**DM Models 2**

**Task 1 Algorithmic Analysis K-Means Clustering with Real World Dataset**

First, download a simulated dataset: kmeans\_data.zip from Modules->Datasets. Then,

implement the K-means algorithm **from scratch**. K-means algorithm computes the distance of a given data point pair. Replace the distance computation function with Euclidean distance, 1- Cosine similarity, and 1 – the **Generalized** Jarcard similarity (refer to:

https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/jaccard.htm).

Q1: Run K-means clustering with Euclidean, Cosine and Jarcard similarity. Specify K= the

number of categorical values of y (the number of classifications). Compare the SSEs of

Euclidean-K-means, Cosine-K-means, Jarcard-K-means. Which method is better? (10 points)

A1.

After the execution from the below screen shot we can see that cosine has the lowest SSE value which makes it the most accurate cluster predictor, so it is also the best method.

Q2: Compare the accuracies of Euclidean-K-means Cosine-K-means, Jarcard-K-means. First,

label each cluster using the majority vote label of the data points in that cluster. Later, compute the predictive accuracy of Euclidean-K-means, Cosine-K-means, Jarcard-K-means. Which metric is better? (10 points)

A2.

Q3: Set up the same stop criteria: “when there is no change in centroid position OR when the SSE value increases in the next iteration OR when the maximum preset value (e.g., 500, you can set the preset value by yourself) of iteration is complete”, for Euclidean-K-means, Cosine-Kmeans, Jarcard-K-means. Which method requires more iterations and times to converge? (10 points)

A3.

Q4: Compare the SSEs of Euclidean-K-means Cosine-K-means, Jarcard-K-means with respect to the following three terminating conditions: (10 points)

• when there is no change in centroid position

• when the SSE value increases in the next iteration

• when the maximum preset value (e.g., 100) of iteration is complete

A4.

Q5: What are your summary observations or takeaways based on your algorithmic analysis? (5points)

A5.