**---------------------------------------------------------------------------------------------**

**DWM EXPERIMENT NO. : 07** **Roll no :50**

**---**-**-----------------------------------------------------------------------------------------**

**Problem definition :**

To perform To perform association rule mining, classification on data sets using the WEKA machine learning toolkit.

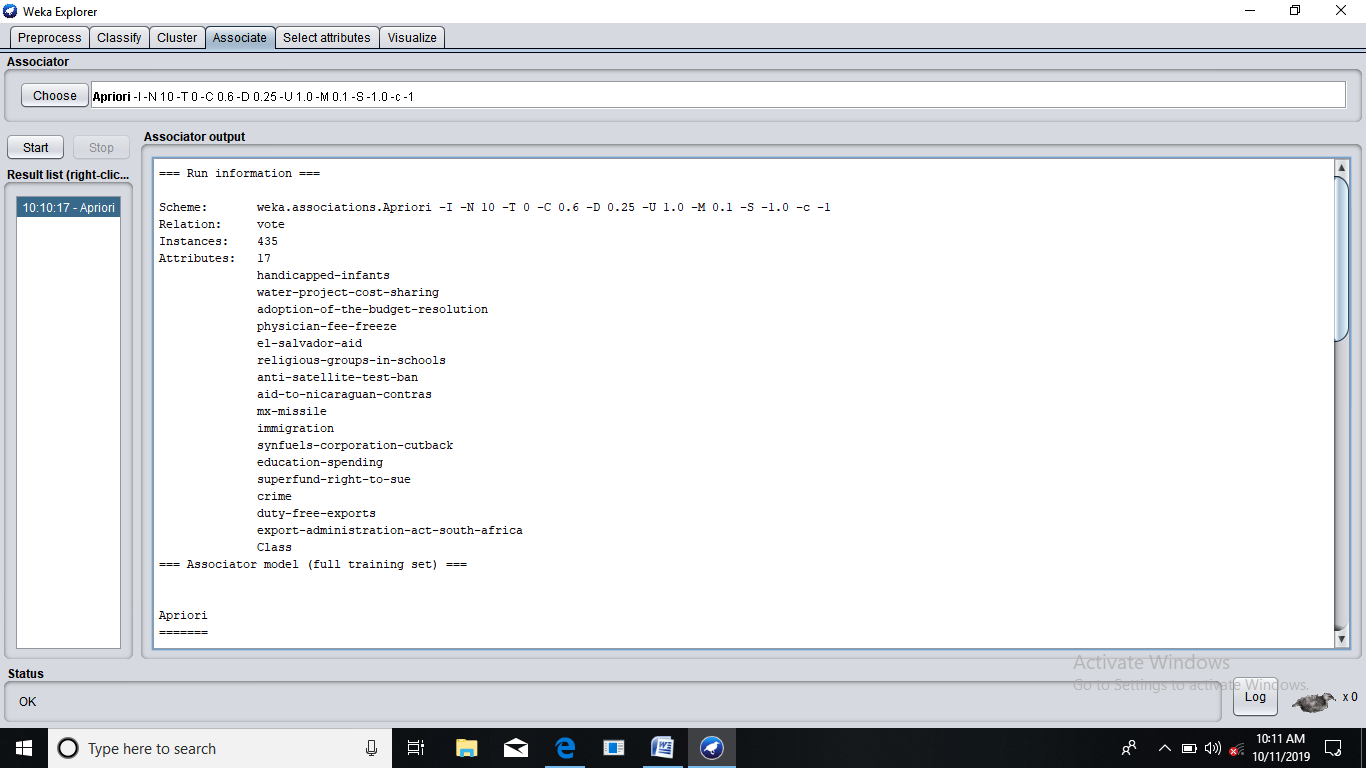
---------------------------------------------------------------------------------------------

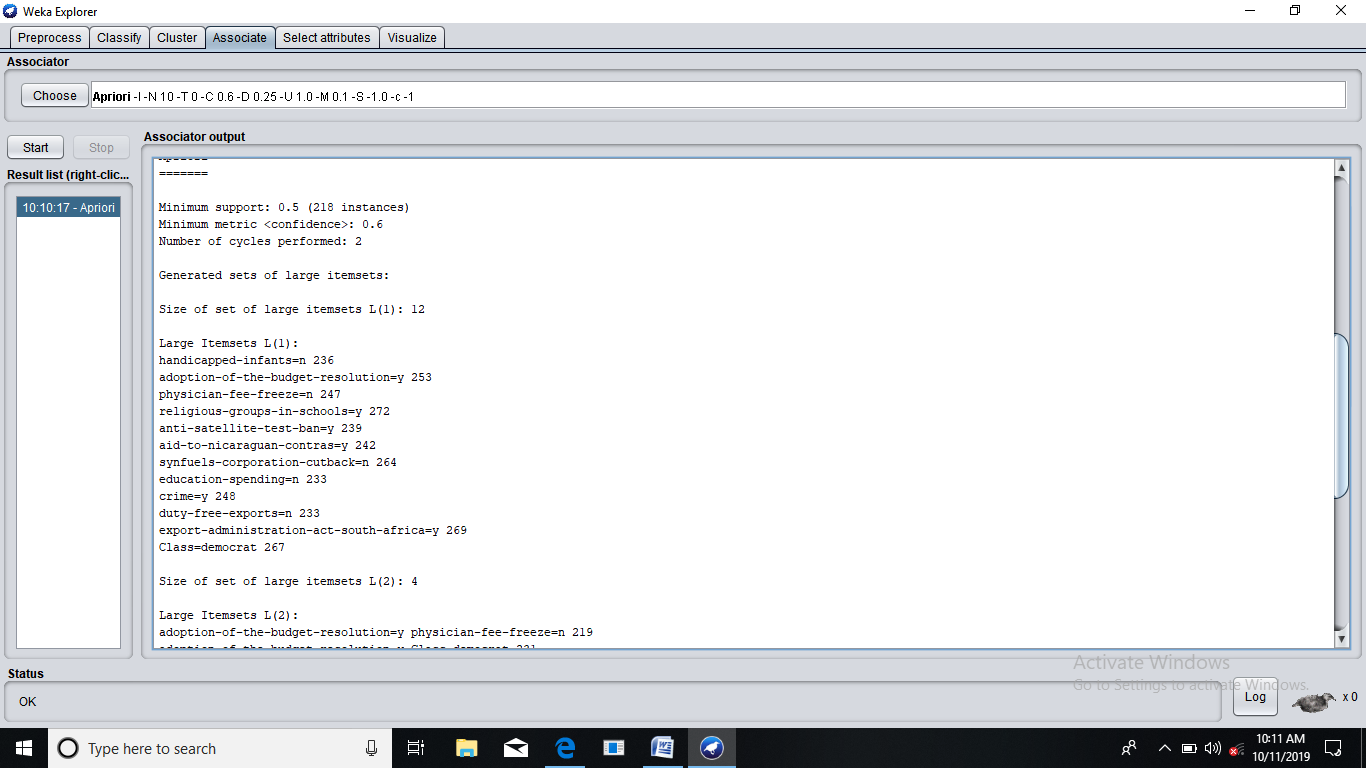
**Association Rule Mining**

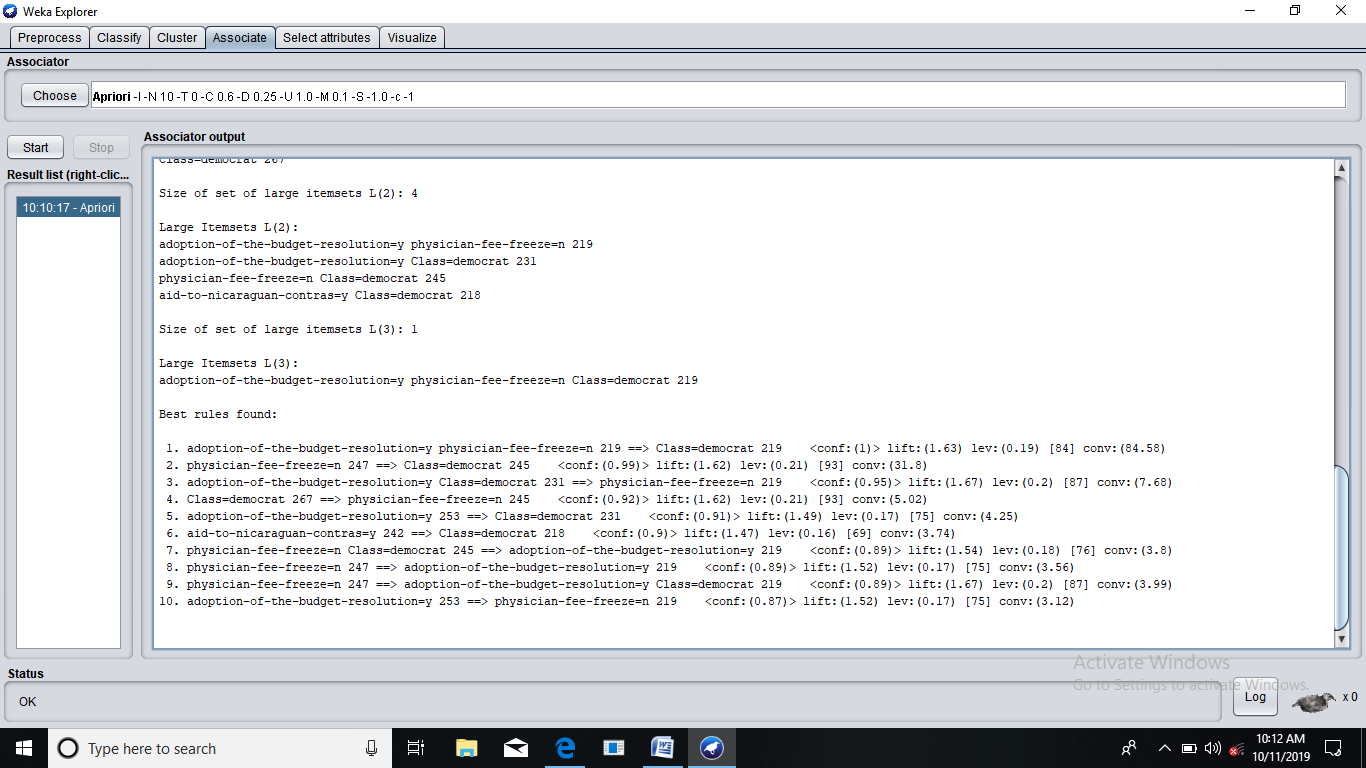
**-----------------------**

**Q1. Perform the following tasks:**

* **Load the ‘vote.arff’ dataset**
* **Apply the Apriori association rule**





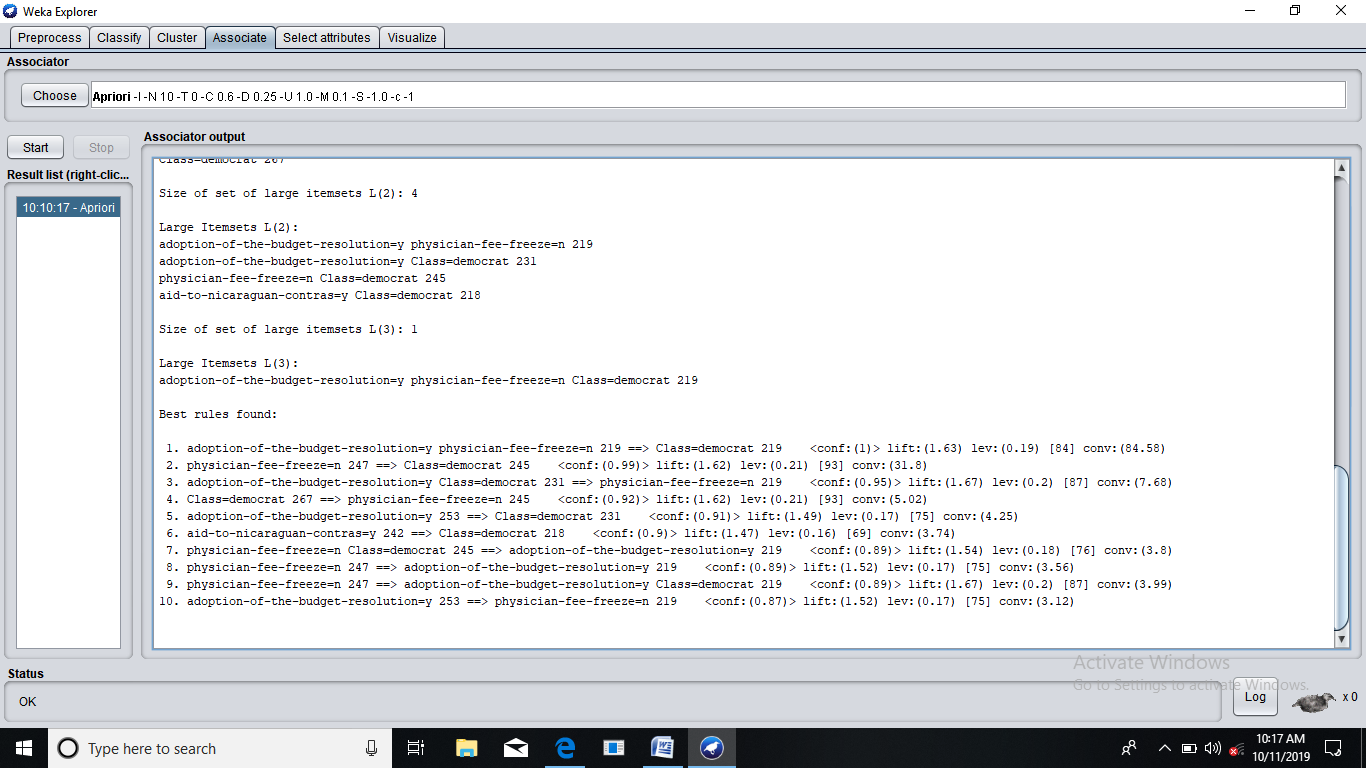


* **What is the support threshold used? What is the confidence threshold used?**

Support threshold : 50%

Confidence threshold : 60%

* **Write down the top 6 rules along with the support and confidence values.**

****

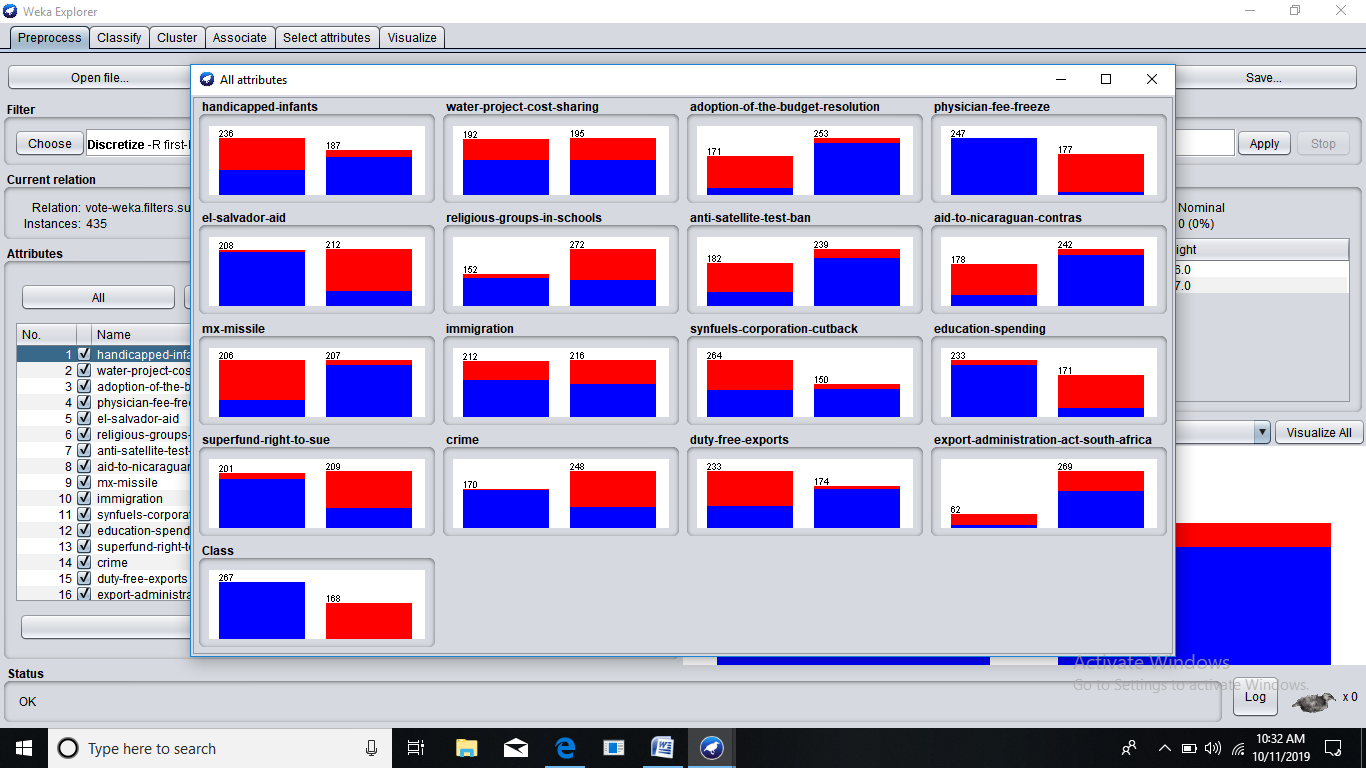
* **What does the figure to the left of the arrow in the association rule represent?**

The LHS of the arrow represent the occurrence of a case, or an event.

* **What does the figure to the right of the arrow in the association rule represent?**

The RHS indicates the implication of LHS or in simple words it indicates what event will occur as well when LHS occurs.

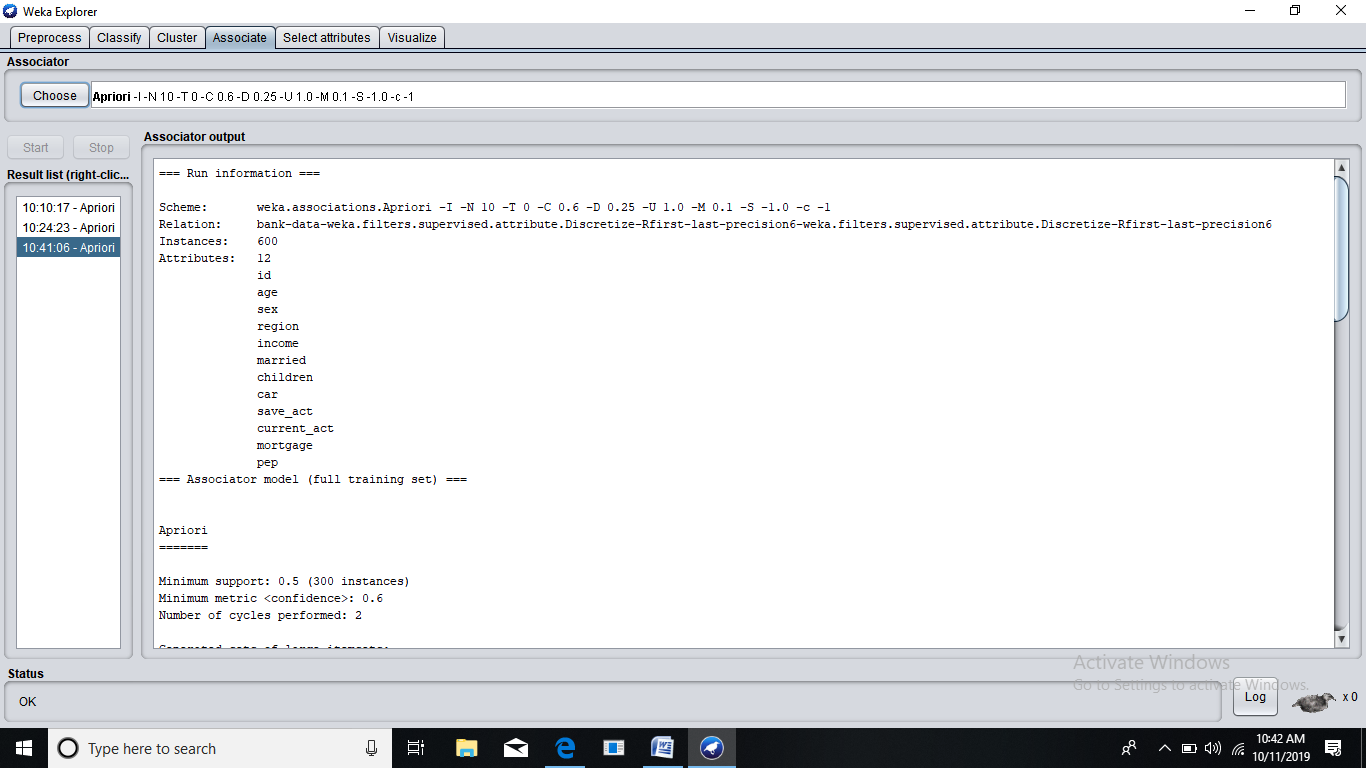
* **For rule 8, verify that numerical values used for computation of support and confidence are in accordance with the data by using the Preprocess panel. Then compute the support and confidence values. Are they above the threshold values?**

****

Yes, the data is in accordance with the computational support and confidence value. In some instances, data is above threshold values too, which helps in improving the association rule mining.

**Q2. Perform the following tasks:**

* + **Load the bank-data.csv file.**
  + **Apply the Apriori association rule algorithm. What is the result? Why?**

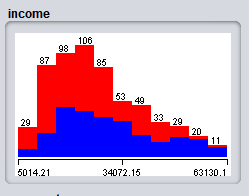
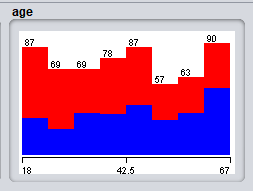
****

We can’t apply the apriori rule algorithm on this dataset as some values are numeric whilst some are nominal. Apriori works only on same type of data set which needs to be nominal. To achieve this, first numeric attributes need to be converted to nominal, using discretization filters.

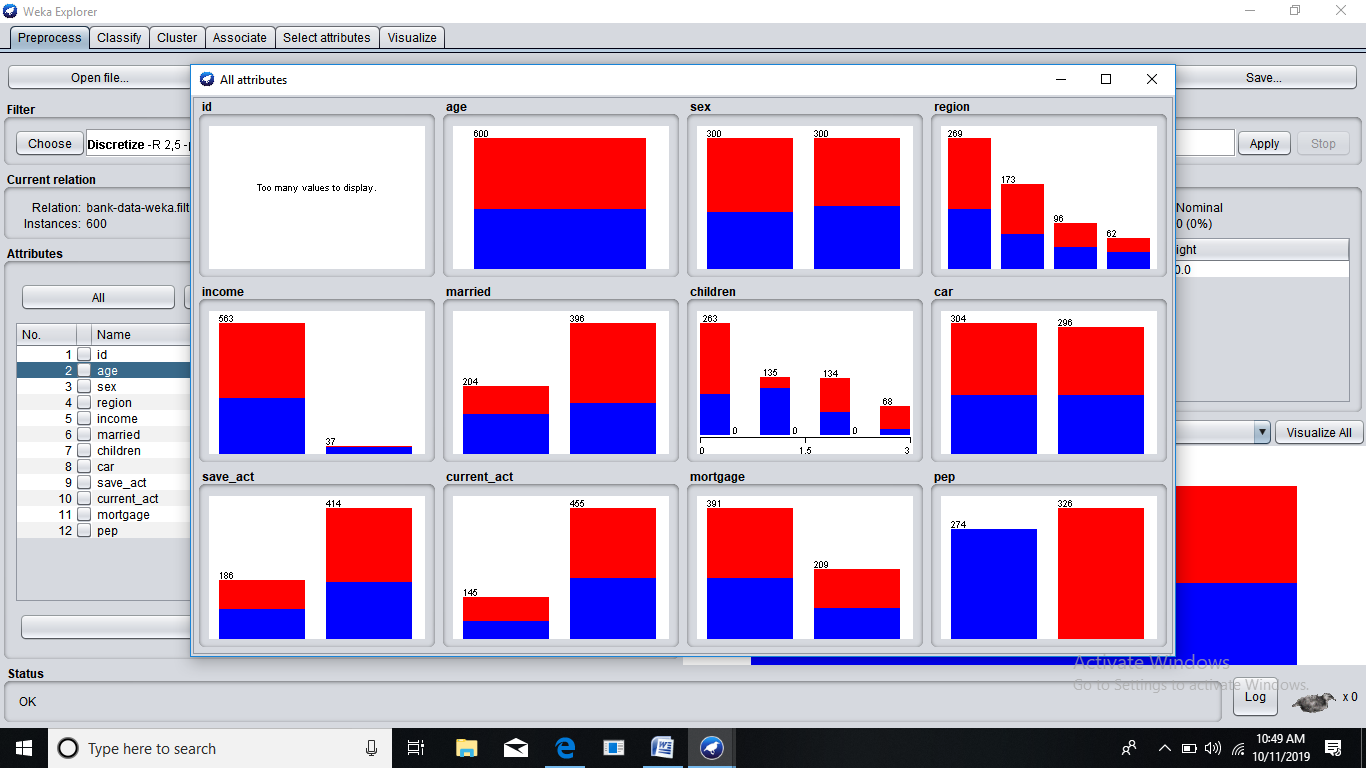
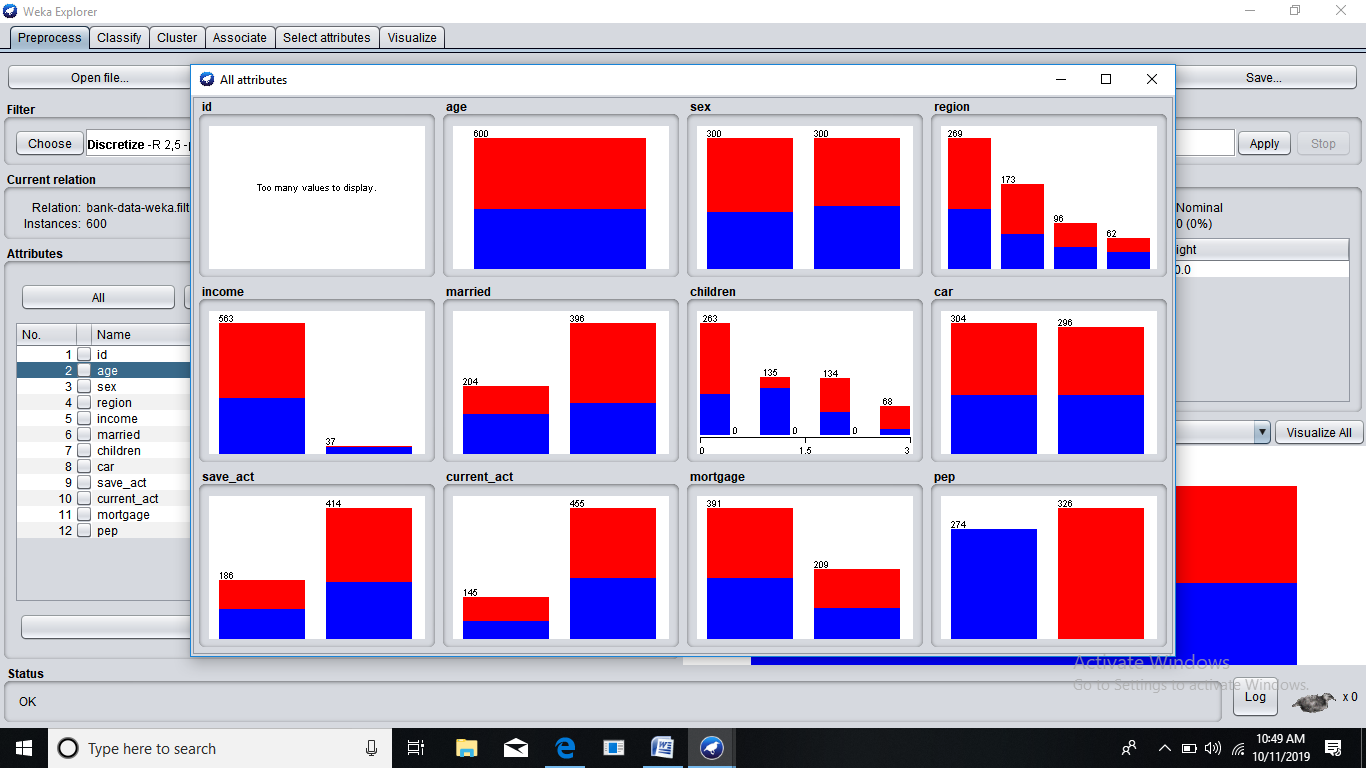
10 rules are generated after applying the association rule algorithm.

* + **Apply the supervised discretization filter to the age and income attributes.**

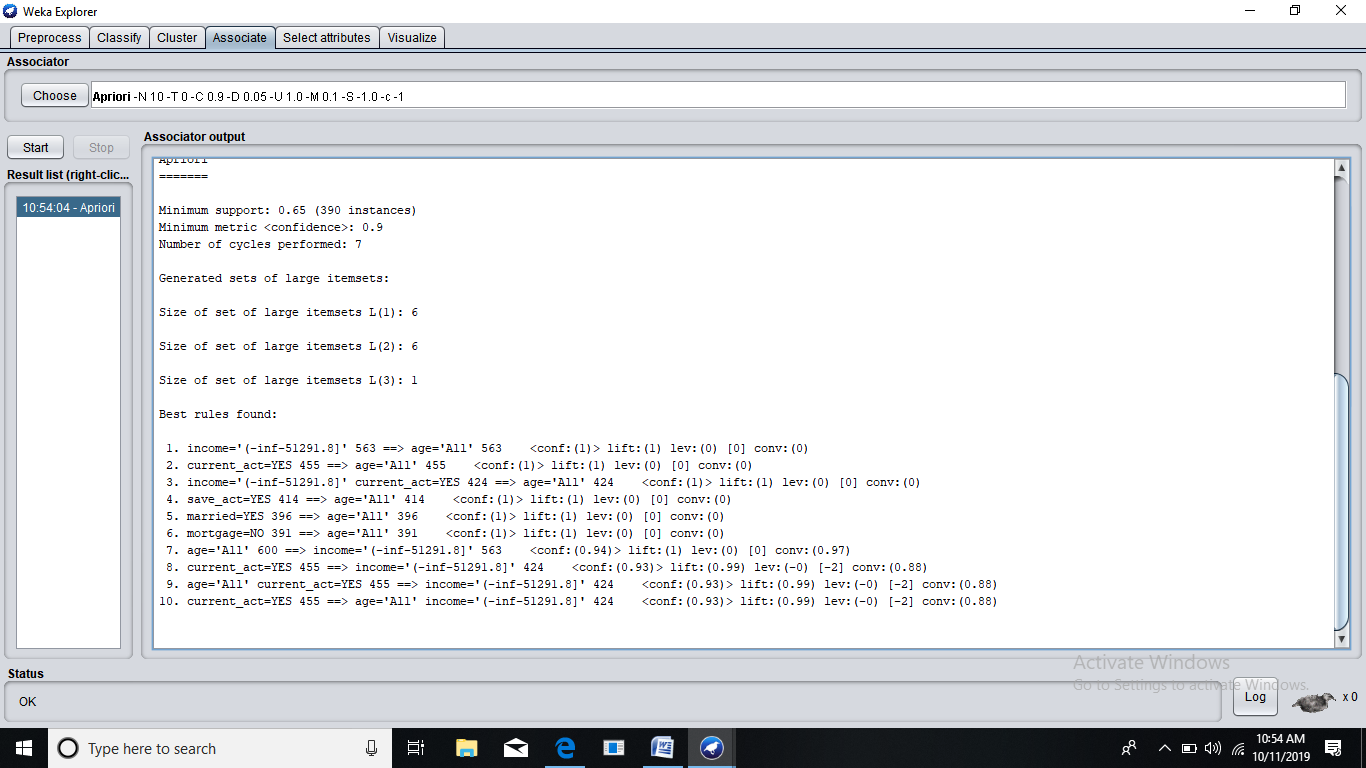
Before discretization :



After discretization :

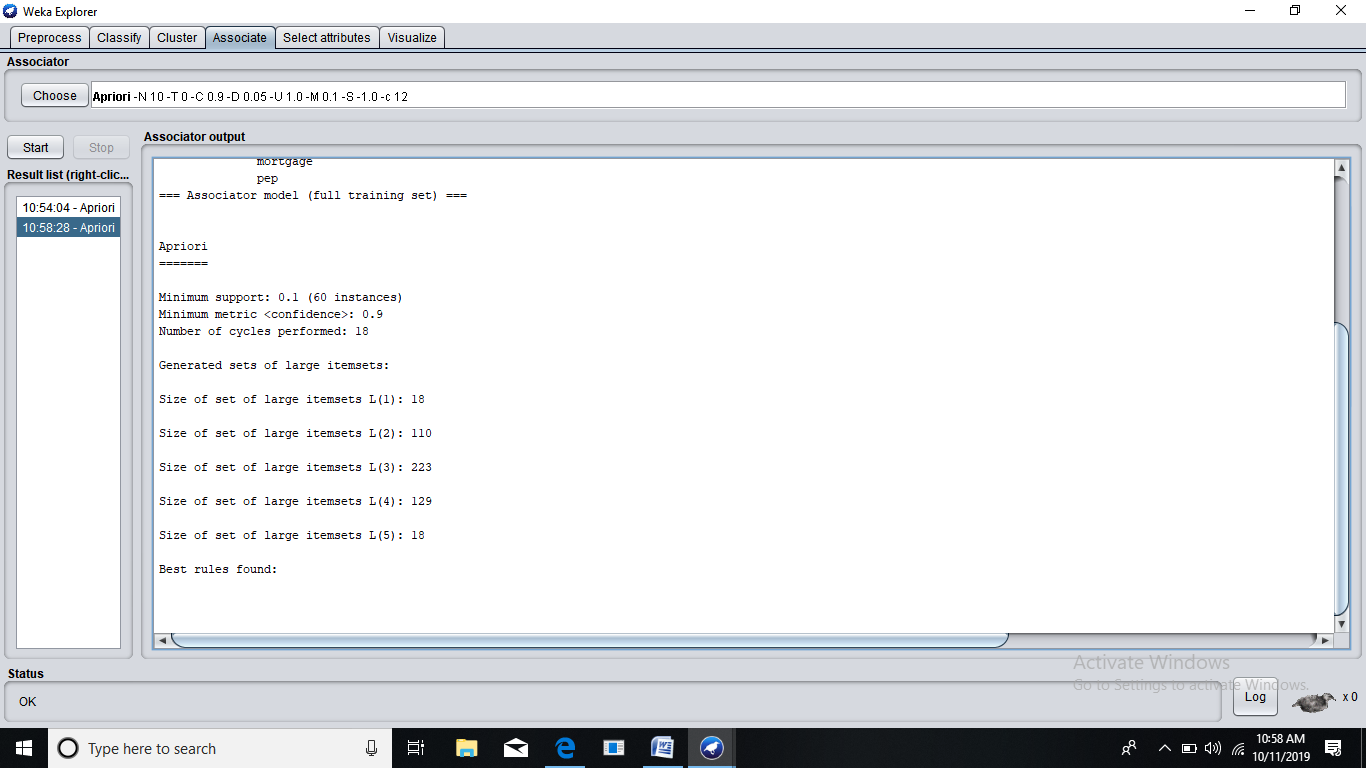


* + **Run the Apriori rule algorithm**
  + **List the rules that were generated.**

****

* + **4) Reload Bank-data.csv. Select only Nominal values.**
  + **5) Go to Associate Tab.**
  + **6) Select Apriori algorithm from “Choose “ button present in Associator**
  + **weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1**
  + **7) Select Start button**
  + **8) now we can see the sample rules.**

**>>NO RULES FOUND.**

****

**-------------------------------------------------------------------------------------**

**Classification**

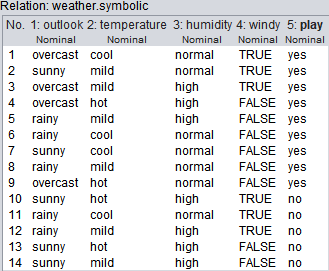
**-------------------------------------------------------------------------------------**

**Q1. Load the ‘weather.nominal.arff’ dataset into Weka and run J48 classification algorithm. Answer the following questions**

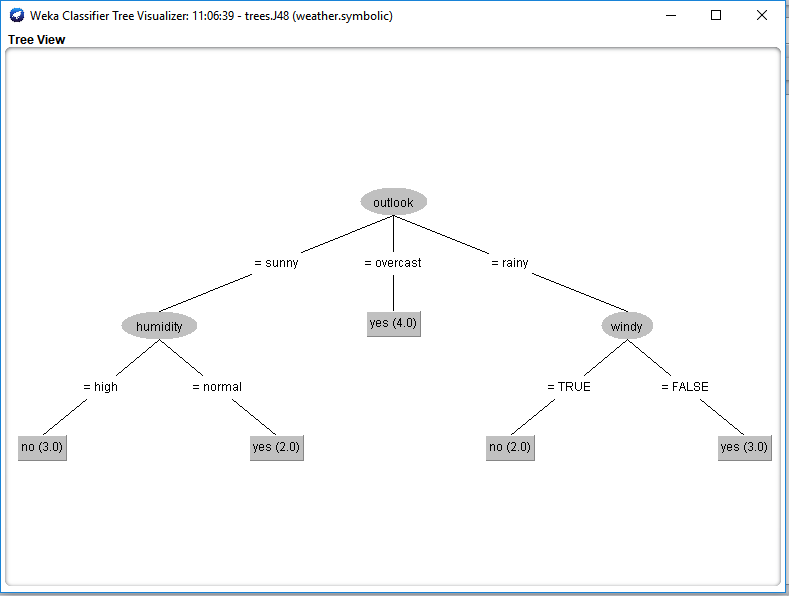
* + **List the attributes of the given relation along with the type details**

****

* + **Create a table of the weather.nominal.arff data**

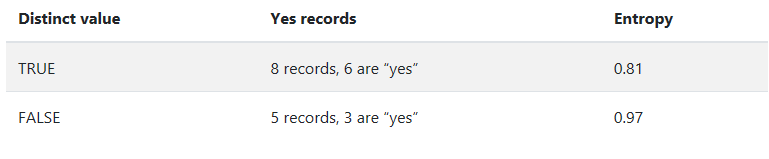
****

* + **Study the classifier output and answer the following questions**
    - **Draw the decision tree generated by the classifier**

****

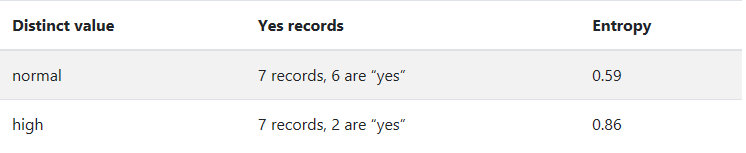
* + - **Compute the entropy values for each of the attributes**

1. **Windy**

****

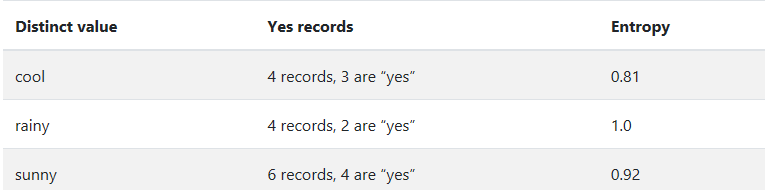
**Entropy = 0.87**

1. **Humidity**

****

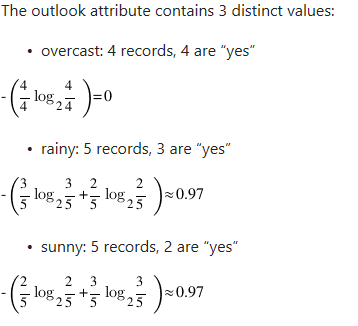
**Entropy = 0.72**

1. **Temperature**

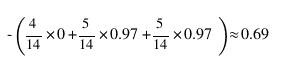
****

**Entropy = 0.91**

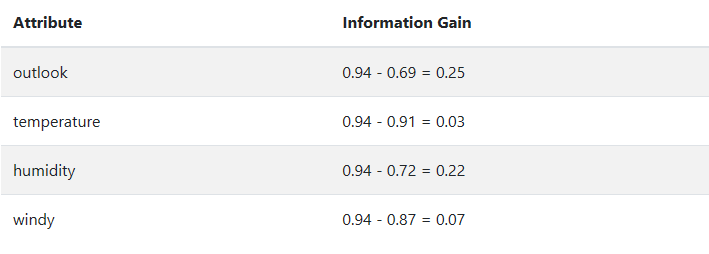
1. **Outlook L=**

****

**Entropy**

****

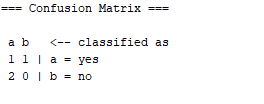
**Information Gain :**

****

* + **What is the relationship between the attribute entropy values and the nodes of the decision 9 tree?**

The higher is the information gain of a node, the more is the node having the decisiveness. The more is an entropy, the more is unpredictivity of a node. Hence, the lesser, the better.

* + **Draw the confusion matrix? What information does the confusion matrix provide?**

****

A confusion matrix provides the scenario statistics of false positives, false negative, true positive, true negative. In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm.

* + **Describe the Kappa statistic?**

The Kappa Statistic is the main metric used to measure how good or bad an attribute measurement system is. In the measure phase of a six sigma project, the measurement system analysis (MSA) is one of the main and most important tasks to be performed.

* + **Describe the following quantities:**
* **TP Rate**: rate of true positives (instances correctly classified as a given class)
* **FP Rate**: rate of false positives (instances falsely classified as a given class)
* **Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class
* **Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)

**Q2. Load the ‘weather.nominal.arff’ dataset in Weka and run the ID3 classification algorithm. What problem do you have and what is the solution?**

ID3 grows very quickly and sooner than you know it, it reaches a huge size with so many splitting points. This causes in a very High Variance model which trains so well on the training data that it often reaches an accuracy of 100% on the training data. But this is due to too much over fitting over the training set. So they lose their ability to generalize and hence when tested on the Testing Set, gives a very low accuracy score.

We can solve this issue with help of cross validation and percentage split to avoid overfitting.