Assignment 4

Machine Learning for Data Science

Kheem Dharmani
MS - DS 22I-0081

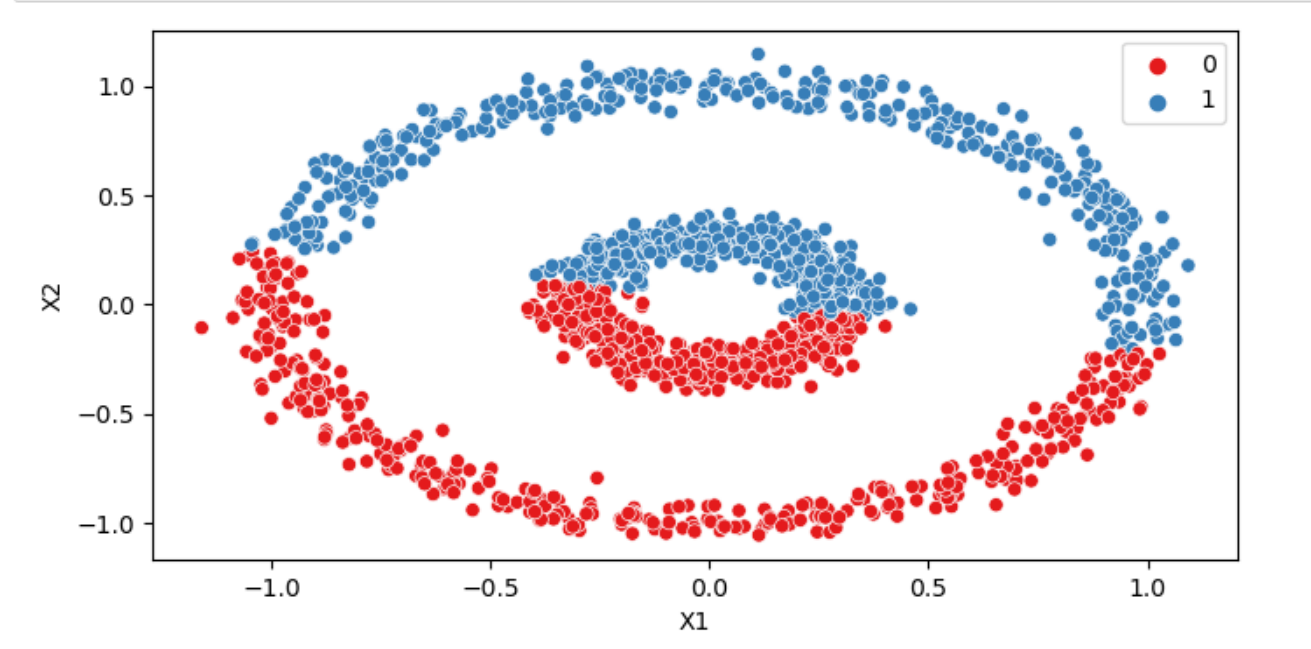
Q#01:

Introduction to DBSCAN:

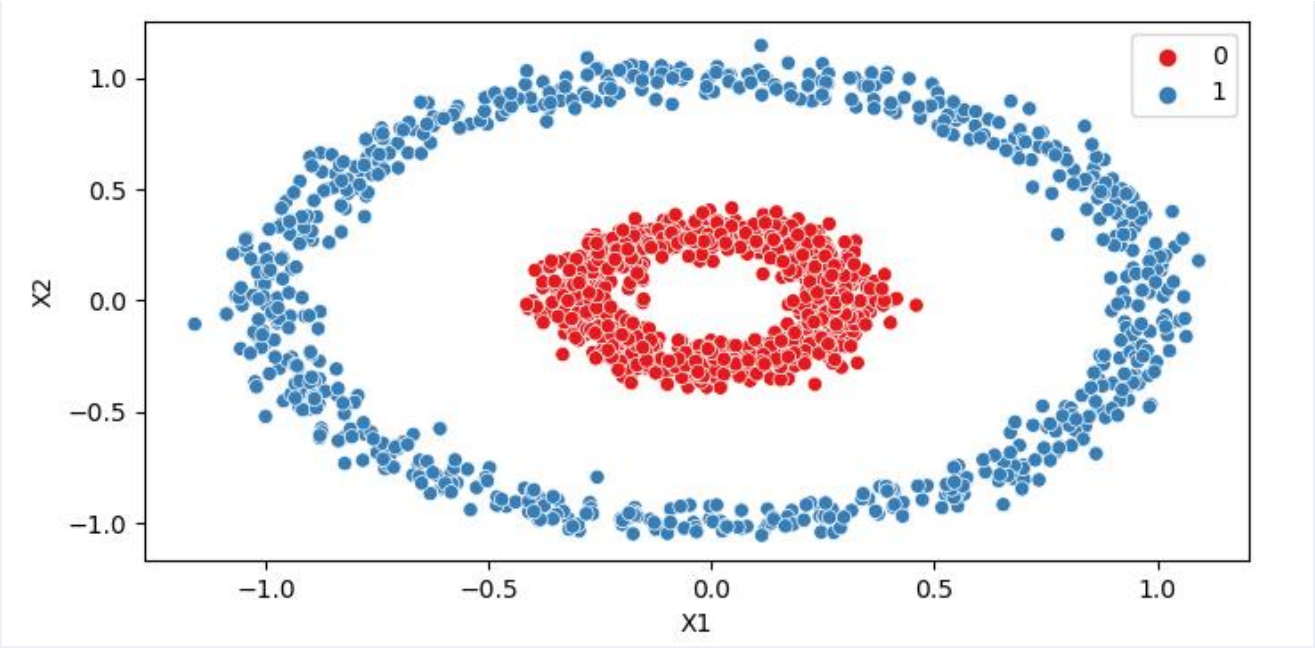
Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996. It is a density-based clustering non-parametric algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away).
DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature. In 2014, the algorithm was awarded the test of time award (an award given to algorithms which have received substantial attention in theory and practice).

Introduction to K-Means:

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.



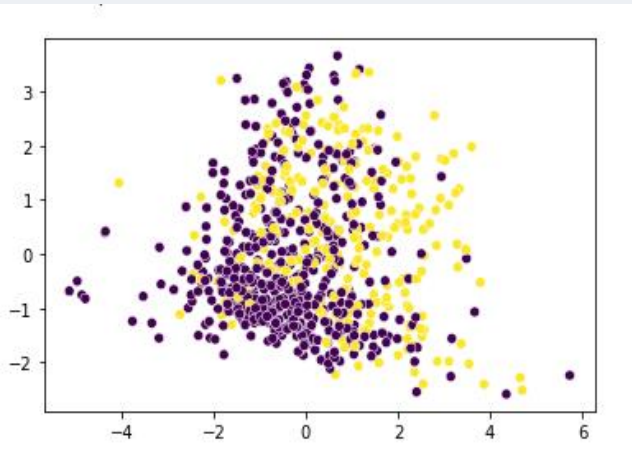
K-Means (2 Clusters)



DBSCAN (2 Clusters)

As we can see the DBSCAN is much more accurate. It is able to capture complex relationships between features. Furthermore, the algorithms were able to spot outliers (Labeled as -1 in the blobs graph).
However, one might think that these points are not outliers, here we should change the default parameters of DBSCAN to take that into account. The parameter that is responsible for that called "epsilon" which decide the range of which n pints is to be considered as neighbors.

Q#02:



There is no such difference in accuracy after applying PCA i.e it remains same 74%