**Implementation of Clustering Algorithms in C to offer Business Insights.**

Report for the Review1 of Capstone

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No** | **Enrollment**  **Number** | **Admission**  **Number** | **Student Name** | **Degree /**  **Branch** | **Sem** |
| 1 | 18021140099 | 18SCSE1140053 | **SUYASH SHARMA** | B.Tech  CSE CNCS | 5th |
| 2 | 19021180103 | 19SCSE1180046 | **KASHIF AHMAD** | B.Tech  CSE AI/ML | 5th |
| 3 | 18021140089 | 18SCSE1140043 | **VARUNDEEP SINGH** | B.Tech  CSE CNCS | 5th |

Under the Supervision of

**Dr. Manoj Kumar, Assistant Professor**



**School of Computing Science and Engineering**

**Greater Noida, Uttar Pradesh Fall 2020 - 2021**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Particulars** | **Page No** |
| 1 | Abstract | 3 |
| 2 | INTRODUCTION | 5 |
| 3 | LITERATURE REVIEW | 6 |
| 4 | OBJECTIVE | 6 |
| 5 | METHODOLOGY | 7 |
| 6 | TOOLS REQUIRED | 11 |
| 7 | References | 13 |

I

**Minor**

**PROJECT TITLE:** Implementation of Clustering Algorithms in C to offer Business Insights

**ABSTRACT**

The project aims to implement clustering algorithms in C to offer business insights. Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. Today many businesses generate a lot of data. Clustering can help businesses to manage their data better – Customer segmentation, grouping web pages, market segmentation and information retrieval are four examples. For retail businesses, data clustering helps with customer shopping behavior, sales campaigns and customer retention. In the insurance industry, clustering is regularly employed in fraud detection, risk factor identification and customer retention efforts. In banking, clustering is used for customer segmentation, credit scoring and analyzing customer profitability.

**Keywords:** Clustering Algorithms, Business Analytics, Customer Segmentation, Marketing Strategies

**INTRODUCTION**

Clustering involves grouping of similar objects into a set known as cluster. Objects in one cluster are likely to be different when compared to objects grouped under another cluster. Clustering is one of the main tasks in exploratory data mining and is also a technique used in statistical data analysis. While clustering is not one specific algorithm, it is a general task that can be solved by means of several algorithms. Some of the popular clustering methods that are used include hierarchical, partitioning, density-based and model-based.[1]

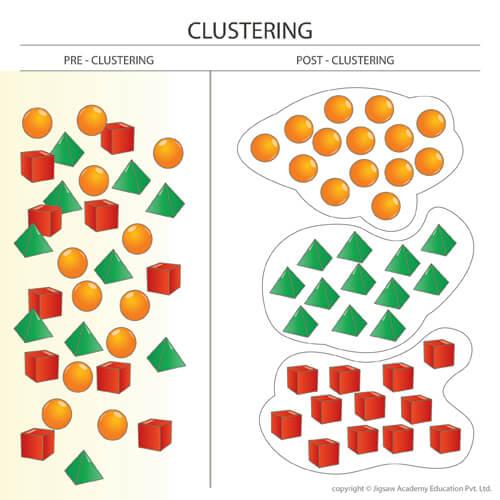


Figure 1: Representation of datasets after and before clustering.

Clustering is an unsupervised machine learning approach, but can it be used to improve the accuracy of supervised machine learning algorithms as well by clustering the data points into similar groups and using these cluster labels as independent variables in the supervised machine learning algorithm

* Different types of methods of Clustering used:

1. Centroid Based Clustering

Algorithm used: K-means Clustering

1. Connectivity Based Clustering

Algorithm used: Hierarchical Clustering

1. Density Based Clustering

Algorithm used: DBSCAN (Density Based Spatial Clustering of Applications with Noise)

* There are two other types of Clustering as well:

1. Subspace clustering

Subspace clustering refers to the task of identifying clusters of similar objects or data records (vectors) where the similarity is defined with respect to a subset of the attributes (i.e., a subspace of the data space).

1. Distribution based methods

It is a clustering model in which we will fit the data on the probability that how it may belong to the same distribution.

Although apart from these, there are there are possibly over 100 published clustering algorithms[2]

Clustering in Business

Personalization, targeting and marketing

This is achieved by looking at specific characteristics of a person and sharing campaigns with them that have been successful with other similar people.

What the problem is: Businesses are trying to get the best return on their marketing investment, it is crucial that they target people in the right way. If they get it wrong, they risk not making any sales, or worse, damaging Customer trust.

In conclusion, Clustering is a powerful technique to explore patterns structures within data and has wide applications is business analytics. There are various methods for clustering. An analyst should be familiar with multiple clustering algorithms and should be able to apply the most relevant technique as per the business needs.[3]

**PROBLEM STATEMENT**

One of the biggest problems with marketing is customer segmentation, businesses develop huge amount of sales data and are clueless as to what to do with it. Businesses need to organize the huge amounts of available data into meaningful structures or break a large heterogeneous population into smaller homogeneous groups.

**LITERATURE REVIEW**

Clustering is a fundamental problem in pattern recognition. It refers to visualization techniques that group data into subsets (clusters), according to a distance measure. Cluster analysis (CA) partitions points of a data set into groups, so that data points within a group are more similar to each other than to points in different groups.[4]

The technique presented subdivides a group into subdivisions of like-minded or like-structured individuals with reference to a given social criterion. This technique consists of three steps: the obtaining of agreement scores of each individual responding to the personal inquiry form with all the subjects; tentative division of the group on the basis of these agreement scores; determination of the pattern that produces the agreement in each subdivision. [5]

The k-means algorithm is well known for its efficiency in clustering large data sets. However, working only on numeric values prohibits it from being used to cluster real world data containing categorical values. [6]

Finding useful patterns in large datasets has attracted considerable interest recently, and one of the most widely studied problems in this area is the identification of clusters, or used by populated regions, in a multi-dimensional dataset. [7] On evaluating the results of cluster analysis, it is common practice to make use of a number of fixed heuristics rather than to compare a data clustering directly against an empirically derived standard, such as a clustering empirically obtained from human informants.[8]

**OBJECTIVES**

Our main objective of this project is to form meaningful clusters from the set of data points given to the algorithms, which would later be reviewed or analyses to offer business insights

* The software should be able to perform on demo data sets given to it as well as new data sets provided at any instance of time.
* Once clusters are segregated, tests can be run on each group with different cases that will help businesses better target their messaging in the future, and reach their target goals.

**METHODOLOGY**

Step1. Acquisition of sales and products data set in desired format.

Step2. Preprocessing of the document is done which may include feature scaling, checking for missing values etc. Entries with missing values could either be dropped or filled with a method like average value of the field or most frequent value, depending on the crucially of the feature.

Step3. Data points will be clustered using three different methods:

1. Centroid Based Clustering

Algorithm used: K-means

ClusteringK points are selected as centroids and other data points are assigning to them based on how far away the point is from each of these points. After each round of Clustering, centroids are moved to the mean of the data points assigned to them, on the basis of these new centroids, data points are reassigned, this process continues until the algorithm converges.

1. Connectivity Based Clustering

Algorithm used: Hierarchical Clustering

It assumed that points that are closer together would be more likely to be related to each other so we choose two points that are closest to each other and cluster them together. Dendrograms are made of the connections and it is required to know when the cut the algorithm off.

1. Density Based Clustering

Algorithm used: DBSCAN (Density Based Spatial Clustering of Applications with Noise)

DBSCAN relates similarity with distance, under this approach two items are clustered together if one is in the neighborhood distance of the other. Algorithm repeats until all core points have been assigned to a cluster.

Step4. These clusters will allow to perform customer segmentation, on the basis of marketing strategies could be developed.

Step5. Testing would be done by making use cases for each module and each of these algorithms.

Step6. Octave is used to show a demo of this algorithm being run on a large data set, also used in initial prototyping and testing.

Flow chart

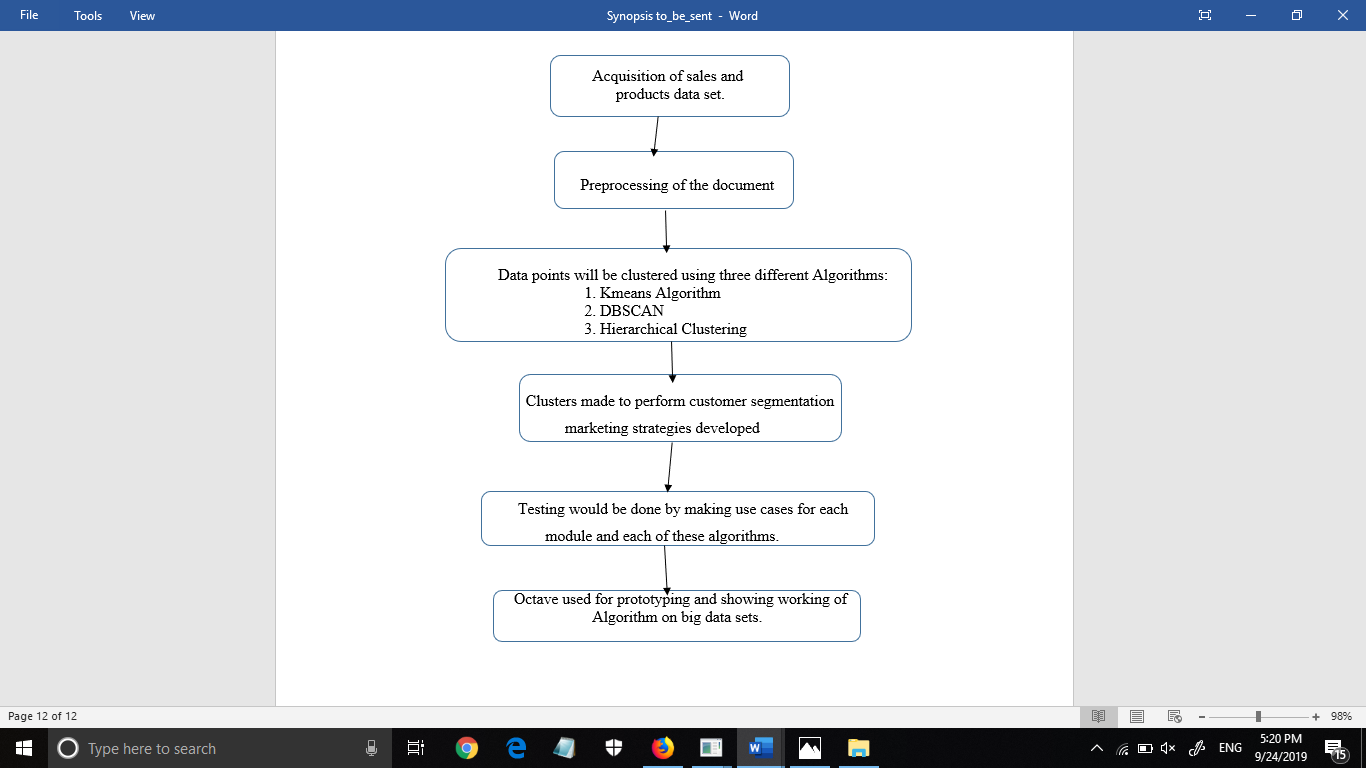


Figure 2. Project Flow Diagram

**ALGORITHMS**

**K means Clustering**

1. Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2} be the set of centres.
2. Select ‘c’ cluster centres or ask from user here c=2.
3. Calculate the distance between each data point and cluster centres.
4. Assign the data point to the cluster centre whose distance from the cluster centre is minimum of all the cluster centres.
5. Recalculate the new cluster centre using mean of data points in the assigned clusters.
6. Recalculate the distance between each data point and new obtained cluster centres.
7. If no data point was reassigned then stop, otherwise repeat from step 3.

**DBSCAN**

1. DBSCAN relates similarity with distance, under this approach two items are similar to each other if the distance between them is small.
2. This Algorithm takes 3 inputs:

* Set of points to be clustered.
* A neighborhood distance that is used to set the cut off distance for similarity in the dataset.
* And a minpts (density) parameter that defines the values at which at a neighborhood of points could be considered dense.

1. For the output the DBSCAN will return the collections of clusters where each cluster is a high-density region in the search space.
2. Each item is represented as a numerical point in some space.
3. Neighborhood of a point is the set of all points that are within a distance set by the neighborhood parameter.
4. Each point is either labelled as a Core, Border or noise point.
5. A core point is any point that has at least minpts in its neighborhood.
6. A border point is a non-core point that has at least one core point in its neighborhood.
7. A noise point is neither a core or border point, they are an outlier in a dataset that do not belong to any particular group.
8. The algorithms picks an unassigned point from the data set and a depth first search starting at that point. Let x be the point being explored. At all the points in x’s neighborhood would be the part of the same cluster.
9. Recursively apply the search onto each unexplored core point in this neighborhood.
10. DBSCAN Algorithm repeats the search until all core points have been assigned to a cluster.

**FLOWCHART**

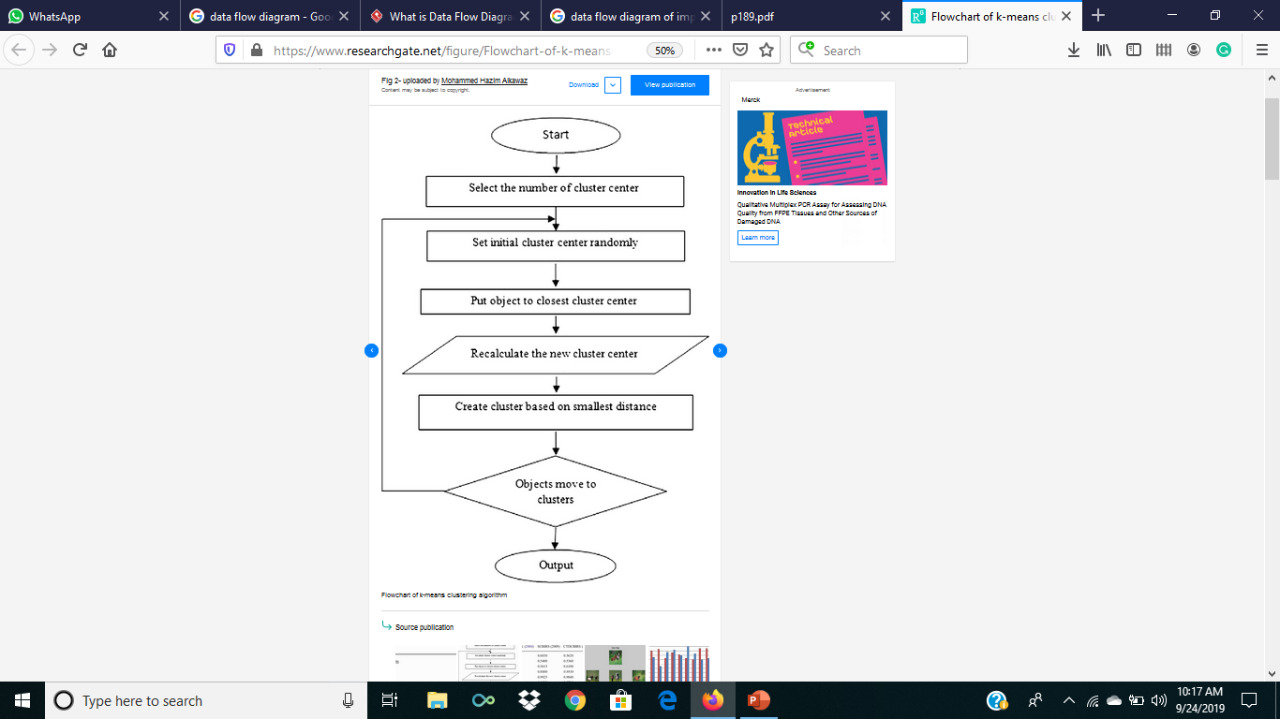


Figure 3. Flowchart of Kmeans Algorithm

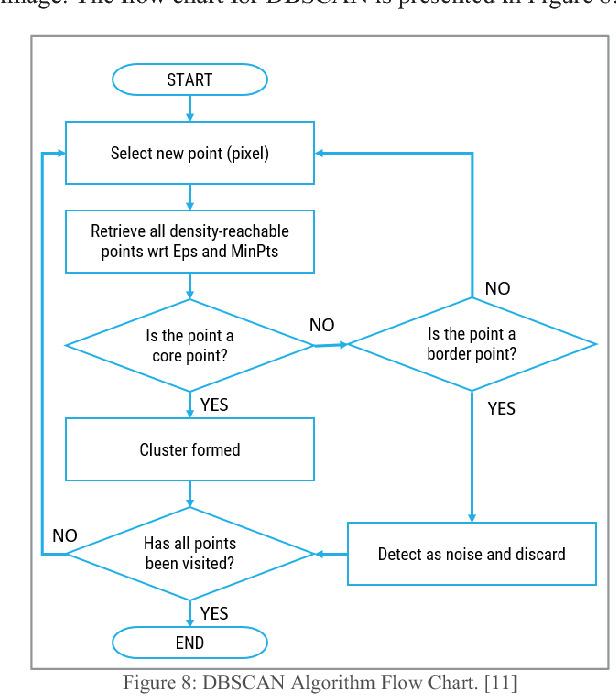


Figure 4. Flowchart of DBSCAN

**SYSTEM REQUIREMENTS**

Software requirements:

* Complier- GCC Complier
* Code Editor-Notepad++
* Testing prototyping Software- Octave used only for prototyping and external test and not implementation of any code that is part of the project

Hardware Requirements:

* RAM: 1GB or above
* Processor: Pentium IV or above

**PERT CHART**

PERT Chart for work division is shown in the chart below along with the work flow, modules, and time assigned to each module.

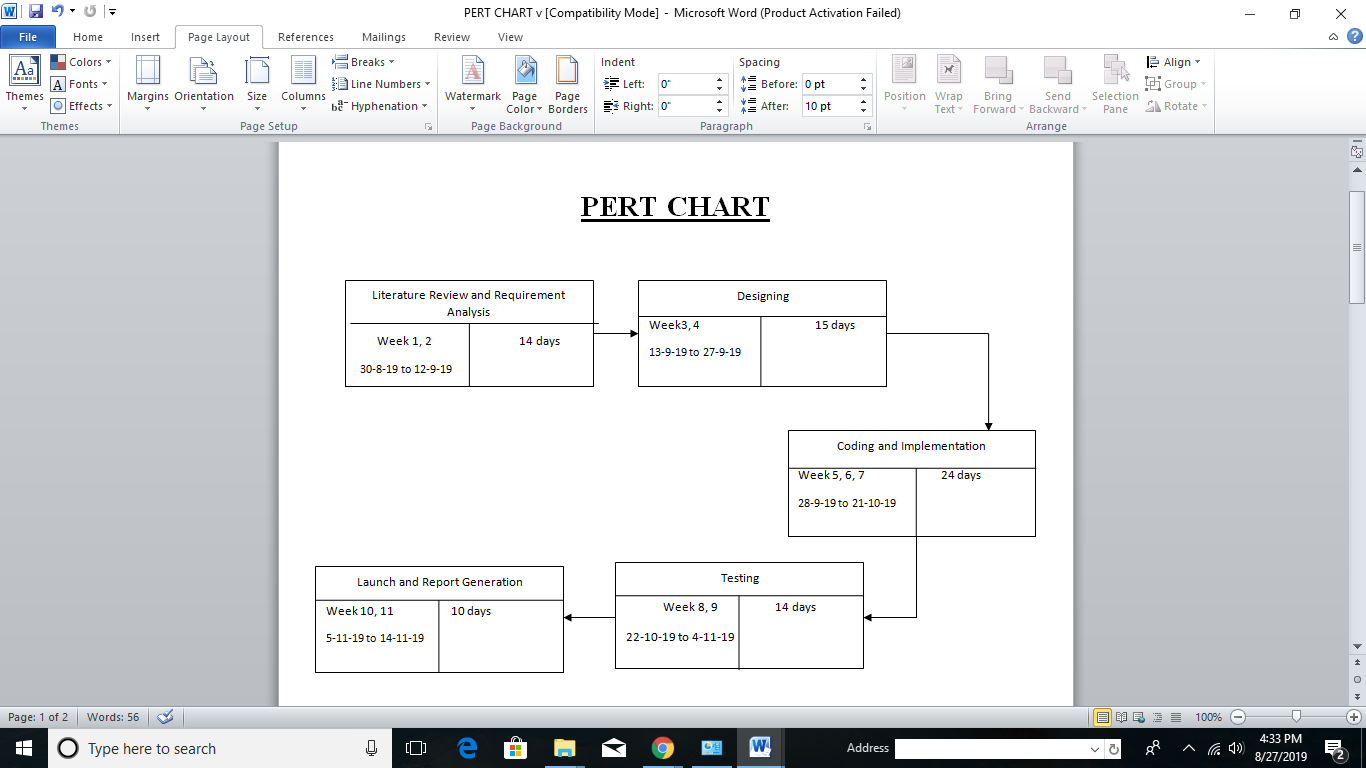


Figure 5. Pert Chart

**REFERENCES**

[1] Wikipedia.org <https://en.wikipedia.org/wiki/Cluster_analysis>

[2] GeeksforGeek <https://www.geeksforgeeks.org/different-types-clustering-algorithm/>

[3] Analyticstraining.com <https://analyticstraining.com/cluster-analysis-for-business/>

[4] Auffarth, B. (July 18–23, 2010). ["Clustering by a Genetic Algorithm with Biased Mutation Operator"](http://www.diva-portal.org/smash/get/diva2:456742/FULLTEXT02). Wcci Cec. IEEE.

[5]  Zubin, Joseph *(1938).* "A technique for measuring like-mindedness". The Journal of Abnormal and Social Psychology*. 33 (4)* [doi](https://en.wikipedia.org/wiki/Digital_object_identifier)*:*[10.1037/h0055441](https://doi.org/10.1037%2Fh0055441)*.*[ISSN](https://en.wikipedia.org/wiki/International_Standard_Serial_Number)[0096-851X](https://www.worldcat.org/issn/0096-851X)*.*

[6] Huang, Z. (1998). "Extensions to the k-means algorithm for clustering large data sets with categorical values". Data Mining and Knowledge Discovery. 2 (3) [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1023/A:1009769707641](https://doi.org/10.1023%2FA%3A1009769707641).

[7] Tian Zhang, Raghu Ramakrishnan, Miron Livny. "An Efficient Data Clustering Method for Very Large Databases." In: Proc. Int'l Conf. on Management of Data, ACM SIGMOD

[8] Pfitzner, Darius; Leibbrandt, Richard; Powers, David (2009). "Characterization and evaluation of similarity measures for pairs of clusterings". Knowledge and Information Systems. Springer. 19 (3):  [doi](https://en.wikipedia.org/wiki/Digital_object_identifier):[10.1007/s10115-008-0150-6](https://doi.org/10.1007%2Fs10115-008-0150-6).

**APPENDIX**

**Kmeans:**

#include<stdio.h>

#include<math.h>

int main()

{

int i, i1, i2, centroid1, centroid2, temp1, temp2, i3, prev\_centroid1, prev\_centroid2;

int data\_set[10];

char product[20];

int cluster1[10], cluster2[10];

printf("\n");

printf("\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Welcome to Customer Segmentation Application \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*");

//User is asked for the name of the product

printf("\nEnter the Name of the product\n");

gets(product);

//User is asked for the number of products bought by 10 customers

for(i=0;i<10;i++)

{

printf("\nEnter number of %s bought by customer %d \n", &product, i+1);

scanf("%d",&data\_set[i]);

}

//Ask user for initial centroids

printf("\n Enter initial centroid1\t");

scanf("%d",&centroid1);

printf("\n Enter initial centroid2\t");

scanf("%d",&centroid2);

do

{

i=i2=i3=0;

//saving previous mean so we can update centroid later

prev\_centroid1= centroid1;

prev\_centroid2= centroid2;

//decision made as to which cluster this data will go to

for(i1=0;i1<10;i1++)

{

//calculating distance to means

temp1=data\_set[i1]-centroid1;

if(temp1<0){temp1=-temp1;}

temp2=data\_set[i1]-centroid2;

if(temp2<0){temp2=-temp2;}

if(temp1<temp2)

{

//near to first mean

cluster1[i2]=data\_set[i1];

i2++;

}

else

{

//near to second mean

cluster2[i3]=data\_set[i1];

i3++;

}

}

//Updating centroids by calculating mean

temp2=0;

for(temp1=0;temp1<i2;temp1++)

{

temp2=temp2+cluster1[temp1];

}

centroid1=temp2/i2;

temp2=0;

for(temp1=0;temp1<i3;temp1++)

{

temp2=temp2+cluster2[temp1];

}

centroid2=temp2/i3;

//Printing clusters

printf("\n");

printf("Type 1 customers in cluster 1\t");

printf("\n");

for(temp1=0;temp1<i2;temp1++)

{

printf("\t");

printf("%d",cluster1[temp1]);

}

printf("\nm1=%d",centroid1);

printf("\n");

printf("\n");

printf("Type 2 customers in cluster 2\t");

printf("\n");

for(temp1=0;temp1<i3;temp1++)

{

printf("\t");

printf("%d",cluster2[temp1]);

}

printf("\nm2=%d",centroid2);

printf("\n");

printf("\n");

}

while(centroid1!=prev\_centroid1&&centroid2!=prev\_centroid2);

return 0;

}