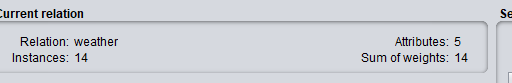
**(Using Weather-train.arff):**

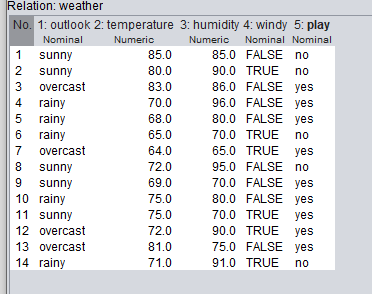
**Question No.1(Data Loading and Analysis):**

1. Size of Dataset= No. of Entries = 14
2. Size Attributes: 5



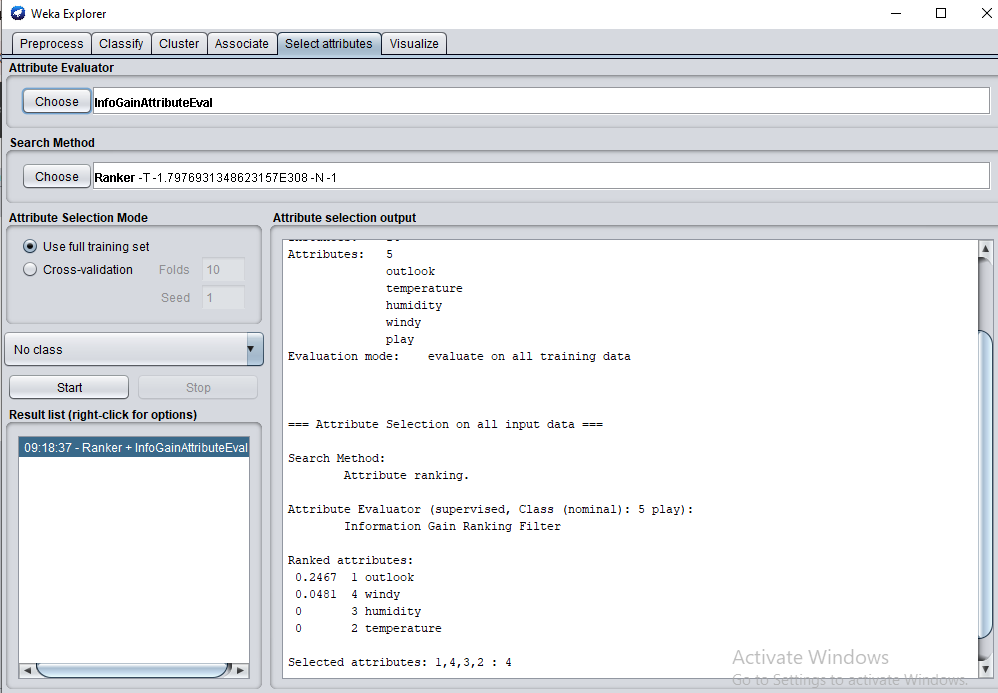
1. Instances(play):

* Positive (yes): 9
* Negative(no): 5



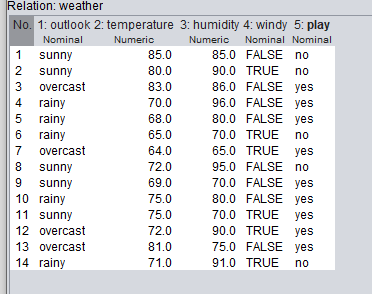
1. Separation of Data:

As the information gain of attribute “Outlook” is highest so we can separate the data on the basis of it.

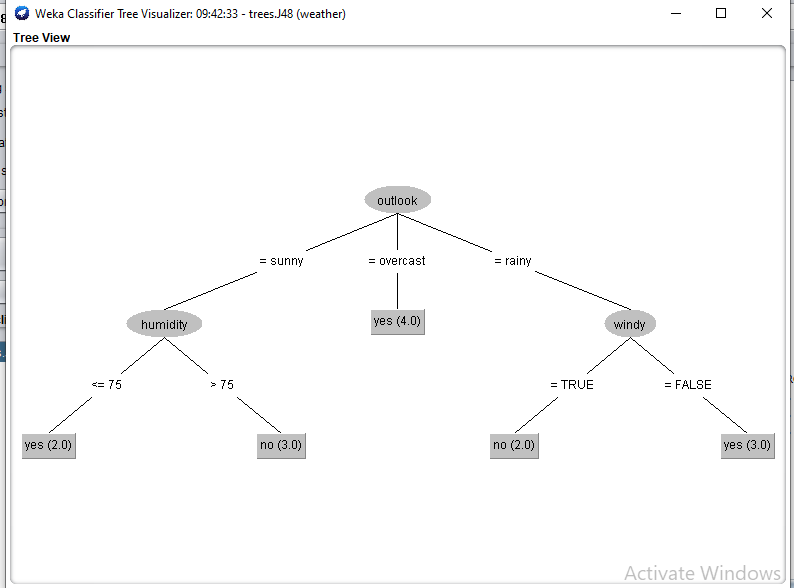


1. Instance (Windy = High):

Number : 6



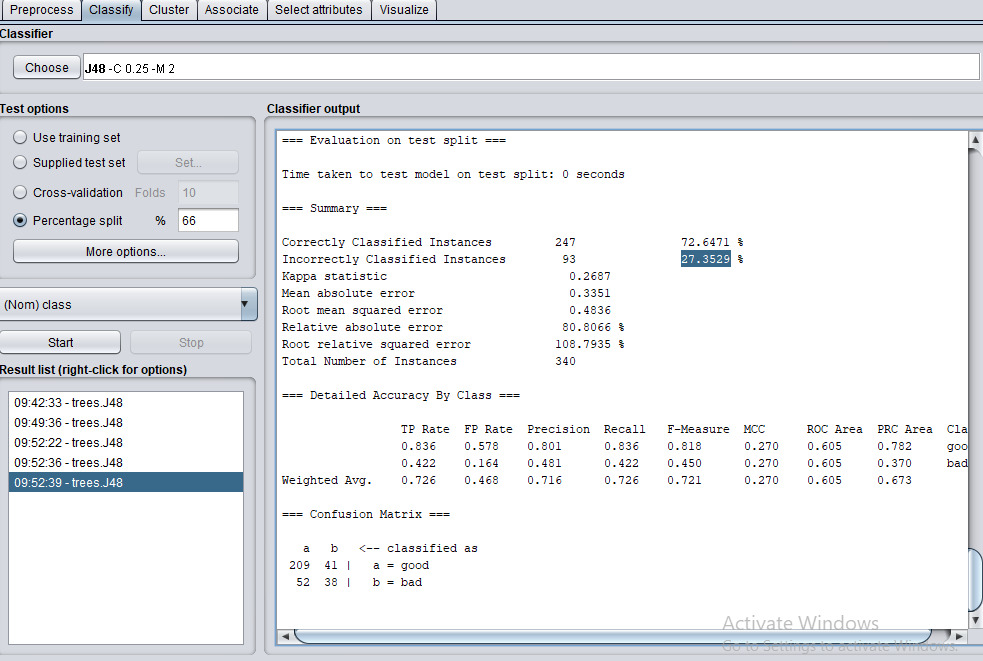
**Question No.2(Load and Analyze Data Set):**

****

**Question No.3(Classification Accuracy):**

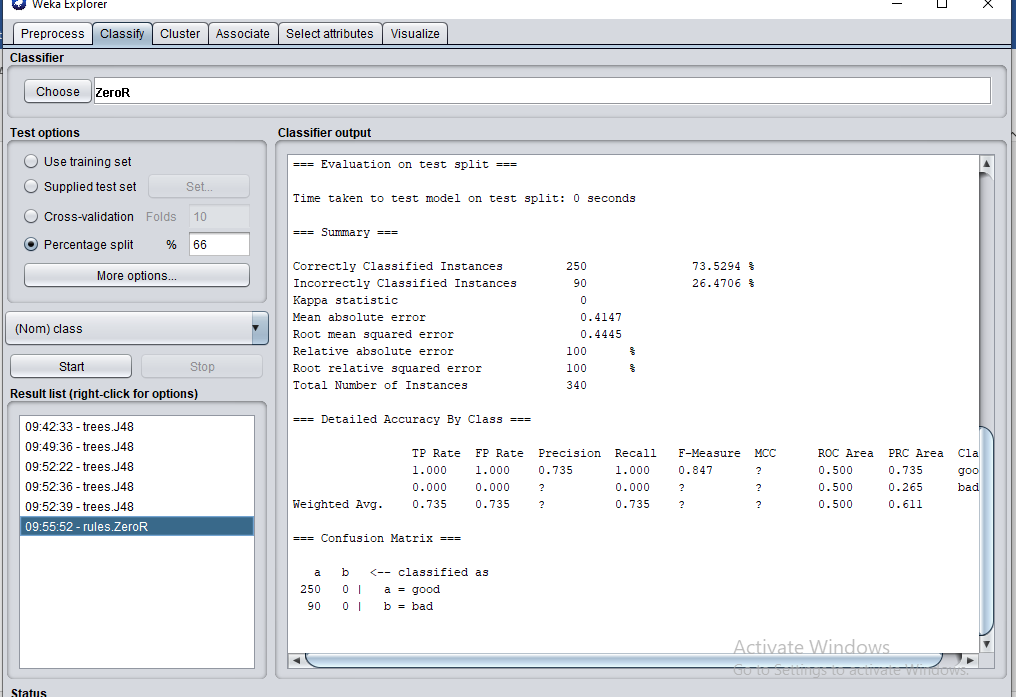
**(Using credit-g.arff)**

72.6471% instances are correctly classified whereas 27.3529% of instances are incorrectly classified. The results are good as more than half of the data is correctly classified.(Using J48 Classifier).



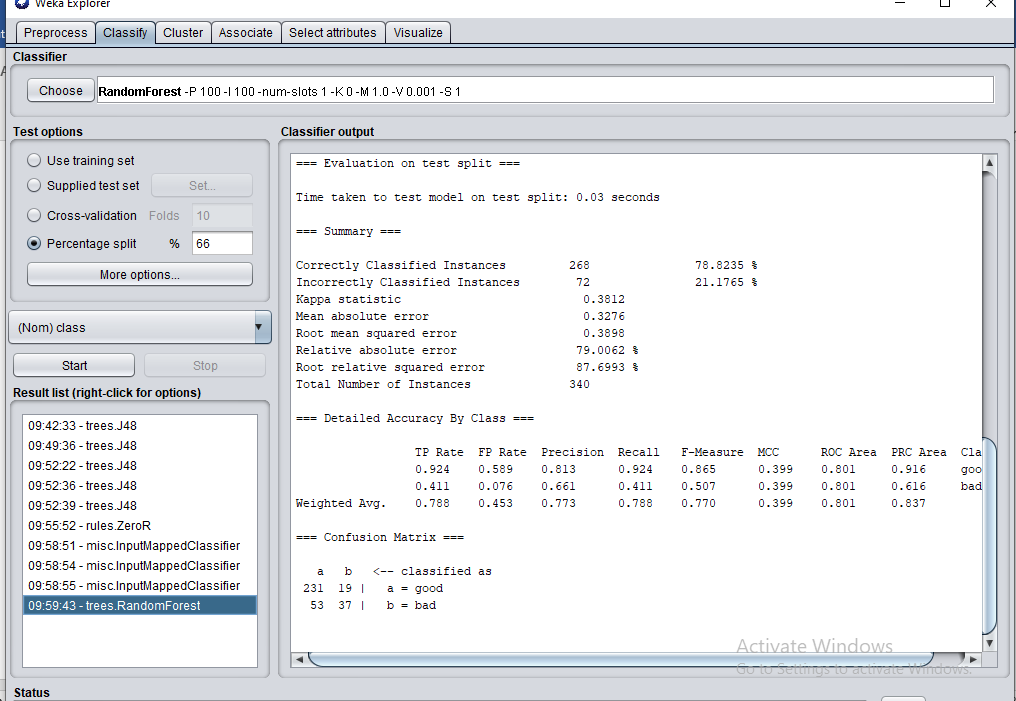
(Using ZeroR classifier)

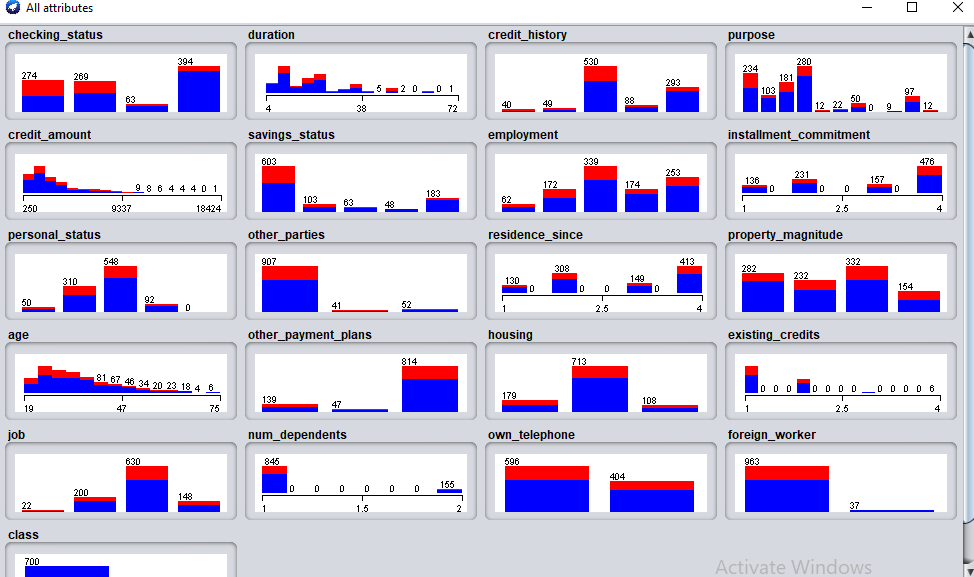
The result is better than the J48 Classifier as the correctly classified instances increases by approx. 1%

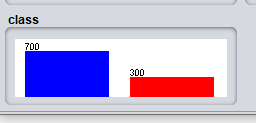


Other Classifier:

We have used Random Forest Classifier and It is better than both the above classifiers as the correctly classified instances increases by 5-6% .

Class Imbalance:





Only the attributes “personal status” , “job” etc are normally distributes. Other attributes like property\_magnitude is mixed. Attributes like “age”, “purpose”, “credit\_amount”, “savings\_status” are right skewed making them imbalance.

The attribute imbalance causes the intervals to be inaccurate because of biasness towards the attributes.

**Why is it worth taking a look at the data before attempting a classification task?**

It is important as if there are any missing values or outliers etc then it will classify the data/attribute incorrectly. So it is important to go through the data so that if there is a need of pre processing so it is performed before the classification