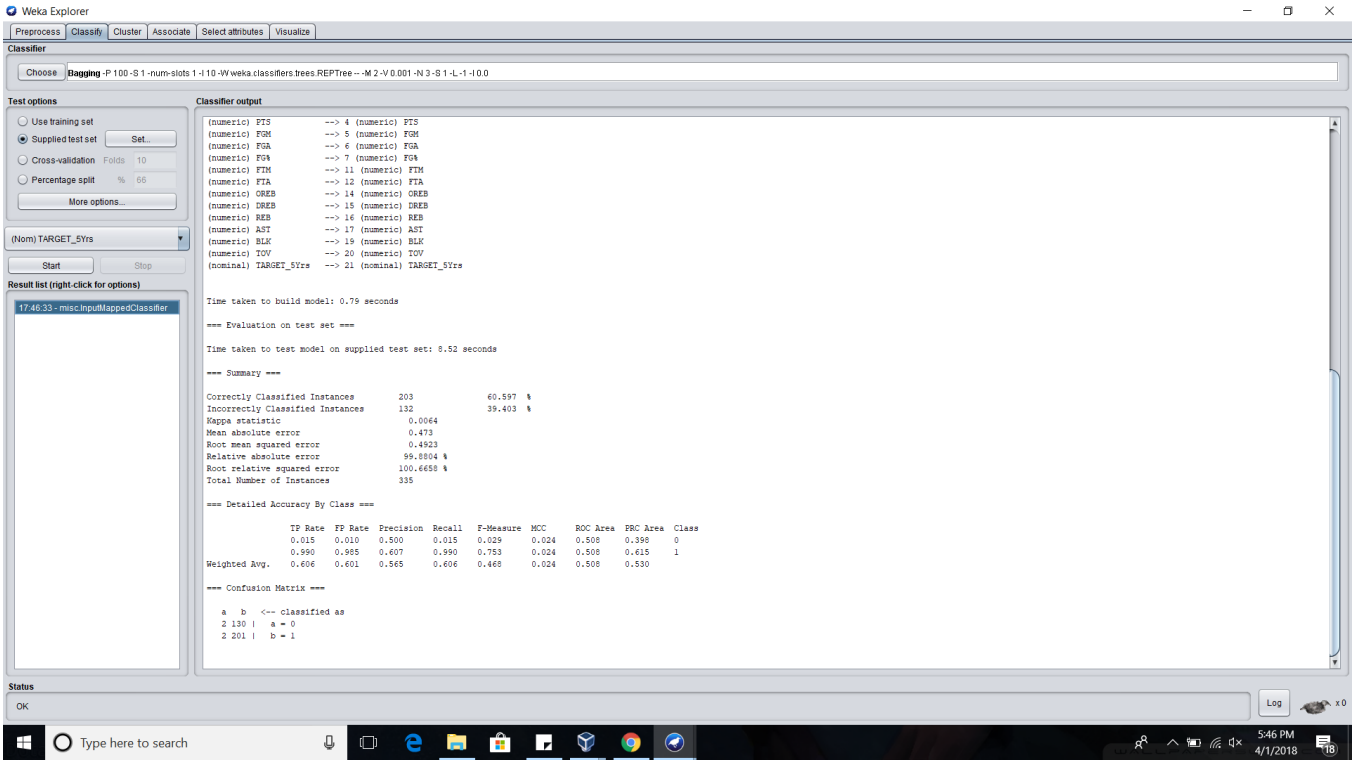
Manual – J48(Test)



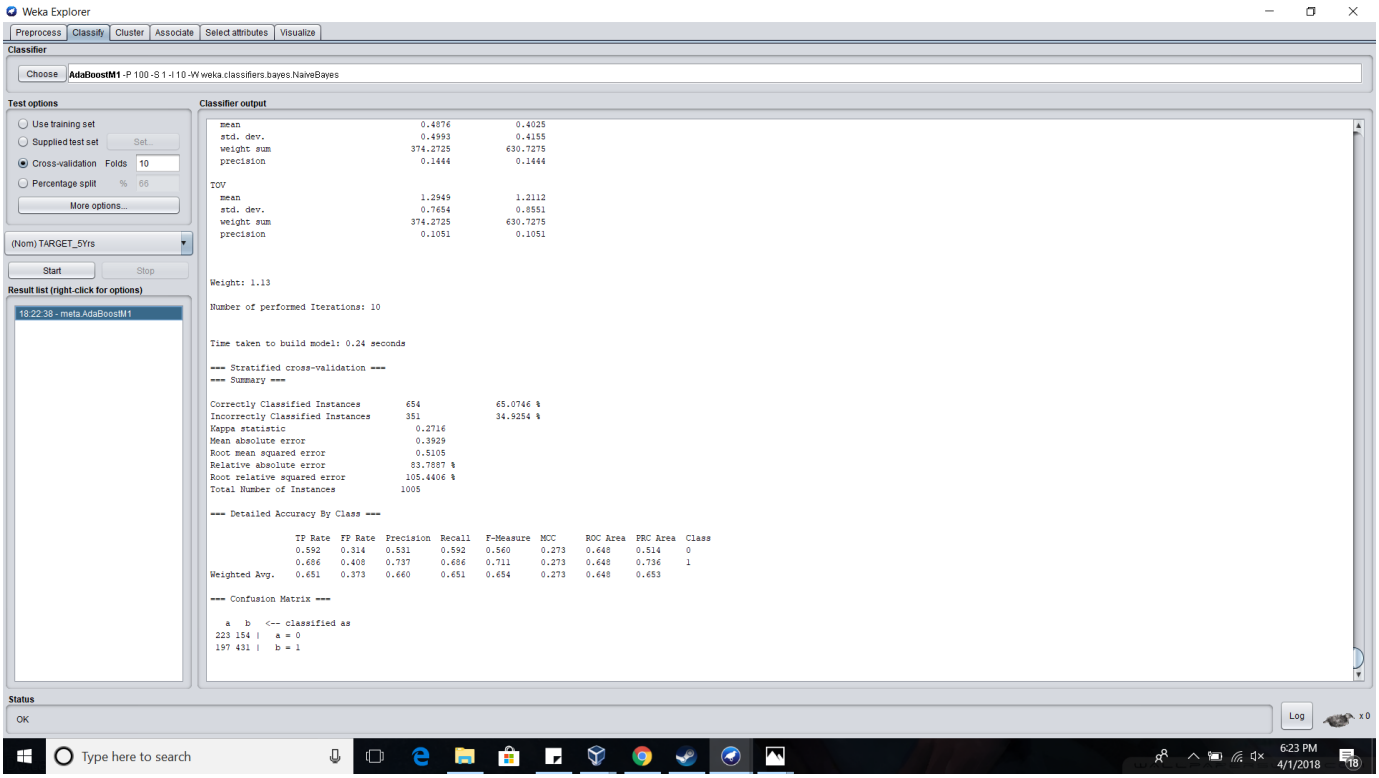
Manual – Bagging(Training)



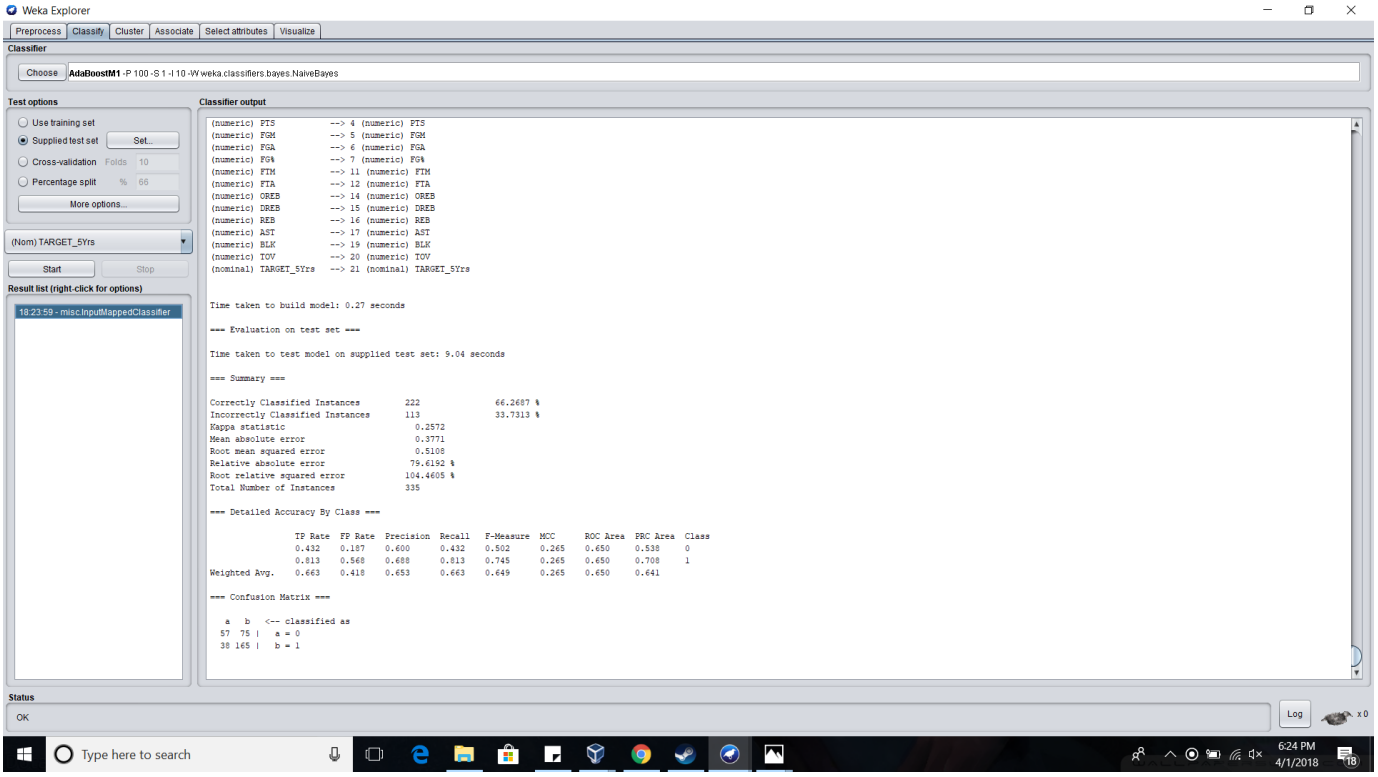
Manual – Bagging(Test)



Manual – AdaBoost(Training)



Manual – AdaBoost(Test)



Performance Comparison

Comparing all the classifiers with respective attribute selection algorithms, AdaBoost classifier with selection algorithms Gain Ratio and Info Ratio, yielded better with 69.85% accuracy while most of the four classifiers did poorly with J48.

TP Rate

Is also called sensitivity or True Positive Rate, it is the proportion that tested positive to, we observe that TP Rate for Info Gain, Adaboost performed highest with 0.69.

FP Rate

Is the probability of falsely rejecting the null hypothesis for a particular test. In the classification done here, we see that it is higher for Bagging over the rest, indicating that in this classifier the probability of falsely rejecting is more and thus a disadvantage over the rest.

ROC Curve

It is a curve where true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points of a parameter, and higher the value or closer to 1, the better it is. In the classifications done we see that Naïve Bayes yields 0.74 being the highest and AdaBoost recording 0.69 and J48 with 0.5 being least.

Based on the observations, AdaBoostM1 is the best-fit classifier for this dataset as it yields better TP Rate, FP Rate and Accuracy compared to other models.

Project Responsibility and Learning

In the course of this project, Venkata Vishnuvardhan Aluri and Aakash T.C split the work based on each phase of the project. Vishnu Aluri does classifier modelling and Dataset retrieval, while Aakash performed pre-processing. Documentation was performed together. There are several major takeaways from the project. Mainly, we get to realize the scope of Data Mining in solving various real world problems. In Addition to that, we got the hands on experience with WEKA and applying various machine-learning algorithms learnt.

References

The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011