Subject: 21DS602

Lab Session: 10

Lab Session Date: 21/12/2022

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Report should be submitted to TurnItIn. Once done, please submit your assignments in Teams.

Main Section (Mandatory):

Please use the data associated with your own project.

Refer:

- **https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html**
- **https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html**
- **https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html**
- **https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html#**

Main Section (Mandatory):

A1. From scatter plot of various feature combinations, study the cluster patterns of your dataset. Merge the data from various classes to perform clustering exercise.

A2. Merge your train & test sets and remove the class labels. Use k-means algorithm with k = 3 or 5 (based on your dataset) to form the clusters.

A2. Determine the ideal k value for your dataset. Determine the clusters for a range of k ϵ [1,31]. Use elbow method to determine the ideal value of k based on average Euclidean distance from cluster center.

A3. Use Agglomerative Clustering for hierarchical clustering of your data.

A4. Plot the dendrogram to visualize the clusters.

A5. Study the various clustering techniques available under sklearn package. Understand the similarities and differences between them.

Optional Section

O1. Study the following papers to get a comprehensive idea about various clustering techniques. Try out some of the techniques described in these papers on your dataset.

- https://www.ijert.org/research/a-comparative-study-of-data-clustering-techniques-IJERTV2IS60712.pdf
- http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.433.3348&rep=rep1&type=pdf
- https://gvpress.com/journals/IJBSBT/vol5_no5/25.pdf
- https://www.sciencedirect.com/science/article/abs/pii/S0167739X97000186

Report Assignment:

1. Describe how you'd be able to determine the number of clusters available from a volume of student reports available in text format (similar to the lab exercise 2 performed for plagiarism detection). [2]
2. With study of the cluster patterns, propose methods by which the number of observations may be reduced from the training dataset. [2]