KDD

Lab 1

Utkarsh Bhangale

20200802124

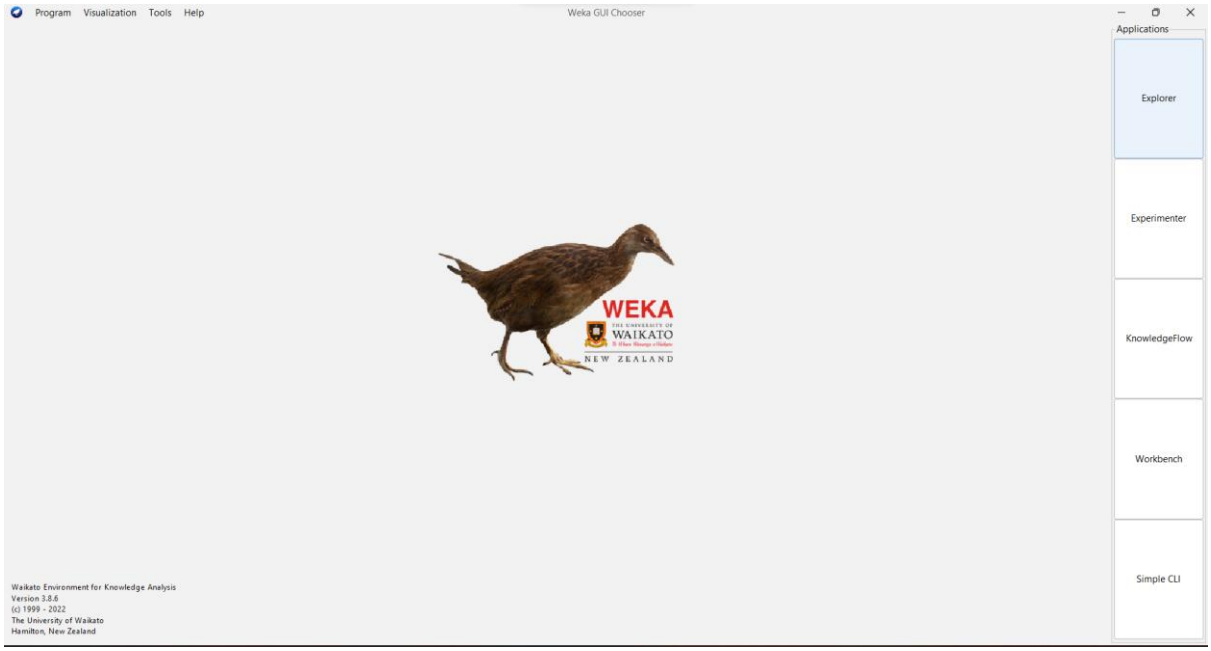
Aim: To explore the open-source WEKA GUI tool and try visualizing the following pre-processing

functions:

1. Classify
2. Cluster
3. Associate

Dataset: https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/weather.nominal.arff

Weka_Application



Observations:

The WEKA GUI Chooser application will start and you will see the following screen-

1. The GUI Chooser application allows you to run five different types of applications as listed here – this tab is also known as the machine learning tabs

a. Explorer

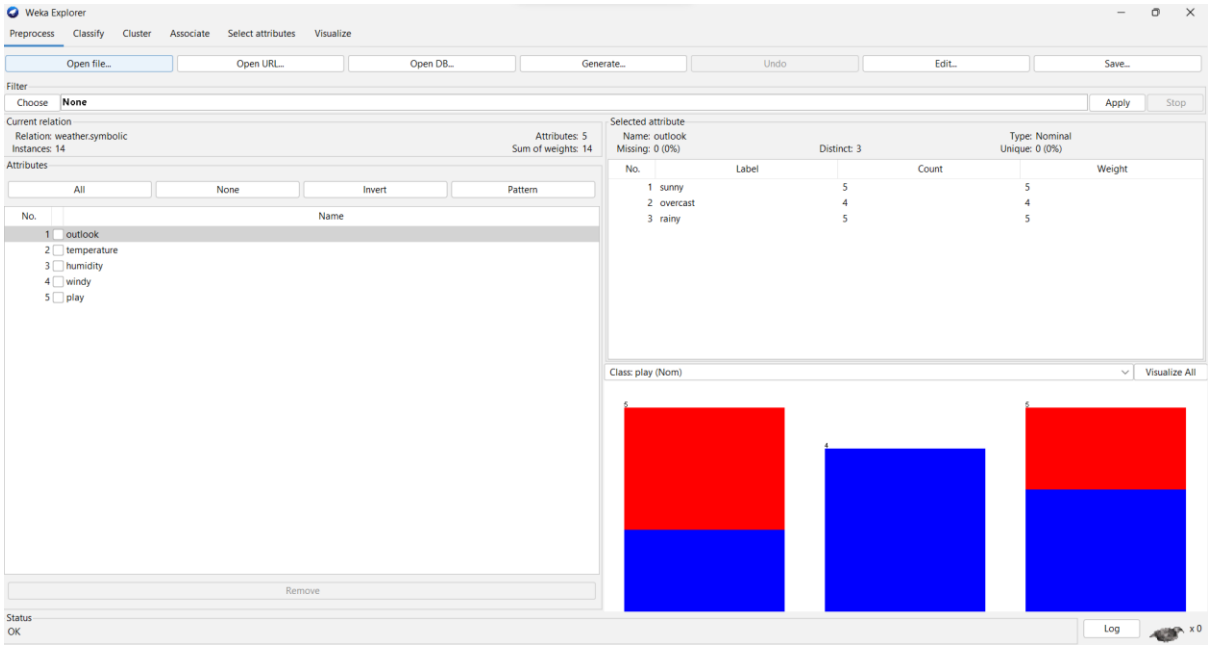
b. Experimenter

c. Knowledge Flow

d. Workbench

e. Simple CLI

ii. We will be using Explorer in this lab.



Observations –

i. After typing the URL of the dataset, you will see the following screen.

ii. Understanding the data:

-There are 14 instances or rows in the table

-The table contains 5 Attributes

-The weather database contains five fields - • Outlook

• Temperature

• Humidity

• Windy

• Play

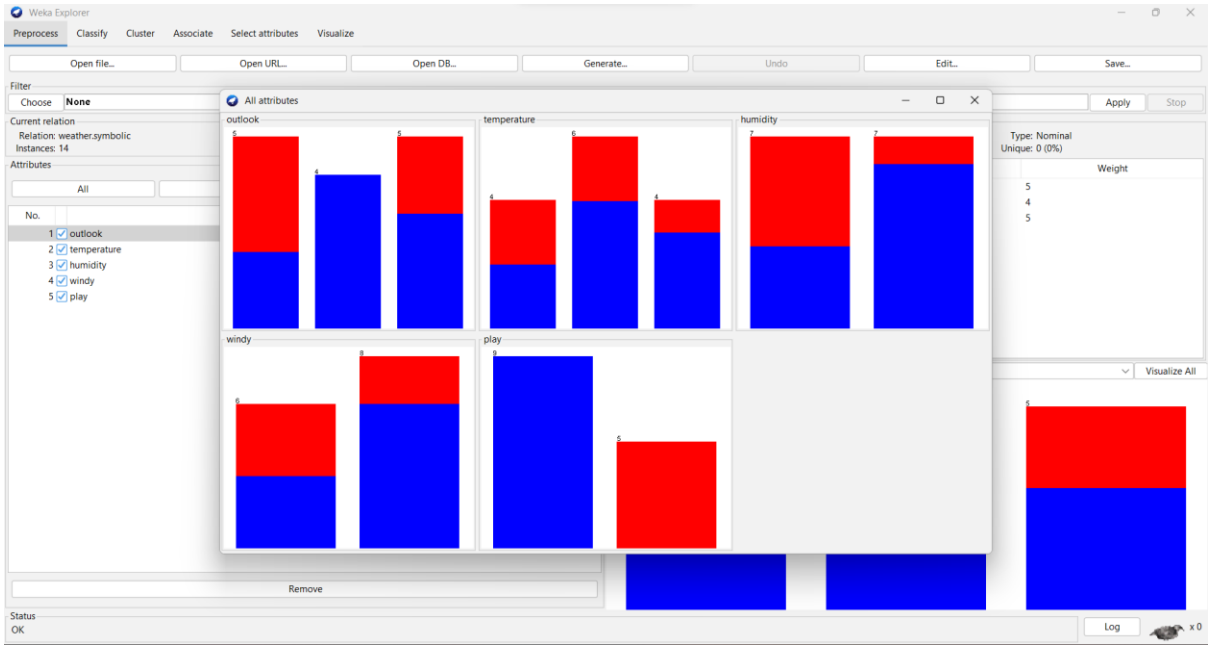
iii. In the Selected attributes sub-window, there is some information about the data.

 The name and the type of the attributes are displayed.

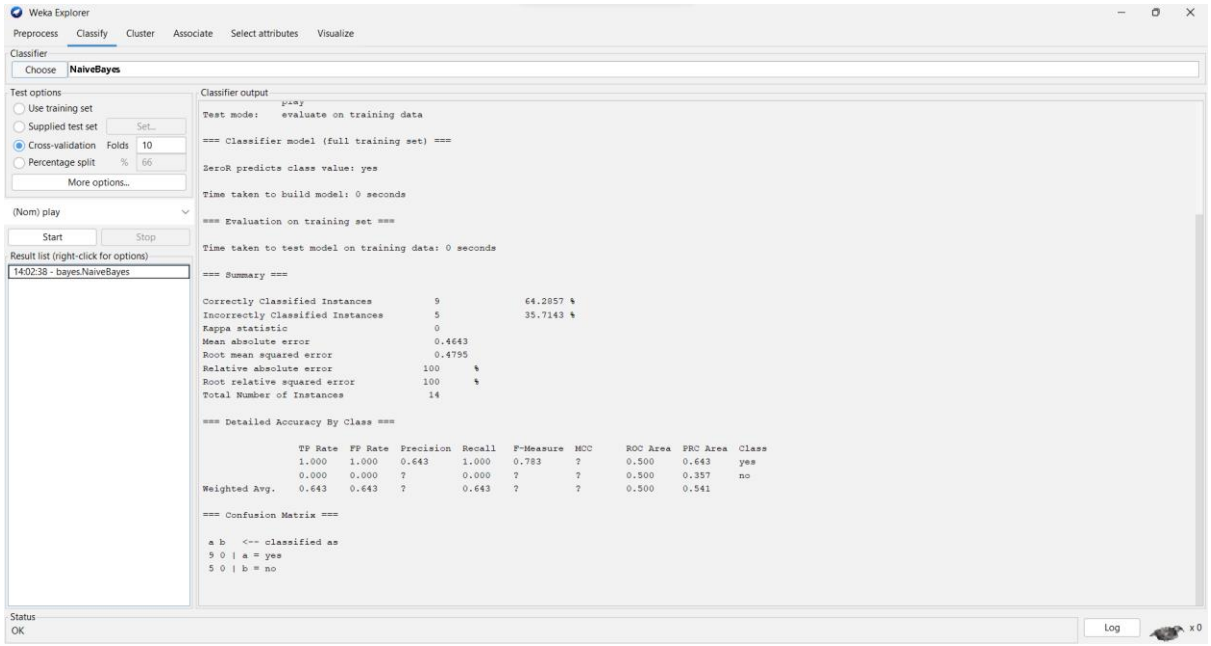
 The type for the Temperature is nominal and for others is categorical.

 There are no Missing Values in the dataset.

Visualize all



Weka - classify



Observations –

i. After using Pre-Processing the data, I used a classifiers algorithm for the dataset.

ii. Clicked on the Classify option which is next to the Pre-process option on the machine learning tab.

iii. In the left most of the window there is a “Test options” includes –

- Use Training set
- Supplied test set

 Cross-Validation Folds (10), (you can set manually how many folds do you want to perform)

 Percentage split (66%), (you can set manually how much percentage split you want to perform)

iv. Then selected J48 method from trees section.

v. Then I started the algorithm and it shows some information in the Classifier output window –

 Number of trees and No. of Leaves are 5 and 8 respectively

 It says that the correctively classified instances are 7 and the incorrectly classified instances are also 7

 Mean Absolute Error is 0.4167

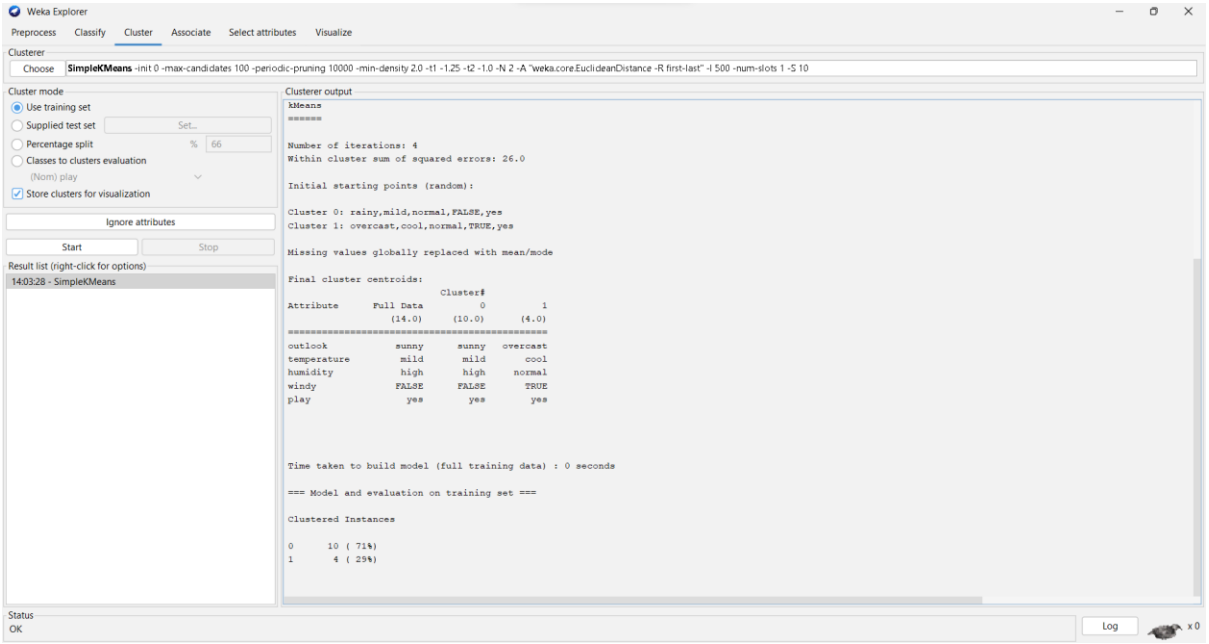
 Root Mean squared Error is 0.5984

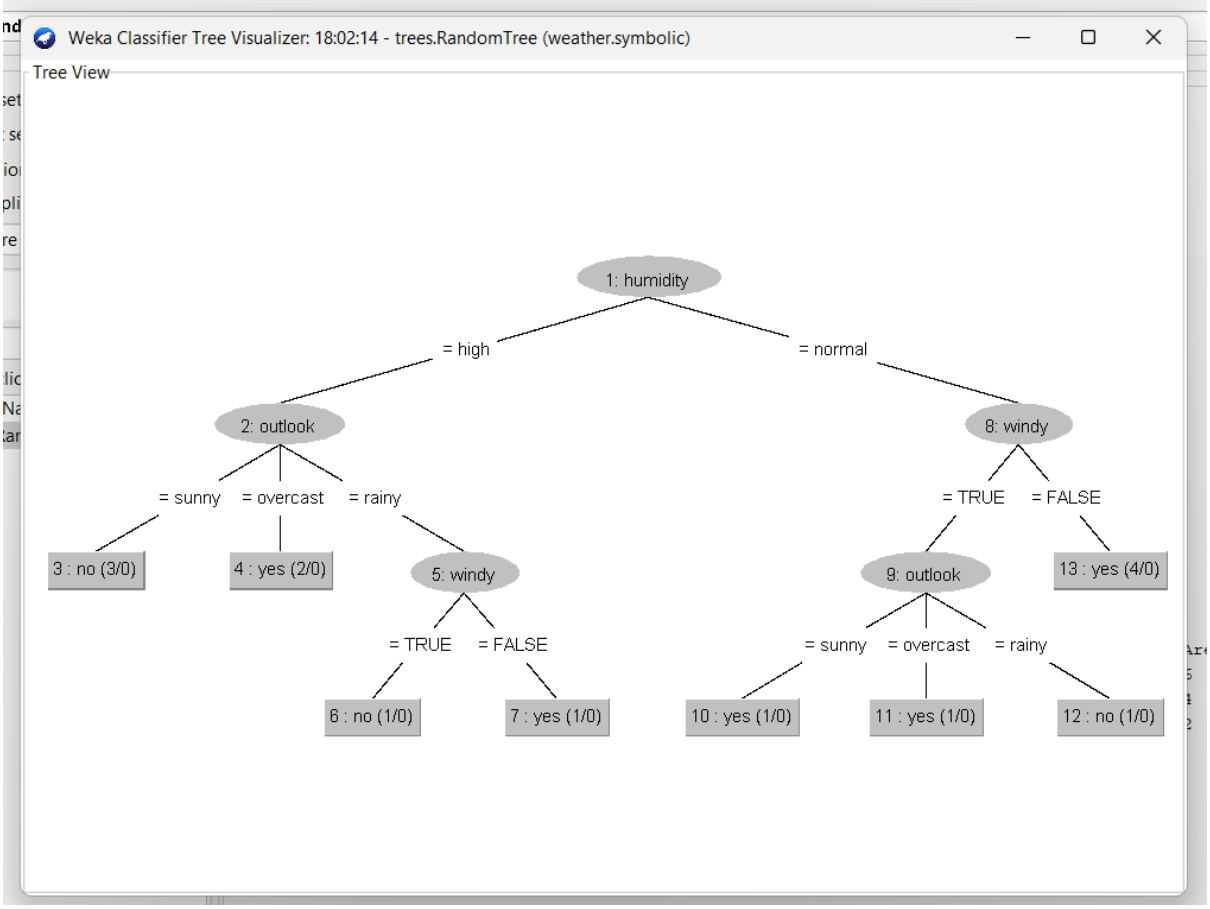
 Relative absolute error is 87.5% and root relative squared error is 121.2987%

 It also showing the confusion matrix and detailed accuracy of the class

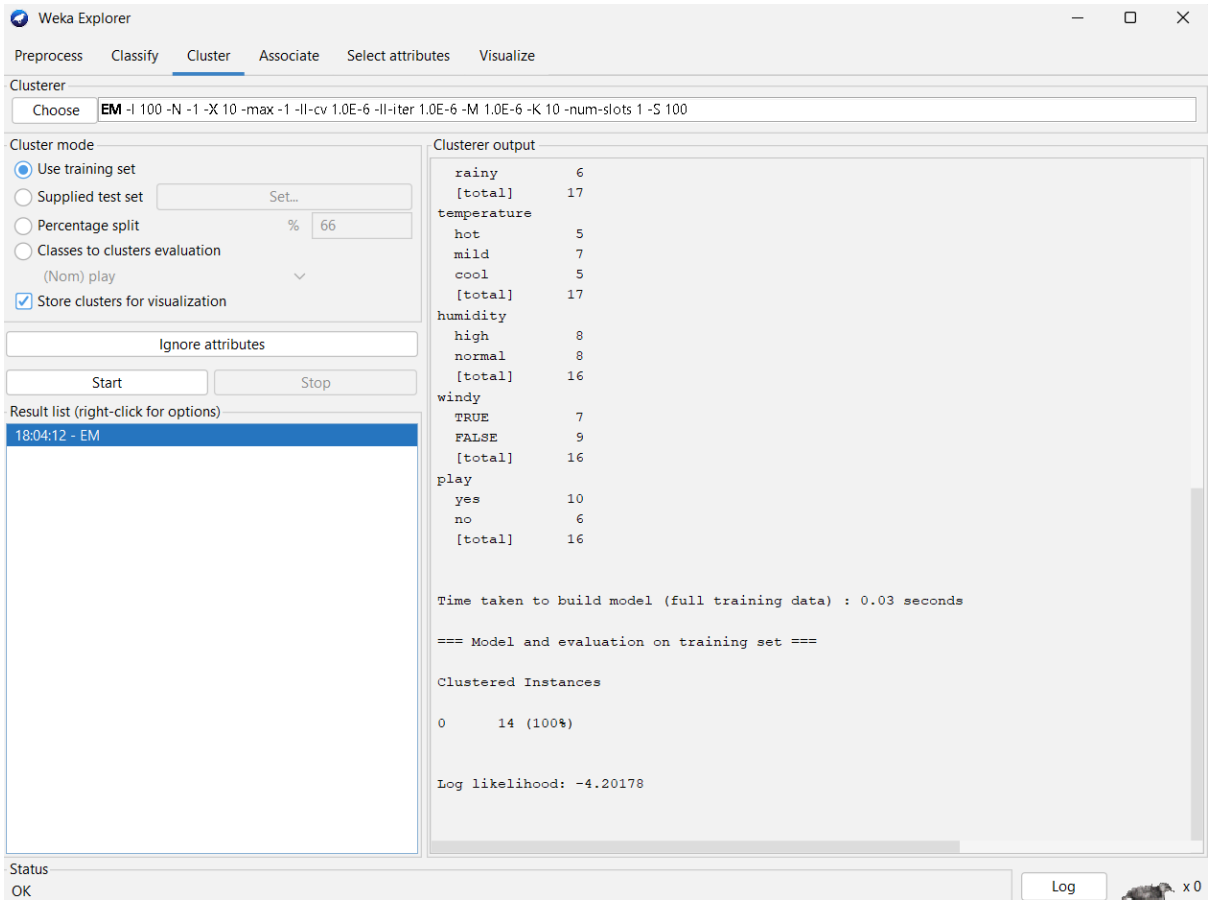
 True Positive and True Negative in the confusion matrix are 5 and 4 respectively

 False Positive and False Negative in the confusion matrix are 2 and 3 respectively





WEKA – Clustering



Observations –

i. Clicked on the cluster tab to apply the clustering algorithms to our loaded data.

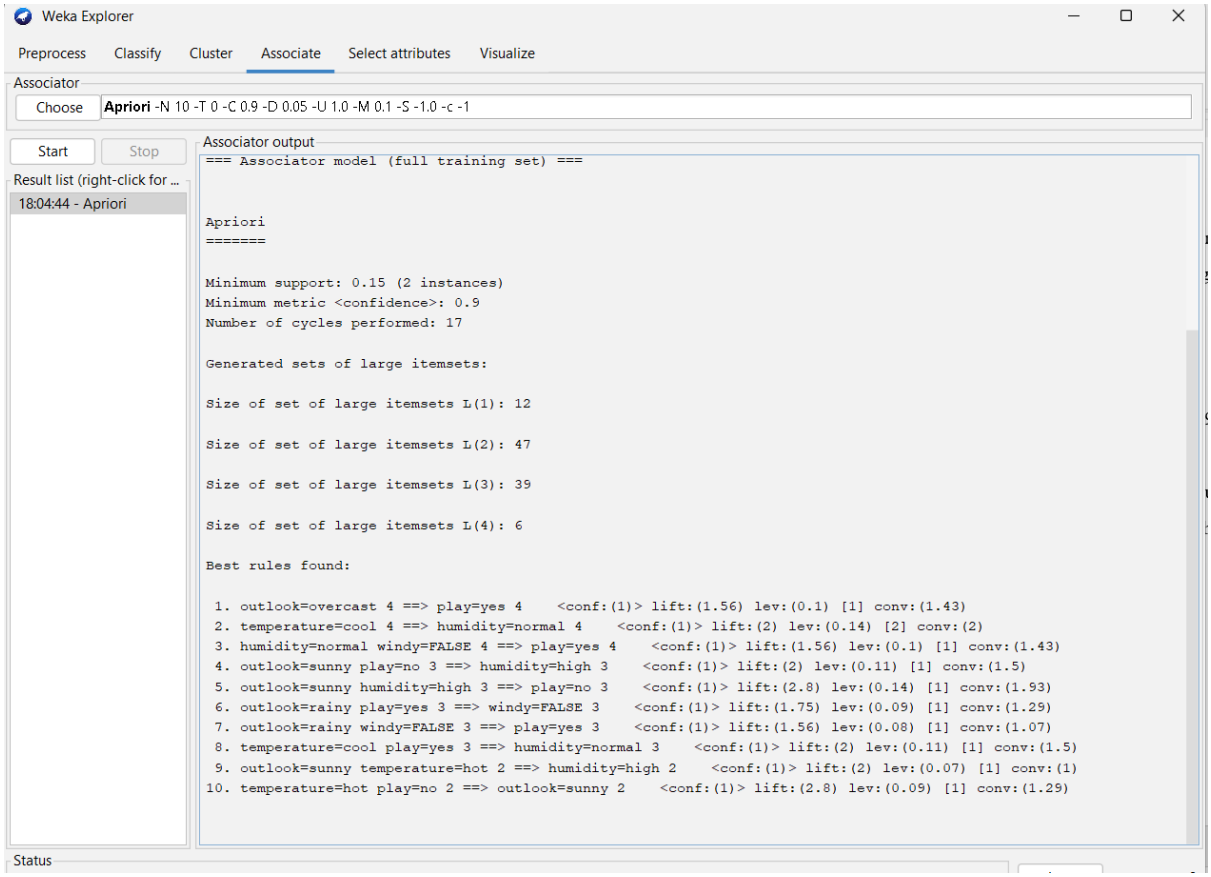
ii. I had kept default method which is EM for clustering algorithm.

iii. And then clicked the start button to process the data.

iv. Examining output –

- Log Likelihood is -3.54934
- Correctly clustered and incorrectly clustered are 9 and 5
- There are 9 Yes and 5 No to cluster
- Clicked on Visualize Clusterer to visualize the clustering of data
- This window shows a clustering relationship with outlook to instance number

WEKA – Associate



Observations –

i. Clicked on the Associate tab and clicked on the choose button to choose method to solve associate.

ii. I selected Apriori method to perform Association rule.

iii. At the bottom you can find the detected best rules of associations. This will help the weather scientist to forecast weather.

iv. At bottom you can also find lev, confidence, conv and lift factors for each of the attributes.

v. There are total of 10 best rules founded for the weather dataset.

Weka - Visualize

Observations –

 Clicked on the Visualize window to visualize the dataset after performing some algorithms on the dataset that was loaded with the URL.

 Each Attribute is correlated with the other and visualized here.

Overall Observations –

1. Loaded the data using the URL in the Weka Application.
2. Pre-processed the data.
3. Later, Classified the data and visualized it using Classify J48 method.
4. Then used the Clustering algorithm to the dataset and visualized it.
5. Used the Apriori method In Associate and found the best rules.
6. At last, With visualization found a correlation between all the attributes.