

**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

Table

Description automatically generated

Based on these results, cluster number 1 has the most similar acting utilities which are: Arizona, Boston, Central, Common, Florida, Kentucky, Madison, Northern, Oklahoma, Southern, Texas, Wisconsin, Virginia.

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

There is an outlier utility in this case (screenshot above), it is utility Consolid.

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

\*Screenshots provided below

Interestingly, the cluster 4 in the screenshot in part A is the same as cluster 2 below which both consist of Idaho, Nevada, Puget, but Nevada becomes an outlier when the max number of clusters is increased (shown in the second screenshot). From the screenshot above cluster 3 seems to be comprised of coastal areas or areas near water as the utilities are Hawaiian, New England, Pacific, San Diego, United.

Table

Description automatically generated

Table

Description automatically generated

*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?

Cluster 2 on tab KMC Clusters which is 4 max clusters had the most healthy cereals out of the clusters. The cereals included Puffed rice, Puffed wheat, Shredded wheat, shredded wheat n bran, shredded wheat spoon size, raisin squares, strawberry fruit wheels, cream of wheat (these are some of the top healthy cereals when the original data is sorted based on low fat and low sodium and low sugars which is what I sorted based on what a healthy cereal would have).

Table

Description automatically generated

1. Cluster with lowest consumer ratings?

Cluster 10, which had the most cereals with the lowest ratings. On tab KMC Clusters3

A screenshot of a computer

Description automatically generated with medium confidence

1. Any other cluster identification?

A picture containing calendar

Description automatically generated

On tab KMC Clusters

The highlighted cluster 3 includes cereals 100% bran, all bran, and all bran extra fiber, while cluster 4 the first record is 100% natural bran which by name would look to be included in cluster 3 but is more similar to higher calorie cereals as what the records after it look to be.

\*For parts A and B, different KMC were used as the number of total clusters would have roughly the same number of records in multiple clusters so I chose to use two outputs that held more records of what was being asked.