Nasri Chiheb

UNIVERSITAT ROVIRA I VIRGILI

Neural and Evolutionary Computation

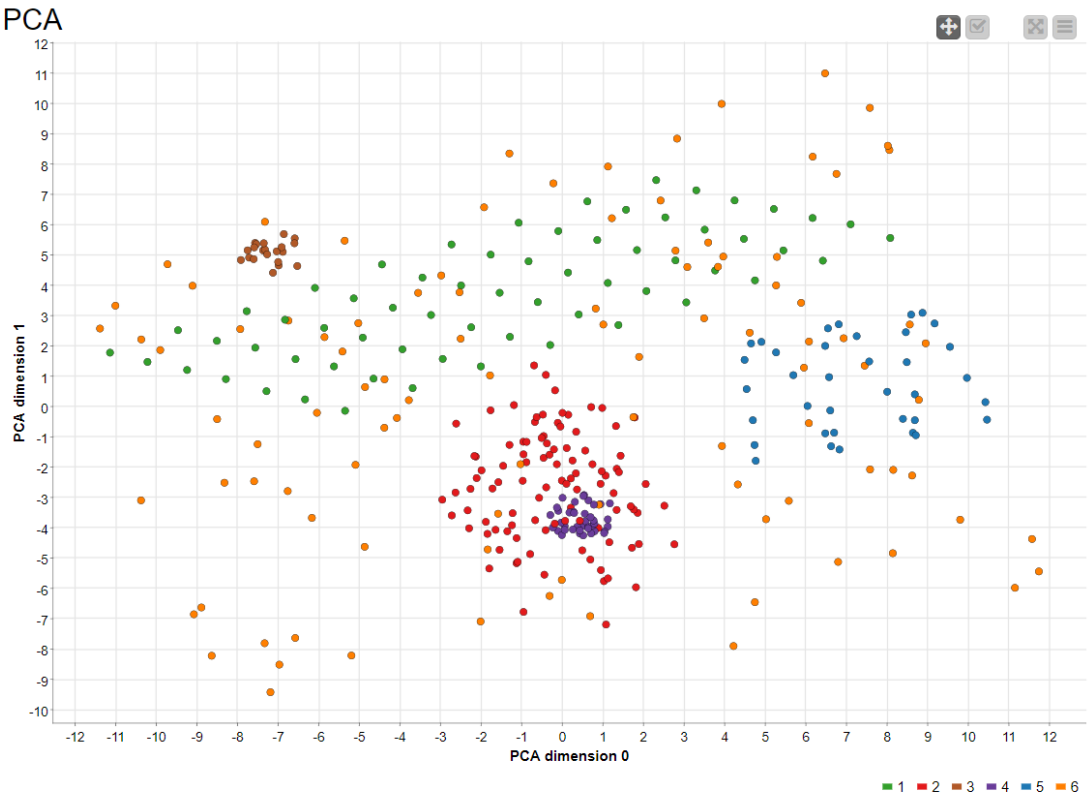
A3: Unsupervised learning with PCA, t-SNE, k-means and SOM

Introduction:

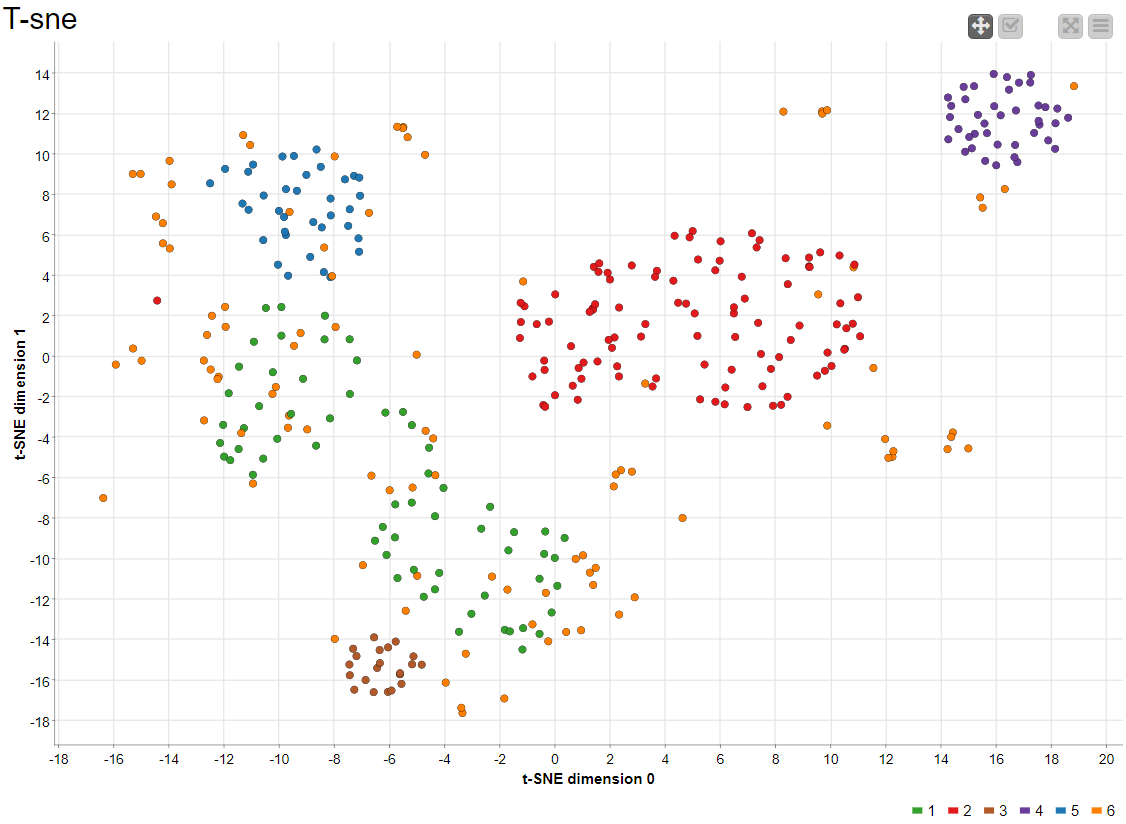
The goal of this assignment is to compare four unsupervised learning method PCA, t-SNE , K-means and SOM. All methods will be compared with the same dataset. The dataset contains 360 patterns of four variables and one class column. In this report we will discuss about the different results after giving a short explanation of the implementation. As done in A2, I will use Knime for the implementation despite for SOM which will be done with R. Please find attached to this report the results and the scripts.

Description of the implementation:

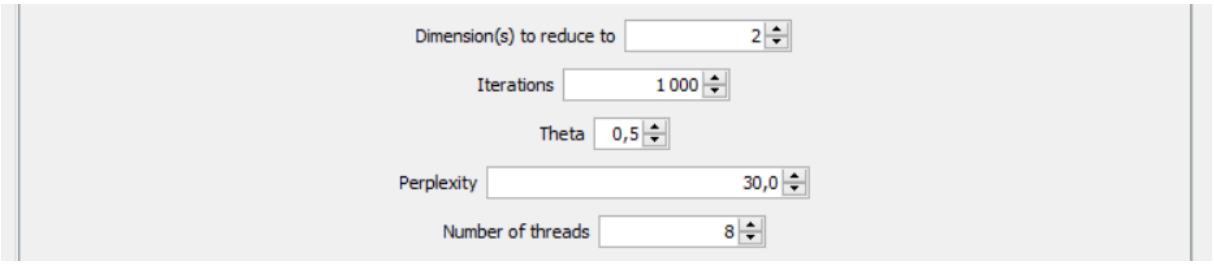
PCA : Implementing PCA is quite fast. Knime nodes for PCA do not offer any configuration. At the end we get this result.



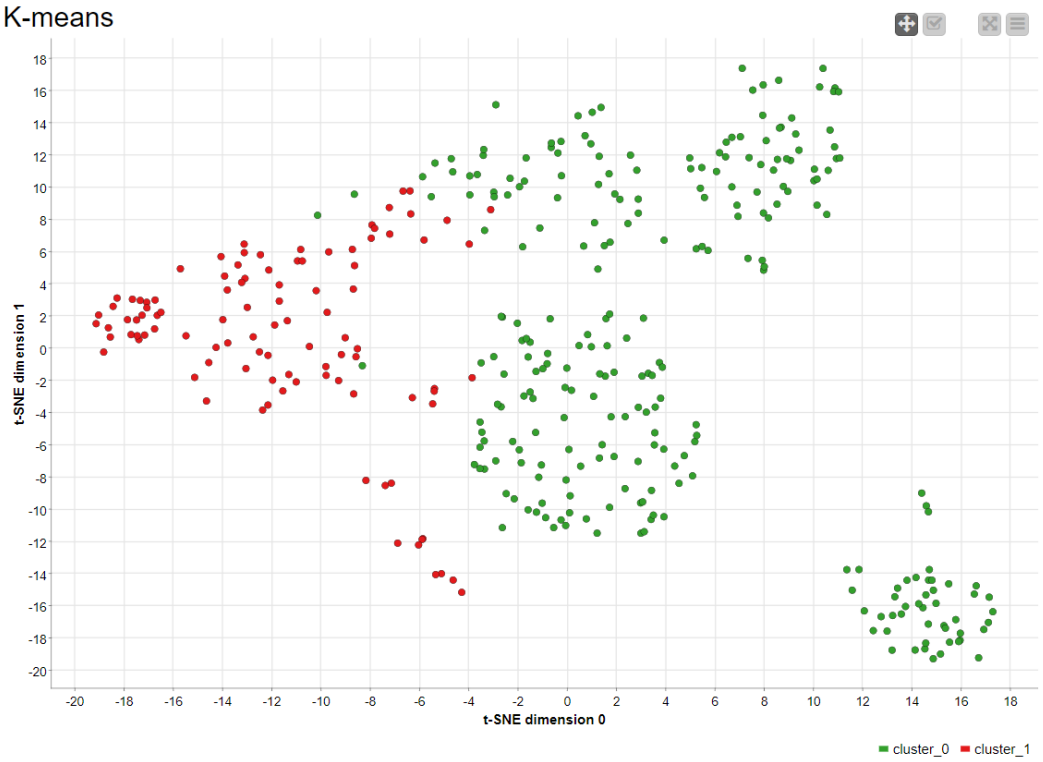
t-SNE : Same as PCA we plot the dataset in two dimensions.

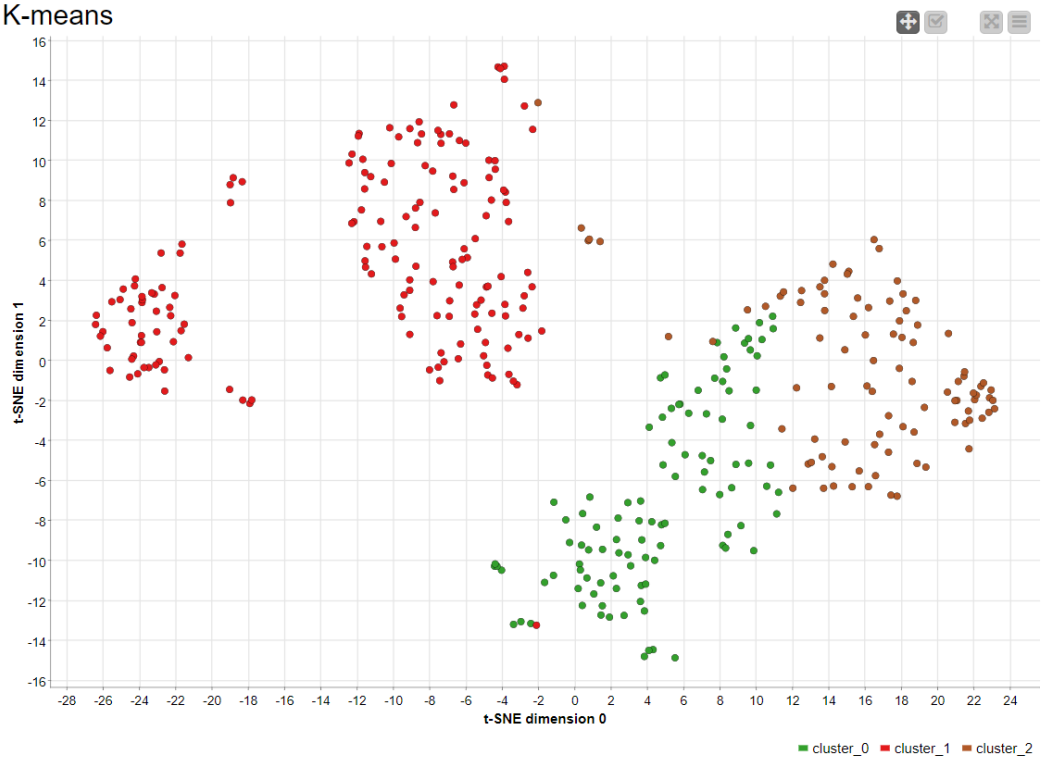


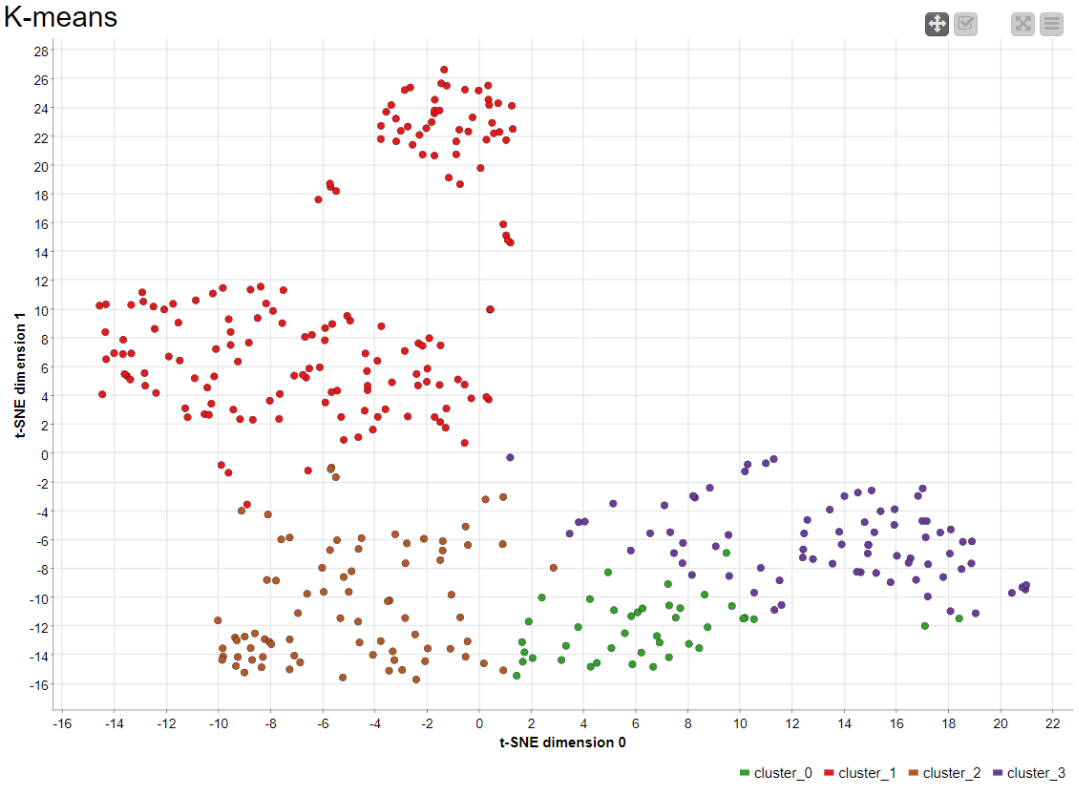
Using these parameters :

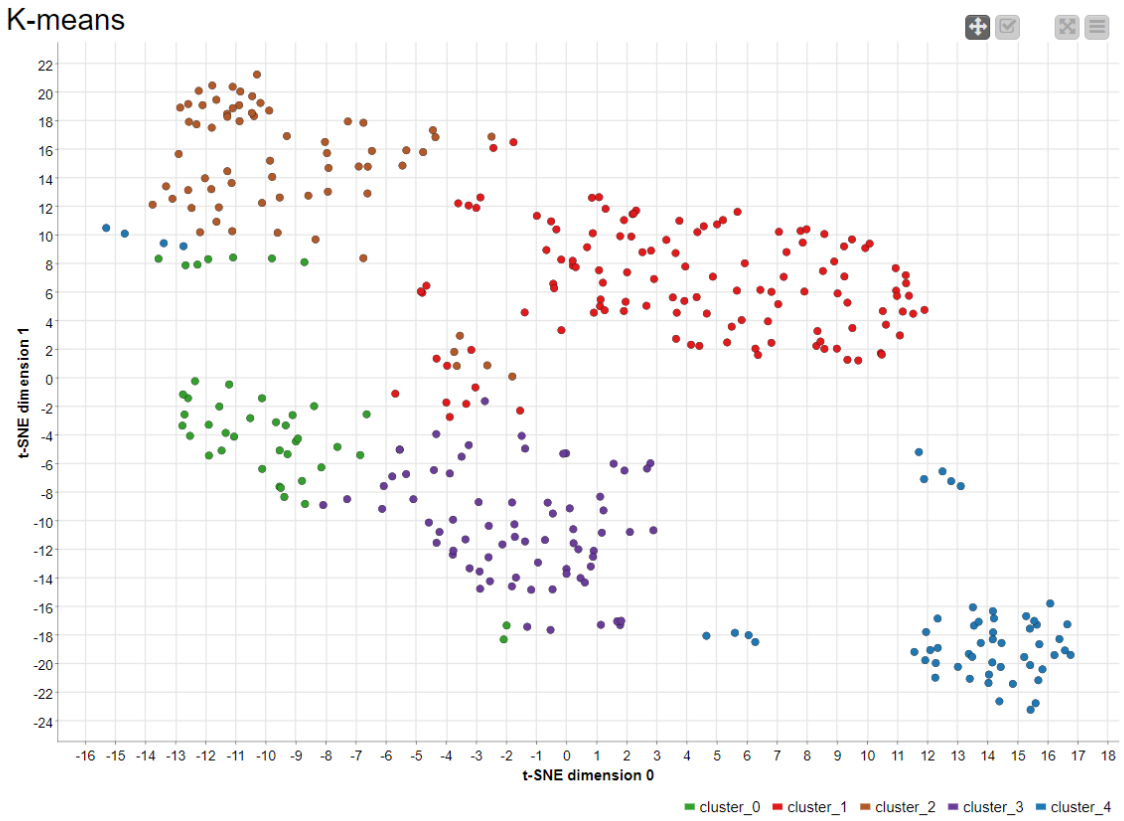


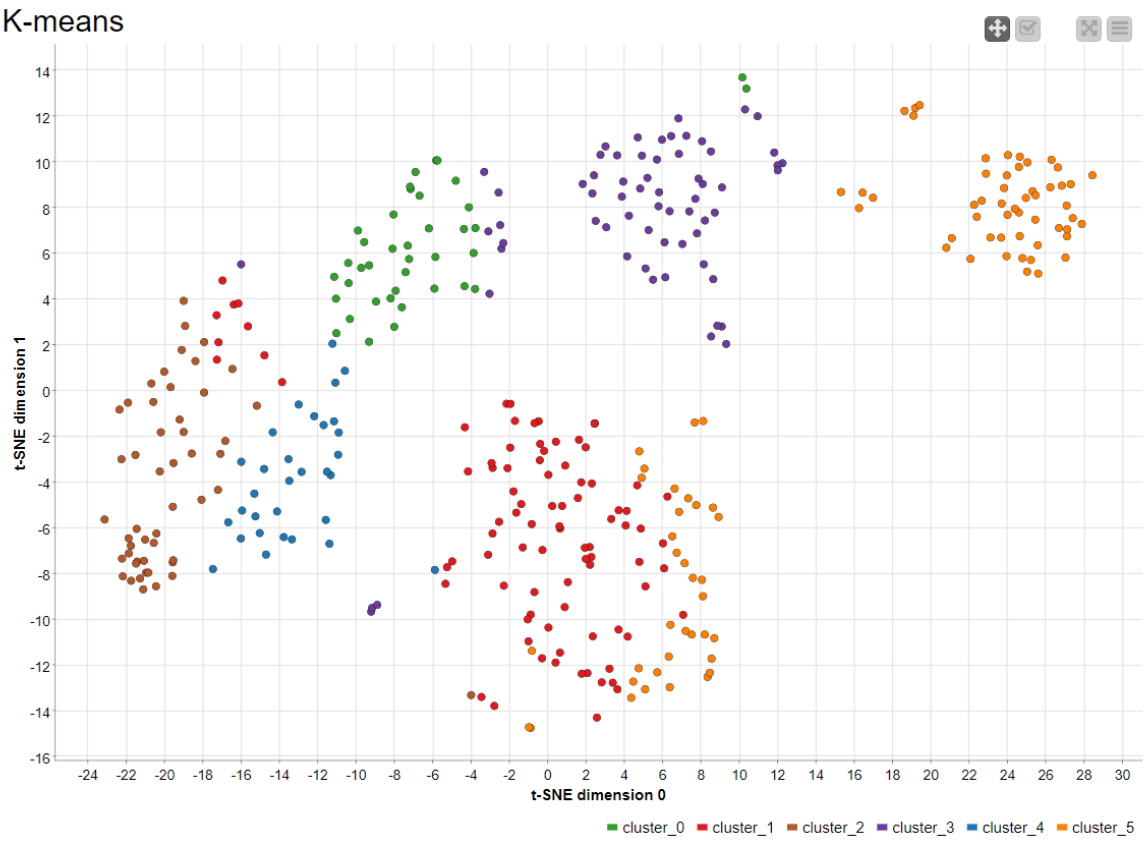
K-means : For k-means we did several plots by varying the number of clusters from 2 to 7. The plots were maid using t-SNE to reduce dimensions to 2 .We will discuss the results in next part.

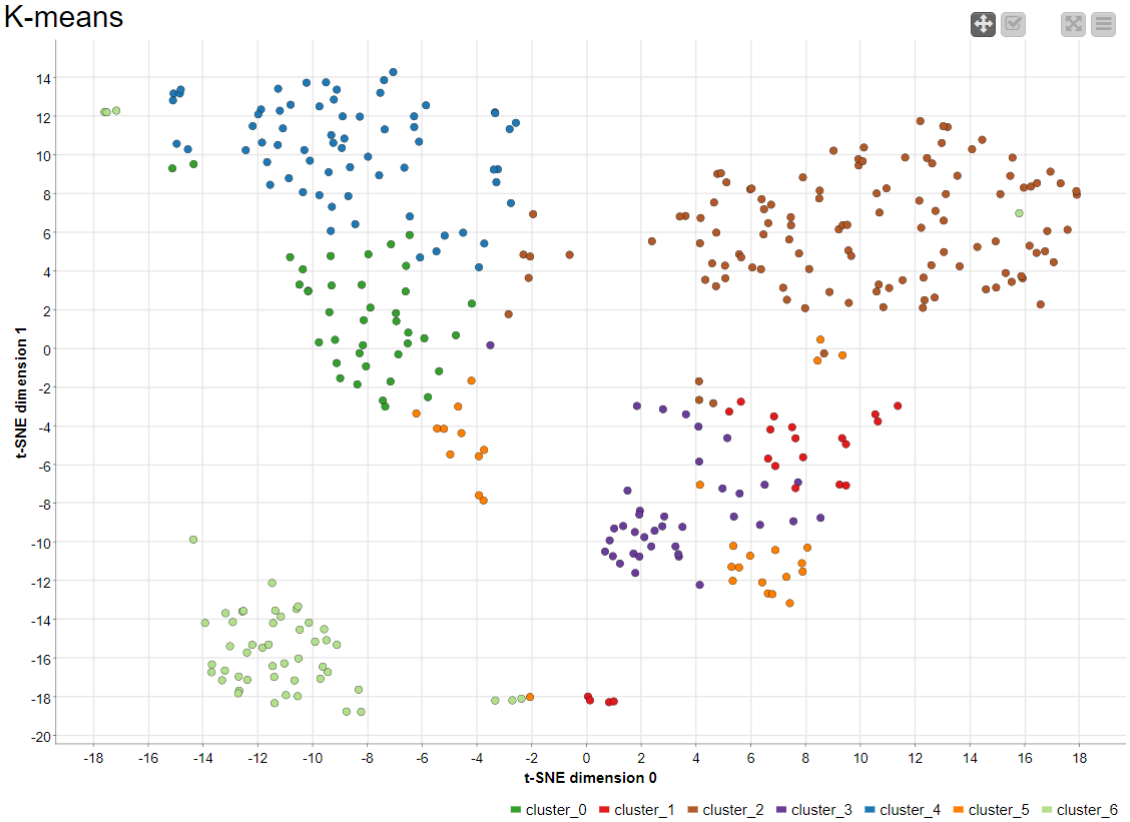




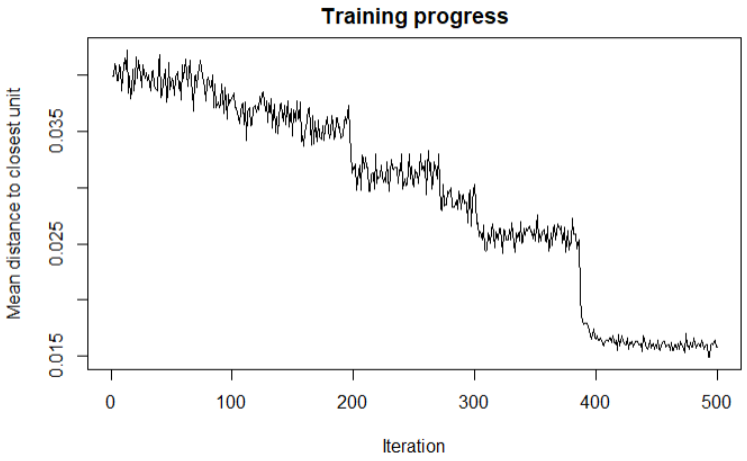

K-means
t-SNE dimension 1
t-SNE dimension 0
cluster_0
cluster_1
cluster_2
cluster_3

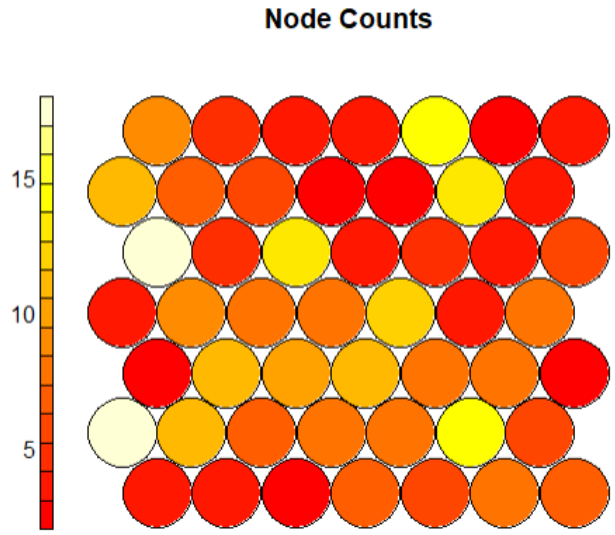

K-means
t-SNE dimension 1
t-SNE dimension 0
cluster_0
cluster_1
cluster_2
cluster_3
cluster_4

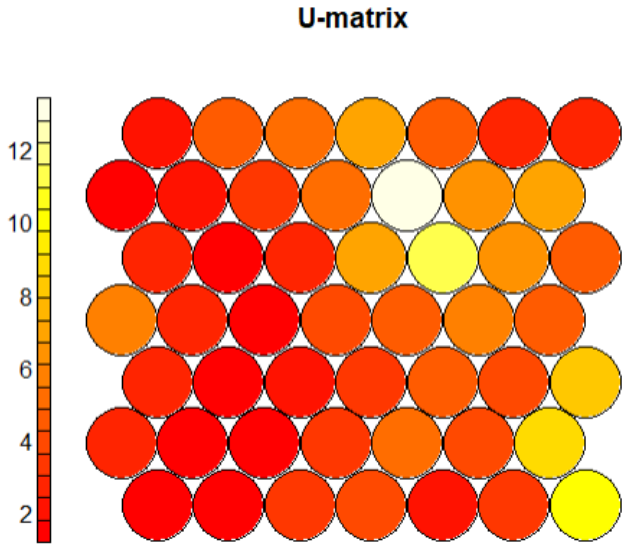

K-means
t-SNE dimension 1
t-SNE dimension 0
cluster_0 cluster_1 cluster_2 cluster_3 cluster_4 cluster_5

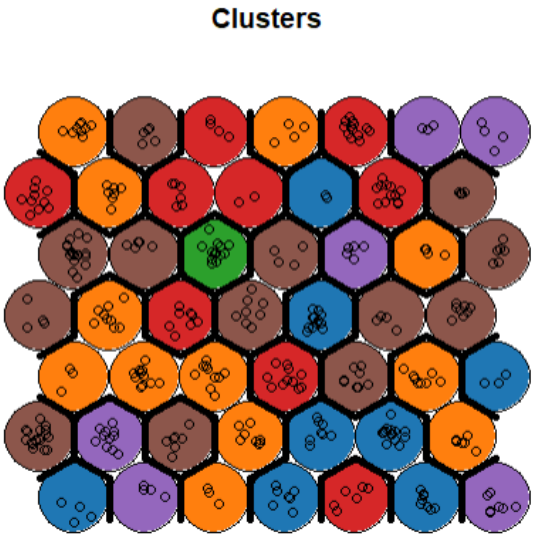

K-means
t-SNE dimension 1
t-SNE dimension 0
cluster_0 cluster_1 cluster_2 cluster_3 cluster_4 cluster_5 cluster_6

SOM : I choose to work with R as SOM implementation in KNIME do not offer visualization yet. The implementation is done using Kohenen package. I use hexagonal map with 8x8 size and bubble neighborhood function. The learning rate is set between 0.05 and 0.01 with 500 iterations. With this configuration we get 0.015 mean distance to closest node and between 5-10 samples per node .At the end we obtain this result.









Results:

For PCA results we can see that clusters are overlapping (e.g. cluster 2 and 4). Overall, the clusters are not well separated. t-SNE gives good results with clear cluster identifications. For k-means we used several numbers of clusters overall, 5 or 6 clusters suits this dataset. For SOM we can see that cluster melt between each other, and no clear pattern is identified.