# Phase 5: Data Modelling

**Team Name:** Deep Diver's

## Team Members:

Vaishnavi Hemant Salaskar vsala2@unh.newhaven.edu

Vedant Chidgopkar vchid2@unh.newhaven.edu

Riddhi Joshi rjosh5@unh.newhaven.edu

## Research Question and Selected Data Set

Today mental health is becoming a more common problem. However, evaluation of mental well-being is extremely important to understanding and providing therapeutic solutions. Diagnostics are complicated tasks and misdiagnosis can result in serious problems if a mental disorder is not properly detected. Can we recognize mental health issues accurately by using data mining techniques?

The data has been collected from Kaggle by Open Sourcing Mental Illness, LTD. Survey data about mental health attitudes are included in this dataset. Which then has been analyzed and pre-processed. The data contains different labels such as age, gender, country, self-employee, family history, work interference, seek help, etc. For better prediction, we have label encoded the data.

## List of Data Mining Techniques Used

- Naïve Bayes
- Decision Tree

## Description of Hardware Used

Weka is used to perform data mining tasks using in-build machine learning tasks. It is used to classify data to know the accuracy of the algorithms performed on the datasets. Classification, Regression, Clustering, Association Rules, and Visualization techniques can be performed in Weka.

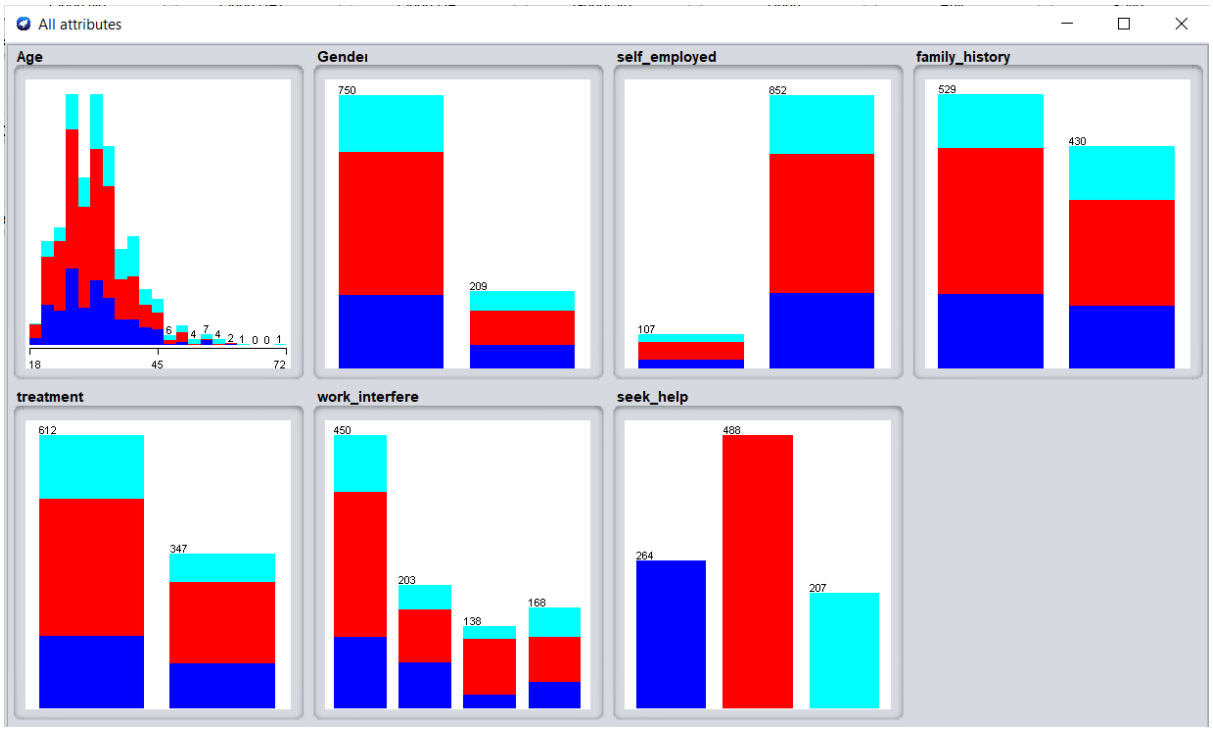
## Visualization

### Histogram

In Weka, we can visualize data in multiple ways. In the histogram, the distribution is done mainly from attributes. A single selected attribute is distributed at a time. By default, that will be the class attribute.

Total Instances: 959

Attributes: 7



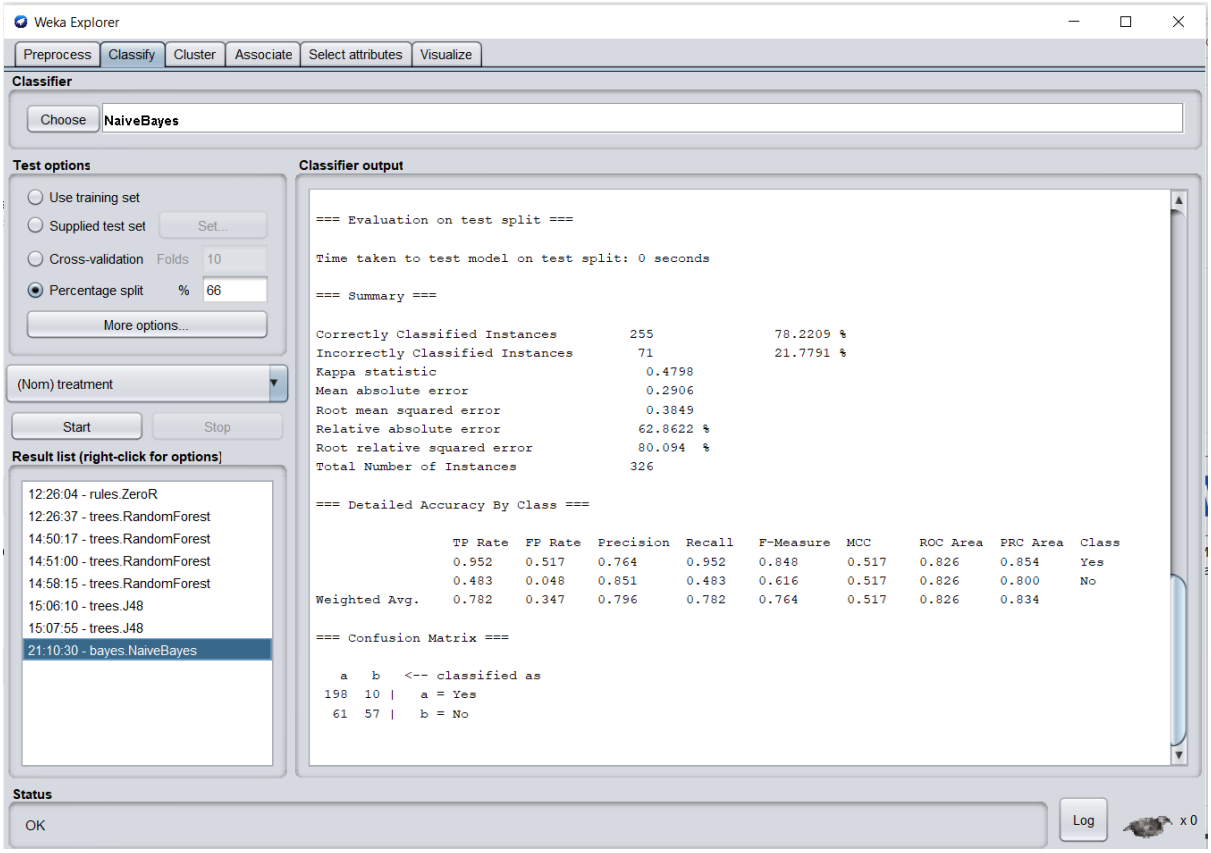

In our dataset, we have noticed that in the Age attribute the Distinct Values are 44, Unique Values are 7 having the Standard Deviation as 7.403, and Mean as 32.356. Furthermore, in the treatment attribute, the Distinct Values are 2 with no Unique Values. The total count of yes is 612 and the total count of no is 347.

## Outcomes

### Naïve Bayes Classification:

Naive Bayes methods are a set of supervised learning algorithms. Naïve Bayes Classification is used for classification task. It is used to classify objects.

Using Bayes.Net, the probability distributions are calculated over all classes, and the evidence arising from the cross folding can be divided into independent parts.

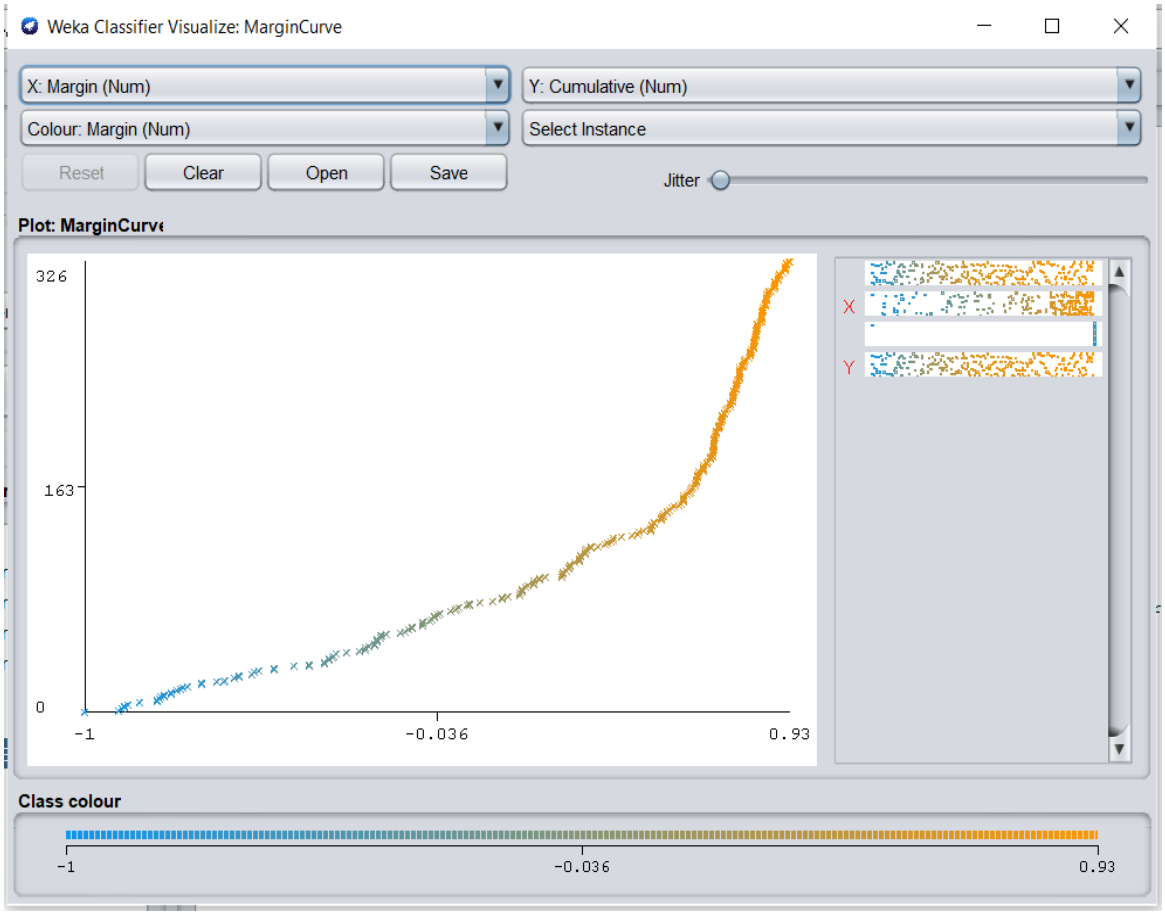



After performing the classification, we observed that Correctly classified Instances are 255 with the accuracy of 78.2209%. And Incorrectly Classified Instances are 71 with inaccuracy of 21.7791%.

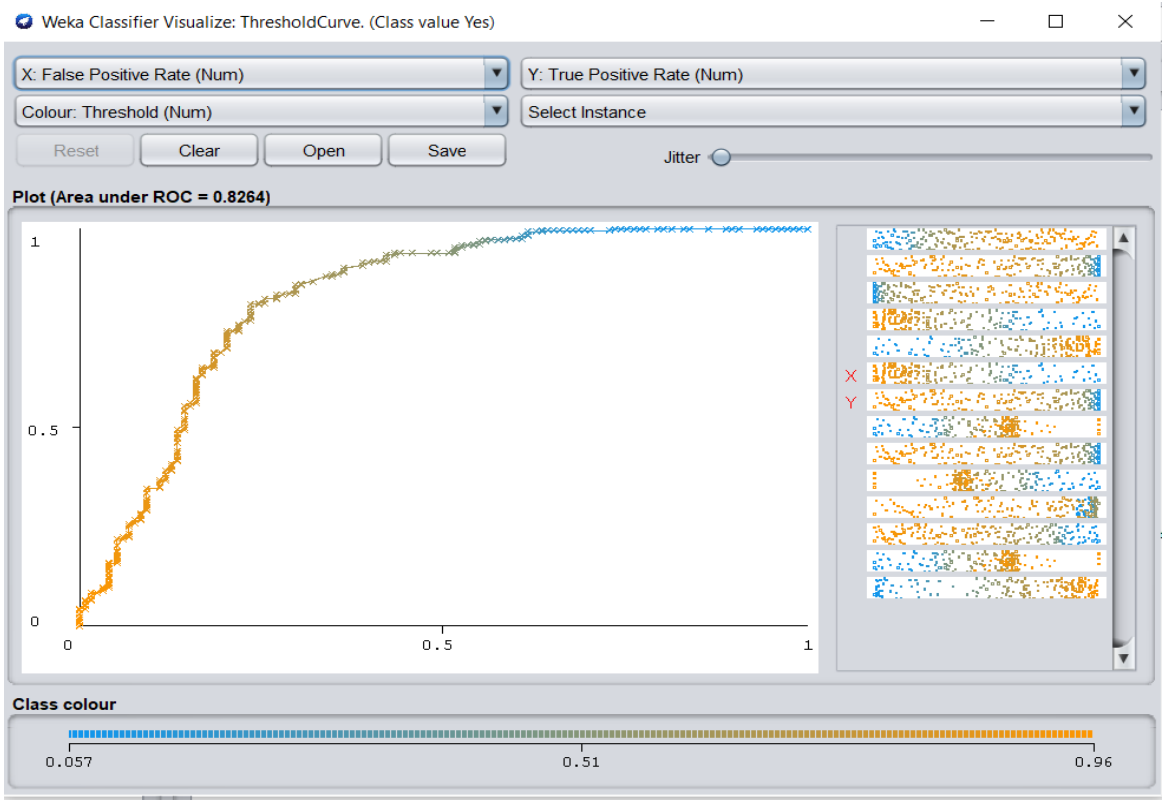
Total Number of Instances are 326

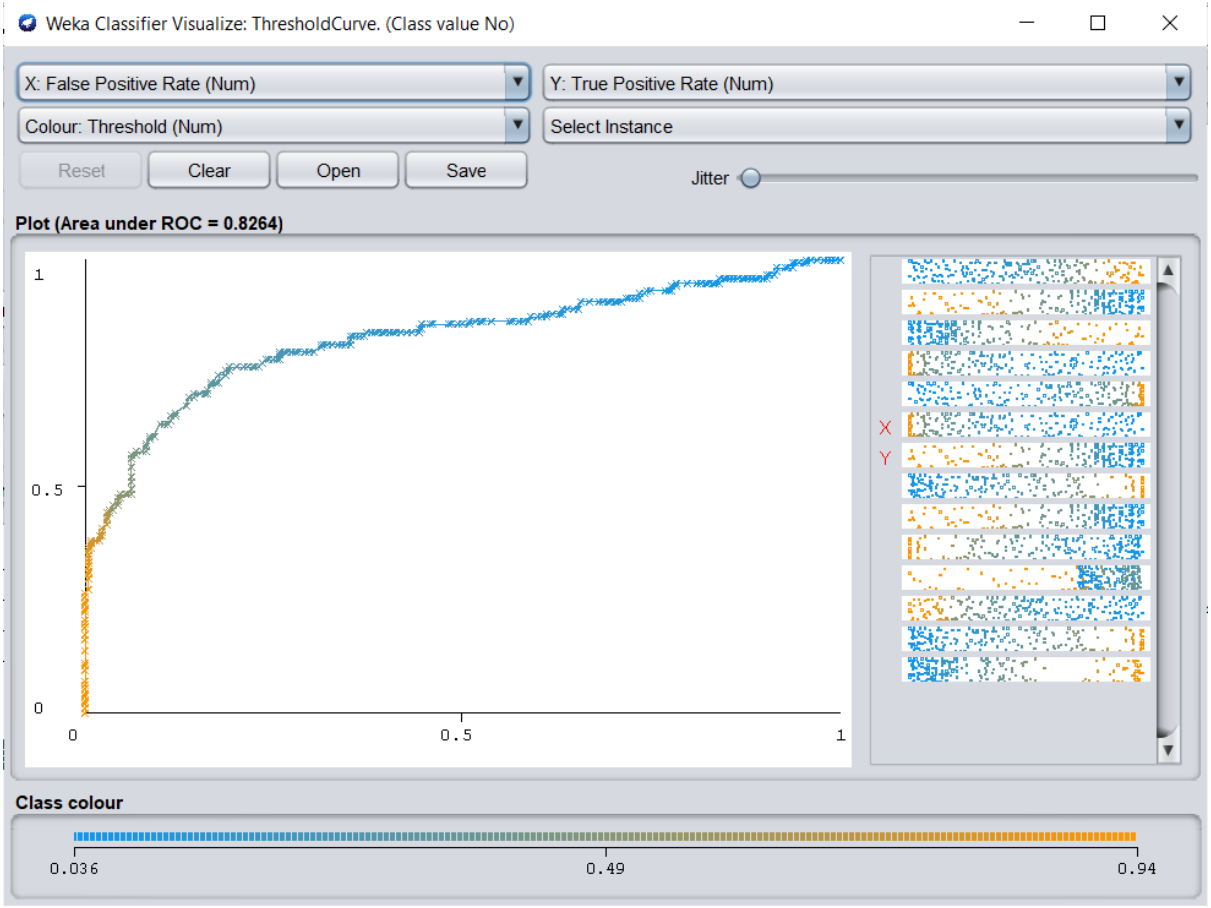
Here, the Precision Value is 0.764 which is pretty excellent value. Also, 198 True(yes) are instances, 61 are False(No) instances and 57 are true negative instances. Furthermore, it is shown that the ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FTR) after classification has been completed.

In Naïve Bayes, we are also calculating margin curve. A margin can be defined as the difference between the probability predicted for the class and the highest prediction for the other classes.




Furthermore, the Bayes Theorem allows you to calculate the threshold curve, which is used to measure the classifier's accuracy independently of the trade-offs the user decides to make.






The Area under ROC for both the threshold is the same 0.8264 which adds a genuine factor to the model.

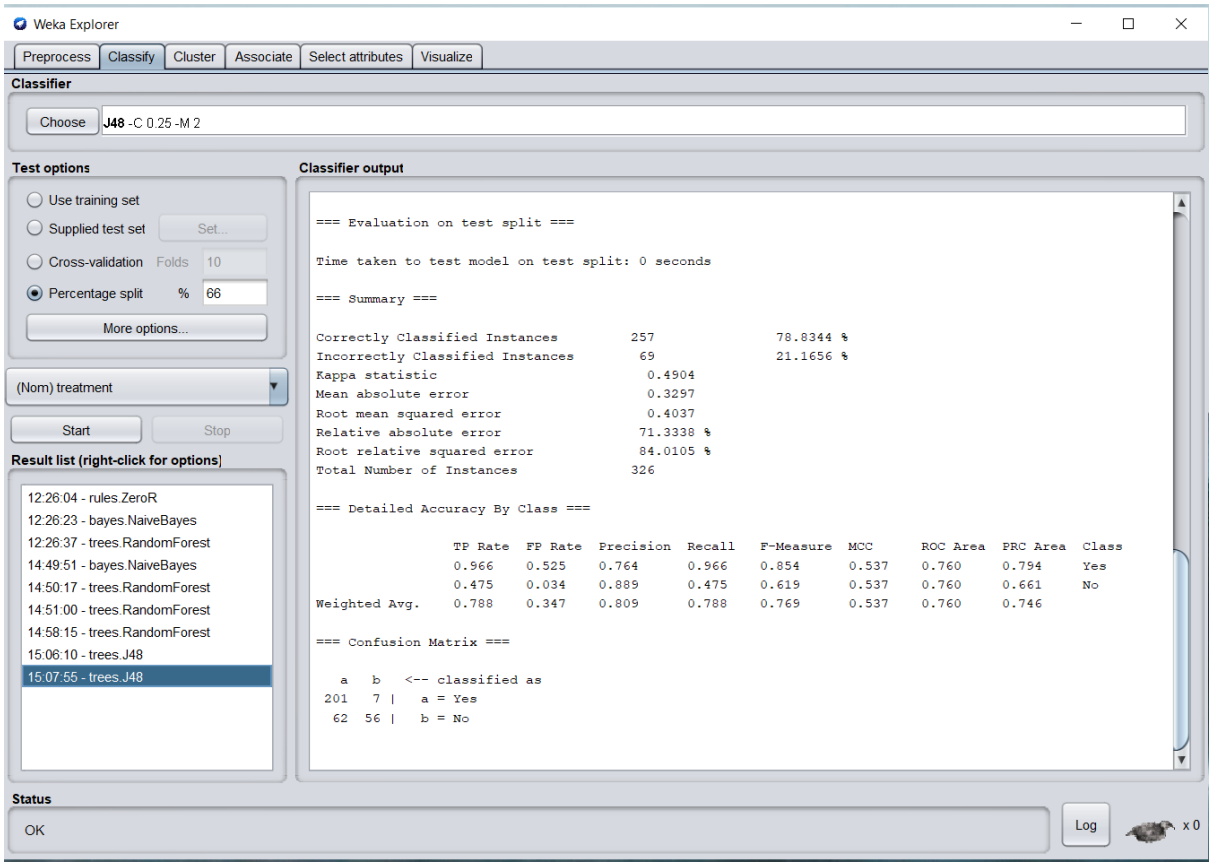
## Decision Tree:

Decision tree is supervised learning which is used for regression and classification. It is non-parametric tree. Using this classification data can be easily understood.

The results we found after the classification from pruned trees are:

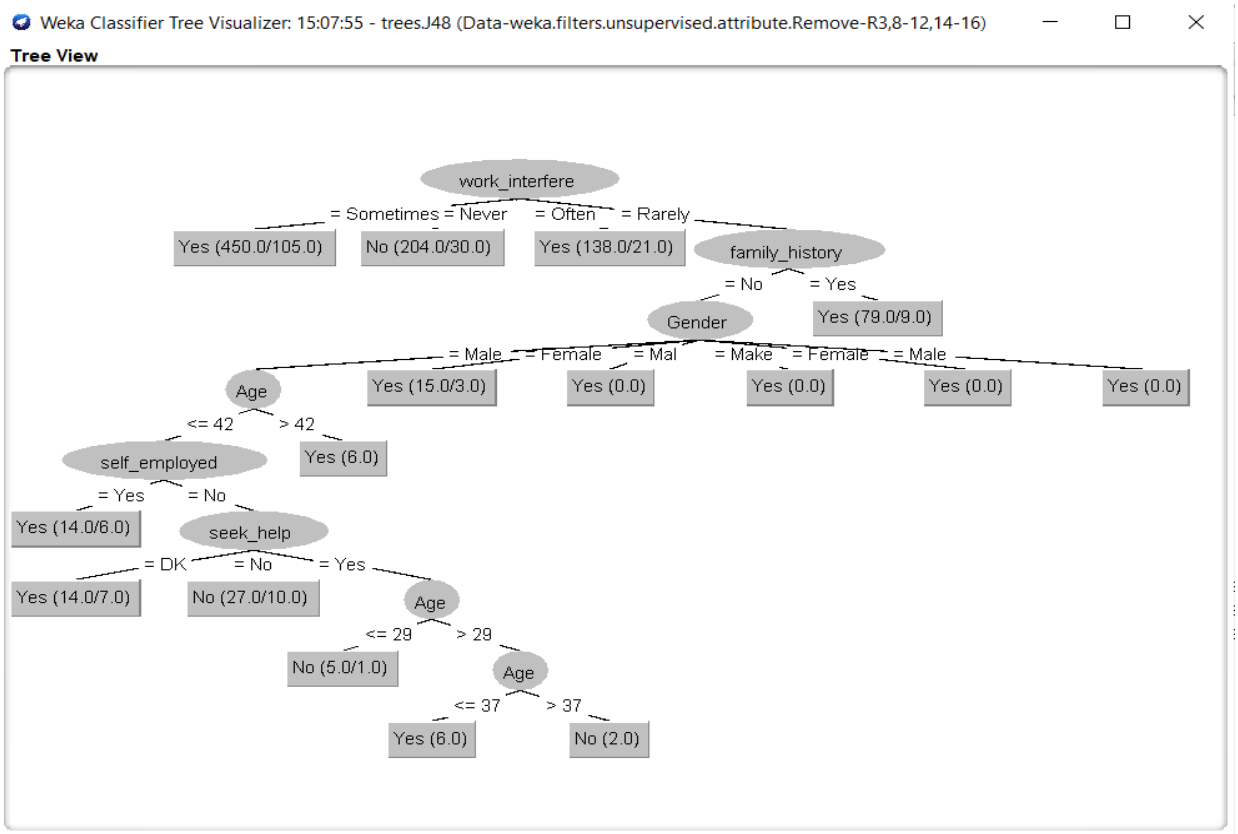
- No. of Instances: - 326
- No. of Attributes: - 7
- Size of the tree - 20 leaves

Total Number of Instances are 326 from which Correctly classified Instances and Incorrectly Classified Instances are 257 and 69 respectively. We also find the Mean absolute error which is 0.3297. Alongside Root mean squared error 0.4037.

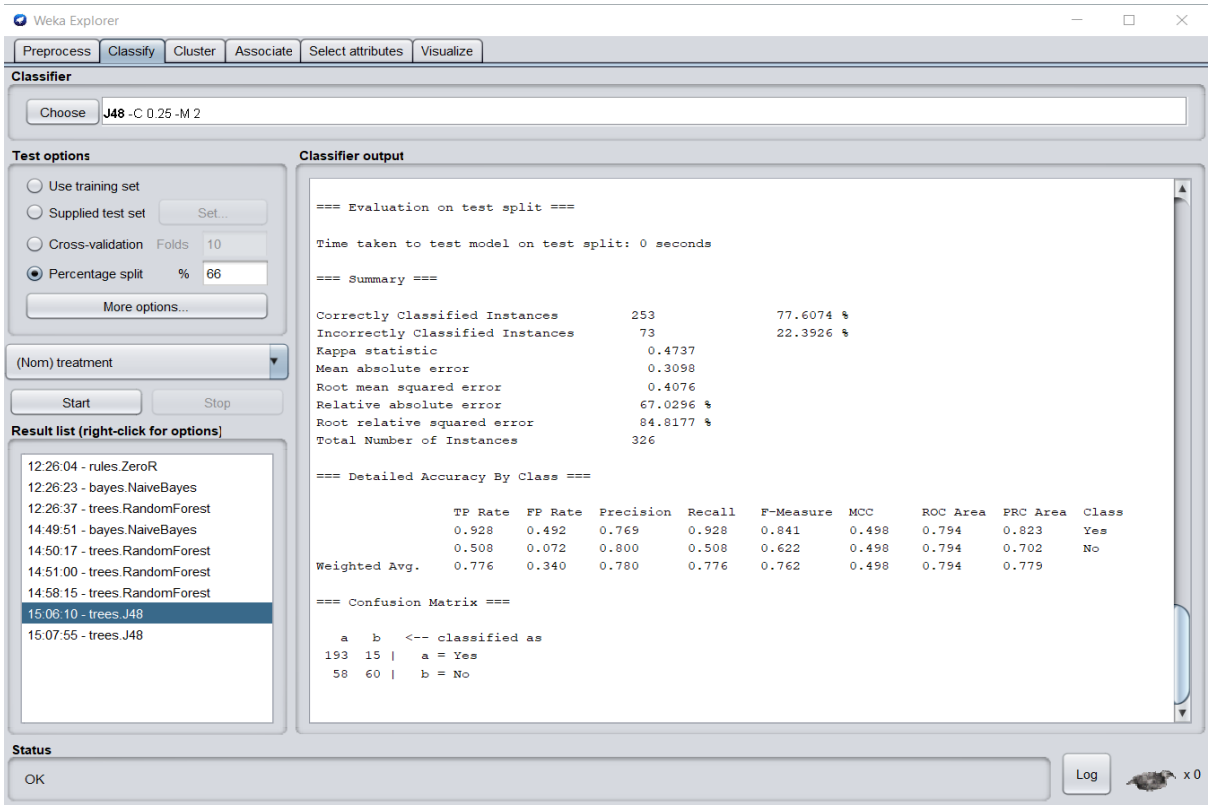



The confusion matrix values for true positive are 201 and for false negative is 56. The precision value is 0.764. We found out recall value for same 0.966.

For decision tree, observations that we found is work interface is a parameter where data is being split and the weightage of attributes are as shown in figure.




- Accuracy with uprunning set to false : - 77.83%
- Accuracy with uprunning set to true : - 78.60%



## Conclusion:

Evaluation, measurement, and comparison of the performance were conducted using the Weka. The result for data using decision tree and Naïve bayes models were observed. The accuracy finding for Naïve bayes is 78.22% with 21.16% of incorrect classification. Whereas for decision tree the accuracy recorded is 78.83% with precision 0.764 and recall 0.966, which is higher as compared Naïve bayes. We found out that for selected dataset the decision tree is providing accurate result and stable model.

## Repository:

https://github.com/vsala2/DataMining