CS 422

Data Mining

Homework Assignment 3

Practicum Questions 1 and 2

Ronit Rudra
A20379221
rrudra@hawk.iit.edu

Course Instructor
Prof. Jawahar Panchal

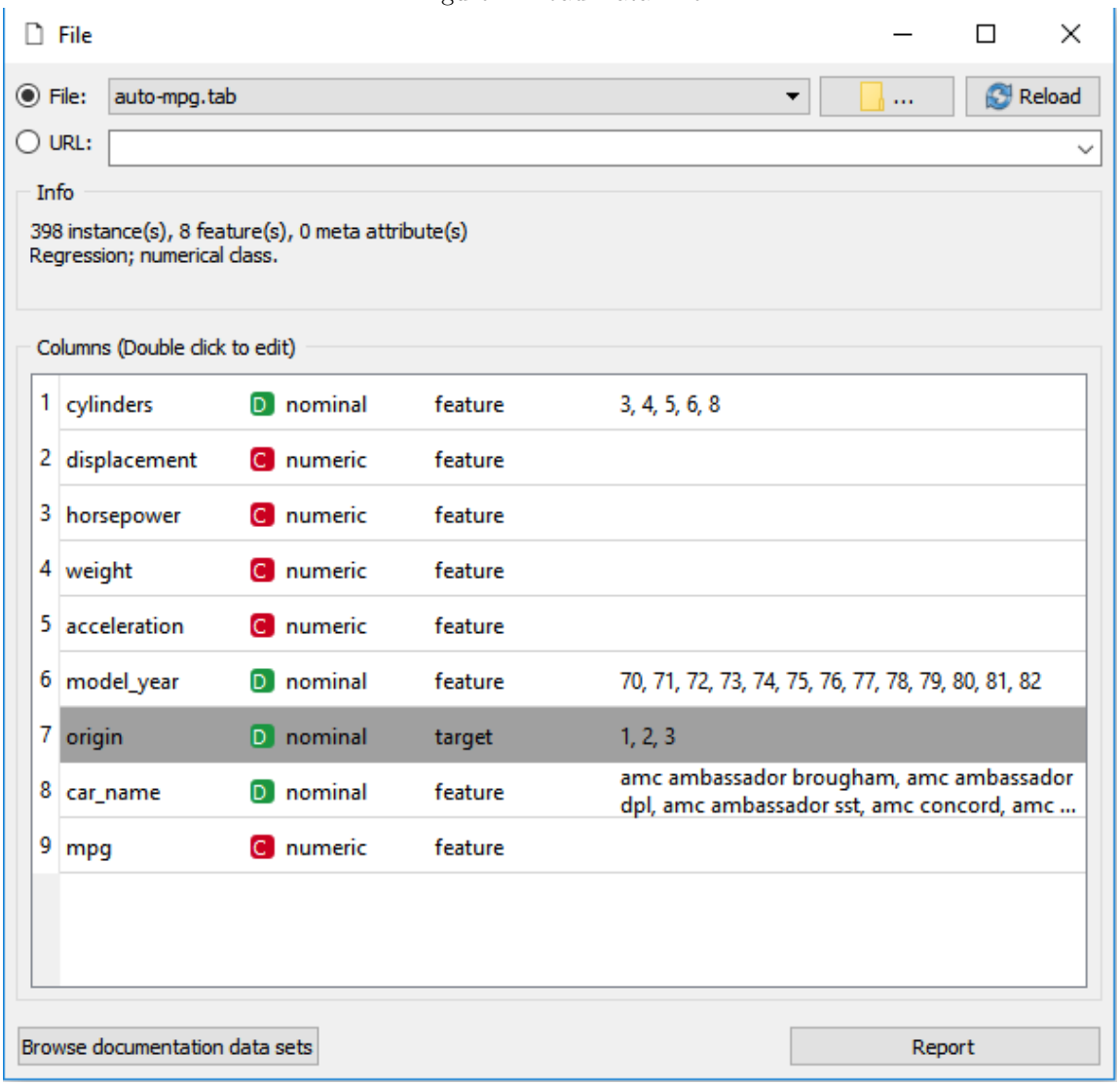
1 Question 1:

1.1 Load the auto-mpg sample dataset into the Orange application - ensure that origin is set as a target attribute type, as it will be used as a class label. Perform a Hierarchical Clustering using Linkage set to Average , after calculating Distances, with Pruning set to a Max Depth of 5. Also, set Selection to Top N with a value of 3. This will result in a shallow tree of depth 5, and a nal cut resulting in 3 clusters. Examine the resulting clusters (C1,C2,C3) via Distributions analysis - is there a clear relationship between the cluster assignment and class label (1,2,3)? What are the probabilities calculated for each value of origin for each cluster? Does changing the Max Depth affect the results in any way?

1.2 Answer:

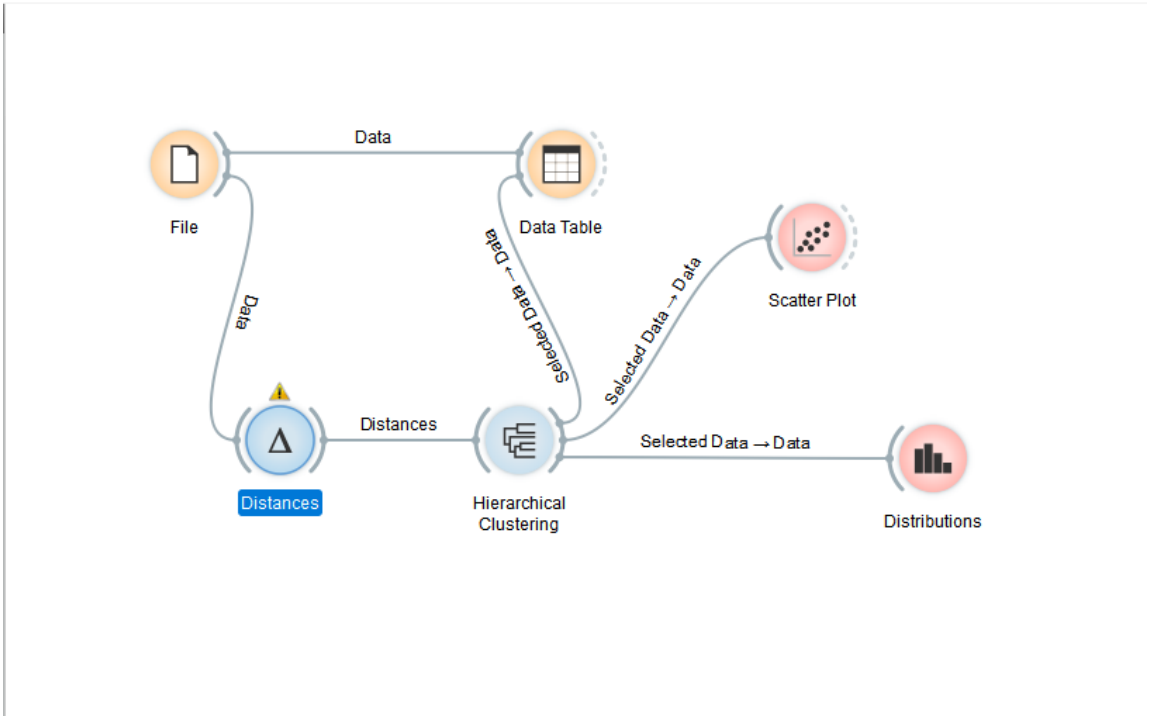
- Loading the dataset and changing *origin* to a target variable.

Figure 1: Load Data File



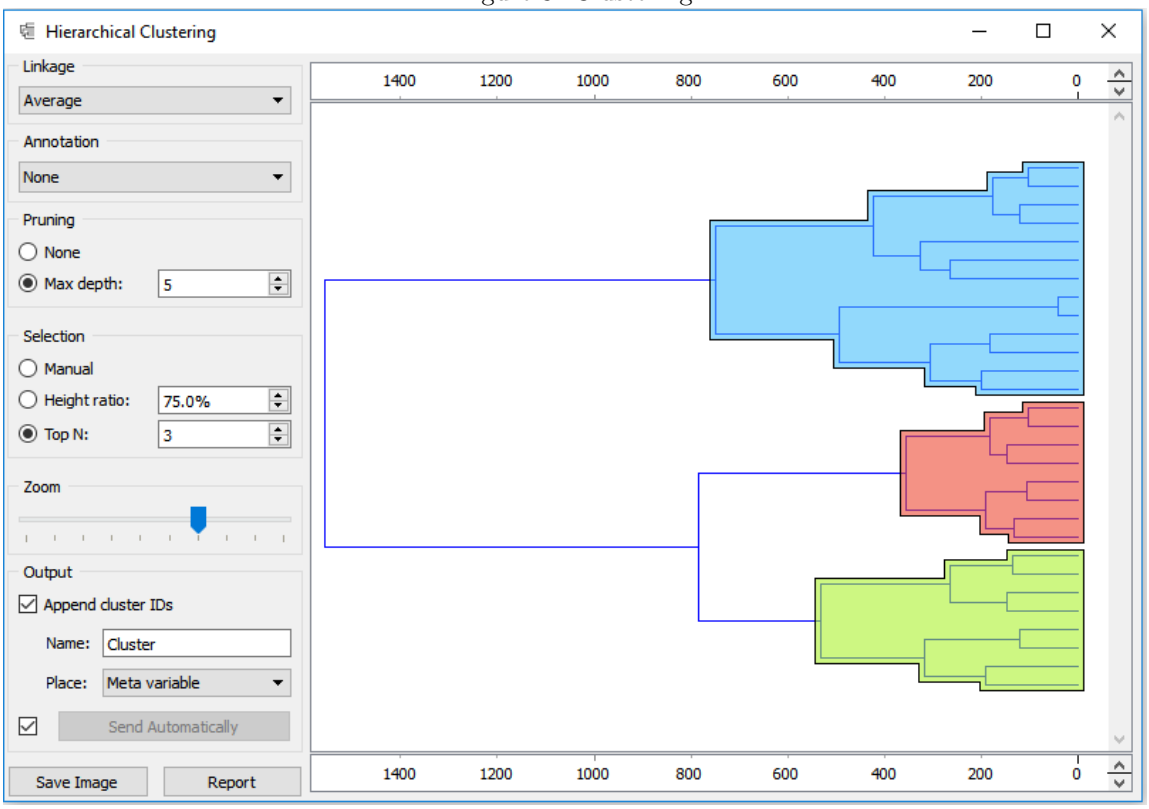
- Creating the appropriate workflow. The euclidean distance metric is calculated and then passed on to the Heirarchial Clustering widget.

Figure 2: Workflow



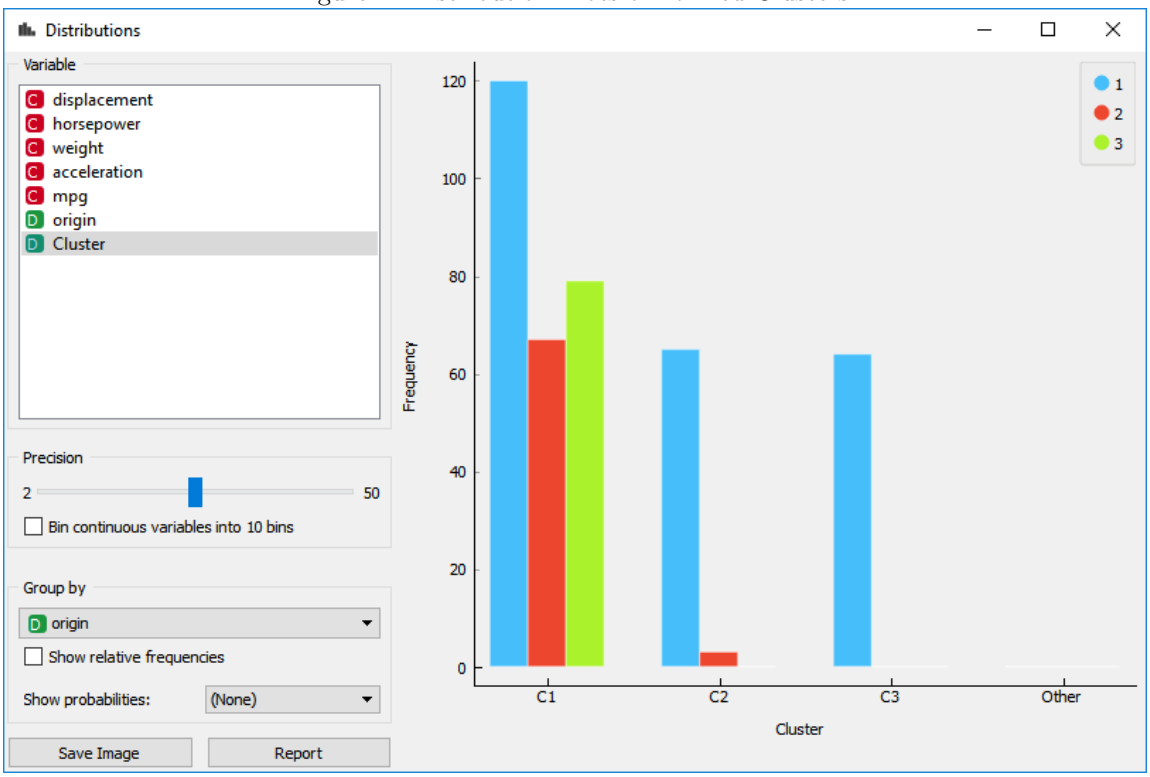
- Setting up parameters of Heirarchial Clustering.

Figure 3: Clustering



- Distribution Analysis.

Figure 4: Distribution Plots of Formed Clusters.



- Including the Probabilities tab we get the probability of each class belonging to a cluster.

Table 1: Class Probabilities

Cluster	Origin 1	Origin 2	Origin 3
1	0.451	0.252	0.297
2	0.956	0.044	0
3	1.0	0.0	0.0

1.3 Observations:

On observation of Figure 4. it is seen that class 2 and 3 are clustered together while class 1 is spread out among all the three clusters. There does not seem to be a definitive relation among the three classes except the fact that class 2 and 3 seem similar. The probability distribution is shown in Figure 5. Changing the *max depth* does not result in a change, only if it is equal to 1 does the number of clusters reduce to 2 which is pretty obvious.

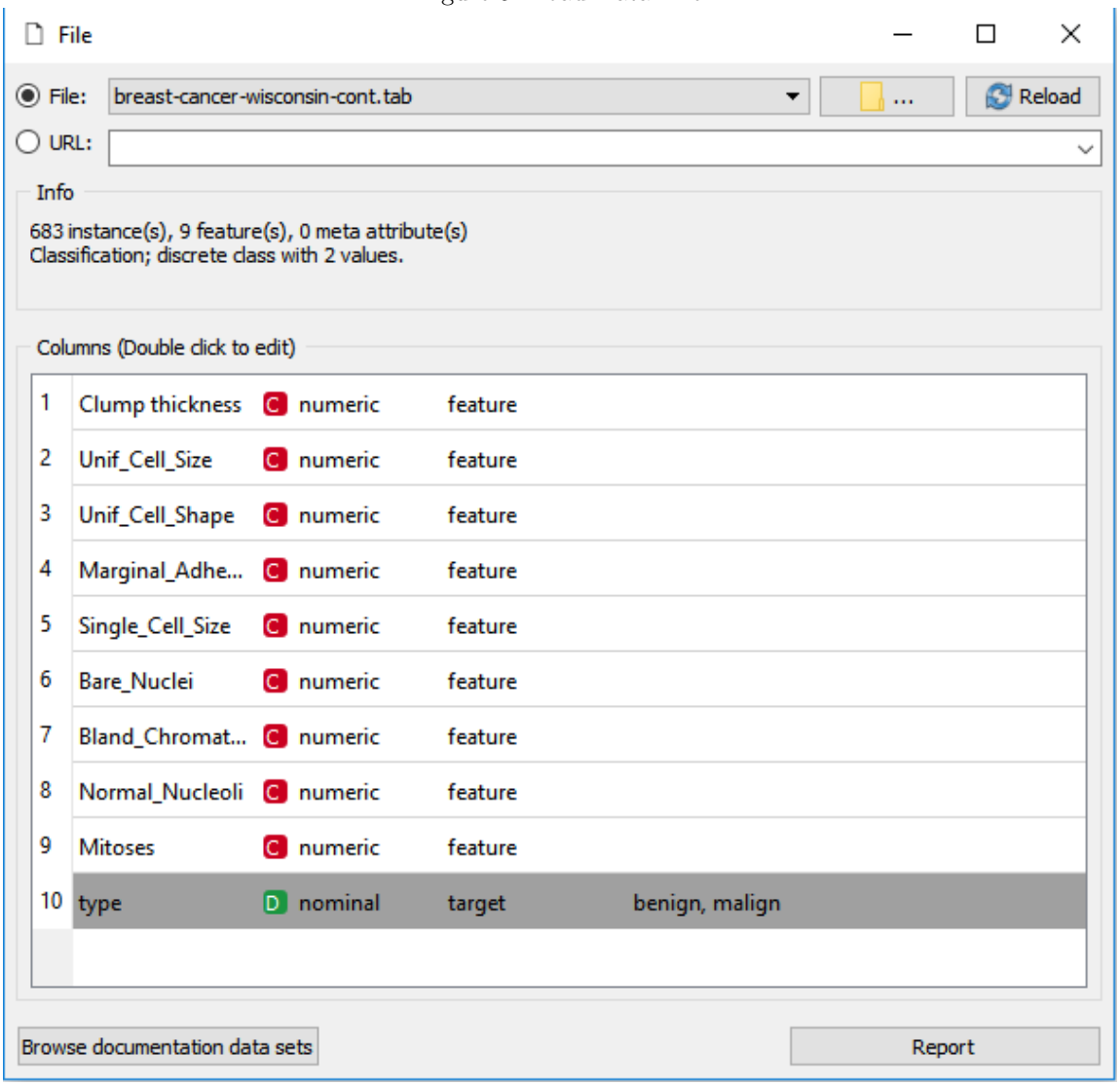
2 Question 2:

2.1 Load the breast-cancer-wisconsin-cont dataset into the Orange application,and run a k-means analysis with the number of clusters Optimized From values for k from 2 to 5. Use Silhouette scoring - what is the score for each value of k? For the best score, what are the coordinates of the centroids? What are the distances between the centroids for the best score?

2.2 Answer:

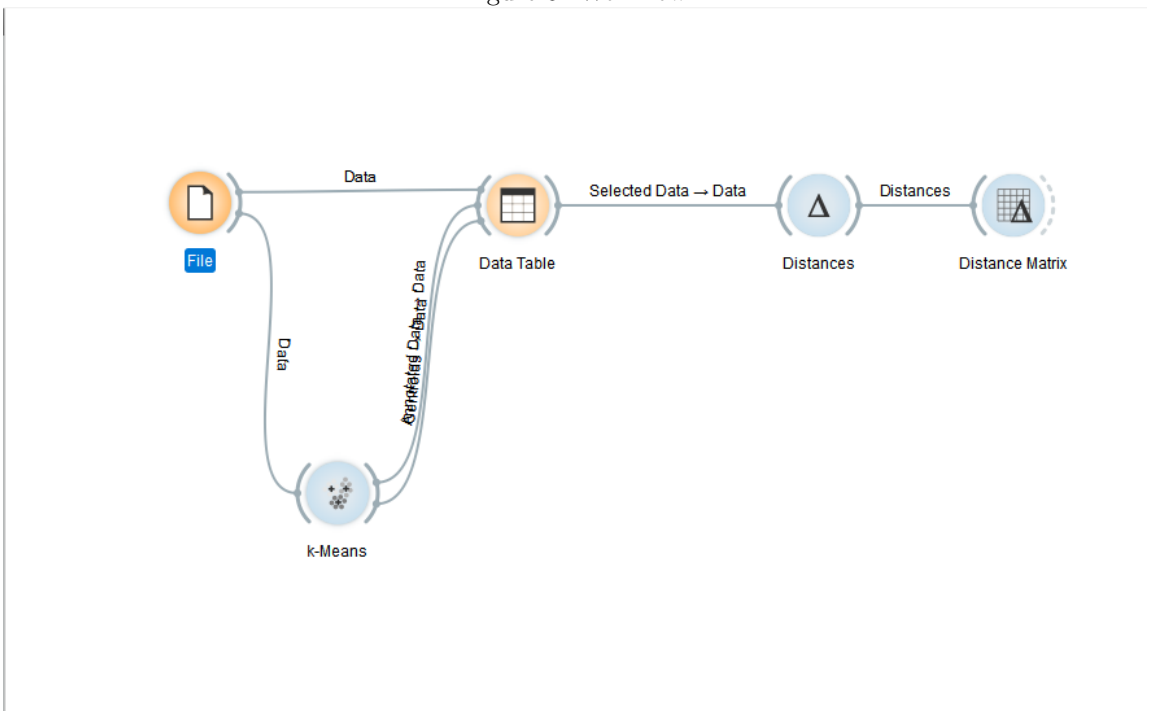
- Loading the dataset.

Figure 5: Load Data File



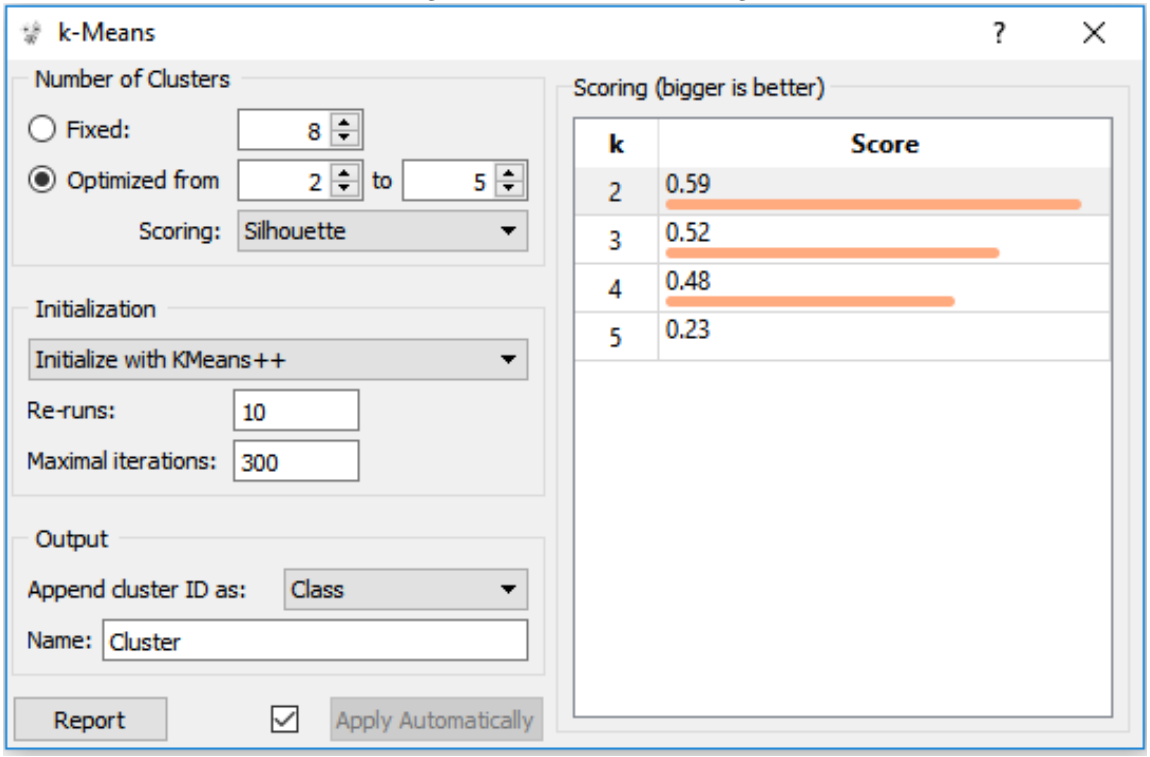
- Creating the appropriate workflow.

Figure 6: Workflow



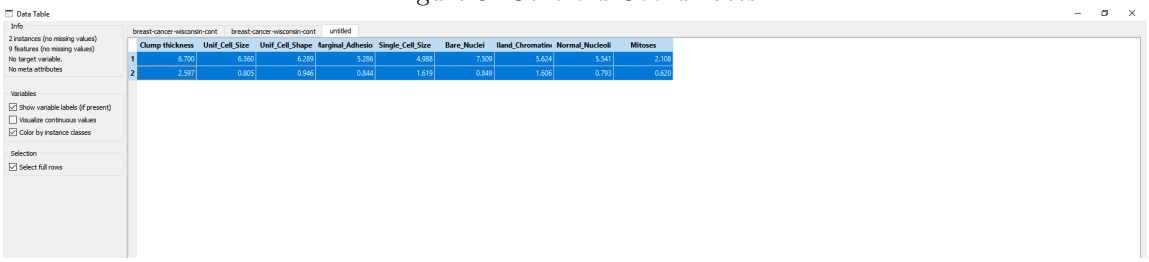
- Performing K-Means Clustering with the given parameters.

Figure 7: K-Means Clustering



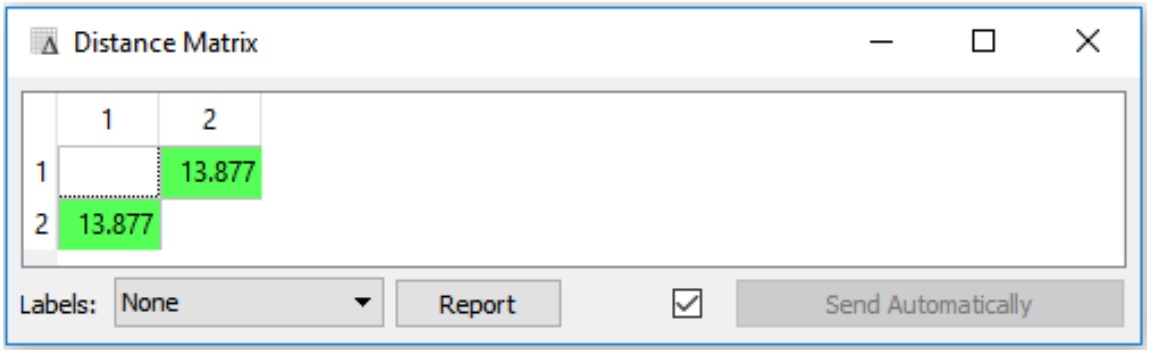
- Passing Centroid Coordinates of model with highest silhouette score to display as table.

Figure 8: Centroid Coordinates



- Calculating euclidean distance between centroids using distance widget and visualized by distance matrix widget.

Figure 9: Centroid Distance



2.3 Observations:

For models with cluster numbers ranging from 2 to 5, the silhouette scores are shown in Figure 8. with model with k=2 wins out with a score of 0.59 . For this model, the centroid coordinates are shown in Figure 9. For the same model, the distance between the centroids is shown in figure 10. and is calculated to be 13.87 .