**Data Mining (CSE 5334)**

**Assignment 4 (Clustering Report)**

**Student details:**

* Kuldip Rameshbhai Savaliya – 1001832000
* Shivani Manojkumar Panchiwala – 1001982478
* Meghaben Ghanshyambhai Patel – 1002006777

**Clustering:**

An abstract object is clustered into classes based on their similarities. There are several types of data objects that can be considered one group. During cluster analysis, the data is first divided into groups based on similarity, then the labels are assigned to the groups. Overclassification by clustering has multiple advantages, including the ability to adapt to change and to distinguish between different groups with useful features.

**Task 1**

**Import necessary packages and library.**

**Text

Description automatically generated**

**Read datsaet.**

**Graphical user interface, application

Description automatically generated**

**Checking for null values.**

**Table

Description automatically generated**

**Elbow method for k =1 to 10.**

The **Elbow Method** is one of the most popular methods to determine the optimal value of k.

**Chart, line chart

Description automatically generated**

From the above graph, In elbow method we can see that the Number of clusters vary from k=1 to 10 and for every value of K SSE is vary according to K. If the Number of clusters increase the SSE is decrease and we can see that for k=4 graph changes rapidly and make an elbow shape.

**Visulization for K-means clusteing**

**K=2**

**Actual Training Labels Vs Predicted Training Labels**

`Chart, scatter chart

Description automatically generated

From the Actual Train Labels vs Predicted Train Lables graph, we can see that red marks are the centroids of the clusters. Model divides the data into two clusters for k=2 which are in the subplot.

**Actual Testing Labels Vs Predicted Testing Labels**

**Chart, scatter chart

Description automatically generated**

**Confusion Matrix**

**Graphical user interface, text, application

Description automatically generated**

**Task 2**

**Hierarchical Agglomerative Clustering**

Hierarchical Agglomerative Clustering is also known as the bottom-up approach. We begin by separating each object into its own group. In the process, close objects or groups are constantly merged. In this way, it continues until all groups are merged into one or until the termination condition is met.

**Calculate pairwise distance**

**Text, letter

Description automatically generated**

**F-1 score for complete and average linkage**

**Text

Description automatically generated**

**Visualization for Hierarchical Agglomerative Clustering**

**Train actual Vs predicted**

**Chart, scatter chart

Description automatically generated**

From the above graph displays the Train Actual Vs Predicted labels for train data. The model divides the data into two clusters for k=2 which are display in the subplot.

**Compare K-means and Hierarchical Agglomerative clustering.**

**Chart, scatter chart

Description automatically generated**

From the above graph we can see that the subplots of actual vs agglomerative predicted values vs kmeans predicted values.

**Confusion matrix for Kmeans and Agglomerative.**

**Graphical user interface, text

Description automatically generated**