Phase 6: Optimization

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
- Random Forest

List of Techniques Used for Optimization

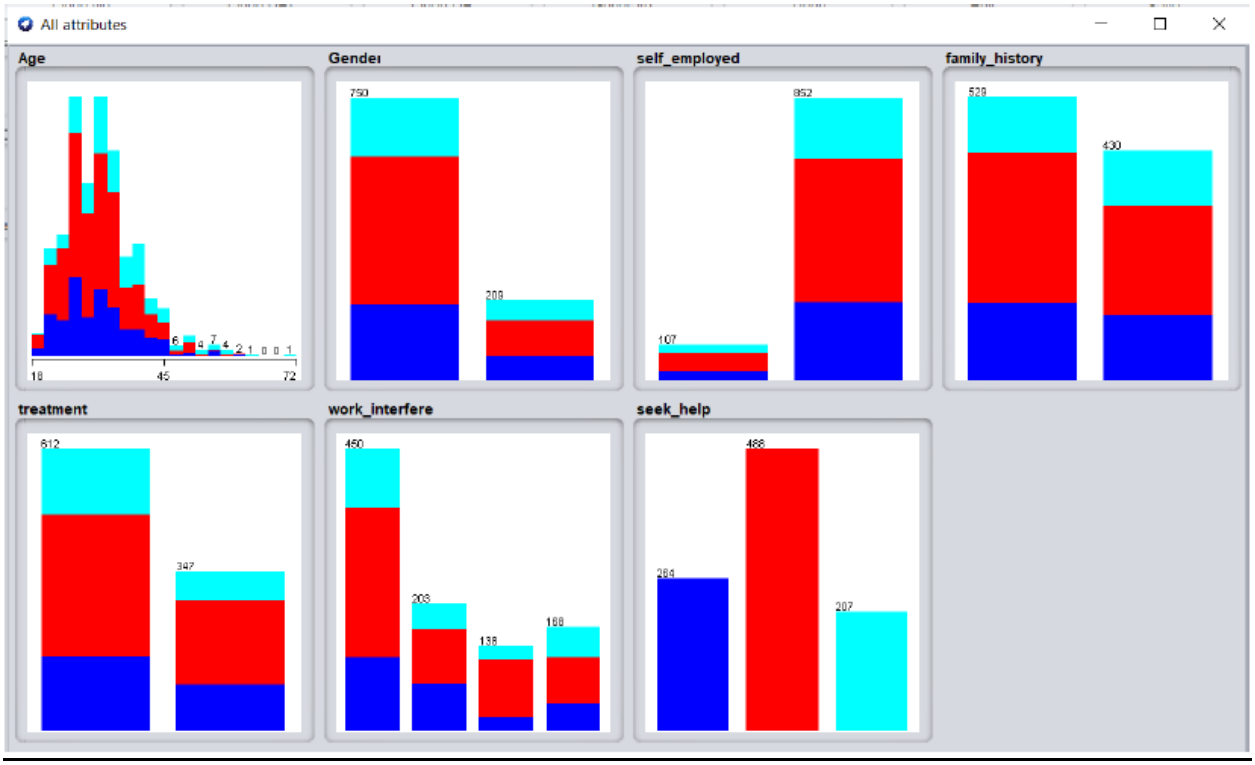
- Multiclass Classifier
- CV Parameter selection
- Random Subspace

Visualization Techniques

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.

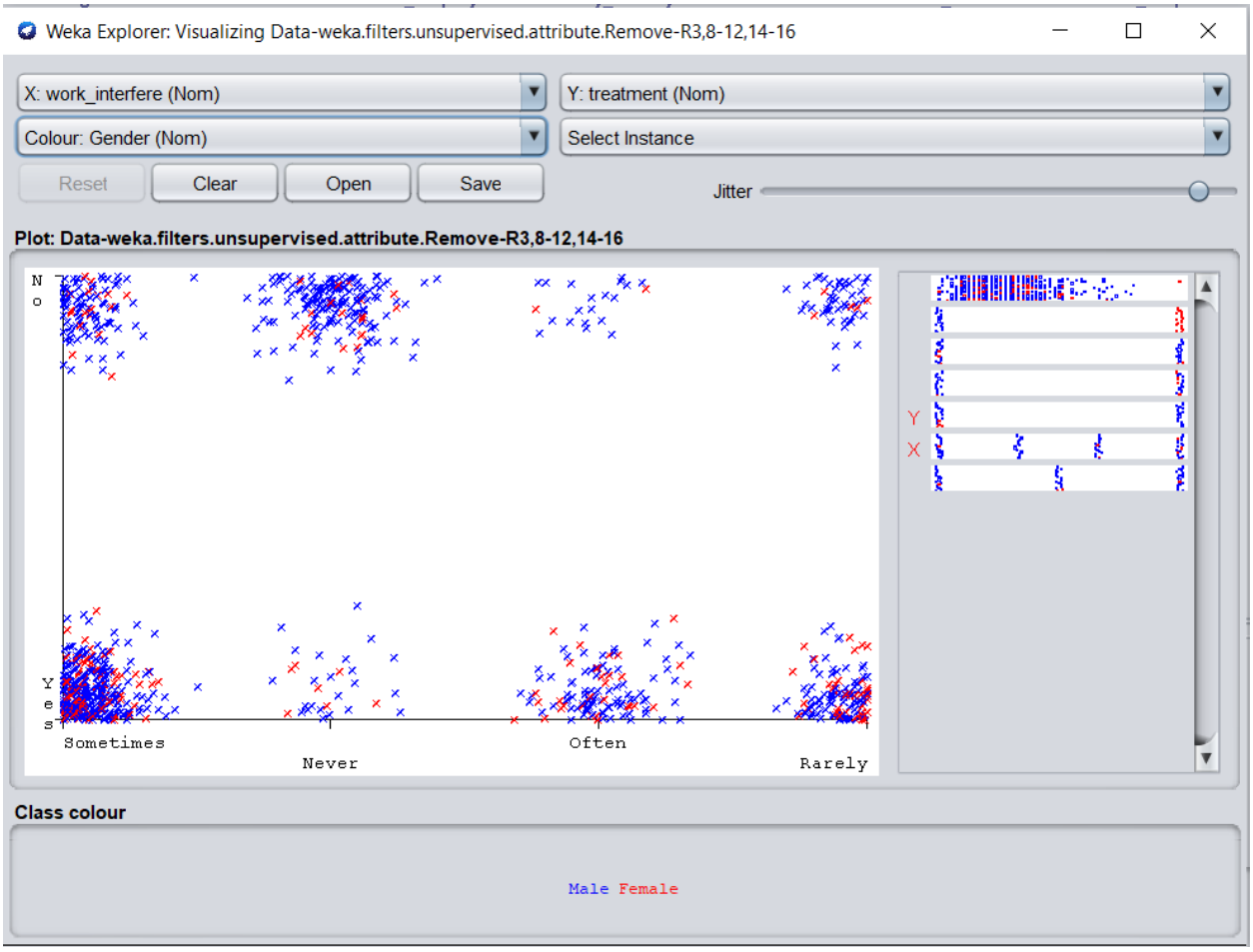
Total Instances: 959
Attributes: 7



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 is 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.

Scatter Plot

In this type of graph, the data is plotted against X and Y axis. We have used 'work_interface' attribute on X axis and 'treatment' on Y axis. In Weka, there is option to select different color to each value of attribute. We have selected blue for 'Male' and red for 'Female'. This helped for better understanding of data.

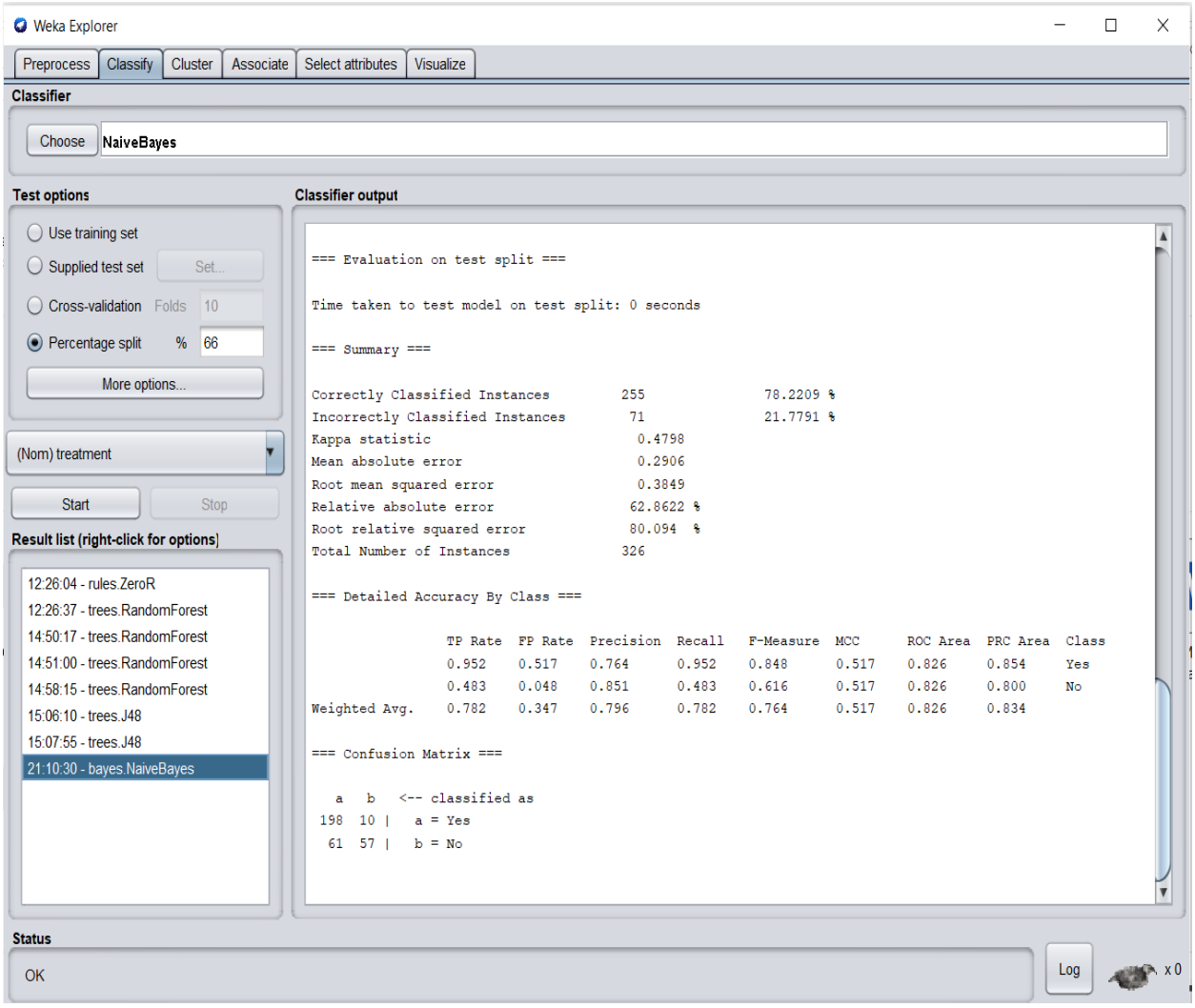


This visualization shows that number of people who has work interfere mostly take treatment. It is also predictable that the number is greater for people who has work interfere but takes treatment often or rarely. These group of people are target for this model.

Data Mining Technique and Optimization Technique used

- **Naïve Bayes Classification:**

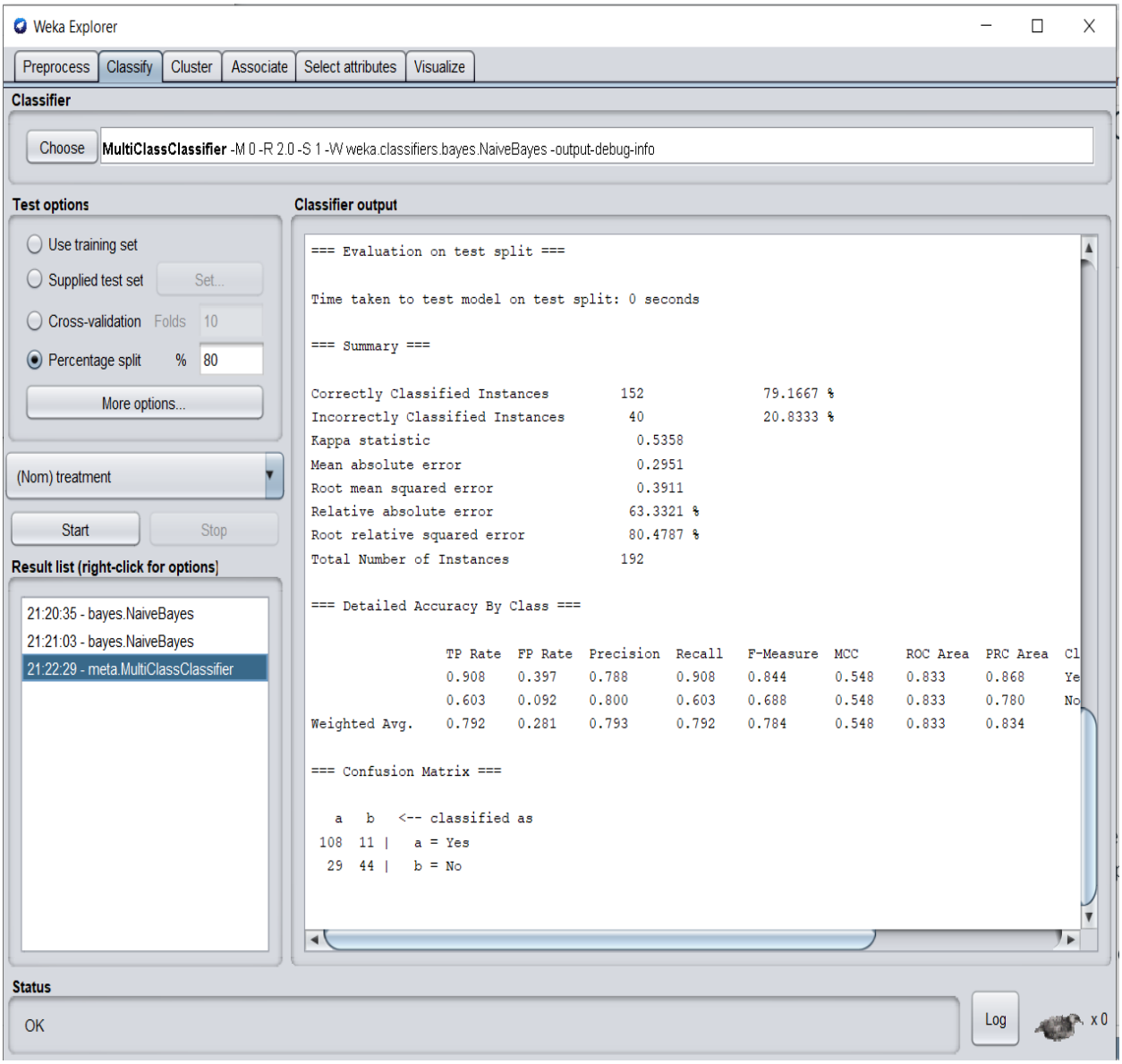
After performing the Naïve Bayes classification, we observed that Correctly classified Instances are 255 with the accuracy of 78.2209%. And Incorrectly Classified Instances are 71 with inaccuracy of 21.7791%.



To make this value optimized, we used Multiclass Classifier optimization technique for this data mining technique.

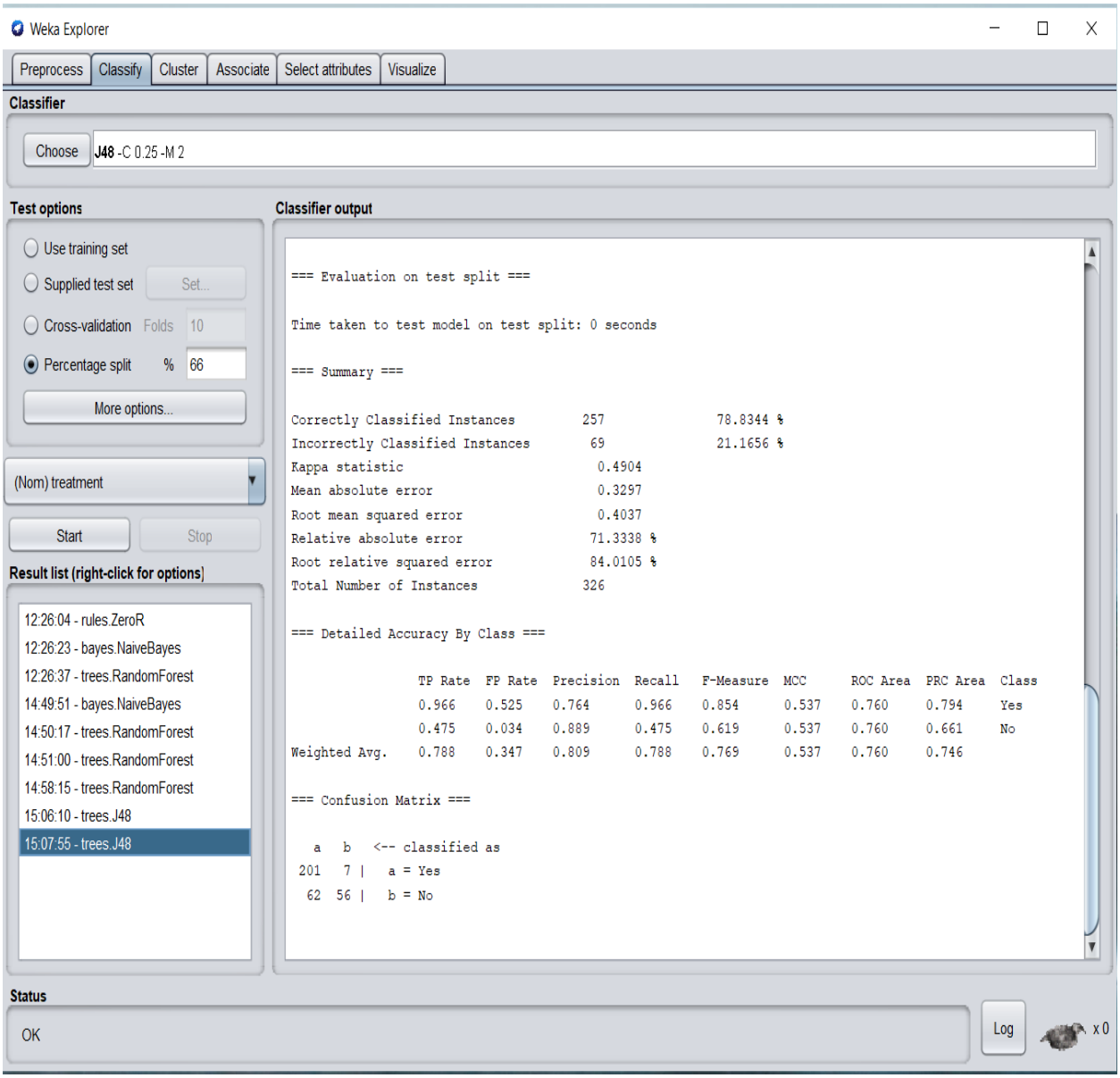
Multiclass Classifier:

For this technique we split 80% of data as training set. By doing this we got accuracy of 79.1667%. Where correctly classified instances are 152. For 40 instances which are incorrectly classified we found inaccuracy rate is 20.8333%. By using MultiClass Classifier, optimization technique for Naïve Bayes Classification, we were able to improve the accuracy rate from 78.22% to 79.166%.



- **Decision Tree:**

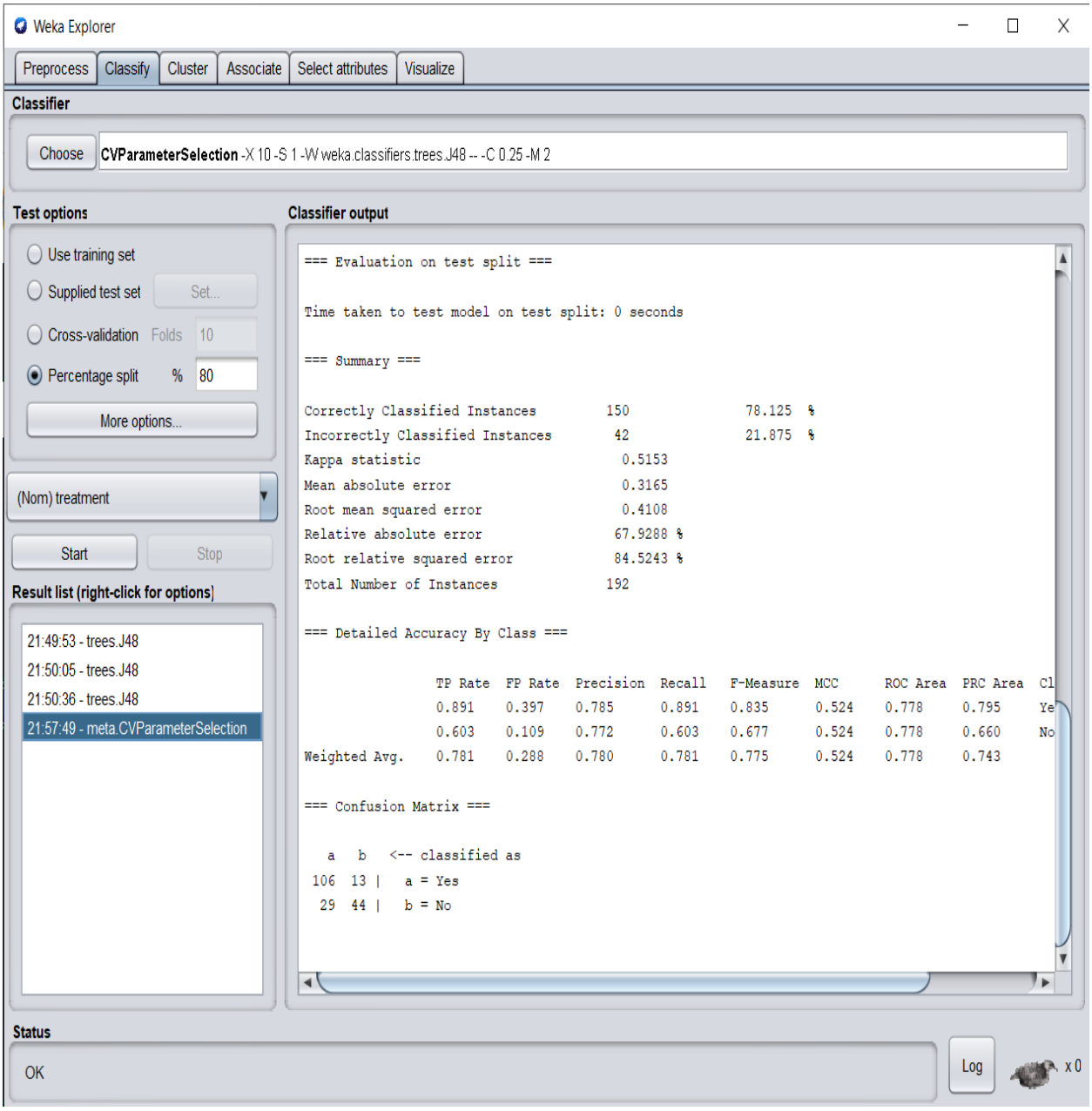
Where Decision Tree Classification results were, Total Number of Instances are 326 from which Correctly Classified Instances and Incorrectly Classified Instances are 257 and 69, respectively. We also find the Mean absolute error which is 0.3297. Alongside Root mean squared error 0.4037. Accuracy with uprunning set to false - 77.83% and with uprunning set to true - 78.8344%.



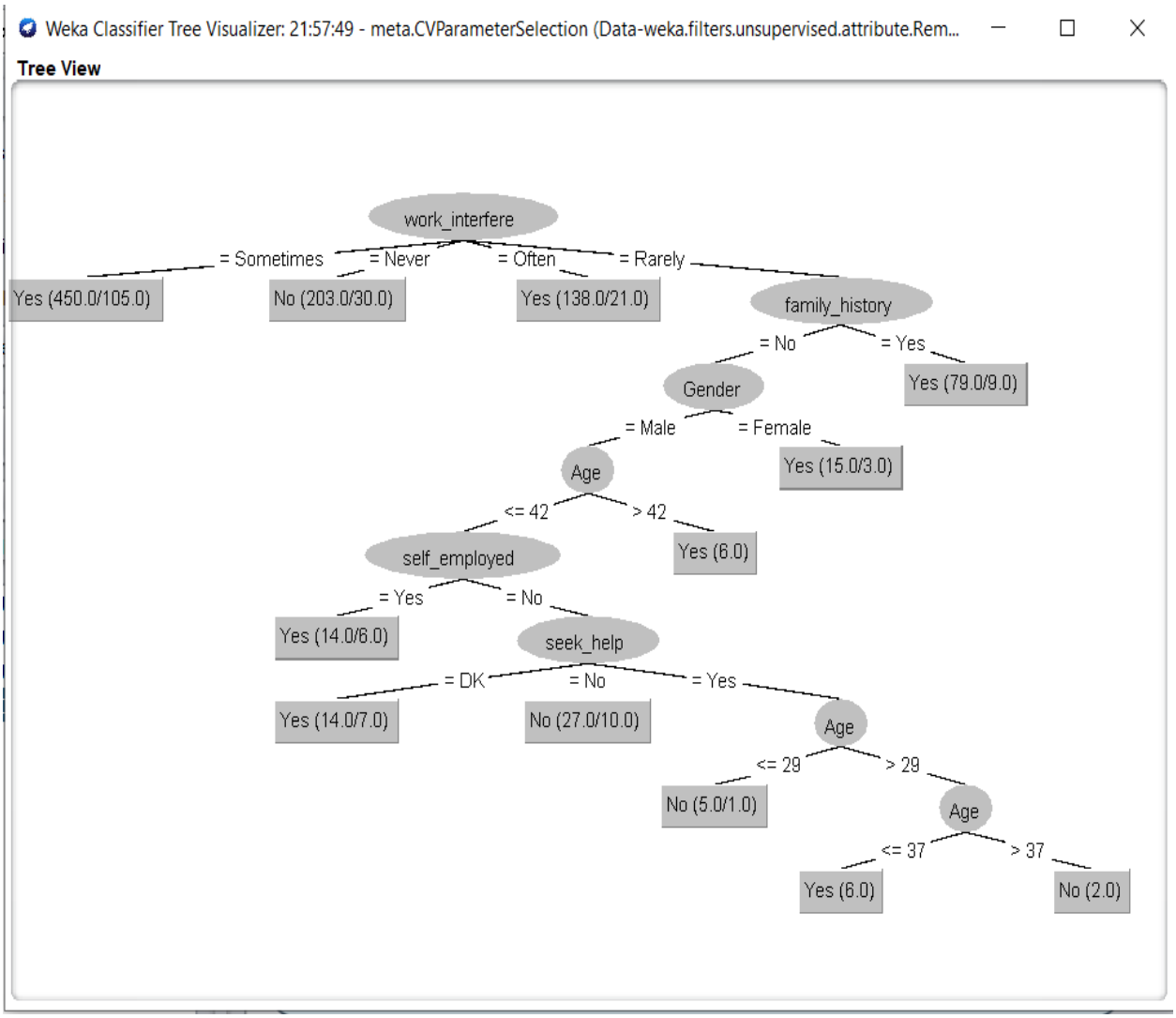
To improve the accuracy for Decision Tree, we used CV Parameter selection optimization technique.

CV Parameter Selection:

We divided 80% of dataset as training set. Obtained accuracy of 78.125% for Correctly Classified 150 Instances. Whereas, 21.875% for Incorrectly Classified 42 Instances. The precision and recall values we found by this technique are 0.78 and 0.781 respectively.



Below is the optimized decision tree we found using CV Parameter Selection. For decision tree, observations that we found is work interfere is a parameter where data is being split and the weightage of attributes are as shown in figure.

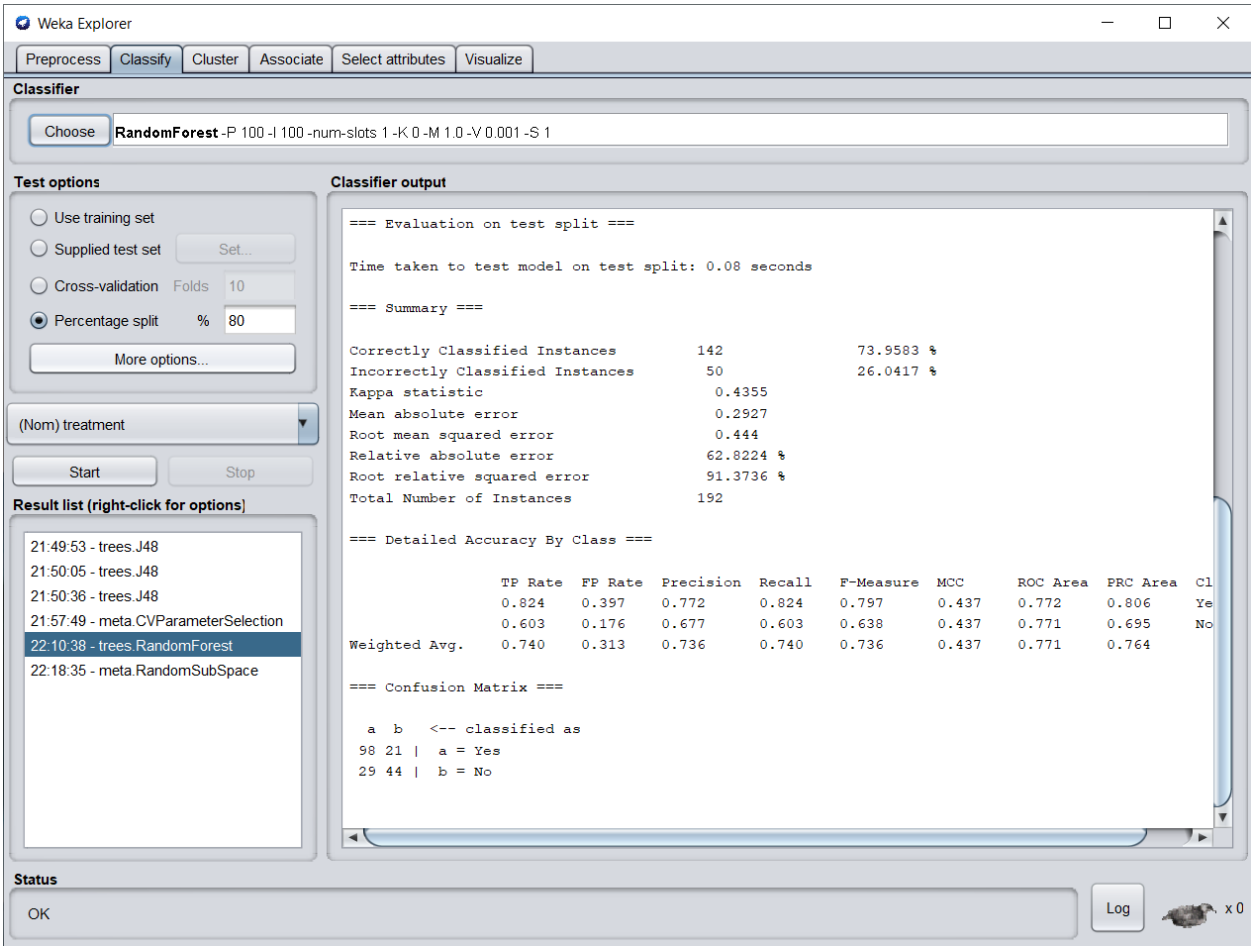


We observed that, Decision Tree has more accuracy 78.8344% than the optimization technique CV Parameter Selection which is 78.125%. We came to the conclusion that the Decision Tree is already optimized.

- **Random Forest:**

After performing the classification, we observed that Correctly classified Instances are 142 with the accuracy of 73.9583%. And Incorrectly Classified Instances are 50 with inaccuracy of 26.0417%.

Total Number of Instances are 192.

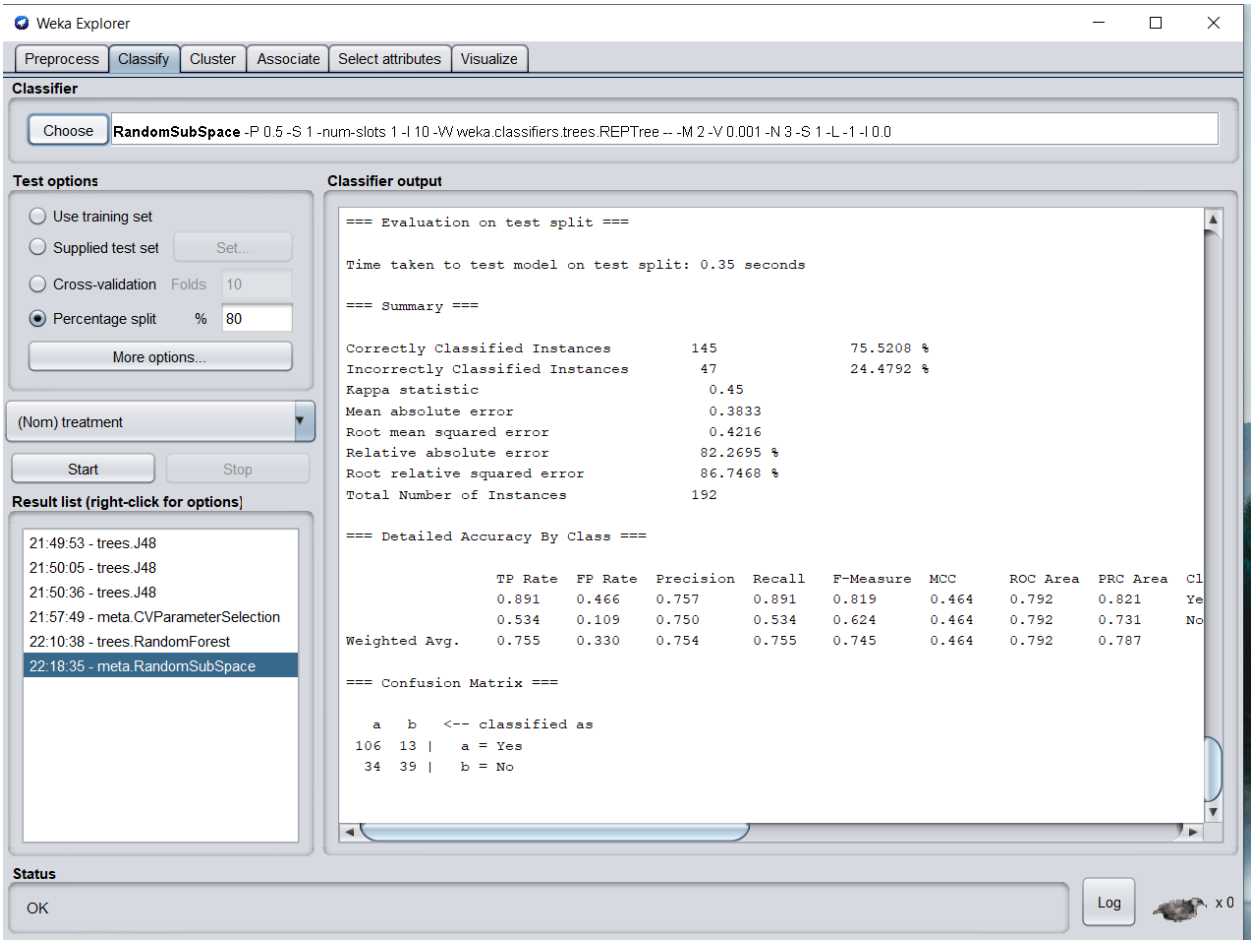


Here, the Precision Value is 0.736. Also, 98 True(yes) are instances, 29 are False(No) instances and 44 are true negative instances.

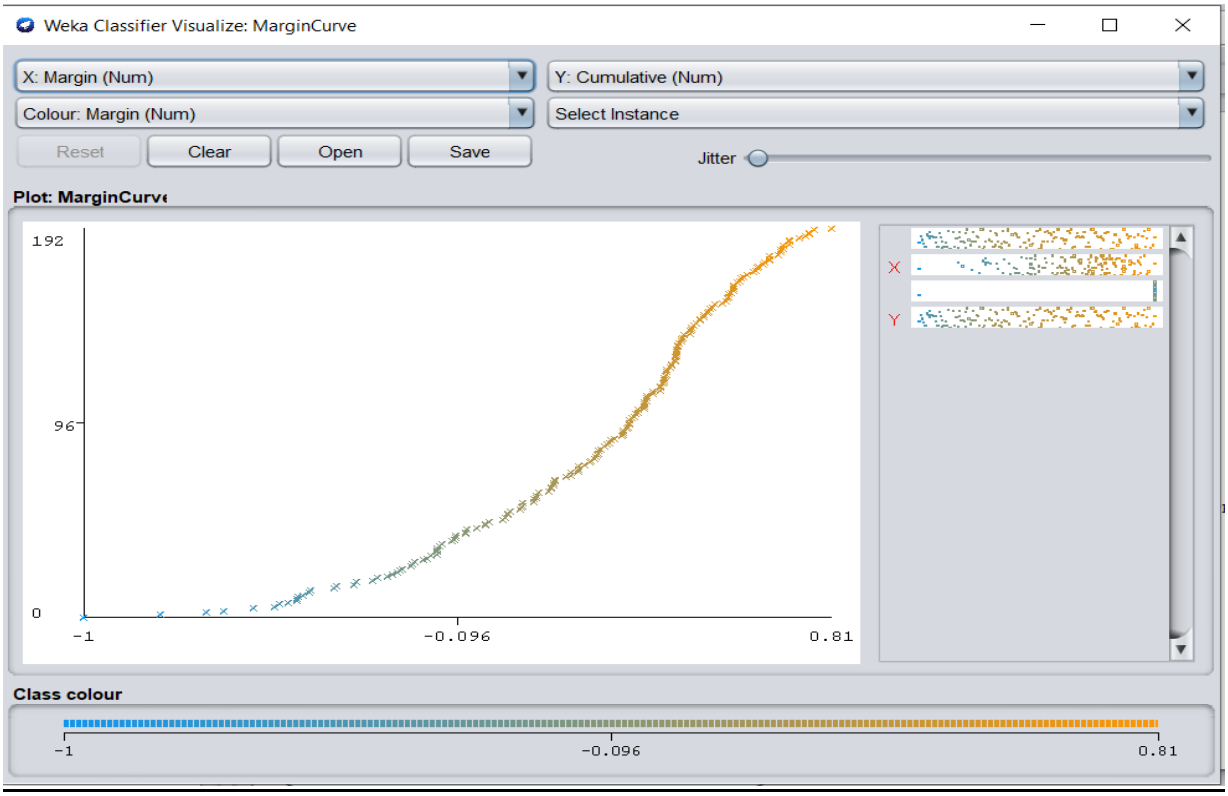
To improve the accuracy for Random Forest, we used Random Subspace Optimization technique.

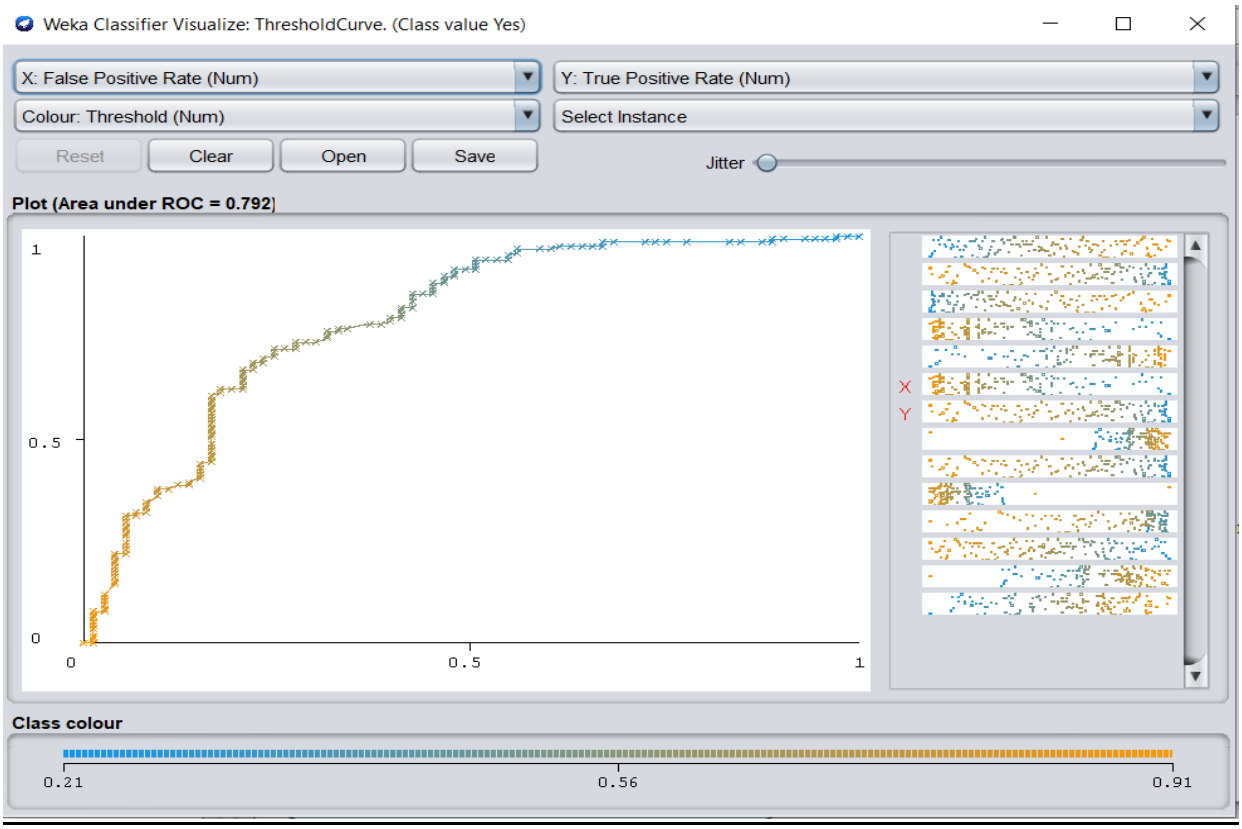
Random Subspace:

We divided 80% of dataset as training set. Obtained accuracy of 75.5208% for Correctly Classified 145 Instances. Whereas, 24.4792% for Incorrectly Classified 47 Instances. The precision and recall values we found by this technique are 0.754 and 0.755 respectively.



In Random SubSpace, we are also calculating margin curve and threshold curve for better understanding. The Area under ROC for threshold is 0.792 model.


Weka Classifier Visualize: MarginCurve
X: Margin (Num)
Y: Cumulative (Num)
Colour: Margin (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: MarginCurve
192
96
0
-1
-0.096
0.81
X
Y
Class colour
-1
-0.096
0.81


Weka Classifier Visualize: ThresholdCurve. (Class value Yes)
X: False Positive Rate (Num)
Y: True Positive Rate (Num)
Colour: Threshold (Num)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot (Area under ROC = 0.792)
1
0.5
0
0
0.5
1
X
Y
Class colour
0.21
0.56
0.91

By using Random Subspace, optimization technique, we were able to improve the accuracy rate from 73.9583% to 75.5208%.

Conclusion:

By performing the optimization phase, we were able to improve the accuracy of the data model for different data mining techniques such as Naïve Bayes, Decision Tree, and Random Forest. We observed the increase in accuracy by using optimization techniques Multiclass Classifier (79.16%), and Random Subspace (75.52%). However, for CV Parameter Selection the accuracy (78.125%) was less than the original Decision Tree Classification which has an accuracy of 78.8344%. Therefore, we conclude that accuracy values are not always final, you can improve accuracy by using different optimization techniques.

Repository:

https://github.com/vsala2/DataMining