**PRACTICAL NO. 7**

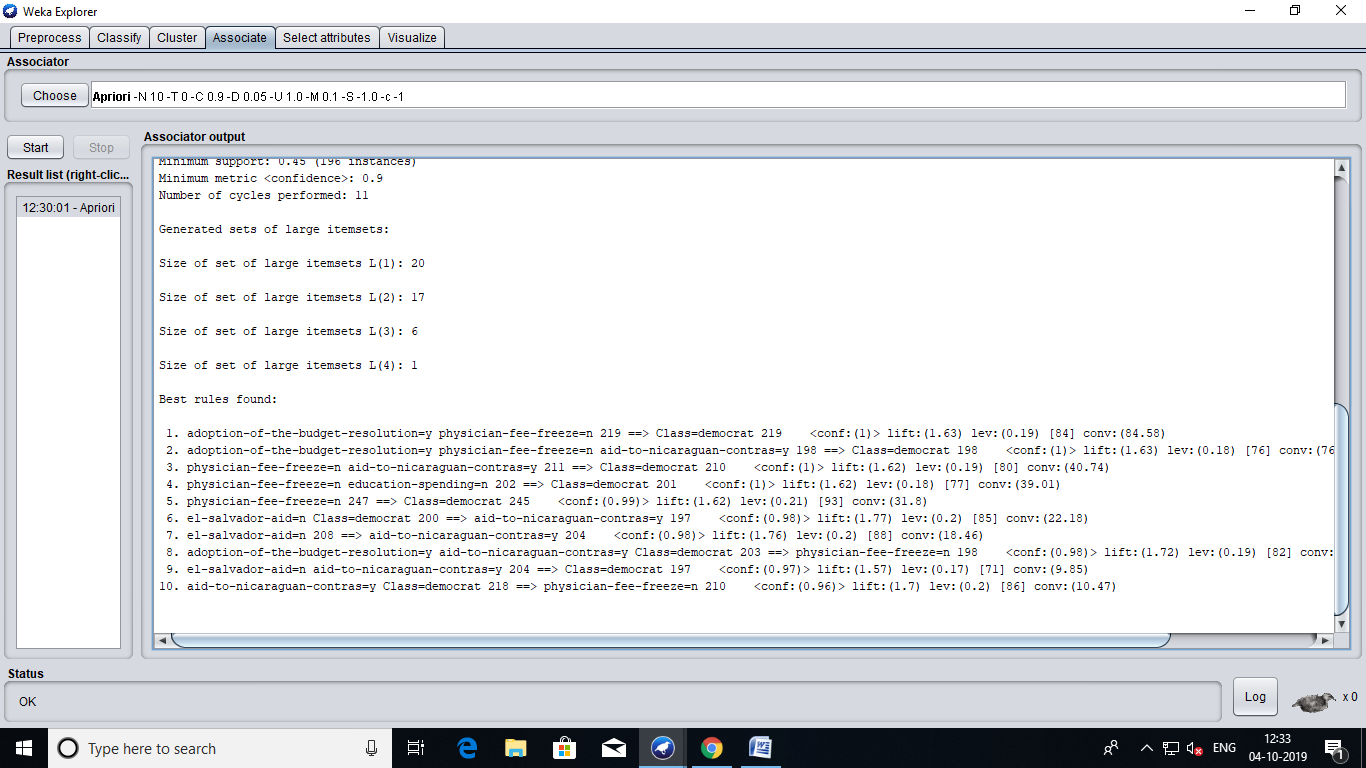
**Aim:** To perform association rule mining, classification and clustering on data sets using the WEKA machine learning toolkit.

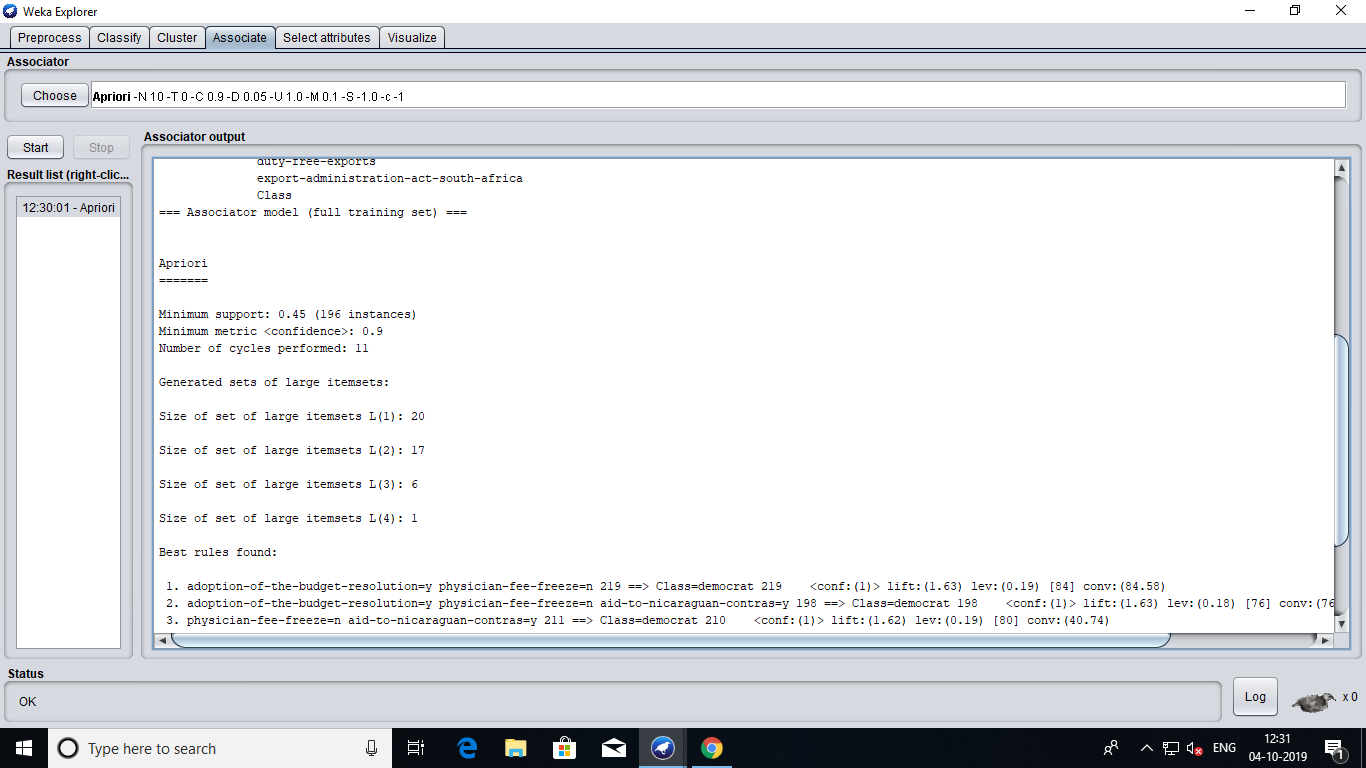
**Association Rule Mining**

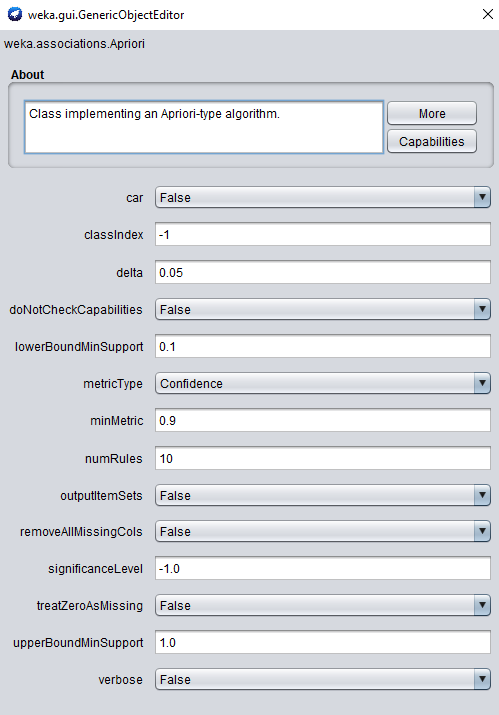
**Q1. Perform the following tasks:**

*1. Load the ‘vote.arff’ dataset*

*2. Apply the Apriori association rule*







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

Minimum Support Threshold : 0.45

Minimum Confidence Threshold : 0.45

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

1. adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 **<conf:(1)>** lift:(1.63) lev:(0.19) [84] conv:(84.58)

2. adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198 **<conf:(1)>** lift:(1.63) lev:(0.18) [76] conv:(76.47)

3. physician-fee-freeze=n aid-to-nicaraguan-contras=y 211 ==> Class=democrat 210 **<conf:(1)>** lift:(1.62) lev:(0.19) [80] conv:(40.74)

4. physician-fee-freeze=n education-spending=n 202 ==> Class=democrat 201 **<conf:(1)>** lift:(1.62) lev:(0.18) [77] conv:(39.01)

5. physician-fee-freeze=n 247 ==> Class=democrat 245 **<conf:(0.99)>** lift:(1.62) lev:(0.21) [93] conv:(31.8)

6. el-salvador-aid=n Class=democrat 200 ==> aid-to-nicaraguan-contras=y 197 **<conf:(0.98)>** lift:(1.77) lev:(0.2) [85] conv:(22.18)

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

It represents the number of instances that satisfy the antecedent.

For example, if

adoption-of-the-budget-resolution=y physician-fee-freeze=n 219

This means that 219 instances have feature adoption-of-the-budget-resolution with value "y" and physician-fee-freeze with value "n".

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

It represents the number of instances that satisfy the consequent.

For example, if

adoption-of-the-budget-resolution=y physician-fee-freeze=n 219 ==> Class=democrat 219 <conf:(1)>

This means that out of 219 instances satisfy the antecedent criteria we get 219 instances classified into class democrat. Hence the confidence is 219/219 = 1.

el-salvador-aid=n aid-to-nicaraguan-contras=y 204 ==> Class=democrat 197 <conf:(0.97)>

Confidence = 197/204 = 0.97

*7. 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?*

Numerical values used for computing confidence in the above example were 197 and 214 and it is in accordance with the data in the preprocess panel as it doesn't exceed any attribute value.

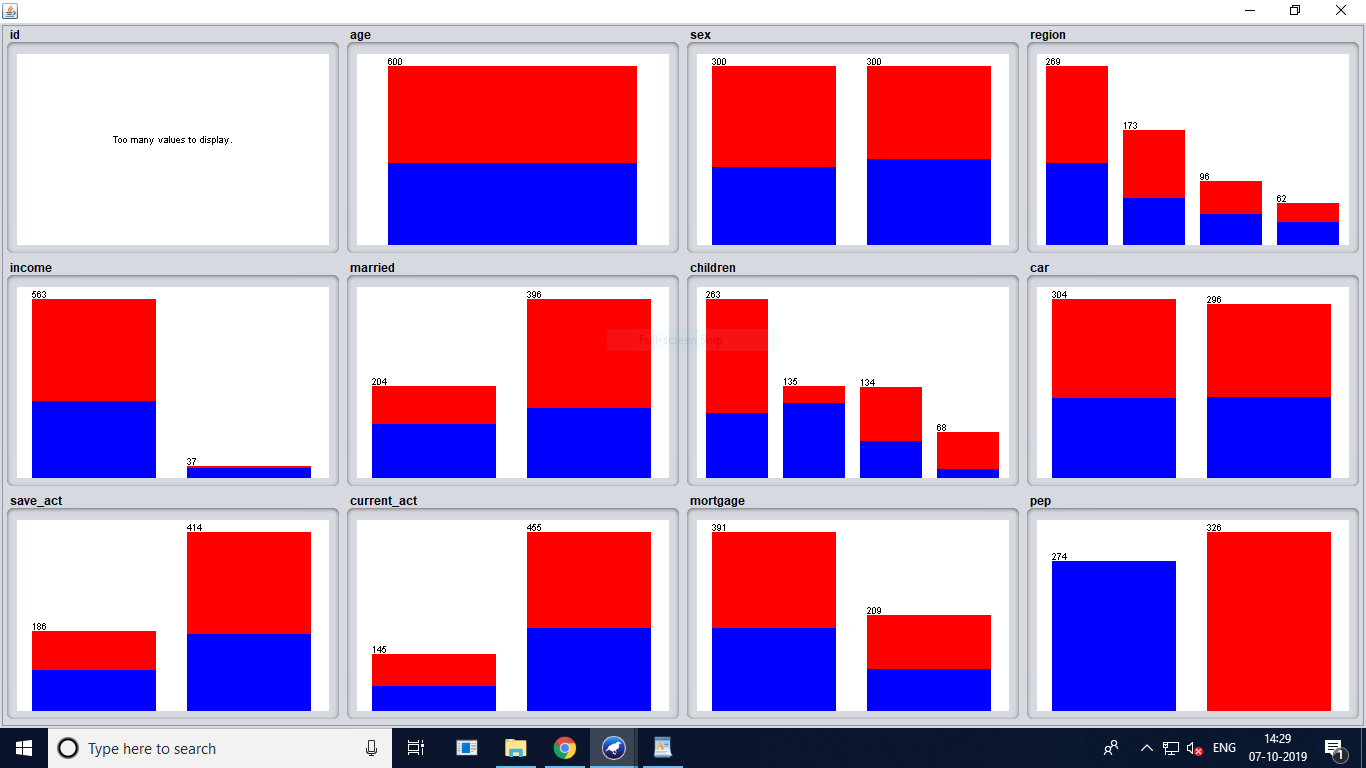
**Q2. Perform the following tasks:**

*1. Load the bank-data.csv file.*

*2. Apply the Apriori association rule algorithm. What is the result? Why?*

No result is obtained as the "Start" button is disabled. It is because Apriori association rule works only on unary, binary or nominal data and not on numerical. In this dataset we have age, income, children as numerical type attribute.

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

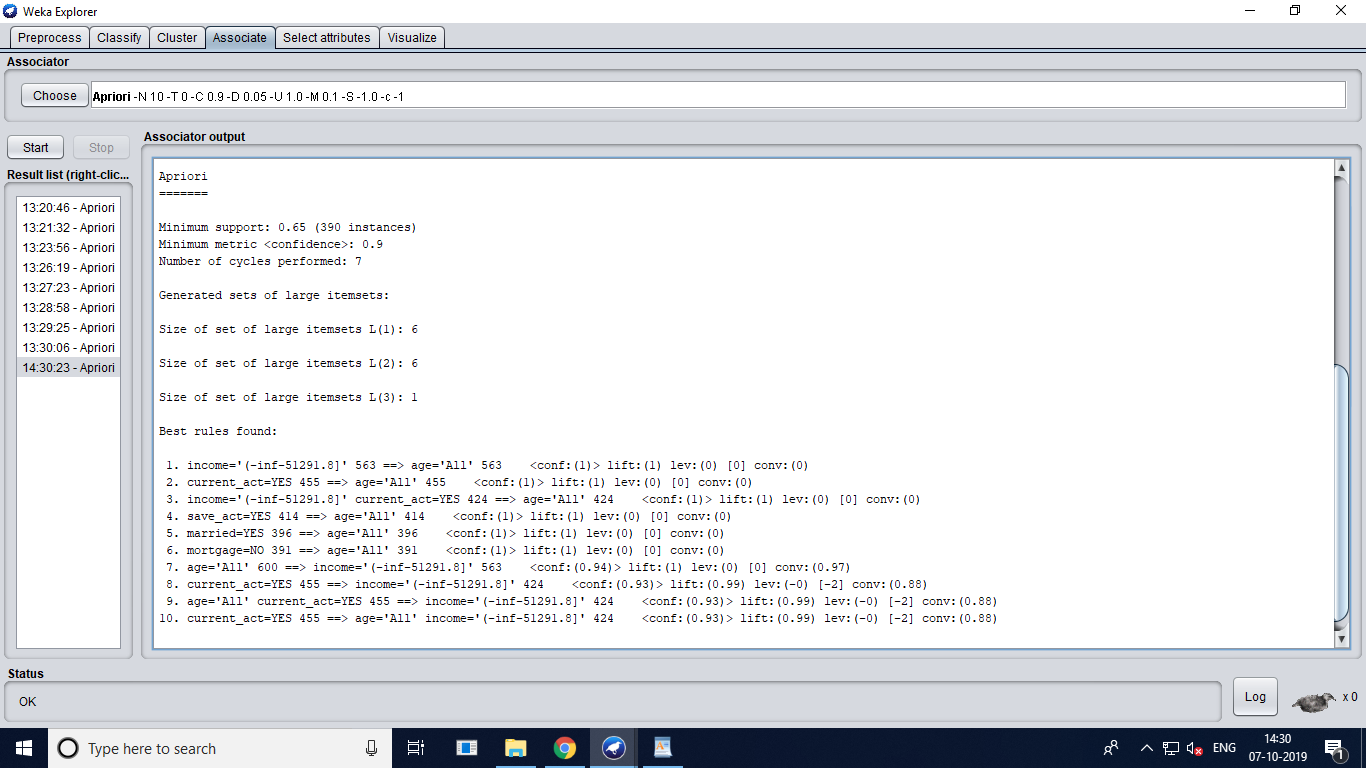
**

*4. Run the Apriori rule algorithm*

Minimum support: 0.65 (390 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 7

**

*5. List the rules that were generated.*

1. income='(-inf-51291.8]' 563 ==> age='All' 563 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

2. current\_act=YES 455 ==> age='All' 455 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

3. income='(-inf-51291.8]' current\_act=YES 424 ==> age='All' 424 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

4. save\_act=YES 414 ==> age='All' 414 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

5. married=YES 396 ==> age='All' 396 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

6. mortgage=NO 391 ==> age='All' 391 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

7. age='All' 600 ==> income='(-inf-51291.8]' 563 <conf:(0.94)> lift:(1) lev:(0) [0] conv:(0.97)

8. current\_act=YES 455 ==> income='(-inf-51291.8]' 424 <conf:(0.93)> lift:(0.99) lev:(-0) [-2] conv:(0.88)

9. age='All' current\_act=YES 455 ==> income='(-inf-51291.8]' 424 <conf:(0.93)> lift:(0.99) lev:(-0) [-2] conv:(0.88)

10. current\_act=YES 455 ==> age='All' income='(-inf-51291.8]' 424 <conf:(0.93)> lift:(0.99) lev:(-0) [-2] conv:(0.88)

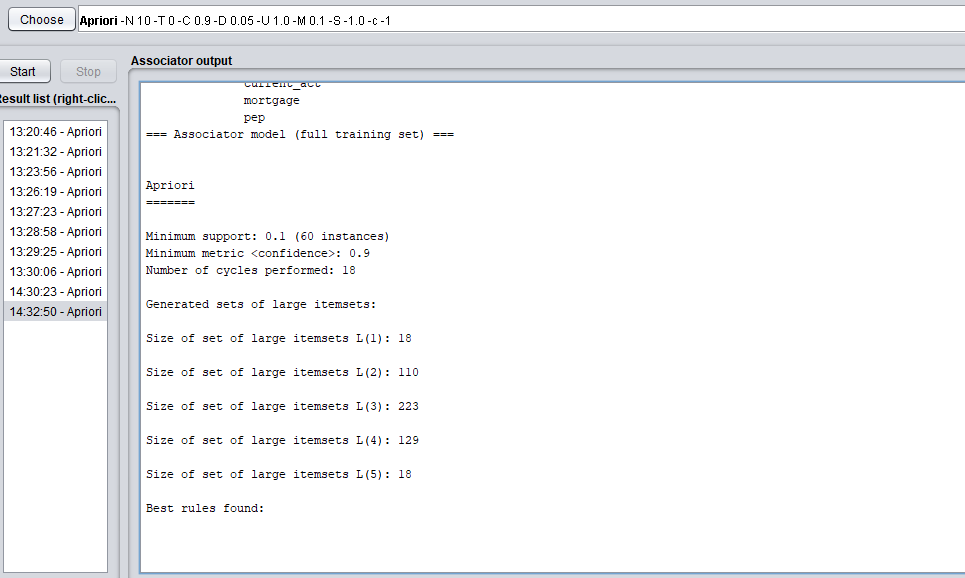
*4) Reload Bank-data.csv. Select only Nominal values.*

We did this by removing the numerical values attribute age, income, children.

*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*

**

No rules found

*7) Select Start button*

*8) now we can see the sample rules.*

By reducing the lowerBoundMinSupport to 0.05 we obtained three rules.

Apriori

=======

Minimum support: 0.05 (30 instances)

Minimum metric <confidence>: 0.9

Number of cycles performed: 19

Generated sets of large itemsets:

Size of set of large itemsets L(1): 18

Size of set of large itemsets L(2): 131

Size of set of large itemsets L(3): 423

Size of set of large itemsets L(4): 512

Size of set of large itemsets L(5): 219

Size of set of large itemsets L(6): 31

Best rules found:

1. region=INNER\_CITY car=YES save\_act=YES mortgage=NO pep=NO 33 ==> married=YES 30 <conf:(0.91)> lift:(1.38) lev:(0.01) [8] conv:(2.81)

2. region=INNER\_CITY save\_act=YES current\_act=YES mortgage=NO pep=NO 50 ==> married=YES 45 <conf:(0.9)> lift:(1.36) lev:(0.02) [11] conv:(2.83)

3. region=SUBURBAN mortgage=NO 40 ==> current\_act=YES 36 <conf:(0.9)> lift:(1.19) lev:(0.01) [5] conv:(1.93)

**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*

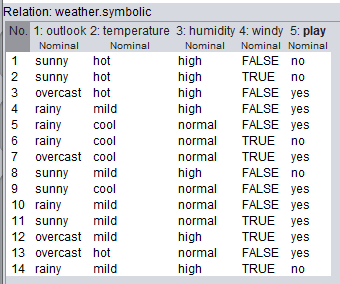
outlook -> sunny, overcast, rainy

temperature -> hot, mild, cold

humidity -> high, normal

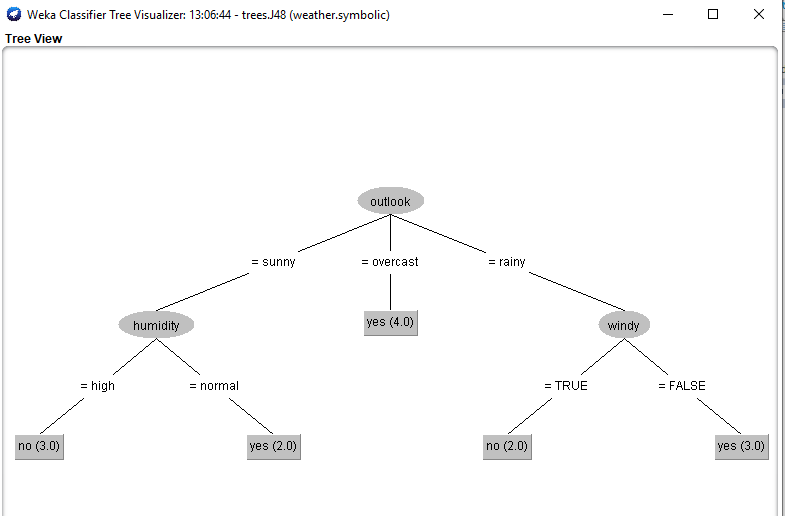
windy -> true , false

*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*

Information Gain Index Attribute

0.2467 1 outlook

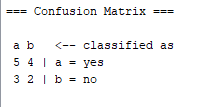
0.1518 3 humidity

0.0481 4 windy

0.0292 2 temperature

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

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



*Describe the Kappa statistic?*

The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. In addition, it takes into account random chance (agreement with a random classifier), which generally means it is less misleading than simply using accuracy as a metric (an Observed Accuracy of 80% is a lot less impressive with an Expected Accuracy of 75% versus an Expected Accuracy of 50%). Computation of Observed Accuracy and Expected Accuracy is integral to comprehension of the kappa statistic, and is most easily illustrated through use of a confusion matrix.

Observed kappa statistic value for the above classification is -0.0426

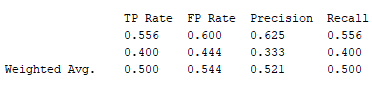
*Describe the following quantities:*

TP Rate -> True positives are data point classified as positive by the model that actually are positive (meaning they are correct)

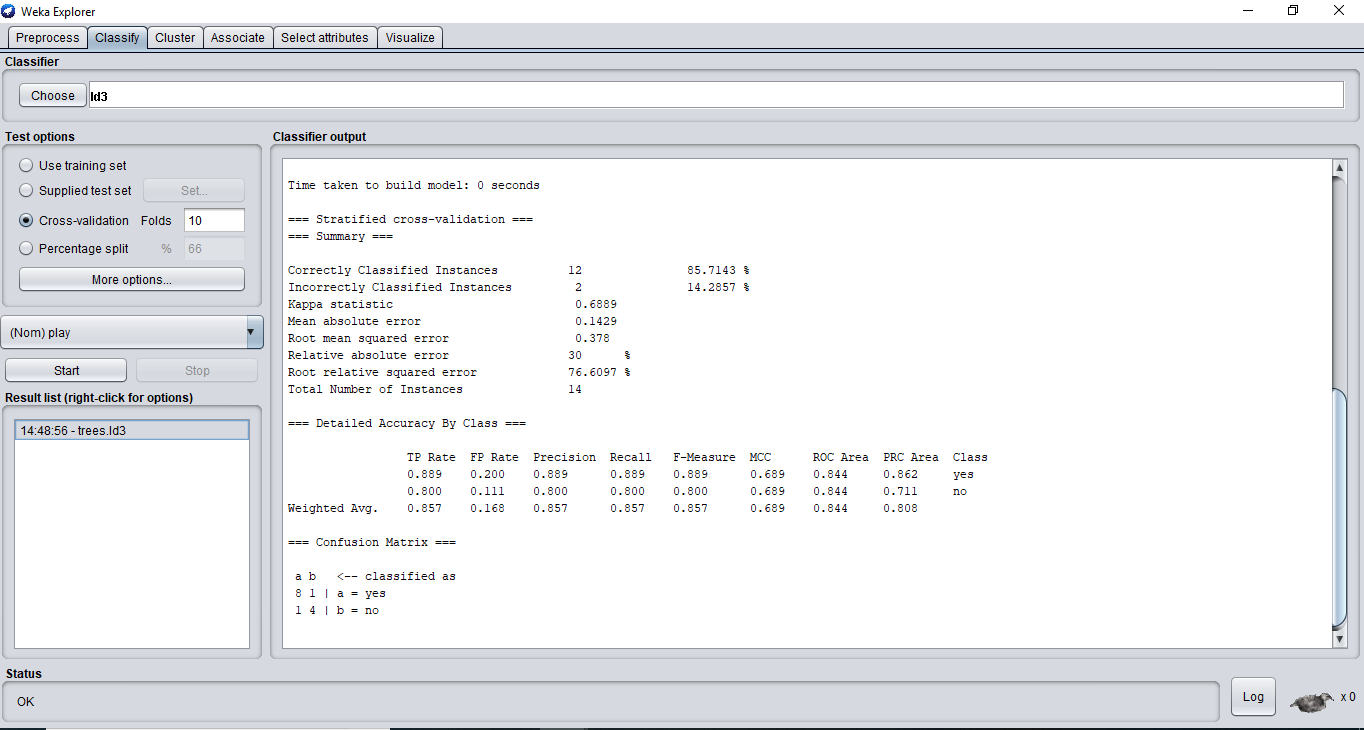
FP Rate -> False positives are cases the model incorrectly labels as positive that are actually negative

Precision -> The ability of a classification model to identify only the relevant data points.Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Recall -> The ability of a model to find all the relevant cases within a dataset. Recall is the number of true positives divided by the number of true positives plus the number of false negatives.



**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?**