**** **Bansilal Ramnath Agarwal Charitable Trust’s**

**Vishwakarma Institute of Information Technology, Pune-48**

**(An Autonomous Institute affiliated to Savitribai Phule Pune University)**

**Department of Computer Science and Engineering (Artificial Intelligence)**

**LAB SUBMISSION**

**Data Science and Machine Learning**

**CAUA22201**

*Submitted by:*

*Avdhoot Nilesh Pimparkar,*

*PRN: 22210178,*

*Roll Number: 281054*

*Second Year*

*Semester II Academic Year 2023-24*

**Assignment: 4**

Aim: To write a program to do following:

We have given a collection of 8 points. P1=[0.1,0.6] P2=[0.15,0.71] P3=[0.08,0.9] P4=[0.16,

0.85] P5=[0.2,0.3] P6=[0.25,0.5] P7=[0.24,0.1] P8=[0.3,0.2]. Perform the k-mean clustering

with initial centroids as m1=P1=Cluster#1=C1 and m2=P8=cluster#2=C2.

Theory:

K-means clustering:

K-means clustering is a popular unsupervised machine learning algorithm used for partitioning data into distinct groups or clusters based on similarities among data points. In the context of the given collection of points, K-means clustering aims to group the points into two clusters, each represented by a centroid. The algorithm starts by randomly initializing the centroids, which serve as the starting points for the clustering process. In this case, the initial centroids are chosen as P1=[0.1,0.6] for Cluster 1 (C1) and P8=[0.3,0.2] for Cluster 2 (C2). Here is the dataset that we have used: “<https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python>”

Next, the algorithm iteratively assigns each data point to the nearest centroid and then updates the centroids based on the mean of the data points assigned to each cluster. This process continues until convergence, where the centroids no longer change significantly or a specified number of iterations is reached.

During each iteration, the distance between each data point and the centroids is calculated using a distance metric, commonly the Euclidean distance. Data points are then assigned to the cluster whose centroid is closest to them. After all data points are assigned, the centroids are recalculated as the mean of all data points assigned to each cluster.

In summary, the K-means clustering algorithm partitions the data into clusters by minimizing the within-cluster sum of squares, where each data point belongs to the cluster with the nearest centroid. This process repeats until convergence, resulting in well-defined clusters that can help in understanding the underlying structure of the data and making predictions or classifications based on those clusters.



Fig. K Means Clustering

Advantages:

1. It is very easy to understand and implement.
2. If we have large number of variables then, K-means would be faster than Hierarchical clustering.
3. On re-computation of centroids, an instance can change the cluster.
4. Tighter clusters are formed with K-means as compared to Hierarchical clustering.

Limitations:

1. It is a bit difficult to predict the number of clusters i.e. the value of k.
2. Output is strongly impacted by initial inputs like number of clusters (value of k).
3. Order of data will have strong impact on the final output.
4. It is very sensitive to rescaling. If we will rescale our data by means of normalization or standardization, then the output will completely change the final output.
5. It is not good in doing clustering job if the clusters have a complicated geometric shape.

Applications:

1. Market segmentation
2. Document Clustering
3. Image segmentation
4. Image compression

Results:

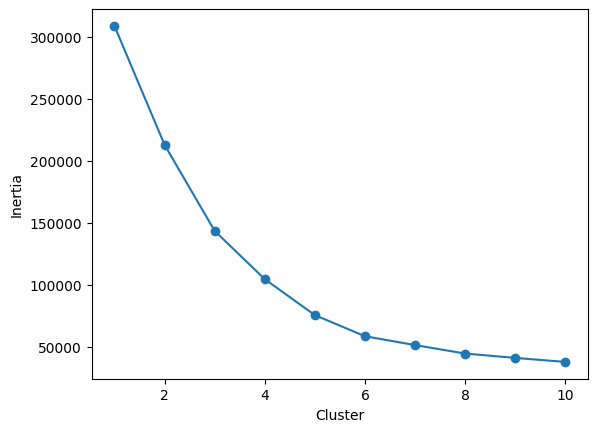
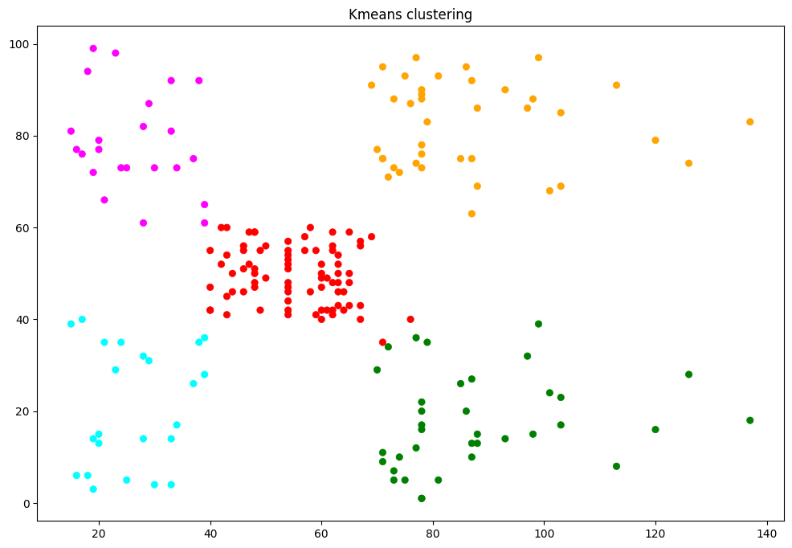
 

Fig. Elbow Curve Fig. Clustering

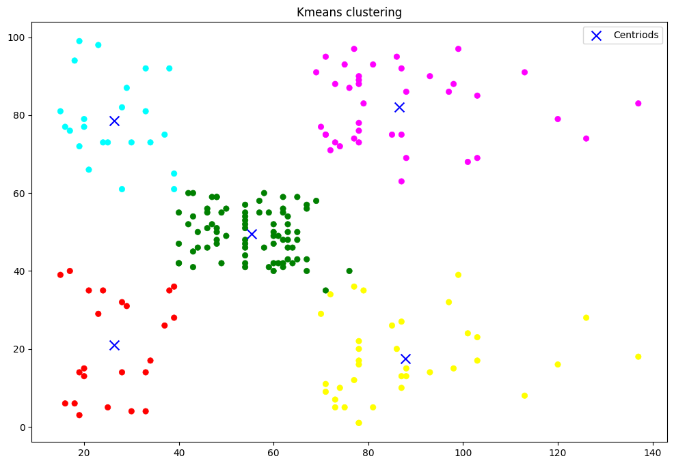


Fig. Clustering with Centroid

Conclusion:

In this assignment, we were able to implement K Means Clustering algorithm to classify the data into clusters.