CS564 Foundations of Machine Learning

ASSIGNMENT 2

Nikunj Pansari

2211MC21

Problem Statement:

- The assignment targets to implement DB Scan algorithms to cluster the 3 datasets with blob, moon and circle structures.

- Apply DB Scan Clustering on the 3 datasets and compare its result with the K-Means algorithm.

```
## Displaying the blob data
dc_blobs.head()
```

	X1	X2
0	4.645333	6.822294
1	4.784032	6.422883
2	-5.851786	5.774331
3	-7.459592	6.456415
4	4.918911	6.961479

```
## Displaying the circles data
dc_circles.head()
```

	X1	X2
0	-0.348677	0.010157
1	-0.176587	-0.954283
2	0.301703	-0.113045
3	-0.782889	-0.719468
4	-0.733280	-0.757354

```
## Displaying the moons data
dc_moons.head()
```

	X_1	X_2
0	0.674362	-0.444625
1	1.547129	-0.239796
2	1.601930	-0.230792
3	0.014563	0.449752
4	1.503476	-0.389164

1. Then, the Data Cleaning take place where the data is checked for NULL and NAN values.

Data Cleaning (Checking NULL or NAN values)

```
## Checking Null Values for all  the data
print('NULL Value for blob : ',dc_blobs.isnull().sum().sum())
print('NULL Value for Circles : ',dc_circles.isnull().sum().sum())
print('NULL Value for moon : ',dc_moons.isnull().sum().sum())
## Checking NAN Values for al the data
print('NAN Value for blob :',dc_blobs.isnull().sum().sum())
print('NAN Value for circles :',dc_circles.isnull().sum().sum())
print('NAN Value for moons:',dc_moons.isnull().sum().sum())
```

```
NULL Value for blob :  0
NULL Value for Circles :  0
NULL Value for moon :  0
NAN Value for blob : 0
NAN Value for circles : 0
NAN Value for moons: 0
```

2. Data Scaling is done, to get of all the attributes within the same range.

```
scaled_dc_blobs
```

	X1	X2
0	0.762231	0.842572
1	0.769209	0.821952
2	0.234057	0.788471
3	0.153158	0.823683
4	0.775996	0.849757
...	...	...
1495	0.572908	0.140031
1496	0.576138	0.115858
1497	0.688508	0.831543
1498	0.802947	0.823933
1499	0.137578	0.852489

1500 rows × 2 columns

```
scaled_dc_circles
```

	X1	X2
0	0.360103	0.484668
1	0.436480	0.046644
2	0.648755	0.428713
3	0.167392	0.153292
4	0.189409	0.136085
...	...	...
1495	0.382814	0.515405
1496	0.551493	0.357260
1497	0.638915	0.430353
1498	0.527028	0.359749
1499	0.418663	0.563172

1500 rows × 2 columns

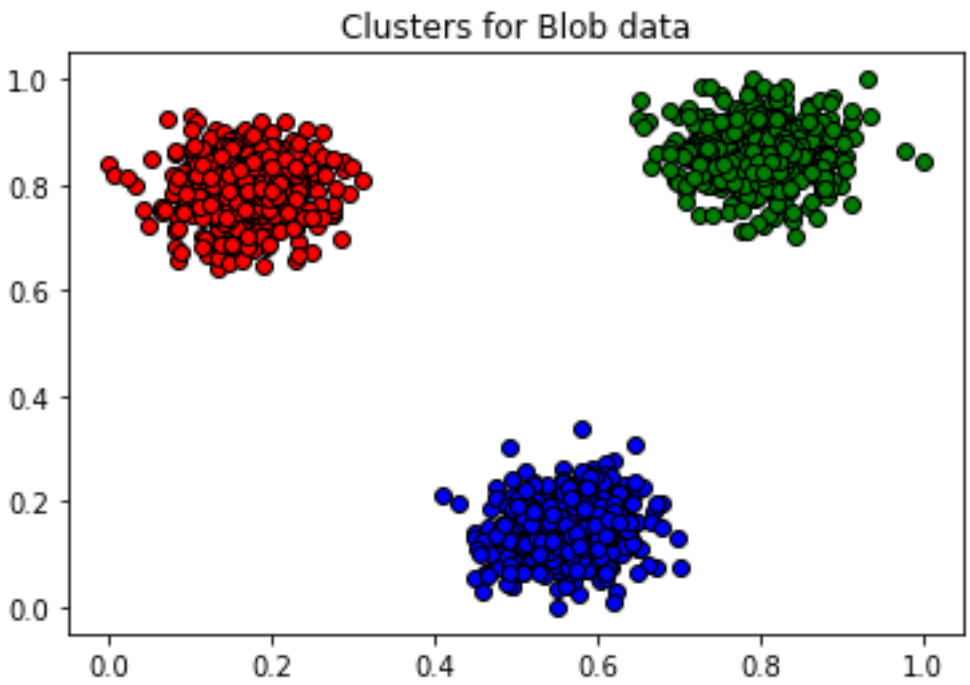
scaled_dc_moon

	X_1	X_2
0	0.557050	0.105134
1	0.829157	0.222315
2	0.846242	0.227466
3	0.351340	0.616799
4	0.815547	0.136863
...	...	...
1495	0.957051	0.466586
1496	0.646851	0.579356
1497	0.109260	0.692265
1498	0.909198	0.270994
1499	0.410185	0.404723

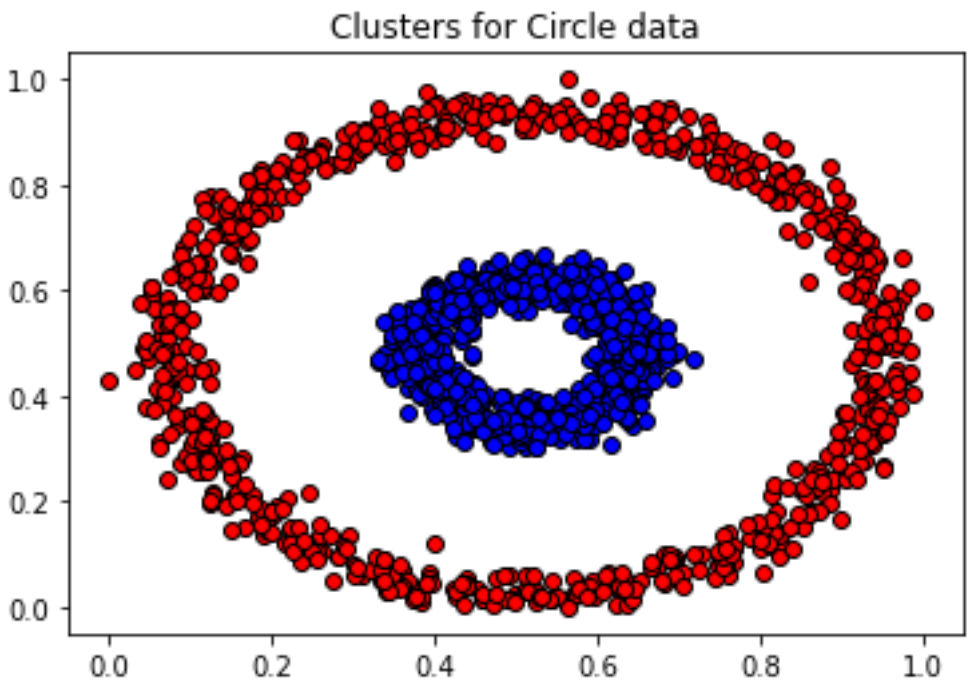
1500 rows × 2 columns

3. **Applying DB Scan Clustering for the 3 datasets.**

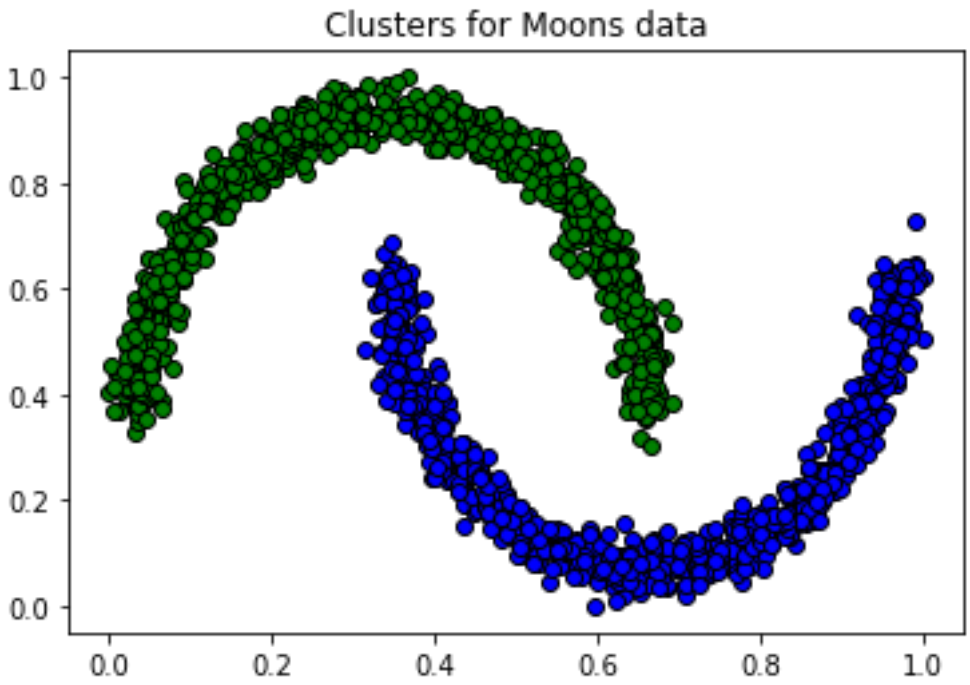
a) For blobs dataset – eps: 0.35, MinPts=10 (3 Clusters)



b) For Circle dataset – eps: 0.112, MinPts=10 (2 Clusters)

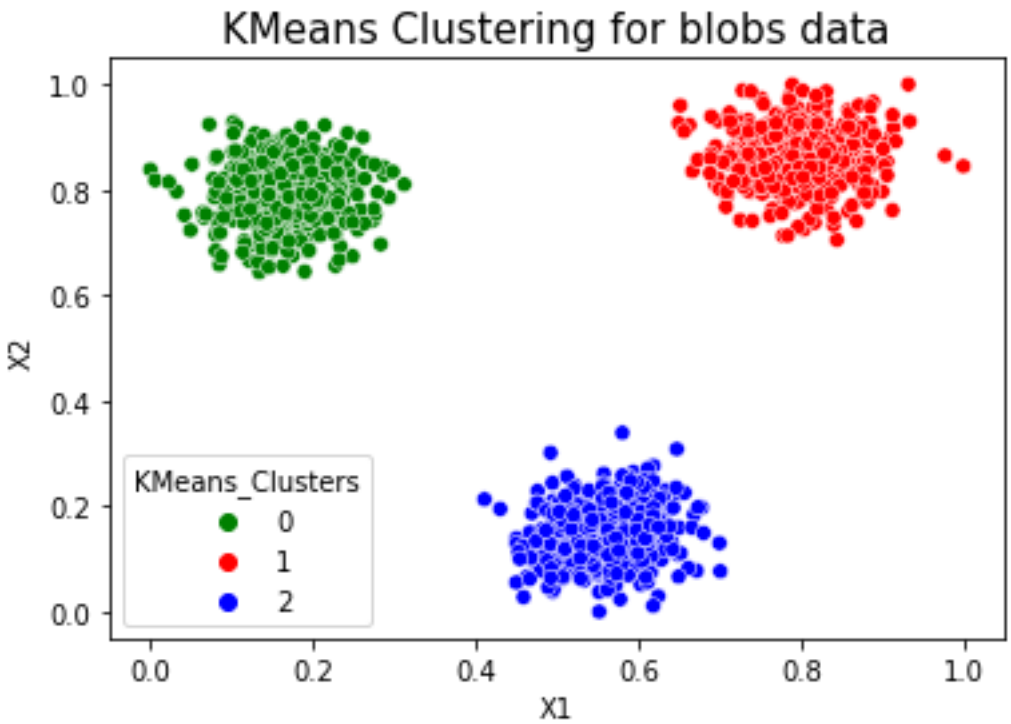


c) For Moons dataset – eps: 0.142, MinPts=10 (2 Clusters)

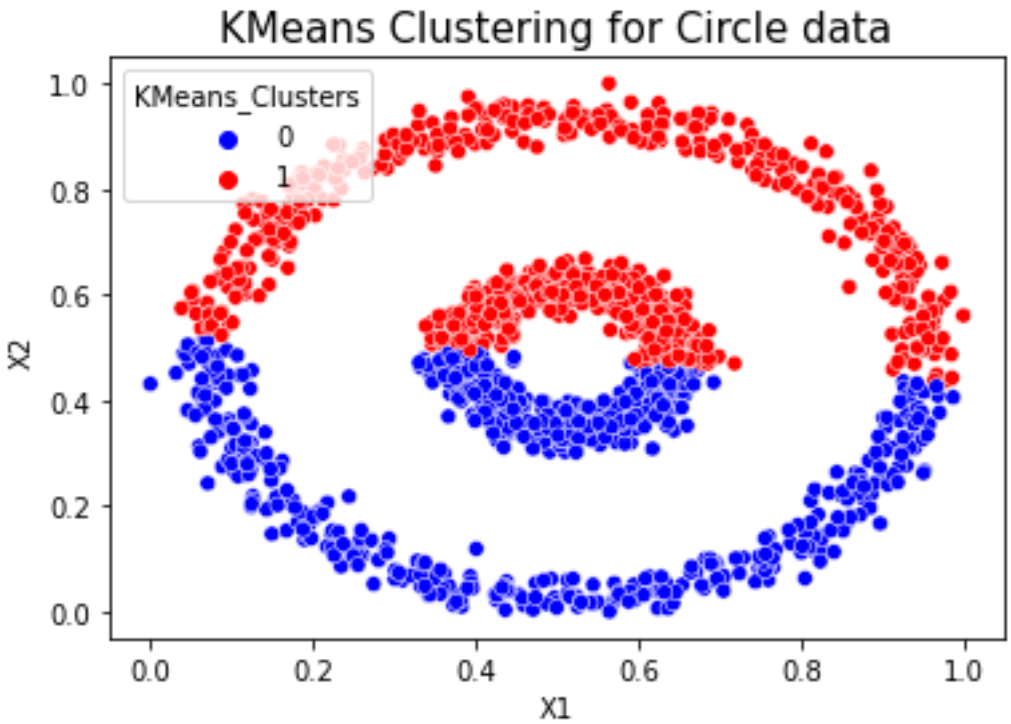


4. Applying K–Means Clustering from the clusters obtained from the DB Scan Clustering.

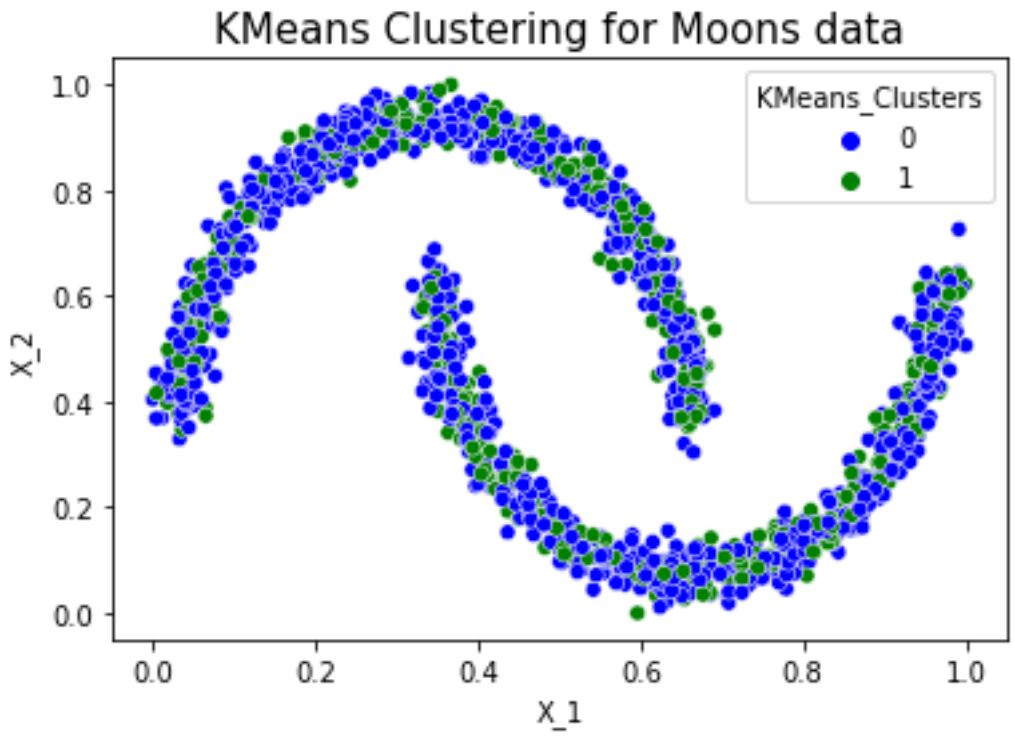
a) For blobs dataset (no of clusters: 3)



b) For Circle dataset (no of clusters: 2)



c) For Moons dataset (no of clusters: 2)



Clustering Algorithm	Silhouette Score
DB Scan	0.86
	0.21
	0.39
K-Means	0.924445
	0.703078
	0.567067

Inference from Results:

- For DB Scan Clustering using the **blob** data, the Silhouette Score is almost comparable to that of the K-means, and the clusters are also well separated. Hence, blob data shows separate and clear clusters.

- For the **circle** data, K-means clustering, the clusters are not well separated though still the Silhouette Score is good for it. But, when we visualise the data, clusters are not well separated. Thus, it is not always necessary to be sure that the good Silhouette Score will give well-separated clusters.

- Same is the case for the **moon** dataset also. Its Silhouette Score is average, but the clusters are not clear and well-separated.