**Report for ML Project [Unsupervised Learning]**

**Team Members:**

Reshma MR (1MS16IS070)

Rishabh Agarwal (1MS16IS071)

Disha Satheesh (1MS15IS033)

**K-Means Algorithm**

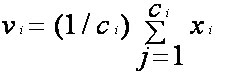
Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select *‘c’* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:



[where,*‘ci’* represents the number of data points in *ith* cluster.]

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

**What is the dataset?**

Thedataset consists of details of various customers purchasing items of their necessity and requirement from an established merchandise outlet which stores the user-data at the time of billing to further analyse and process it to cater the needs of the customers in a more efficient way.

It includes the details such as: Name, Gender, Age, a rough idea of person’s annual capital, and a normalised quantity called Spending Score generated by the system based on the amount of shopping done and the frequentness of the customer.

After initial pre-processing of the dataset, we are using the Age and Spending Score as the key features to disintegrate the customers into various clusters to analyse which age group of people could be the potential customers in future.

**Observation:**

Here, we understand how the centers of each cluster is being calculated. We see how the iterative process of updating the centers goes on till we get the values which would be most suitable for the given data points of the various clusters.

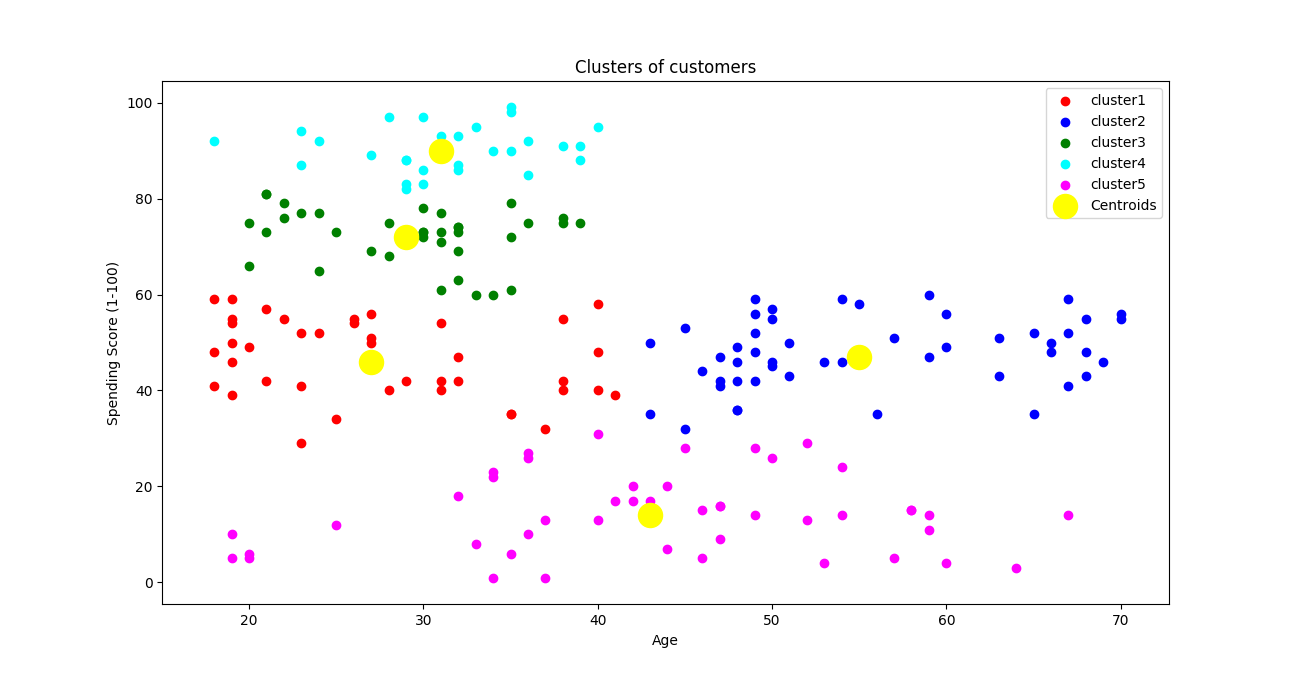
Cluster 1: Clients of age between 15-45 and average spending thus can be labelled as “careful” spenders.

Cluster 2: Clients with age more than 45 and average spending thus can be labelled as “standard” spenders.

Cluster 3: Clients between age group of 20-40 but relatively less spending than cluster 4 thus can be labelled as “potential targets” spenders.

Cluster 4: Clients with similar age of that of cluster 3 but having high spending score thus can be labelled as “careless” spenders of the enterprise.

Cluster 5: Clients with age spread over the entire range and low spending thus can be labelled as “sensible” spenders.



**Result:**

K-means Clustering algorithm along with various Data visualisation techniques is being used highly by various business enterprises for better understanding the demography of their retail clients, influencing the markets and catering to their needs in best possible way to keep their business afloat in this era of competition and prime quality products and service providers.