Experiment 1

Aim

To perform data pre-processing task using Weka data mining tool.

Theory

WEKA is an open-source software provides tools for data pre-processing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems

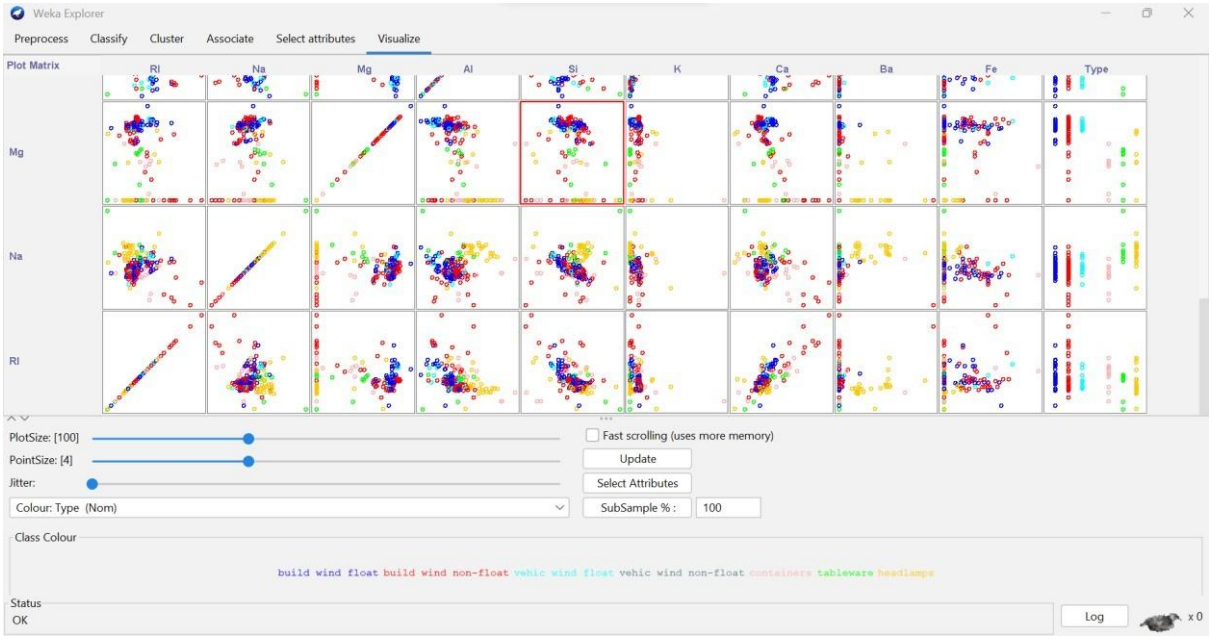
Tasks performed through Weka:

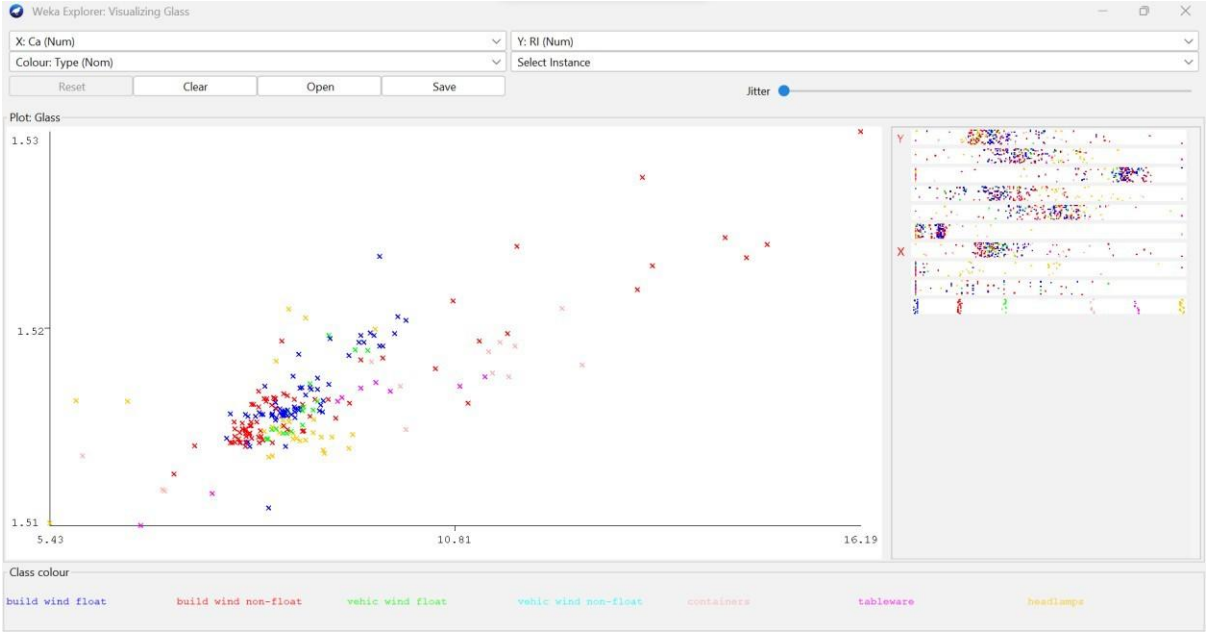
1. Pre-processing: Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format. The steps involved in this include Data Cleaning, Data Transformation and Data Reduction.
2. Classification: Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.
3. Clustering: The process of making a group of abstract objects into classes of similar objects is known as clustering. In the process of cluster analysis, the first step is to partition the set of data into groups with the help of data similarity, and then groups are assigned to their respective labels.
4. Association Rule: Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently an itemset occurs in a transaction. A typical example is a Market Based Analysis.
5. Select Attributes: The attribute selection task essentially consists in selecting a subset of originally available attributes to be subsequently used for model creation. A search strategy is needed for any attribute selection technique that is based on evaluating attribute subsets rather than single attributes.
6. Visualization: Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

TASKS

Pre-processing activities to be observed in Weka (performed on the ‘glass’ dataset) :

1. Visualization : Visualize scatter plot for all the attributes from dataset selected from Weka. Determine correlation if any using these plots for different datasets.

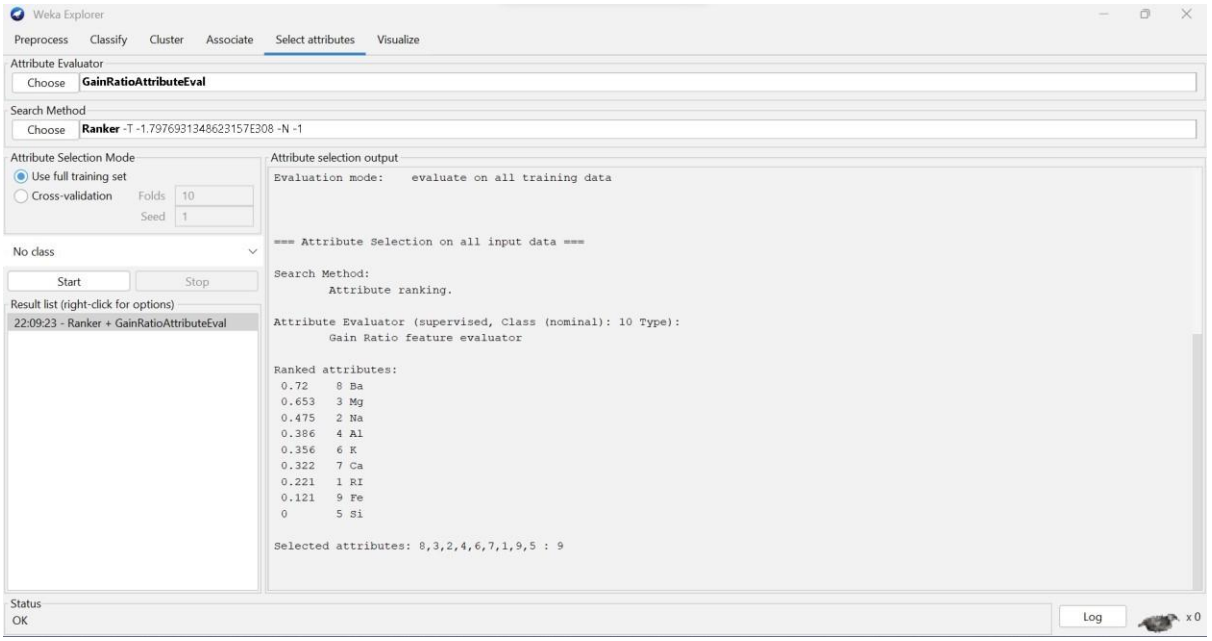




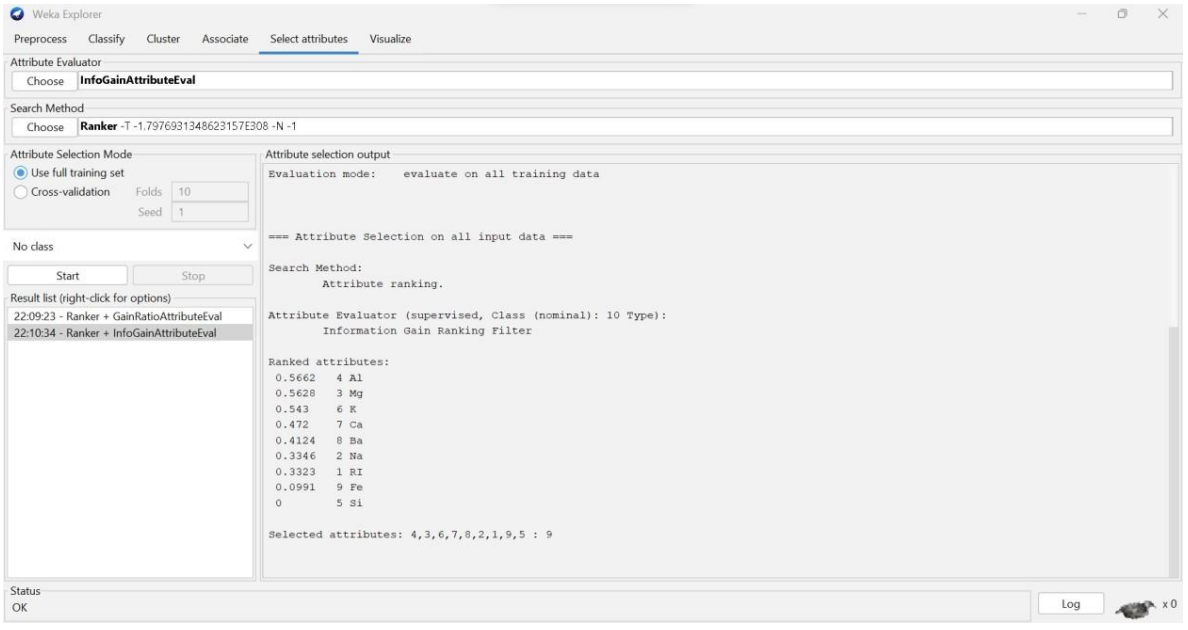
Observation – There is positive correlation in this graph which means that the relationship between two variables(types) tend to move in the same direction.

2. Select Attributes : Apply suitable feature selection filter like GainRatio etc to choose relevant attributes from the list of attributes. Observe the ranks / priority provided by the filter.

GainRatioAttributeEval –

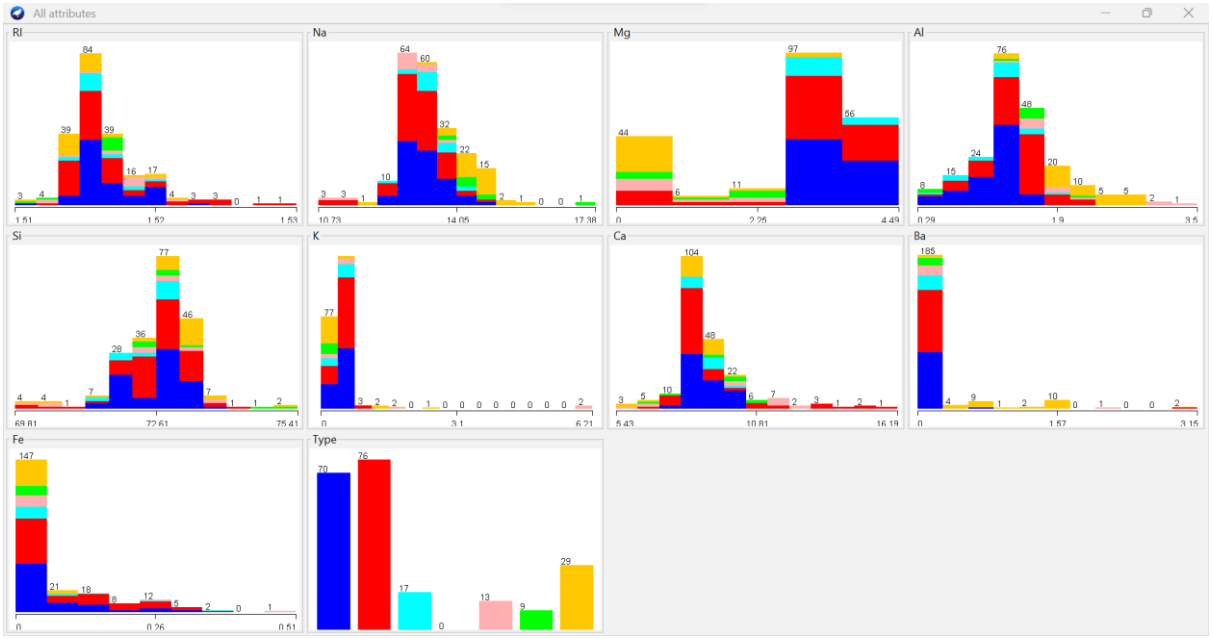


InfoGainAttributeEval –



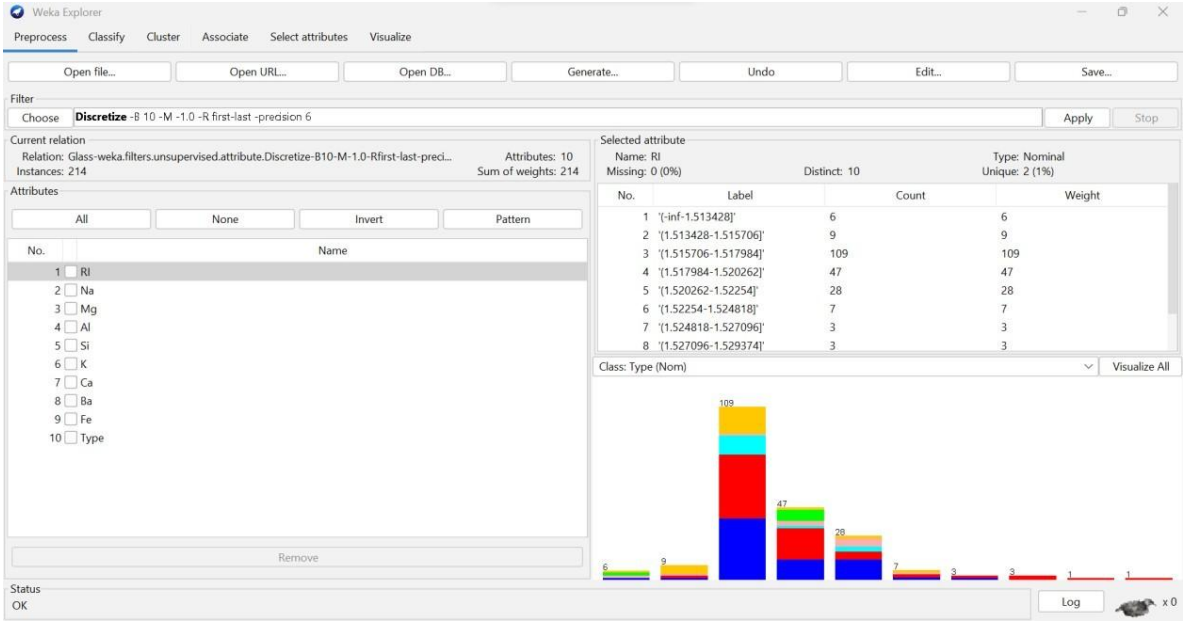
3. Pre-processing :

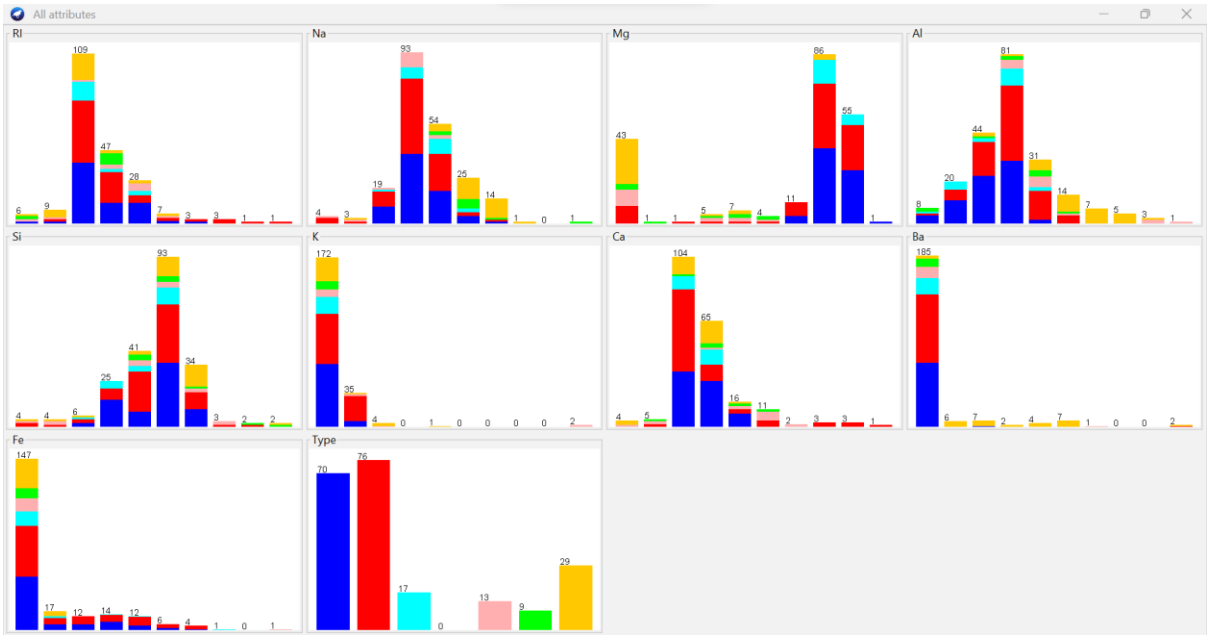
a) Visualize All: Select this button to visualize histograms of all attributes.



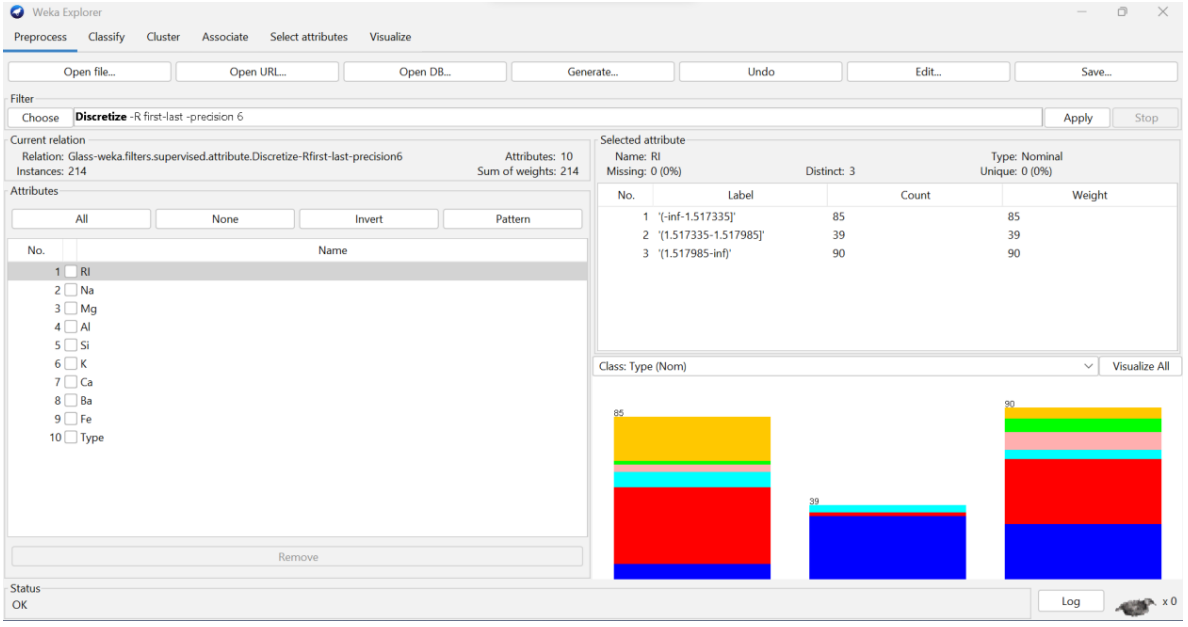
b) Filter: Choose Discretization under Unsupervised and Supervised methods. Observe the discretization and the outliers.

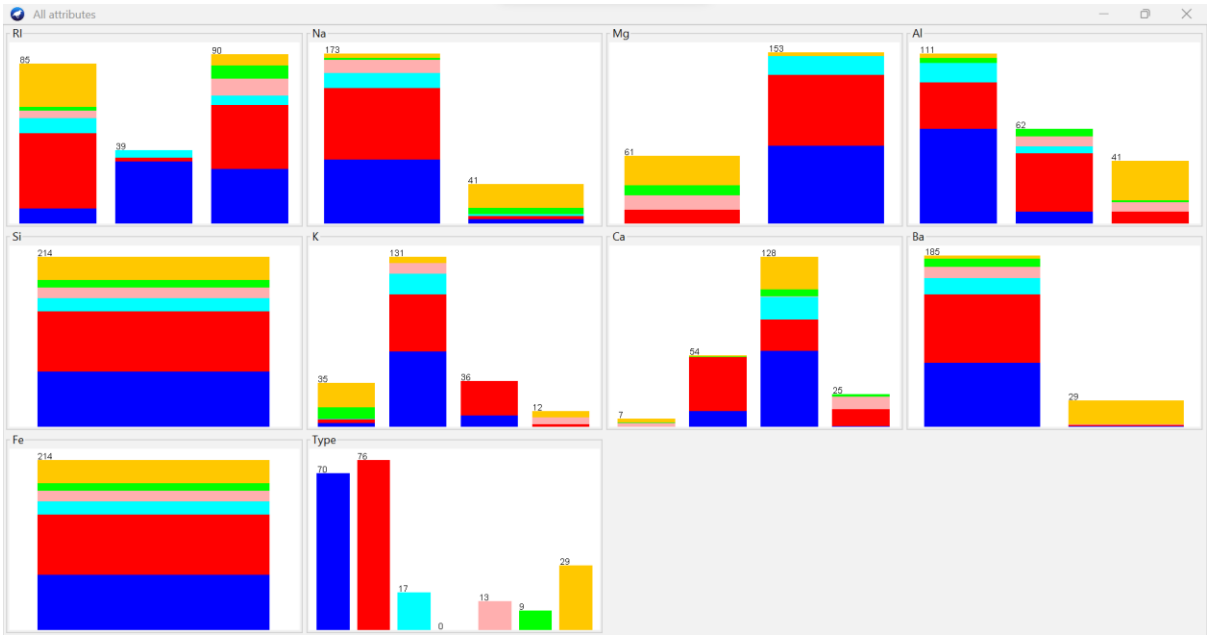
Discretization under Unsupervised method –



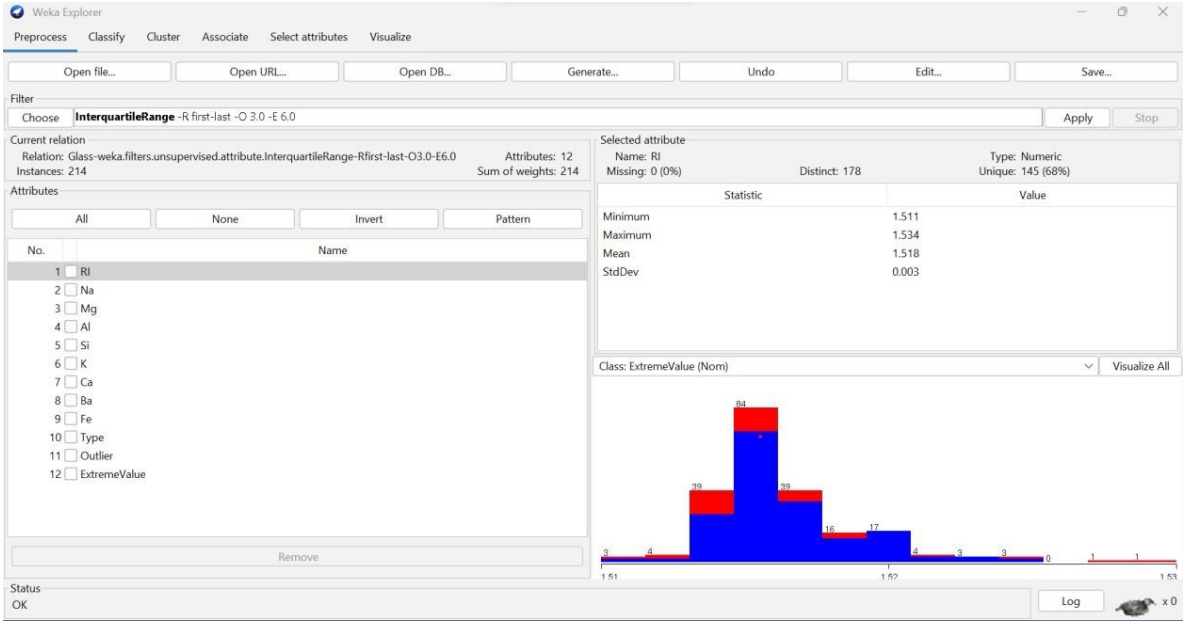


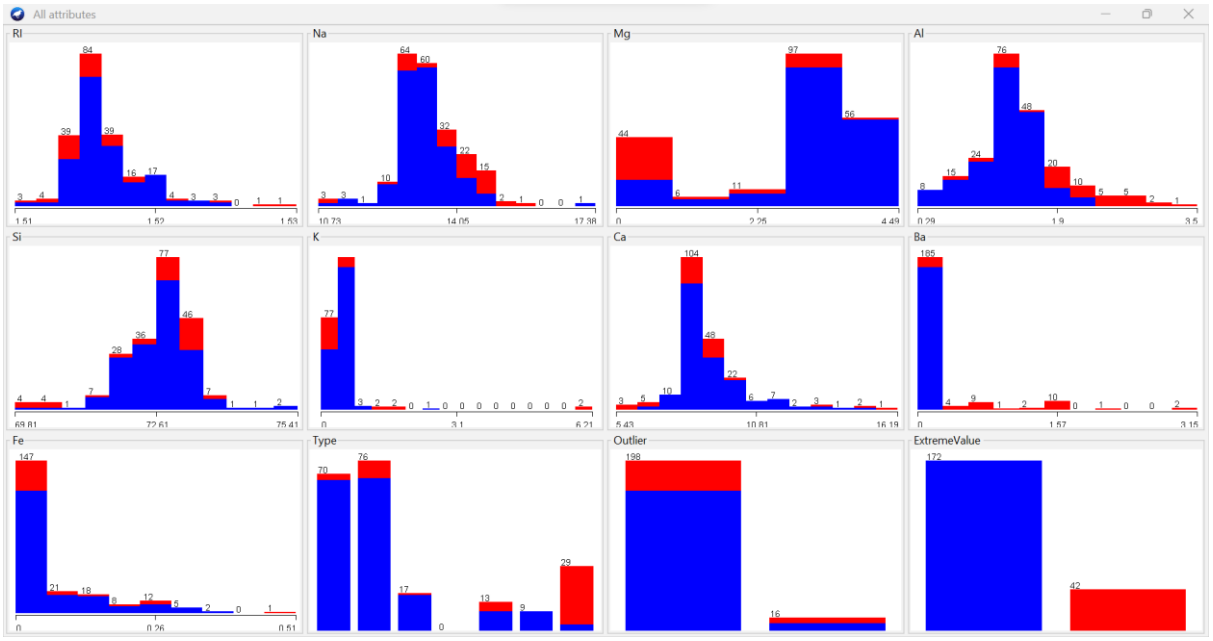
Discretization under Supervised method –



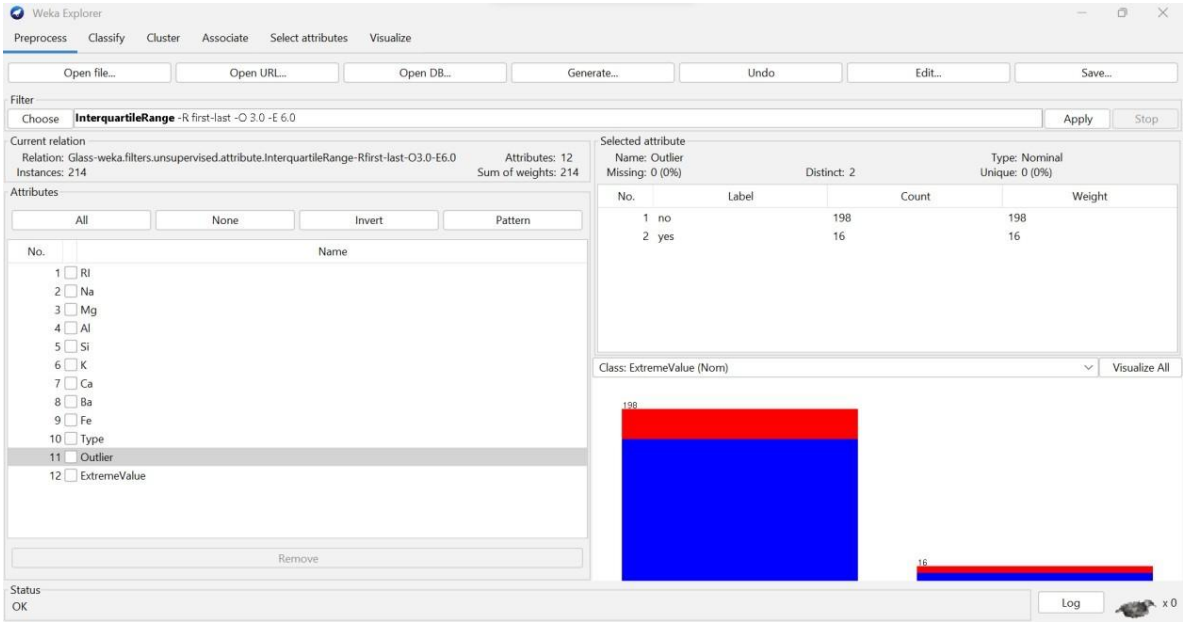


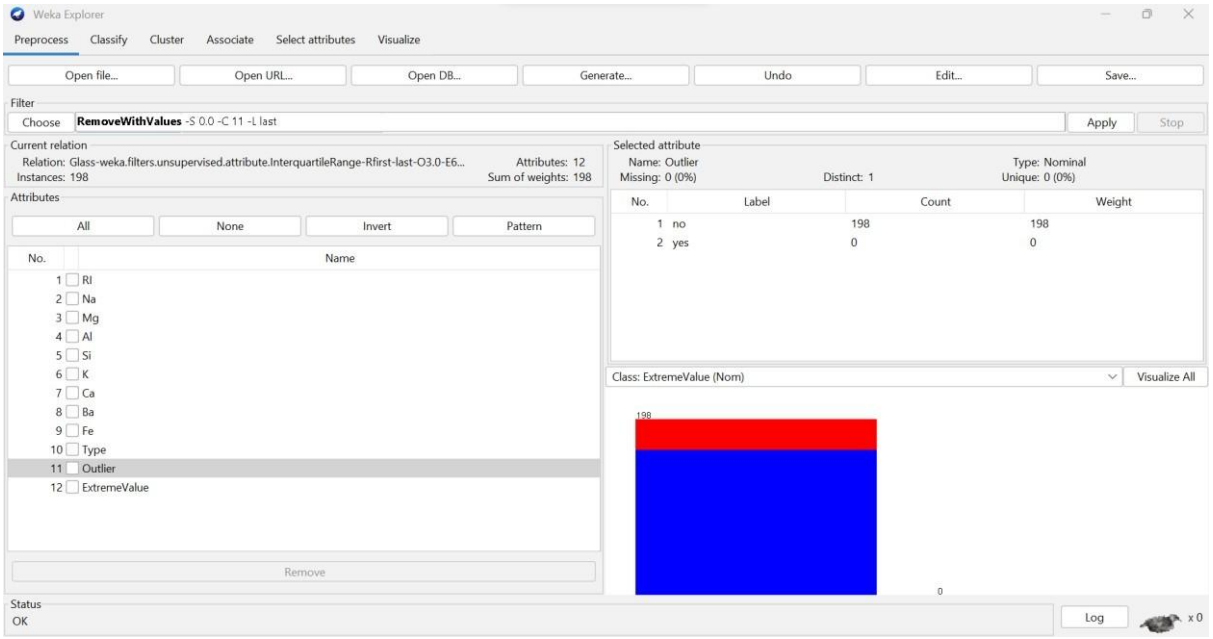
c) IQR: Observe the IQR values for a selected attribute. Observe the outlier and extreme values.



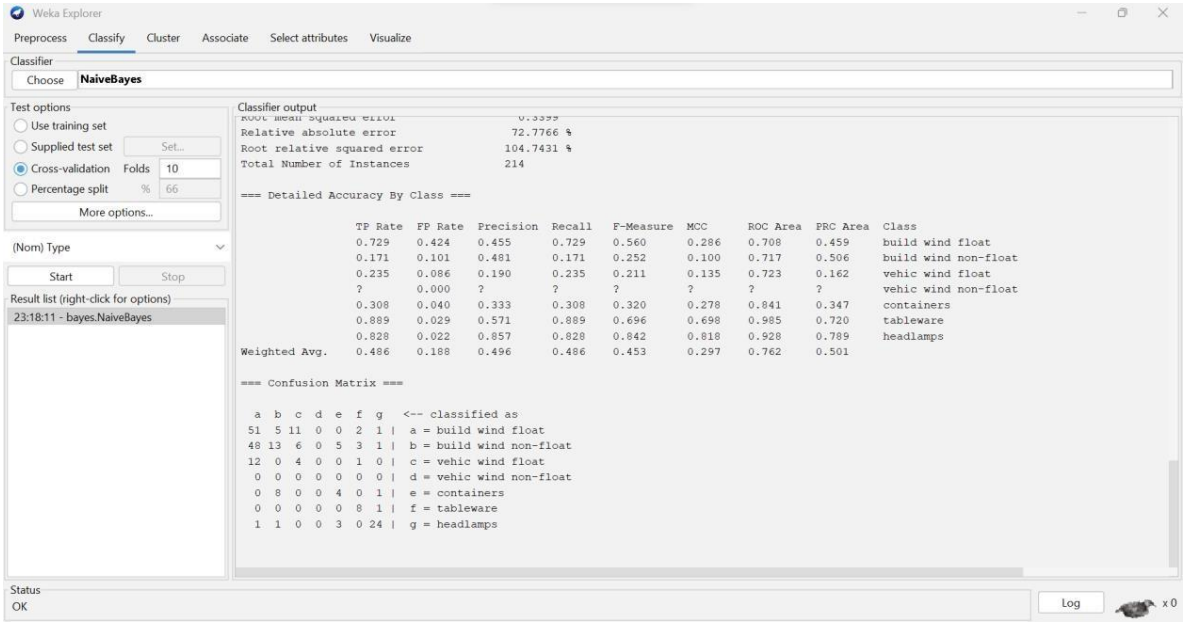


d) Removethevalue: Remove instances with outlier values and show the screenshots of dataset before and after the removal.

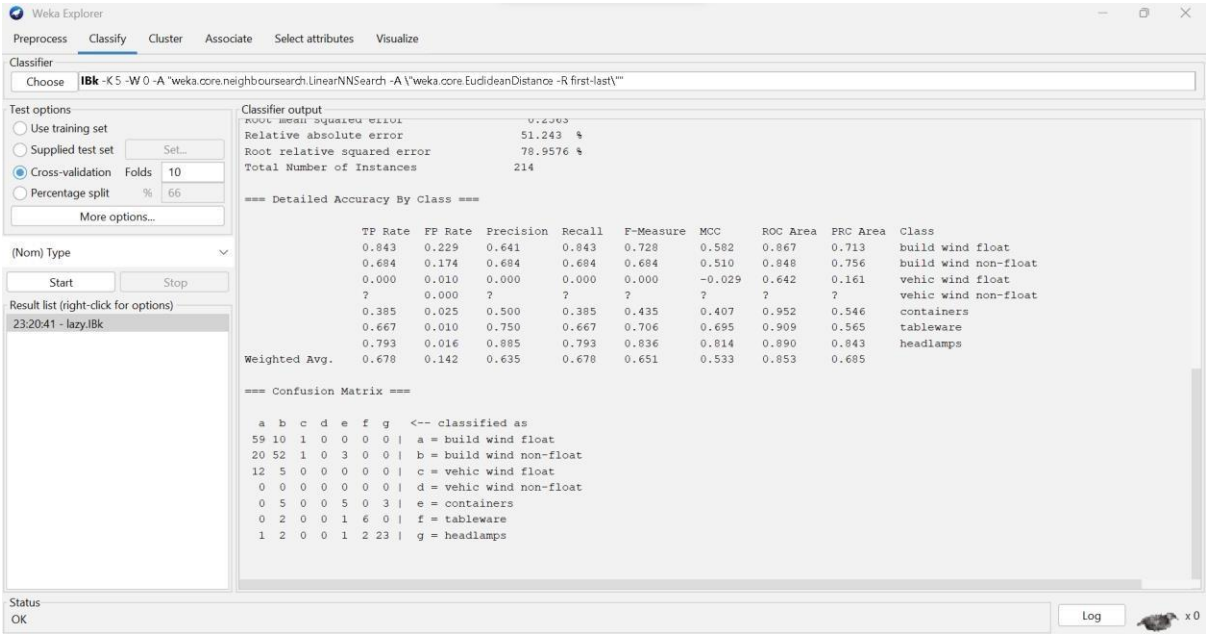




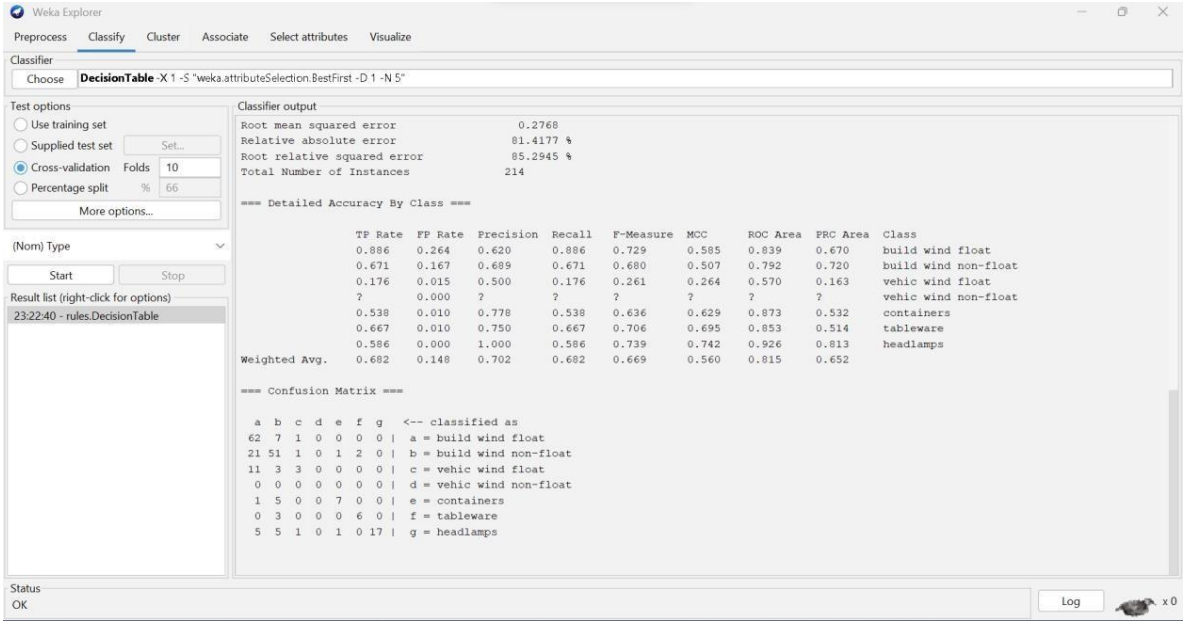
4. Classification : Perform NB, kNN and DT/rule-based classification. Naive Bayes –



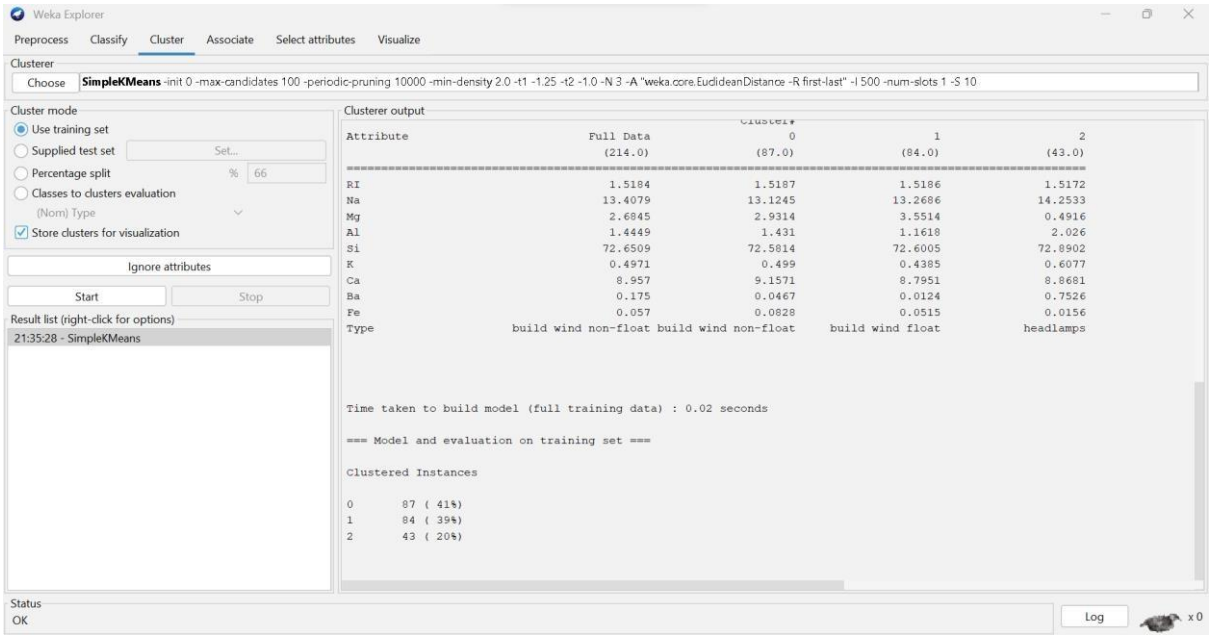
kNN –



DecisionTable –

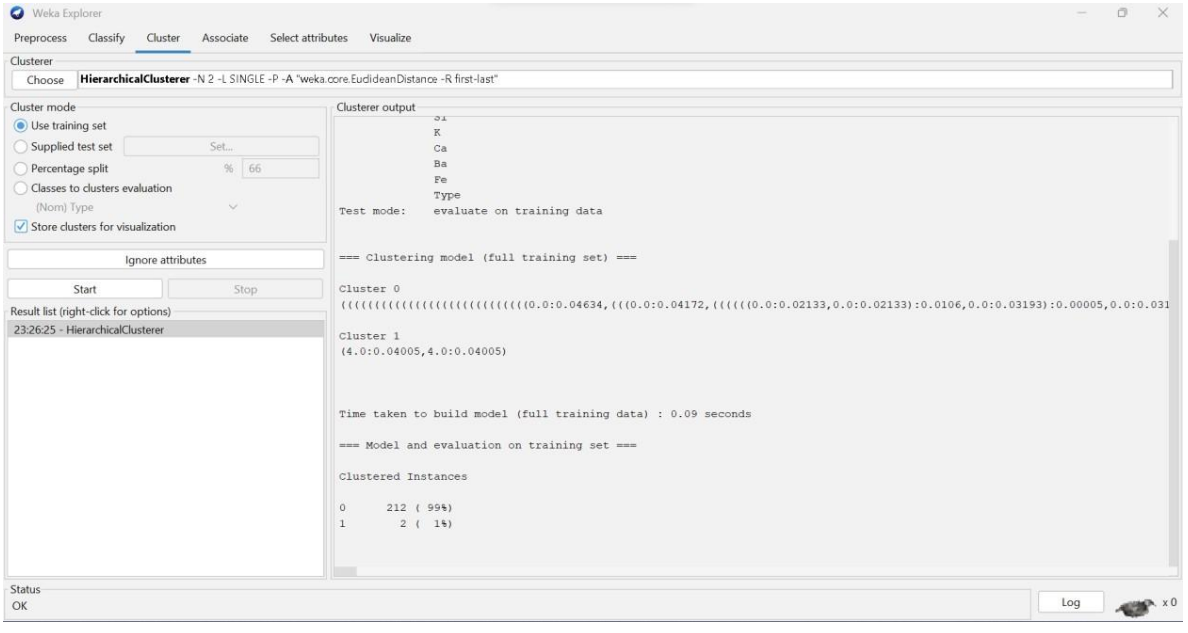


5. Clustering : Perform kmeans, hierarchical clustering and explain the output. KMeans –



Observation – SimpleKMeans Clustering is an Unsupervised Learning Algorithm. It groups the unlabelled dataset into different clusters. We have to define the number of pre-defined clusters that need to be created in the process and then discovered the categories of groups in the unlabelled dataset on its own without the need for any training. Here, I defined three clusters and the output shows all the instances divided into the three different clusters.

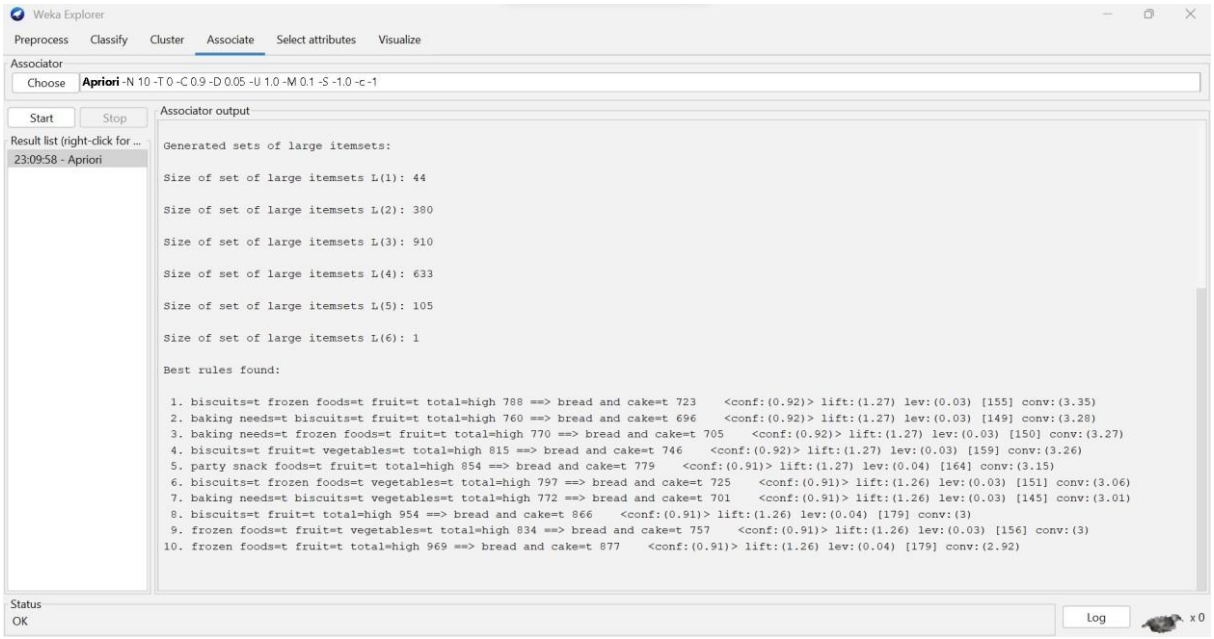
HierarchicalCluster –



Observation - Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabelled datasets into a cluster. In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram. In this algorithm, we do not need to have knowledge about the predefined number of clusters. Here, I set the number of clusters to be 2 and the output shows the instances are divided into the respective two clusters.

6. Association rule mining : Perform apriori algo and show the rules

created. (Performed on the 'supermarket' dataset)



Conclusion :

Weka tool is useful and convenient tool to perform various preprocessing and data mining algorithms.