# PRACTICAL NO. 9

**Aim:** To perform classification and Clustering using the data mining toolkit

**Classification:**
**Solve the following questions:**

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

1. List the attributes of the given relation along with the type details

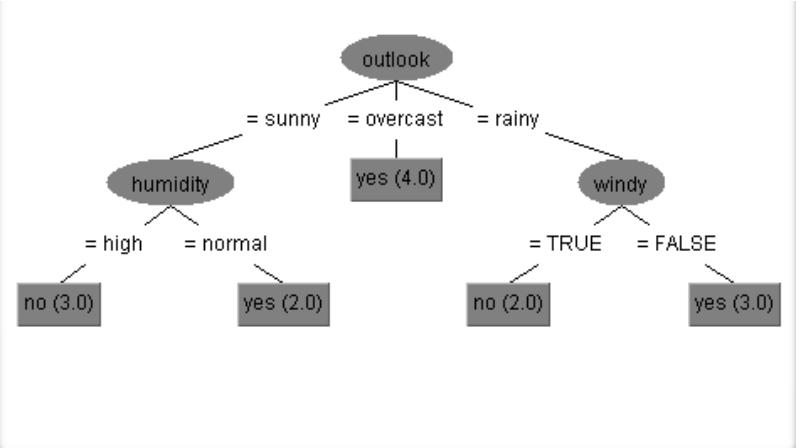
| Attributes | Type |
|---|---|
| Outlook | Nominal |
| Temperature | Nominal |
| Humidity | Nominal |
| Windy | Nominal |
| Play | Nominal |

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

| No. | 1: outlook | 2: temperature | 3: humidity | 4: windy | 5: **play** |
|---|---|---|---|---|---|
| | Nominal | Nominal | Nominal | Nominal | Nominal |
| 1 | sunny | hot | high | FALSE | no |
| 2 | sunny | hot | high | TRUE | no |
| 3 | overcast | hot | high | FALSE | yes |
| 4 | rainy | mild | high | FALSE | yes |
| 5 | rainy | cool | normal | FALSE | yes |
| 6 | rainy | cool | normal | TRUE | no |
| 7 | overcast | cool | normal | TRUE | yes |
| 8 | sunny | mild | high | FALSE | no |
| 9 | sunny | cool | normal | FALSE | yes |
| 10 | rainy | mild | normal | FALSE | yes |
| 11 | sunny | mild | normal | TRUE | yes |
| 12 | overcast | mild | high | TRUE | yes |
| 13 | overcast | hot | normal | FALSE | yes |
| 14 | rainy | mild | high | TRUE | no |

**3. Study the classifier output and answer the following questions**

**1.Draw the decision tree generated by the classifier**




3.Compute the entropy values for each of the attributes
Ans: **For outlook –**
H(class)= -9/14 log2 9/14 -5/14 log2 5/14 = 0.941
H(Yes)= -2/9 log2 2/9 – 3/9 log2 3/9 – 4/9 log2 4/9 = 1.531
H(No)= -3/5 log2 3/5 – 2/5 log2 2/5 -0 = 0.9708
IG(Outlook)= 0.941 – (9/14*1.531 -5/14*0.9708) = 0.247

**For temperature-**
H(class)= -9/14 log2 9/14 -5/14 log2 5/14 = 0.941
H(Yes)= -2/9 log2 2/9 – 3/9 log2 3/9 – 4/9 log2 4/9 = 1.531
H(No)= -2/5 log2 2/5 – 1/5 log2 1/5 --2/5 log2 2/5 = 1.5217
IG(temperature)= 0.941 – (9/14*1.531 -5/14*0.9708) = 0.029

**For humidity-**
H(class)= -9/14 log2 9/14 -5/14 log2 5/14 = 0.941
H(Yes)= -3/9 log2 3/9 – 6/9 log2 6/9 = 0.9182
H(No)= -4/5 log2 4/5 – 1/5 log2 1/5 = 0.7218
IG(humidity)= 0.941 – (9/14*0.9182-5/14*0.7218) = 0.152

**For windy-**
H(class)= -9/14 log2 9/14 -5/14 log2 5/14 = 0.941
H(Yes)= -3/9 log2 3/9 – 6/9 log2 6/9 = 0.9182
H(No)= -3/5 log2 3/5 – 2/5 log2 2/5 -0 = 0.9708
IG(windy)= 0.941 – (9/14*0.9182 -5/14*0.9708) = 0.048

**3.What is the relationship between the attribute entropy values and the nodes of the decision 9 tree?**
Ans: The nodes with highest entropy value is chosen as the root if the decision tree.

**4. Draw the confusion matrix? What information does the confusion matrix provide?**
Ans: Confusion matrix
a b <-- classified as
9 0 | a = yes

0 5 | b = no

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.
The confusion matrix shows the ways in which your classification model is confused when it makes predictions.It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

**5. Describe the Kappa statistic?**

Ans: Cohen's kappa statistic measures interrater reliability (sometimes called interobserver agreement). Interrater reliability, or precision, happens when your data raters (or collectors) give the same score to the same data item.This statistic should only be calculated when:
1.Two raters each rate one trial on each sample, *or*.
2.One rater rates two trials on each sample.

The Kappa statistic varies from 0 to 1, where.

- 0 = agreement equivalent to chance.
- 0.1 – 0.20 = slight agreement.
- 0.21 – 0.40 = fair agreement.
- 0.41 – 0.60 = moderate agreement.
- 0.61 – 0.80 = substantial agreement.
- 0.81 – 0.99 = near perfect agreement
- 1 = perfect agreement.

**6. Describe the following quantities:**

1. TP Rate

Ans: Recall is the TP rate ( also referred to as sensitivity) answers the question what fraction of those that are actually positive were predicted positive?

2. FP Rate

Ans:

3. Precision

4. Recall

2. Load the 'weather.nominal.arff' dataset in Weka and run the Id3 classification algorithm. What problem do you have and what is the solution?

3. Load the 'weather.arff' dataset in Weka and run the OneR rule generation algorithm. Write the rules that were generated.

4. Load the 'weather.arff' dataset in Weka and run the PRISM rule generation algorithm. Write down the rules that are generated.

**Clustering:**

**Answer the following Questions:**

***Part 1***

**1. Perform the following tasks:**
**1. Load the bank data set in Weka. Explain the characteristics of this dataset.**

(**Dataset:** The bank dataset can be used to answer various questions such as which age group are more likely to buy the personal equity plan)

2. Write down the following details regarding the attributes:
1. names
2. types
3. values.

**3. Run the Simple K-Means clustering algorithm on the dataset. Consider no. of clusters as 6.**What do you mean by seed value?
Ans: Seed value is used to generate random data. The seed value is randomly generated

**1. What are the number of instances and percentage figures in each cluster?**
Ans: The following are number of instances and percentage figures in each cluster
Cluster 0: ID12614,25,FEMALE,RURAL,14505.3,NO,3,NO,YES,YES,NO,NO
Cluster 1: ID12131,61,FEMALE,RURAL,22942.9,YES,2,NO,YES,YES,NO,NO
Cluster 2: ID12190,54,FEMALE,INNER_CITY,31095.6,YES,2,NO,NO,YES,NO,YES
Cluster 3: ID12485,36,FEMALE,TOWN,26920.8,YES,0,NO,NO,YES,NO,NO
Cluster 4: ID12203,42,MALE,INNER_CITY,15499.9,YES,0,YES,NO,YES,YES,YES
Cluster 5: ID12597,50,MALE,TOWN,40972.9,NO,2,YES,YES,YES,YES,YES

**2. What is the number of iterations that were required?**
Ans: 18 iterations were required

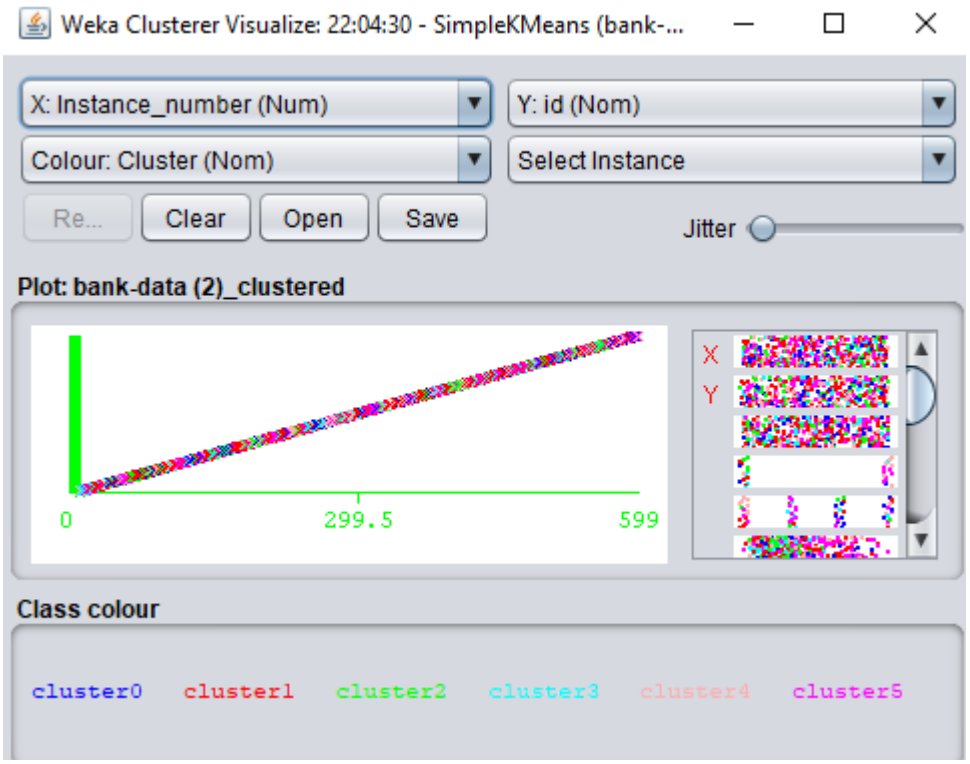
**3. What is the sum of squared errors? What does it represent?**
Ans: The sum of squared errors is 1955.4146. It represents the variation within a cluster. If all cases within a cluster are identical then sum of squared errors would be equal to 0.

**4. What do you mean by centroid? What does it represent? Tabulate the characteristics of the centroid of each cluster.**
Ans: A centroid is a data point (imaginary or real) at the center of a cluster. These centroids are used to train a knn classifier. The resulting classifier is used to classify the data thereby produce an initial randomized set of clusters.

| Attribute | Full Data (600.0) | 0 (74.0) | 1 (164.0) | 2 (71.0) | 3 (58.0) | 4 (99.0) | 5 (134.0) |
|---|---|---|---|---|---|---|---|
| id | ID12101 | ID12107 | D12103 | ID12101 | ID12104 | ID12102 | ID12108 |
| Sex | Female | Female | Female | Female | Female | Male | Male |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| age | 42.395 | 42.9324 | 43.7744 | 39.0282 | 37.3103 | 38.404 | 47.3433 |
| region | Inner_City | Rural | Inner_City | Inner_City | Town | Inner_City | Town |
| Income | 27524.0312 | 28838.76 | 28586.406 | 20463.1273 | 20600.85 | 25720.037 | 33568.392 |
| married | YES | NO | YES | YES | YES | YES | NO |
| Children | 1.0117 | 1.973 | 0.628 | 0.6901 | 1.6207 | 0.899 | 0.9403 |
| Car | NO | NO | NO | NO | NO | YES | YES |
| Save_act | YES | YES | YES | NO | NO | NO | YES |
| Current_act | YES | YES | YES | YES | YES | YES | YES |
| Mortgage | NO | NO | NO | NO | NO | YES | NO |
| Pep | NO | NO | NO | YES | NO | YES | YES |

5. Visualize the results of this clustering (let the X-axis represent the cluster name, and the Y-axis represent the instance number)



1. **Is there a significant variation in age between clusters?**

Ans: Yes, the age attribute is distributes uniformly across all clusters.

2. **Which clusters are predominated by males and which clusters are predominated by females?**

Ans: Clusters 2 and 4 are predominated by females and clusters 1 and 5.

3. **What can be said about the variation of income between clusters?**

Ans: Mostly clusters are distributed in the range of income 5014.21 - 34072.155.

4. **Which clusters are dominated by married people and which clusters are dominated by unmarried people?**

Ans: Married people dominate clusters 3,cluster 1 and cluster 5 whereas unmarried people dominate clusters 0 and cluster 4.

5. **How do the clusters differ with respect to the number of children?**

Ans: Mostly children in the age bracket if 0-1 fall in cluster 2 and cluster 1. Children above the age of 1.5 mostly dominate clusters 0 and cluster 4. Very few children above the age of 3 fall in cluster 3 and cluster 3.

6. **Which cluster has the highest number of people with cars?**

Ans: Cluster 4 and cluster 5 has the highest number of people with cars.

**7. Which clusters are predominated by people with savings accounts?**

Ans:Clusters 1 and 5 are predominated by people with savings accounts.

**8. What can be said about the variation of current accounts between clusters?**

Ans:Cluster 1 is dominated by people with current accounts whereas cluster 2 is dominated by people without any current account.

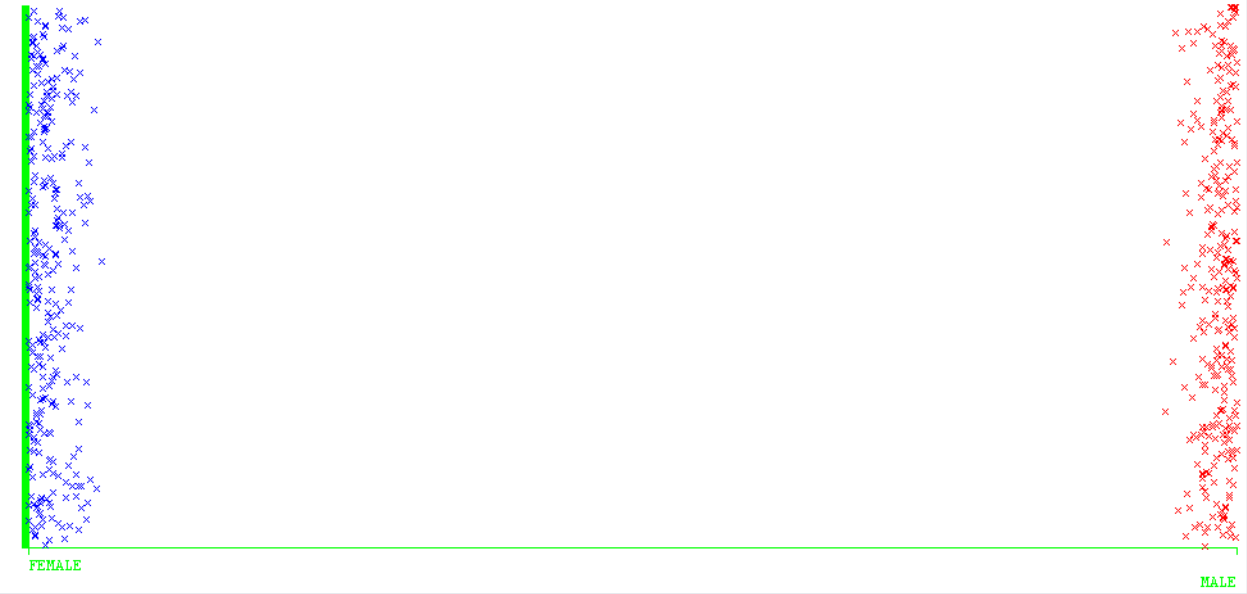
**9. What can be said about the variation of mortgage holdings between clusters?**

Ans: Maximum people with mortgage holdings belong to cluster 2 and cluster 1 whereas people without any mortgage holdings belong to cluster 4 and cluster 1.

**10.Which clusters comprise mostly of people who buy the PEP product and which ones are comprised of people who do not buy the PEP product?**

Ans: Clusters 2 and 5 comprise mostly of people who buy the PEP product and clusters 1,3 and 4 are dominated by people who do not buy PEP products.

**7.Select "sex" attribute as the color dimension and visualize the result, the distribution of males and females in each cluster. (Note: In this case, by changing the color dimension to other attributes, we can see their distribution within each of the clusters.)**




**7. Finally, save the resulting data set which included each instance along with its assigned cluster. To do so, we click the "Save" button in the visualization window and save the result as the file "bank-kmeans.arff".**

4. Run the SimpleKMeans algorithm for values of K (no. of clusters) ranging from 2 to 12. Tabulate the sum of squared errors for each run. What do you observe about the trend of the sum of squared errors?

| K | Sum of squared errors |
|---|---|
| 2 | 2335.27 |
| 3 | 2165.47 |
| 4 | 2047.72 |
| 5 | 2047.39 |
| 6 | 1955.41 |
| 7 | 1920.95 |
| 8 | 1840.22 |
| 9 | 1778.17 |
| 10 | 1752.55 |
| 11 | 1725.64 |
| 12 | 1667.11 |

Observation- As the number of clusters increase, the sum of squared errors decreases. The number of clusters is inversely proportional to the sum of squared errors.

**5. Do you see any differences in your ability to evaluate the characteristics of clusters generated for K=6 versus K=12? Why does this difference arise?**

Ans: The sum of squared errors for K=6 is 1955.41 whereas K=12 the sum of squared errors is 1667.11. As we can observe, the sum of squared errors has decreased drastically. This difference

occurs because by increasing the number of clusters we are trying to decrease the error. A good cluster is the one which has minimum value of sum of squared errors

***Part 2***

**Use DBSCAN clustering algorithm on the same dataset and observe the results. What can you conclude about DBSCAN and simple k means algorithm?**
Ans: DBSCAN clustering algorithm is not available in Weka tool.

***Part3***

**Run both the algorithms on another dataset and conclude the results.**

Result of SimpleKMeans Clustering Algorithm on iris.arff dataset

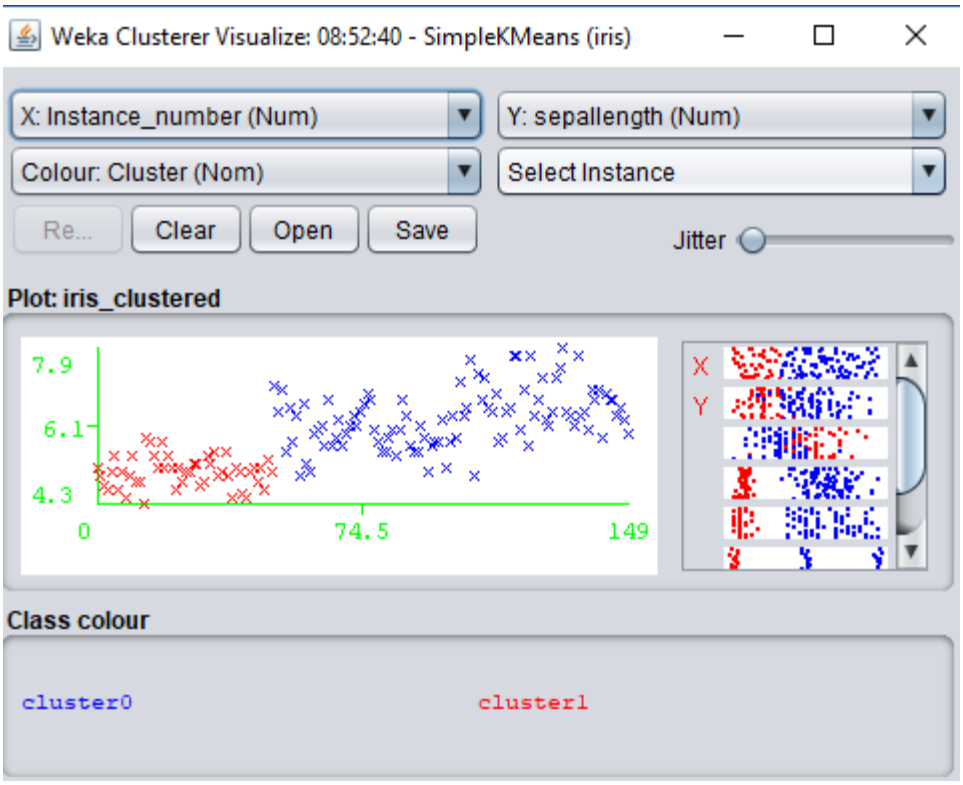
```
Number of iterations: 7
Within cluster sum of squared errors: 62.1436882815797

Initial starting points (random):

Cluster 0: 6.1,2.9,4.7,1.4,Iris-versicolor
Cluster 1: 6.2,2.9,4.3,1.3,Iris-versicolor

Missing values globally replaced with mean/mode

Final cluster centroids:
                                        Cluster#
Attribute               Full Data             0               1
                          (150.0)         (100.0)          (50.0)
=================================================================
sepallength                5.8433           6.262           5.006
sepalwidth                  3.054           2.872           3.418
petallength                3.7587           4.906           1.464
petalwidth                 1.1987           1.676           0.244
class                 Iris-setosa Iris-versicolor     Iris-setosa
```


Weka Clusterer Visualize: 08:52:40 - SimpleKMeans (iris)
X: Instance_number (Num)
Y: sepallength (Num)
Colour: Cluster (Nom)
Select Instance
Re...
Clear
Open
Save
Jitter
Plot: iris_clustered
7.9
6.1
4.3
0
74.5
149
X
Y
Class colour
cluster0
cluster1