

**DATAMINING**

**FINAL TERM PROJECT**

**OPTION 1**

**PRIYANTH DASARI**

NJIT ID: 313-64-126

UCID: pd272

Email:pd272@njit.edu

Computer science Department

**DATA MINING**

**FINAL TERM PROJECT REPORT**

**Aim of the Project:**

The project is to perform a supervised data mining i.e. classification on a data set using two algorithms.

The algorithms which I chose are:

* J48
* Naïve Bayes

Tool used:

* Weka for both the algorithms

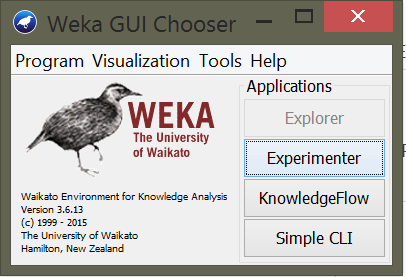
The data set I, preferred is

Program execution steps:

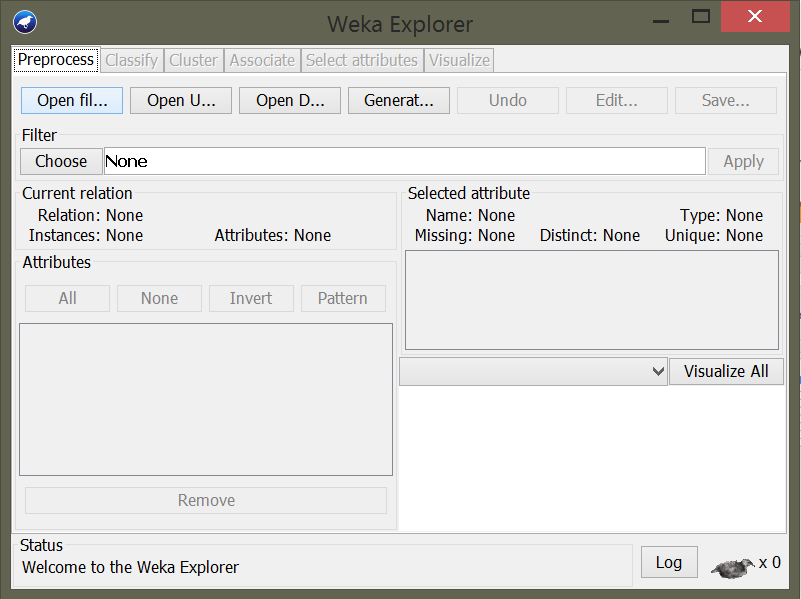
1. First save the data set in a notepad and save the file with a name you prefer with .arff extension so that you can open the file with weka
2. You can also save the data in excel in .csv format and export it to weka
3. Open the weka tool
4. Then click on Explorer button
5. The application open up for you
6. Now click on the open file button in the preprocess tab and chose the data set from your Pc.
7. Now go to the classify tab and you can chose a classifier you want
8. Here in this case I chose J48 it under the tree tab
9. Now chose the use training set option under the test options
10. Now click on the start button and your test starts and the results are displayed on the right side of the window
11. Repeat the same procedure by choosing the Naïve Bayes algorithm and you the results are displayed on the right side of the window.
12. Now chose the option Cross-validation using 10 fold option under test options and then click on the start button to see the classification results.
13. To get experimental results that show the comparison of classification accuracies between the two classification algorithms you need to go the experimenter tab when you open the tool
14. Now click on the new button to start a new project
15. Now click on Add new button under data sets option and add the data you prefer.
16. Keep the number of folds to 10 under cross validation tab and number of repetitions to 1 under iteration control tab.
17. Now click on the Add new button under Algorithms tab so that you can chose the J48 and naïve bayes algorithm for the test.
18. Now go to the run tab and then click on start button in the run tab.
19. The test begins to start and it displays if there any errors.
20. After the test is completed go to the analyze tab and then click on the experiment button.
21. Now you need to configure the test options as you prefer and click on the perform test button so that the result is displayed accordingly on the right side of the window.

**Execution and implementation of the project with pictures:**

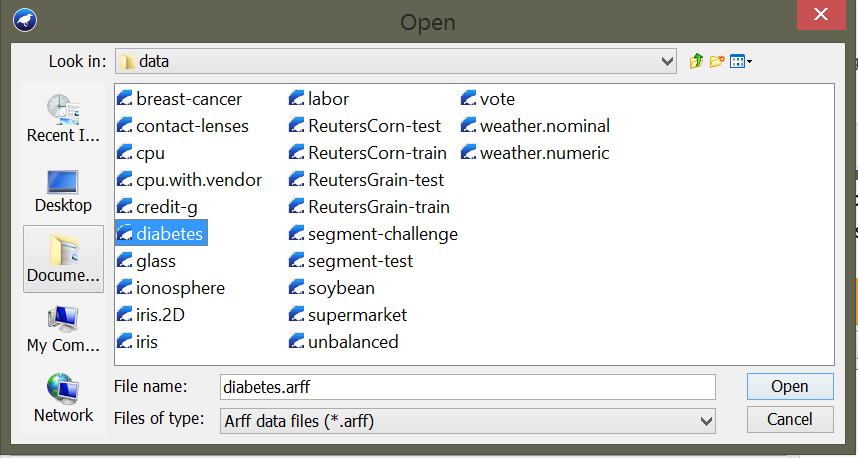
1. My data set is diabetes.arff
2. Simple take a data set in notepad separated by comma symbol and save the file with .arff extension.
3. Open the weka application and click on the explorer button



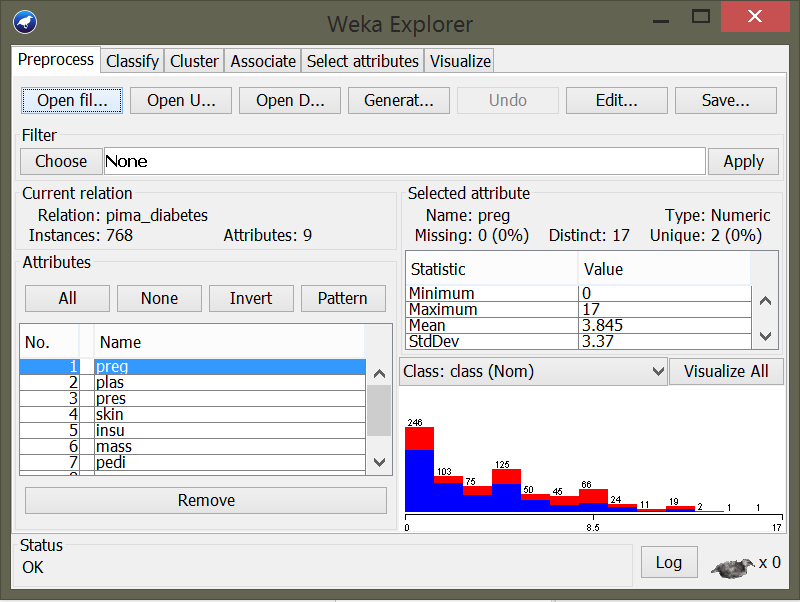
1. A new explorer window opens up as shown below:



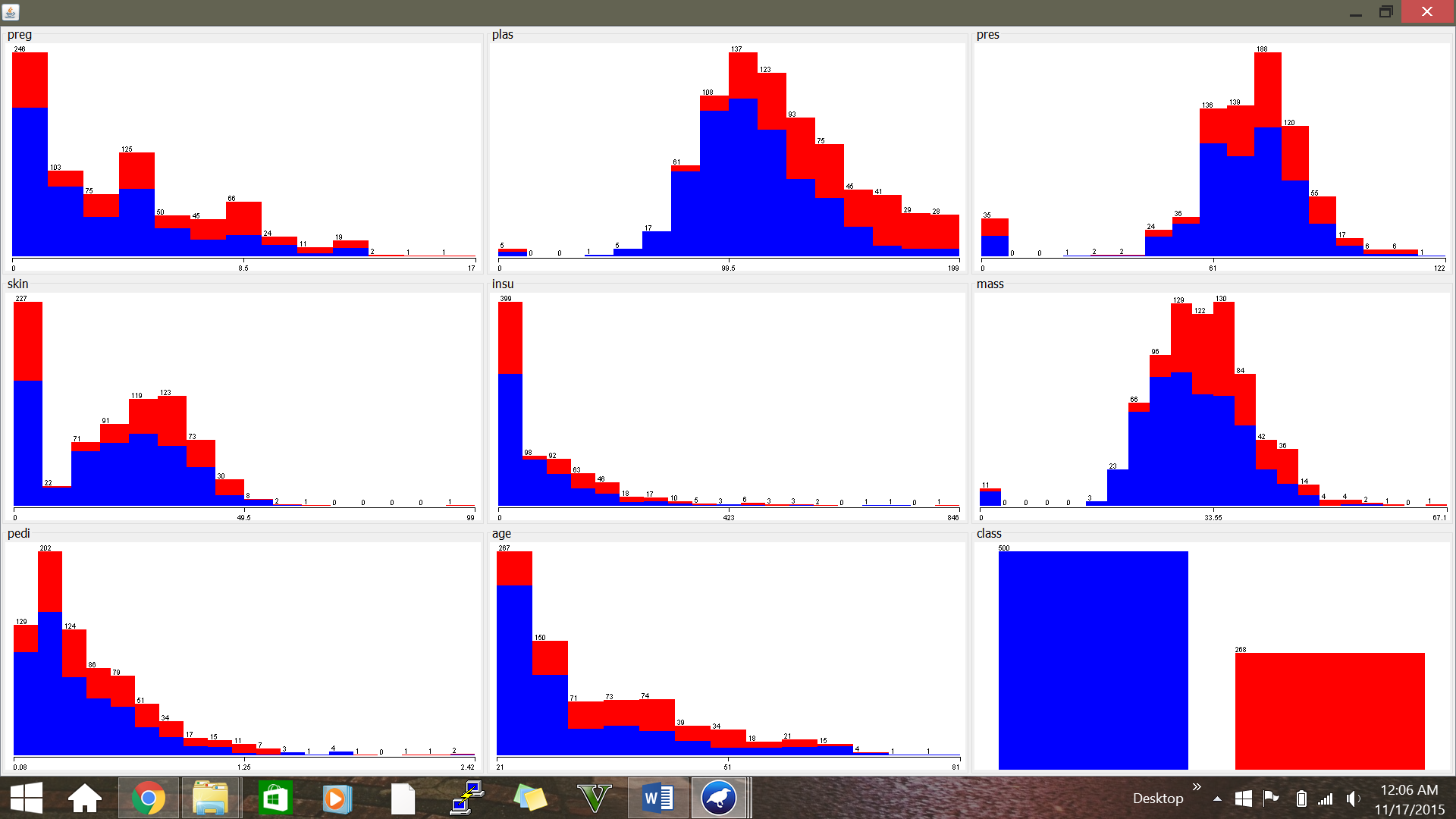
1. Now in the preprocess tab click on the open file menu then dialog window opens up asking the user to select the desired data set as shown in the below picture



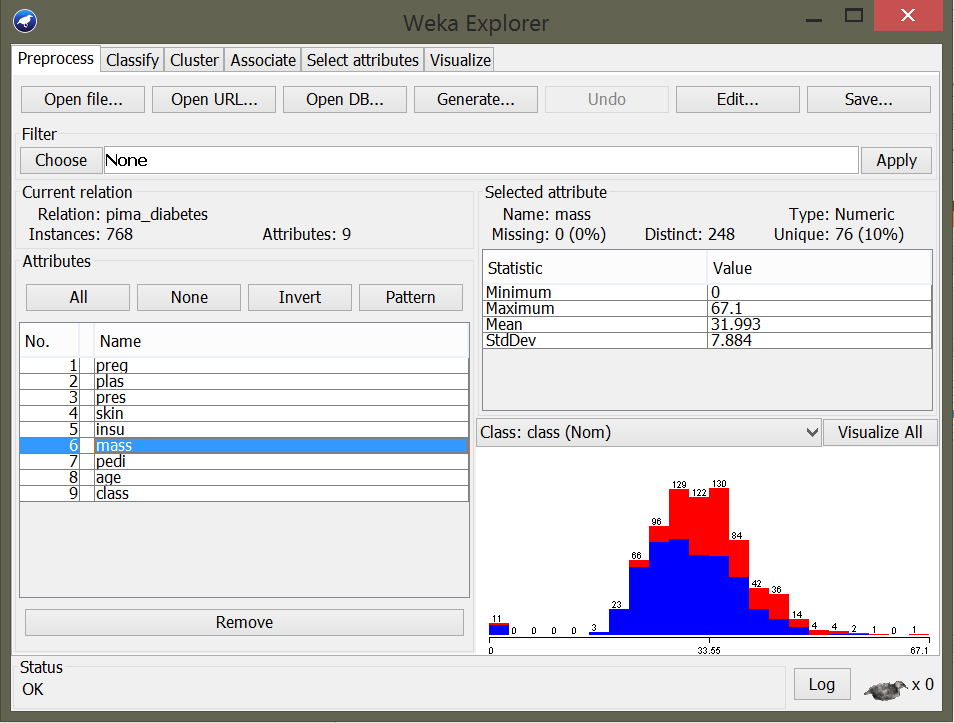
1. When you select the desired data set, it is loaded into the weka tool as shown in the below picture



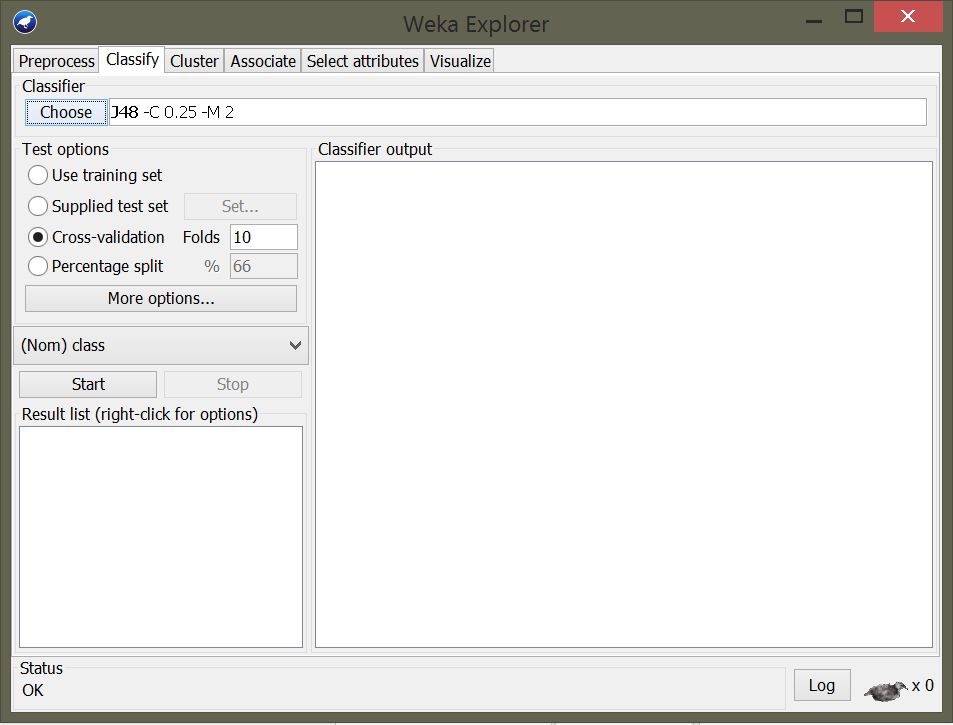
1. Now if you click the visualize all button you can view all the bar graphs of all the attributes as shown below



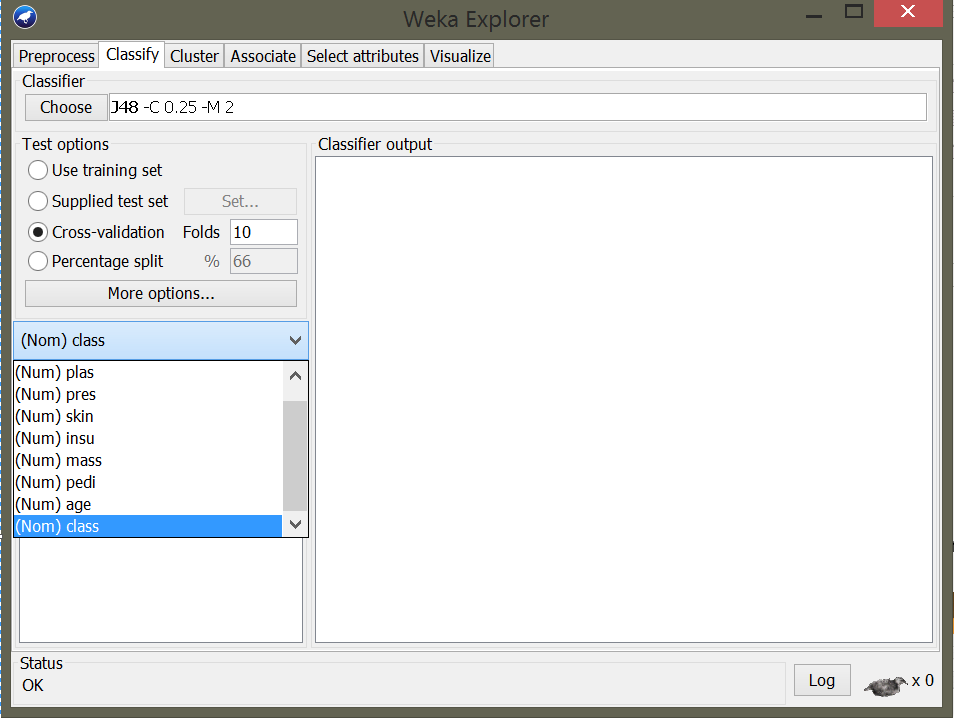
1. If you want to visualize a specific class attribute you can select the attribute on the left side of the window as shown below, as of now I ‘ve selected the attribute mass



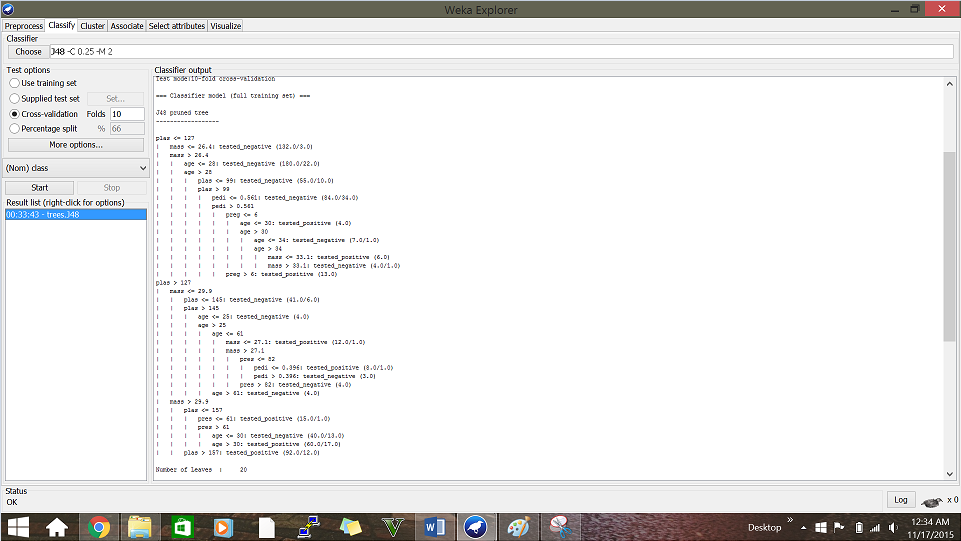
1. Now go to the classify tab and choose a classifier. Now click on choose button and select the J48 Classifier algorithm under tree

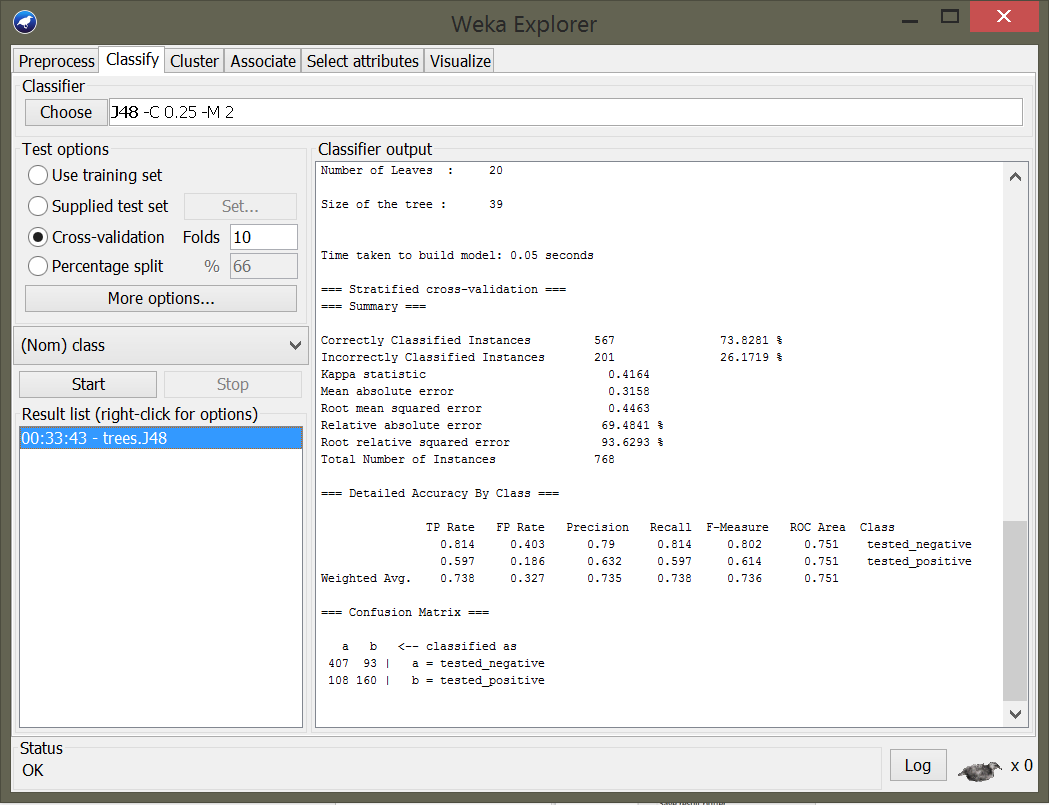


1. Now select the use training set option under the test options and select the nom class in drop down menu and click on start button to start the J48 classification.

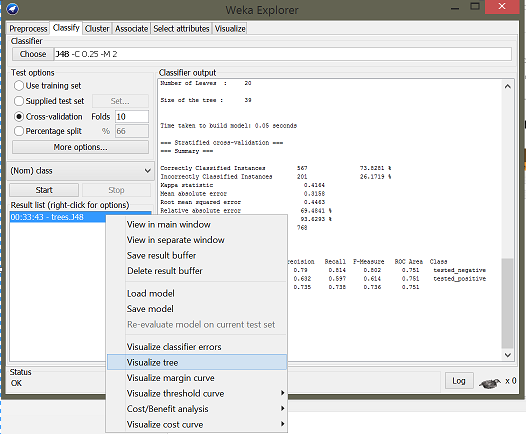


1. After the classification is done you can see the results on the right side of the window and the results displays instances, attributes, J48 pruned tree, no. of leaves, size of tree, evaluation of training set, accuracy of class and confusion matrix

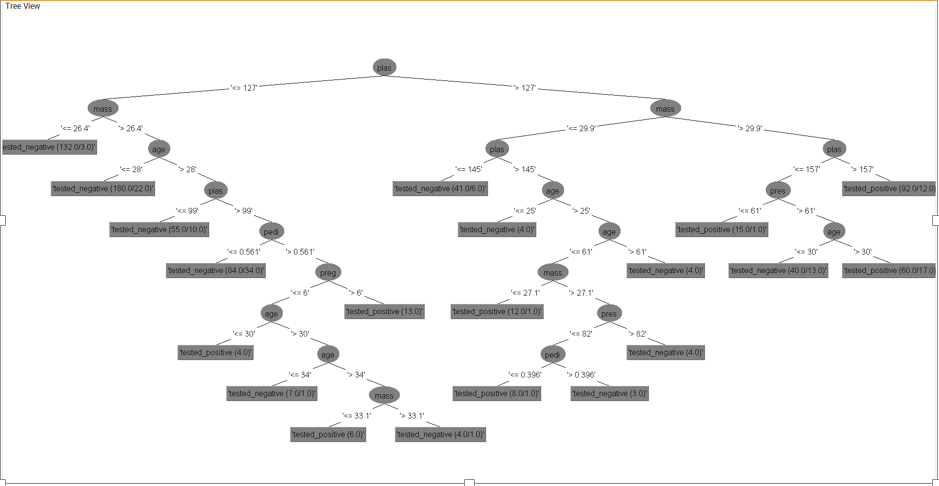




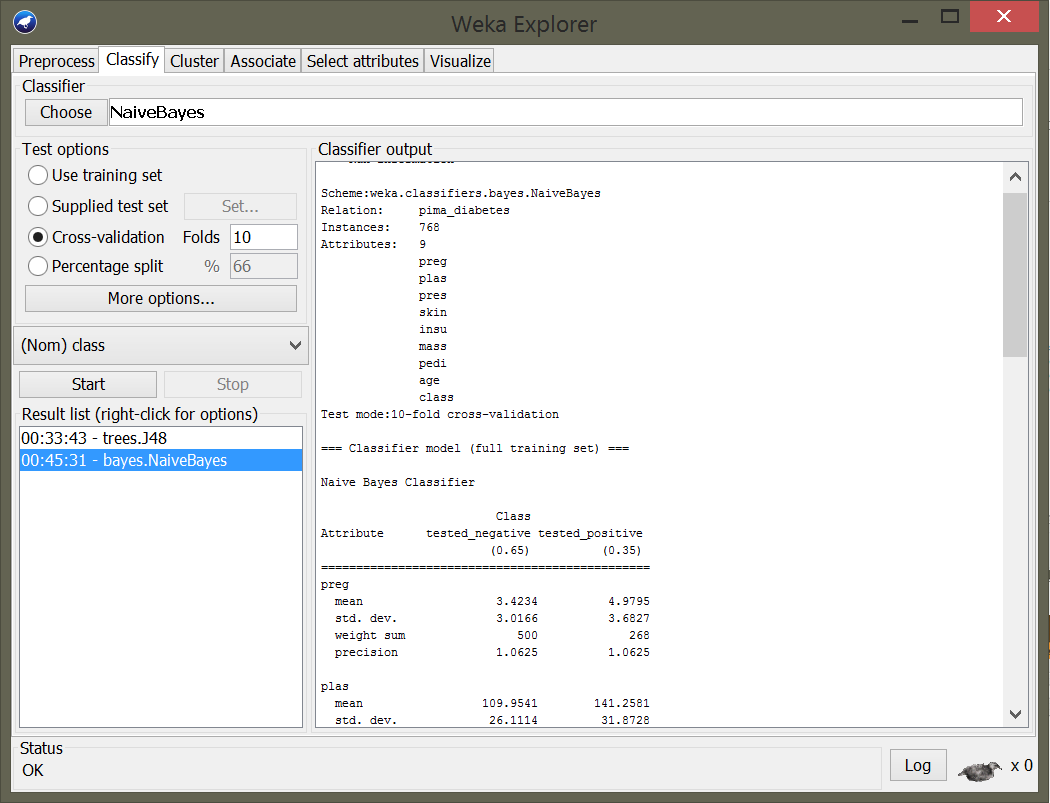
1. Right click on the tree J48 under results list and click on visualize tree as shown below

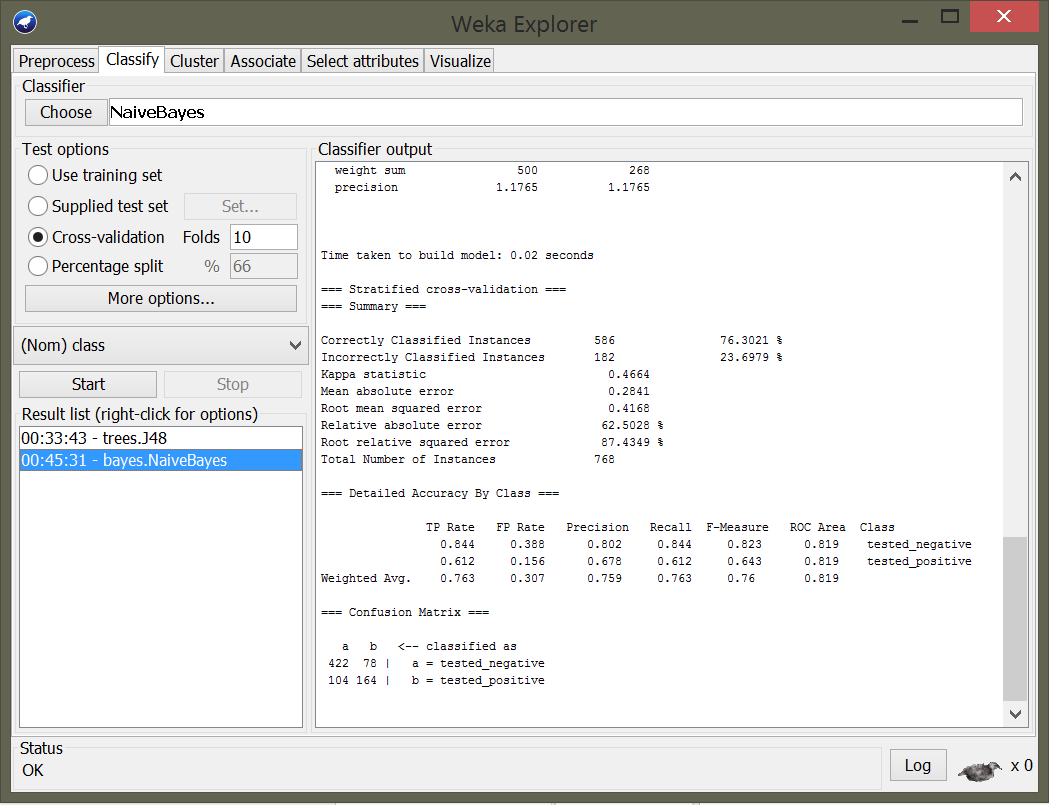


1. You can visualize the J48 tree as shown below

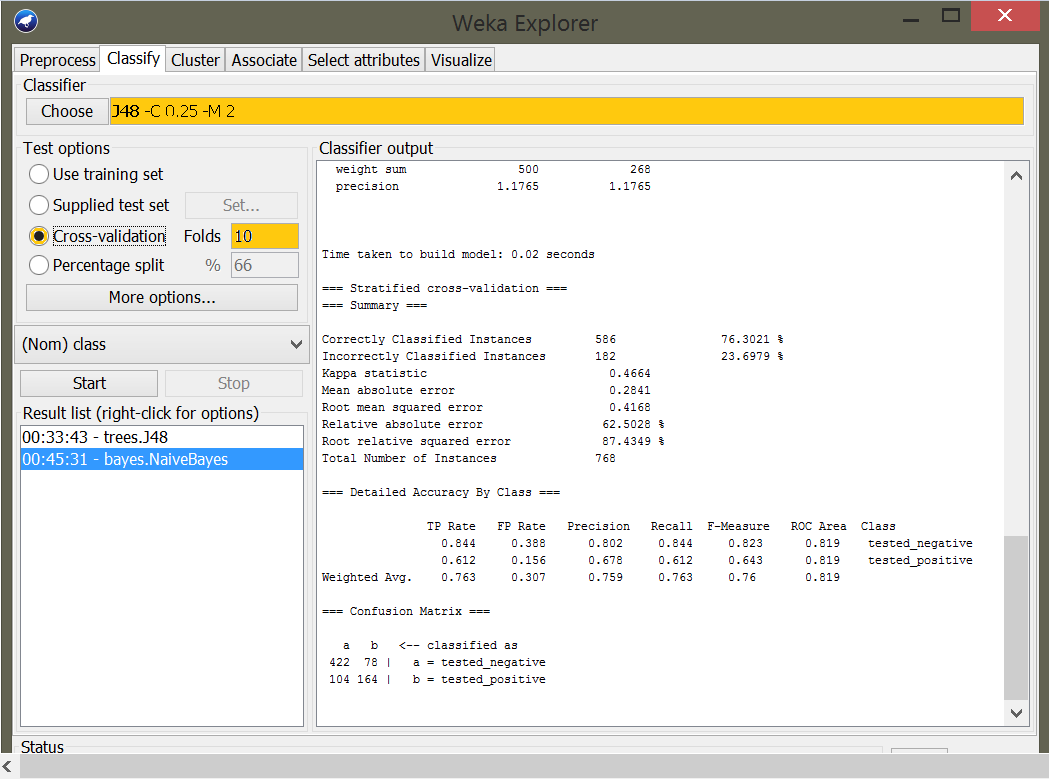


1. Now choose naïve bayes classifier under bayes as shown below and repeat the same procedure to get the classification results of naïve bayes classification algorithm as shown below

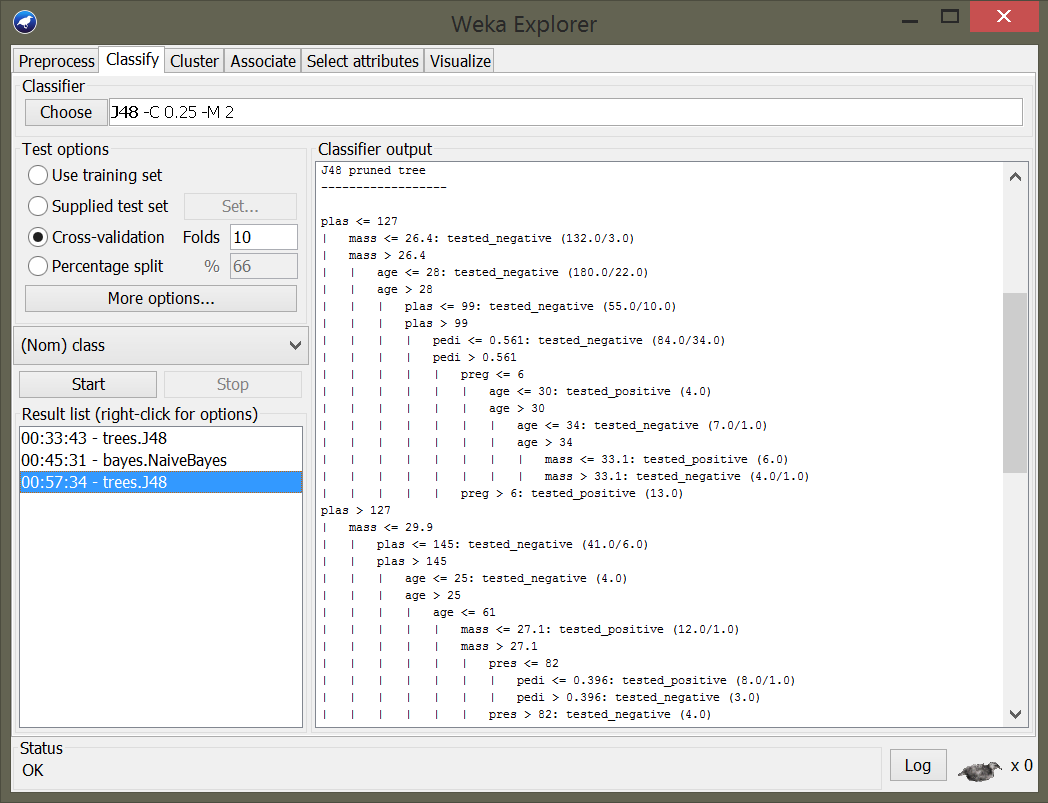


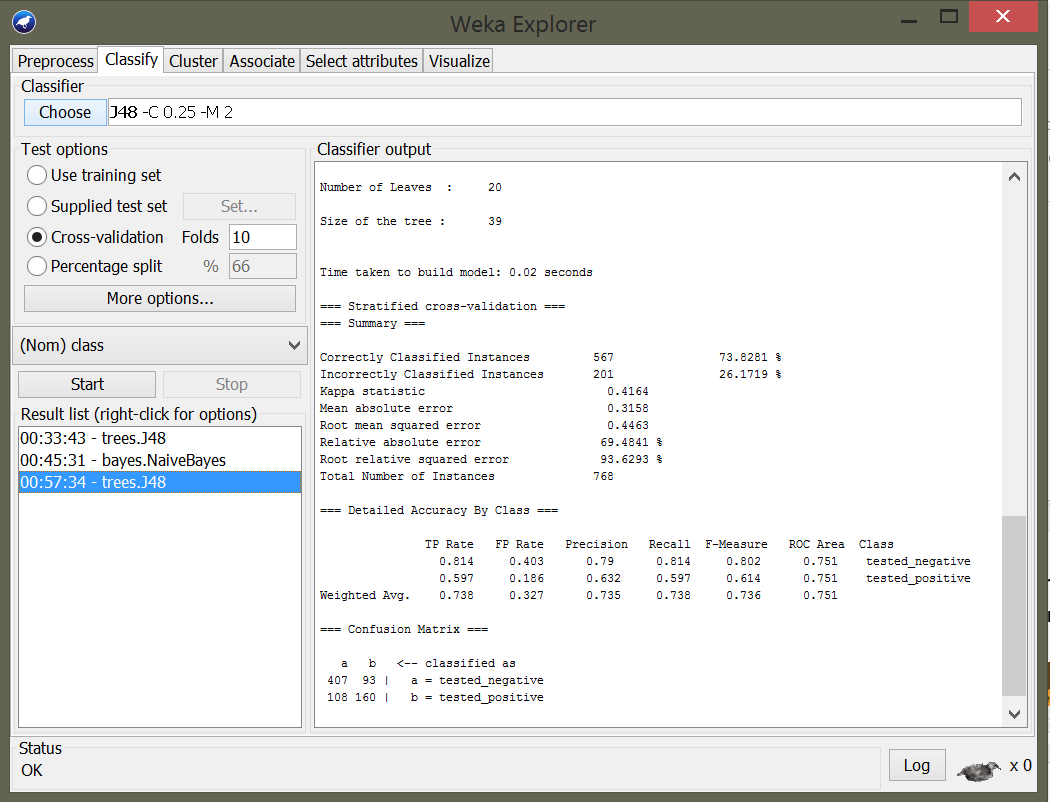


1. Now we need to calculate the 10 fold – cross validation the J48 classification algorithm, For that to happen first we need to choose the J48 classification algorithm and then select the Cross- validation option under test options with 10 folds and then click on start button to see the results of 10 fold cross - validation of J48 algorithm.

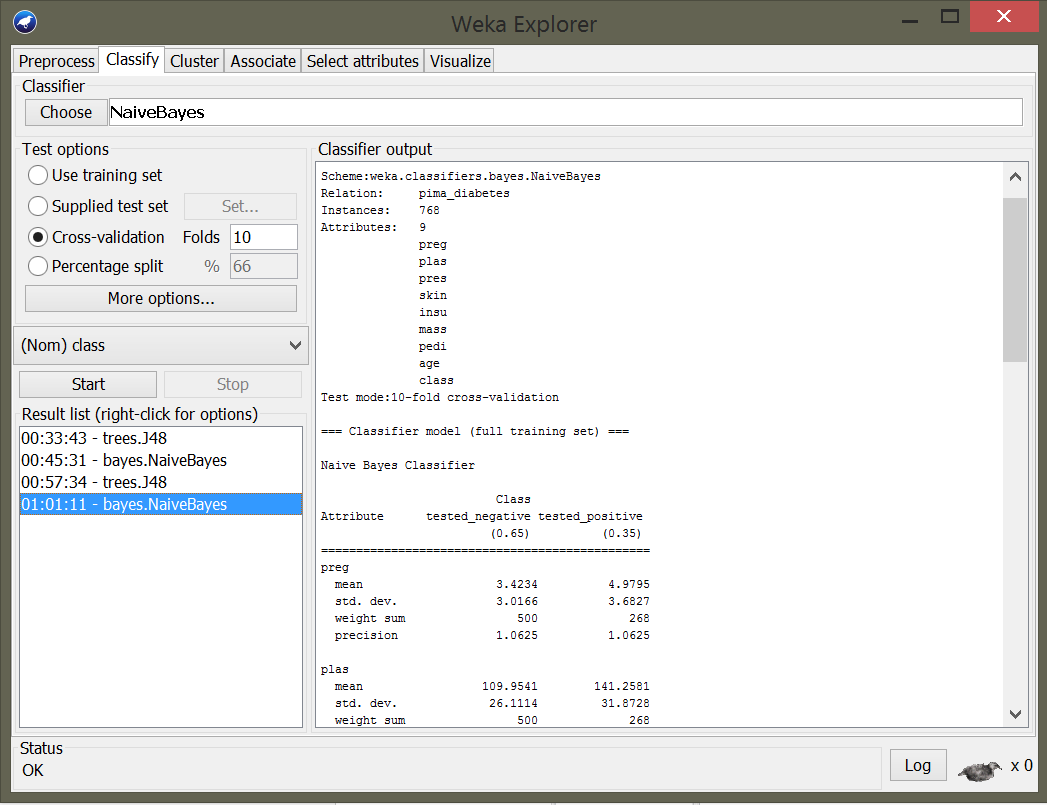


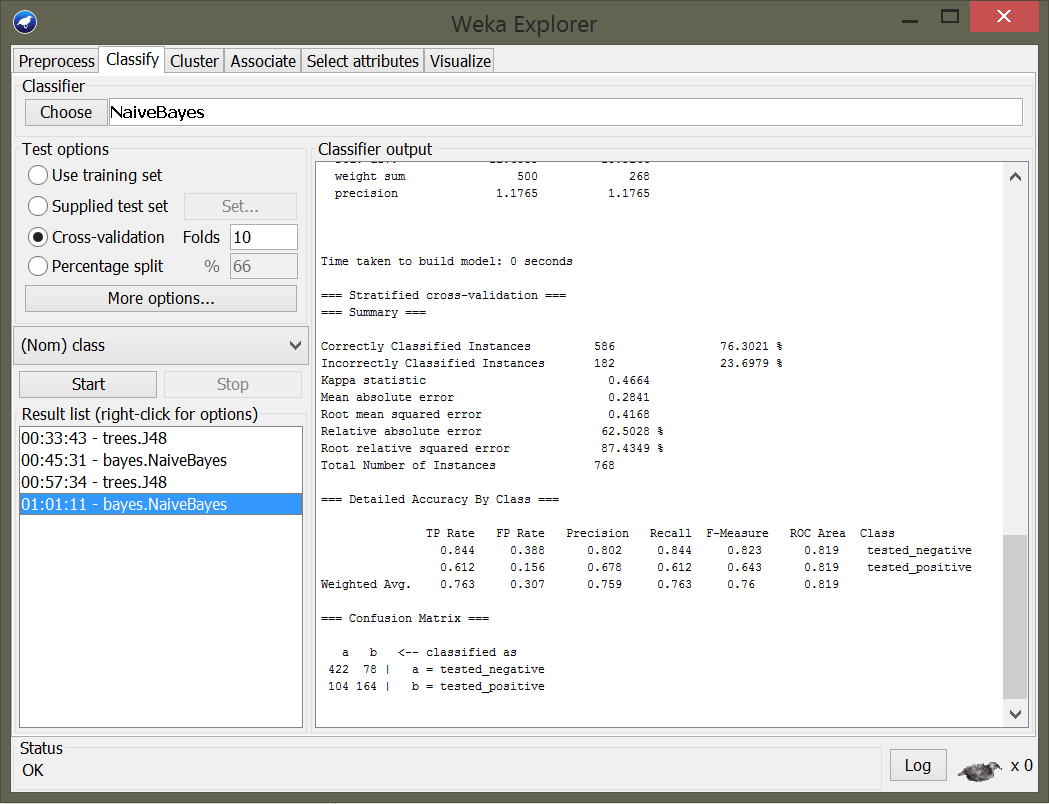
1. Now If we click the start button the results for 10 fold cross - validation of J48 algorithm is shown on the right side of the window as below





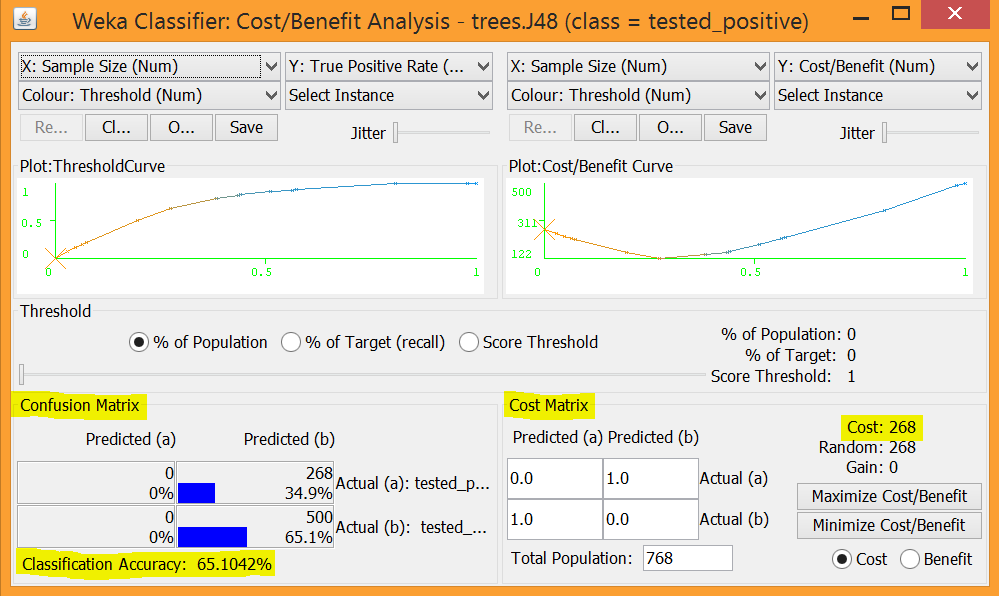
1. Now choose the naïve bayes classification algorithm and click on the start button to see the results of naïve bayes classification algorithm on right side of the window as below



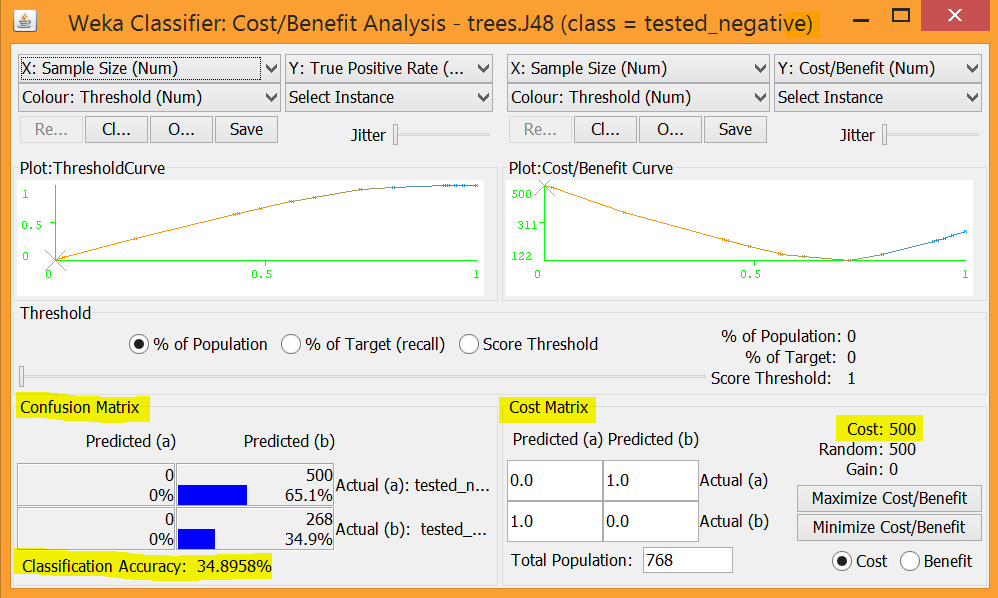


**Comparison of classification accuracies using ROC and AUC between the two classification algorithms ­­­­­­**

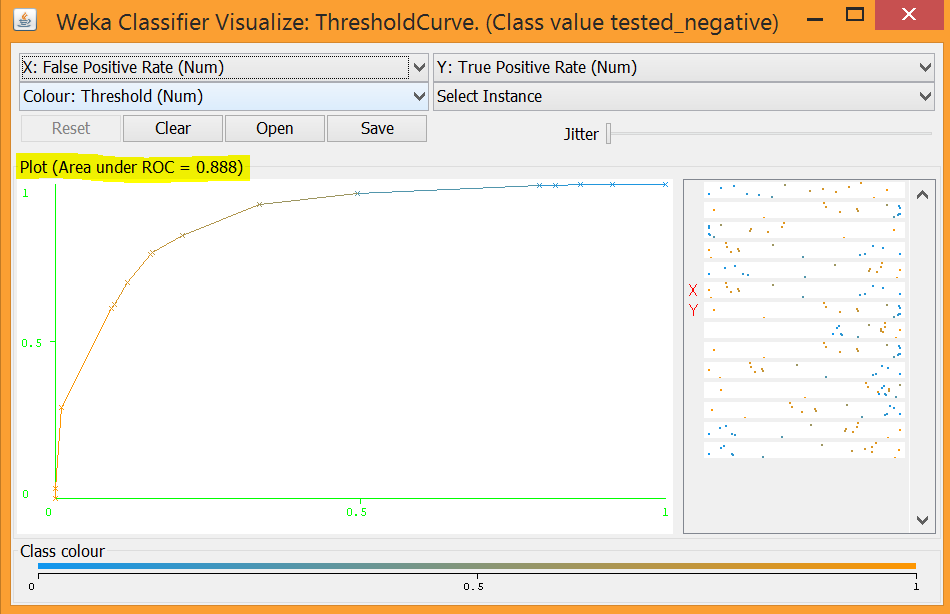
1. Cost/benefit analysis of J48 cclassification algorithm tested positive



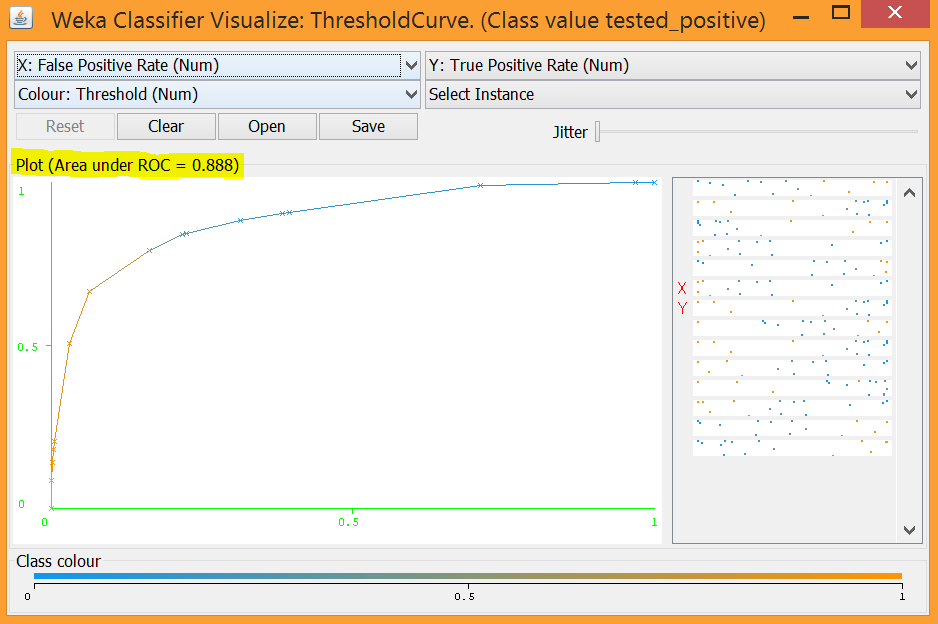
1. Cost/benefit analysis of J48 cclassification algorithm tested negative



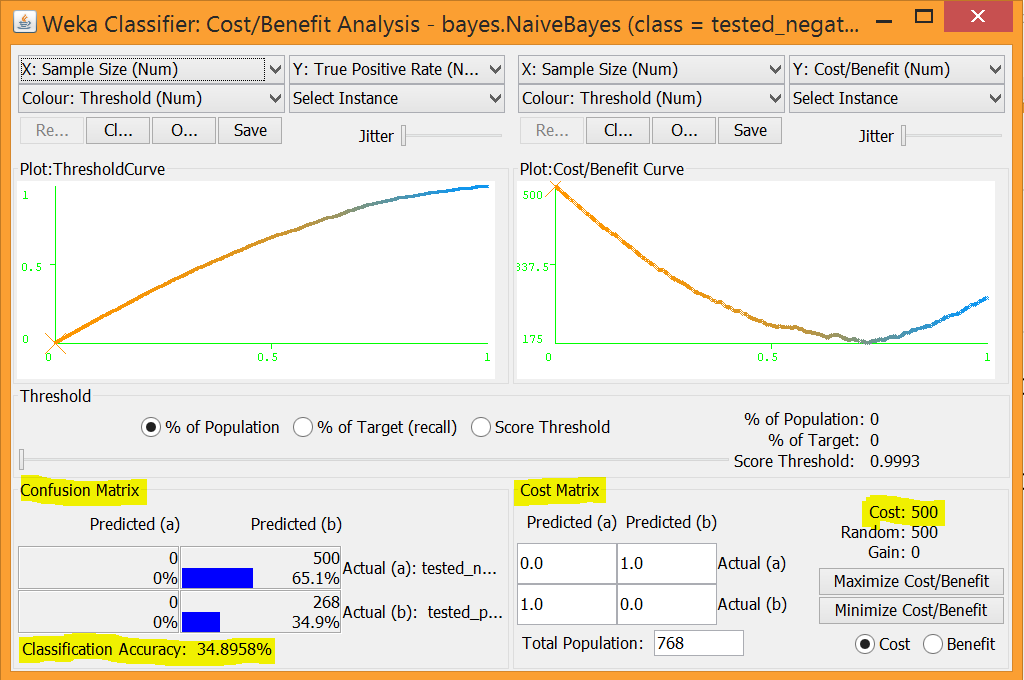
1. Threshold curve / Area under ROC for J48 classification algorithm tested negative



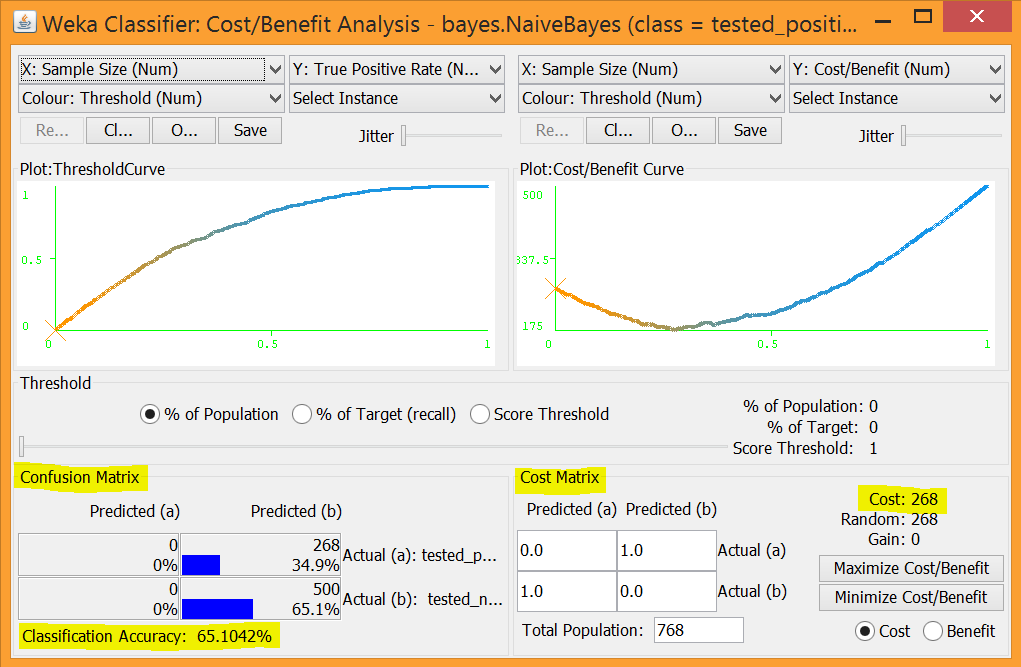
1. Threshold curve / Area under ROC for J48 classification algorithm tested positive



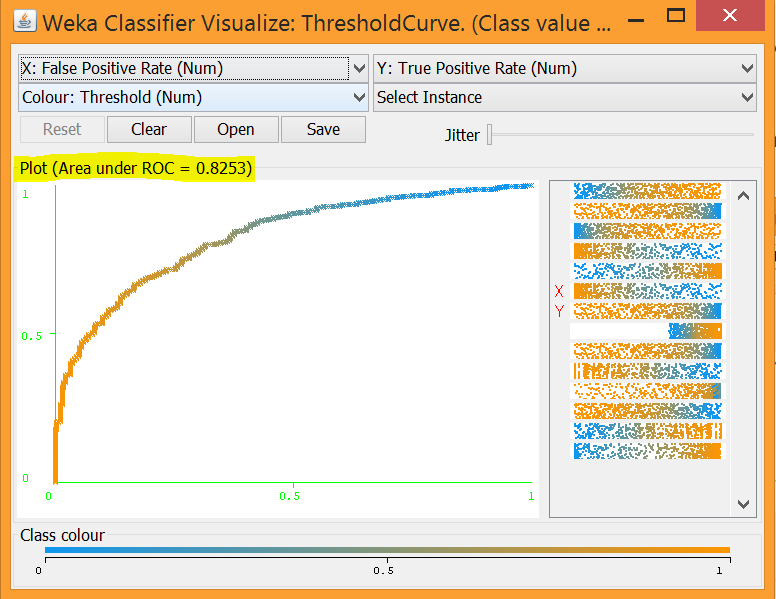
1. Cost benefit analysis of naïve bayes classification algorithm tested negative



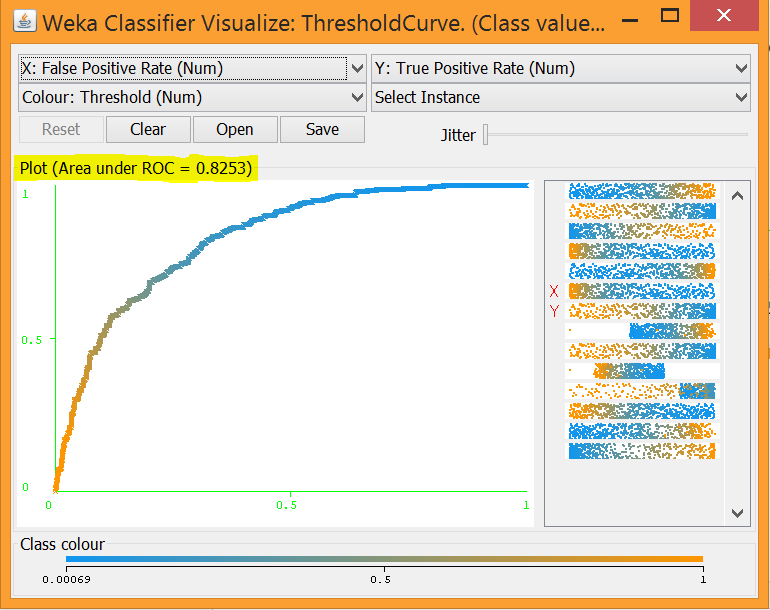
1. Cost benefit analysis of naïve bayes classification algorithm tested positive



1. Threshold curve / Area under ROC for naïve bayes classification algorithm tested negative

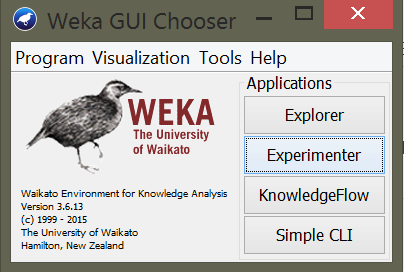


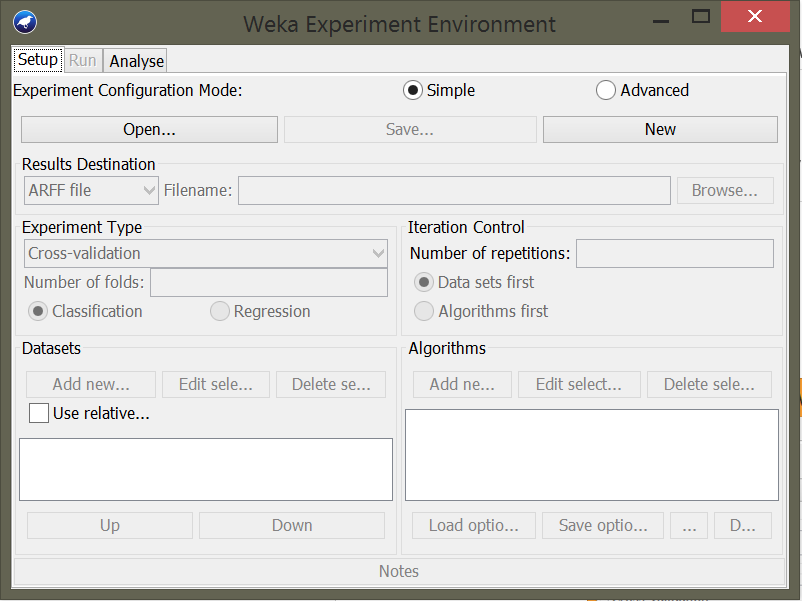
1. Threshold curve / Area under ROC for naïve bayes classification algorithm tested negative



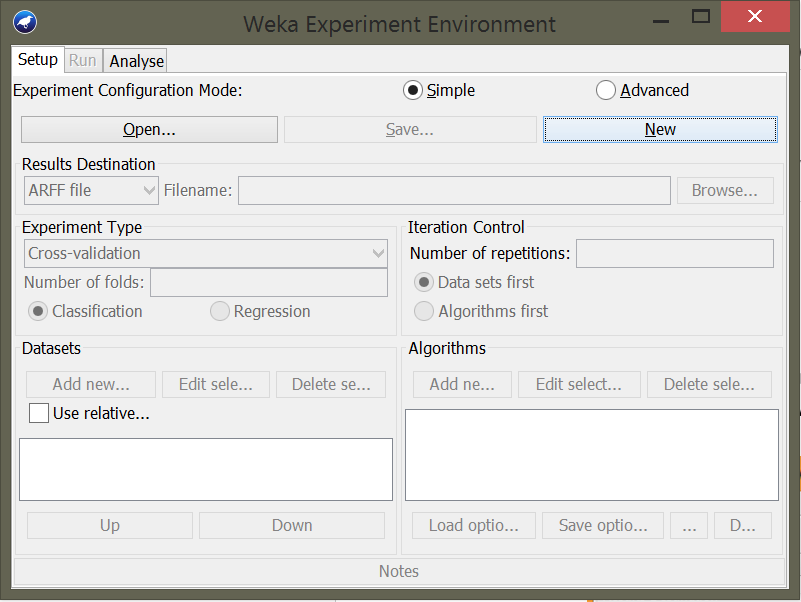
**Comparison of results of J48 and Naïve bayes classification algorithms in Experimenter**

1. You have to go the Experimenter section as shown below

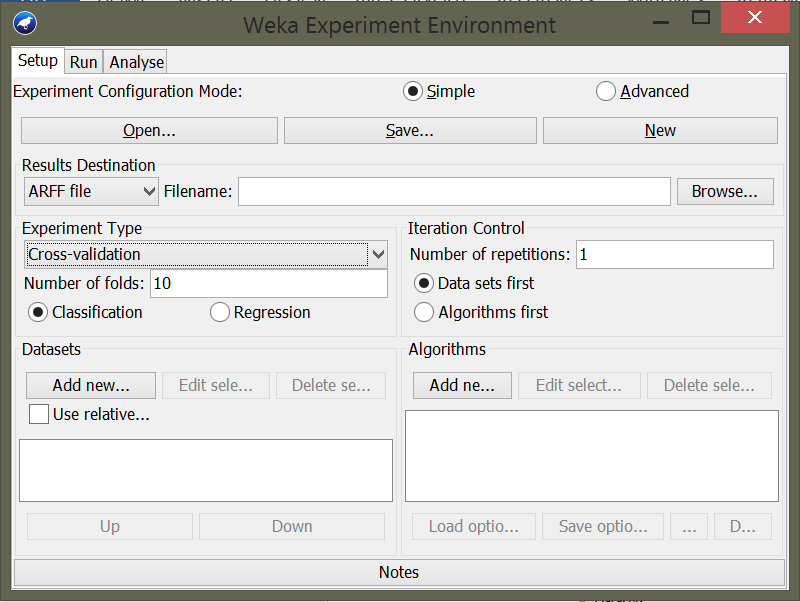




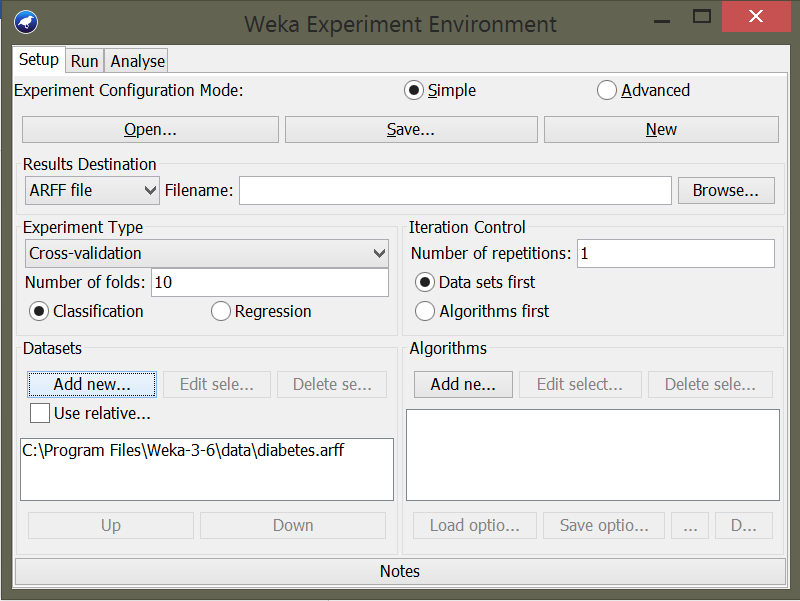
1. Now click on the new button to start a new project as shown below



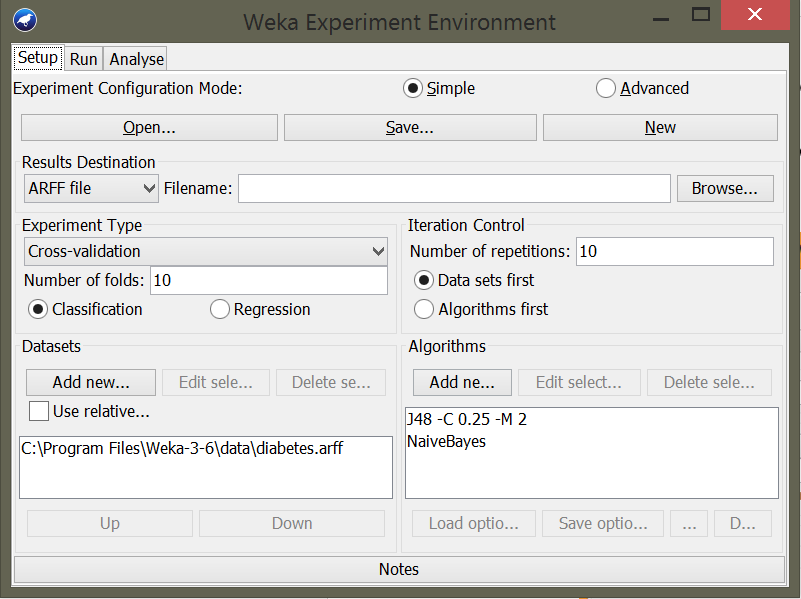
1. You have to set experiment type to 10- fold cross validation and number of repetitions to one as shown below



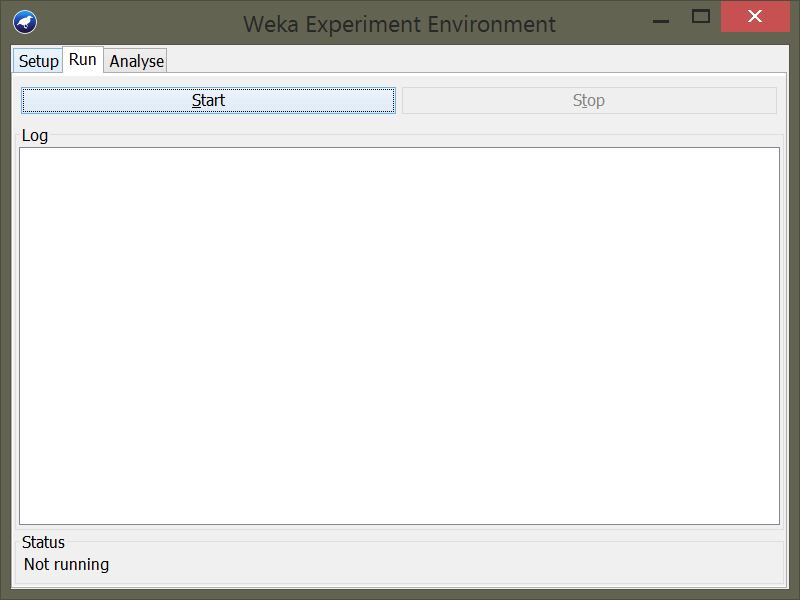
1. Now click on add new button below the data sets option to add new data sets. You can add as many data sets as you need



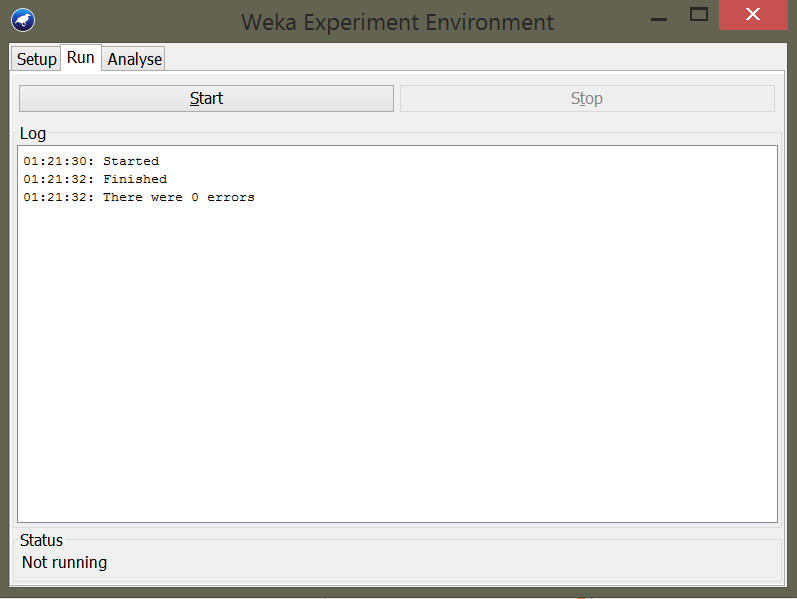
1. Now click on add new button below algorithms to add the desired classification algorithms. I’ve added naïve bayes and J48 classification algorithms as shown below



1. Now that you are ready with the set up you can go the run tab beside the setup tab and click on the start button to start the test as shown below

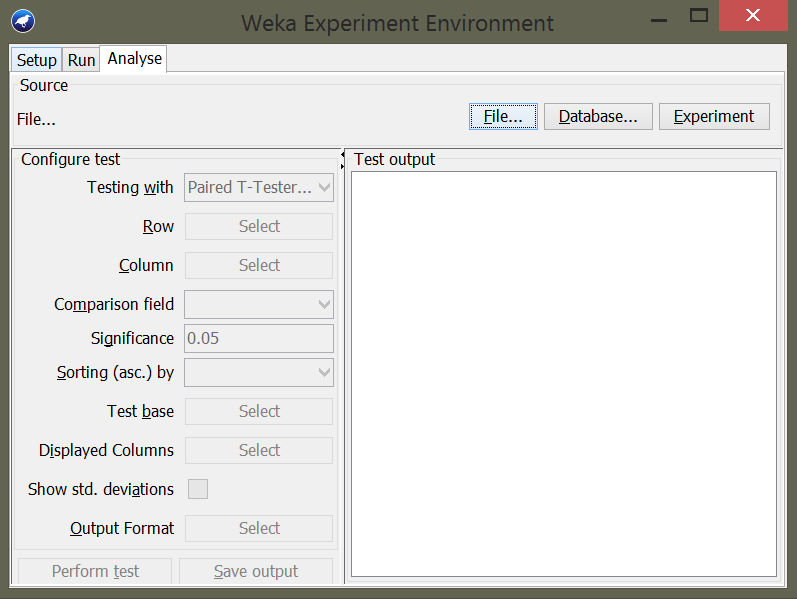


1. As you click on the start button the test starts to begin and the log results are displayed on the right side of the window

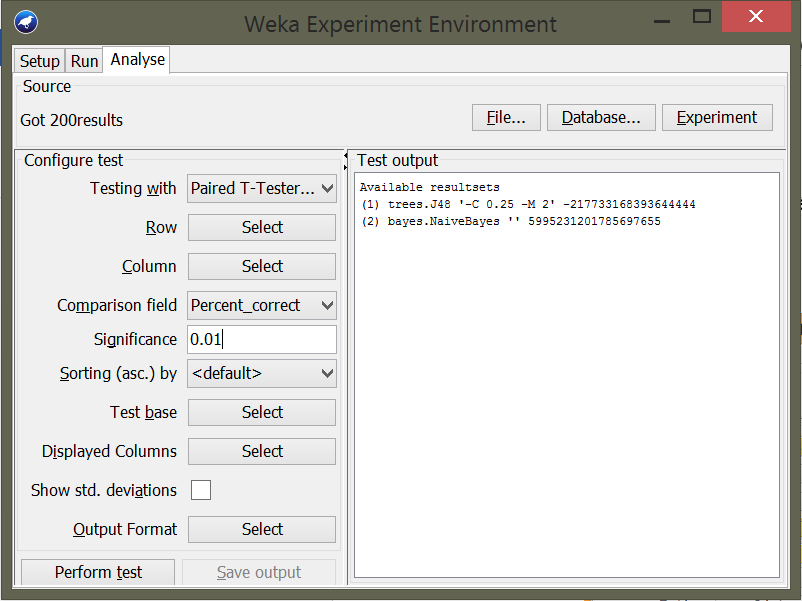


The log results are displayed showing that we have no errors so we can proceed to analyse section as there are no errors

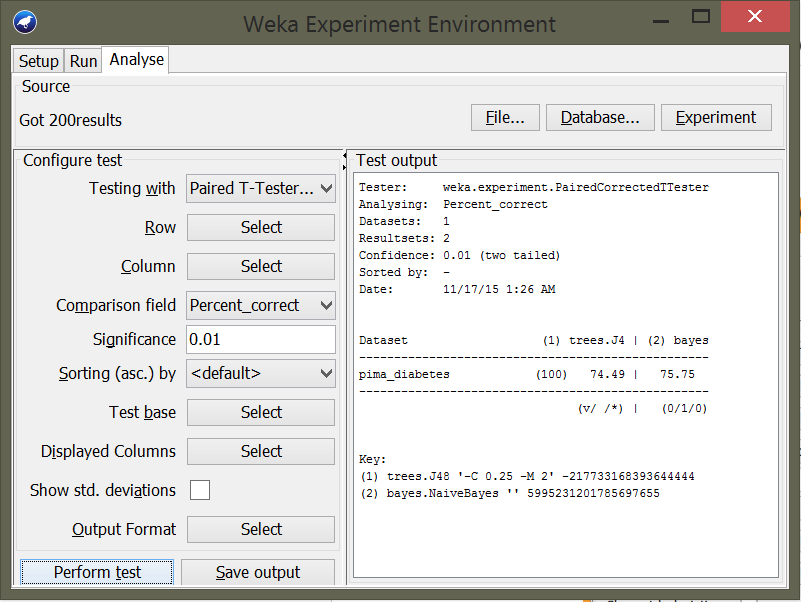
1. Now go the analyse tab beside the run tab to start the analysis of classification algorithms as shown below

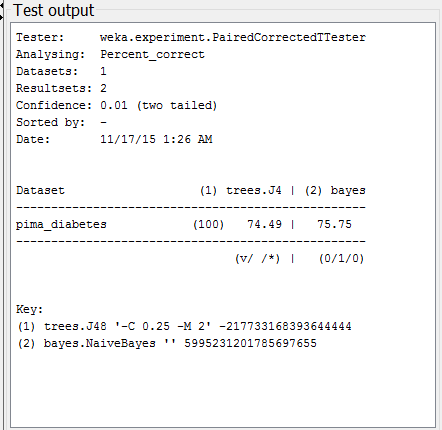


1. Now when you click on the Experiment menu it displays the available result sets for analysis and to see the performance results of two classification algorithms you need to select the parameters for the analysis and then click on perform test button as shown below



1. As you click on the perform test button the results are displayed on the right side of the window





**Source code for classification algorithms**

1. J48 classification algorithm:

<https://svn.cms.waikato.ac.nz/svn/weka/trunk/weka/src/main/java/weka/classifiers/trees/J48.java>

1. Naïve bayes:

<https://svn.cms.waikato.ac.nz/svn/weka/trunk/weka/src/main/java/weka/classifiers/bayes/NaiveBayes.java>



**Links to download weka software tool and dataset**

1. Weka tool link:

<http://www.cs.waikato.ac.nz/~ml/weka/downloading.html>

1. Data set link:

<https://svn.cms.waikato.ac.nz/svn/weka/branches/waikato/wekadocs/data/diabetes.arff>

