**Data Mining (CSE 5334)**

**Assignment 4 (College dataset 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.**

**Text

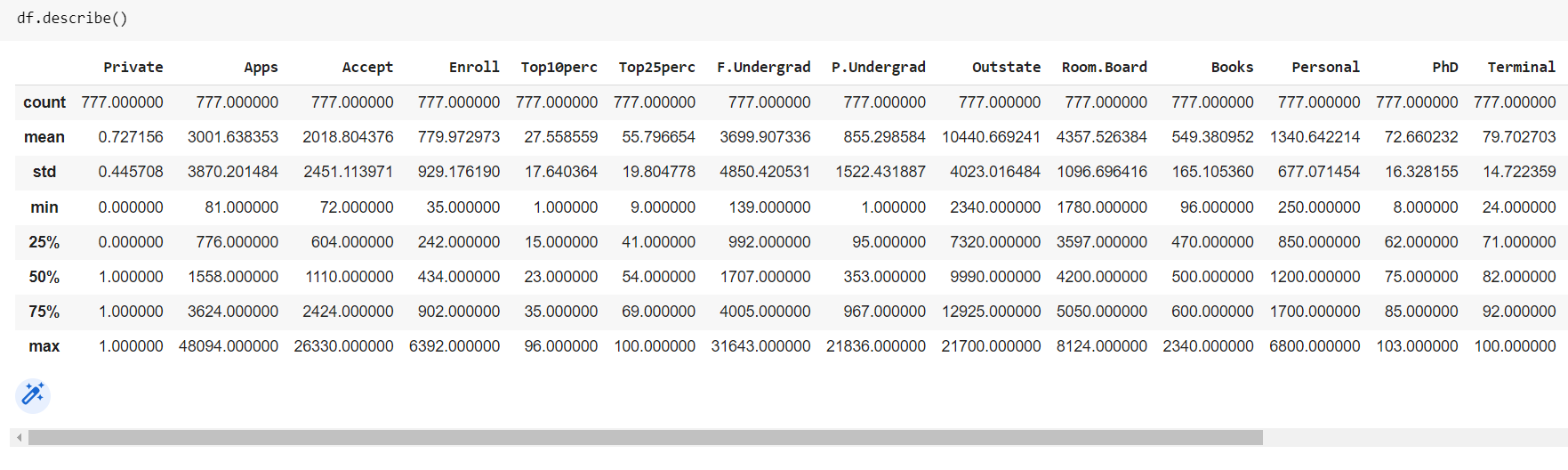
Description automatically generated with low confidence**

**Checking for null values.**

**A picture containing table

Description automatically generated**

**After processing dataset**

****

**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

In the Elbow method we are varying the number of clusters K and for each value of K sum ofsquared errors between each point and centroid in a cluster. Here in the above plot, we can see that as the number of clusters increases the SSE starts decreasing.

Here we took the Optimal value of K as 3 because before 3 variation changes rapidly but after 3 it slows down leading to an elbow formation in the curve.

**Visualization for K-Means Clustering**

**For k =2**

**Visualizing the Predicted training samples Vs Actual training samples**

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Visualizing the Predicted Test labels Vs Actual Test labels**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The above graph is the scatter plot of predicted labels versus actual labels for testing data where I have run Kmeans on the whole dataset and assigned the actual labels dividing the data in three clusters for k=2. Moreover, I have then predicted the labels. Black Cross are the centroids of the clusters respectively. The model divides the data into three clusters for k=3 which isdisplayed in the first subplot of graph.

**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.

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

**Text

Description automatically generated**

**Visualization for Hierarchical Agglomerative Clustering**

**Visualize using the best model Vs Actual Test labels**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

**Compare K-Means Clustering and Hierarchical Agglomerative Clustering**

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

The above graph is the scatter plot of predicted labels by kmean and Agglomerative clustering versus actual labels for traning data. The actual labels are divided in two clusters. Moreover,

I have then predicted the labels. Black Cross are the centroids of the clusters respectively.

The model divides the data into two clusters for kmeans no. of clusters 2 &Cosine Complete method which is displayed in the first two subplot of graph.

**Confusion matrix for Kmeans and Agglomerative.**

**A picture containing table

Description automatically generated**

Graphical user interface, text, application, email

Description automatically generated