Foundation Of Machine Learning

Assignment-02

Preprocessing Activities:

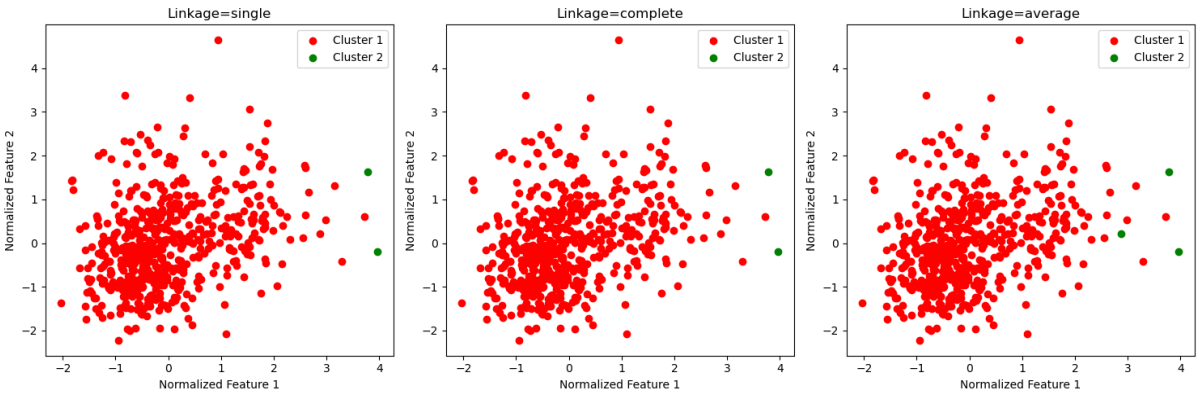
Data Cleaning: No null values were found in the dataset, indicating that there are no missing data points. This is a significant achievement as missing data can lead to biassed or unreliable results in subsequent analyses.

Data Formatting: All data entries have been reviewed and confirmed to be in the correct format. This step ensures consistency and compatibility with the chosen analysis or modelling techniques.

Hierarchical Agglomerative Clustering:

Hierarchical clustering is a data analysis technique that groups similar data points into clusters or a hierarchical structure. It can be agglomerative (bottom-up) or divisive (top-down). In agglomerative clustering, it starts with individual data points as clusters and merges them based on proximity using methods like single, complete, or average linkage. This process continues until all data points belong to a single cluster. A dendrogram visualizes the hierarchy, and cutting it at a certain level yields clusters. Hierarchical clustering has the advantage of not requiring the number of clusters to be predefined and provides a hierarchy view of data. However, it can be computationally expensive and sensitive to noise

Output for respective linkage : here Normalised Feature 2 indicates texture_mean and feature 1 indicates radius_mean



Cluster 1: 567 points
Cluster 2: 2 points

Cluster 1: 567 points
Cluster 2: 2 points

Cluster1: 566
Cluster2:3

DBSCAN:

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm used in data analysis and machine learning. It groups data points based on their density in the feature space, making it effective for finding clusters with irregular shapes and varying densities. DBSCAN identifies core points (dense central points), border points (within a certain distance of core points), and noise points (isolated points). It doesn't require specifying the number of clusters in advance, making it flexible. However, it may require parameter tuning and can be computationally intensive for large datasets. DBSCAN is valuable for tasks like spatial data analysis, image segmentation, and anomaly detection.

The **outputs of the dataset** are visually represented in the adjacent figure, which provides a graphical representation of the data's patterns, clusters, or any relevant insights.