**ASSIGNMENT 6**

Student Name: Uday Kumar Kuppam

Student ID: 700739768

CRN: 13469

GitHub Link: <https://github.com/KUPPAM-700739768/ASSIGNMENT6_ML_700739768>

Video Link: <https://youtu.be/bkMkdmAZJ9w>

1. (Provide only mathematical solutions for this question) Six points with the following attributes are given, calculate and find out clustering representations and dendrogram using Single, complete, and average link proximity functions in hierarchical clustering technique.

Question1 Document -



🡪Import the required libraries

Graphical user interface, text, application, email

Description automatically generated

Assigning the data points to variables

Graphical user interface, text

Description automatically generated

single link of hierarchical clustering, the proximity of two clusters is defined as the minimum of the distance (maximum of the similarity) between any two points in the two different clusters. Figure dendrogram below shows the result of applying the single link technique to our example data set of six points

Chart, box and whisker chart

Description automatically generated

the complete link of hierarchical clustering, the proximity of two clusters is defined as the maximum distance (minimum of the similarity) between any two points in the two different clusters. Figure dendrogram below shows the results of applying MAX to the sample data set of six points.

Chart

Description automatically generated

The group average version of hierarchical clustering, the proximity of two clusters is defined as the average pairwise proximity among all pairs of points in the different clusters. This is an intermediate approach between the single and complete link approaches.

Chart, box and whisker chart

Description automatically generated

* 1. 2) Use CC\_GENERAL.csv given in the folder and apply: a) Preprocess the data by removing the categorical column and filling the missing values.
  2. b) Apply StandardScaler() and normalize() functions to scale and normalize raw input data.
  3. c) Use PCA with K=2 to reduce the input dimensions to two features.
  4. d) Apply Agglomerative Clustering with k=2,3,4 and 5 on reduced features and visualize result for each k value using scatter plot.
  5. e) Evaluate different variations using Silhouette Scores and Visualize results with a bar chart.

Importing the required libraries

Graphical user interface, text, application, email

Description automatically generated

Reading the CSV file using read method

Graphical user interface, text

Description automatically generated with medium confidence

Describing the data features using describe method

Graphical user interface, table

Description automatically generated

Table

Description automatically generated

When categorical columns, such as string or numerical ones, have missing values, the most prevalent category might be used to fill the gaps. Initially check for Null dataset

Table

Description automatically generated

Fill null values and check for any null values

Table

Description automatically generated

Gradient correlation

Graphical user interface, table

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

* 1. reduce the input dimensions to two features and visualize the data in terms of PC1 and PC2

Chart, bubble chart

Description automatically generated

Chart

Description automatically generated

Agglomerative Clustering 3 on reduced features and visualize result for each k value using scatter plot.

Chart

Description automatically generated

Agglomerative Clustering 4 on reduced features and visualize result for each k value using scatter plot.

Chart

Description automatically generated

Agglomerative Clustering 5 on reduced features and visualize result for each k value using scatter plot.

Chart

Description automatically generated

Graphical user interface, application

Description automatically generated