**Methodology**

**Data Preprocessing**

We did not perform Data Preprocessing techniques like discretization, aggregation, and feature selection since the dataset consisted only of character type and all attributes were equally important for model creation.

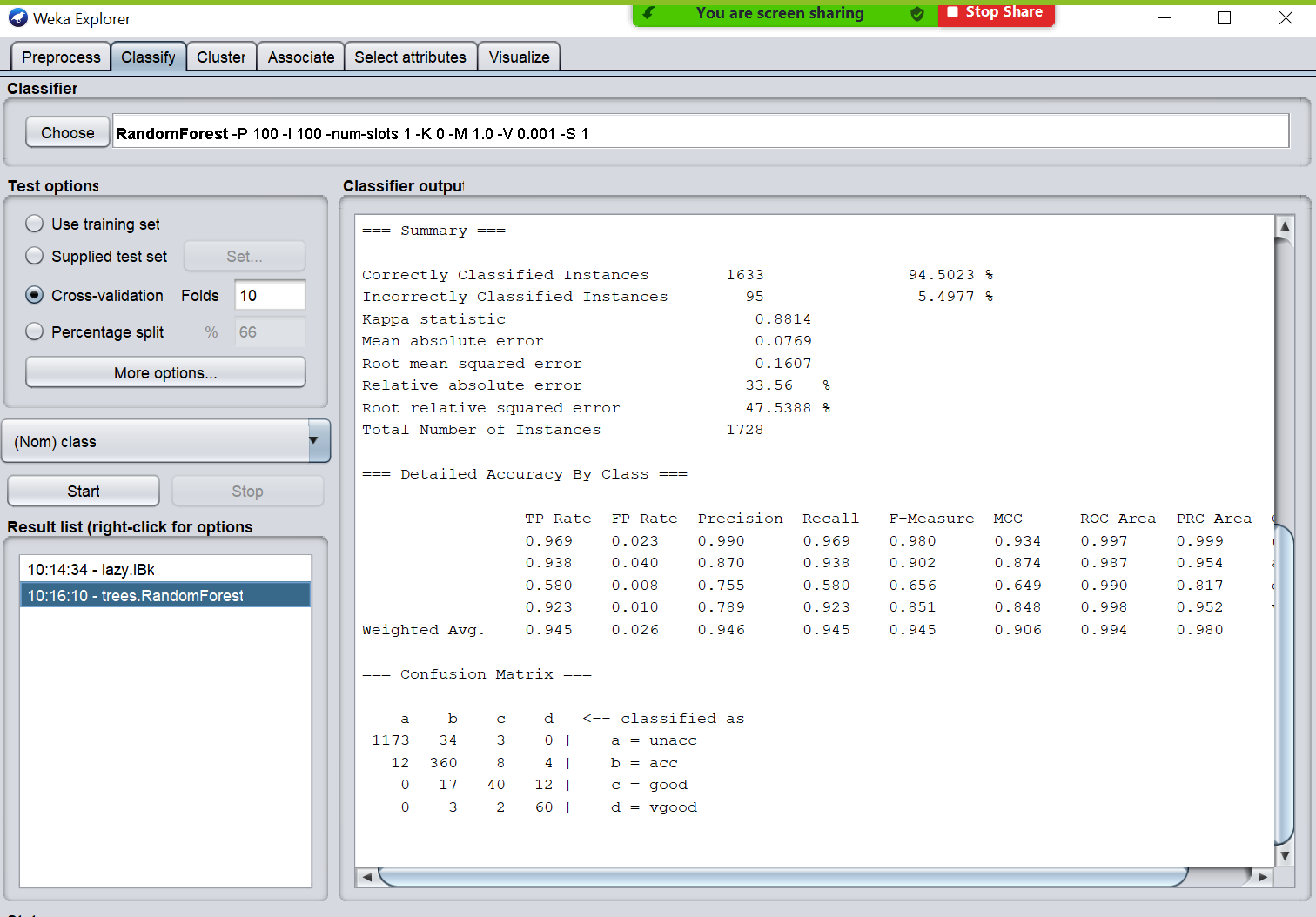
**Mining the data**

Data mining techniques

We have chosen Random Forest and KNN Data mining techniques

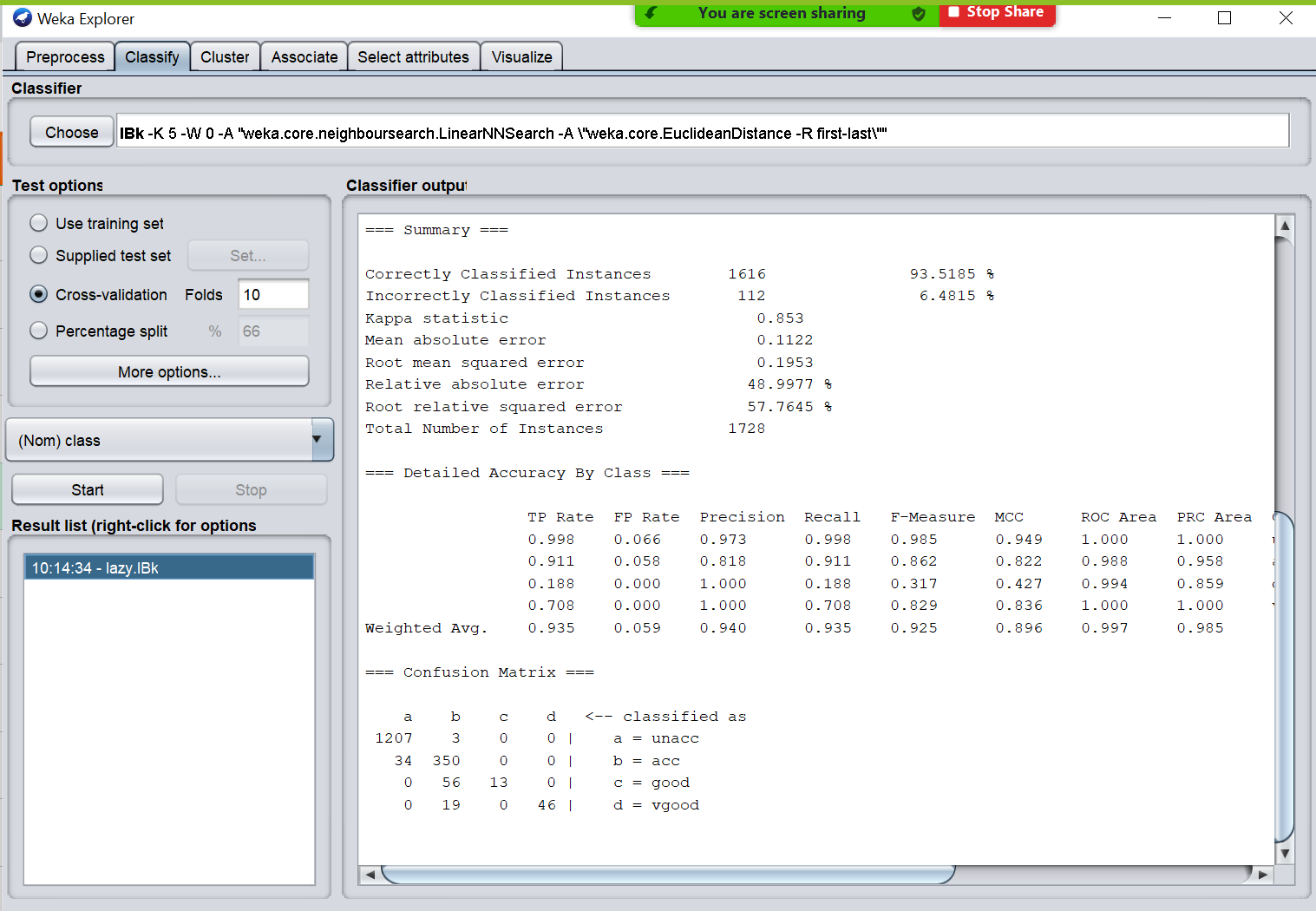
Random Forest

Random Forest is an ensemble classification Data mining technique which builds multiple Decision trees and takes the average of all the Decision trees to predict the class. We tuned parameters like batchsize, seeds and a few others, but at the end, the result like correctly classified instances, ROC Area and confusion matrix did not change. Random Forest has the highest accuracy among many other tested algorithms, and accuracy alone isn’t enough to say that it suits our dataset. Based on the below figure, all the measures like TP Rate, FP rate, Precision, F-Measure and a few others prove that Random Forest is a good fit for our dataset.



K-Nearest Neighbors

K-Nearest Neighbors is a classification Algorithm, as the name suggests the class predicted based on the given number[K] of Nearest Neighbors. We tuned parameters like K-value, nearestNeighbourSearchAlgorithm, and a few others, but the results did not change unless we give extreme K-values, which is not suggested for K-Nearest Neighbors Algorithm. Has the second highest accuracy after Random Forest among other tested algorithms. We also got a precision of 1 for classes good and verygood, also the ROC Area is 1 for classes unacceptable and verygood. We believe that IBk model is also a good fit for our dataset as it illustrates satisfactory model performance values.



**Models Performance**

We used Correctly classified instances(accuracy), ROC Area and Confusion Matrix to evaluate the Model Performance.

Correctly classified instances(accuracy) is the measure of the number of correctly classified instances comparing the model’s predicted class to the actual class after the model is trained.

Receiver operator Characteristics curve (ROC) is a graph showing performance of a classification model which is obtained by plotting false positives rate against true positives rate for different thresholds. Greater the area under the curve (AUC), better is the efficiency.

Confusion Matrix is the table that also describes the Models performance, comparing all the predicted classes with the actual class, each row in the confusion matrix represents the actual class, and each column represents the predicted class. The diagonal elements of the matrix are the correctly classified instances for the given class, the rest are misclassified instances.

**Motivation behind choosing the Models Performance**

We believe that the measures which have been chosen depicts the true performance of the Models. ROC Area gives a better idea about accuracy. Confusion Matrix gives a clear picture of the Models Performance, as it reveals how many instances are incorrectly classified and which class they belong to.

**Performance measures estimation**

We chose Cross validation and gave 10 as the input for the number of folds. Cross Validation will ensure that all instances in the dataset are being used as Test and Training data, by this way the entire dataset will be used for both training and testing purposes.

**Performance of Different modes**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Options** | **Training set** | **Percentage Split - 66%** | **10-fold cross validation** |
| IBk | 100% | 92.675% | 93.457% |
| Random Forest | 100% | 92.517% | 94.502% |

The training set results are always expected to be higher and most of the time the expected accuracy is 100% since the same set is used for both training and testing purposes, which evidently result in higher accuracy and misleading. The Percentage Split is done at 66%, which means that 66% of the dataset is used for training purposes and 34% of the dataset is used for testing purposes. Although Percentage split may seem a better mode to evaluate the performance, we may include a few outliers in the training set to make overfit and model or underfit the model be excluding some crucial instances while training the model. Cross validation is by far the best way to evaluate the model as it uses the entire dataset for training and testing purposes.