JUNAID GIRKAR

60004190057

TE Comps A4

LAB EXPERIMENT NO. 01

Aim: Perform data Pre-processing task using Weka data mining tool

Theory:

WEKA - an open source software provides tools for data preprocessing, 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:

Preprocessing:

Classification:

Clustering:

Association Rule:

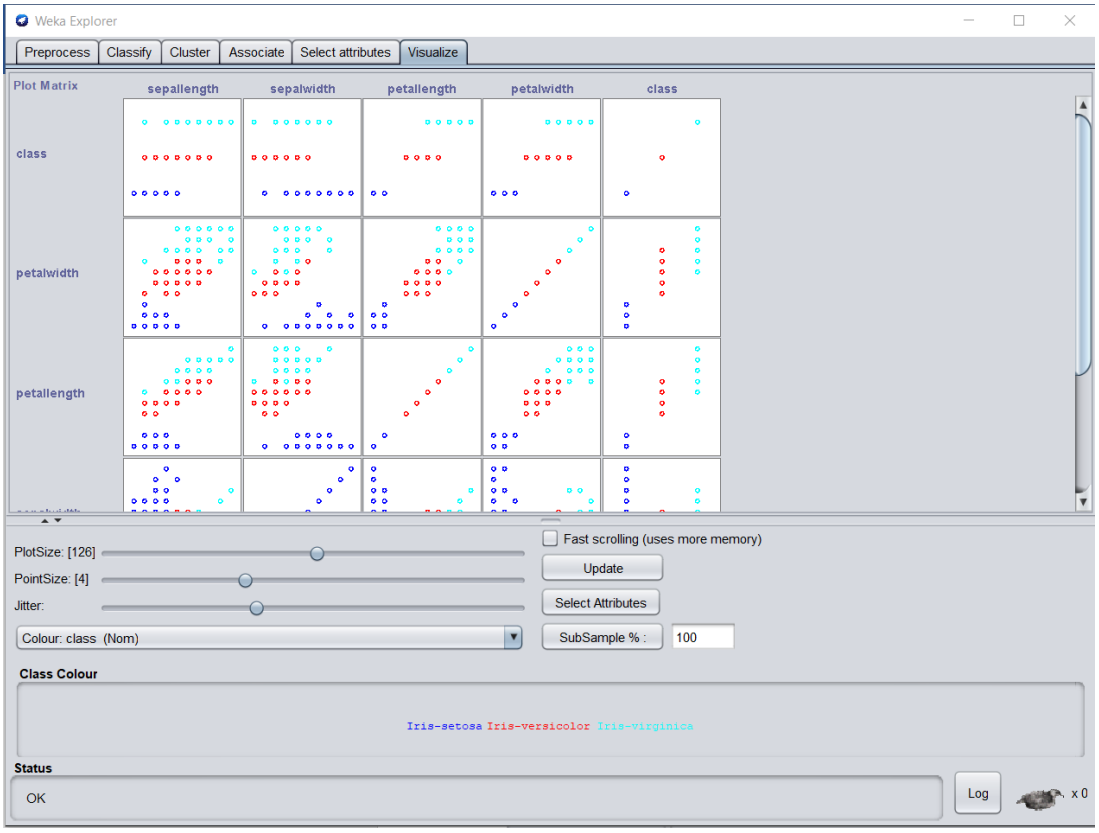
Select Attributes:

Visualization:

Preprocessing activities to be observed in Weka:

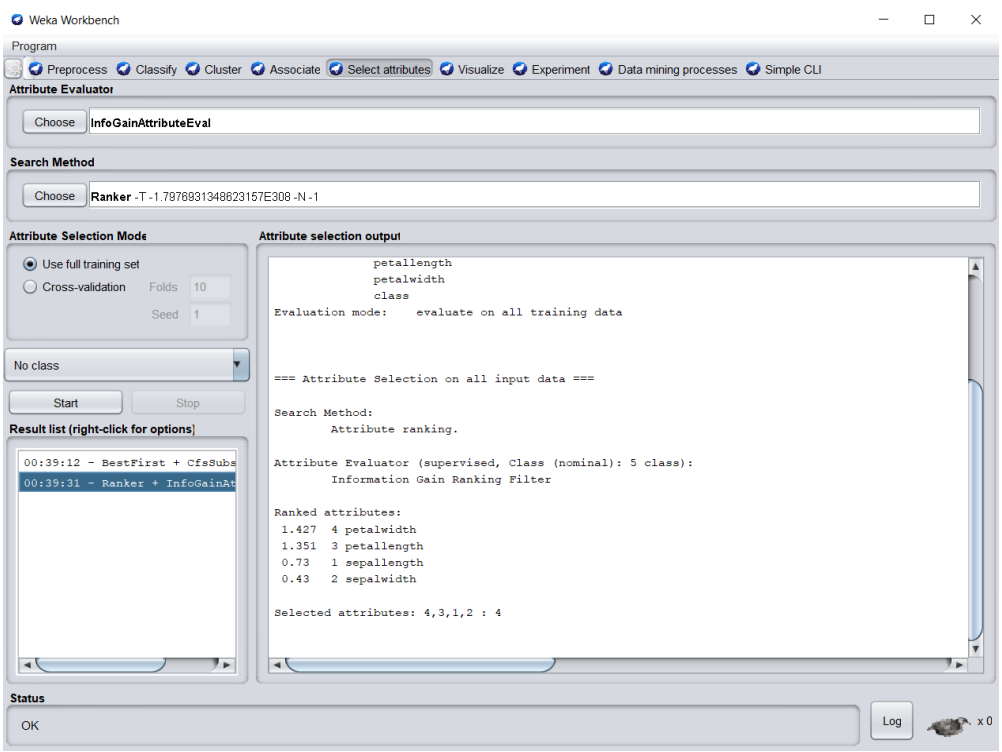
1. **Visualization:** Visualize scatter plot for all the attributes from dataset selected from Weka. Determine correlation if any using these plots for different datasets

   The visualize tab is for reviewing the pairwise scatterplot matrix of each attribute plotted against every other attribute in the loaded dataset. It is useful to get an idea of the shape and relationship of attributes that may aid in data filtering, transformation and modeling.
   Increase the point size and the jitter and click the “Update” button to set an improved plot of the categorical attributes of the loaded dataset.



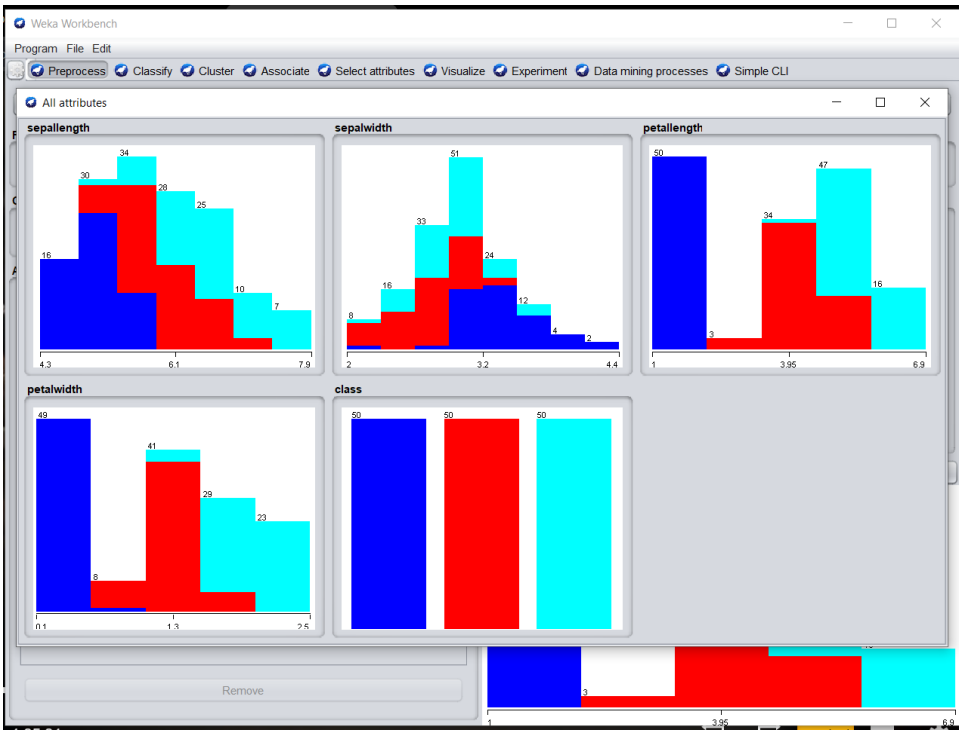
2. **Select Attributes:** Apply suitable feature selection filter like GainRatio etc to choose relevant attributes from the list of attributes. Observe the ranks / priority provided by the filter.

   The **select attributes** tab is for performing feature selection on the loaded dataset and identifying those features that are most likely to be relevant in developing a predictive model.



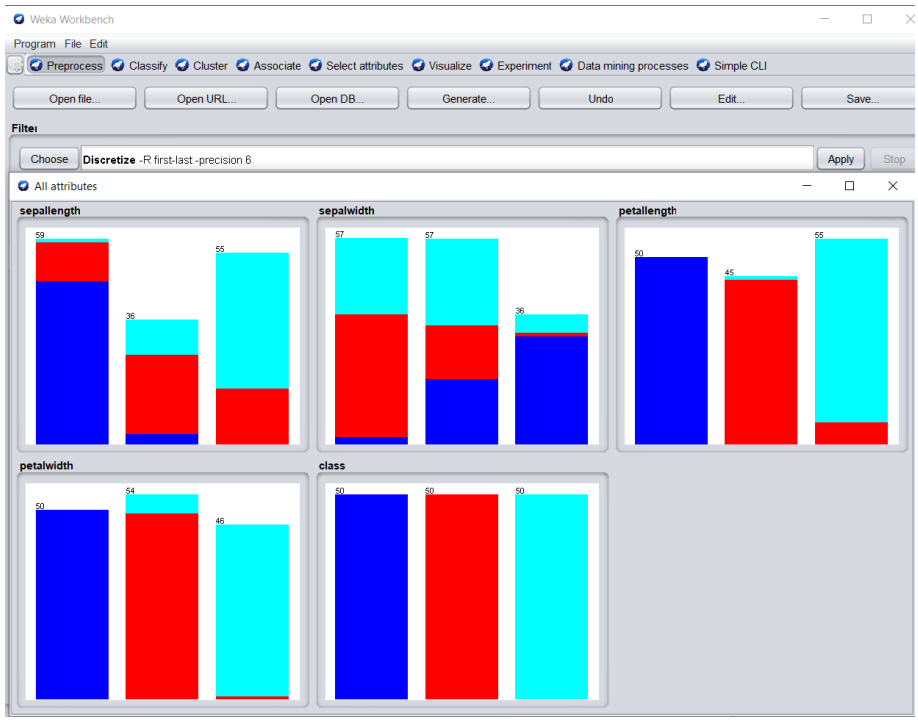
3. Preprocessing:

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

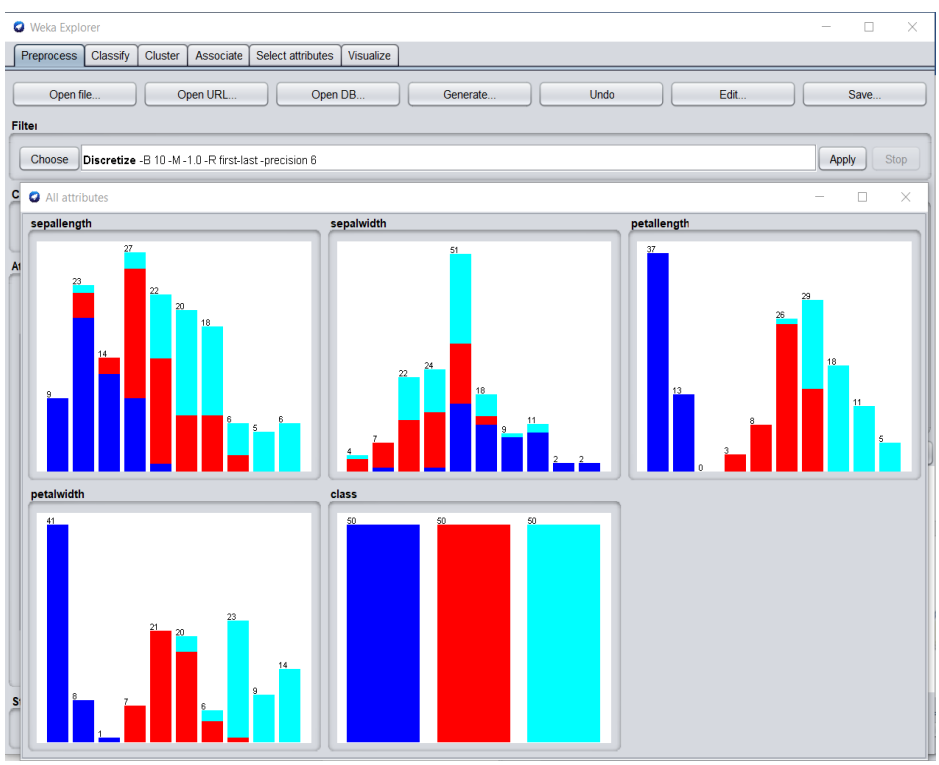


b. Filter: Choose Discretization under Unsupervised and Supervised methods. Observe the discretization and the outliers.

SUPERVISED DISCRETIZE FILTERING

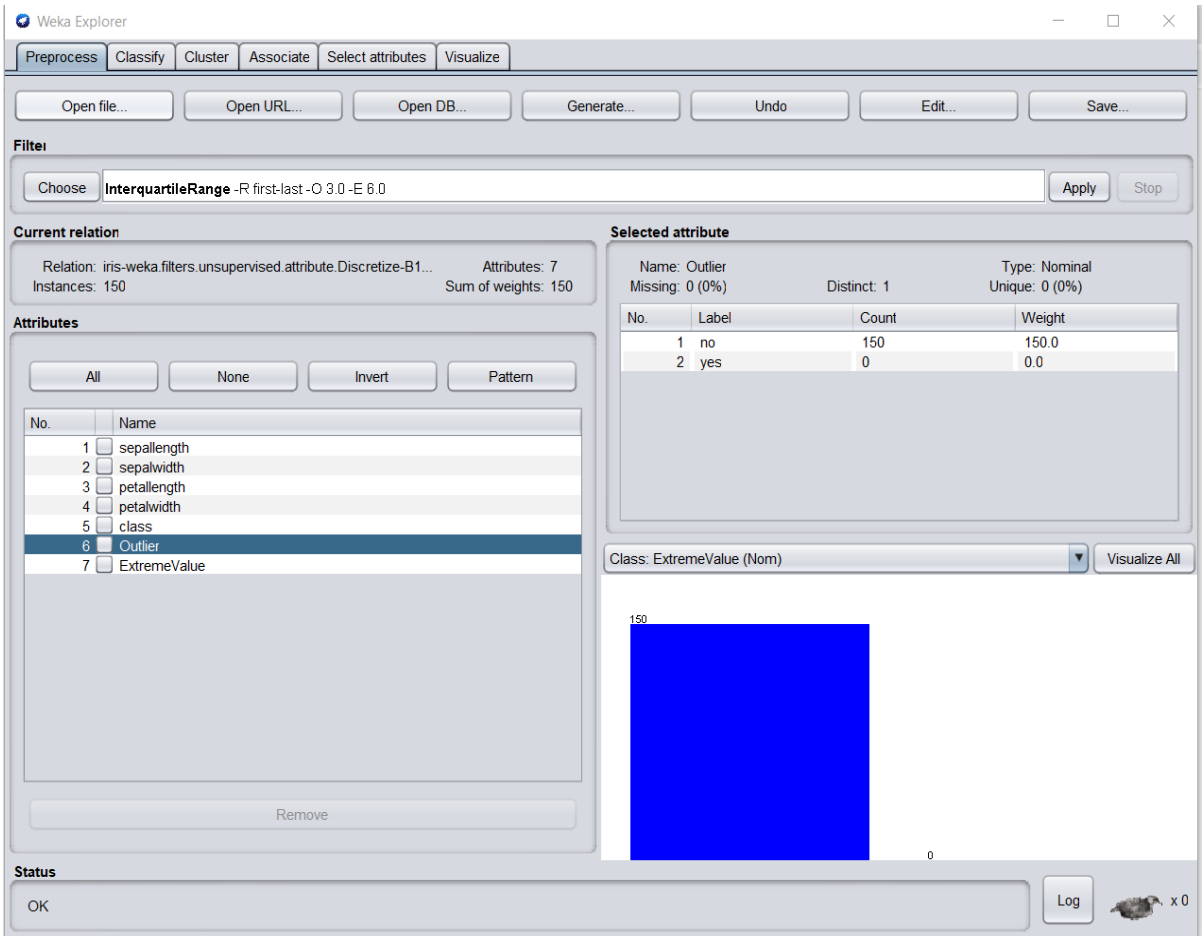


UNSUPERVISED DISCRETIZE FILTERING

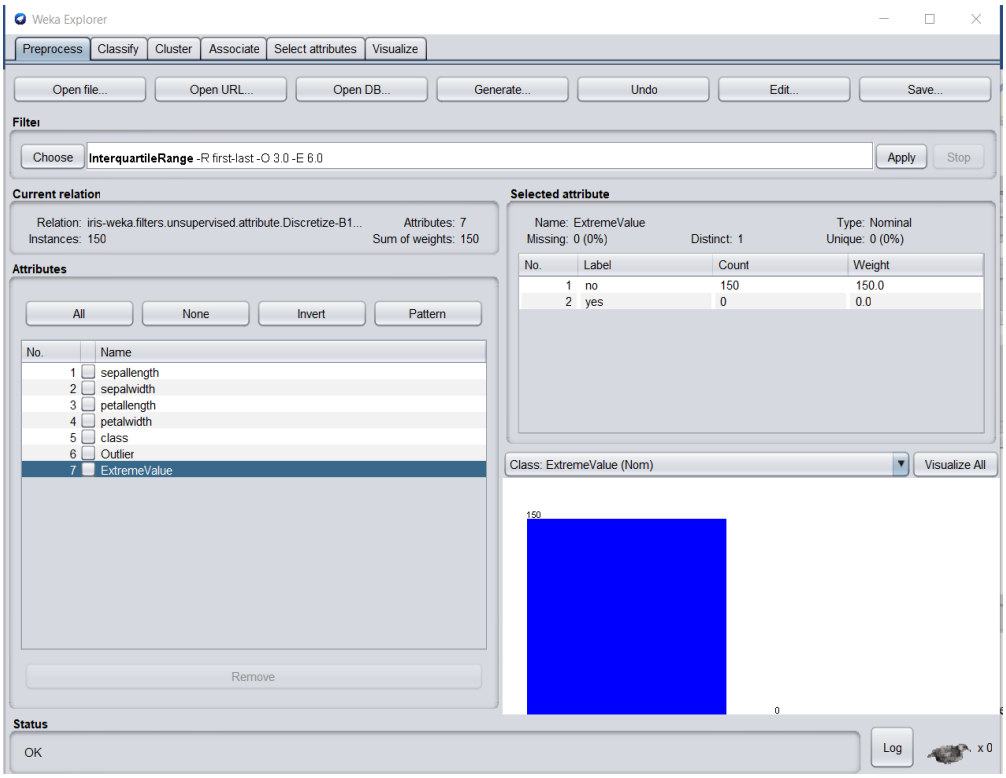


c. **IQR:** Observe the IQR values for a selected attribute. Observe the outlier and extreme values

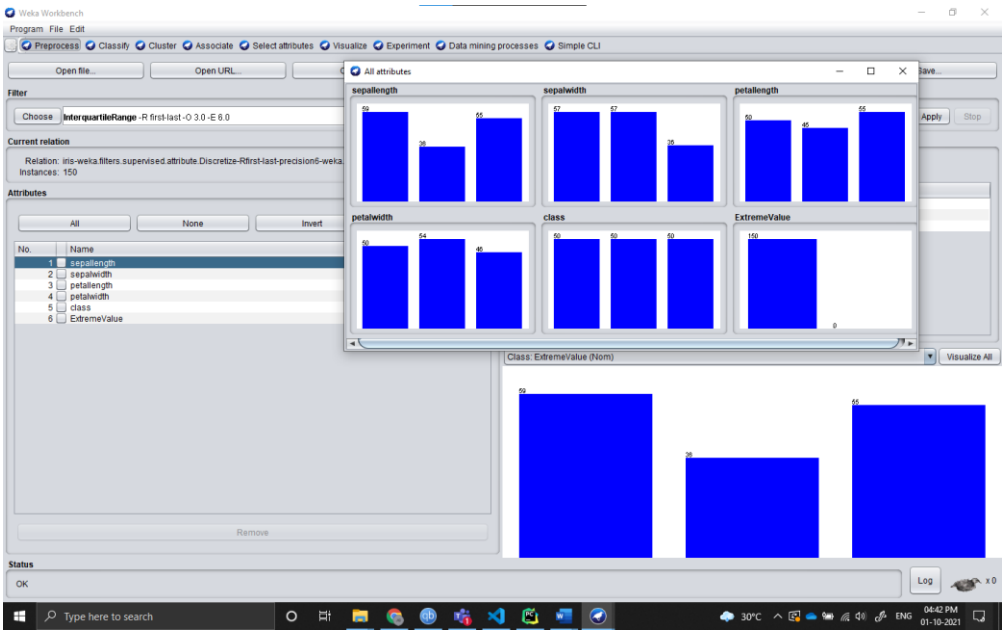
OUTLIER



EXTREME VALUE

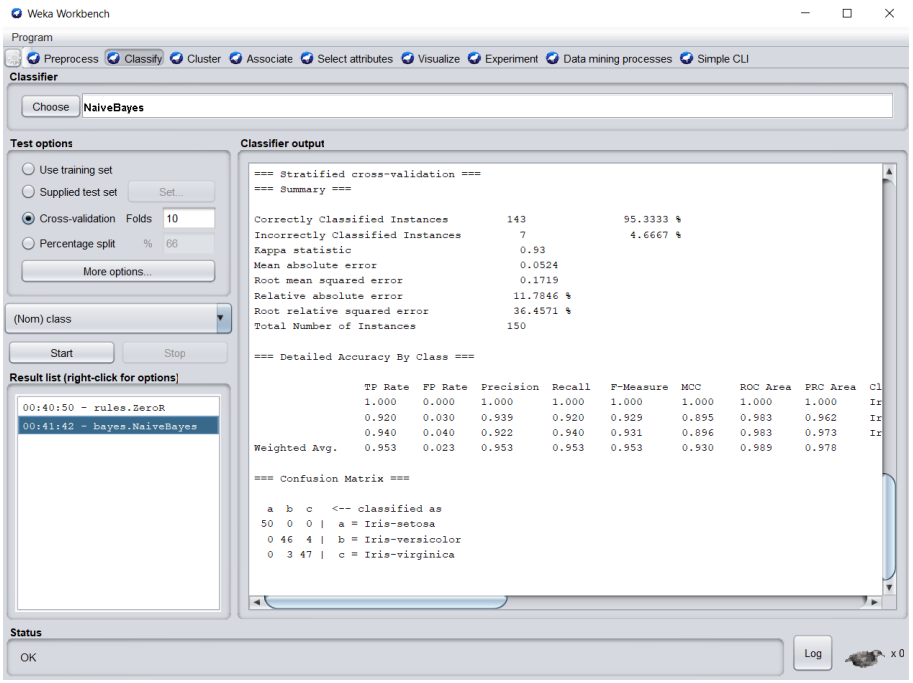


d. **Removethevalue:** Remove instances with outlier values and show the screenshots of dataset before and after the removal.



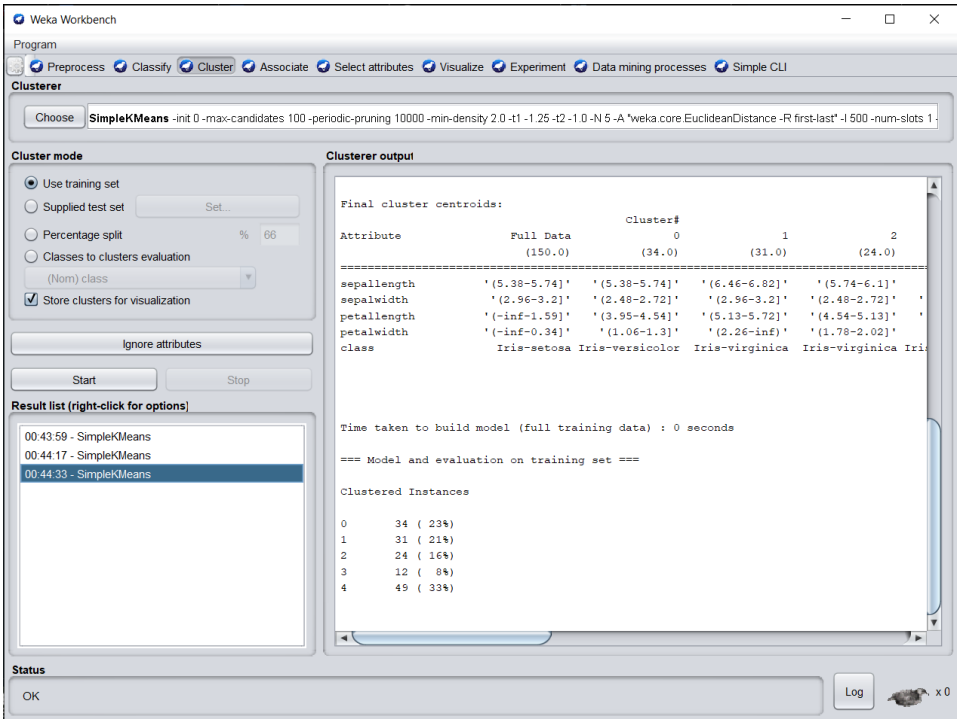
4. **Classification:** Perform NB, kNN and DT/rule based classification

   The **classify** tab is for training and evaluating the performance of different machine learning algorithms on your classification or regression problem. Algorithms are divided up into groups, results are kept in a result list and summarized in the main Classifier output.



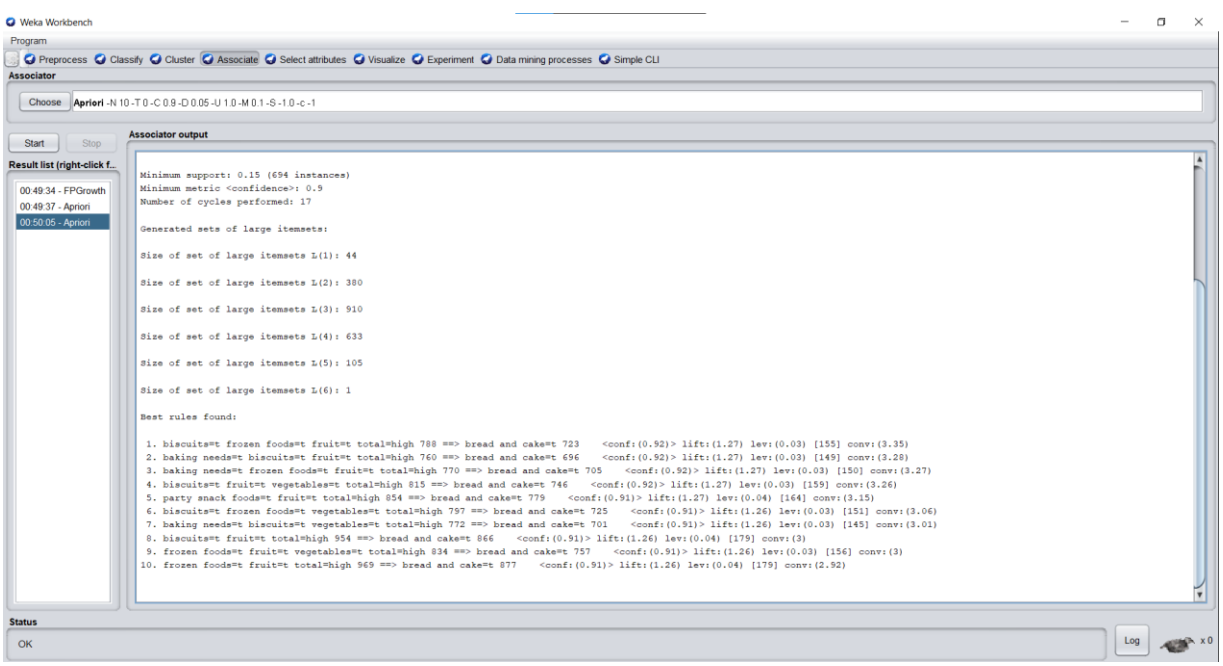
5. **Clustering:** Perform kmeans, hierarchical clustering and explain the output

   The **cluster** tab is for training and evaluating the performance of different unsupervised clustering algorithms on your unlabeled dataset. Like the Classify tab, algorithms are divided into groups, results are kept in a result list and summarized in the main Clusterer output.



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

The **associate** tab is for automatically finding associations in a dataset. The techniques are often used for market basket analysis type data mining problems and require data where all attributes are categorical.



CONCLUSION:

We learnt about the Weka tool and how to do data analysis with it. We used 2 different databases: Iris petals and Supermarket.

We tried both the supervised and unsupervised learning algorithms. We can easily visualize with charts how the data transforms when we filter it using different algorithms.

We also used the select attribute to find out which attribute is ranked best for classification. We implemented different clustering and classification algorithms.

In the second database i.e., the supermarket one, we implemented the associate function where we found the different associations in a dataset.