

**ISM 6136 – Datamining/Predictive Analytics**

**Class Assignment 9**

**10 points**

**TASK: Performing Clustering – Data Mining Task using XLMiner or RapidMiner**

1. For the Public Utilities Dataset: 22 US Utility firms and 8 variables. Using the **Hierarchical clustering algorithm** (Try out 3 different number of clusters values and determine the following:

a) The cluster with maximum no. of utilities that are operating in a similar manner based

on the 8 variables

b) Any outlier utilities that are not combined with another one to form a cluster?

c) Any other cluster# identification and talk about the number and which utilities are under it?

*Note: This is an example where clustering would be useful as a study to predict the cost impact of deregulation. To perform the requisite analysis, economists would be required to build a detailed cost model of the various utilities. It would save a considerable amount of time and effort by clustering similar types of utilities, building a detailed cost model for just one typical utility in each cluster, then scaling up from these models to estimate results for all utilities.*

1. Using **K-means clustering algorithm** on the cereal dataset - find out the following and and explain along with screen shots for each of the answer. Try out at least 3 different number of clusters to determine the following:

Explain how many clusters you had to create to get these answers.

1. Cluster of ‘healthy cereals (low fat, low salt etc)’. Which cereals are part of that cluster?
2. Cluster with lowest consumer ratings?
3. Any other cluster identification?

**If using XLMIner - Submit the two Excel spreadsheets and this Word Document (Question answers along with screen shots) on Canvas.**

**If using RapidMiner – Submit the screen shots of your Design and Results (Graph and cluster selection) and answers of the questions asked on this Word document.**