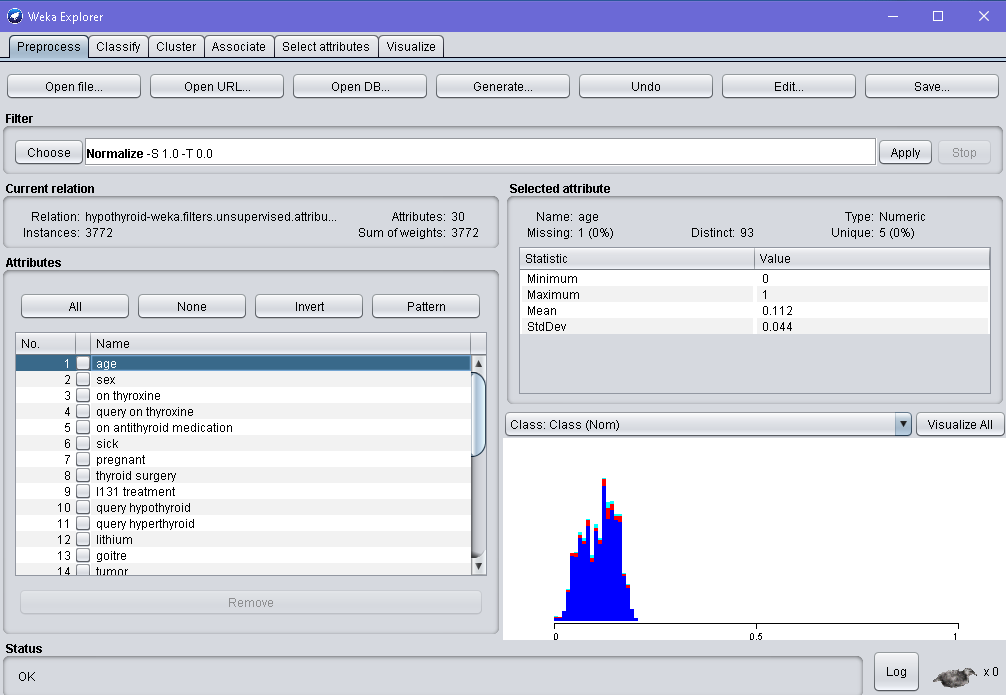
PROJECT 3

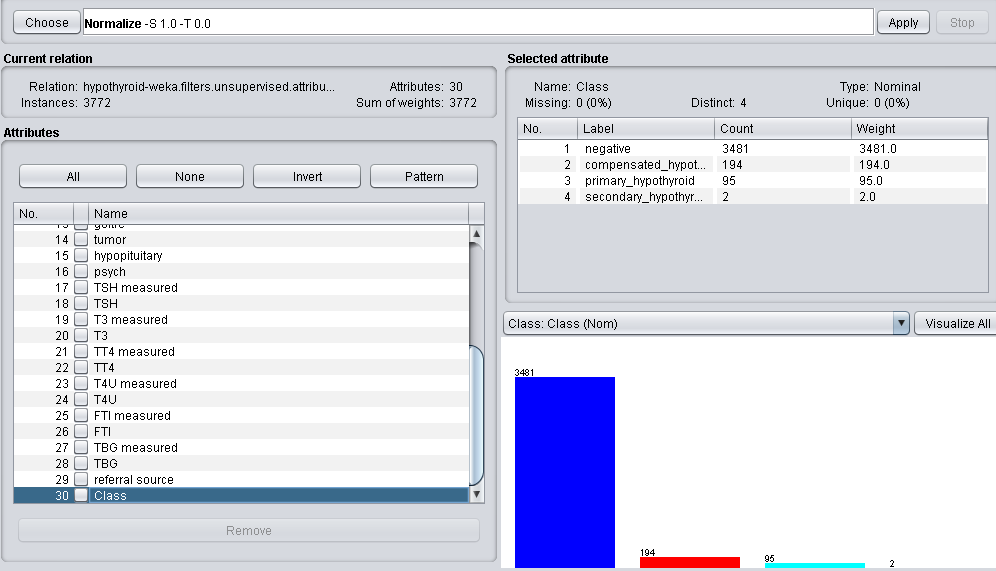
Name: Sidra

Dataset: Hypothyroid- most common disease found in humans. This data set has 30 attributes and 3772 attributes with both nominal and numeric values.

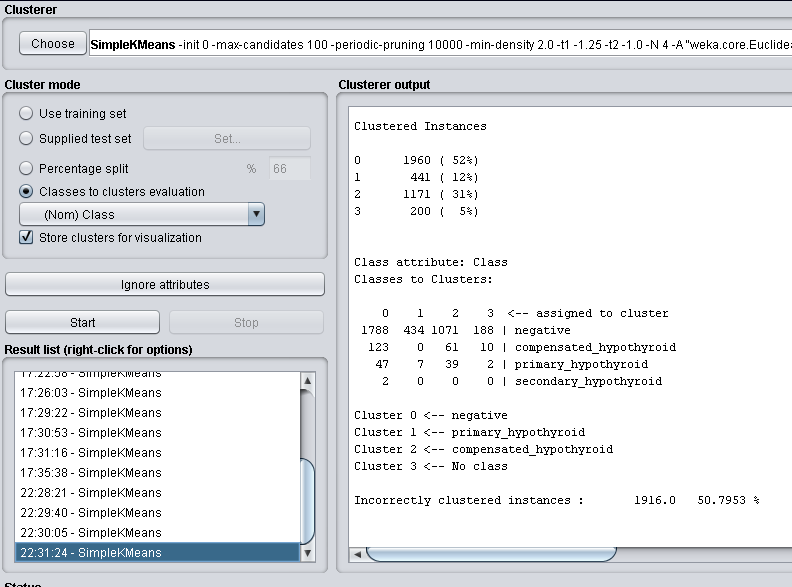
Purpose: To find patterns by different algorithms and to improve the error rate so that one could identify the relation of attributes with thyroid i.e. dependent attributes.

1-First we are normalizing values since we have numeric data values.



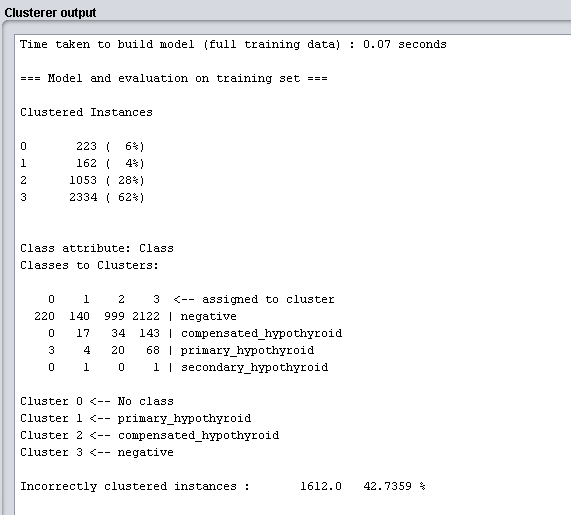


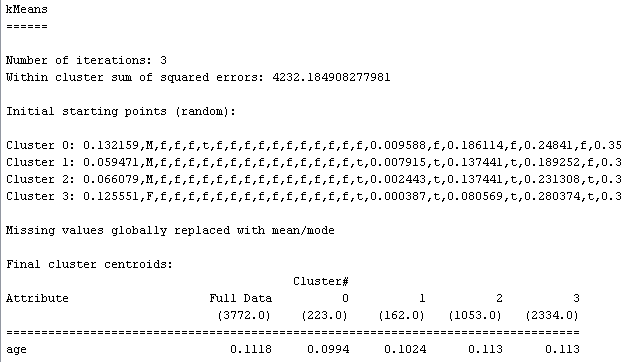
2- Choosing simpleKmeans algorithm while setting k=4, seed=11 and choosing classes to cluster evaluation. I Ignored class label from this data set.



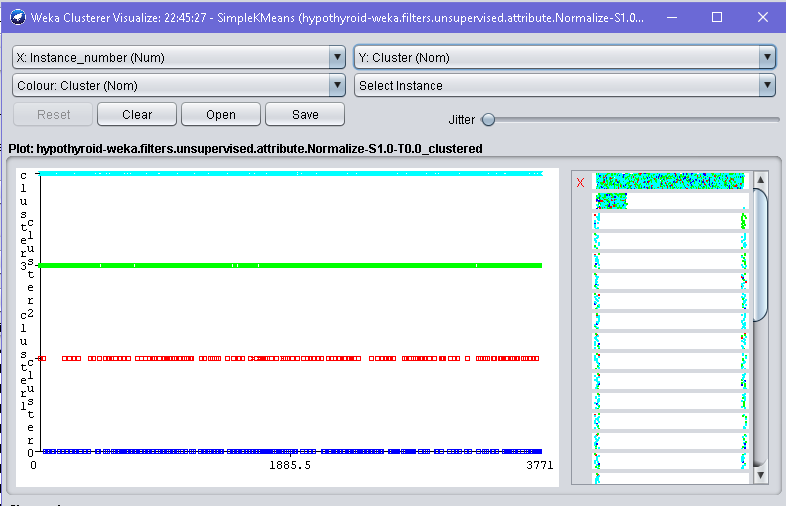
3-increasing seed to get reduced incorrect instances and total squared error:

K=4, seed=20

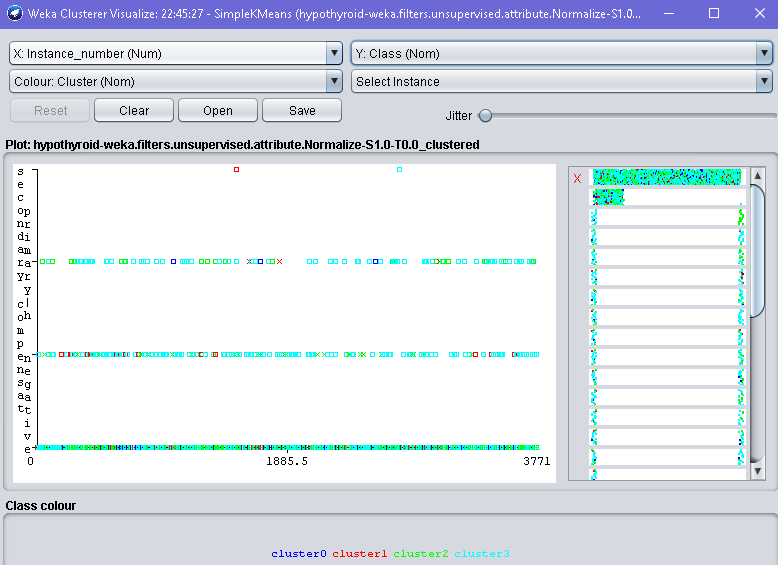




4-Visualizing the cluster by right clicking and then I can adjust it to my needs by selecting any combination of two attributes,

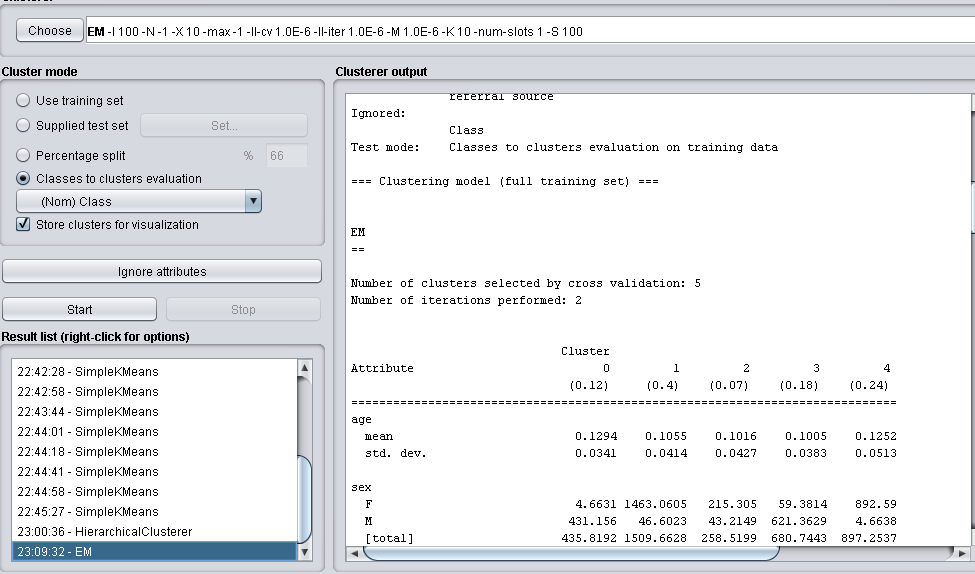


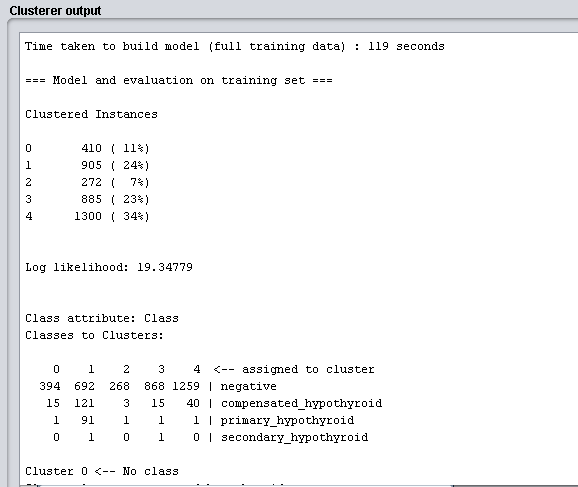
And this was the visualization by selecting instance\_number and class

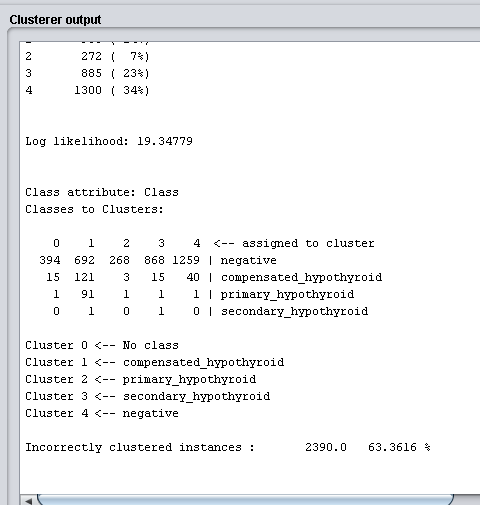


5-Now trying to implement EM clustering, where seed=100 by default and the algo will decide the number of clusters, since it is equal to -1. We’ll see that how algo works on the same data set.

By EM clustering we get the probability for each attribute, but it has more error than our kmeans algo.







The manually handled cluster i.e. Kmeans showed great patterns in our case.