**INFO 5810 Assignment – 5**

**Section – 1:**

I have added the data requirements to the new excel file and added the new excel file renamed as Tejamanikanta\_assignment\_5 and imported to the RapidMiner app.

As mentioned in the question I have added select attributes to the process tab with a parameter filtered in reviews.

A screenshot of a computer

Description automatically generated

I have added process documents for processing the document with tokenizing, transforming and filtering the words.

A screenshot of a computer

Description automatically generated

Added replace missing values and changed the Nominals to text for easy processing.

A screenshot of a computer

Description automatically generated

Added Clustering of K-means with k value = 4.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Final result of the cluster Model:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Section - 2:**

These are attributes in the dataset.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Added screenshots of the results.

A screenshot of a computer

Description automatically generated

These are clusters in clustering:

A screenshot of a computer

Description automatically generated

{success, ashma, test, sinus, gave, cant} describes the patients having breathing issues. Highest value is 0.056.

A screenshot of a computer

Description automatically generated

Highest value is 0.200 and this is acne issues.

A screenshot of a computer

Description automatically generated

Highest value is 0.063 and cluster describes anemia.

Highest value is 0.094 and cluster describes vomiting.

A screenshot of a computer

Description automatically generated

Changing k value 4 to 5

A screenshot of a computer

Description automatically generated

Highest value is 0.176 and cluster describes acne.

A screenshot of a computer

Description automatically generated

Highest value is 0.071 and cluster describes pregnant.

A screenshot of a computer

Description automatically generated

Highest value is 0.125 and cluster describes skin infections.

A screenshot of a computer

Description automatically generated

Highest value is 0.085 and cluster describes anemia

A screenshot of a computer

Description automatically generated

Highest value is 0.096 and cluster describes nausea.

A screenshot of a computer

Description automatically generated

**Section – 3:**

**Extended Definition and Application of Clustering**

Clustering, a central technique in data analysis, involves the grouping of a set of objects in a way that those in the same cluster exhibit high similarity compared to those in other clusters (Jain, Murty, & Flynn, 1999). It's a major method in machine learning and data mining for statistical data analysis, applicable across diverse fields like image and pattern recognition, bioinformatics, and market research.

The primary objective of clustering is to categorize datasets into groups where members of each group are as similar as possible and as different as possible from members in other groups. This process often involves measuring similarity in terms of various metrics like Euclidean or Manhattan distance, depending on the nature of the data (Xu & Wunsch, 2005).

Different types of clustering algorithms exist, including hierarchical, partitioning, density-based, and grid-based clustering. Each type has specific advantages and limitations, and the choice depends on the data characteristics and the analysis's goals.

In practical applications, clustering is used extensively. For example, in marketing, it helps in segmenting customers based on purchasing behavior, demographics, and interests, thereby enabling tailored marketing strategies. In bioinformatics, clustering is instrumental in classifying genes and proteins with similar functions. It's also crucial in areas like environmental science for analyzing and categorizing different ecological regions based on various environmental parameters.

**References:**

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys (CSUR)*, *31*(3), 264–323.

Xu, R., & Wunsch, D., 2nd. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. <https://doi.org/10.1109/TNN.2005.845141>

**RapidMiner Operators and Their Usage in clustering:**

**Select Attributes:**

Definition and Usage: The 'Select Attributes' operator in RapidMiner is used for selecting and manipulating dataset attributes. It's essential for identifying variables relevant to the clustering process, such as selecting specific customer demographics in market segmentation (RapidMiner, n.d.).

**Process Documents from Data:**

Definition and Usage: This operator is critical in text mining, converting raw text into a structured format suitable for analysis. It's used in clustering tasks involving textual data to extract features essential for analysis, such as word frequency or term importance (RapidMiner, n.d.).

**K-Means Clustering:**

Definition and Usage: The K-Means clustering operator in RapidMiner segments a dataset into K distinct subgroups. The choice of the 'k' value is a strategic decision influenced by the nature of the dataset and the specific goals of the analysis. This operator is particularly useful in scenarios where there is a need to identify distinct groupings within a dataset, such as customer segmentation based on purchasing patterns (RapidMiner, n.d.).

**Cluster Distance Performance:**

Definition and Usage: This operator evaluates the effectiveness of clustering. It measures the quality of the clusters formed in terms of their compactness and the separation between them. This evaluation is crucial in guiding adjustments to the clustering process, such as tweaking the 'k' value or modifying other parameters to optimize the results (RapidMiner, n.d.)