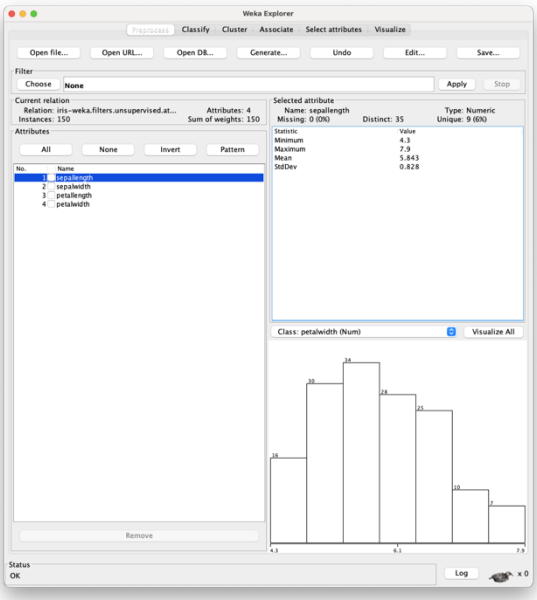
LAB EXERCISE

CLUSTERING WITH WEKA EXPLORER

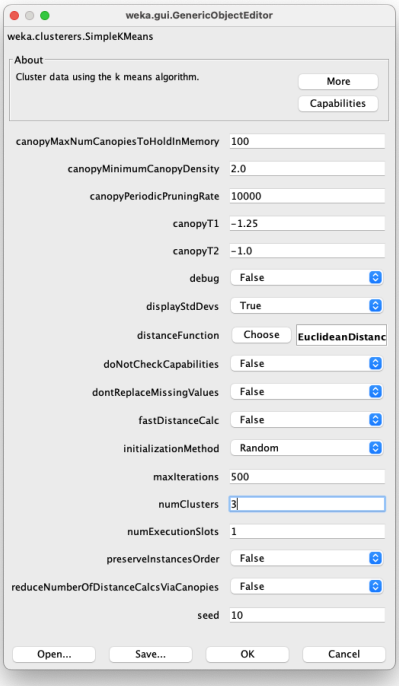
TASK 1

For this exercise, you will use WEKA's SimpleKMeans unsupervised clustering algorithm and Hierarchical Agglomerative Clustering.

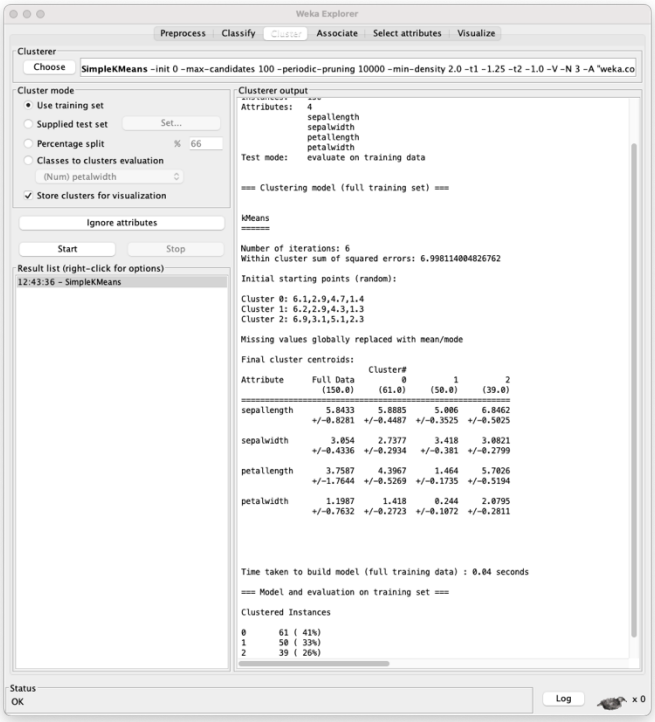
i. Open the WEKA Explorer and load the numerical form of the dataset, such as *weather.numerical.arff or iris.arff* dataset. For this exercise use *iris.arff.*
ii. Remove the Class attribute as you do not want the value of this attribute to affect the clustering. (Alternatively use weka.gui.GenericObjectEditor to set the attributes)
iii. Click on Cluster and choose the SimpleKMeans algorithm.
iv. Set displayStdDevs to True. This will give us the domain standard deviation of each attribute as well as the within-class attribute standard deviations.
v. We know there should be three distinct clusters, set numClusters to 3.
vi. Click on Start to begin the data mining session. The output should include attribute mean and standard deviation values for each cluster as well as the total number of instances assigned to each cluster.
vii. Next, you must decide if the clusters are interesting. Try to increase the cluster numbers (k) and visualize the assignment.
viii. Repeat the exercise using Hierarchical Agglomerative Clustering algorithm. Try all combinations of distance function (Euclidean or Manhattan distance) and linkage type (single, complete, average) seen in lectures.



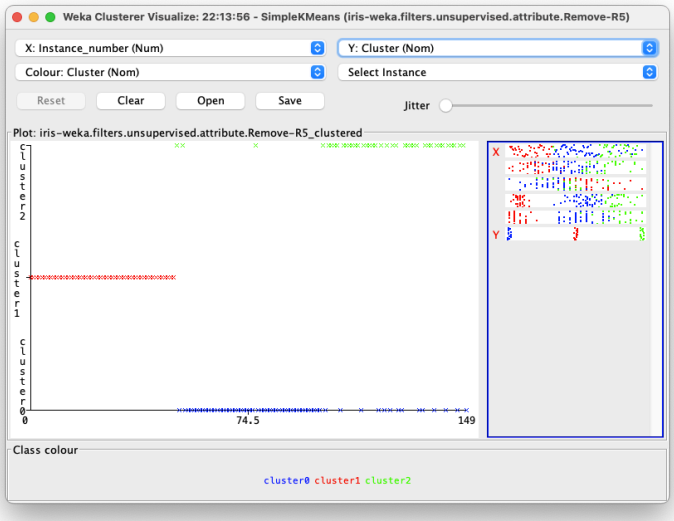
Firstly, we need to remove the class label from the dataset.



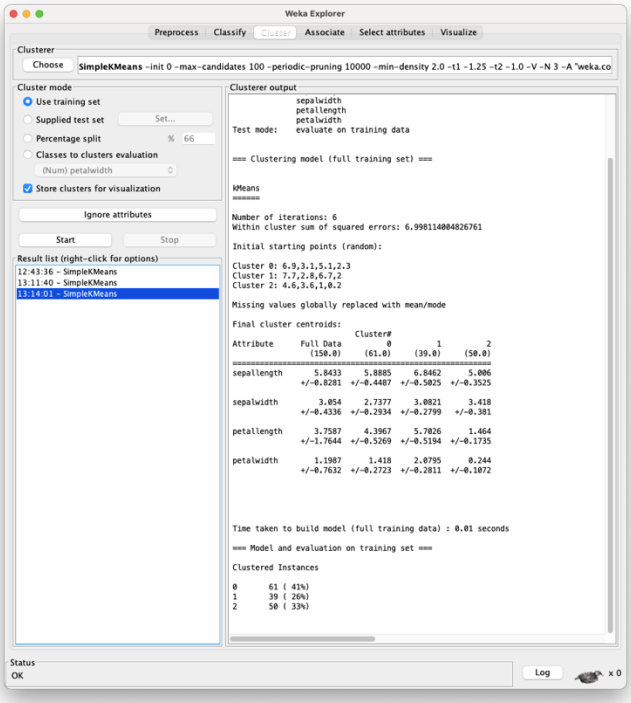
Then, we need to configure the k-means to set the number of cluster as 3 and set the displayStdDev to true.



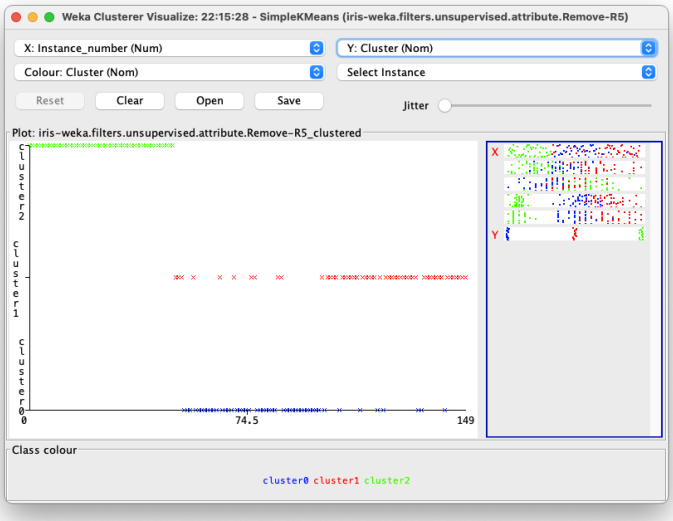
Later, we can run the k-mean algorithm. The result show, the number of instance = 6, the number of instances for each cluster, together with the mean, and standard deviation.



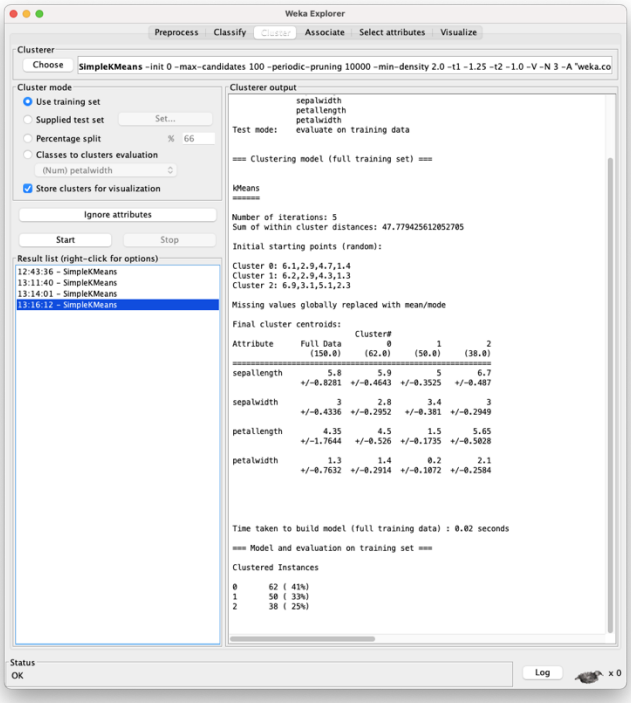
We can visualize the result of the clustering.



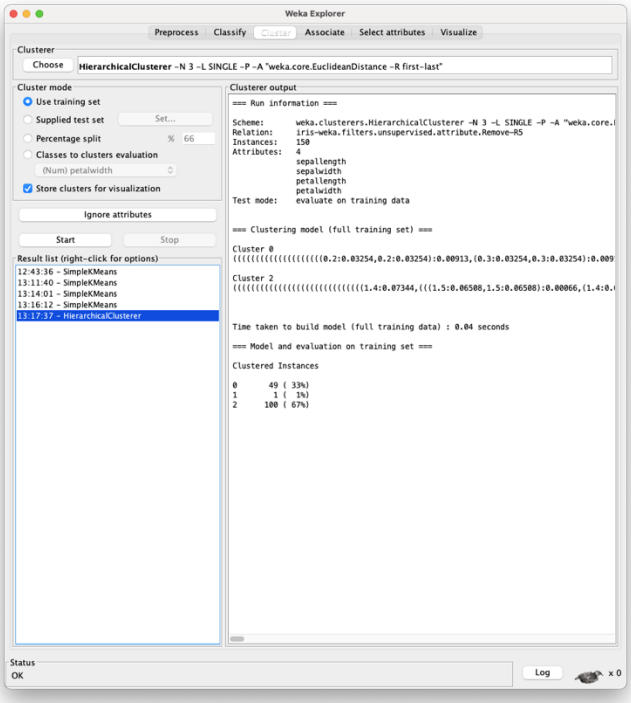
We can try to experiment to by increasing the number of seeds in order to view changes of the result.



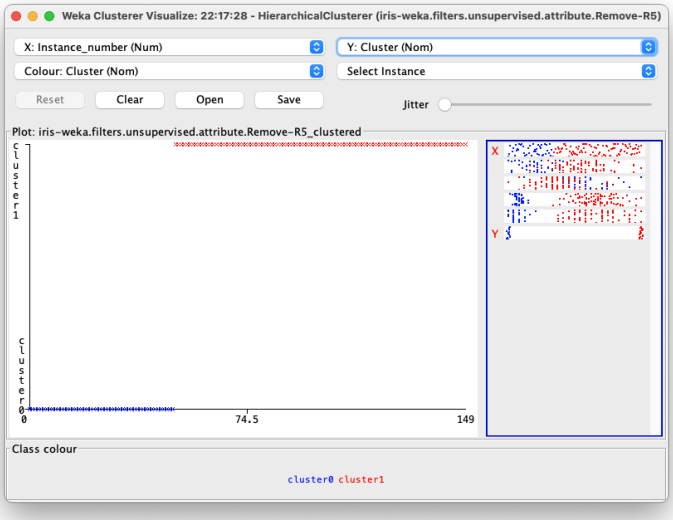
We can also visualize the result. As we can see, the cluster is different from the previous result.



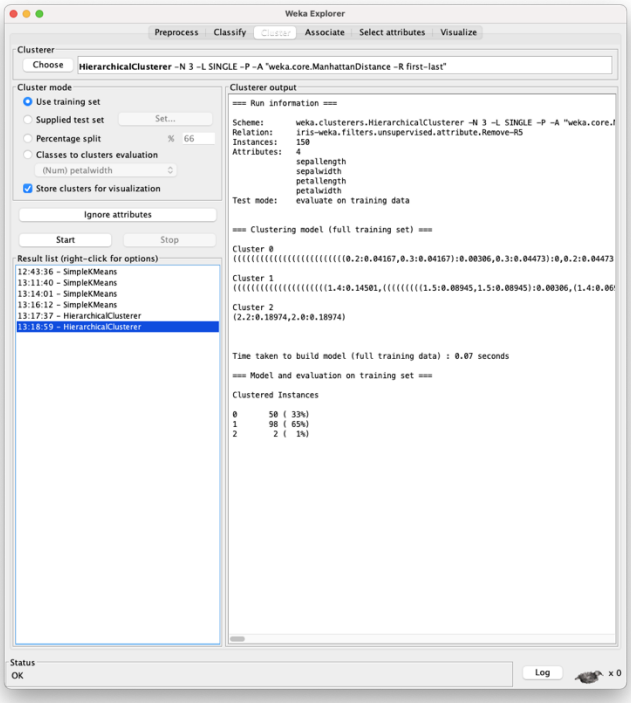
We can also change the distance function to ManhattanDistance.



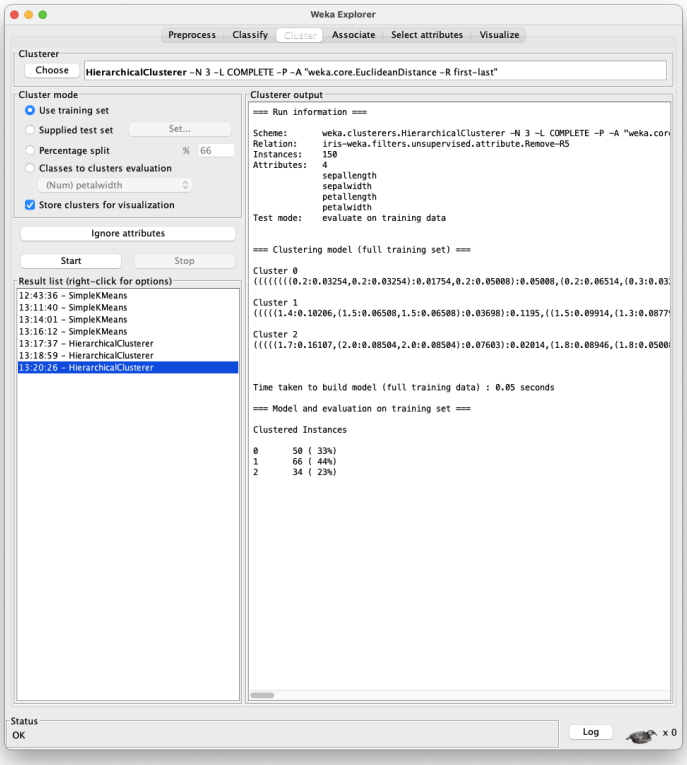
Next, we can also try using Hierarchical Agglomerative Clustering algorithm.



The visualization result for the Hierarchical Cluster.



We can also change the linkage to other type as such average and complete.



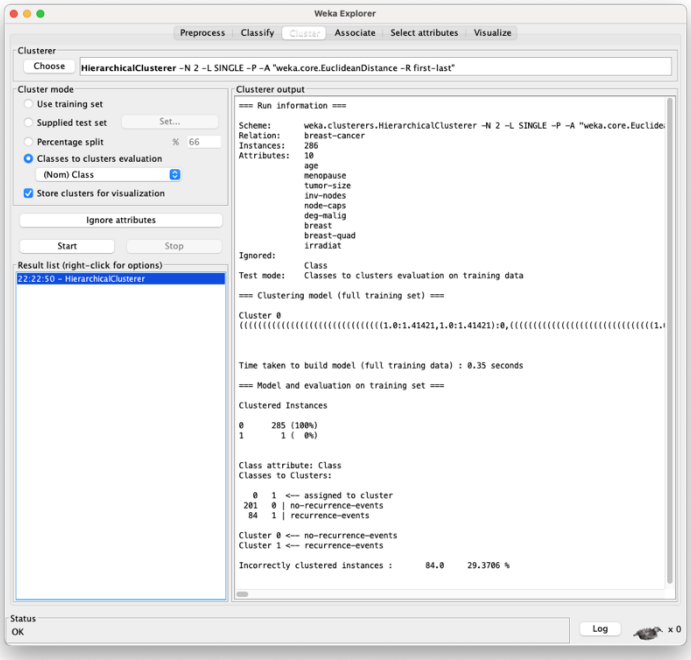
Hierarchical Agglomerative Clustering algorithm using complete link type.

TASK 2

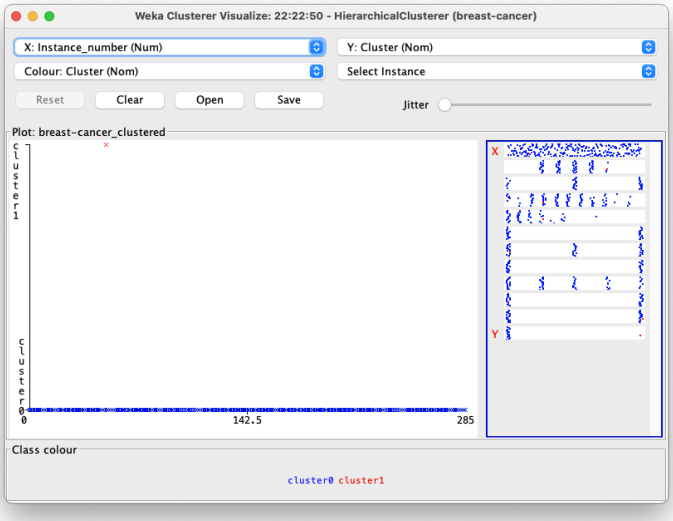
Use the breast cancer data (breast-cancer.arff) provided with Weka or available from UCI. We will use the "class" attribute to see if we can find clusters that align with this class (cluster mode is "classes to clusters evaluation"). The clustering algorithm can be used to perform supervised training also.

i. Use Hierarchical Agglomerative Clustering to cluster the data into two clusters. Try all combinations of distance function (Euclidean or Manhattan distance) and linkage type (single, complete, average). Which methods perform best for accuracy of predicting the class value (whether or not there is recurrence)?

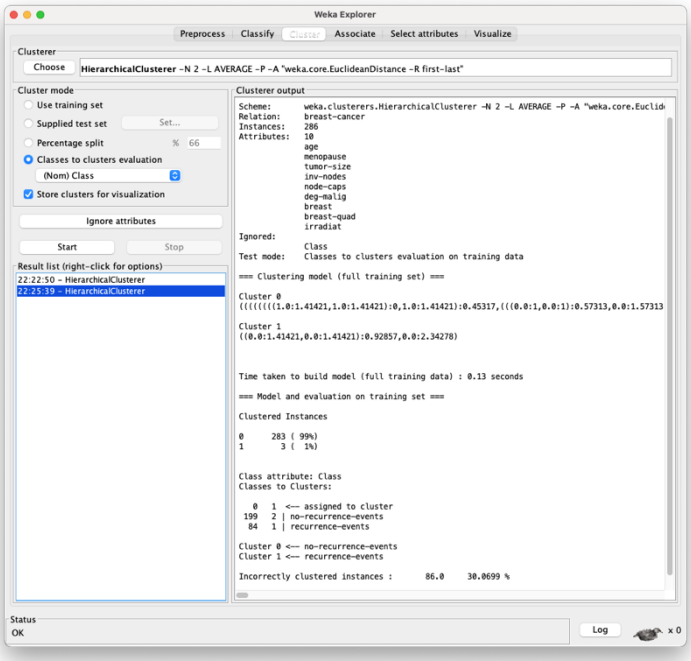
ii. Use SimpleKMeans algorithm and check the accuracy.



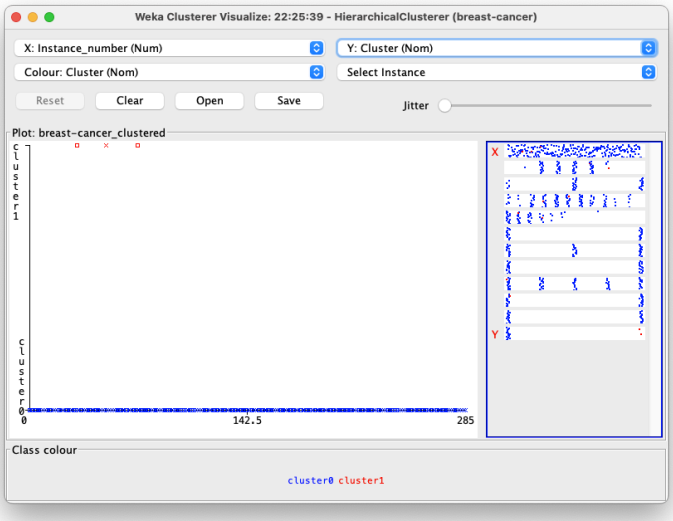
Hierarchical Agglomerative Clustering using Euclidean distance, link type Single.



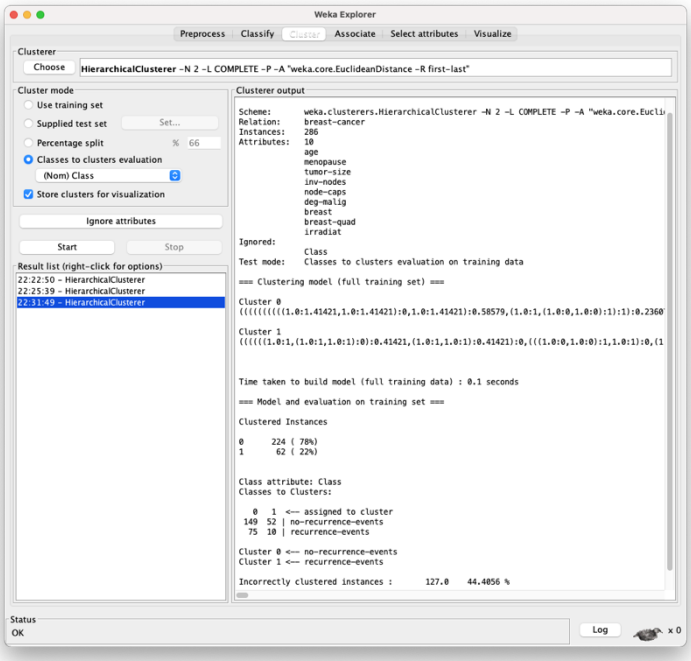
Visualization of the result



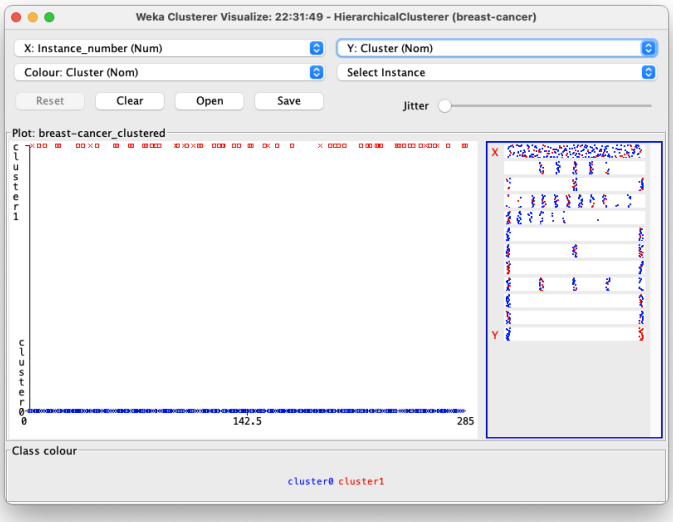
Hierarchical Agglomerative Clustering using Euclidean distance, link type Average.



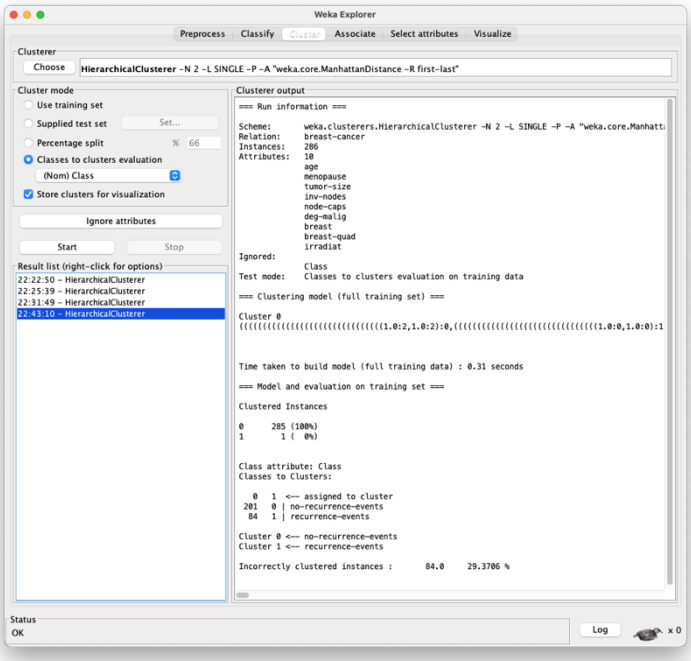
Visualization of the result. Try tanya Dr, kenapa dekat cluster to dia ada square dengan X. Ada maksud ke?



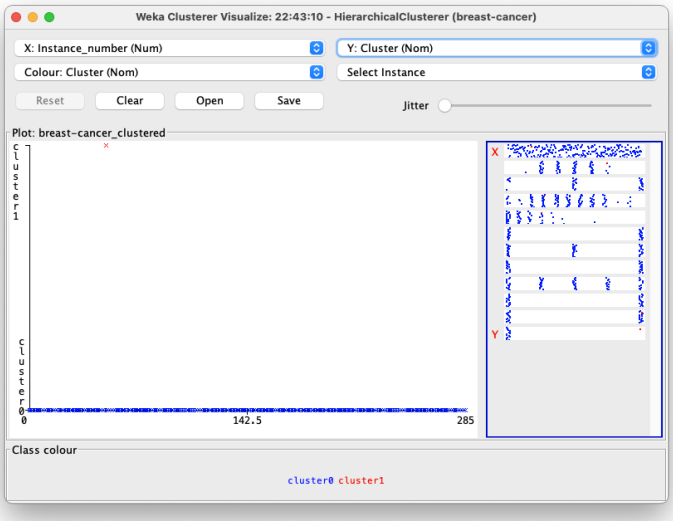
Hierarchical Agglomerative Clustering using Euclidean distance, link type Complete.



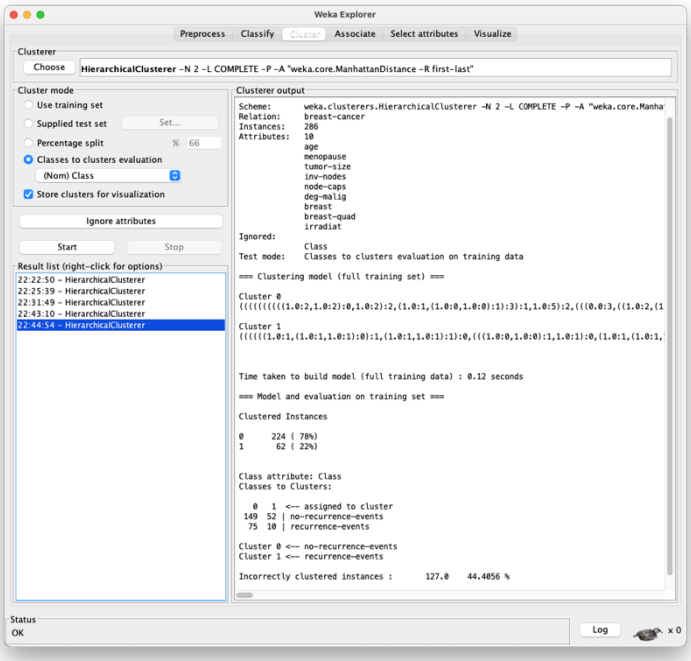
Visualization of the result



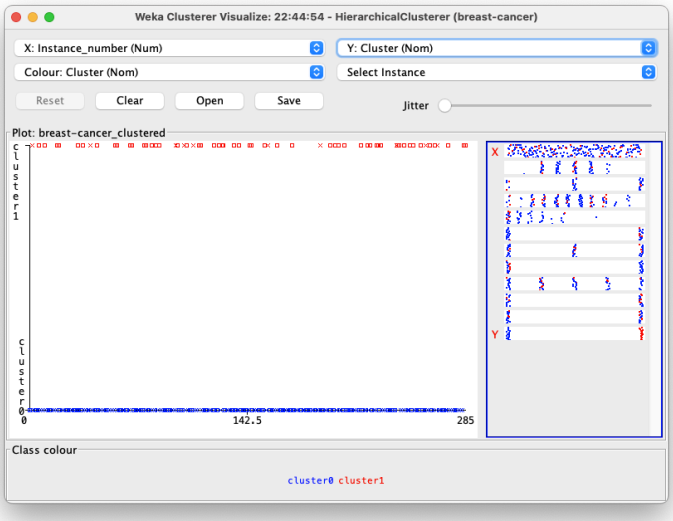
Hierarchical Agglomerative Clustering using Manhattan distance, link type Single.



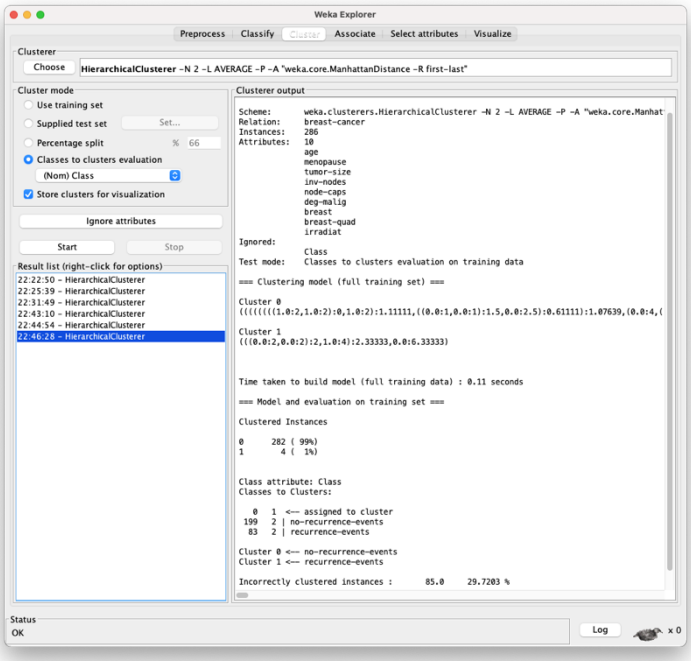
Visualization of the result



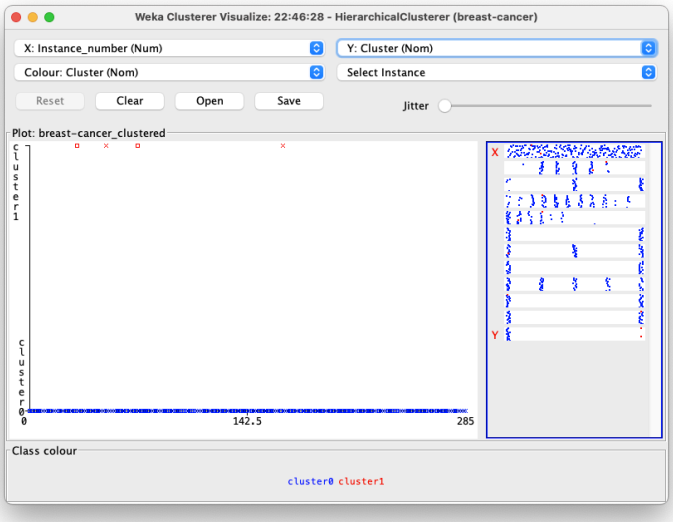
Hierarchical Agglomerative Clustering using Manhattan distance, link type Complete.



Visualise the result

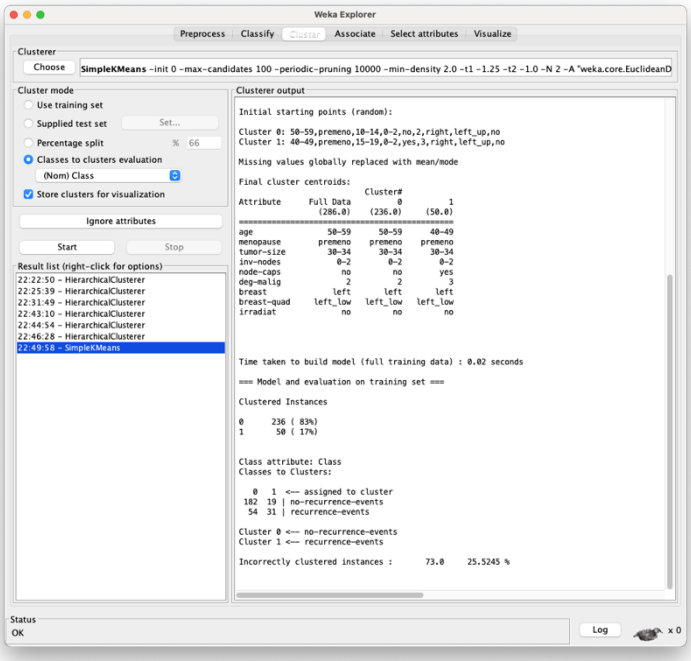


Hierarchical Agglomerative Clustering using Manhattan distance, link type Average.

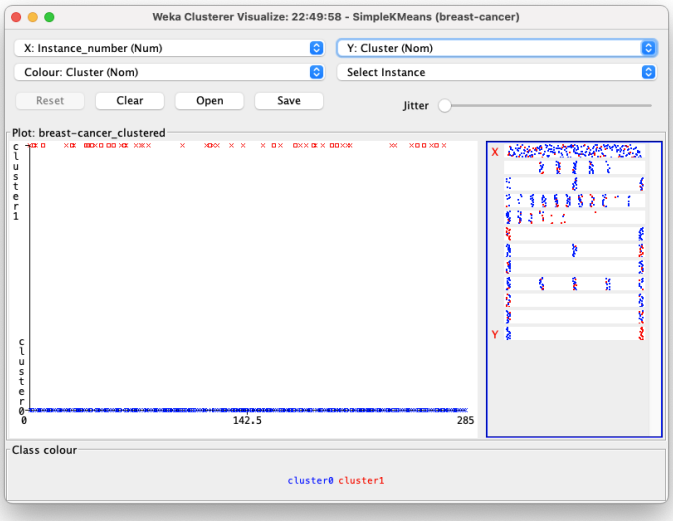


Visualize the result

Based on the clustering using Hierachical Clusterer, it is shown that Hierarchical Clusterer using Euclidean distance and single linktype has the highest accuracy, 29.3706%



Simple kmeans clustering



Visualize the result

It is shown that Simple Kmeans perform better compared to Hierarchical Clustering

TASK 3

Repeat the task using other algorithm (NB, ANN or kNN) and compare the accuracy of models.

Other references:

- ✓ https://www.ibm.com/developerworks/library/os-weka2/
- ✓ https://www.youtube.com/watch?v=zHbxbb2ye3E
- ✓ https://www.youtube.com/watch?v=QXOkPvFM6NU
- ✓ https://www.youtube.com/watch?v=HCA0Z9kL7Hg (WEKA)
- ✓ https://www.youtube.com/watch?v=9aODdNSAaul&list=PLm4W7_iX_v4OMSgc8xowC2h70s-unJKCp&index=19 (WEKA)