CA Assignment 2
Data Clustering
Implementing clustering algorithms

1. A well-liked unsupervised machine learning approach called K-means clustering is used to organise data points into clusters based on their similarity. It is a straightforward but efficient algorithm that divides the data into K clusters iteratively, where K is a user-specified threshold for the number of clusters to produce.

   The pseudo code works as follows:

   ```
   1. Initialize k centroids randomly
   2. Assign each data point to the nearest centroid
   3. Repeat until convergence:
       a. Compute the mean of each cluster
       b. Update the centroid to be the mean
       c. Assign each data point to the nearest centroid
   4. Return the final set of k centroids and their respective clusters
   ```

   A set of K cluster centroids and the cluster assignments for each data point are the output of the K-means method. The algorithm seeks to reduce the separation between the data points and the centroids that are allocated to them. However, it also has some limitations such as being sensitive to the initial random selection of centroids and requiring the user to specify the number of clusters.

2. K-means++ is a K-means clustering method version that enhances the initial cluster centroids selection, potentially producing superior clustering outcomes. The main goal of the K-means++ algorithm is to choose the initial K cluster centroids in a way that maximises the likelihood that the centroids chosen are representative of the data and are spaced apart from one another.

   When compared to the regular K-means algorithm, the K-means++ approach can achieve faster convergence and better clustering outcomes since it makes sure that the starting centroids are well distributed and representational of the data.

After deciding here on initial centroids, the K-means method continues as usual, assigning data points to the closest centroid and updating the centroids until convergence.

The pseudo code works as follows:

Input: Data points X, number of clusters K

Output: Cluster centroids C

```
1. Choose the first centroid $c_1$ uniformly at random from X
2. for i = 2 to K do:
    a. Calculate the distance between each data point x in X and its nearest centroid $c_{i-1}$, and let $D(x)$ be the minimum distance.
    b. Choose a new centroid $c_i$ from X with probability proportional to $D(x)^2$
3. Initialize the cluster centroids C to $c_1, c_2, ..., c_K$
4. Repeat until convergence:
    a. Assign each data point in X to its nearest centroid in C
    b. Update each centroid in C to be the mean of the assigned data points
5. Return the final cluster centroids C
```

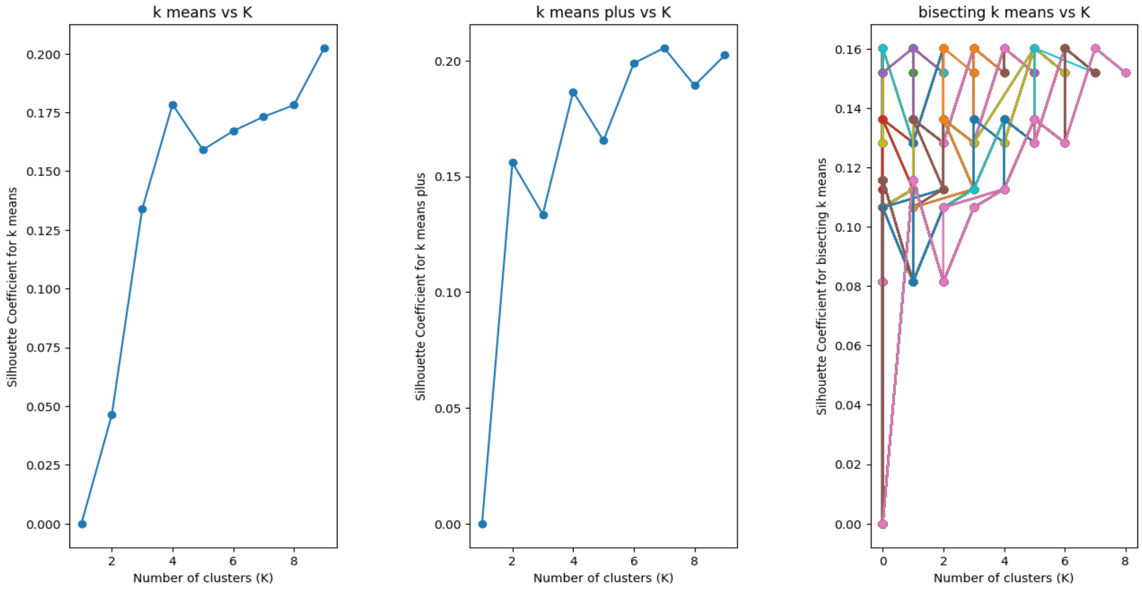
3. A K-means algorithm version called bisecting K-means clustering divides a dataset into K clusters iteratively. Once the largest cluster has been repeatedly divided into two smaller ones until K clusters have been produced, it starts with all the data points belonging to a single cluster.

   The Bisecting K-means algorithm produces a set of K clusters and the assigned clusters for every data point. In order to create well-separated clusters with a high level of intra-cluster similarity, the algorithm splits the largest cluster into smaller ones iteratively.

When the normal K-means algorithm fails to find effective clusters because of local minima or subpar initial centroids, bisecting K-means clustering can be helpful. Compared to the traditional K-means technique, it is likewise more computationally expensive, although it frequently produces better clustering outcomes.

The Pseudo code is as follows:

- Initialize the algorithm by randomly assigning each data point to a cluster.
- Select the cluster with the largest sum of squared distances to its centroid.
- Apply the k-means algorithm to this cluster with k=2 to obtain two subclusters.
- Calculate the sum of squared distances for each subcluster.
- Select the subcluster with the largest sum of squared distances and repeat steps 3-5 until k clusters are obtained.



7. According to the different clustering techniques used in this Assignment. K-means ++ could classify the dataset into different clusters effectively than the other two variations of k-means algorithms. This is because it makes sure that the starting centroids are well distributed and representational of the data. Bisecting K-means has the lowest silhouette score compared to other two ,that means the data is not clustered properly using this algorithm.