**ML Assignment-6**

**Sai Sree Chikkala**

**#700726877**

**GitHub Link:**

**Video Link:**

**Question-1**

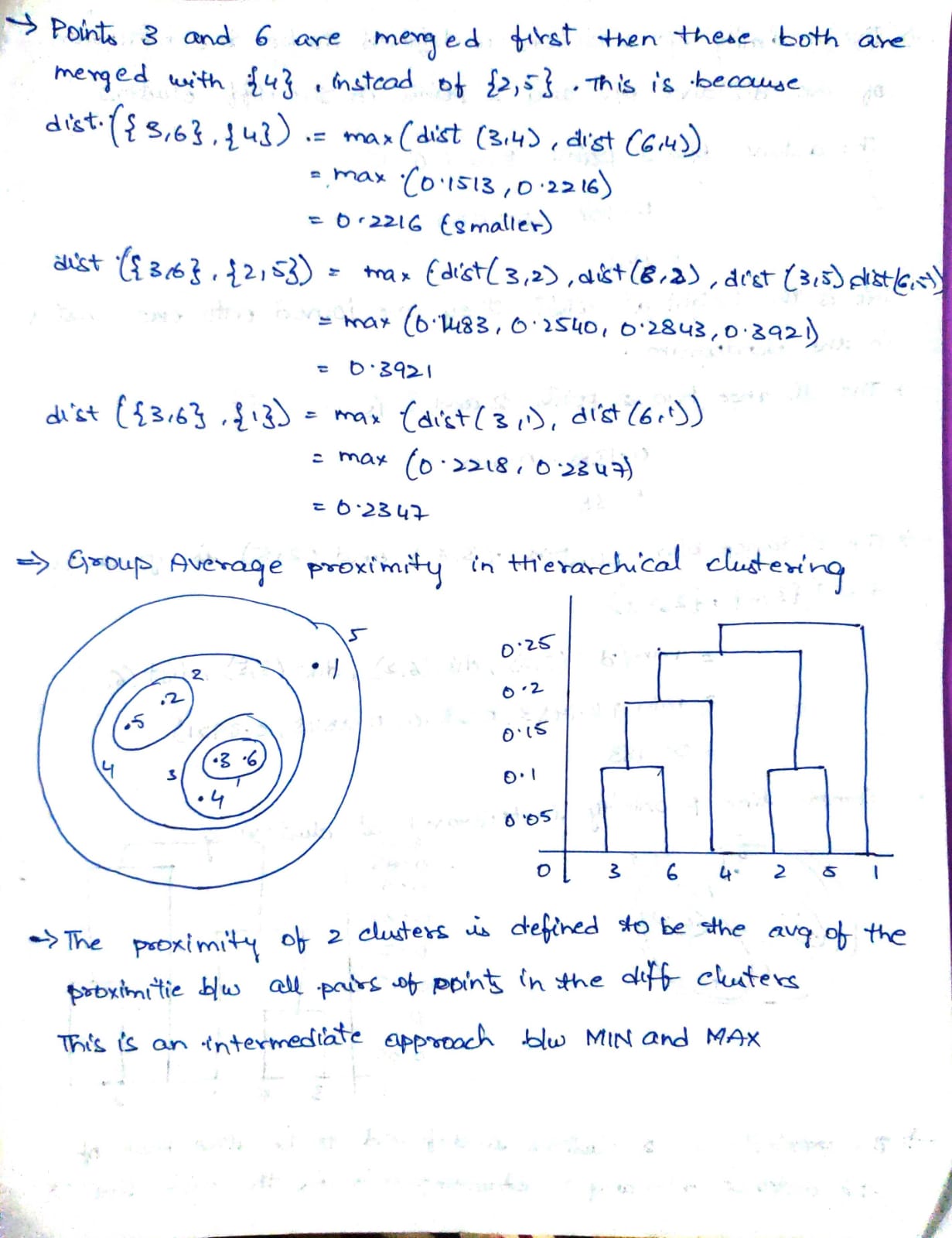
1. Six points with the following attributes are given, calculate and find out clustering representations and dendrogram using Single, complete, and average link proximity function in hierarchical clustering technique.

Table

Description automatically generated

Diagram

Description automatically generated



A piece of paper with writing

Description automatically generated with medium confidence

**Question-2**

1. Use CC\_GENERAL.csv given in the folder and apply:

**Graphical user interface, text, application

Description automatically generated**

1. Preprocess the data by removing the categorical column and filling the missing values.

* One approach to fill these missing values can be to replace them with the most common or occurring class. We can do this by taking the index of the most common class which can be determined by using value counts() method

Graphical user interface

Description automatically generated

1. Apply StandardScaler() and normalize() functions to scale and normalize raw input data.

* Normalization is the process of scaling individual samples to have unit norm.
* This process can be useful if you plan to use a quadratic form such as the dot-product or any other kernel to quantify the similarity of any pair of samples.

X\_normalized = preprocessing.normalize(X\_scaled\_df)

* Converting the numpy array into a pandas DataFrame

X\_normalized = pd.DataFrame(X\_normalized)

Graphical user interface, text, application

Description automatically generated

1. Use PCA with K=2 to reduce the input dimensions to two features

* Principal Component Analysis (PCA) is an unsupervised linear transformation technique that is widely used across different fields, most prominently for feature extraction and dimensionality reduction.
* PCA helps us to identify patterns in data based on the correlation between features. In a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

Graphical user interface, text, application

Description automatically generated

1. Apply Agglomerative Clustering with k=2,3,4 and 5 on reduced features and visualize result for each k value using scatter plot.

* Agglomerative Clustering is one of the most common hierarchical clustering techniques.

Chart, bubble chart

Description automatically generated

**Building and visualizing the different clustering models for different values of k a) k = 2**

Chart, scatter chart

Description automatically generated

**Building and visualizing the different clustering models for different values of k a) k = 3**

Chart

Description automatically generated

**Building and visualizing the different clustering models for different values of k a) k = 4**

Chart

Description automatically generated

**Building and visualizing the different clustering models for different values of k a) k = 5**

Chart

Description automatically generated

1. Evaluate different variations using Silhouette Scores and Visualize results with a bar chart.

Graphical user interface, application

Description automatically generated