

**ISM 6136 – Datamining/Predictive Analytics**

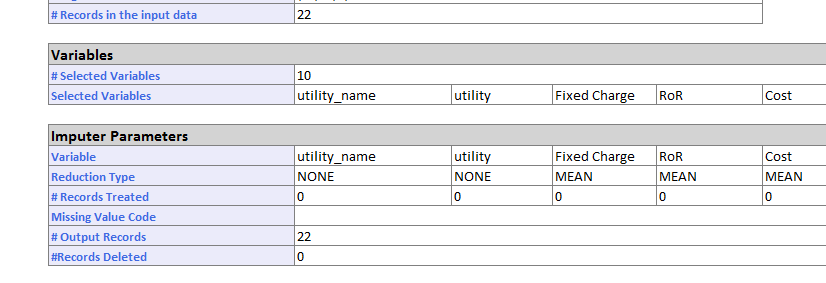
**Class Assignment 8**

**5 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 or K-means clustering algorithm (Try out 3 different number of clusters and determine the following:

No missing value



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?

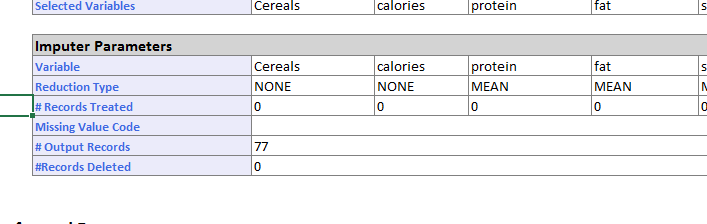
Answers are provided in the table below. Performed hierarchical clustering algorithm. Input data was normalized.

|  |  |  |  |
| --- | --- | --- | --- |
| **Question** | **Cluster #1**  **No. of leaves=10**  **No. of clusters=4** | **Cluster #2**  **No. of leaves=10**  **No. of clusters=6** | **Cluster #3**  **No. of leaves=10**  **No. of clusters=8** |
| **Part a) Maximum no. of utilities operating at a similar manner** | Cluster 1 has maximum (13) number of utilities. This includes Arizona, Boston, Central, Common, Florida, Kentucky, Madison, Nothern, Okhlahoma, Southern, Texas, and Virginia. | Cluster 1 has maximum (7) number of utilities. This includes Arizona, Central, Florida, Kentucky, Okhlahoma, Southern, and Texas. | Cluster 2 has maximum (6) number of utilities. This includes Boston, Common, Madison, Northern, Wisconsi and Virginia. |
| **Part b) Outlier utilities not combined with another to form a cluster** | Record 5 or Consolid is an outlier that is not combined with any cluster. | Record 5 or Consolid and  Record 17 or San Diego  Are outliers that are not combined with any cluster. | Record 5 or Consolid,  Record 11 or Nevada and  Record 17 or San Diego  Are outliers that are not combined with any cluster. |
| **Part c) other cluster# identification -number and utilities are under it** | Cluster 3 has four utilities – Hawaiian, New England, Pacific and San Diego.  Cluster 4 has three utilities – Idaho, Nevada and Puget. | Cluster 2 has six utilities – Boston, Common, Madison, Northern, Wisconsin and Virginia.  Cluster 4 has four utilites- Hawaiian, New England, United and Pacific.  Cluster 5 has three utilities – Idaho, Nevada and Puget. | Cluster 1 has five utilities – Arizona, Florida, Okhlahoma, Southern and Texas.  Cluster 3 has two utilities – Central and Kentucky.  Cluster 5 has four utilities – Hawaiian, New England, Pacific and Untied.  Cluster 6 has two utilities – Idaho and Puget. |

*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 **both** the Hierarchical and K-means clustering algorithm - 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:

No missing values.



Using Hierarchical algorithm (explain how many clusters you had to create to get these

answers)

1. Which is the cluster of ‘healthy cereals (low fat, low salt etc)’. Which cereals are part of that cluster?

To answer this question, columns A (cereals), column K(shelf) and column L(rating) were not considered. Input data was normalized.

Number of clusters

Trial 1 – 4 clusters, 10 leaves 🡺 Tab in excel is highlighted- HC\_Cluster

Trial 2 – 9 clusters, 10 leaves

Trial 3- 12 cluster, 10 leaves

Trial 4 – 18 clusters, 10 leaves

Trial 5 – 3 clusters, 10 leaves

Increasing the clusters splits one of most healthy cereal (All Bran with Extra Fiber) as an outlier. Keeping the number of clusters to 3 or 4 yields the following result. Hence, it can be said the cluster of healthy cereals are:

1. 100% Bran

2. All Bran

3. All Bran with Extra Fiber

Cluster 1= healthy cereals

|  |  |  |  |
| --- | --- | --- | --- |
| **Record ID** | **Cluster** | **Sub-Cluster** | **Column1** |
| **Record 1** | 1 | 1 | 1 is 100%\_Bran |
| **Record 3** | 1 | 1 | 2 is 100%\_Natural\_Bran |
| **Record 4** | 1 | 3 | 3 is All-Bran |
| **Record 2** | 2 | 2 | 4 is All-Bran\_with\_Extra\_Fiber |
| **Record 5** | 2 | 4 | 5 is Almond\_Delight |
| **Record 6** | 2 | 4 | 6 is Apple\_Cinnamon\_Cheerios |
| **Record 7** | 2 | 4 | 7 is Apple\_Jacks |

Definition of “healthy” cereal – high protein, low fat, low sodium, high fiber, low carbs, low sugar, high potassium, high vitamins.

1. Which is the cluster with lowest consumer ratings?

To answer this question, only column L(rating) was considered. With 7 clusters, it can be seen that cluster # 6, as shown below, forms a cluster for all cereals rates less than 25%.

Tab in excel is highlighted- HC\_Cluster6

|  |  |  |
| --- | --- | --- |
| **Record 11** | 6 | 6 |
| **Record 13** | 6 | 6 |
| **Record 15** | 6 | 7 |
| **Record 19** | 6 | 7 |
| **Record 32** | 6 | 7 |
| **Record 36** | 6 | 7 |

|  |  |
| --- | --- |
| Cereals (record no) | rating |
| Cap'n'Crunch (11) | 18.04285 |
| Cinnamon\_Toast\_Crunch (13) | 19.82357 |
| Honey\_Graham\_Ohs (36) | 21.87129 |
| Count\_Chocula (19) | 22.39651 |
| Cocoa\_Puffs (15) | 22.73645 |
| Golden\_Grahams (32) | 23.80404 |

1. Any other cluster identification?

With 4 clusters, cluster #3 reveals cereals high in “vitamins”

|  |  |  |
| --- | --- | --- |
| **Record 39** | 3 | 6 |
| **Record 40** | 3 | 6 |
| **Record 54** | 3 | 6 |
| **Record 70** | 3 | 6 |
| **Record 71** | 3 | 10 |
| **Record 72** | 3 | 6 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cereals (record no) | calories | protein | fat | sodium | fiber | carbo | sugars | potass | vitamins | shelf | rating |
| Total\_Whole\_Grain (72) | 100 | 3 | 1 | 200 | 3 | 16 | 3 | 110 | 100 | 3 | 46.65884 |
| Product\_19 (54) | 100 | 3 | 0 | 320 | 1 | 20 | 3 | 45 | 100 | 3 | 41.50354 |
| Just\_Right\_Crunchy\_\_Nuggets (39) | 110 | 2 | 1 | 170 | 1 | 17 | 6 | 60 | 100 | 3 | 36.52368 |
| Total\_Corn\_Flakes (70) | 110 | 2 | 1 | 200 | 0 | 21 | 3 | 35 | 100 | 3 | 38.83975 |
| Total\_Raisin\_Bran (71) | 140 | 3 | 1 | 190 | 4 | 15 | 14 | 230 | 100 | 3 | 28.59279 |
| Just\_Right\_Fruit\_&\_Nut (40) | 140 | 3 | 1 | 170 | 2 | 20 | 9 | 95 | 100 | 3 | 36.47151 |

Using K-means algorithm (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?

Trial 1 – 4 clusters, 10 iterations- Tab in excel is highlighted- KMC\_Clusters

Trial 2- 7 clusters, 10 iterations

Trial 1 and 2 both grouped the most healthy cereals in one cluster. From trial 1, cluster 3 produced the same results as from part a) with the following “healthy” cereals

1. 100% Bran- Record 1

2. All Bran -Record 3

3. All Bran with Extra Fiber – record 4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Record 1** | 3 | 7.03430605 | 5.60757149 | 1.14964993 | 5.15942326 |
| **Record 3** | 3 | 6.73851709 | 5.7885504 | 1.4266418 | 5.15719285 |
| **Record 4** | 3 | 8.47238544 | 7.39478194 | 1.88177688 | 7.33679499 |

1. Cluster with lowest consumer ratings?

To answer this question, only column L(rating) was considered.

Trial 1 – 7 clusters, 10 iterations

Trial 2 – 10 clusters, 10 iterations ==🡺 Maximum clusters allowed in the XLminer version is 10 for K-means. Could not increase further.

With 10 clusters, it can be seen that cluster # 10, as shown below, forms a cluster for all cereals rates less than 29%.

Tab in excel is highlighted- KMC\_Cluster3

|  |  |
| --- | --- |
| **Record 11** | 10 |
| **Record 13** | 10 |
| **Record 15** | 10 |
| **Record 19** | 10 |
| **Record 30** | 10 |
| **Record 32** | 10 |
| **Record 36** | 10 |
| **Record 38** | 10 |
| **Record 43** | 10 |
| **Record 71** | 10 |
| **Record 74** | 10 |

|  |  |
| --- | --- |
| Cereals (record no) | rating |
| Cap'n'Crunch (11) | 18.04285 |
| Cinnamon\_Toast\_Crunch (13) | 19.82357 |
| Honey\_Graham\_Ohs (36) | 21.87129 |
| Count\_Chocula (19) | 22.39651 |
| Cocoa\_Puffs (15) | 22.73645 |
| Golden\_Grahams (32) | 23.80404 |
| Lucky\_Charms (43) | 26.73452 |
| Trix (74) | 27.7533 |
| Fruity\_Pebbles (30) | 28.02577 |
| Total\_Raisin\_Bran (71) | 28.59279 |
| Honey-comb (38) | 28.74241 |

1. Any other cluster identification?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Record 2** | 4 |  |  | Record# | Cereals | calories |
| **Record 5** | 4 |  |  | 47 | Mueslix\_Crispy\_Blend | 160 |
| **Record 6** | 4 |  |  | 45 | Muesli\_Raisins,\_Dates,\_&\_Almonds | 150 |
| **Record 8** | 4 |  |  | 46 | Muesli\_Raisins,\_Peaches,\_&\_Pecans | 150 |
| **Record 11** | 4 |  |  | 40 | Just\_Right\_Fruit\_&\_Nut | 140 |
| **Record 12** | 4 |  |  | 50 | Nutri-Grain\_Almond-Raisin | 140 |
| **Record 13** | 4 |  |  | 71 | Total\_Raisin\_Bran | 140 |
| **Record 14** | 4 |  |  | 8 | Basic\_4 | 130 |
| **Record 20** | 4 |  |  | 52 | Oatmeal\_Raisin\_Crisp | 130 |
| **Record 28** | 4 |  |  | 2 | 100%\_Natural\_Bran | 120 |
| **Record 29** | 4 |  |  | 11 | Cap'n'Crunch | 120 |
| **Record 35** | 4 |  |  | 13 | Cinnamon\_Toast\_Crunch | 120 |
| **Record 36** | 4 |  |  | 28 | Fruit\_&\_Fibre\_Dates,\_Walnuts,\_and\_Oats | 120 |
| **Record 37** | 4 |  |  | 29 | Fruitful\_Bran | 120 |
| **Record 42** | 4 |  |  | 35 | Great\_Grains\_Pecan | 120 |
| **Record 45** | 4 |  |  | 36 | Honey\_Graham\_Ohs | 120 |
| **Record 46** | 4 |  |  | 49 | Nut&Honey\_Crunch | 120 |
| **Record 47** | 4 |  |  | 53 | Post\_Nat.\_Raisin\_Bran | 120 |
| **Record 50** | 4 |  |  | 59 | Raisin\_Bran | 120 |
| **Record 52** | 4 |  |  | 5 | Almond\_Delight | 110 |
| **Record 53** | 4 |  |  | 6 | Apple\_Cinnamon\_Cheerios | 110 |
| **Record 57** | 4 |  |  | 7 | Apple\_Jacks | 110 |
| **Record 58** | 4 |  |  | 12 | Cheerios | 110 |
| **Record 59** | 4 |  |  | 14 | Clusters | 110 |
|  |  |  |  |  |  |  |

The above is cluster#4 of high-calories cereals from KMC\_Clusters Excel spreadsheet (4 clusters). Mapping it to the original data

**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.**