Name: Sakshi

Roll No. : 21CSU419

Semester/Section: 6th DS-B DS5

Submitted To: Ms. Ankita

1. Project Title: Cluster Analysis Using Different Clustering Algorithms

2. Dataset Description and Link: The project utilizes datasets generated from the sklearn datasets library. The datasets include blobs, moons, and circles, each containing 1500 samples with 2 features. Artificial noises have been added to these datasets to simulate real-world scenarios. The datasets can be accessed through the sklearn.datasets module.

3. Project Objectives:

- To compare the performance of different clustering algorithms, including KMeans, KMedoids, AGNES, BIRCH, and DBSCAN, on various datasets.

- To evaluate the effectiveness of these algorithms in clustering datasets with different shapes, sizes, and densities.

- To determine which clustering algorithms are more suitable for specific types of datasets based on their characteristics.

- To explore the impact of preprocessing techniques and dimension reduction on clustering performance.

4. Workflow and Summary: The project follows a structured workflow:

1. **Data Acquisition:** Acquiring datasets from the sklearn datasets library, including blobs, moons, and circles, each with 1500 samples and 2 features.

2. **Exploratory Data Analysis (EDA):** Conducting EDA to understand the characteristics of the datasets and visualize potential clusters.

3. **Data Preprocessing:** Preprocessing the datasets, including feature scaling, to prepare them for clustering algorithms.

4. **Dimension Reduction:** Applying dimension reduction techniques to reduce the number of features while retaining information.

5. **Clustering Algorithms:** Applying KMeans, KMedoids, AGNES, BIRCH, and DBSCAN algorithms to the datasets.

6. **Performance Evaluation:** Evaluating the performance of the algorithms using metrics such as Silhouette score, Adjusted Rand Index (ARI) score, and Normalized Mutual Information (NMI) score.

7. **Results and Conclusion:** Analyzing the results to determine the most effective clustering algorithms for different dataset characteristics and summarizing the findings.

In summary, the project aims to compare and evaluate different clustering algorithms on various datasets to understand their effectiveness in clustering data with different characteristics.