Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)
NAAC Accredited with "A" Grade (CGPA : 3.18)

A.Y. 2022-2023

LAB EXPERIMENT NO. 01

Name: Parth Parekh **SAP ID:60004200006**

Branch:Computer Engineering **Batch:A1**

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

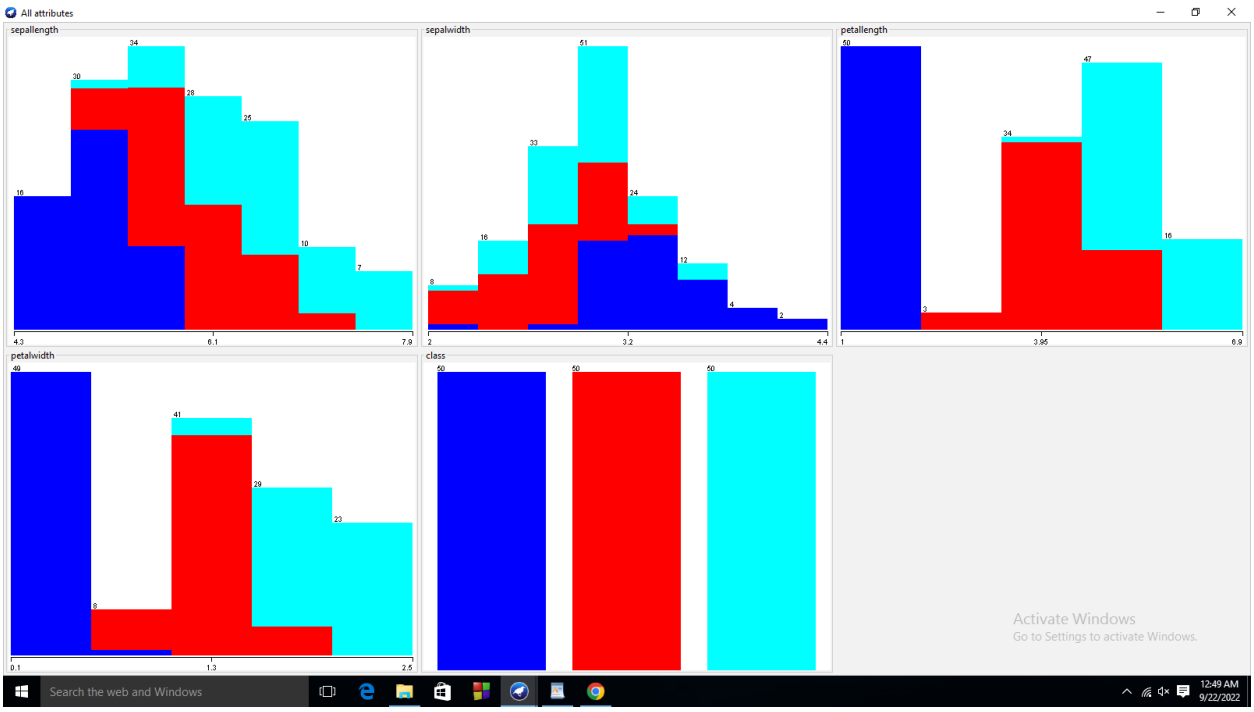
Theory:

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:

1. **Preprocessing:**

A. Visualize all:

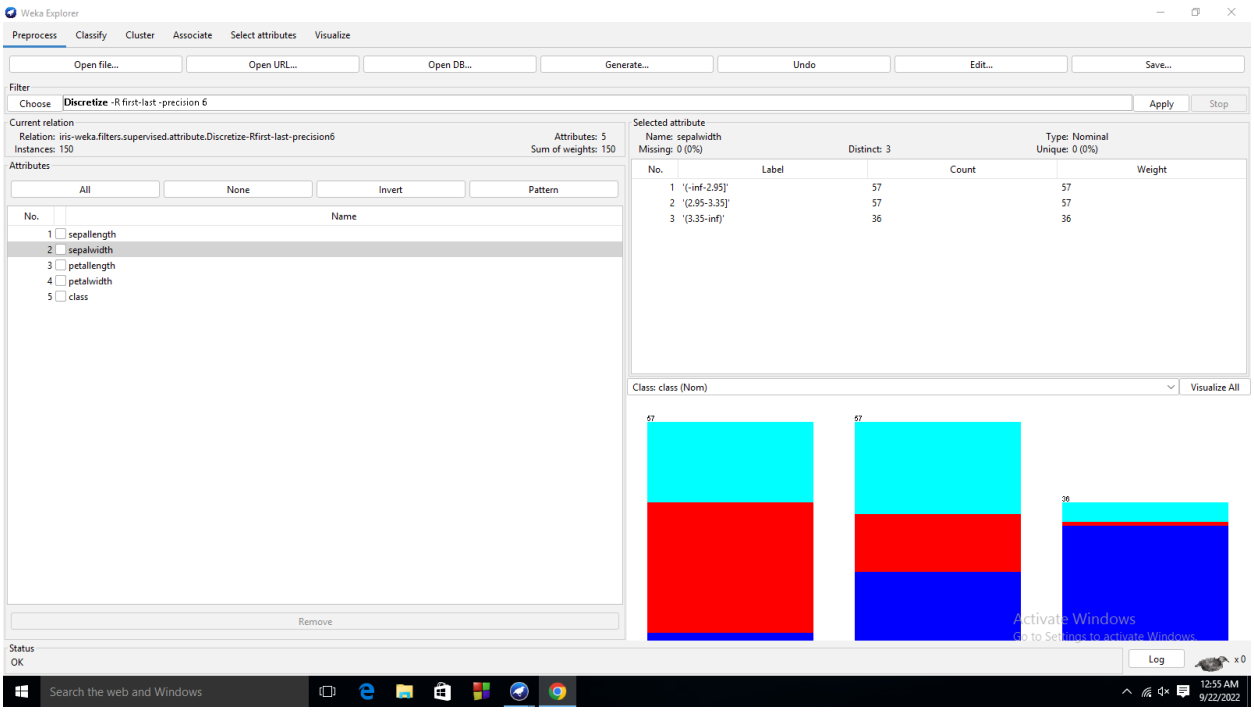


This is histogram visualization of all attributes without any filter for iris dataset.We can infer that labels in sepal length and sepal width are closely related in terms of frequency distribution and

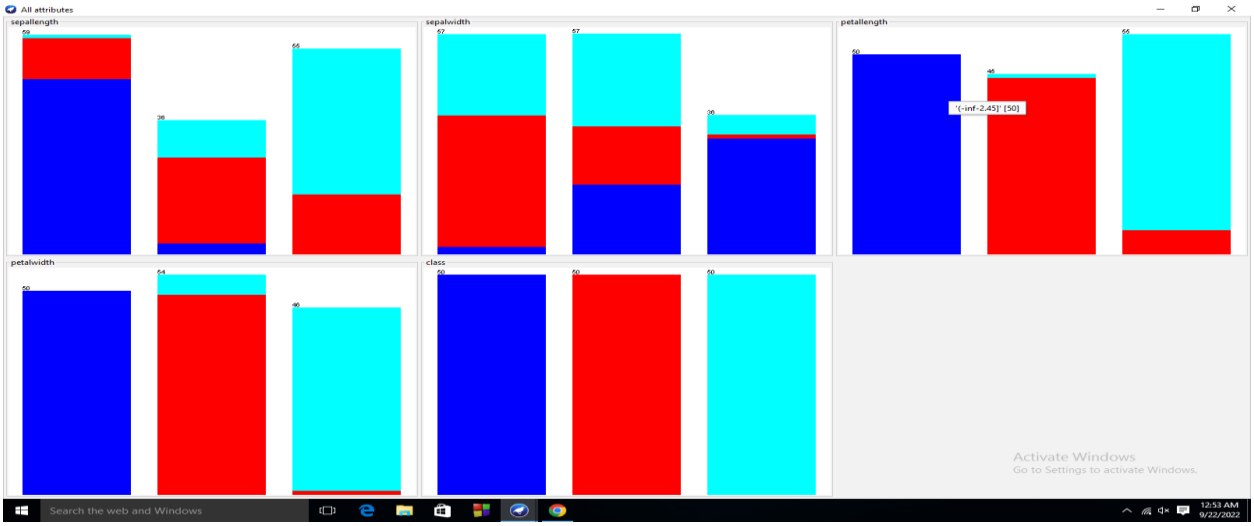
sepal length is skewed-right and sepal width is skewed-left whereas label in petal length and petal width are not so closely related in terms of frequency distribution.

B. Filter:

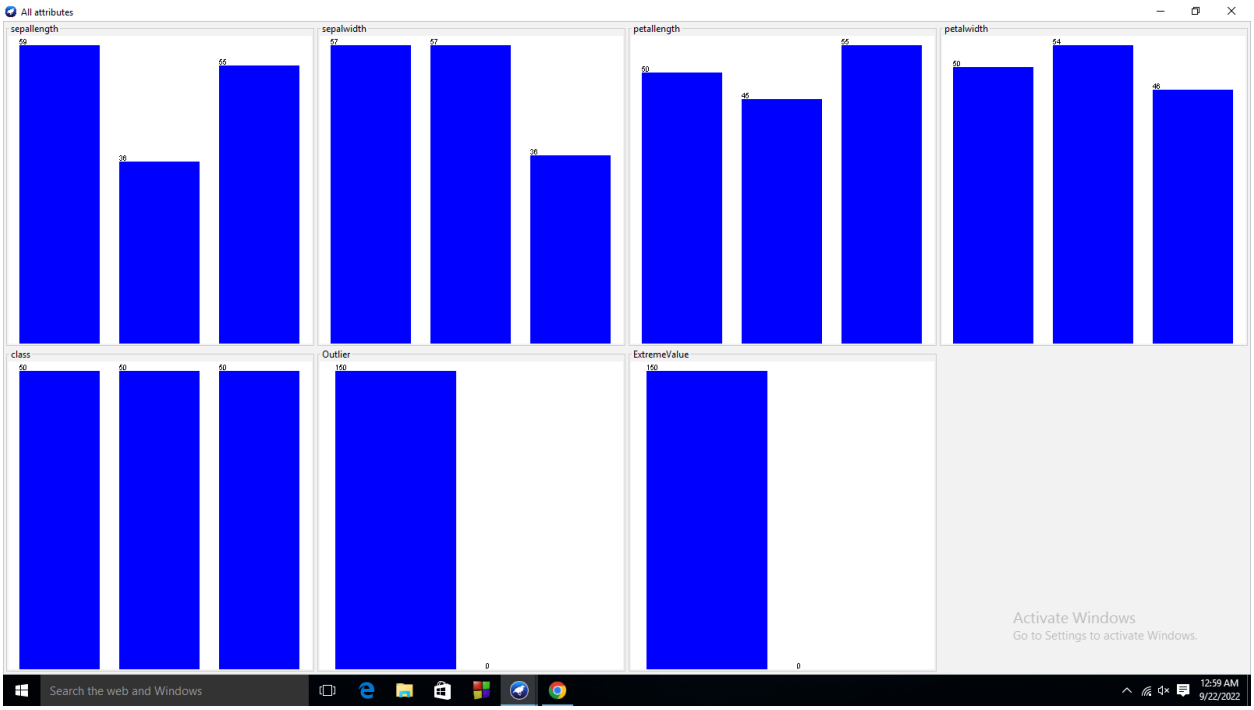
i) Discretization under supervised learning



This is Preprocessing window of discrete supervised class (nominal datatype) where we can see it created three labels of different numerical intervals.

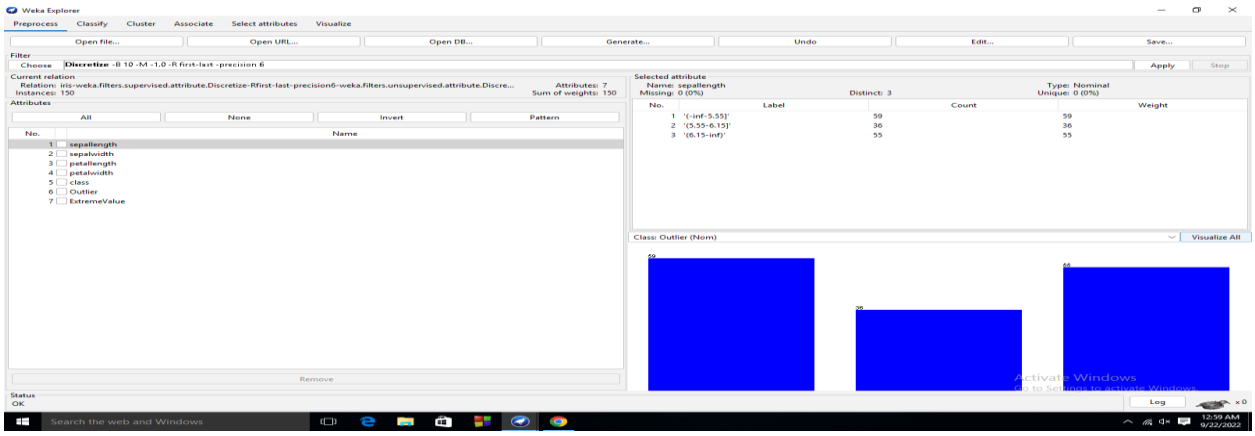


This is stacked column chart of discretized supervised learning



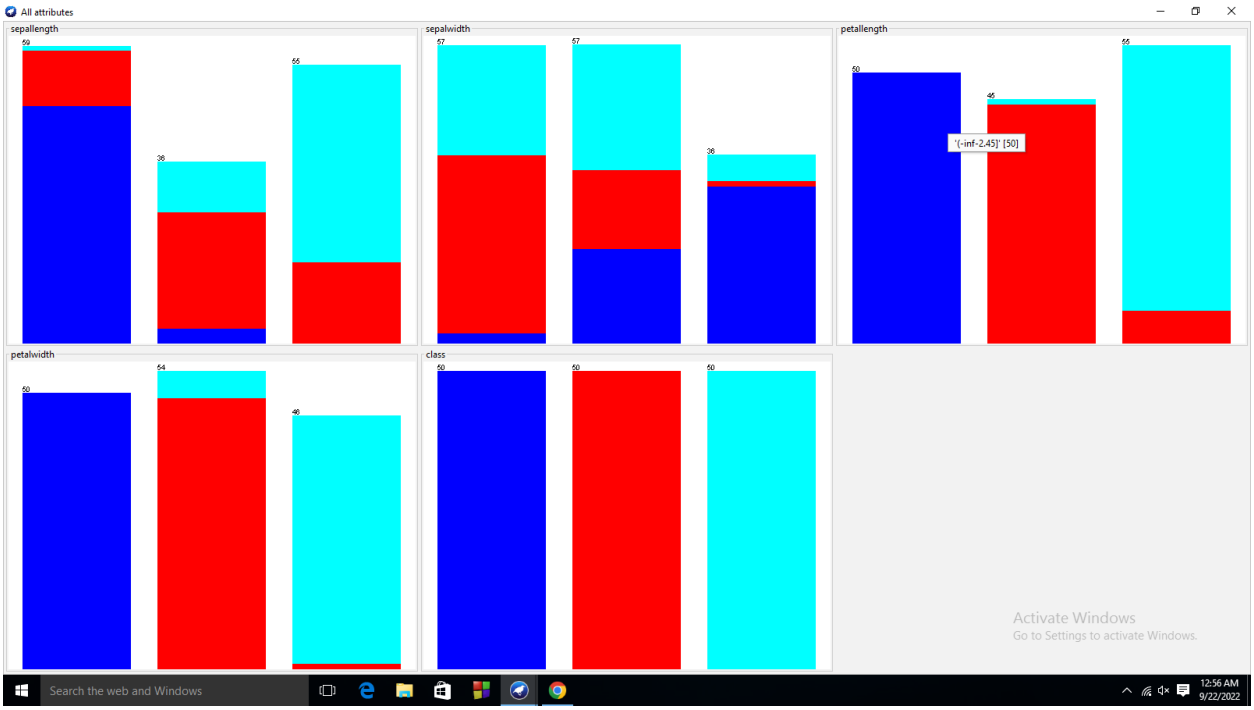
Above picture shows the extreme and outlier values of different attributes extreme values for discretize supervised learning . Insights from this is that extreme value for sepal length is 59 and two numerical intervals have same extreme value of 57 in sepal width.Outlier and Extreme values are 150.

ii) Discretization under unsupervised learning

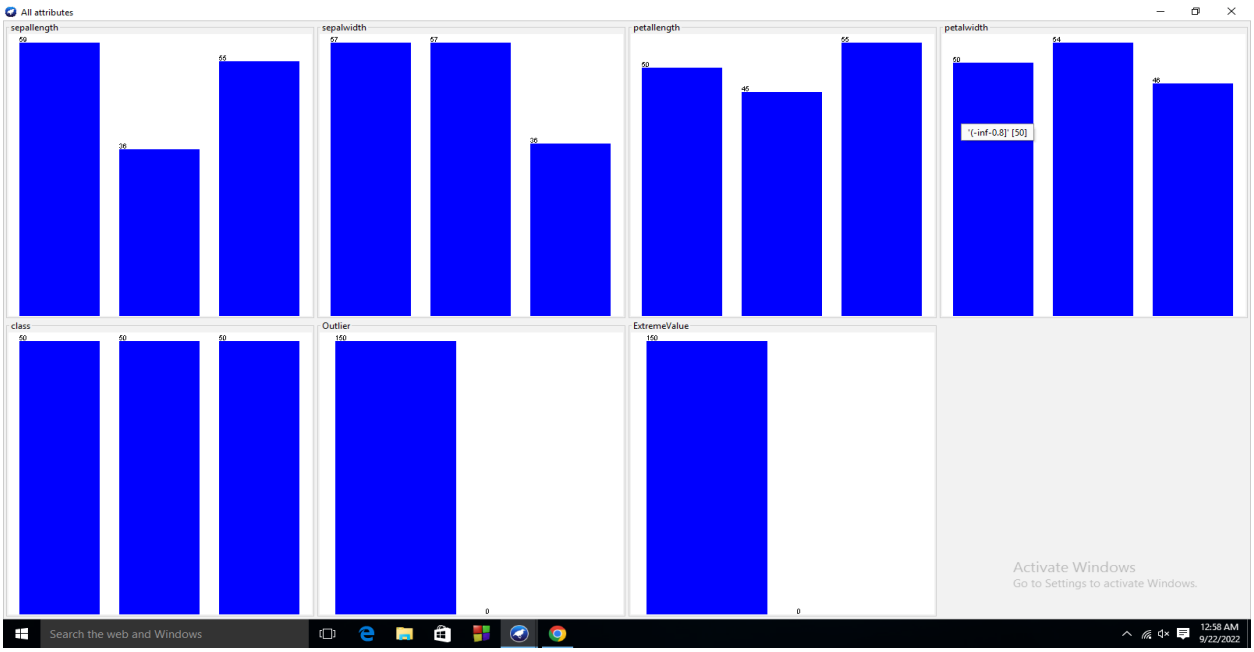


Preprocessing window of discretize unsupervised window where outlier for first bin is 59 which is the highest.

A.Y. 2022-2023

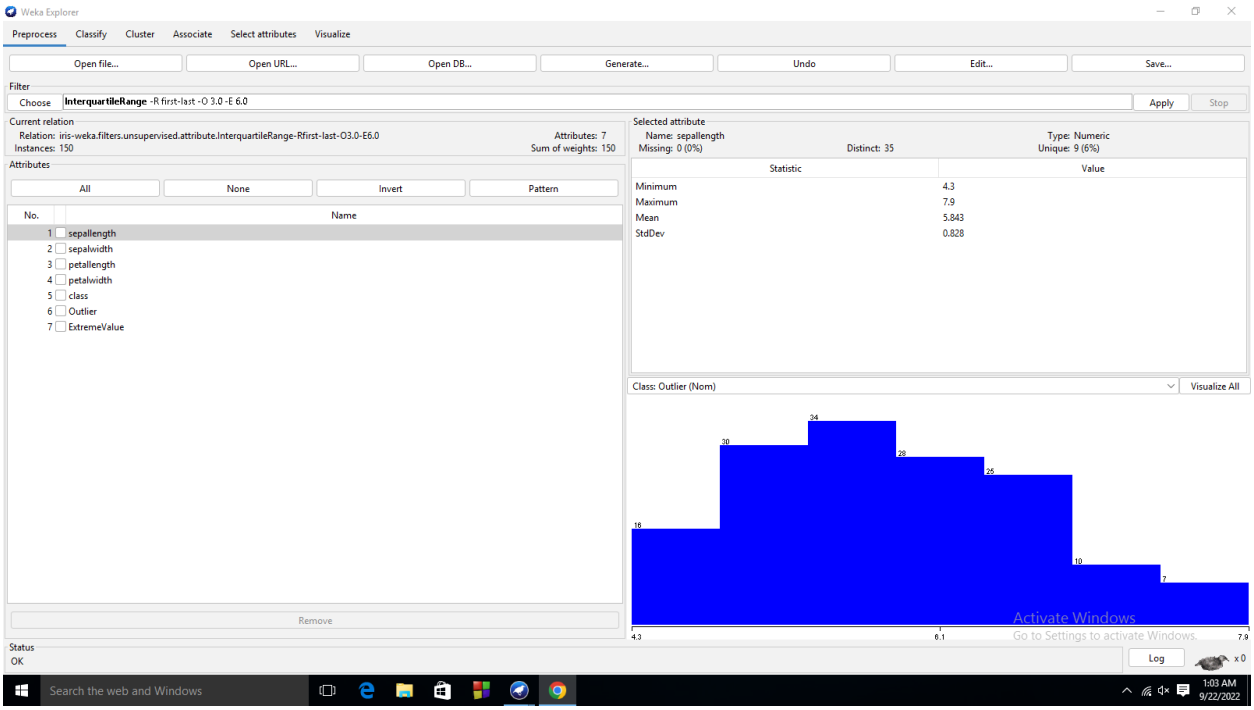


This is stacked column chart of discretized unsupervised learning

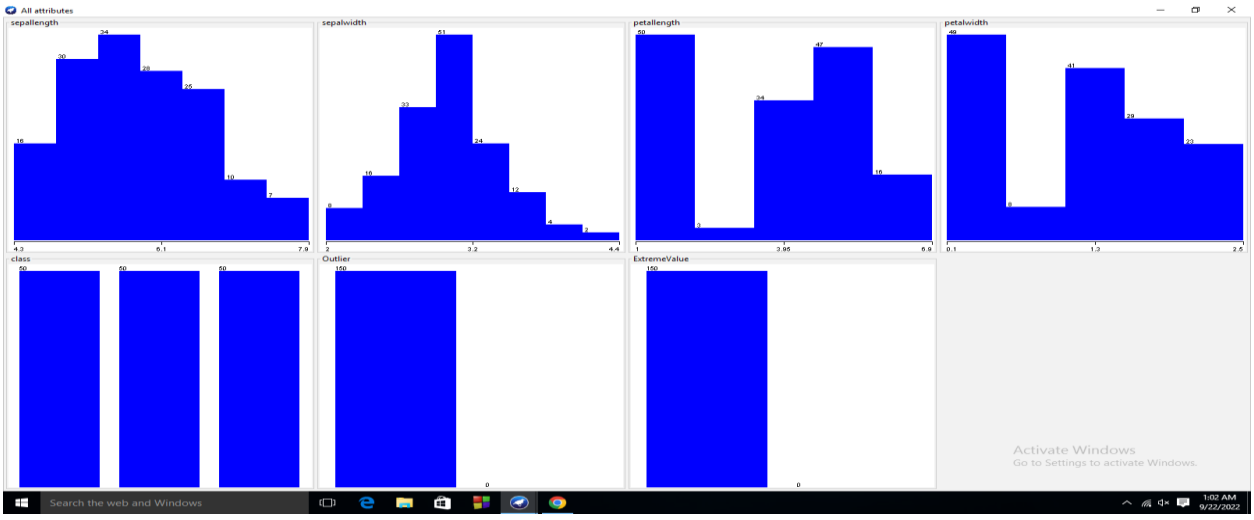


Above picture shows the extreme and outlier values of different attributes extreme values for discretize unsupervised learning . Insights from this is that extreme value for petal length is 55 and that of the petal width is 54.Outlier and Extreme values are 150

C. IQR:



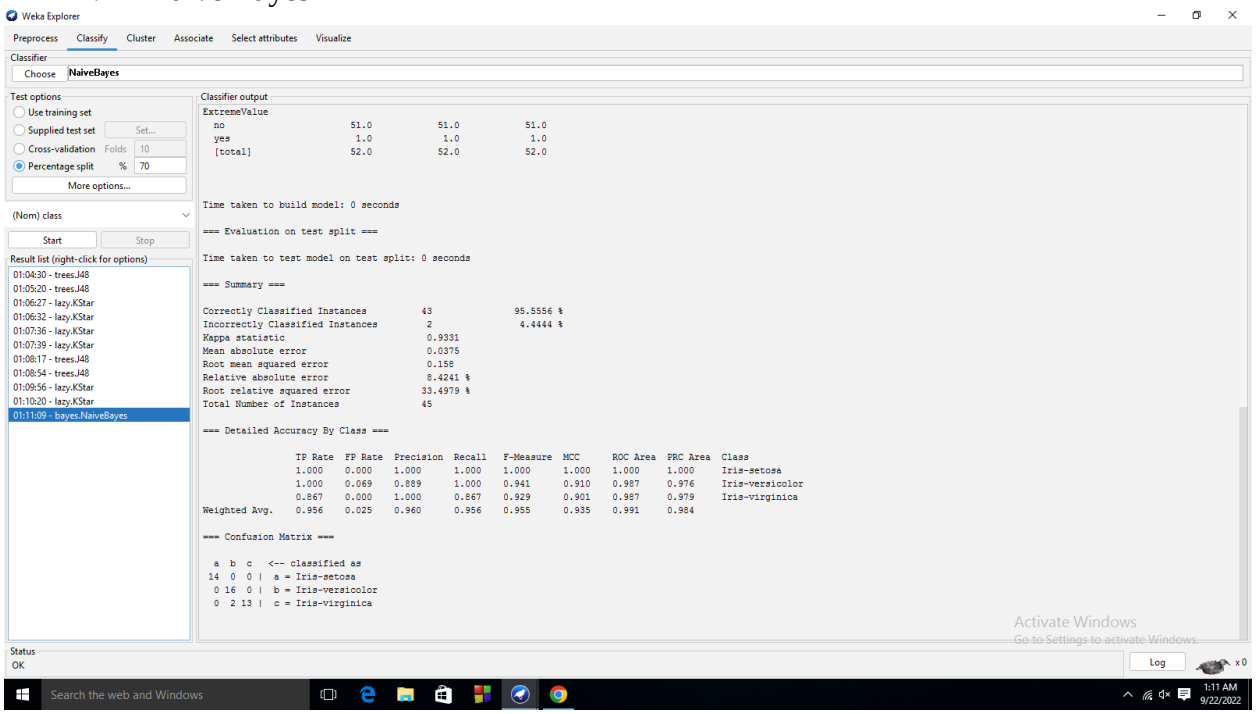
This is the preprocessing window of IQR with sepal length as chosen attribute whose outlier histogram is skewed-right with mode 54 and it has minimum value of 4.3 and maximum value as 7.9 .



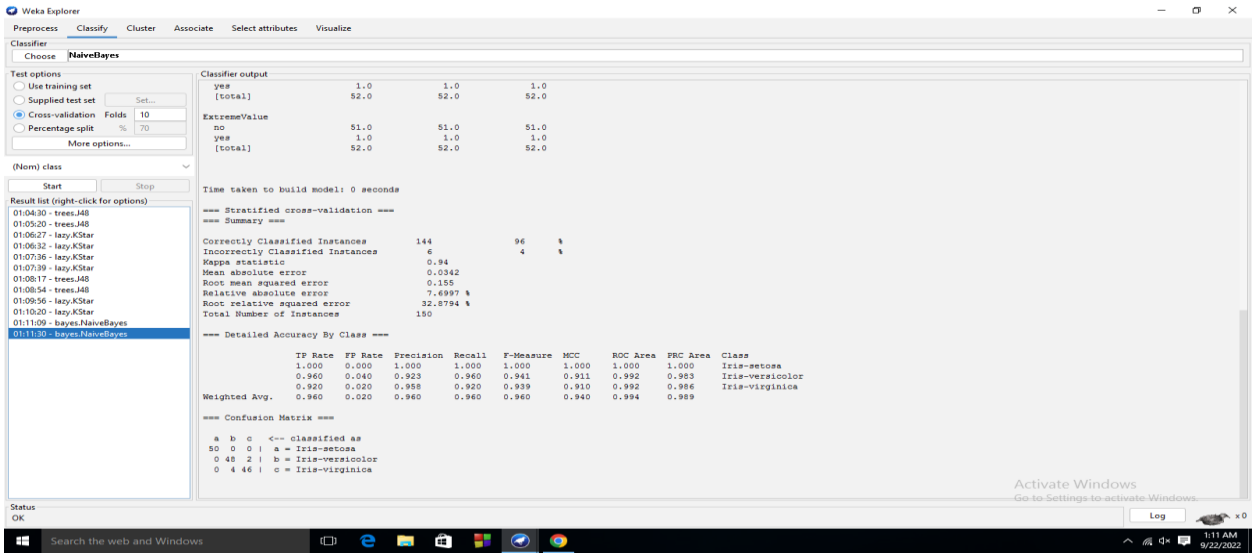
This is IQR outlier and extreme histogram visualization of all attributes with sepal length and sepal width being unimodal and petal length and petal width being multi modal.

2. Classification:

i. Naïve Bayes

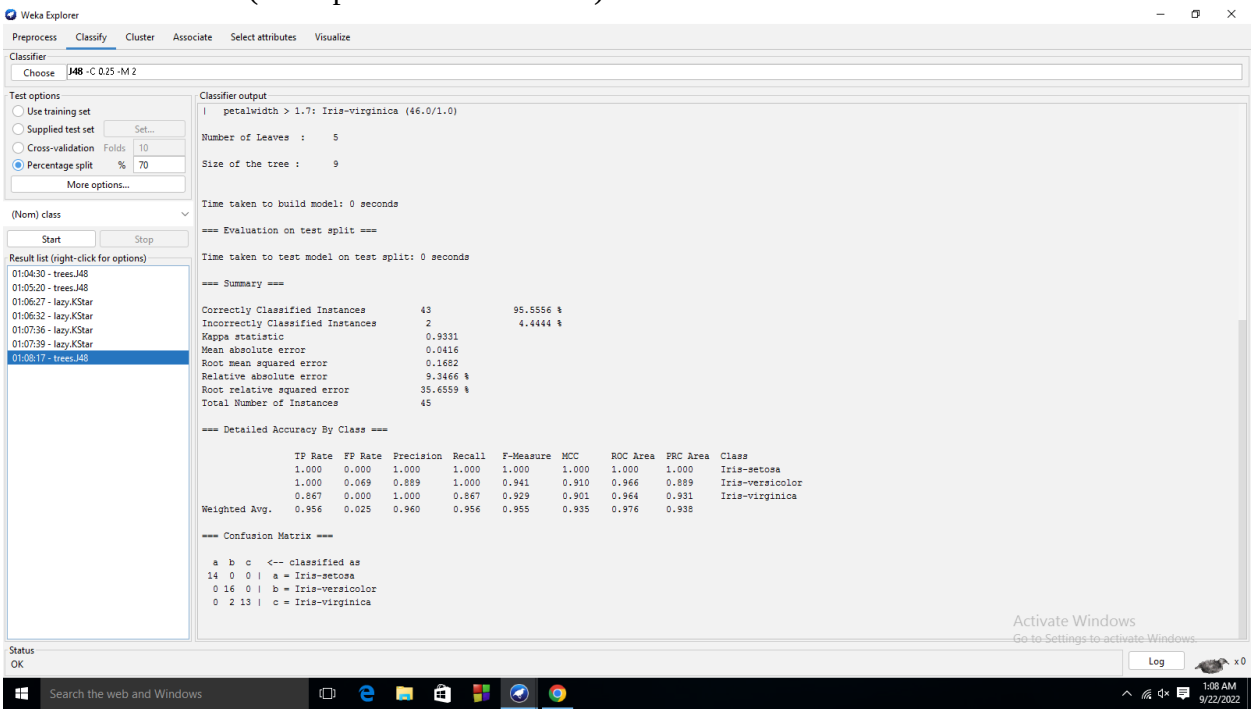


This is Naïve Bayes percentage split of 70% where correctly classified instances was 95.556 With mean absolute error as 0.0375

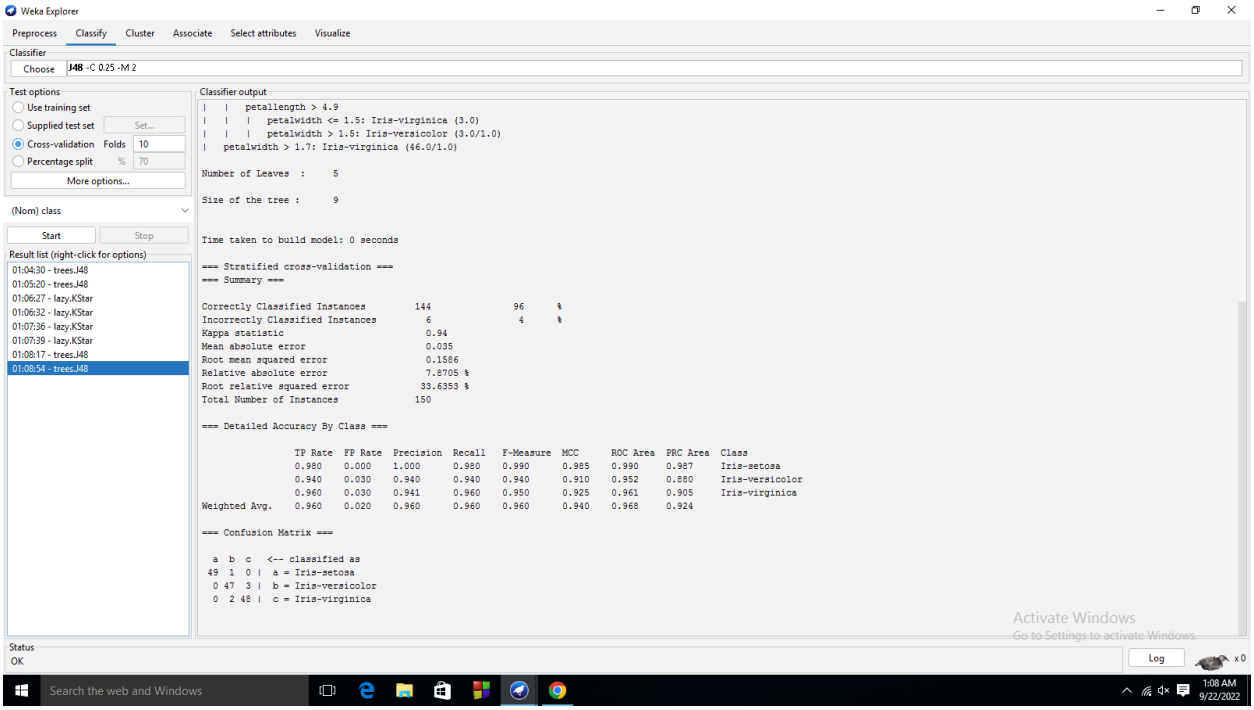


This is naïve Bayes cross validation with 10 folds in which correctly classified instances has increased to 96% with mean absolute error reduced to 0.342.

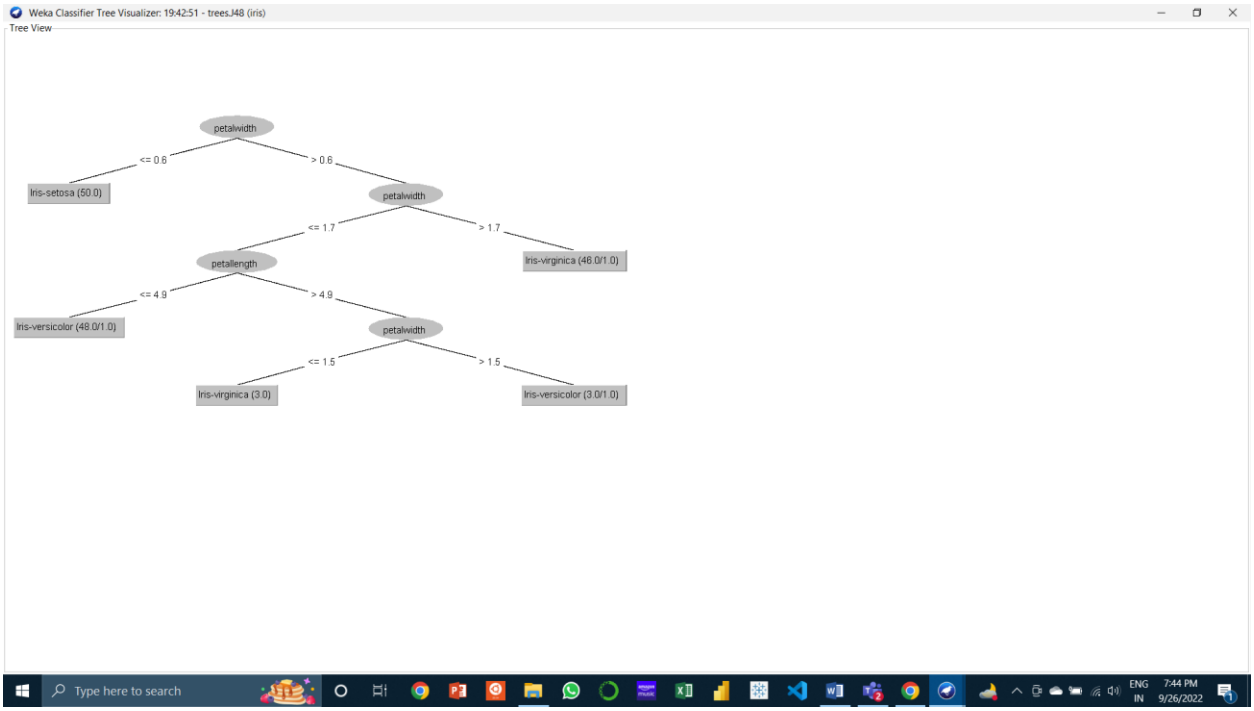
ii. J48 (Example of Decision Tree):



This is J48 performance split of 70/30 with correctly classified instances of 95.556 and mean absolute error of 0.0416

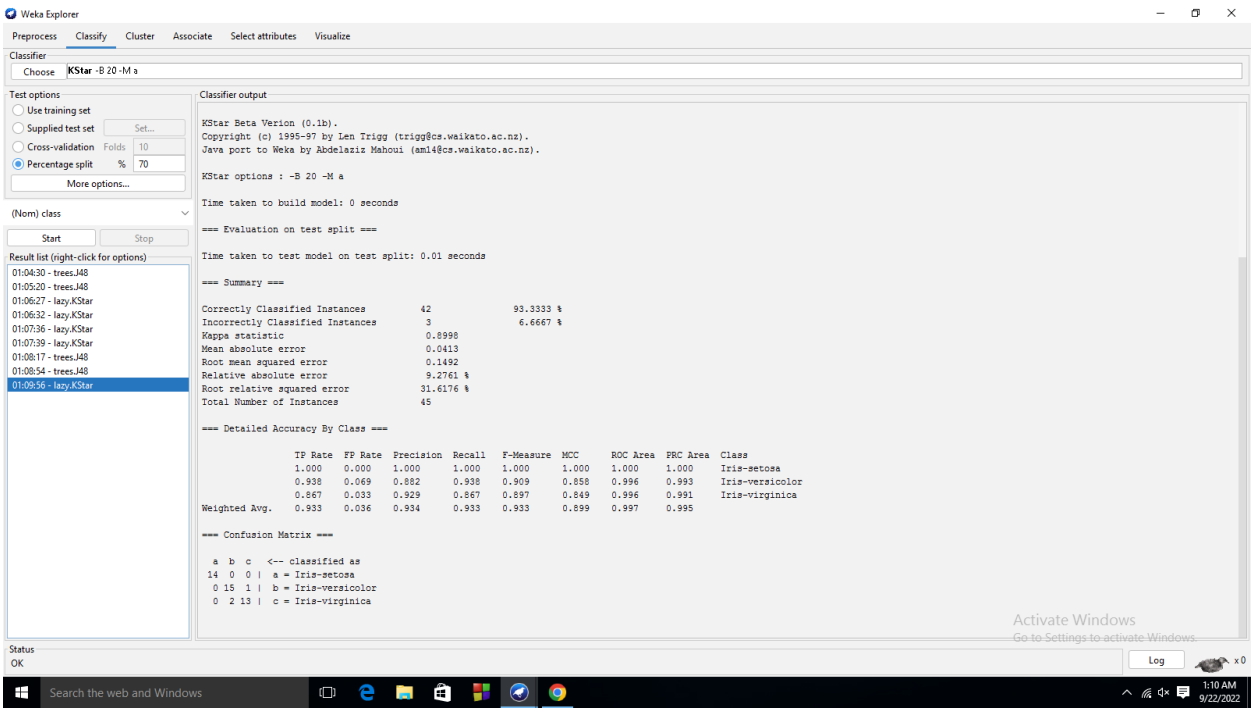


This is J48 cross validation with 10 folds and correctly classified instances at 96% and mean absolute error as 0.035

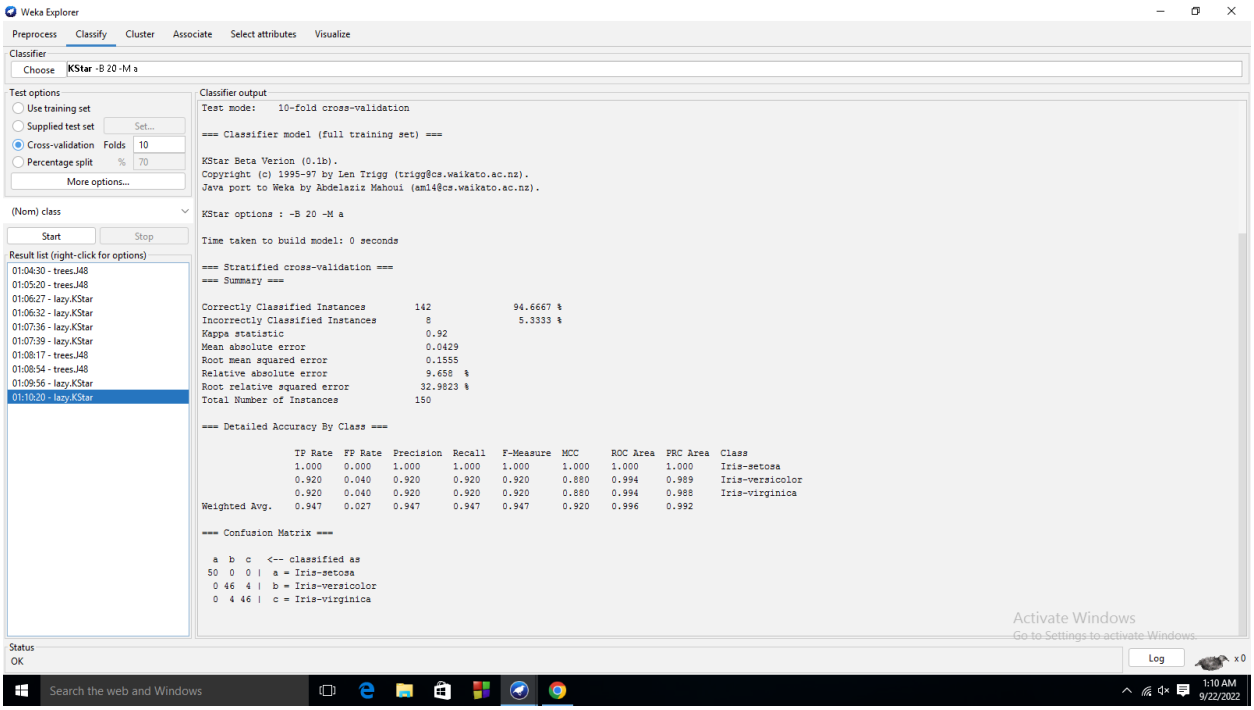


This is the visualization of J48 tree

iii. K-star:



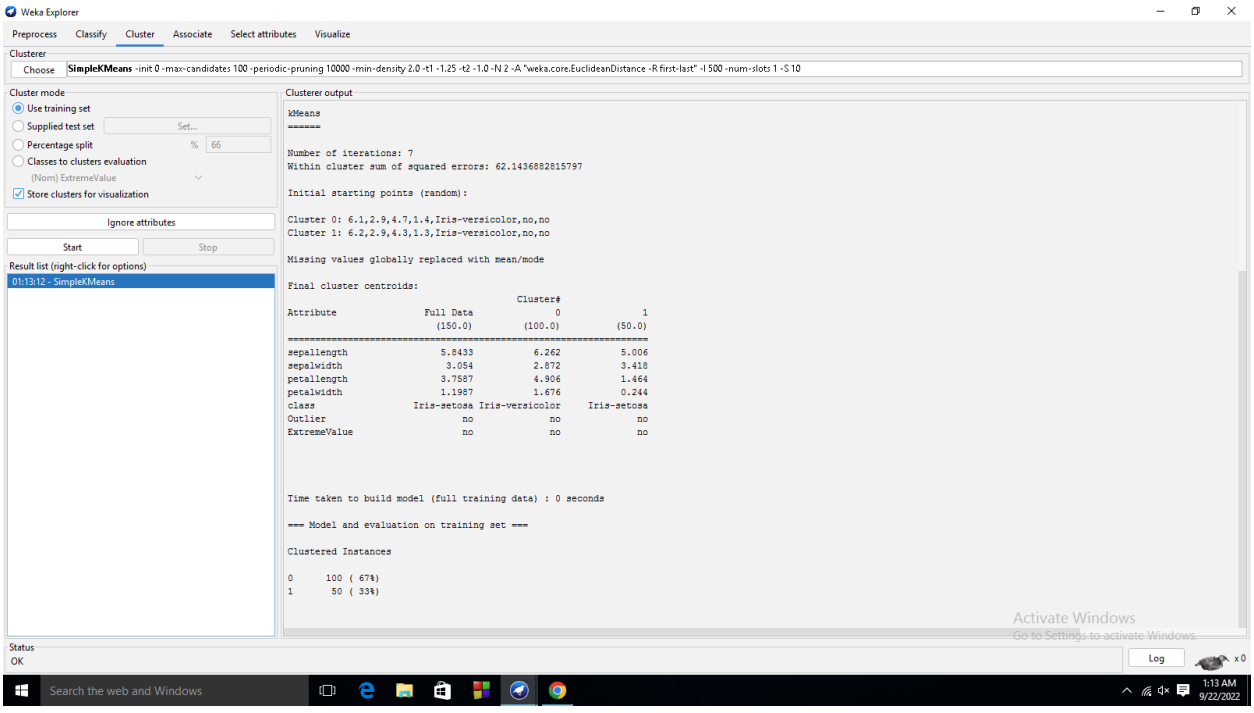
This is K-Star percentage split 70/30 with correctly classified instances at 93.33% with mean absolute error being 0.0413



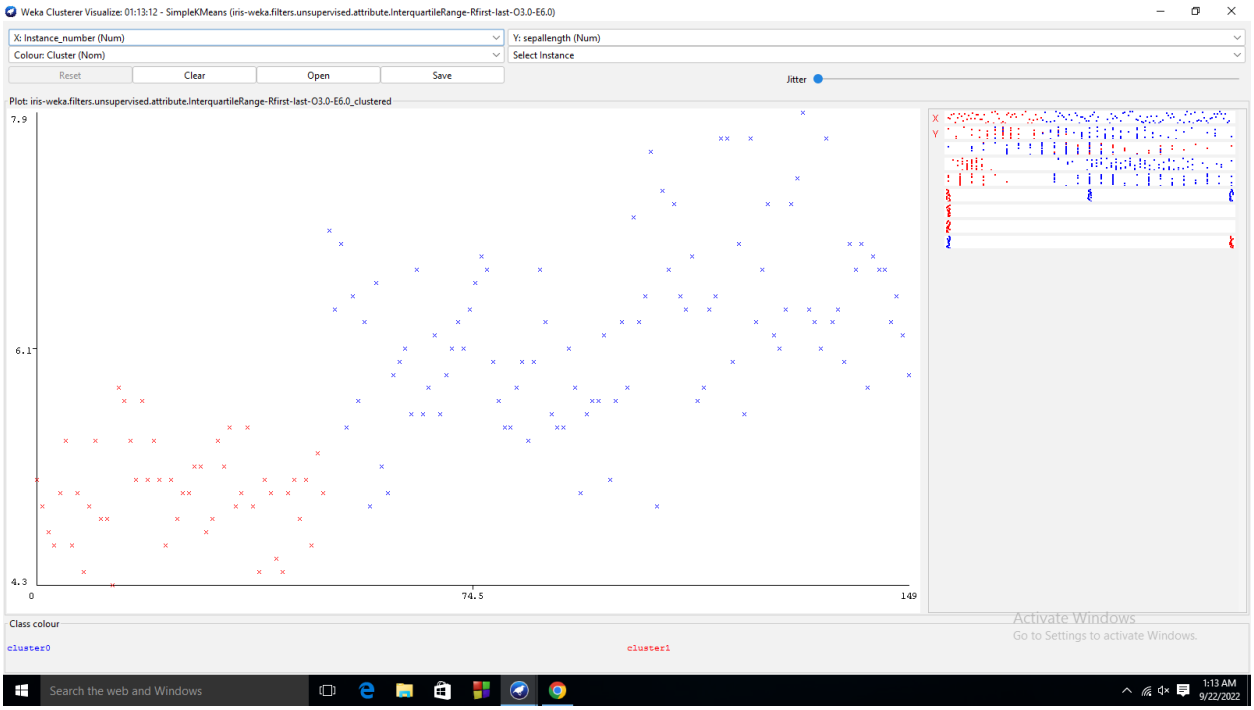
This is the K-Star cross-validation with 10 folds having correctly classified instances at 94.67% and mean absolute error being 0.0429

3. Clustering:

i. K-Means

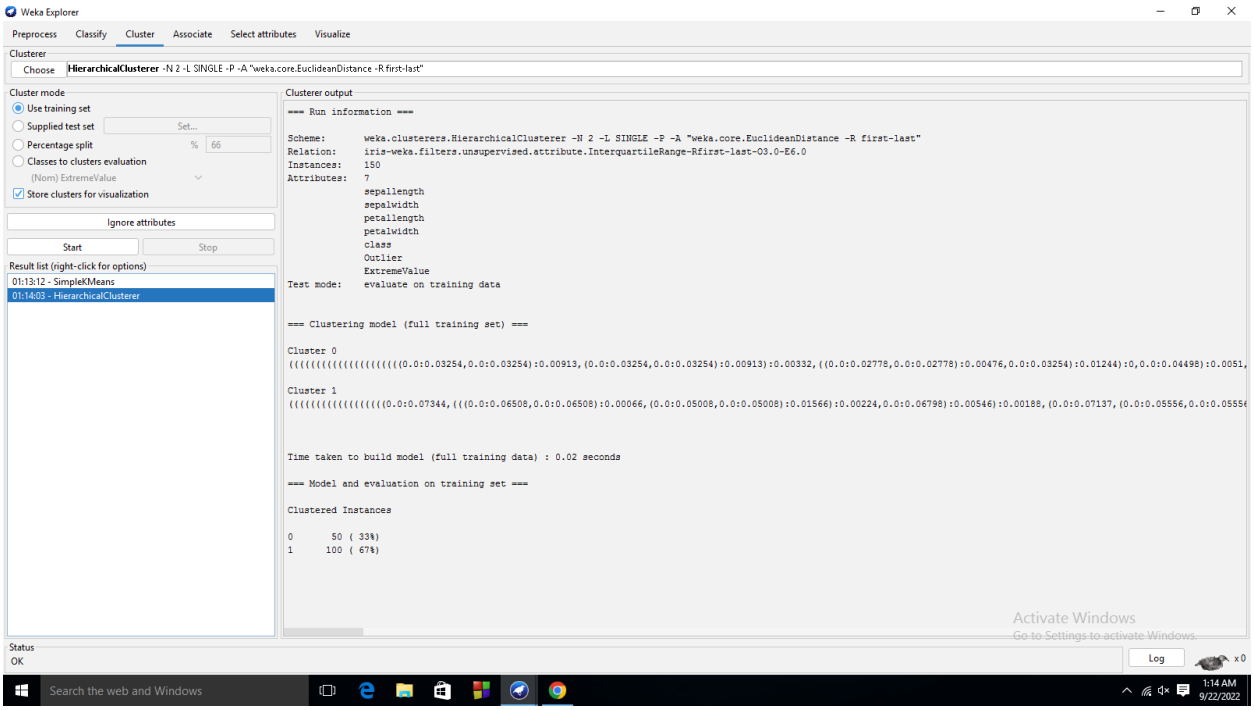


This is output of the model of K-Means with cluster 0 having 67% and cluster 1 having 33%

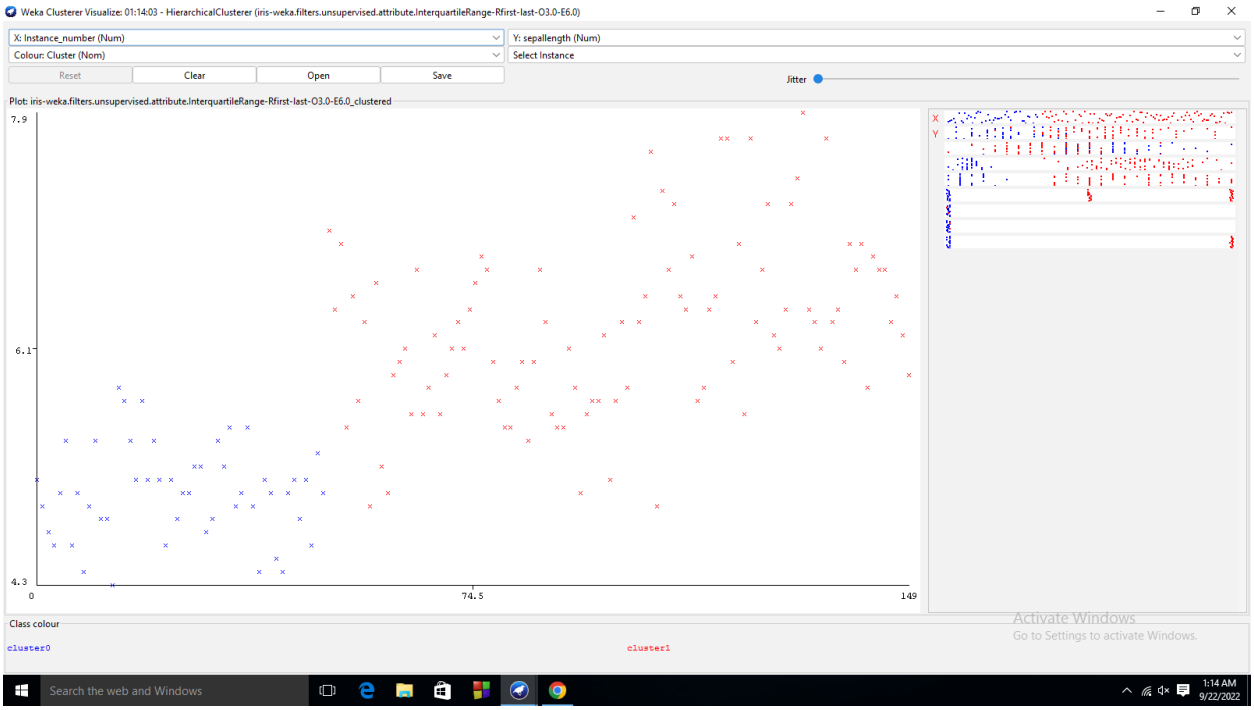


This is the scatter plot for K-Means cluster assignment where we can conclude that there are two clusters one at the beginning and second in the middle also there is a upward trend.

ii. Hierarichal Cluster:



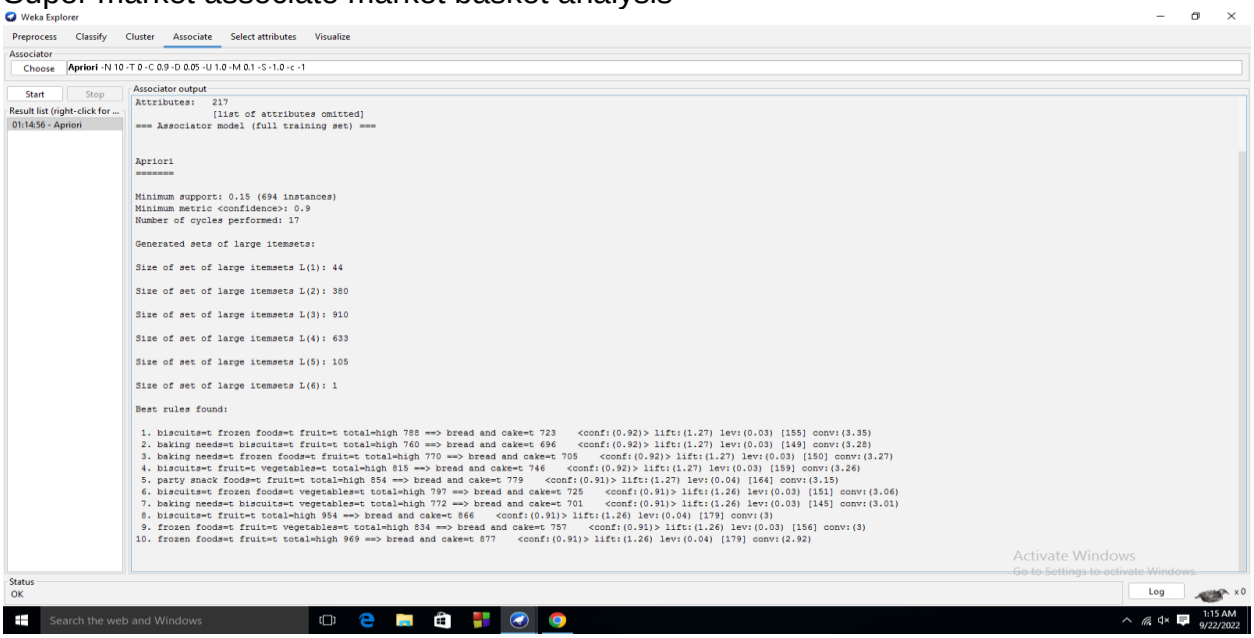
The percentage split of cluster is same as above.



This is scatter plot of hierarichal cluster assignment . According to me, data is evenly distributed with upward trend and two cluster forming one at the start and other at middle.

4. Association Rule:

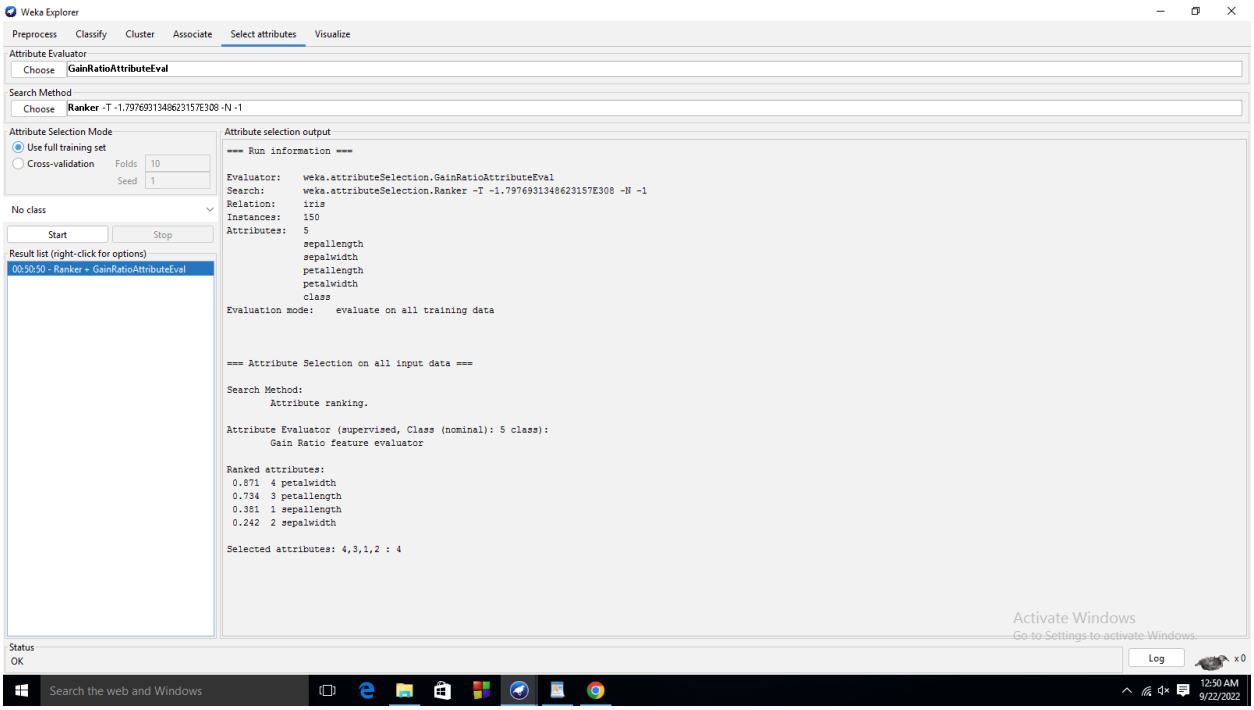
Super market associate market basket analysis



Confidence level of top 10 rules is either 0.92 or 0.91 .We found that there is high relation between biscuits,frozen foods and fruits and many more.

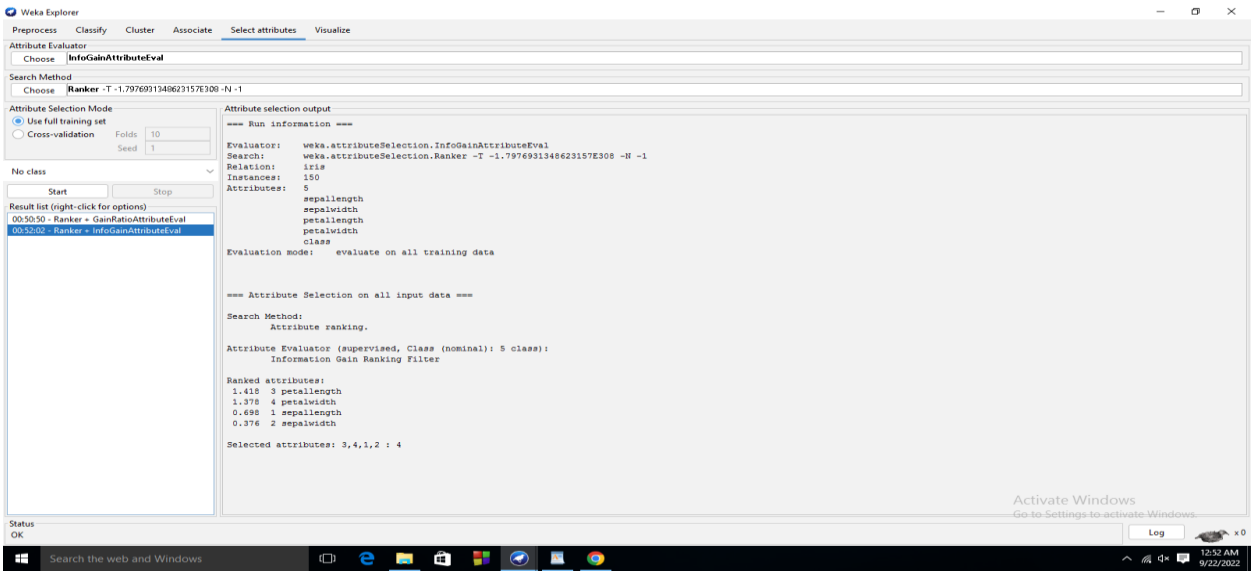
5. Select Attributes:

i. Gain Ratio:



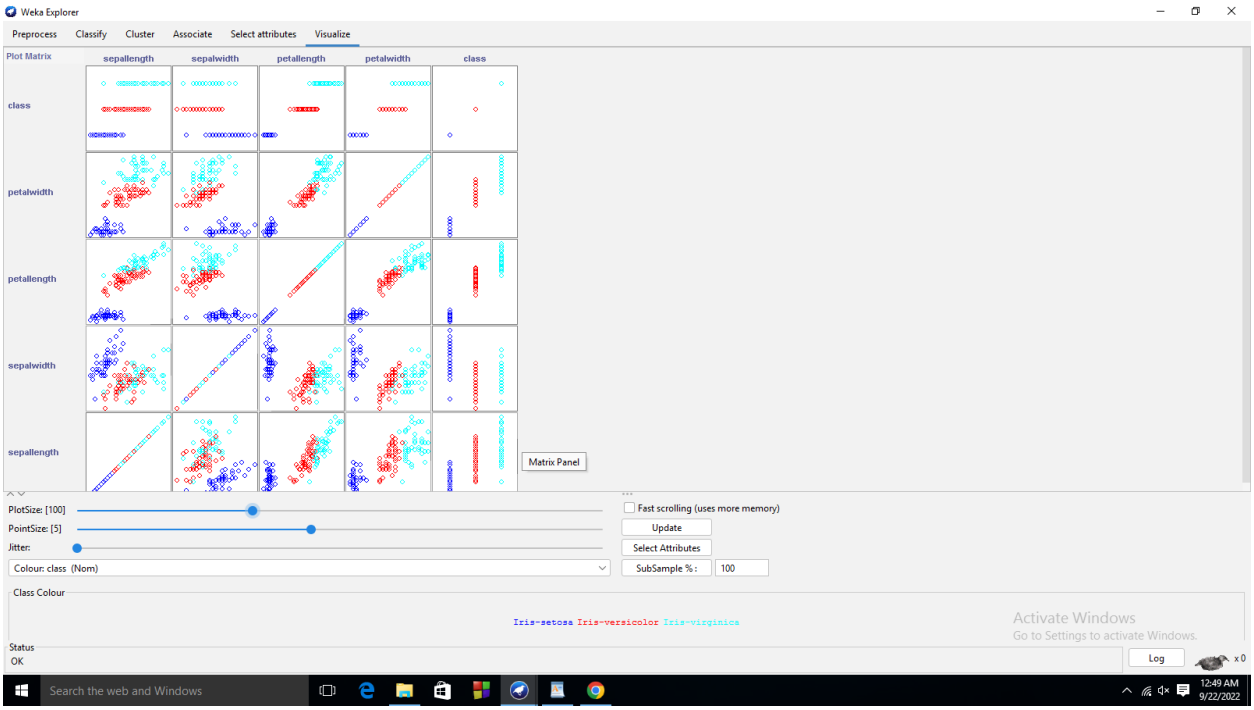
Petal width has the highest rank with gain ratio 0.871 and sepal width has the lowest rank with gain ratio 0.242

ii. Info Gain Ratio



Petal length has the highest info gain with 1.418 and sepal width with least of 0.376

6. Visualization:



Petal width and Petal length are related by upward trendline
For lower value of petal length we find values of sepal width to be more and we find it to be downward line.
Sepal length and Sepal Width are closely related with all the clusters in close proximity.

Conclusion:

J48 ,Naïve Bayes had highest correct classified instances with 96%.

Confidence level of top rules was around 0.92