Yiduo Feng

CS 699

project

Date: 6/21/2022

**Analysis of 2018 BRFSS Survey Data**

CS 699 Final Project (Prof. Jae Young Lee, Summer 2022)

**Dataset Details**

This data set is collected from the CDC website in 2018 on the diseases and physical conditions of people who participated in the survey. Each row in this dataset represents information about one participant, including 11933 tuples and 108 attributes.

**Objective**

Divide the dataset into two parts: training dataset and test dataset, use five different attribute selection methods to reduce the dataset, and take out some attributes for classification. Each attribute selection method will perform five different classifications and will eventually test the total 25 classifier models, and we will find the best one with the best prediction.

**Attribute selection methods 1:**

Apply an attribute selection method on the *project-training.arff* dataset to select a subset of attributes. Save it as *reduced training dataset-1*. The selected attributes are shown in the picture below.

Attribute Evaluator: **CfsSubsetEval**

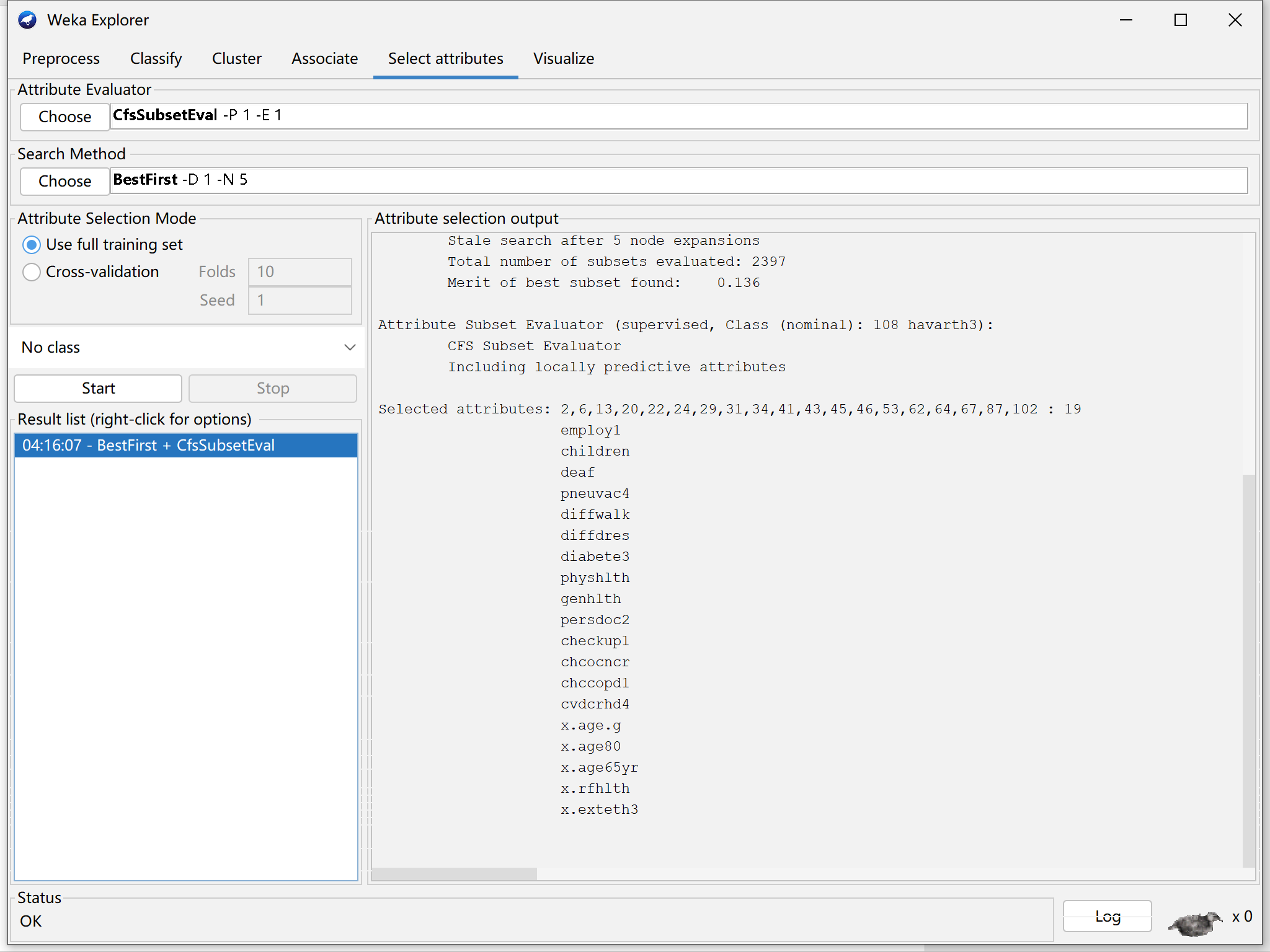
The value of a subset of attributes is evaluated by the predictive efficiency of each feature as well as its redundancy.

Search Method : **BestFirst**

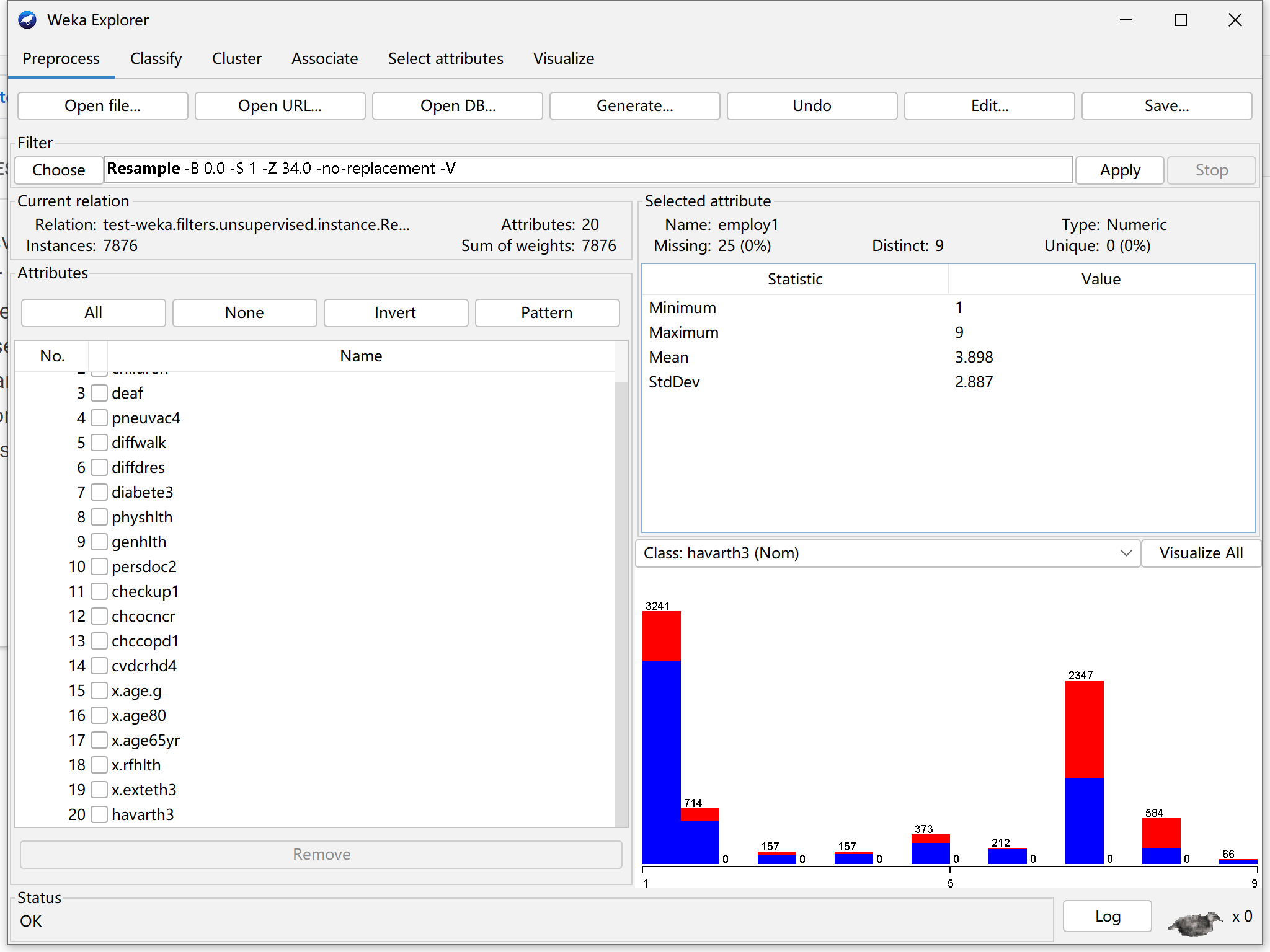
Model the dataset through a greedy algorithm, select the n best features, and select the features that yield the model with the best performance.

By using this attribute selection method, we got 19 attributes.

They are employ1, children, deaf, pneuvac4, diffwalk, diffdres, diabete3, physhlth, genhlth, persdoc2, checkup1, chcocncr, chccopd1, cvdcrhd4, x.age.g, x.age80, x.age65yr, x.rfhlth, x.exteth3.

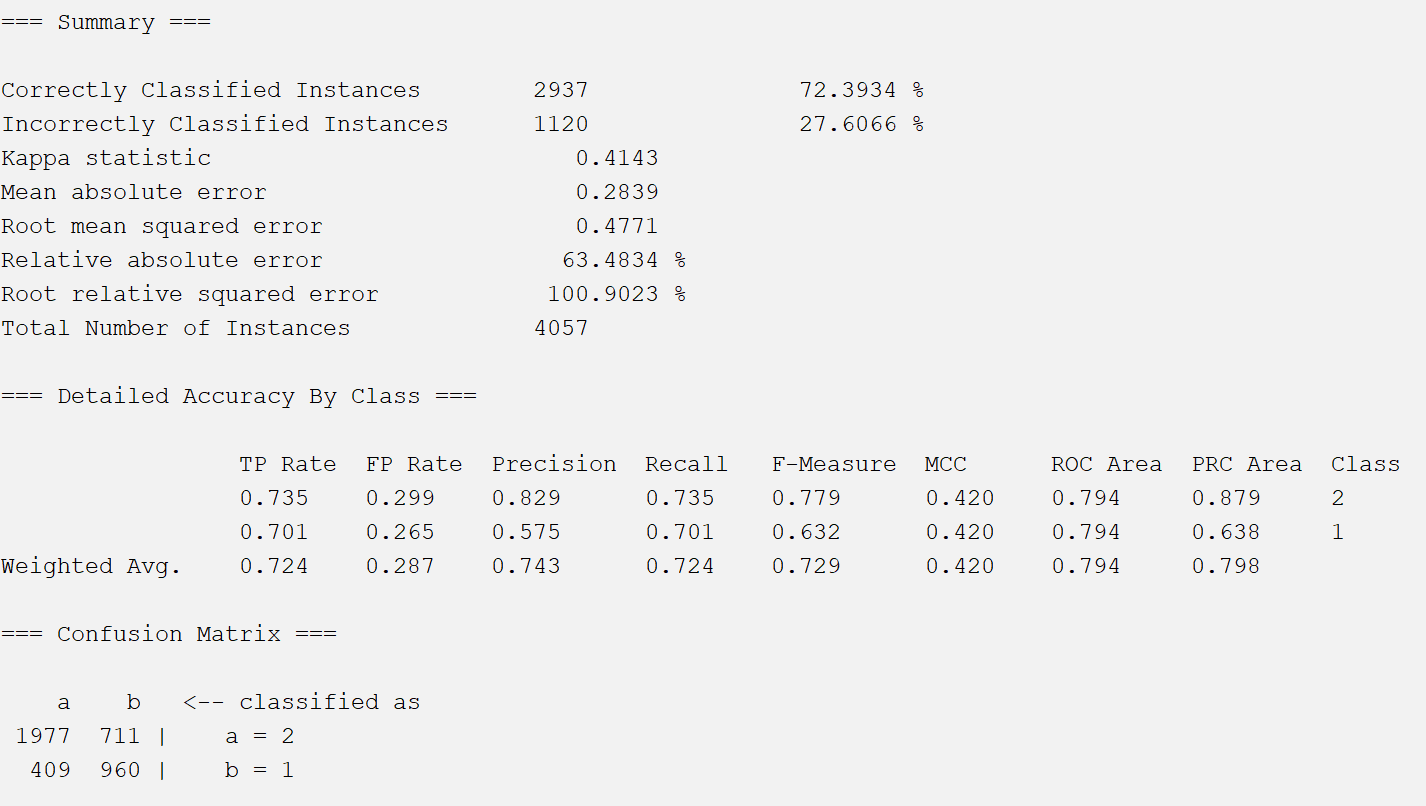


From the project-test. arff dataset, select only those attributes that are in the reduced training dataset, and save it as reduced test dataset-1.



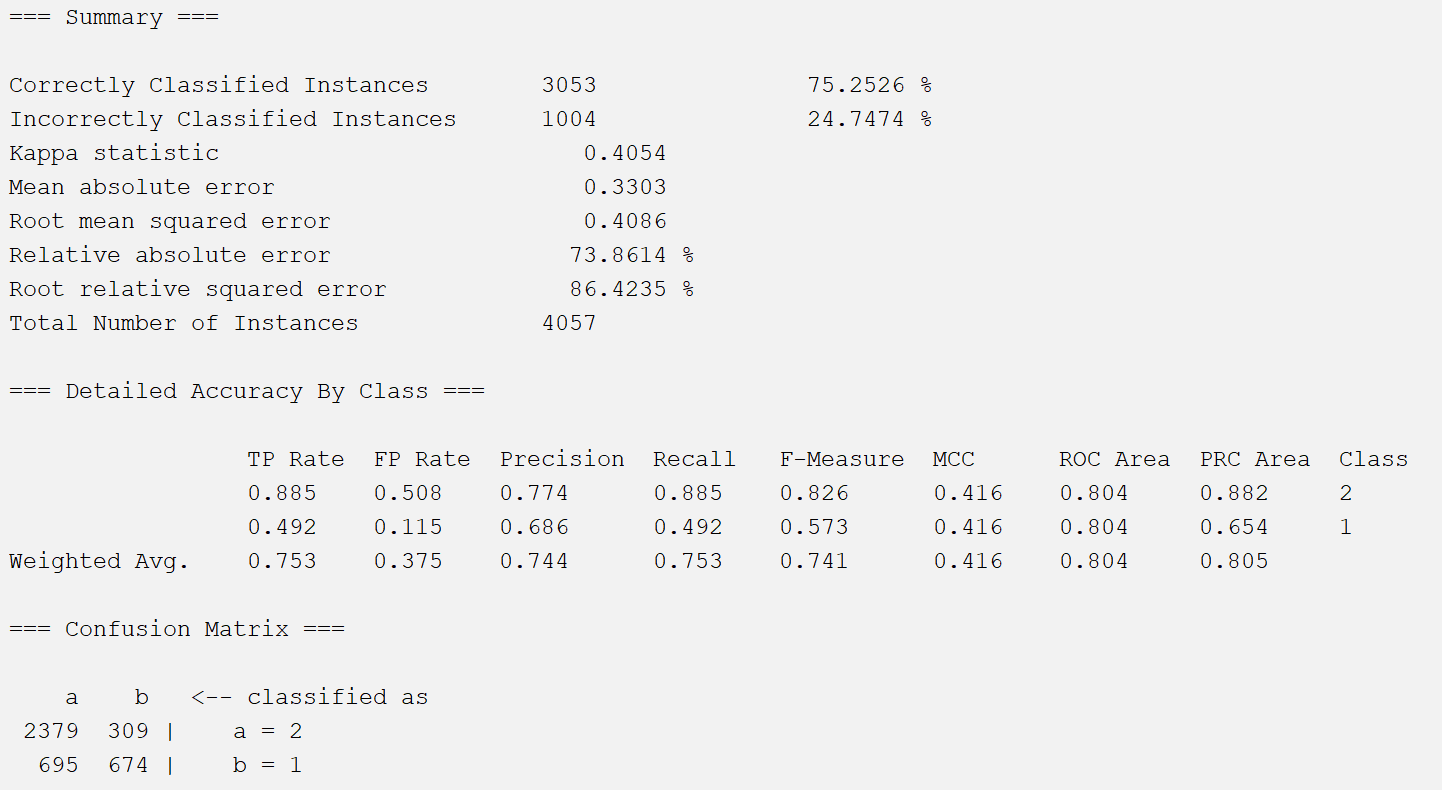
1. NaiveBayes

Based on Bayes' theorem, use Bayesian formula to calculate the relationship between attributes, features and classes for classification.



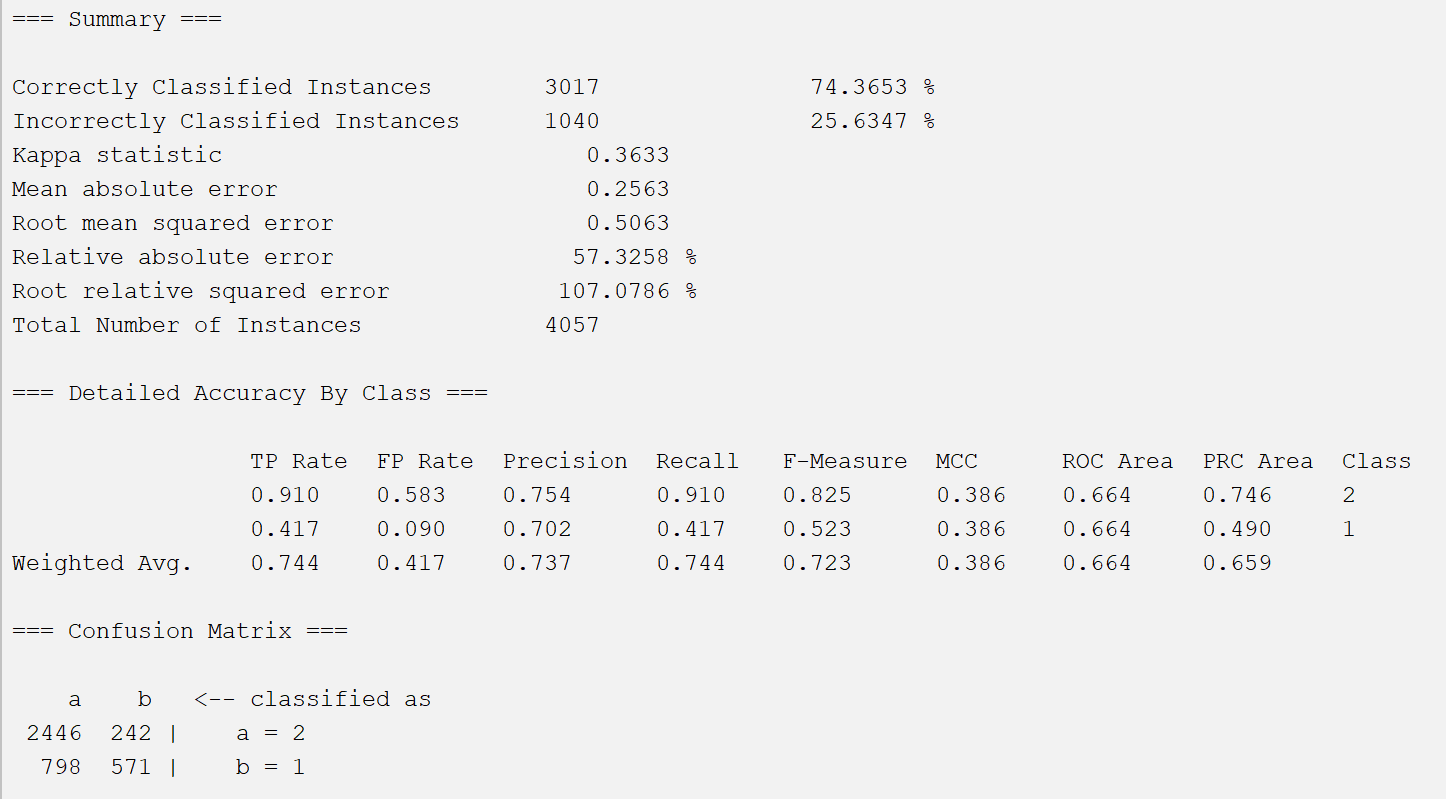
1. logistic regression

Like many other machine learning algorithms, logistic regression is borrowed from statistics. Logistic regression is the preferred method for binary classification tasks. It outputs a discrete binary result between 0 and 1. Base on there are four actions in this data set, and Logistic regression only have 2 kind of results, we consider action "Allow" is 0, and others is 1.



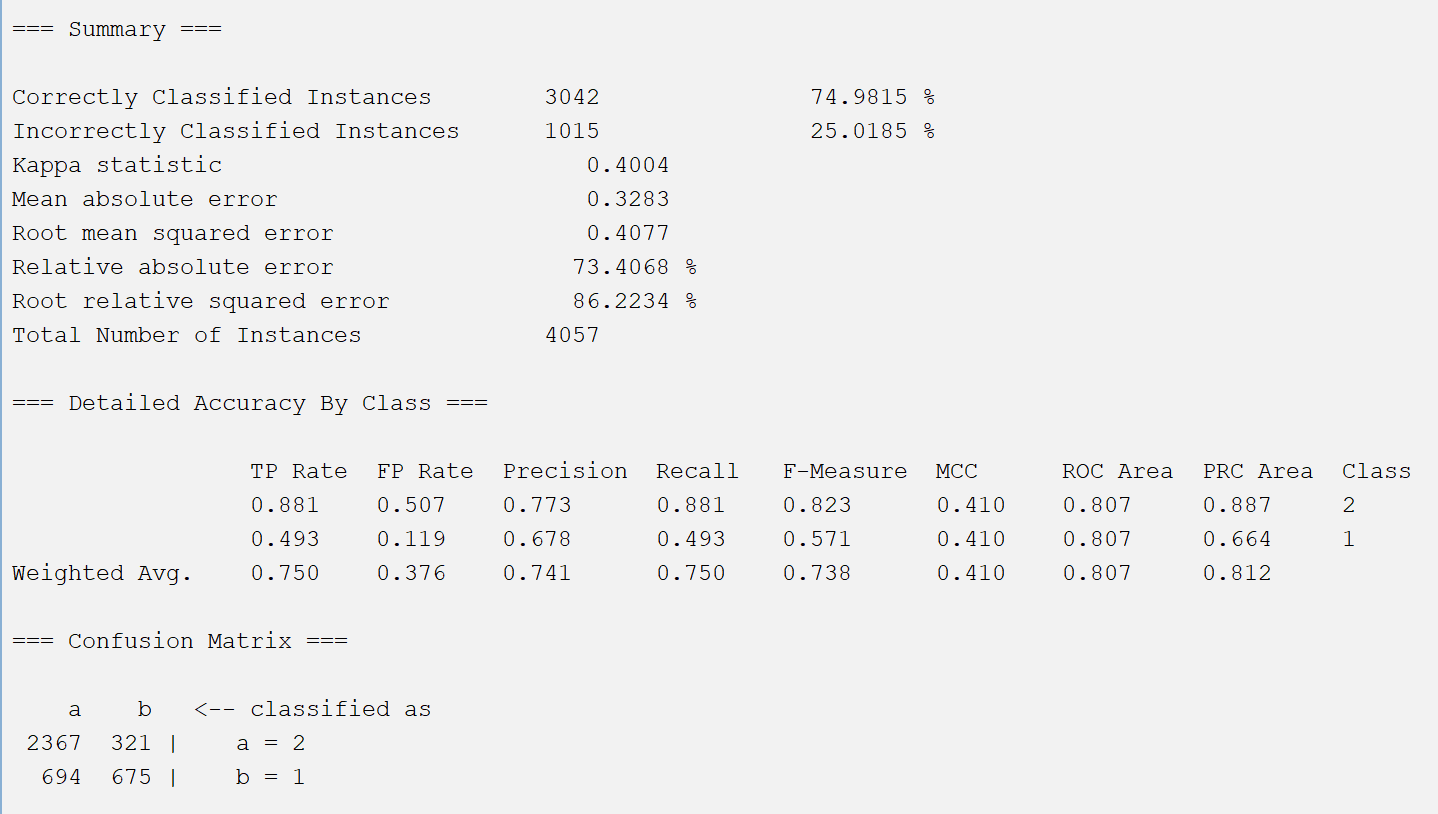
1. SGD

The full name of sgd algorithm is stochastic gradient descent, which has the advantage of faster convergence than batch gradient descent. The word "random" in the name of the algorithm is the central spirit of the algorithm.



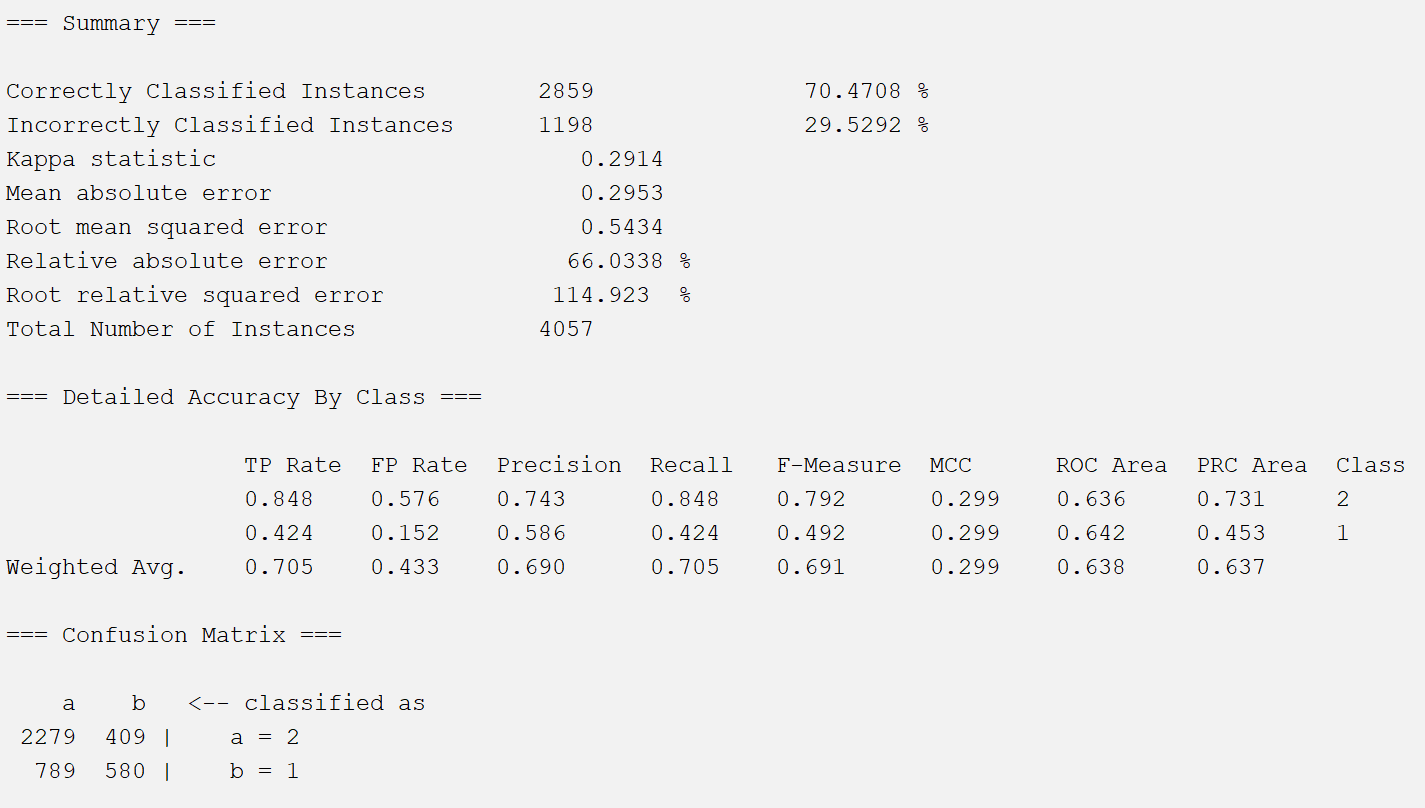
1. SimpleLogistic

Similar to Logistic, SimpleLogistic uses LogitBoost, while Logistic uses a Ridge estimator.



1. VotedPerceptron

The method is based on the intuition that since a prediction vector produces more correct predictions, it should have a larger weight.



**Attribute selection methods 2:**

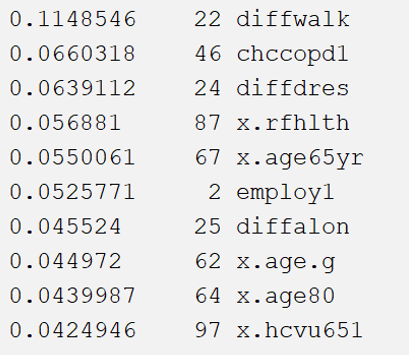
Attribute Evaluator: **GainRatioAttributeEval**

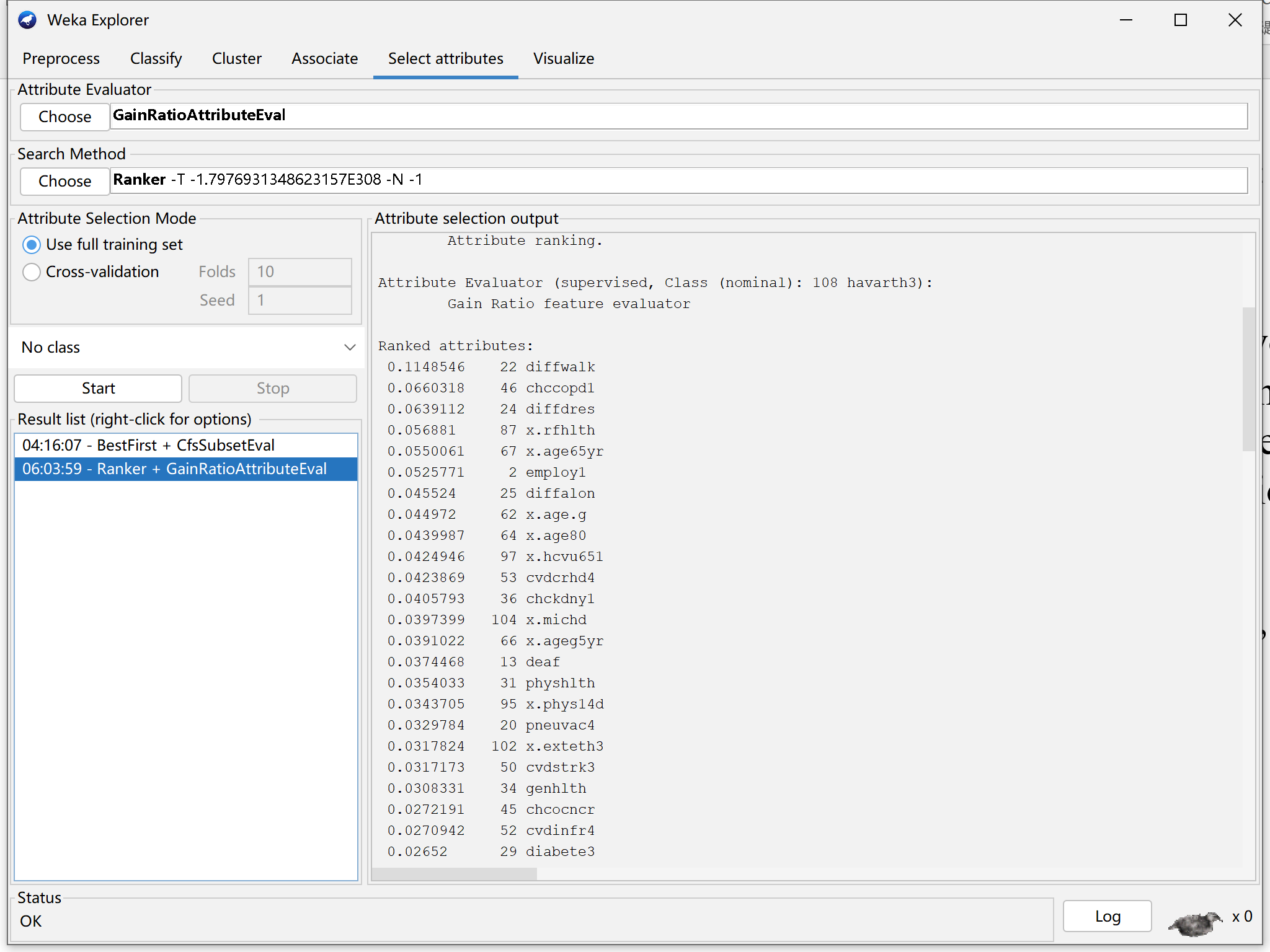
By calculating the gain ratio relative to the class, the value of the attribute is evaluated. Search Method : **Ranker**

Rank attributes by their assessed value.

Because this method provides all properties and it ranks properties by value. Here I choose the top ten attributes to reduce the dataset.

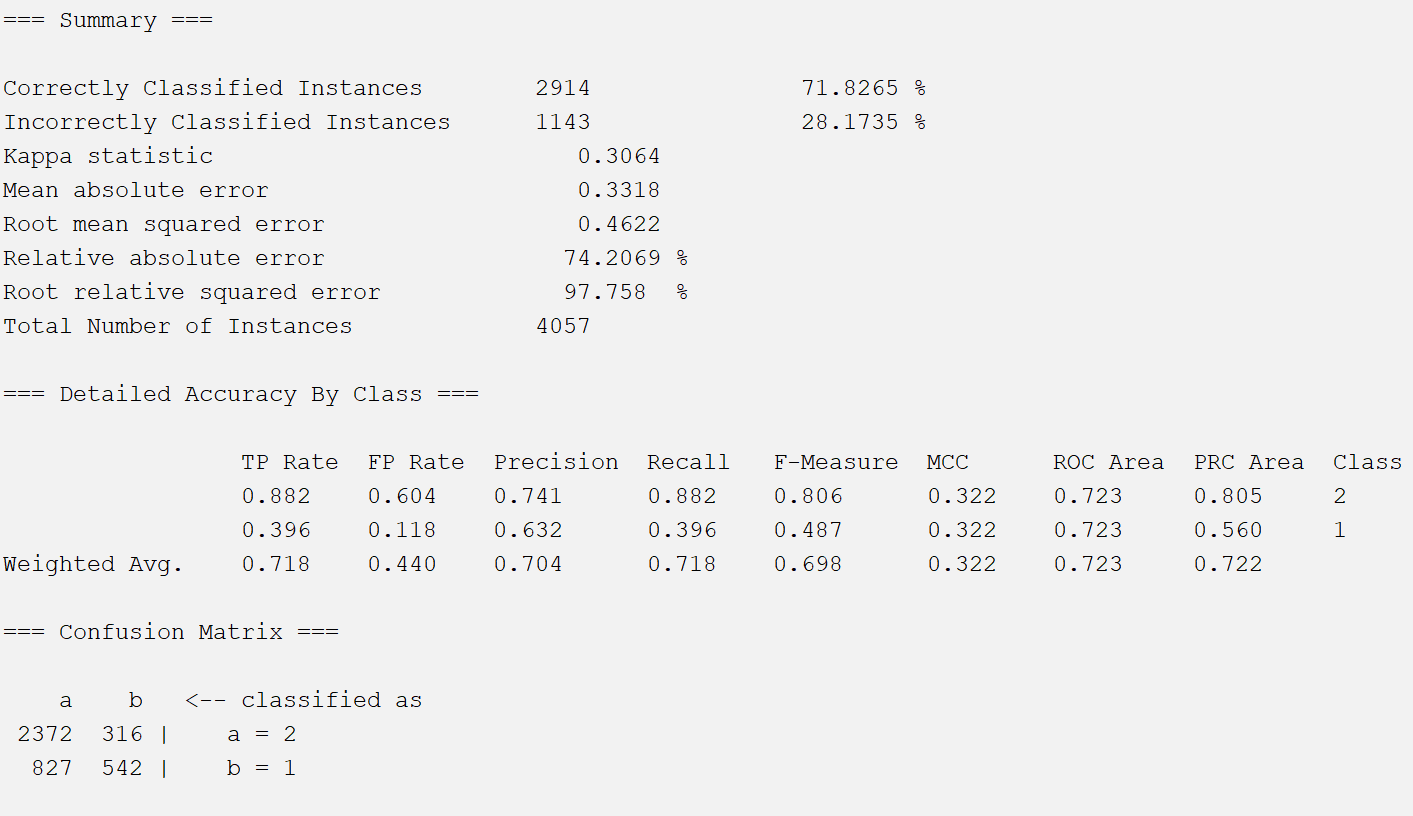
The attributes are:





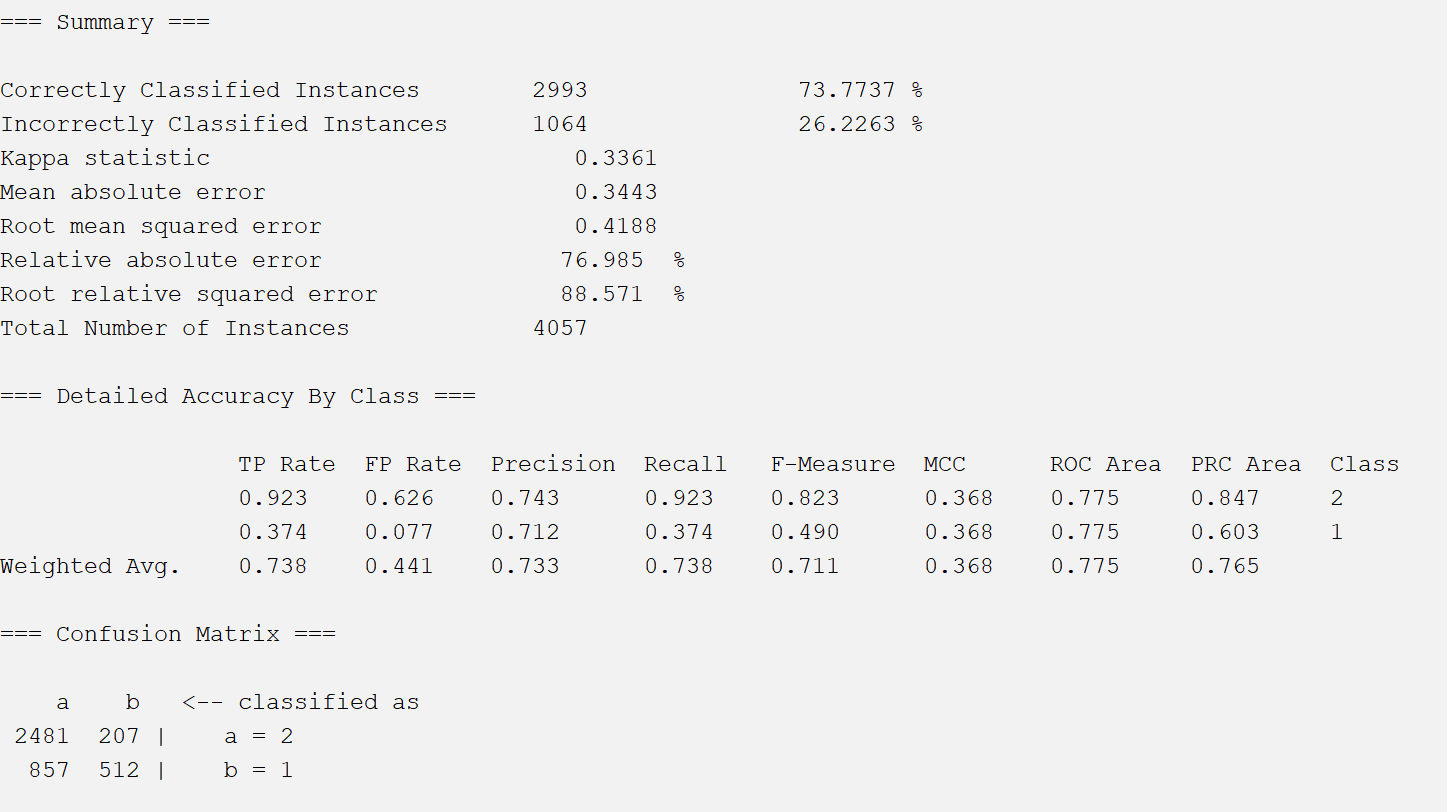
1. IBk

Given a training data set, for a new input instance, find the K instances closest to the instance in the training data set, and most of these K instances belong to a certain class, classify the input instance into this class.



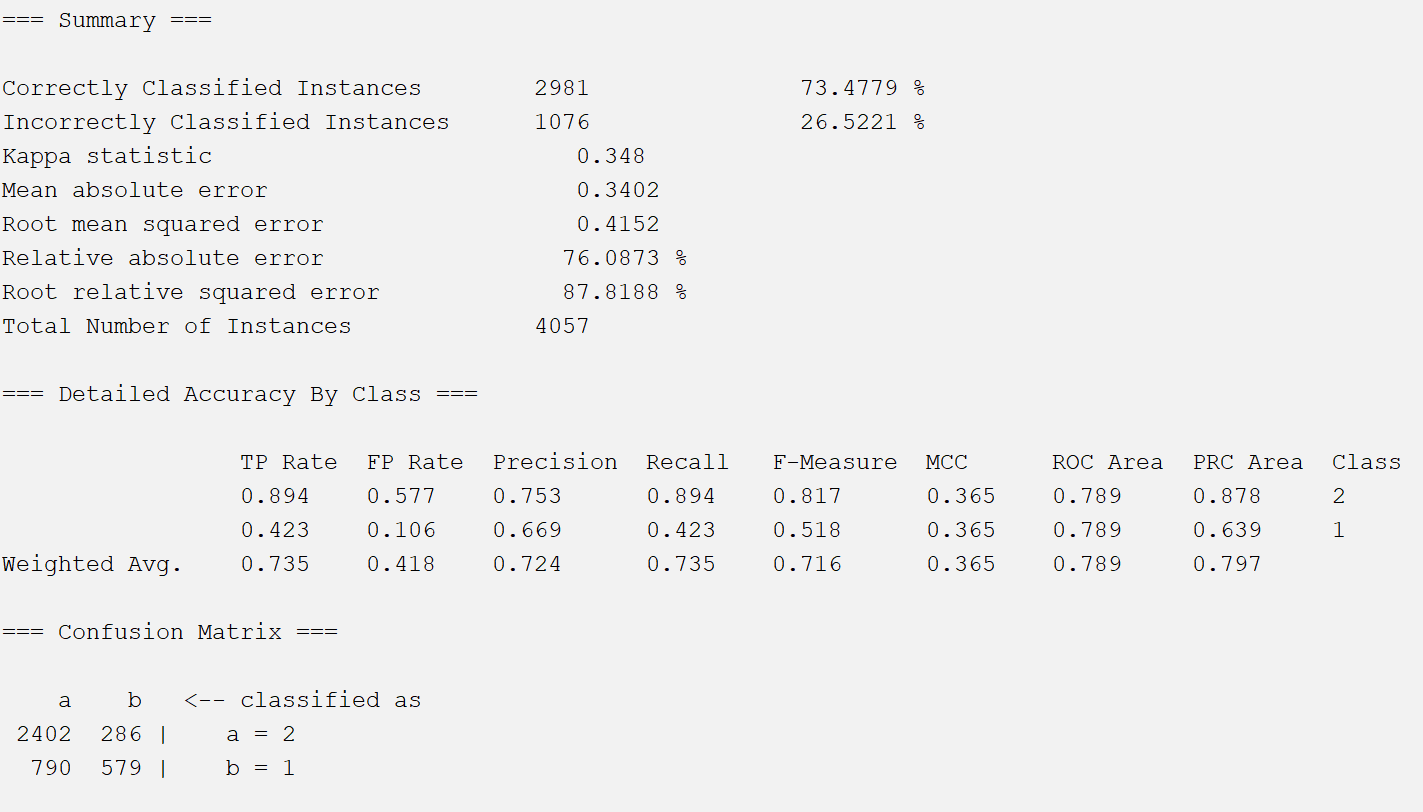
1. AdaBoostM1

According to the classification error rate of the current classifier, the distribution of sample weights is adjusted to ensure that the weight of misclassified samples increases and the weight of correctly classified samples decreases; and the weight of the current classifier in the final decision is adjusted.



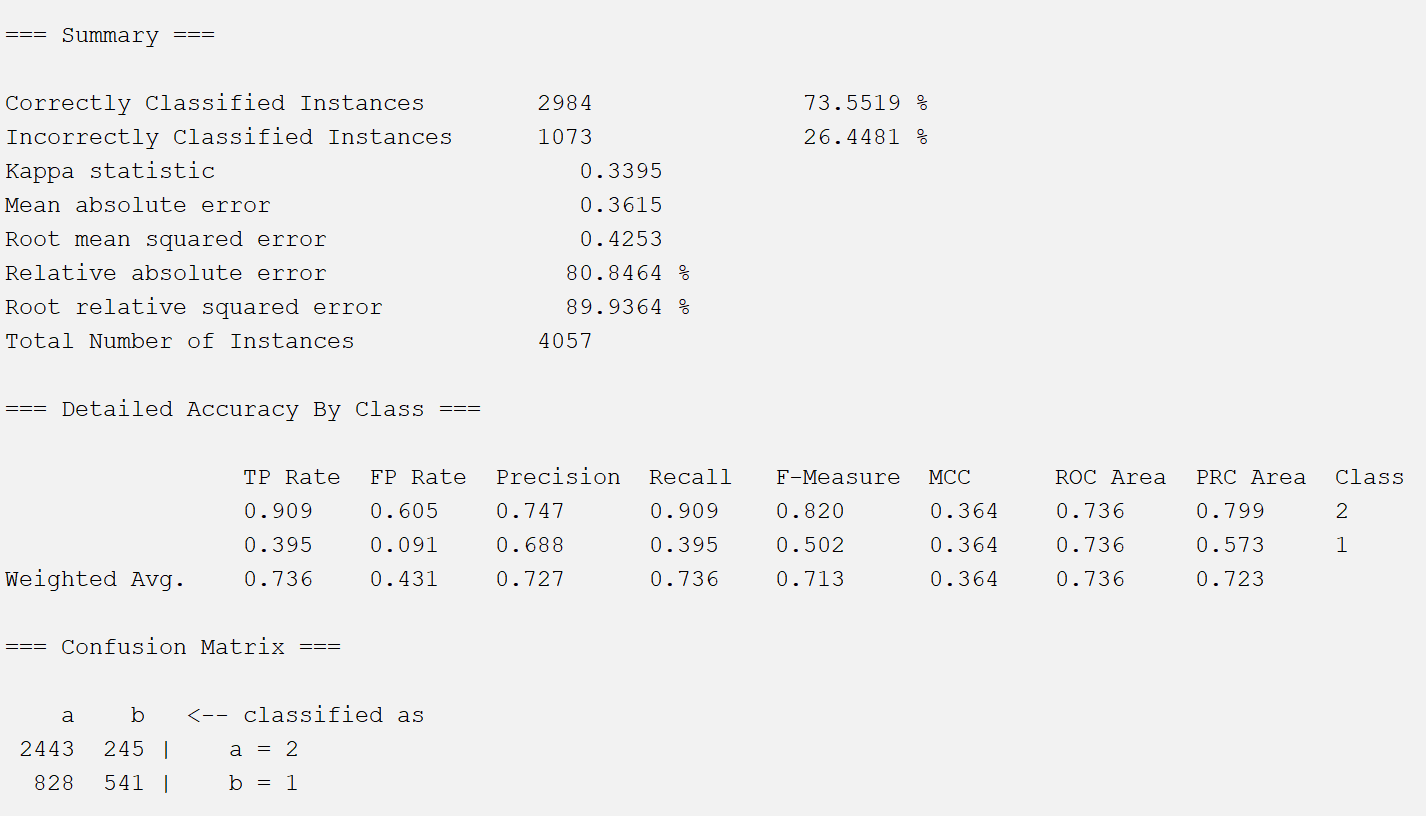
1. Bagging

Techniques for reducing generalization error by combining several models. The main idea is to train several different models separately, and then let all models vote on the output of the test examples.



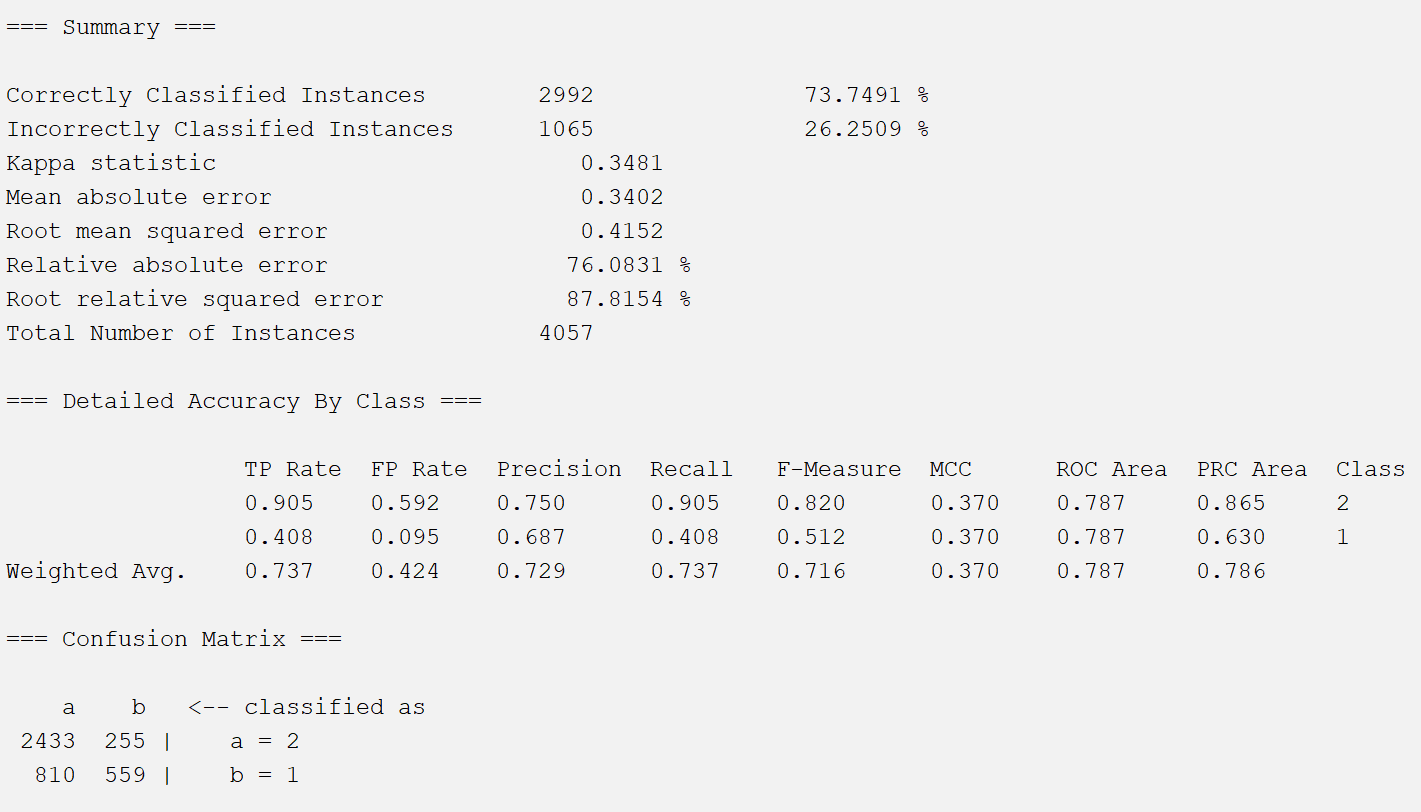
1. FilteredClassifier

A class that runs a classifier on data that has passed the filter.



1. LogitBoost

The method of integrating several classifiers into one classifier, before the boosting algorithm was generated, there were two more important methods of integrating multiple classifiers into one classifier.



**Attribute selection methods 3:**

Attribute Evaluator: **InfoGainAttributeEval**

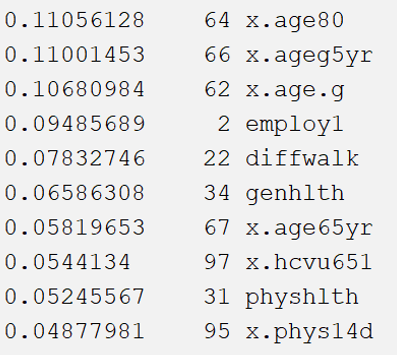
The value of the attribute is obtained by calculating the information gain.

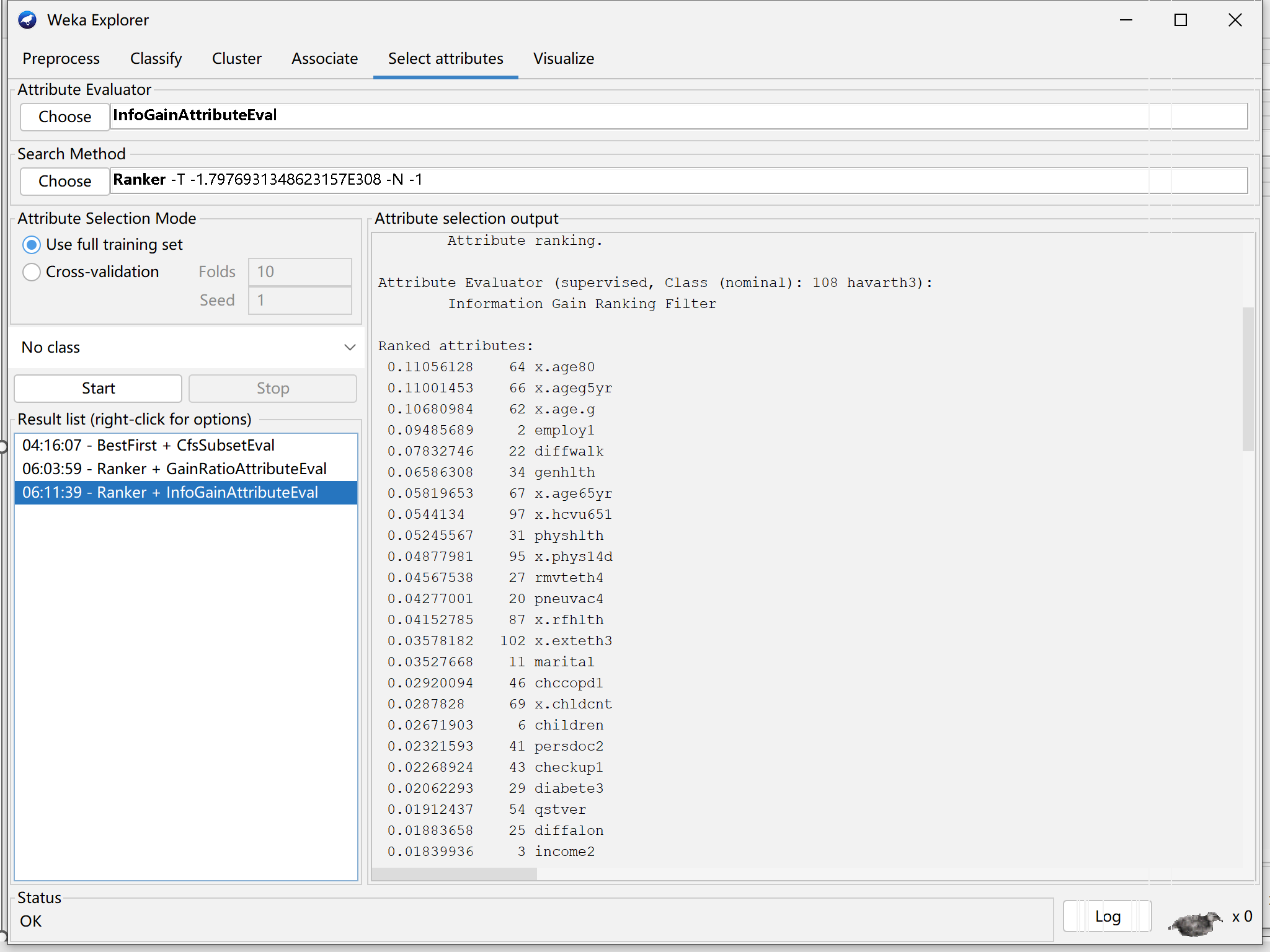
Search Method : **Ranker**

Rank attributes by their assessed value.

Because this method provides all properties and it ranks properties by value. Here I choose the top ten attributes to reduce the dataset.

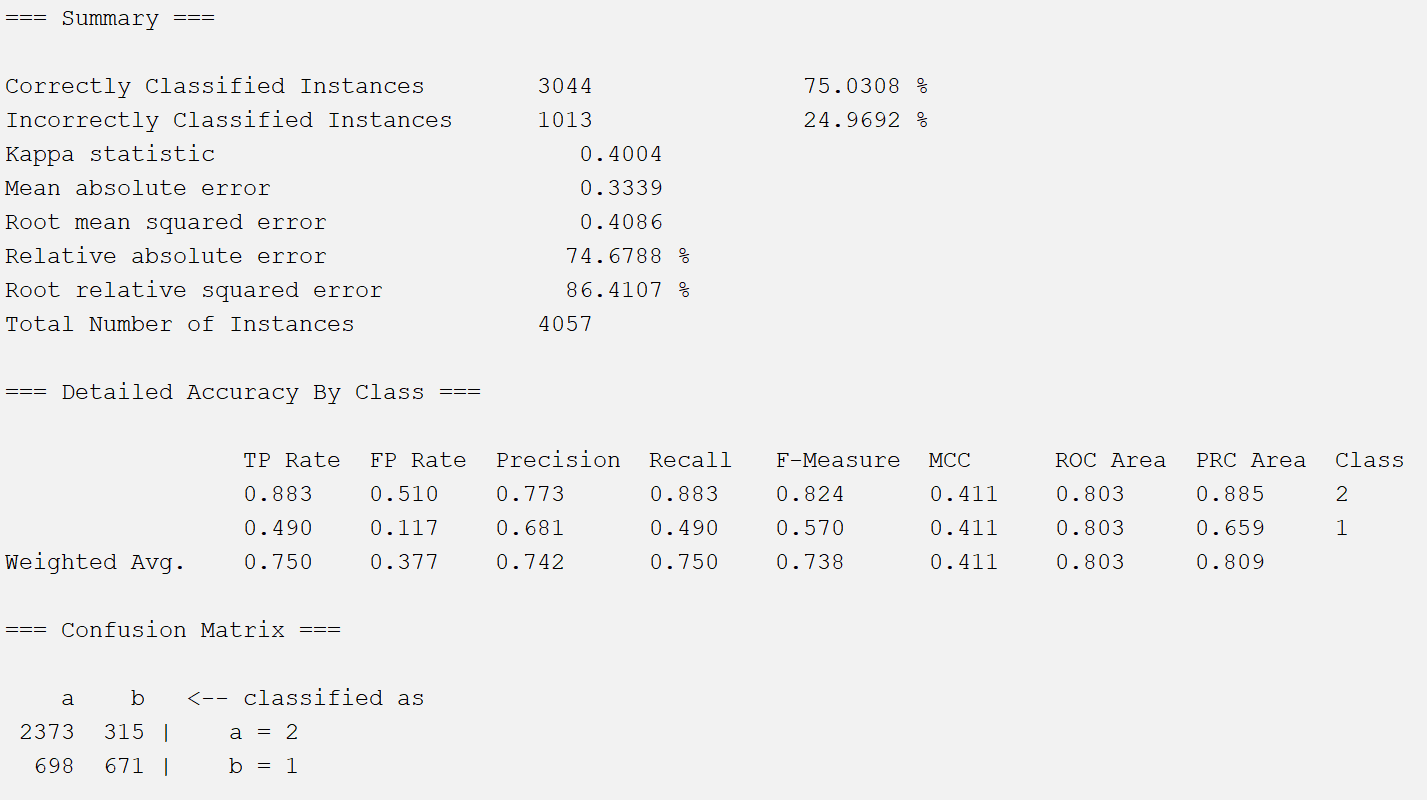
The attributes are:





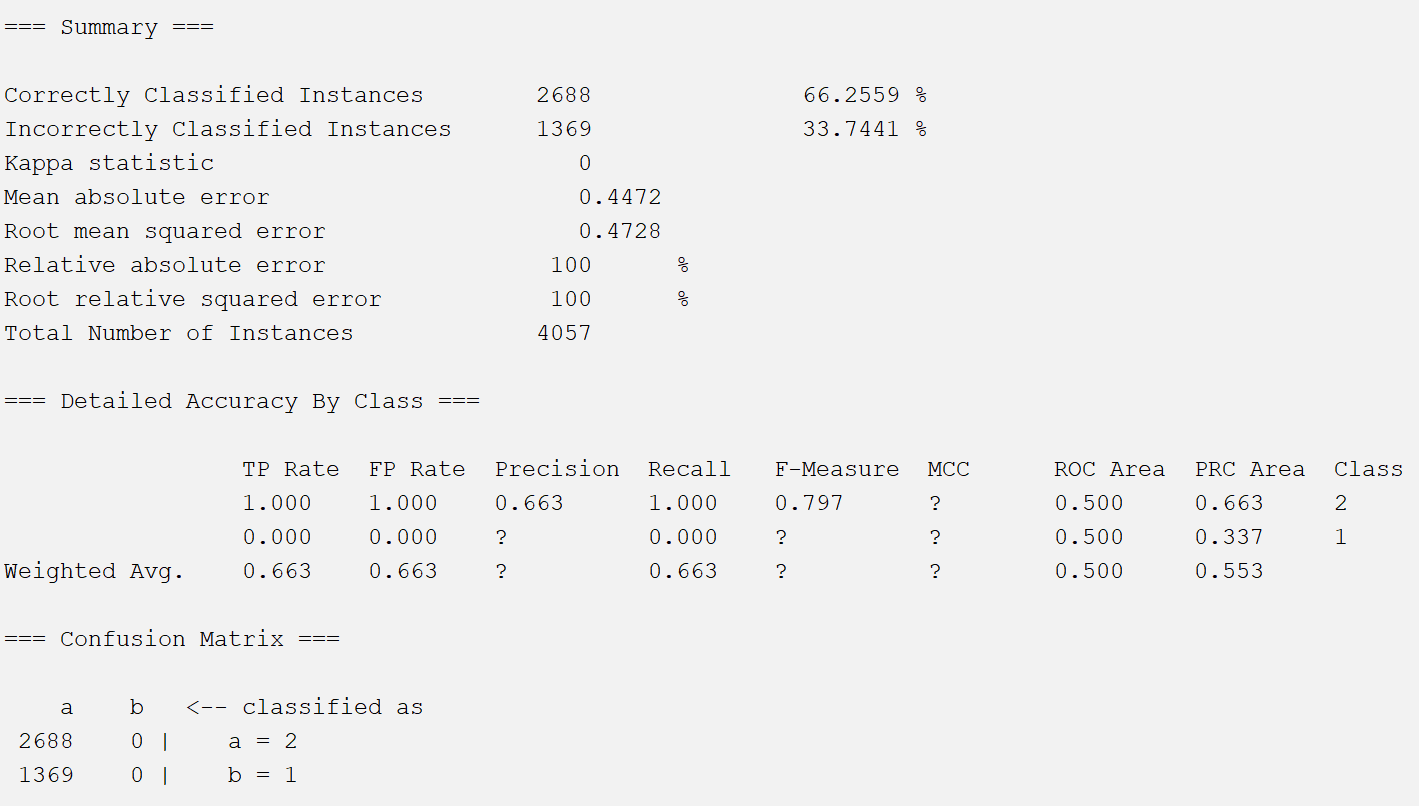
1. MultiClassClassifier

A meta-classifier for processing multi-class datasets with binary classifiers.



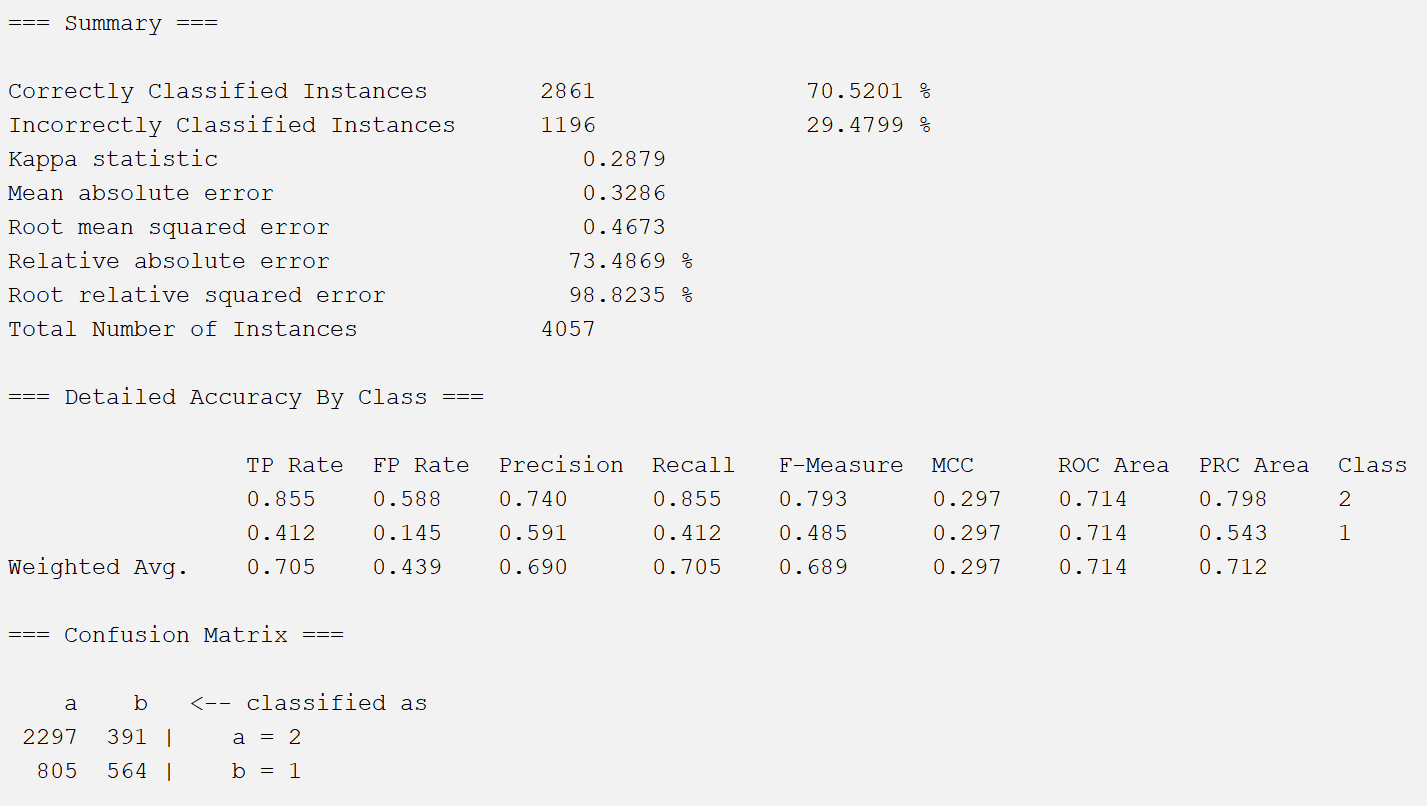
1. MultiScheme

Select a classifier class from multiple classifiers to obtain cross-validation and performance on training data.



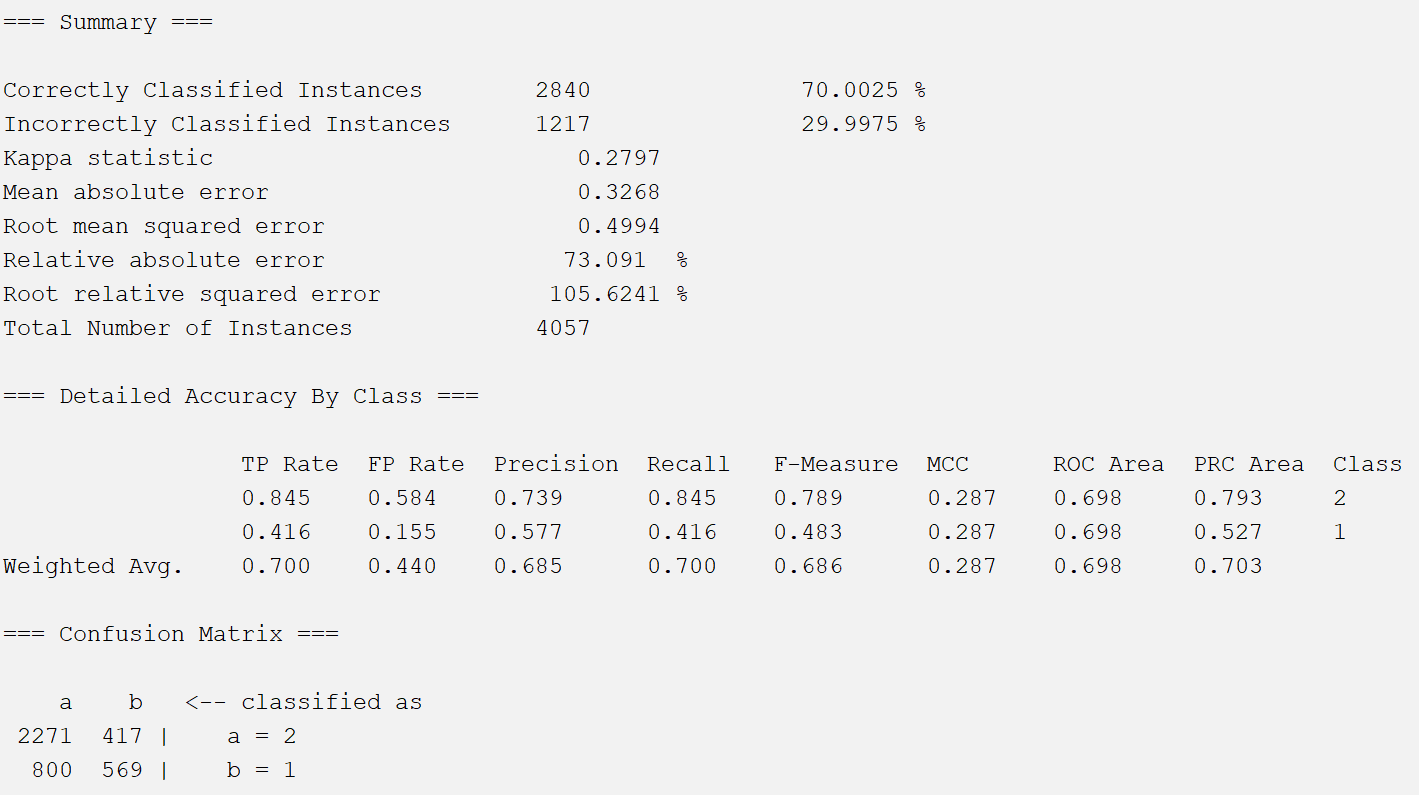
1. RandomCommittee

Build an ensemble of random classifiers.



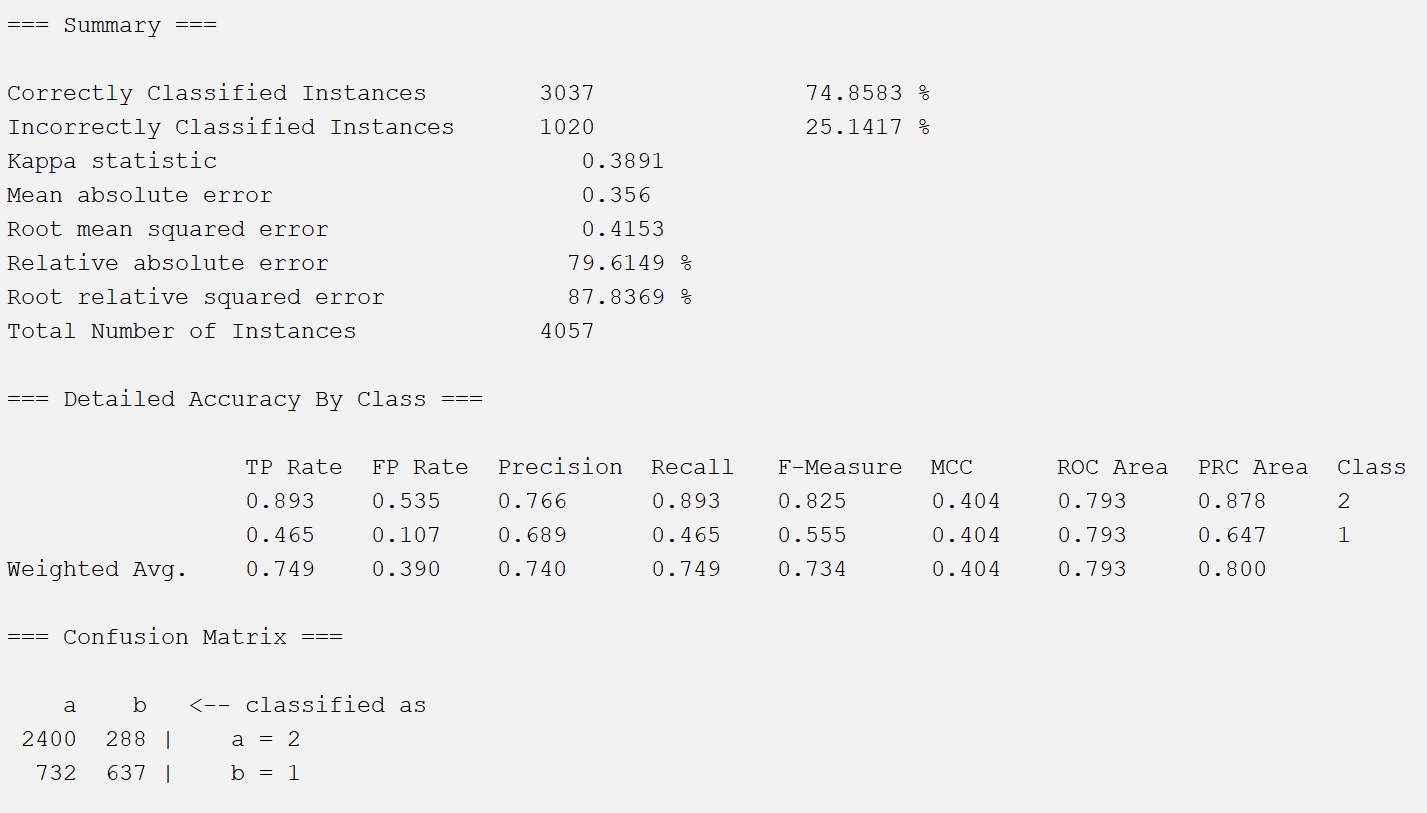
1. RandomizableFilteredClassifier

Build an ensemble of random classifiers.



1. RandomSubSpace

It is a type of ensemble learning. Random subspace reduces the correlation between each classifier by training each classifier with a random subset of features instead of all features.



**Attribute selection methods 4:**

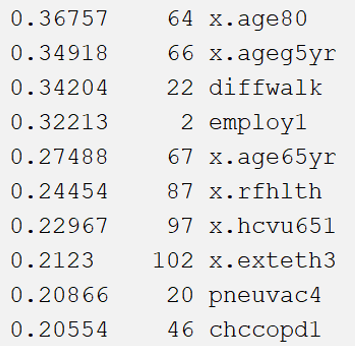
Attribute Evaluator: **CorrelationAttributeEval**

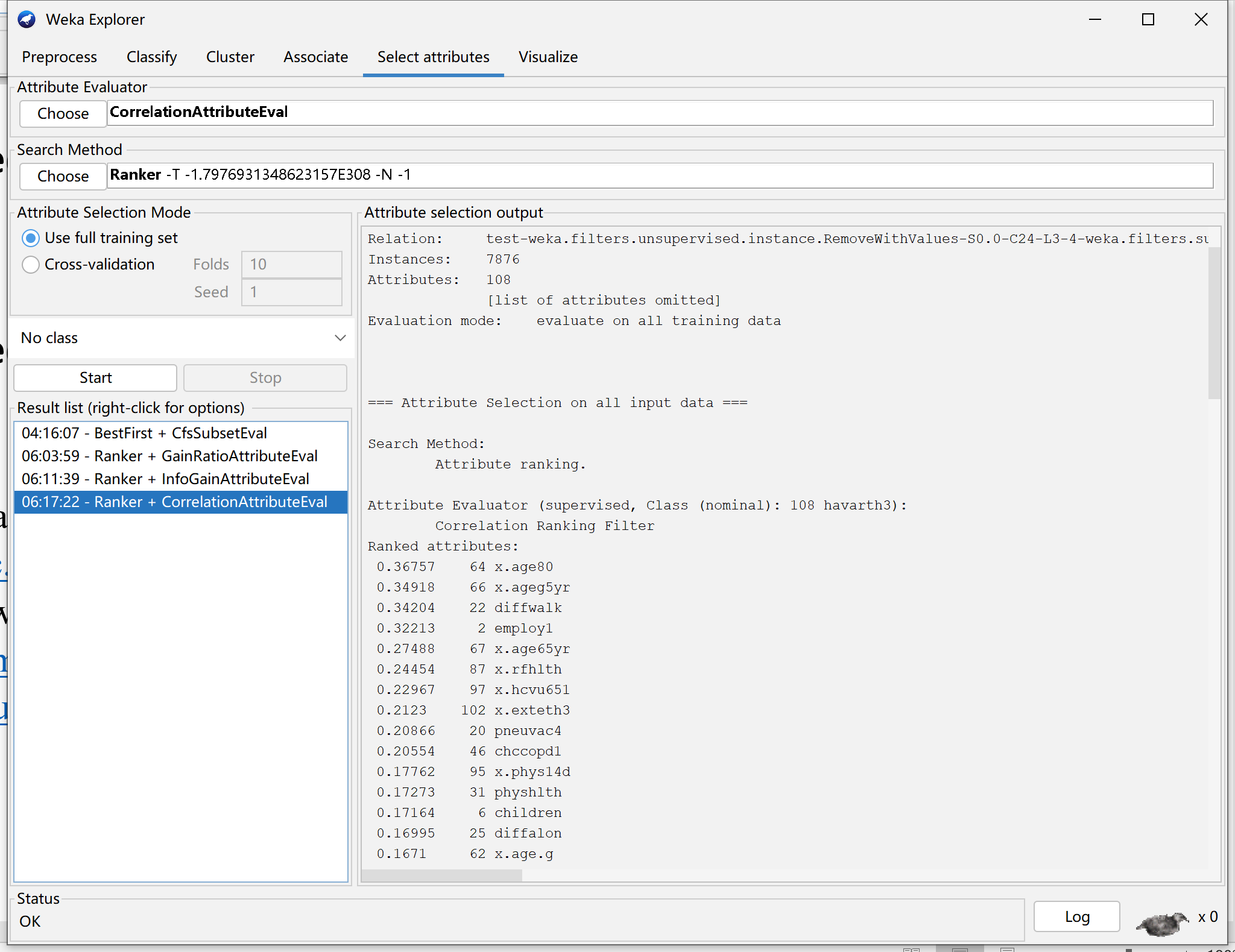
By calculating the correlation relative to the class, the value of the attribute is evaluated. Search Method : **Ranker**

Rank attributes by their assessed value.

Because this method provides all properties and it ranks properties by value. Here I choose the top ten attributes to reduce the dataset.

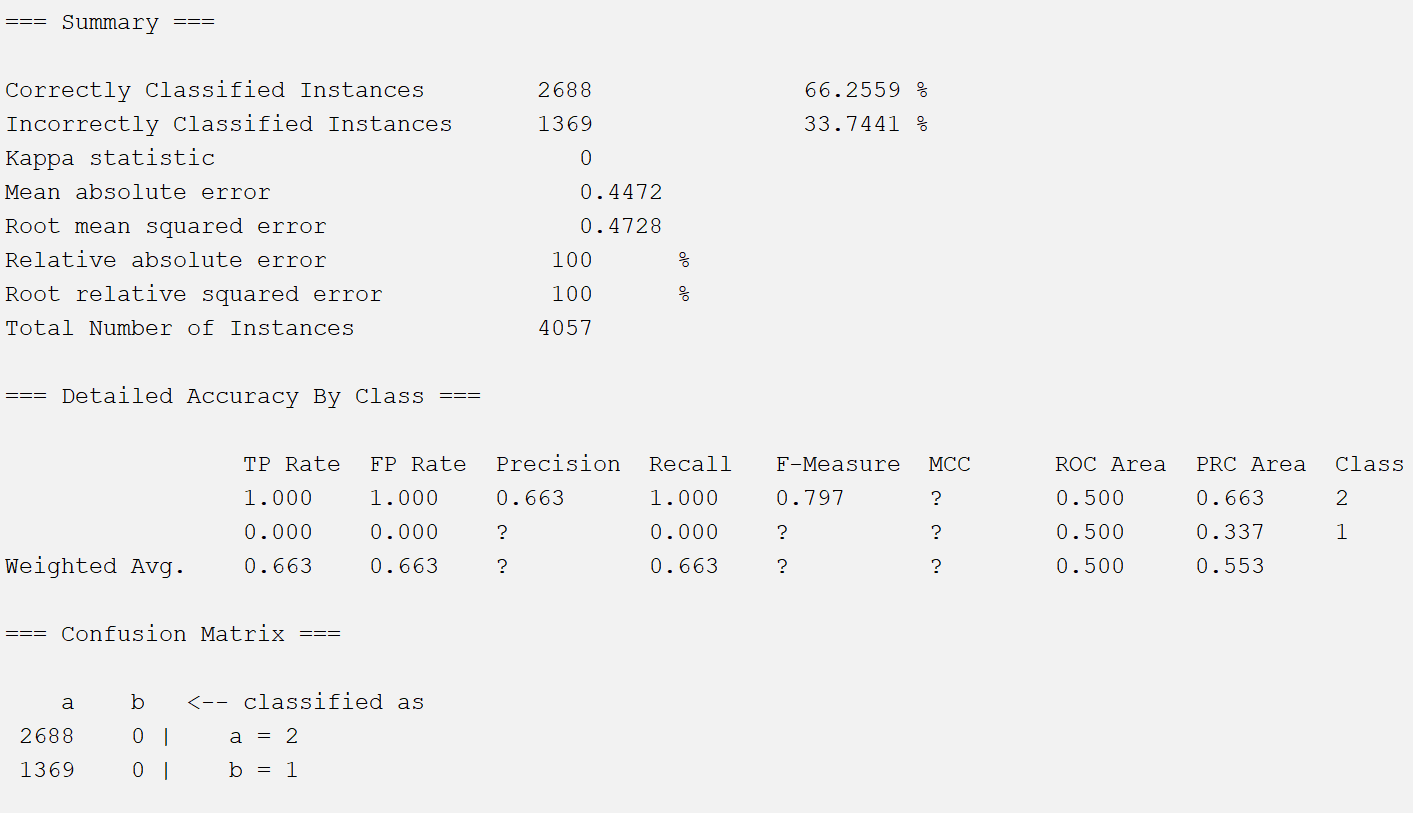
The attributes are:





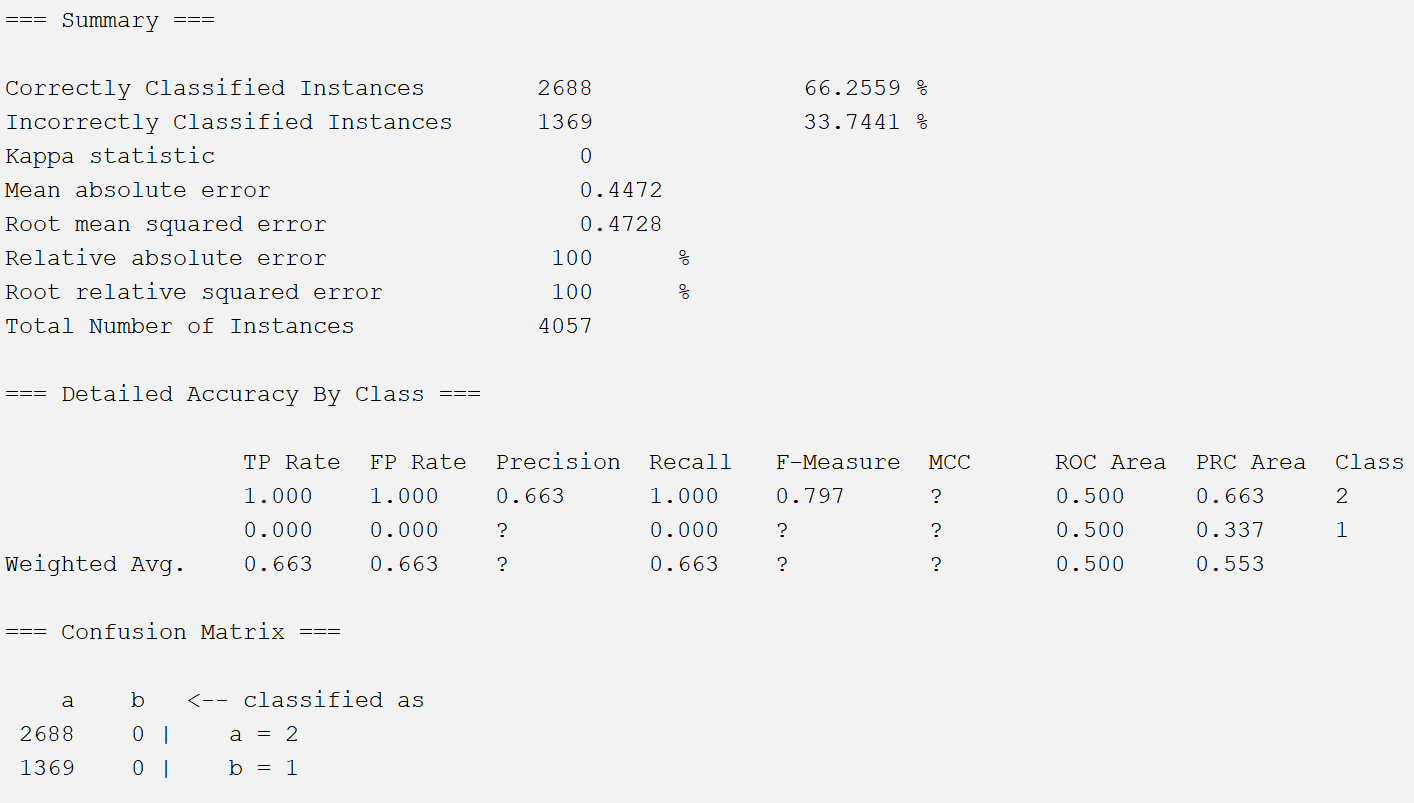
1. Stacking

Take the answer of the base model as input, and let the secondary learner learn to organize the assignment of weights to the answer of the base model.



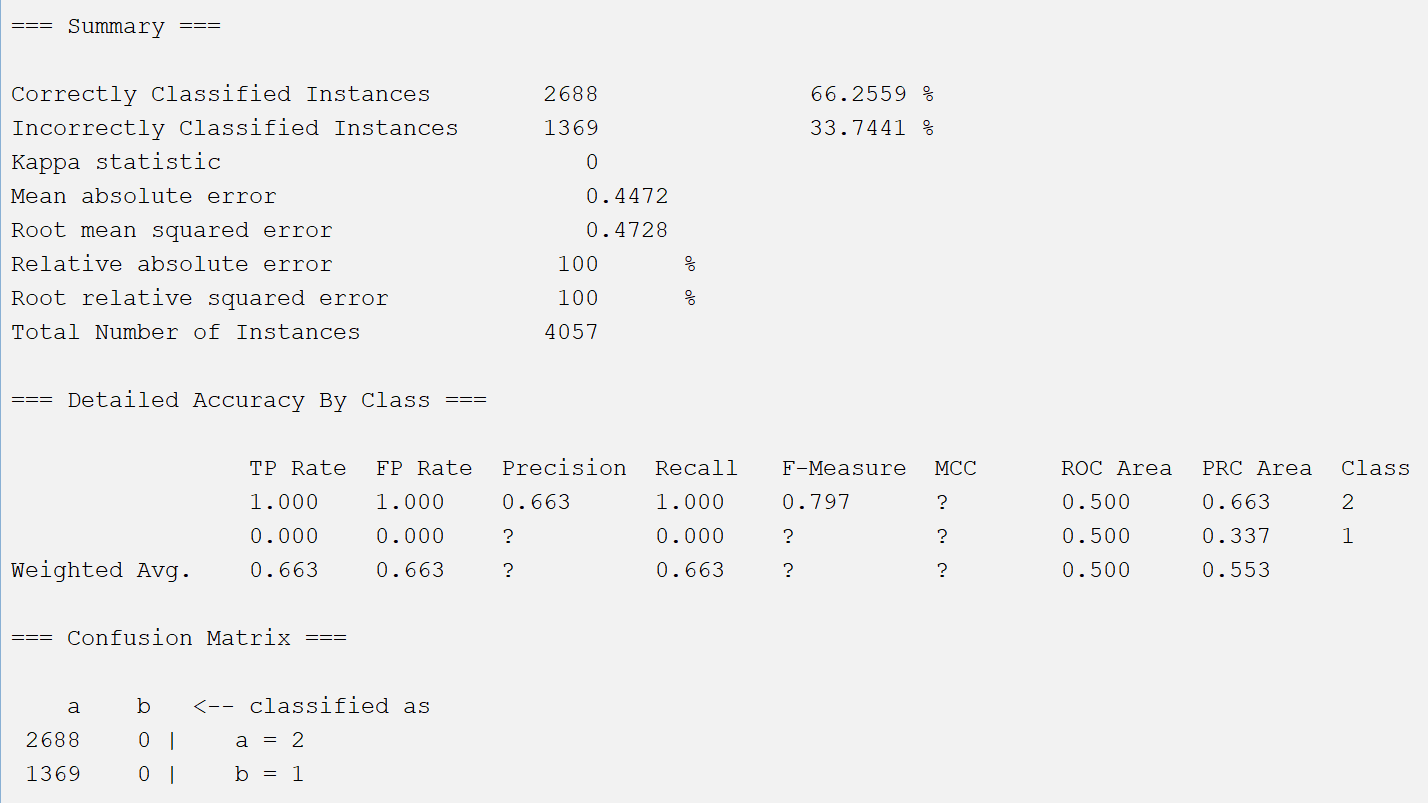
1. Vote

If an element appears, if it is equal to the candidate element, the votes of the candidate element are increased by 1; if it is not equal to the candidate element, the votes of the candidate element are decreased by 1.



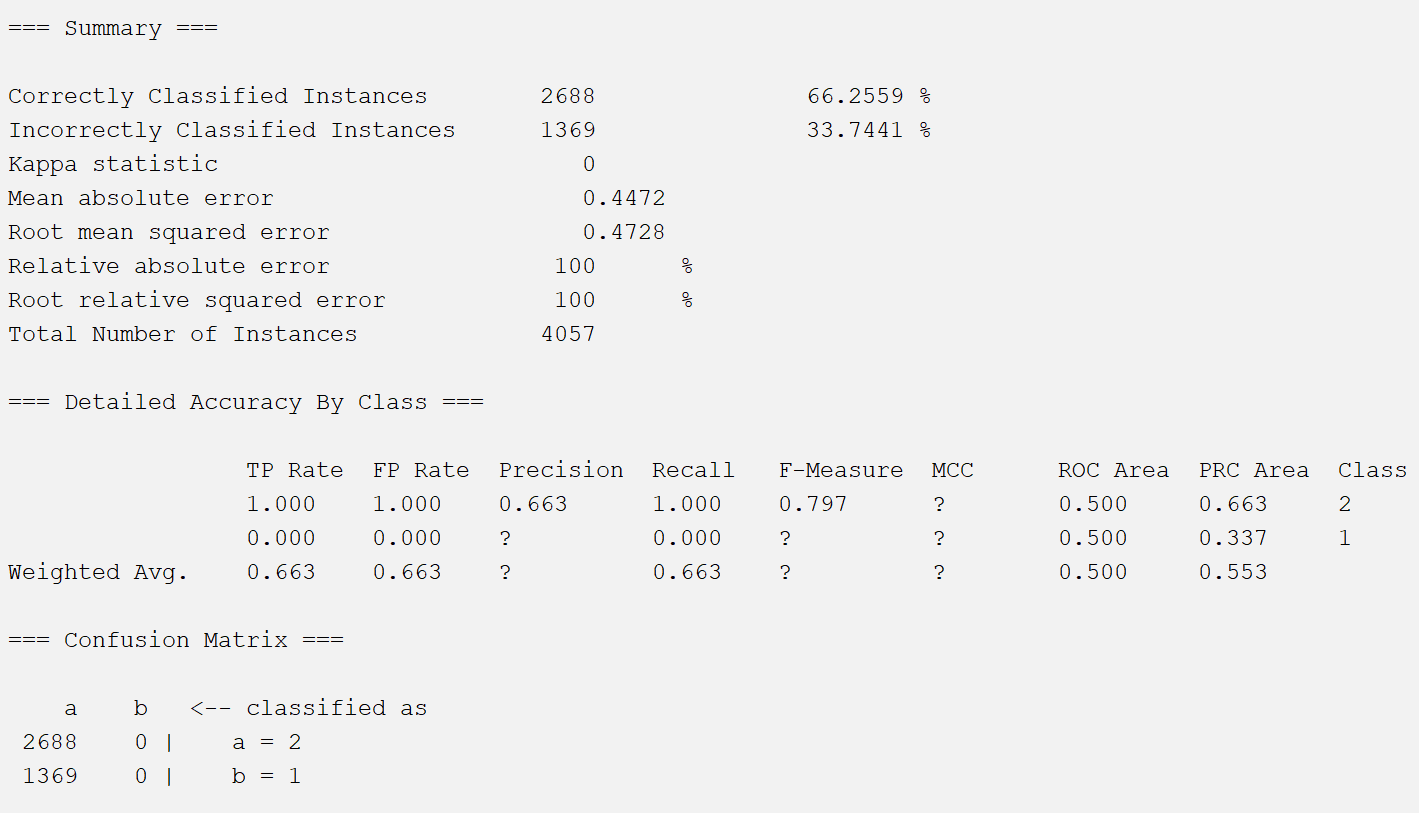
1. WeightedInstancesHandlerWrapper

A wrapper around a classifier to enable weighted instance support.



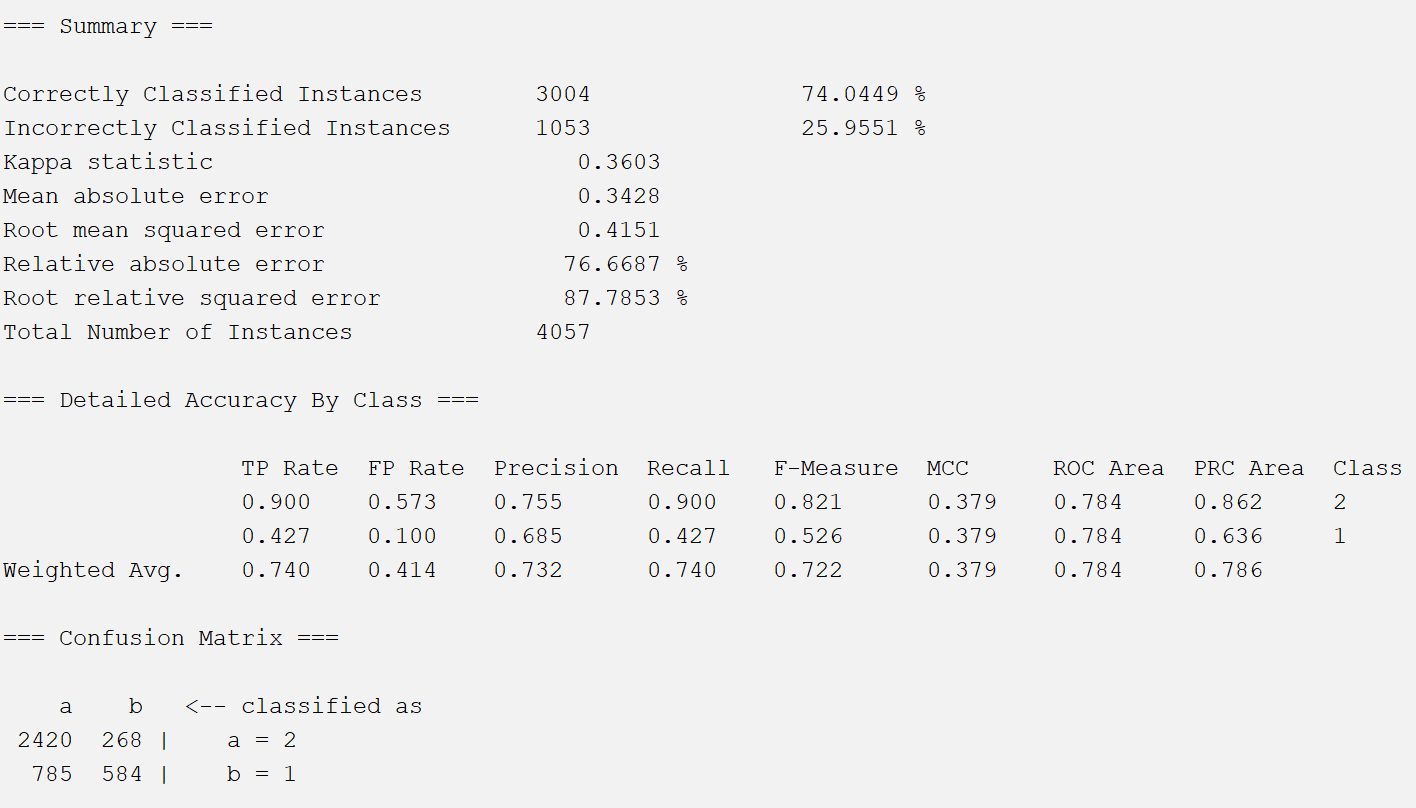
1. InputMappedClassifier

The incompatibility problem is resolved by establishing a mapping between the structures on which the classifier is built.



1. DecisionTable

The process of recursively selecting the most characteristic feature and dividing the training data according to this feature so that there is a best classification process for each sub-data set.



**Attribute selection methods 5:**

Attribute Evaluator: **SymmetricalUncertAttributeEval**

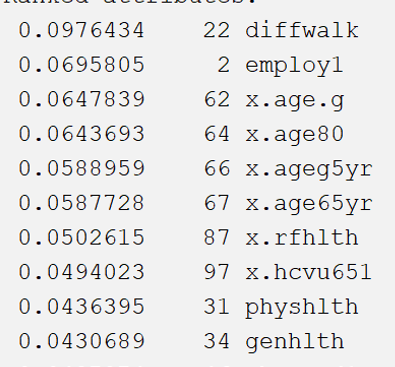
By calculating the Symmetrical uncertainty relative to the class, the value of the attribute is evaluated.

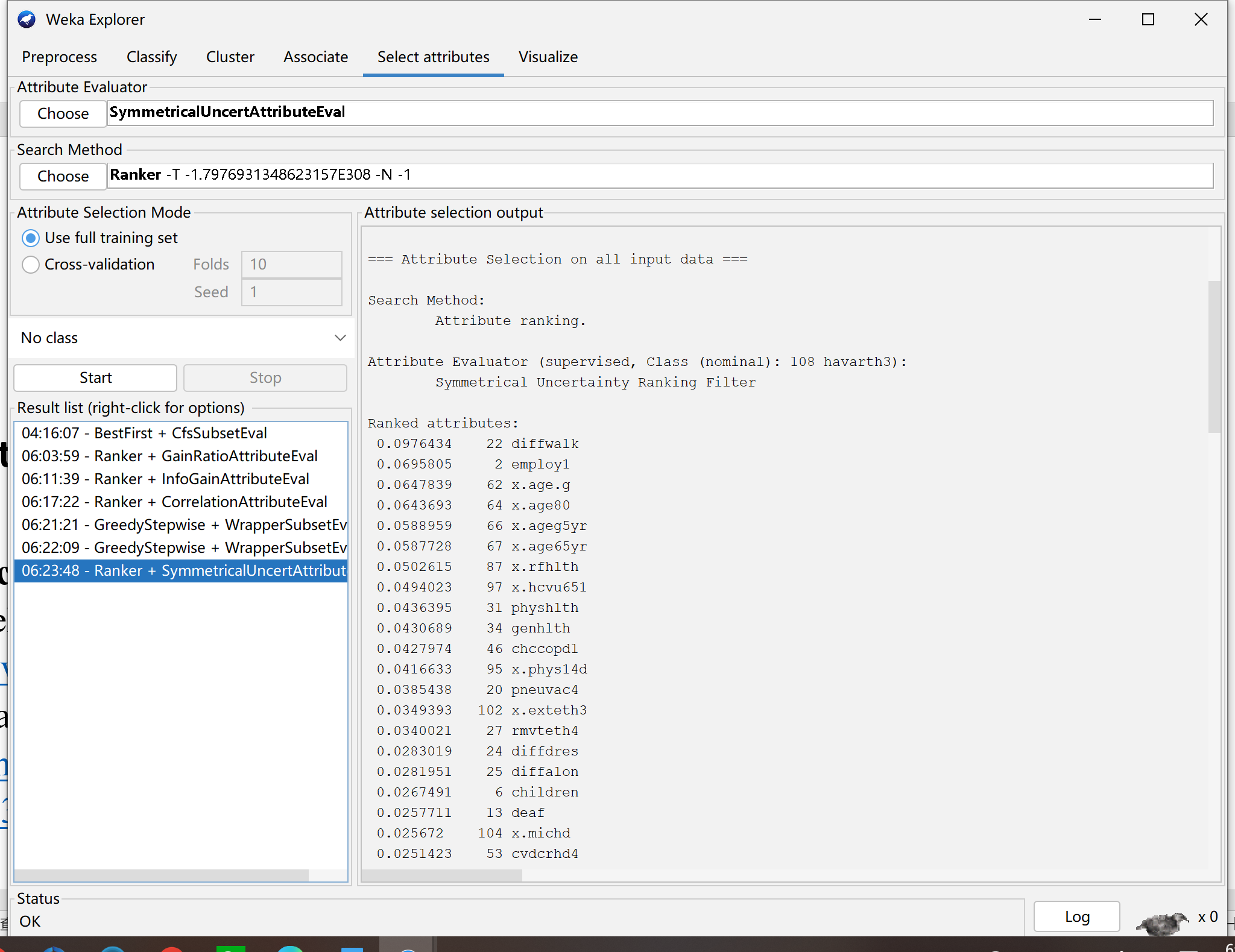
Search Method : **Ranker**

Rank attributes by their assessed value.

Because this method provides all properties and it ranks properties by value. Here I choose the top ten attributes to reduce the dataset.

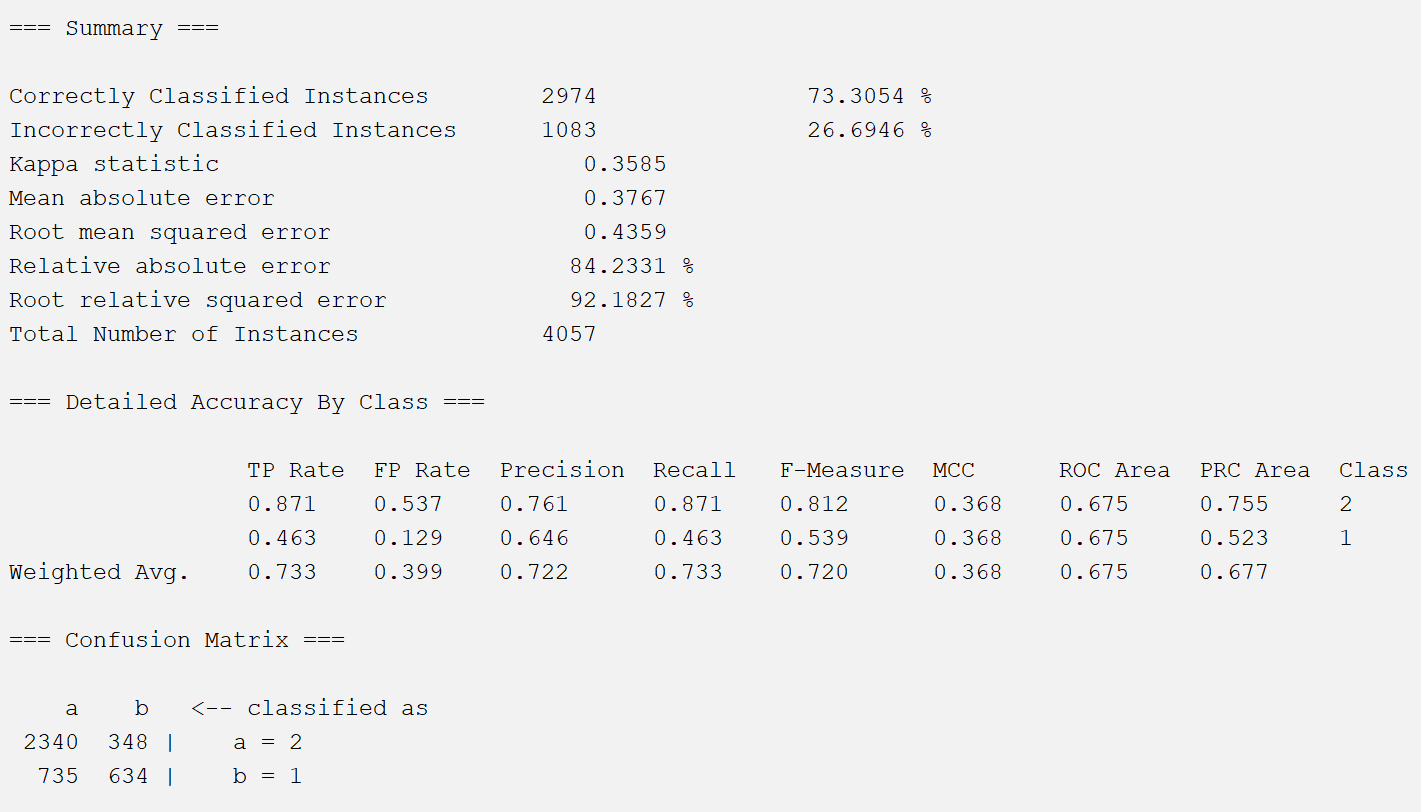
The attributes are:





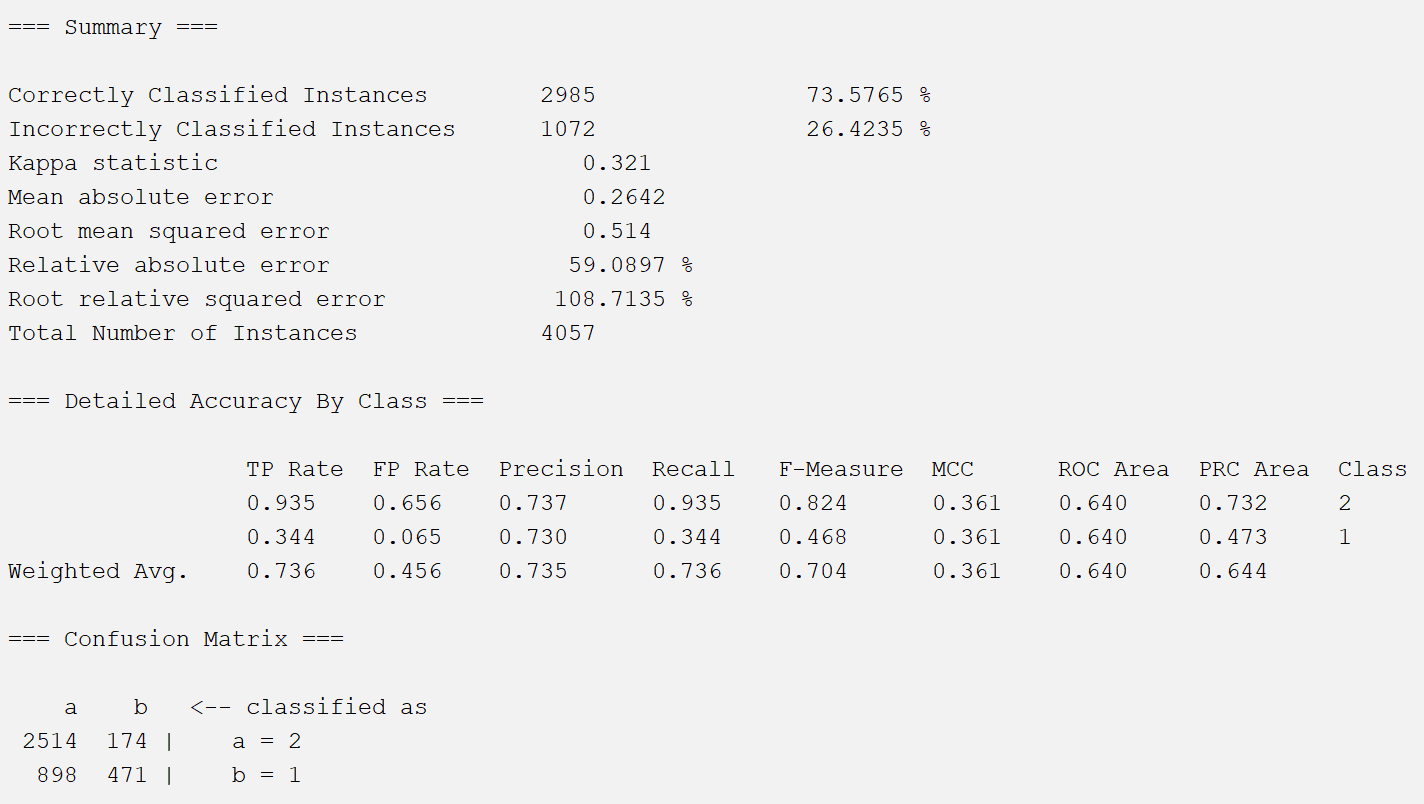
1. JRip

Generate a rule that randomly divides non-covered instances into a growing set and a pruning set, specifying that each rule in the rule set has two rules to generate



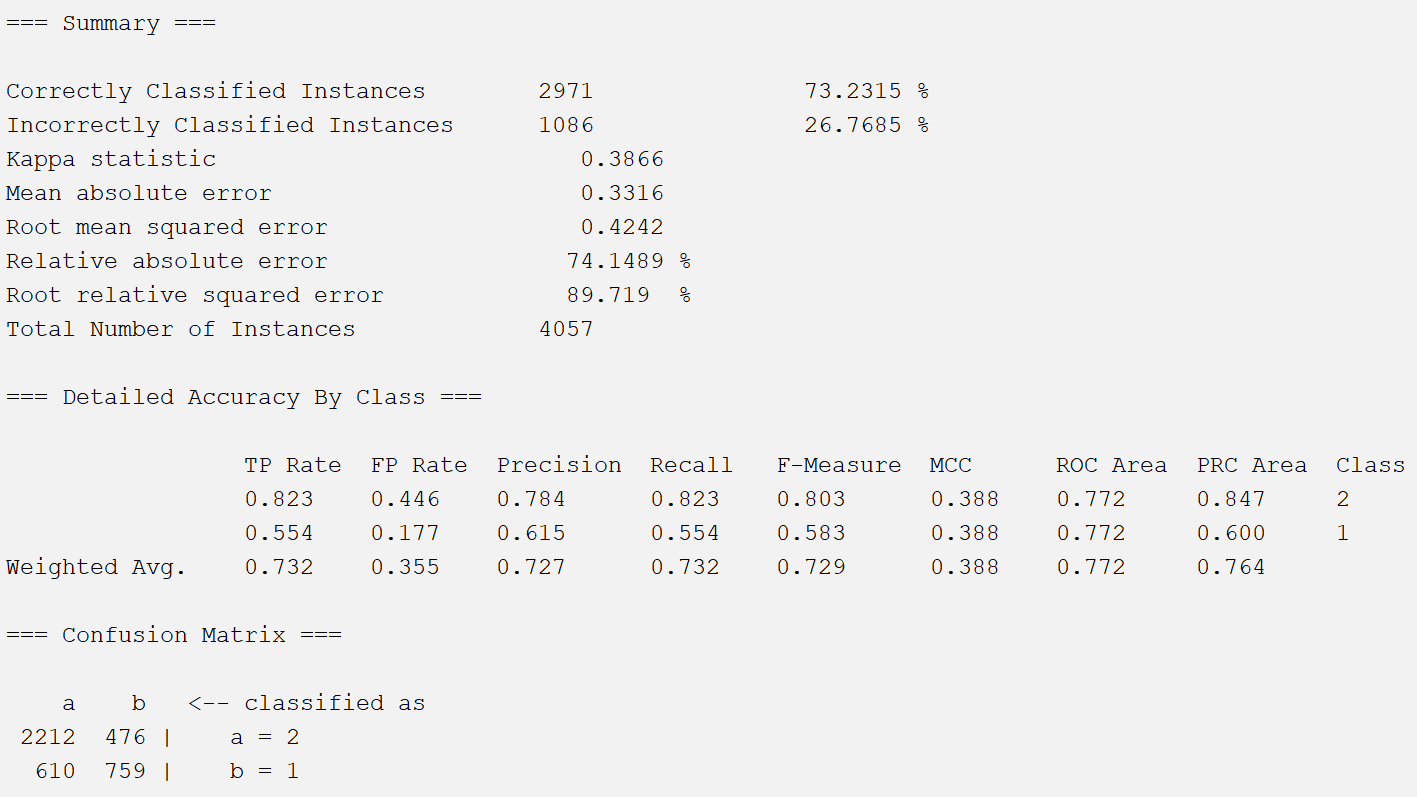
1. OneR

Find the feature with the best classification effect through the training set, use it as the classification basis, and use other features to calculate the error rate to realize the classification algorithm.



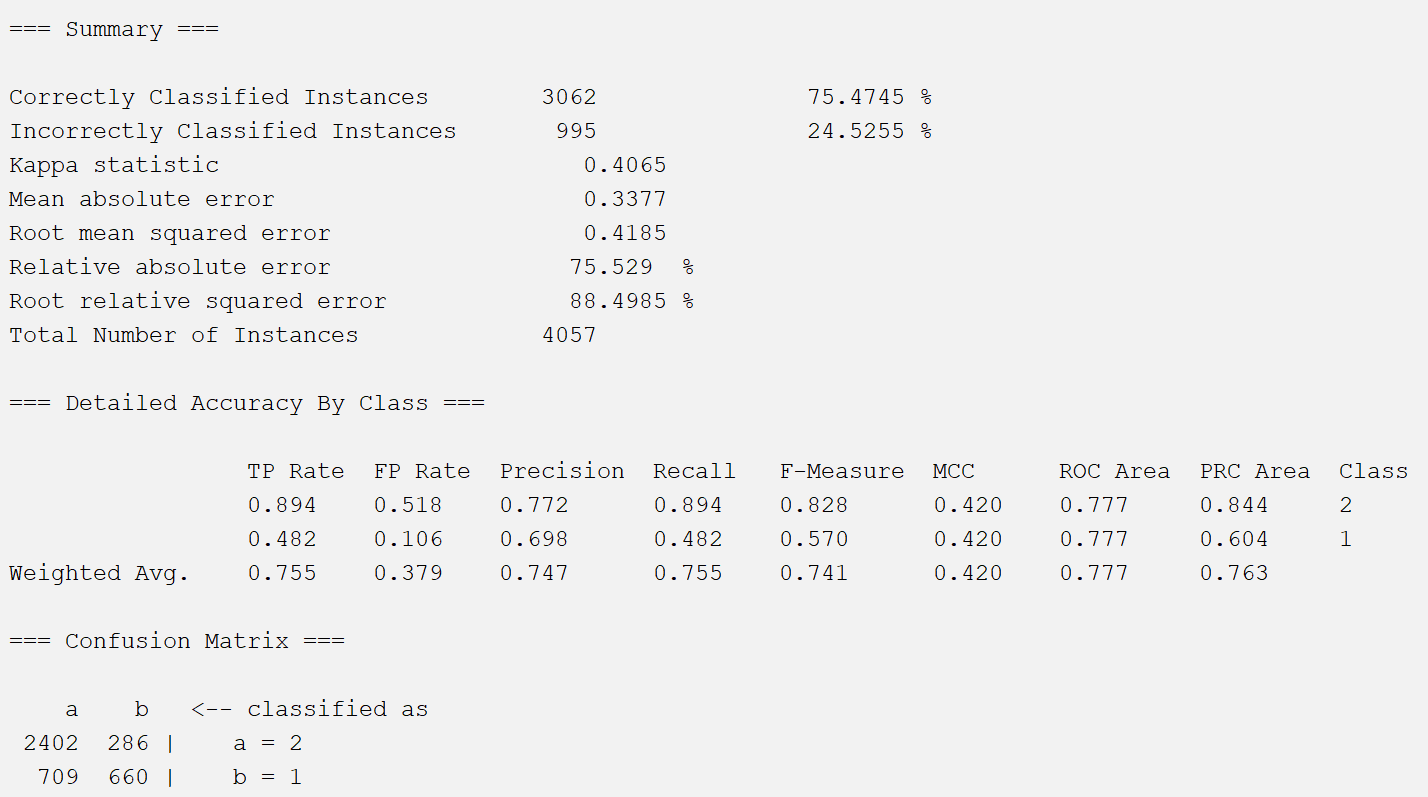
1. PART

Create a rule to remove the instances covered by the rule, and then recursively create rules for the remaining instances until there are no remaining instances.



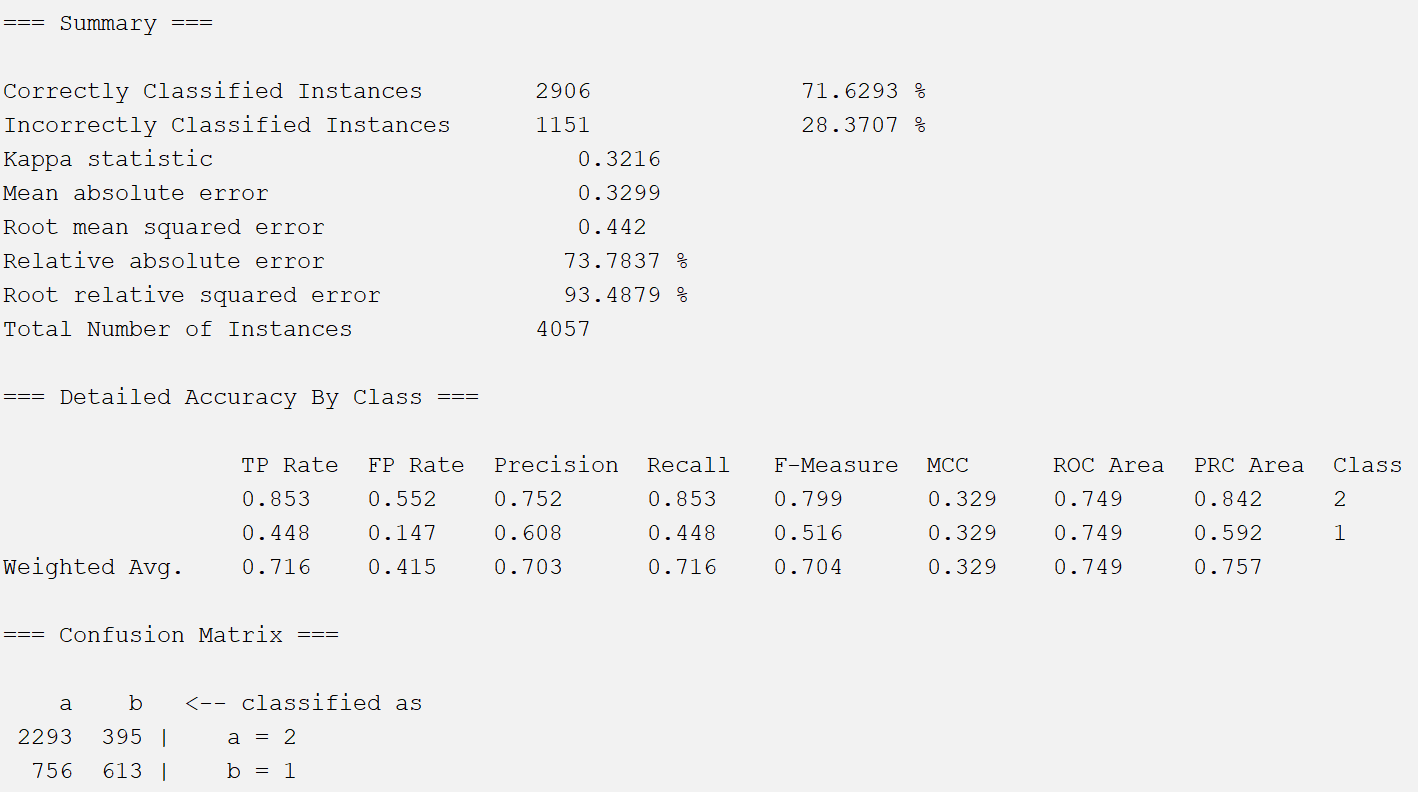
1. J48

Based on top-to-bottom strategy, recursive divide-and-conquer strategy, select an attribute to place at the root node, generate a branch for each possible attribute value, divide the instance into multiple subsets, each subset corresponds to a branch of the root node , then recursively repeat the process on each branch. Stop when all instances have the same classification.



1. RandomForest

Random forest is composed of many decision trees, and there is no relationship between different decision trees. When we carry out the classification task, new input samples enter, and each decision tree in the forest will be judged and classified separately. Each decision tree will get its own classification result, which one of the classification results of the decision tree is classified. At most, then the random forest will treat this result as the final result.



**Conclusion**

According to all of the results of 25 classification models. Since in the five attribution selections, the number of 1s and the number of 2s in the class are quite different, we mentioned in the class discussion that ROC uses FP / (TN + FP), ROC will be affected by both FP and TN, so ROC is not accurate enough when the data differs greatly, so we ignore the value of ROC Area for this dataset, instead, we focus on PRC.

If we only look at the total accuracy, the top three are CfsSubsetEval method with logistic regression 75.26%, InfoGainAttributeEval with MultiClassClassifier 75.03% and SymmetricalUncertAttributeEval method with J48 75.47%. J48 has highest precision, recall, F-measure and MCC. Precision shows that J48 predicts the largest number of true positive samples TP/(TP+FP) among the positive samples, while Recall shows that TP/(TP+FN) positive samples are successfully predicted to be the largest number of positive samples. Based on the formula of F-measure, F-measure combines the results of precision and recall. The higher the F-measure, the more effective the test method is.

Thus, SymmetricalUncertAttributeEval attribute selection method with a J48 algorithm gave me the best performance.

**References:**

CDC. Behavioral Risk Factor Surveillance System

<https://www.cdc.gov/brfss/annual_data/annual_2018.html>

Blackboard. how-to-split-dataset-to-training-and-test.docx.

<https://onlinecampus.bu.edu/bbcswebdav/pid-10341678-dt-content-rid-70620893_1/courses/22sum1metcs699so1/course/syllabus/Supplements/stratified-split.pdf>