**Group 16**

**CO 544 Project Report**

**Milestone 2 (Final Predications to Kaggle)**

*We used two methods for predictions***,**

1. WEKA
2. Python

Basic Procedure

1)Used weka to build the model using train data and predict test data .

For validation we used 10 fold cross validation. Which can divide the particular data set to 10 sets and validate .

(Train using data set always showed more accuracy than the validation, because train set uses all data to build the model.)

2)Uploaded predicted test result to kaggle

3)Choose the best algorithm from the accuracy

4)Improved (tuned the ) choosen algorithms further using python because is more flexible than weka.

**Method 1(Using WEKA)**

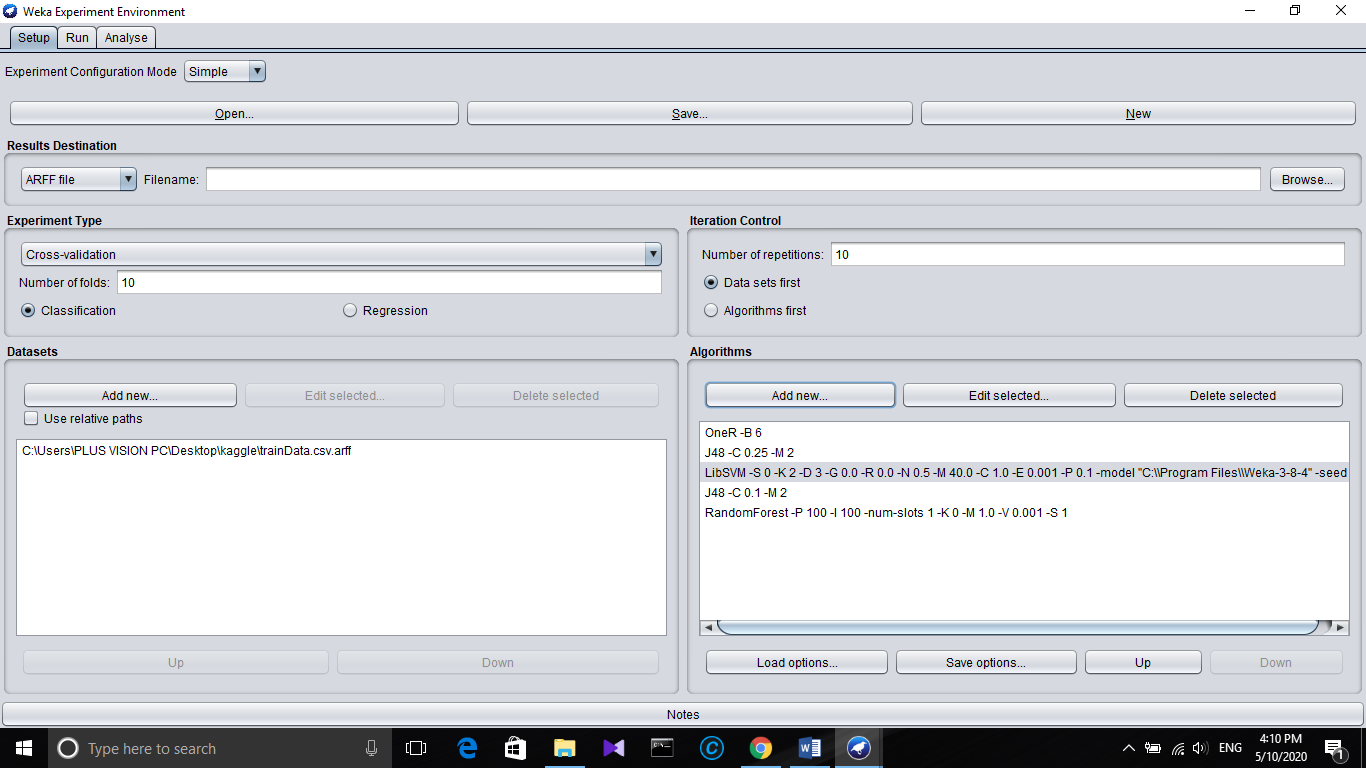
Why WEKA?

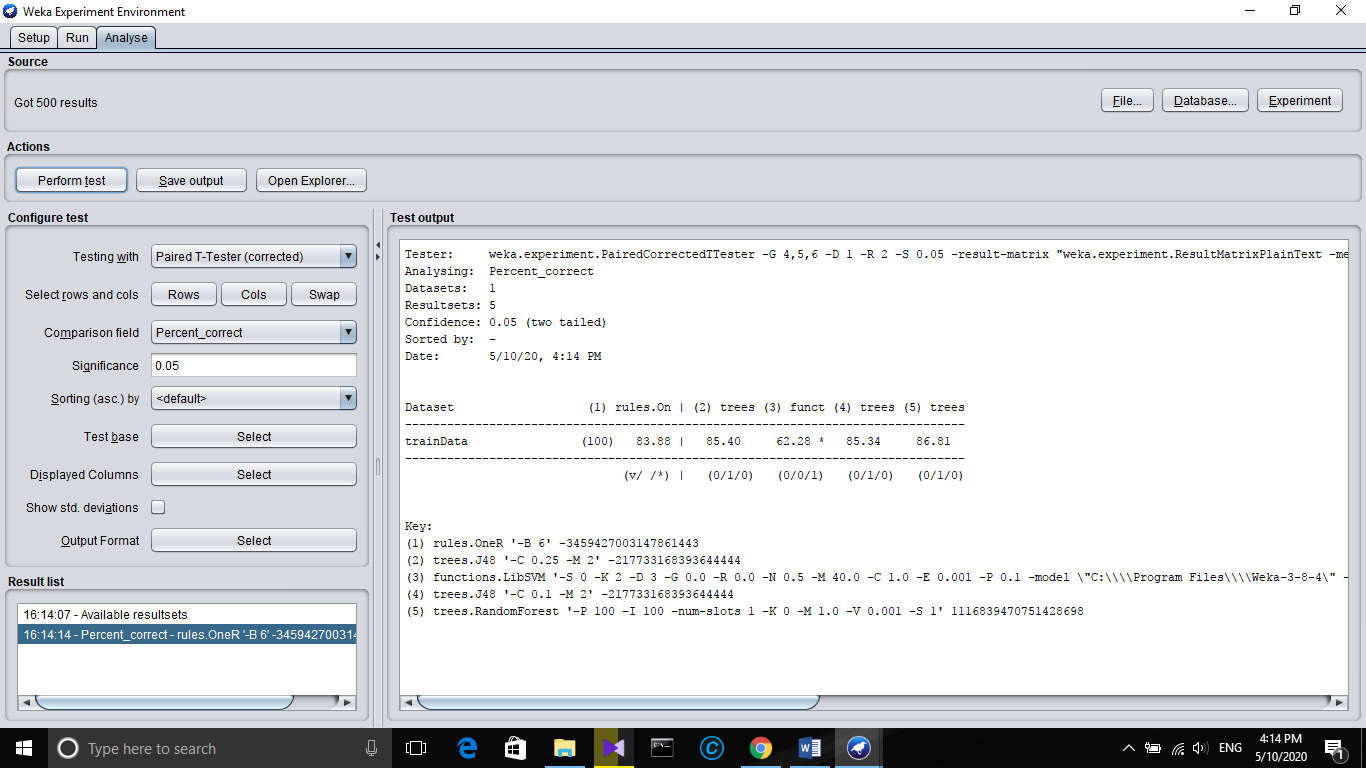
Weka gives a GUI which made comparisions easy for the training data.

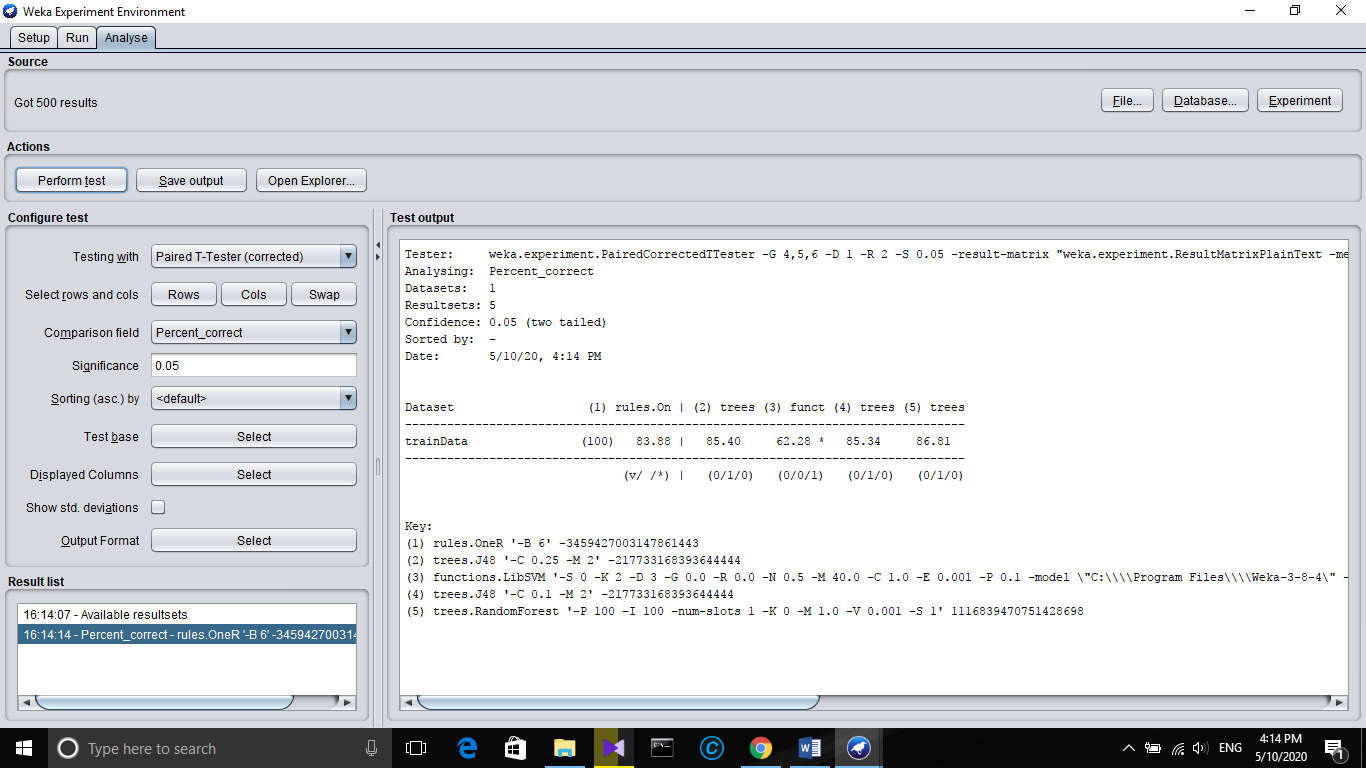
We tried various algorithms supervised classification algorithms on the training data. We can get a percentage of correctly classified instances in Weka GUI.

* K-nearest neighbors
* Support Vector Machines
* Decision Trees
* Random Forest

**How to compare the algorithms?**

1. **Weka -> Experimenter**
2. **Choose the train data set**
3. **Choose 10 folds cross validation. Repetition =10**
4. **Algorithms- J48(Confidence factor=0.1 and 0.25),SVM, Random forest,1R**

**5.Analyse using “Paired T-tests”**

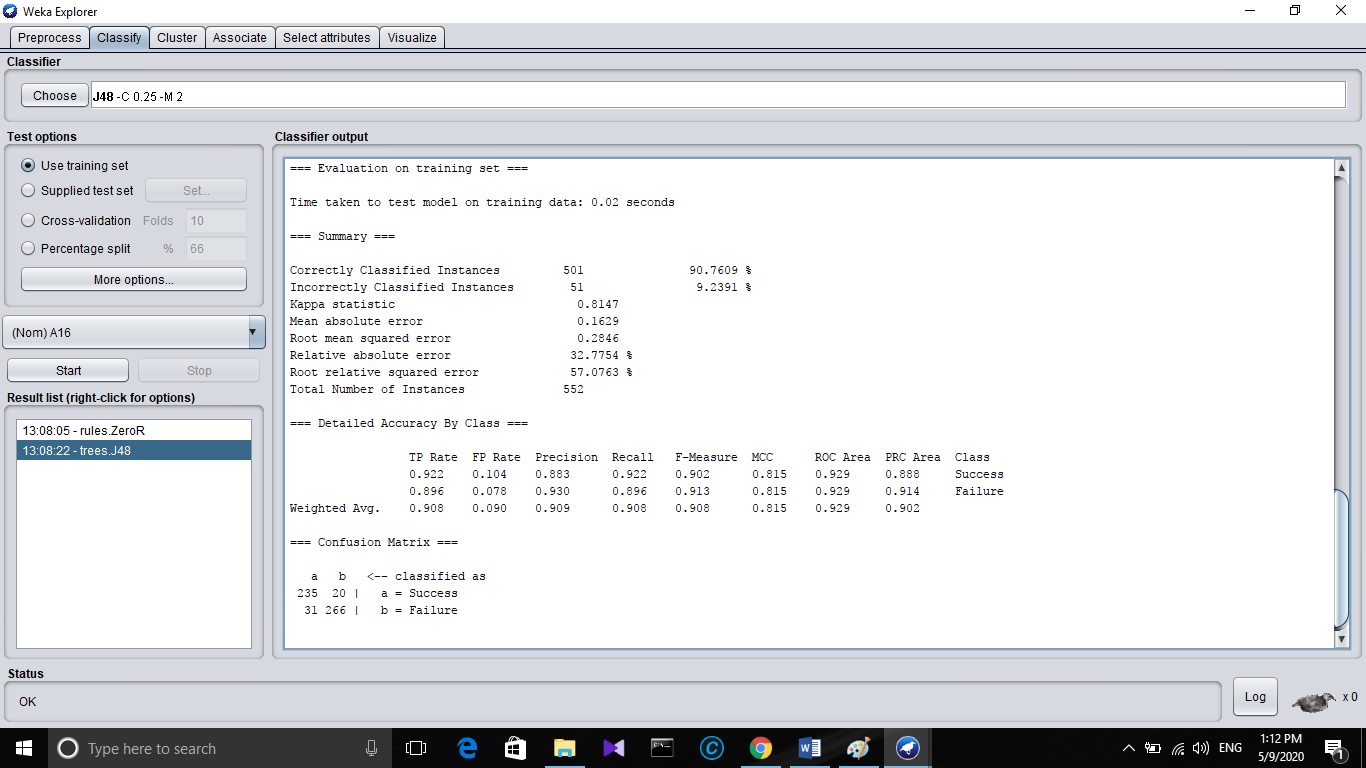


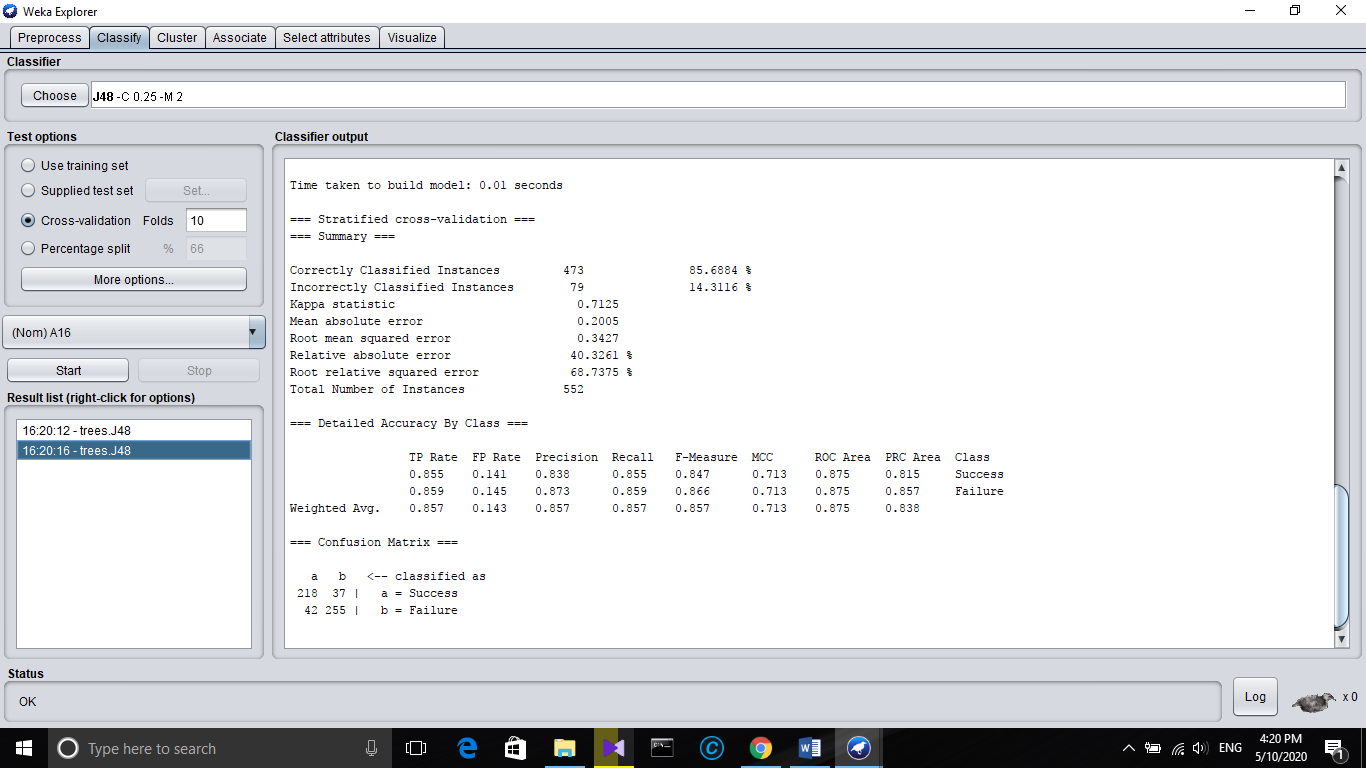
**6.From the above results it is clearly concluded that Random Forest and J48 tree gives the best accuracy.So they were used to build model in explorer.**

**(i)J48(C4.5)**

Confidence factor =0.25

1)train set

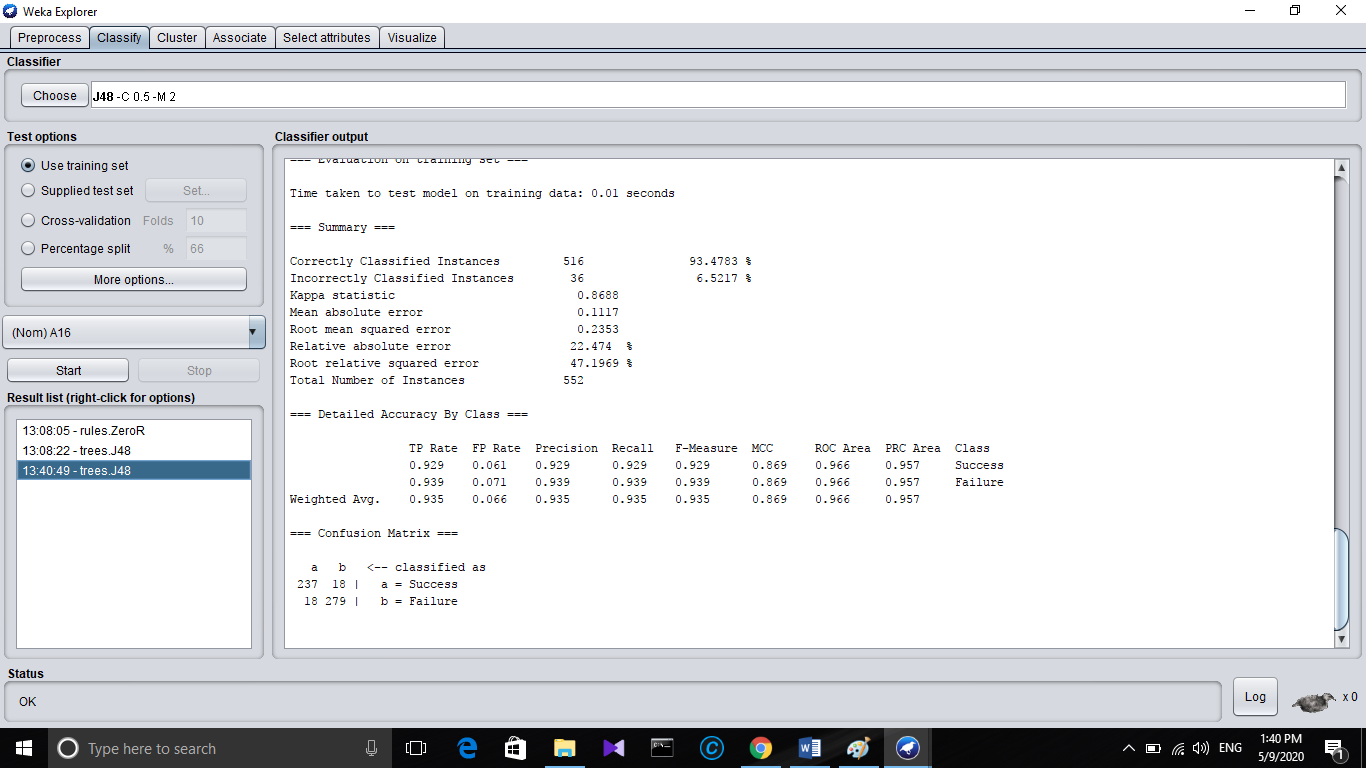


**2) cross validate**

**Remarks**:Even though Cross validation shows less accuracy than test the tree was same because at the end cross validation too uses all data points to build up the model

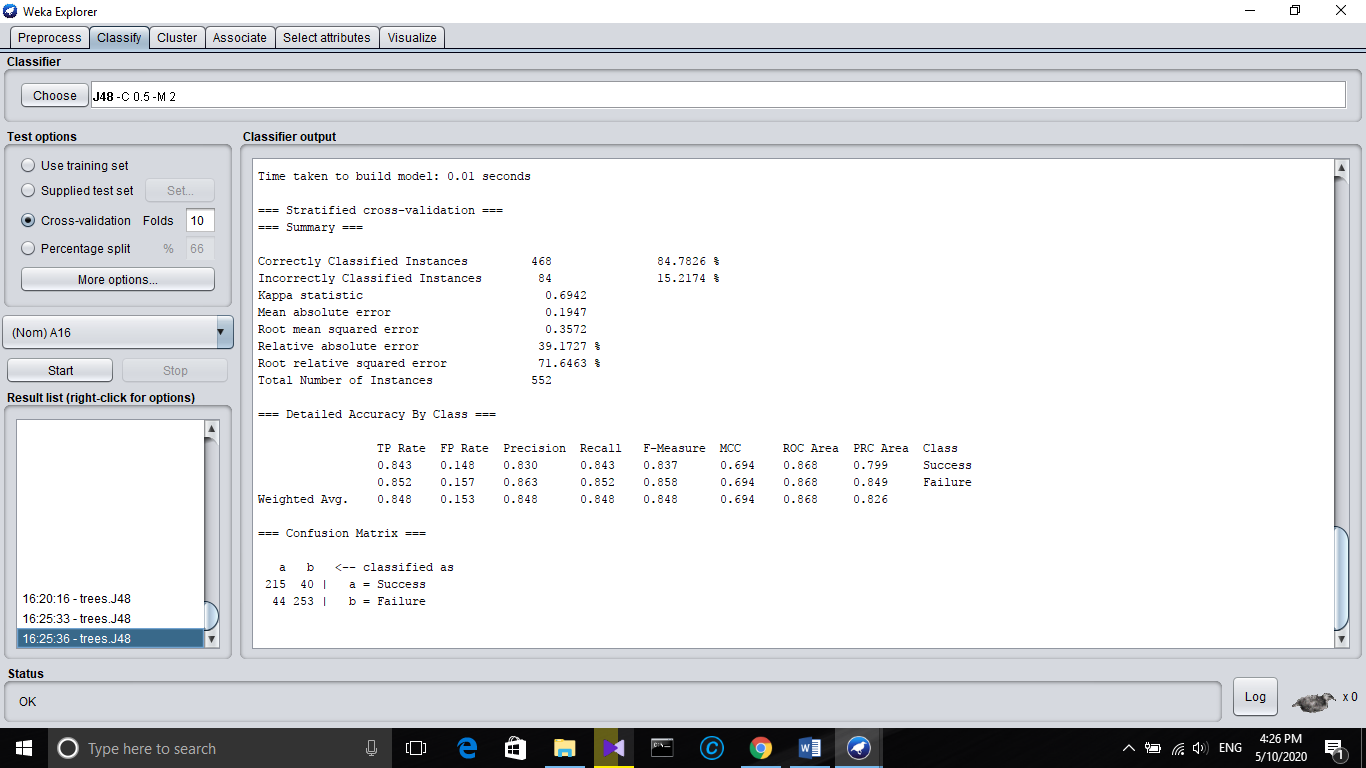
C4.5 (**J48**) is an **algorithm** used to generate a **decision tree** developed by Ross Quinlan. Extension of ID3

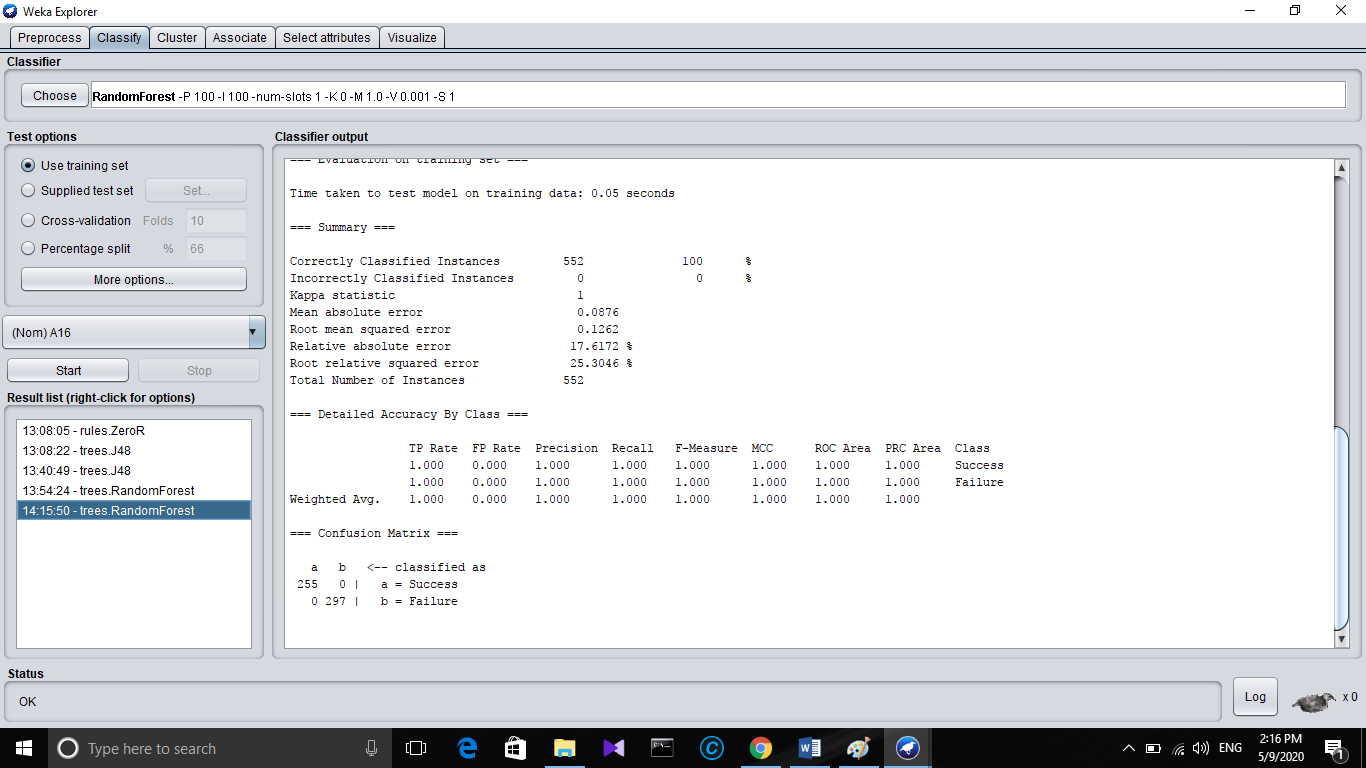
Confidence Factor=0.5

1. Train set

**Remarks**:Increasing the confidence factor increased the accuracy.

2)Cross validation



(ii ) Random forest

Remarks:Very high accuracy. BagsizePercentage and batch size and nmber of slots were changed to get the highest accuracy with lowest mean absolute error.

**Step:**

By uploading the results of the above algorithms to the kaggle we got the above decision tree and the Random forest algorithm gave the the best accuracy for test data set (in the public leader board).

Therefore we choose J48 classifier and Random Forest classification with python for further development.

**Python**

**Why python?**

Eventhogh weka gives GUI and a great envioranment for predoctions, it is less flexible and so as we explore we got to know that the Python in ML gives more degree of freedom when exploring and tune and tweeking the algoithms.

Following is the python procedure +source code

**Brief Procedure of how weka used to predict**

**Milestone 1(Before kaggel introduced)**

**Prediction margin**-

defined as the difference between the probability predicted for the actual class and the highest probability predicted for the other classes. A margin of 1 means that the correct class is predicted with 100% confidence (very good), a margin of -1 means that an incorrect class is predicted with 100% confidence (very bad).

**Procedure**

**1)**.Both Training and Test data sets were preprocessed and made to ARFF formats. For test data sets Attribute names were given and the Class (Which is to be predicted is names as “A16”)

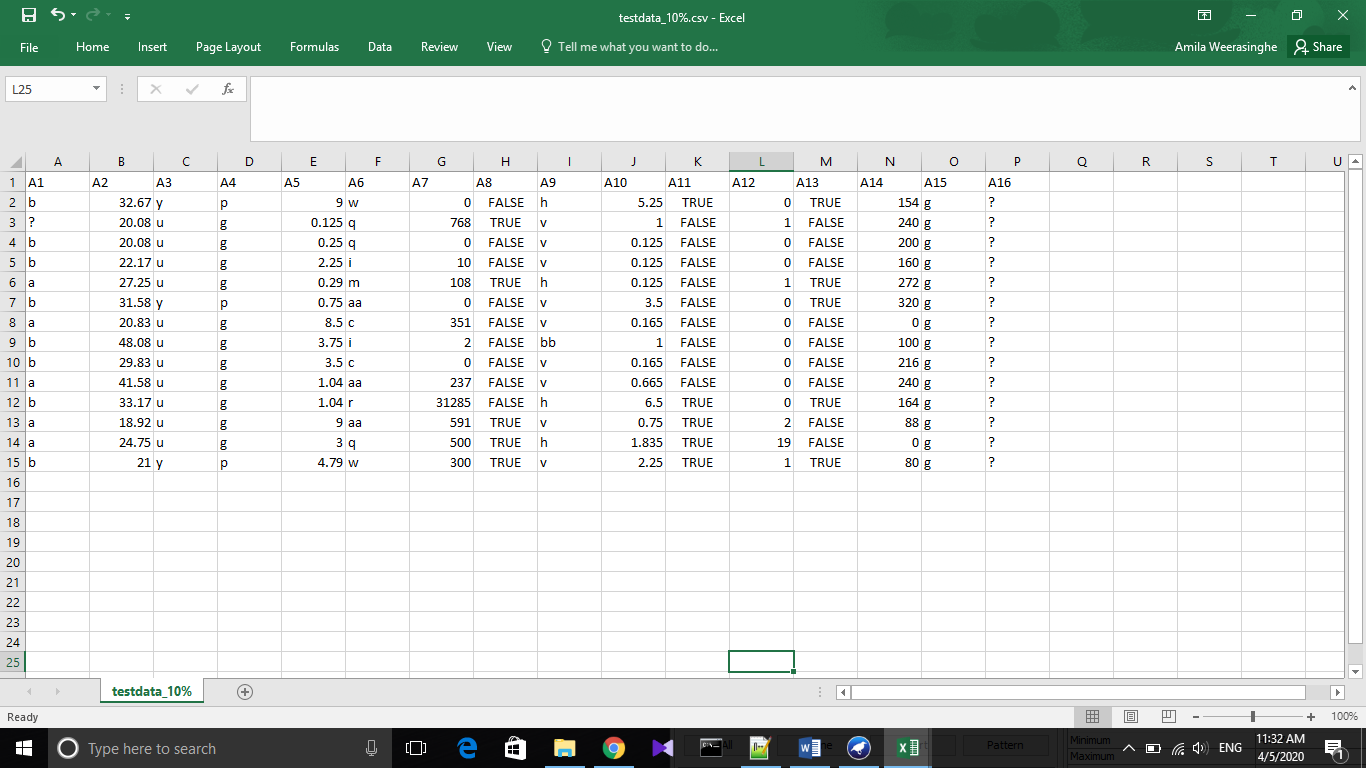


Figure 1: Modified Test data set

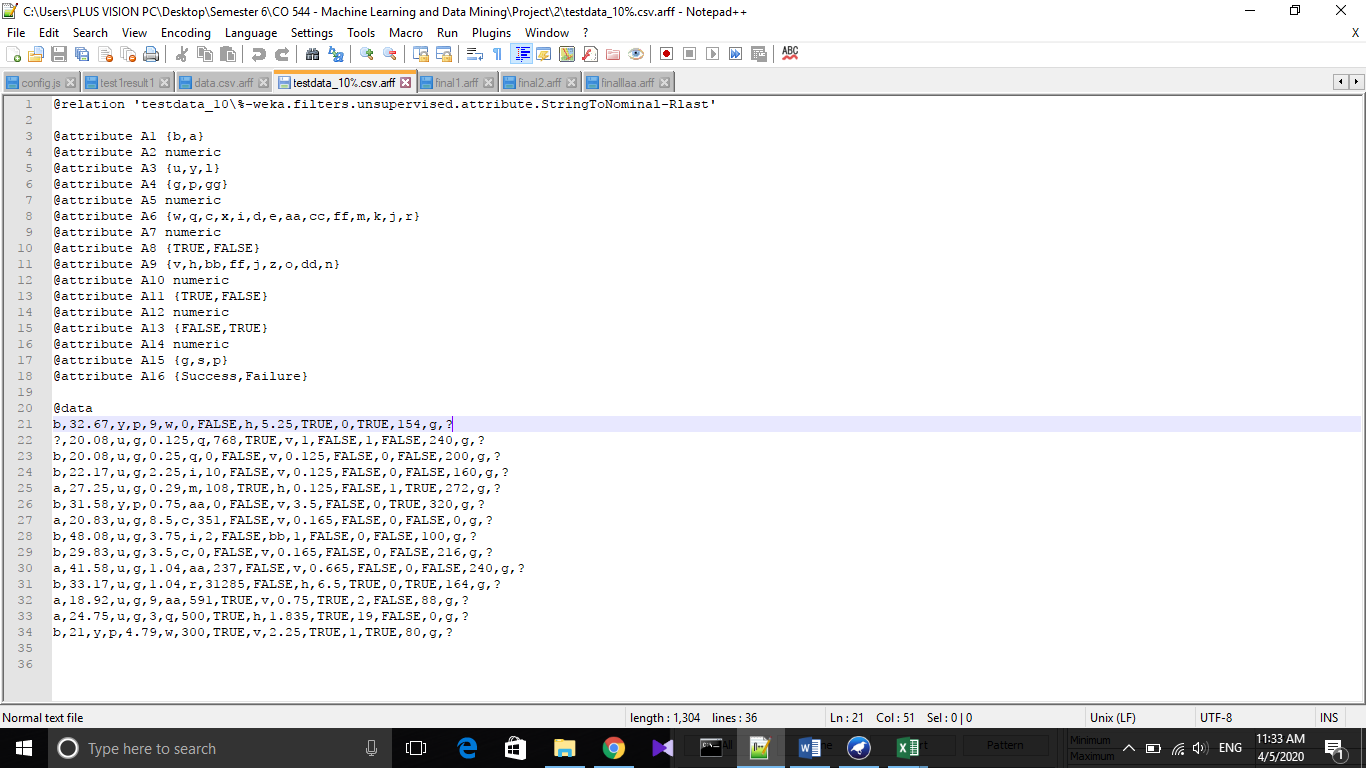
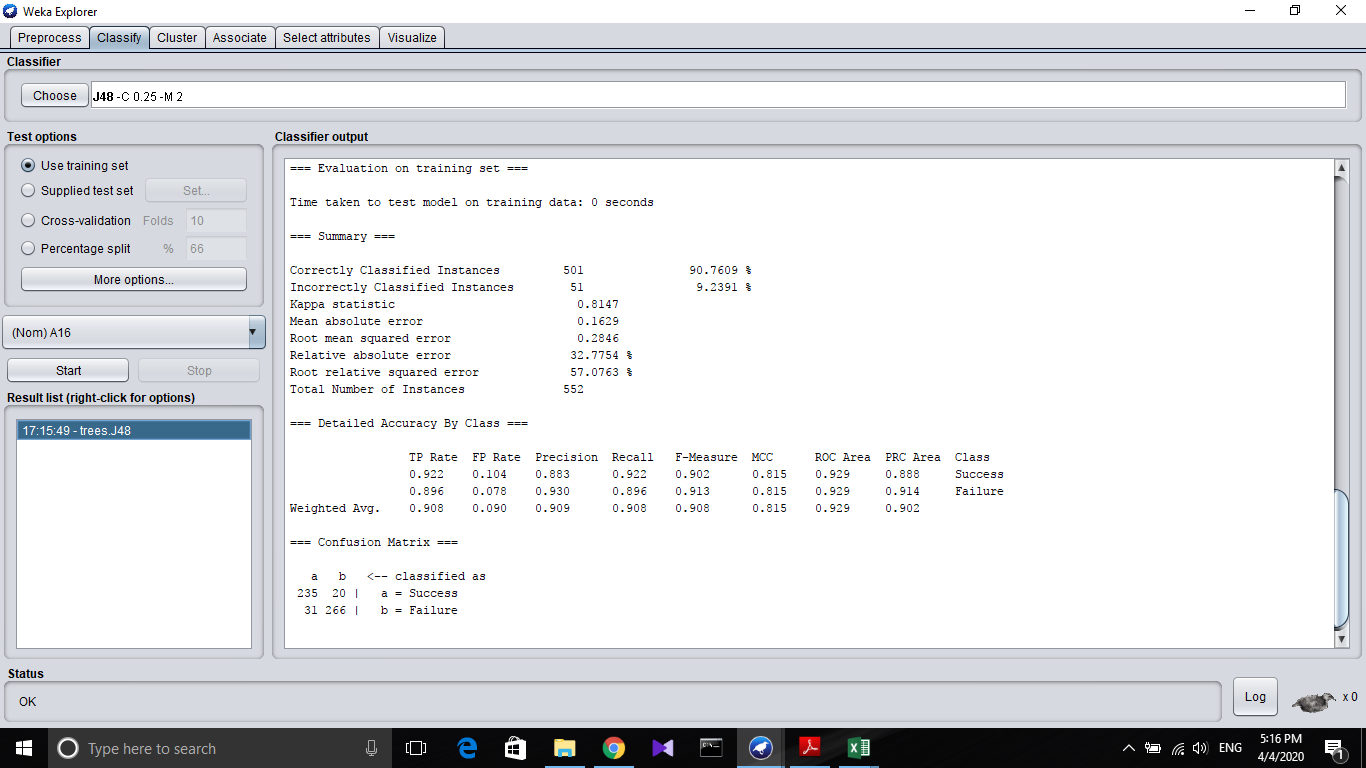
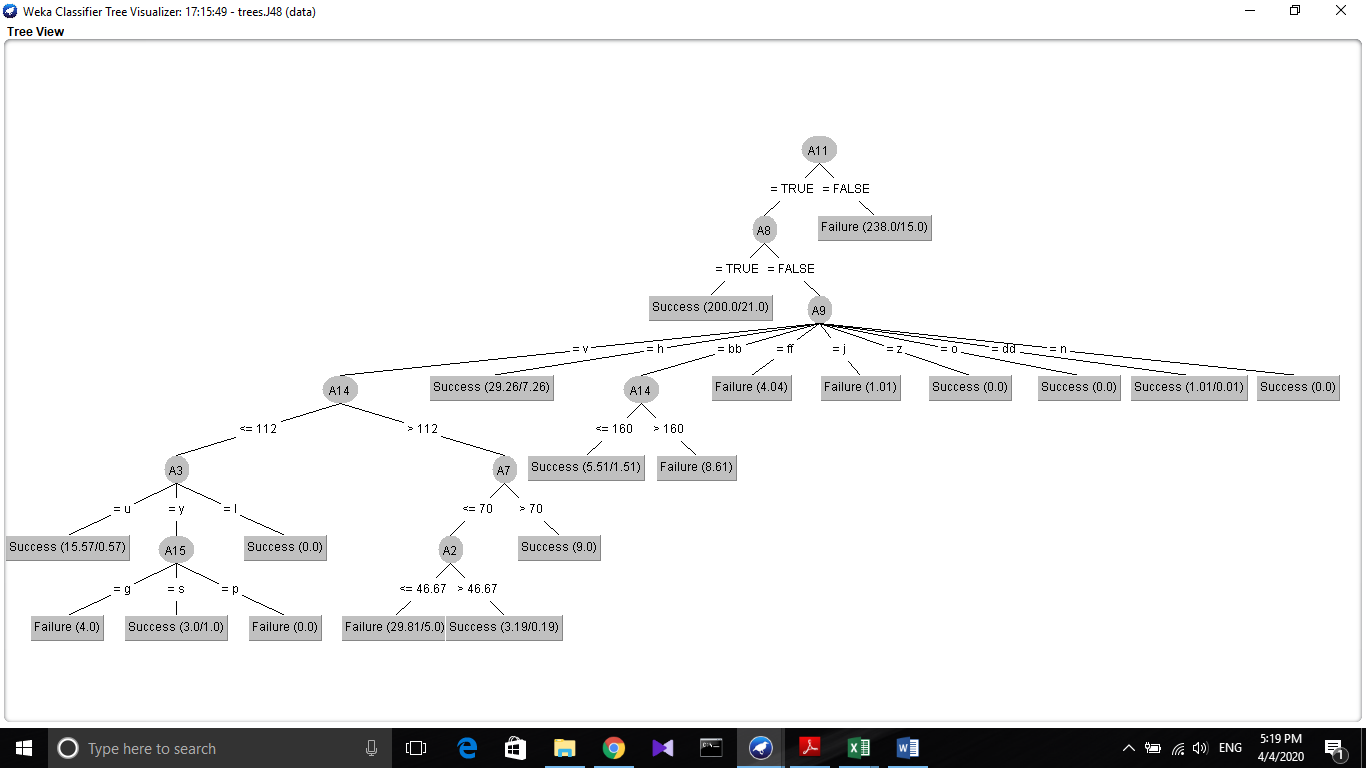


Figure 2: Modified test data in ARFF format

**2)**Training Data set classified Using J48 tree(Use as training data set)

**Correctly Classified 90.7609%**

Visualised tree(for training data set)



**3) Supplied the test data set and classified Using J48 algorithm.**

**Results were saved.**

**Results**

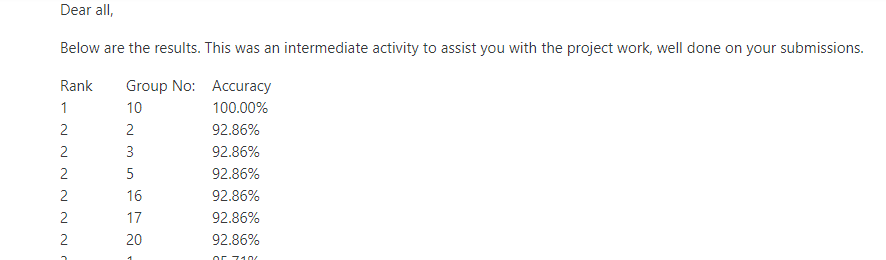
****

Figure 3: Final Predictions results on Feels

Our predictions were 92.86% accurate ranked 2.

**Conclusion**

**Prediction margin:** This is defined as the difference between the probability predicted for the actual class and the highest probability predicted for the other classes. A margin of 1 means that the correct class is predicted with 100% confidence (very good), a margin of -1 means that an incorrect class is predicted with 100% confidence (very bad).