**Machine Learning**

**CSL313**

Project Report



Faculty name: Ms. Ankita Student name: Sakshi

Roll No.: 21CSU419

Semester: 6th

Group: DS-B DS-5

**Department of Computer Science and Engineering**

**The NorthCap University, Gurugram- 122001, India**

**Session 2023-24**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **S.No** |  | **Page No.** |
| **1.** | **Project Description** | **iii** |
| **2.** | **Problem Statement** | **iii** |
| **3.** | **Analysis**  **3.1 Hardware Requirements**  **3.2 Software Requirements** | **iv** |
| **4.** | **Design**  **4.1 Data/Input Output Description:**  **4.2 Algorithmic Approach / Algorithm / DFD / ER diagram/Program Steps** | **v** |
| **5.** | **Implementation and Testing (stage/module wise)** | **vii** |
| **6.** | **Output (Screenshots)** | **ix** |
| **7.** | **Conclusion and Future Scope** | **xii** |

**Project Description**

The project focuses on conducting cluster analysis using different clustering algorithms on various datasets. It aims to compare the performance of algorithms such as KMeans, KMedoids, AGNES, BIRCH, and DBSCAN on datasets with different shapes, sizes, and densities. The project also explores the impact of preprocessing techniques and dimension reduction on clustering performance.

**Problem Statement**

The objective is to determine which clustering algorithms are more suitable for specific types of datasets based on their characteristics. This involves evaluating the effectiveness of the algorithms in clustering datasets with varying complexities and identifying the most appropriate algorithms for different clustering scenarios.

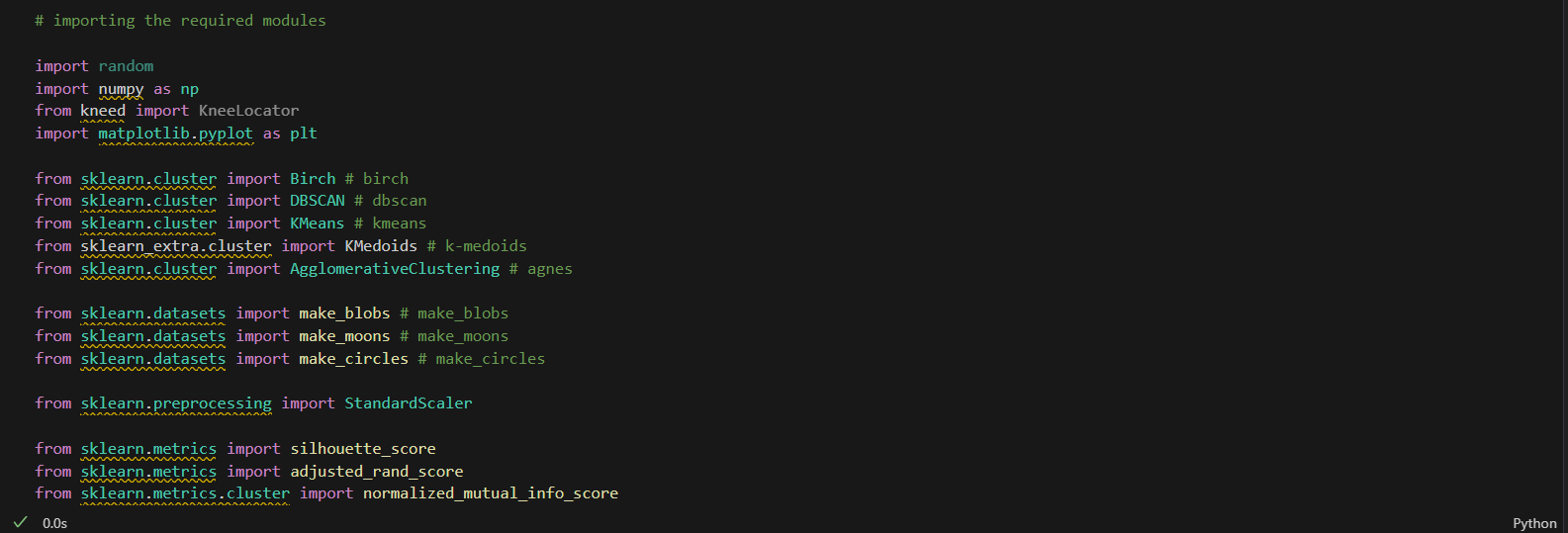
**Analysis**

**3.1 Hardware Requirements**

* A computer with sufficient processing power and memory to handle dataset manipulation and algorithm execution.

**3.2 Software Requirements**

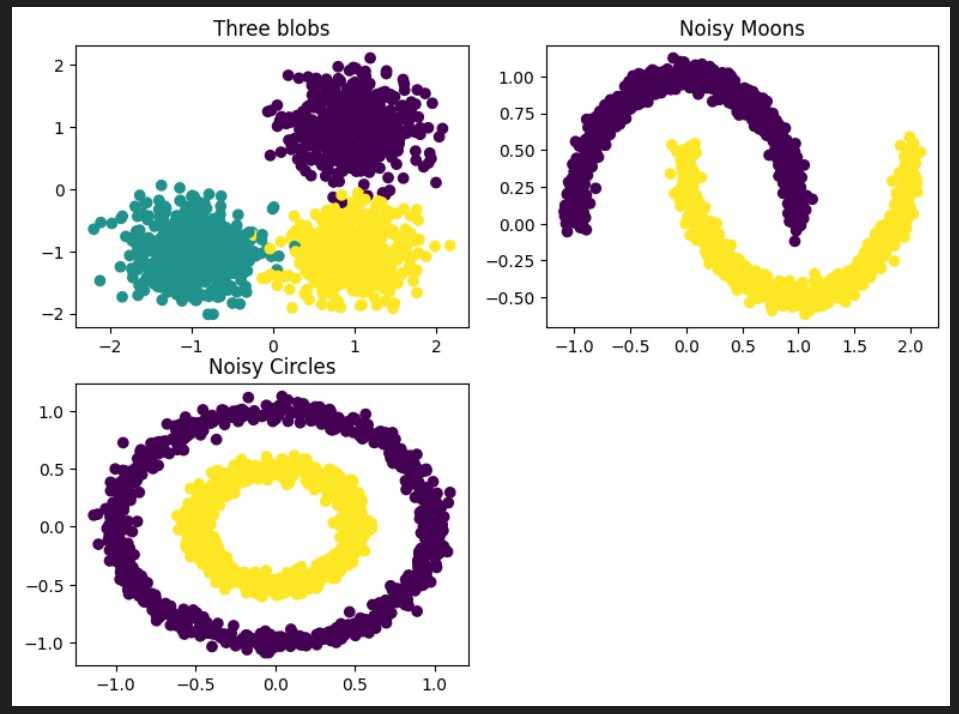
* Python programming language
* Jupyter Notebook or similar IDE for code development
* Scikit-learn library for clustering algorithms
* Matplotlib and Seaborn for data visualization
* Pandas for data manipulation
* Numpy for numerical computations

****

**Design**

**4.1 Data/Input Output Description:**

* **Input**: Datasets generated from the sklearn datasets library, including blobs, moons, and circles, each containing 1500 samples with 2 features.



* **Output**: Performance metrics such as Silhouette score, Adjusted Rand Index (ARI) score, and Normalized Mutual Information (NMI) score, to evaluate the clustering algorithms.

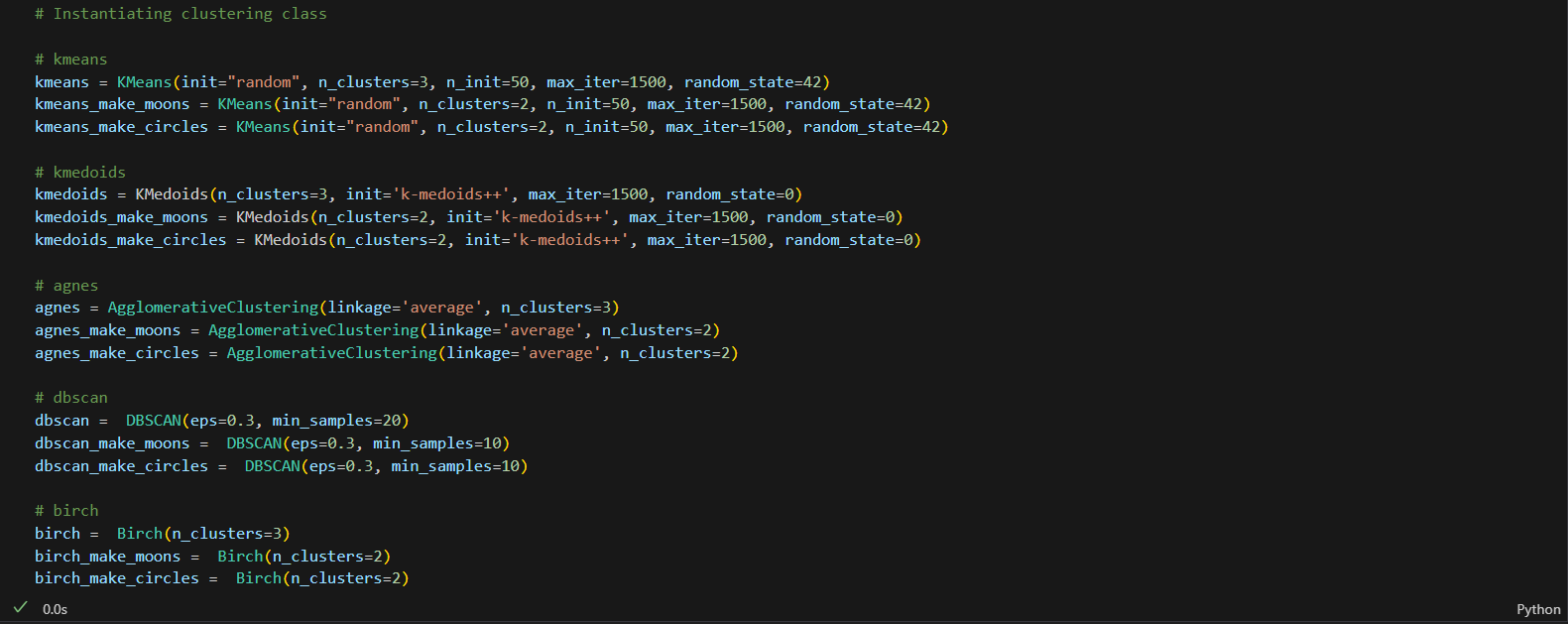
**4.2 Algorithmic Approach / Algorithm:**

1. **Data Acquisition**: Acquiring datasets from sklearn datasets library.
2. **Exploratory Data Analysis (EDA)**: Understanding dataset characteristics and visualizing potential clusters.
3. **Data Preprocessing:** Preprocessing datasets, including feature scaling.
4. **Dimension Reduction:** Applying dimension reduction techniques to reduce the number of features.
5. **Clustering Algorithms**: Applying KMeans, KMedoids, AGNES, BIRCH, and DBSCAN algorithms.
6. **Performance Evaluation:** Different evaluating metrics are used to show the performance of different algorithms on different datasets. Silhouette score, Adjusted Rand Index (ARI) score and Normalized Mutual Information (NMI) score have been used. Short description about these metrics is given below:

* Silhouette score’s value ranges from -1 to 1. 1 means clusters are well apart from each other and clearly distinguished. 0 Means clusters are indifferent or the distance between clusters is not significant and -1 Means clusters are assigned in the wrong way.
* ARI score’s value ranges from 0 to 1. 0 means random labelling and 1 means perfect labelling.
* NMI score’s value also ranges from 0 to 1 and have same meaning like ARI.

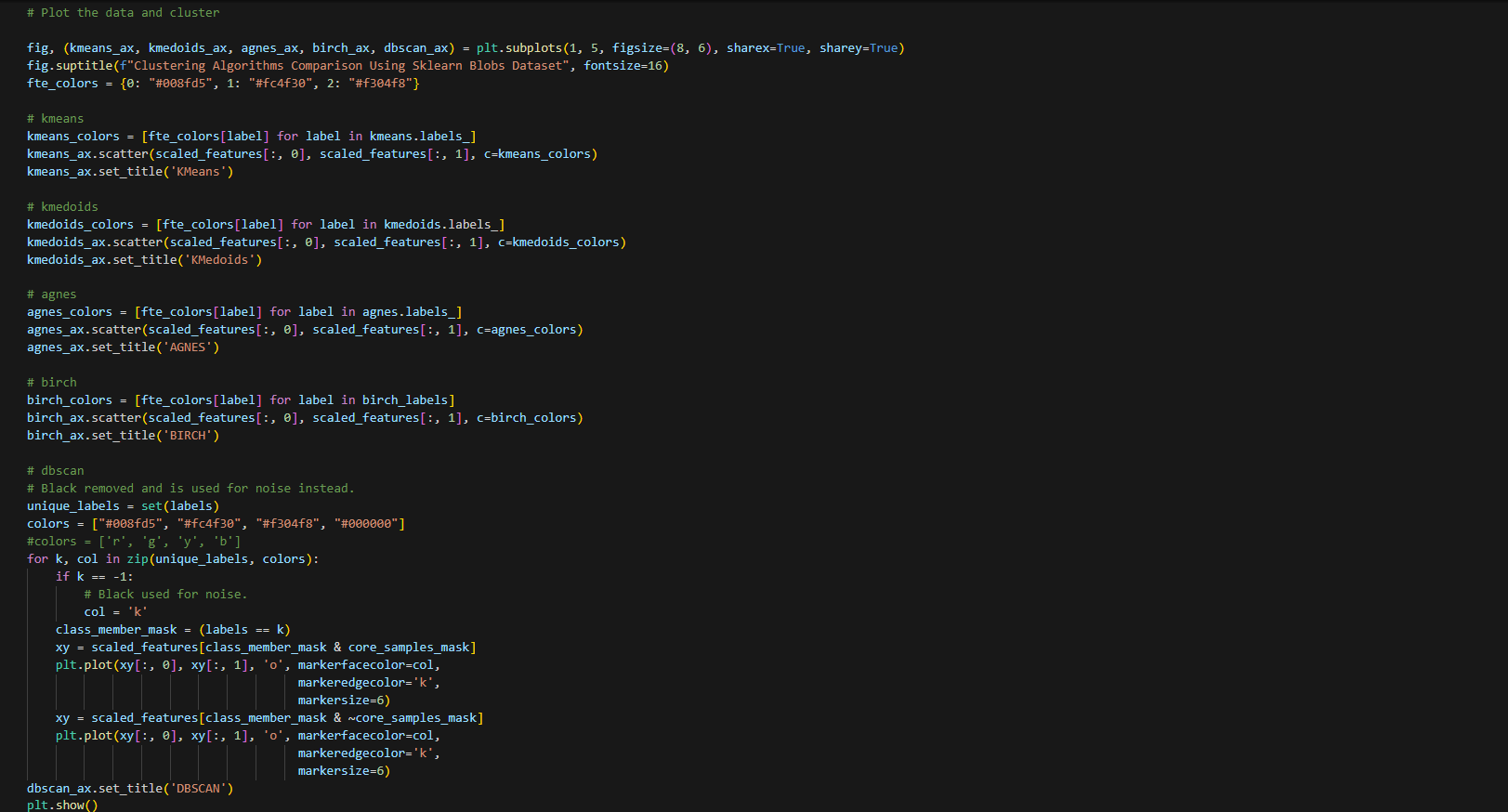
**Implementation and Testing (stage/module wise)**

* Implementing the above algorithmic steps in Python using appropriate libraries.

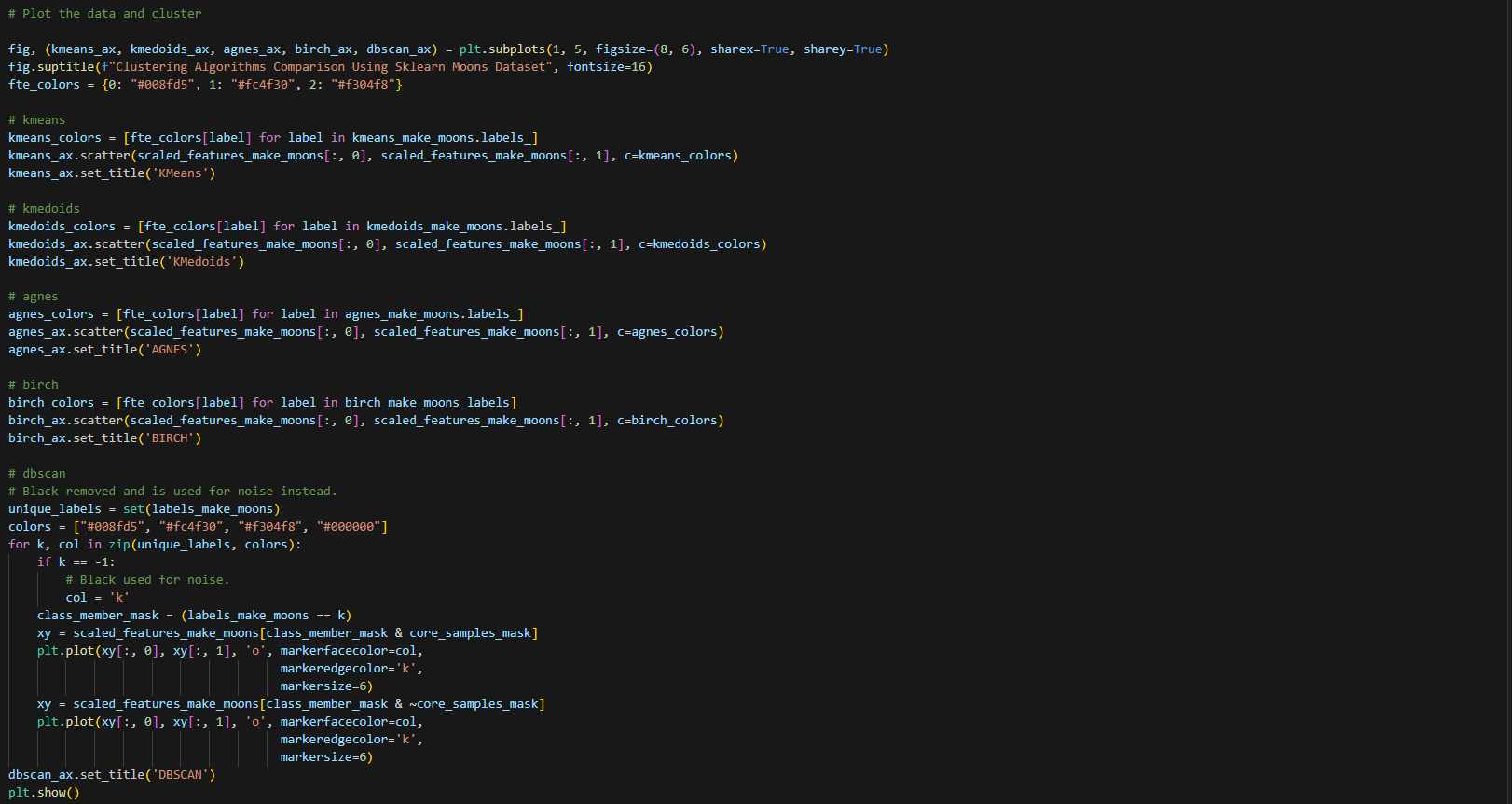


* Testing the implementation on the datasets and verifying the results.

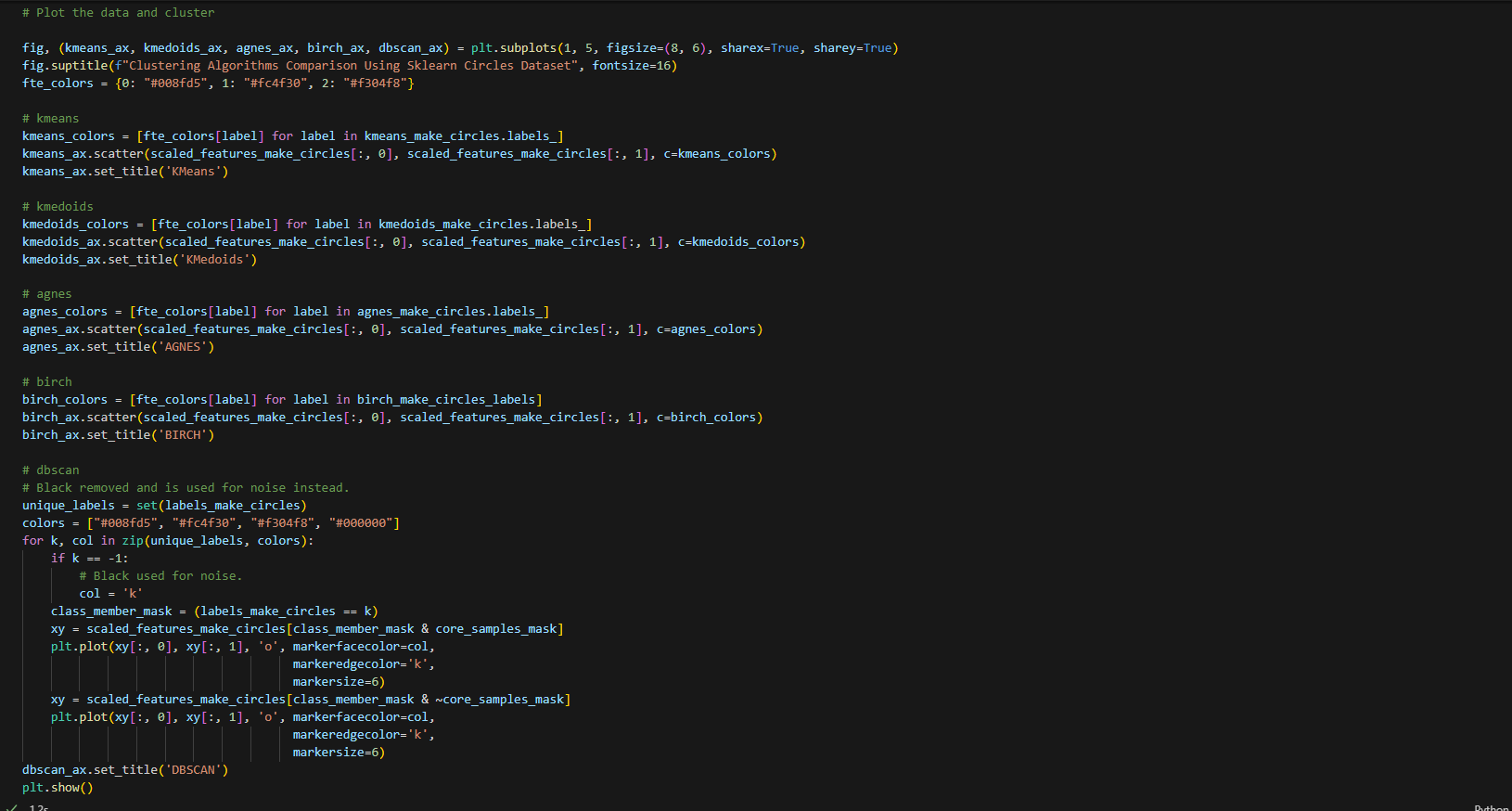
Blob dataset:

****

Moons Dataset:

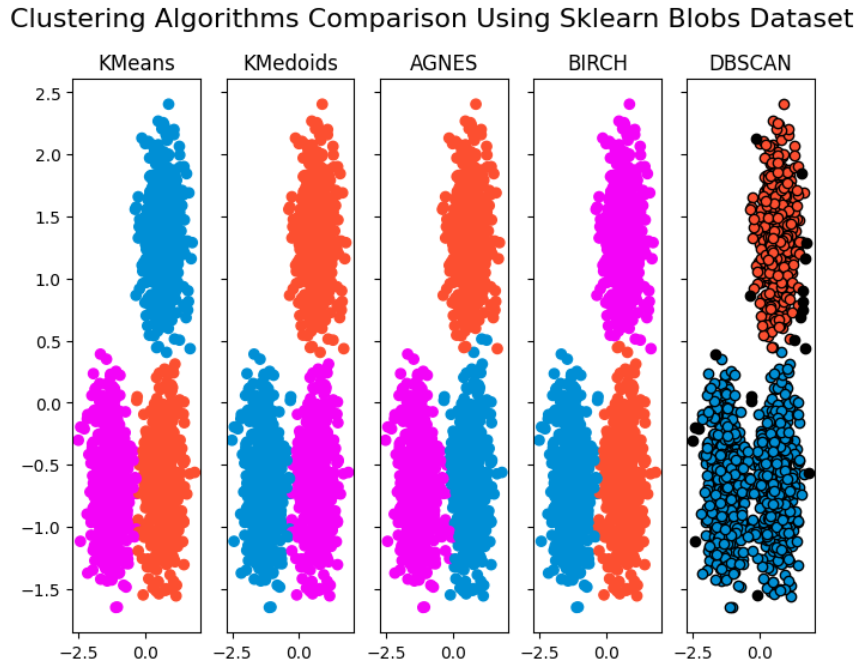


Circles Dataset:

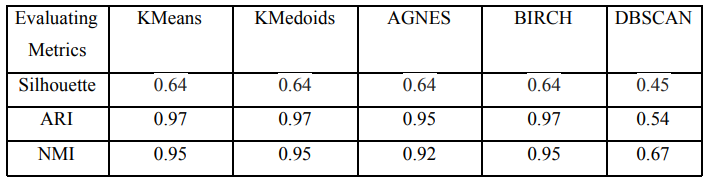


**Output (Screenshots)**

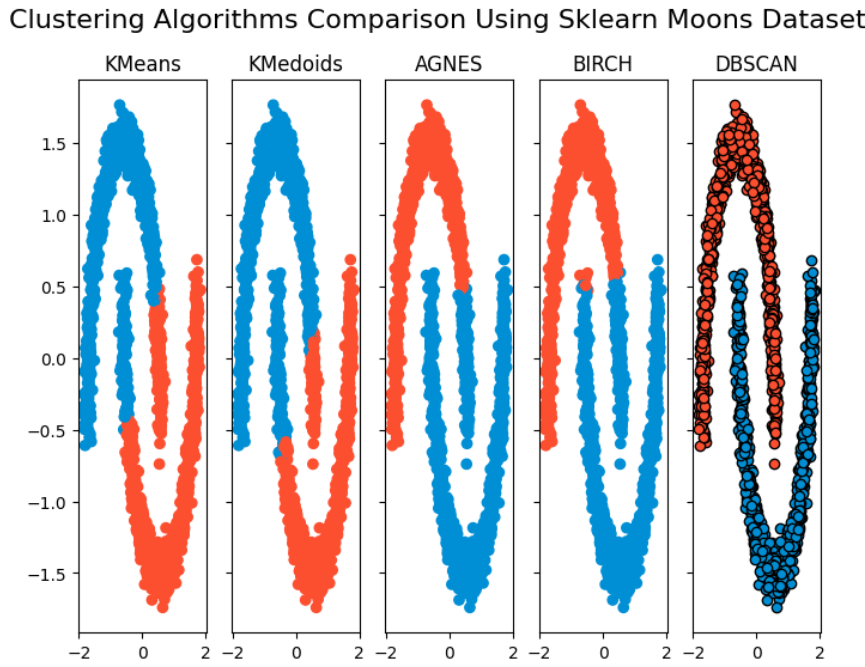
**Sklearn Blobs Dataset:**

****

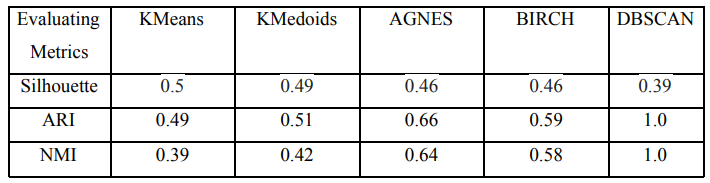
ARI score and NMI score show that all algorithms perform well on this dataset but silhouette score fails to show it. However, the table shows that all algorithms perform well on this dataset.

****

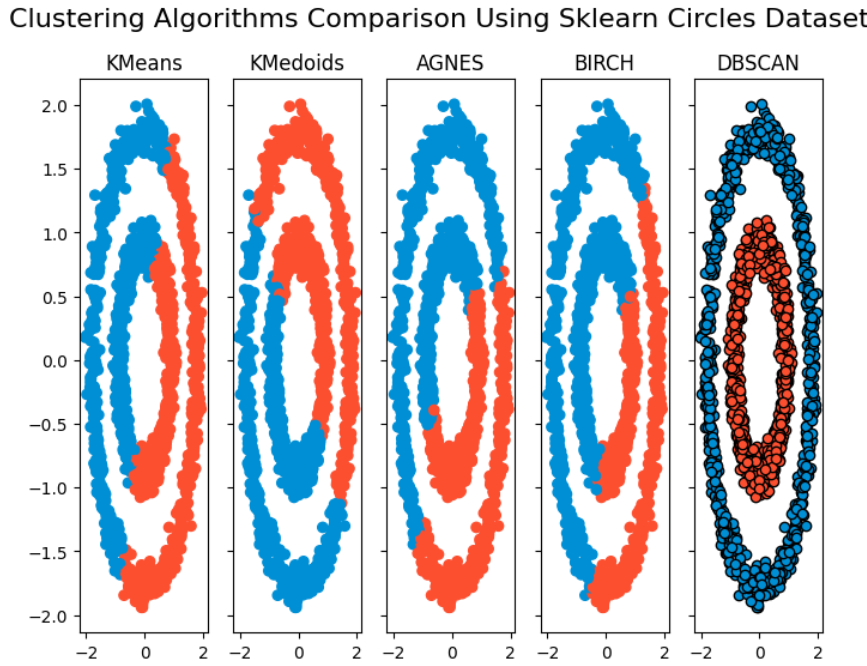
**Sklearn Moons Dataset:**

****

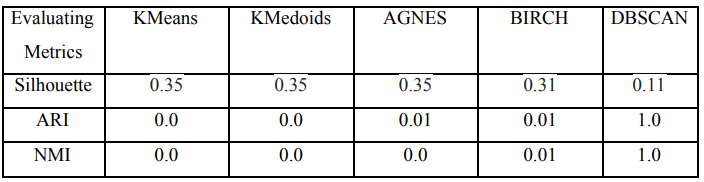
All the metrics show that DBSCAN performs very well on this complex clustering shapes and other algorithms fail.

****

**Sklearn Circles Dataset:**

****

It shows DBSCAN performs very well and other algorithms fail to cluster this dataset.

****

**Conclusion and Future Scope**

The project concludes by identifying the most effective clustering algorithms for different dataset characteristics. Future scope includes exploring other clustering algorithms, evaluating performance on more diverse datasets, and investigating the impact of different preprocessing techniques and parameter settings on clustering results.