Minor Project-1
Final Report

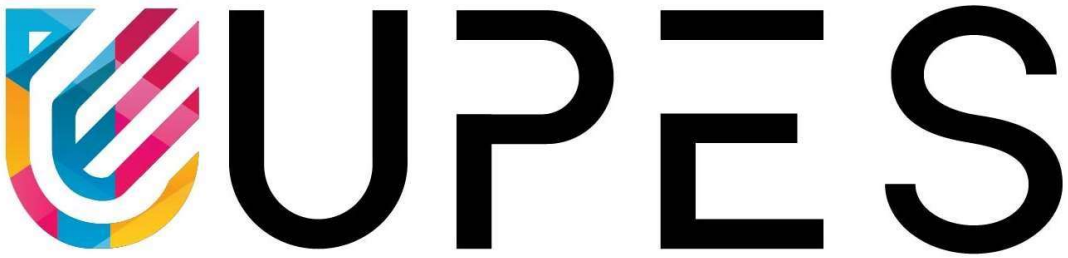
For

A Comparative Study of K-means and G-means Clustering Algorithms on the heart disease dataset

7th December 2022

Prepared by

Specialization	SAP ID	Name
BAO	500083441	Hiya Chopra
BAO	500083281	Shivam Gupta
BAO	500083687	Shrey Gilotra



Department of Informatics
School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES, DEHRADUN- 248007.
Uttarakhand

Table of Contents

Revision History

Date	Change	Reason for Changes	Mentor Signature

WEEK	DATE	DISCUSSION
1	24 Aug 2022	Deciding topic for Minor project -1 with mentor.
2	31 Aug 2022	Finalizing Sub areas for project, Datasets
3	7 Sep 2022	Researched about various Data mining clustering techniques on Heart disease dataset.
4	14 Sep 2022	Worked on Data Preprocessing techniques on our dataset.
5	21 Sep 2022	Shortlisted best algorithms for use
6	28 Sep 2022	Learned about K-means algorithm
7	5 Oct 2022	Implemented K-means on our dataset
8	12 Oct 2022	Testing K-means model over scatter coefficient
9	19 Oct 2022	Validated errors and analysed algorithm issues
10	26 Oct 2022	Started working on G-means model
11	2 Nov 2022	Documenting SRS and report for mid-sem presentation
12	9 Nov 2022	Analysing mistakes focused by mid-sem evaluation
13	15 Nov 2022	Implemented and Tested G-means algorithm
14	23 Nov 2022	Comparitive analysis of both our algorithms
15	28 Nov 2022	Testing the algorithms against various test cases and scenarios
16	2 Dec 2022	Documenting the Report file
17	5 Dec 2022	Adding and editing Final features
18	7 Dec 2022	Final Documentation submission

1. INTRODUCTION:

Data is abundant in every sector; we receive millions of data entries a day. Data mining focuses on finding necessary information hidden in this data. This advancement in technology can be benefited the healthcare sector, as the digital world fits an entire hospital inside a small computer. Cardiovascular diseases are one of the major concerns among people these days.
Clustering a form of unsupervised learning, is a common partitioning algorithm. In layman's terms, it is used to group similar data together separately from un-similar data. We are going to apply this process to our heart disease database to differentiate people into groups, based on their health portfolios. We are clustering the people based on their lifestyle, symptoms, and habits. Some of the criteria used are age, sex, smoking, cholesterol, etc. We are trying to group people as either positive or negative based on their health profiles. This can be done by many algorithms, but we are taking k-means and g-means into consideration. We will be comparing both based on their efficiency.

1.1. PURPOSE:

This project is inclined towards heart disease, one of the major branches in the healthcare sector. According to WHO, 17.9 million people (approx.) die every day due to cardiovascular diseases. Data is abundant, we are trying to find a fast and efficient way to help the healthcare sector process and analyze this data. K-means and G-means are two effective techniques to work with this data and produce results. We will be comparing both techniques through research and experimentation to generate desired results.

1.2. TARGET BENEFICIARIES:

The target beneficiaries are people who are curious to know, whether they are suffering from cardiovascular diseases and who can enter their data through our program.

1.3. PROJECT SCOPE:

This project involves prediction modeling in terms of heart disease with the use of Data Mining. This allows the healthcare sector to perform data mining techniques to track patients' performance and the risks involved. Also, it helps to keep an update on someone who wishes to keep track of his/her cardiovascular activity. Data generated from the healthcare sector can be used to detect many diseases in a pre-mature state which could bring a change and puts a reminder to every individual not to fall under bad circumstances. This project focuses on research-based cardiovascular data to perform clustering techniques on heart disease datasets.

1.4. REFERENCES:

An enhanced k-means and g-Means clustering algorithm for pattern discovery in health care.

[1] Singh, R. and Rajesh, E., 2019. Prediction of heart disease by clustering and classification techniques Prediction of Heart Disease by Clustering and Classification Techniques. International Journal of Computer Sciences and Engineering.

- This research paper includes analysis and extraction of such meaningful heart disease information which is complex enough to be explored with a computing tool. This tells us about data mining technologies benefiting healthcare organizations for grouping patients having similar types of diseases or health issues. With the help of K-means and G-means algorithms, we would be performing clustering.

[2] Haraty, R. A., Dimishkieh, M., & Masud, M. (2015). An enhanced k-means clustering algorithm for pattern discovery in healthcare data. International Journal of distributed sensor networks, 11(6), 615740

- This research paper implies K-means clustering which is a non-hierarchical method that partitions datasets into several different groups. The research paper primarily focuses on how the K-means algorithm is used to perform clustering on the healthcare dataset.

[3] Gomanth D Reddy (2021). Heart disease clustering using K-Mean analysis. International Journal of Advance Research, Ideas, and Innovations in Technology, 7(4) www.IJARIIT.com.

- This research paper is about how the Kmeans algorithm is applying on a heart disease dataset

2. PROJECT DESCRIPTION:

2.1. Data/Data Structure-

This database is multidimensional containing 76 attributes, but to get our desired output efficiently we need only a subset of them, i.e., 14 attributes. Son we have pre-processed this data to get its desired output. The Cleveland database is the only one that has been used by ML researchers to this date. The "goal" field refers to the presence of heart disease in the patient. All data is stored in a .data type file. Only 14 attributes were used:

1. age in years
2. sex (1 = male; 0 = female)
3. cp: chest pain type
   -- Value 1: typical angina
   -- Value 2: atypical angina
   -- Value 3: non-anginal pain -- Value 4: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholesterol in mg/dl
6. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

7. restecg: resting electrocardiographic results

-- Value 0: Normal

-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST Elevation or depression of > 0.05 mV)

-- Value 2: showing probable or definite left ventricular hypertrophy By Estes' criteria

8. thalach: maximum heart rate achieved
9. exang: exercise-induced angina (1 = yes; 0 = no)
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment

-- Value 1: upsloping

-- Value 2: flat

-- Value 3: down sloping

12. ca: number of major vessels (0-3) colored by fluoroscopy
13. thal: 3 = normal; 6 = fixed defect; 7 = reversible defect, thal-> thalassemia
14. num: diagnosis of heart disease (angiographic disease status)

-- Value 0: < 50% diameter narrowing

-- Value 1: > 50% diameter narrowing

In this project, we are using four data sheets, Cleveland (303 instances), Hungarian (294 instances), Switzerland (123 instances), and Long Beach VA (200 instances). Our entire database runs on the same attributes as mentioned above, inclusive of 920 tuples.

2.2. SWOT ANALYSIS-

Strengths-

- There is always scope for finding more information in a dataset, this is where we use data mining algorithms commonly known as knowledge discovery in databases (KDD). In our project, we are grouping similar sets of individuals in a group, signifying the underlying pattern.
- Forecasting Trends involves prediction modeling. We are taking inputs from the user, which can generate output based on our dataset.
- Our database is a global database, so it is not biased to any region in particular.
- Healthcare data are received from various healthcare service providers including sensory environment to provide better healthcare services. This data contains details about patients, medical tests, and treatment.
- Expedited Decision Making

Weaknesses-

- Multidimensionality integrating noisy data. To overcome this weakness, we have pre-processed our data and have used only 14 attributes.

- The accuracy of our model depends on the number of iterations, which can also be infinite.
- Inappropriate User Input leading towards redundancy.

Opportunities-

- Development in health care sectors and artificial intelligence, focusing majorly on cardiovascular diseases and prediction modeling.

Threats-

- Inaccuracy in algorithm
- Incompleteness of data by the user.
- Inconsistency in attributes
- Data not updated timely with the changing trends
- Improbable sources

2.3. Project Features:

Here, we are using 14 attributes out of 76 attributes to run our model using K-means and G-means clustering algorithms. We are running both of them individually to find flaws, strengths, and precision in each. This result will be compared and we can figure out which algorithm can be used to generate accurate results.

One of the most well-renowned methodologies used in clustering is the K-means algorithm.

The algorithm takes 'k' as input, determining the number of clusters to be formed. It gives the local optima of the squared error function. Sometimes choosing the centroids randomly cannot give fruitful results.

To improve upon the existing K-means clustering, an extended version is now being used known as the G-means clustering algorithm. Instead of automatically allocating the number of clusters, G means runs K means with increasing 'k' hierarchically until the test accepts the hypothesis that the data assigned to each k means center is gaussian.

2.4. Design and Implementation Constraints:

The working of the K-Means algorithm is explained in the below steps:

Step 1: Select the number K to decide the number of clusters.

Step 2: Select random K points or centroids. (It can be other than the input dataset).

Step 3: Assign each data point to its closest centroid, which will form the predefined K clusters.

Step 4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third step, which means reassigning each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step 4 else go to FINISH.

Step-7: The model is ready.

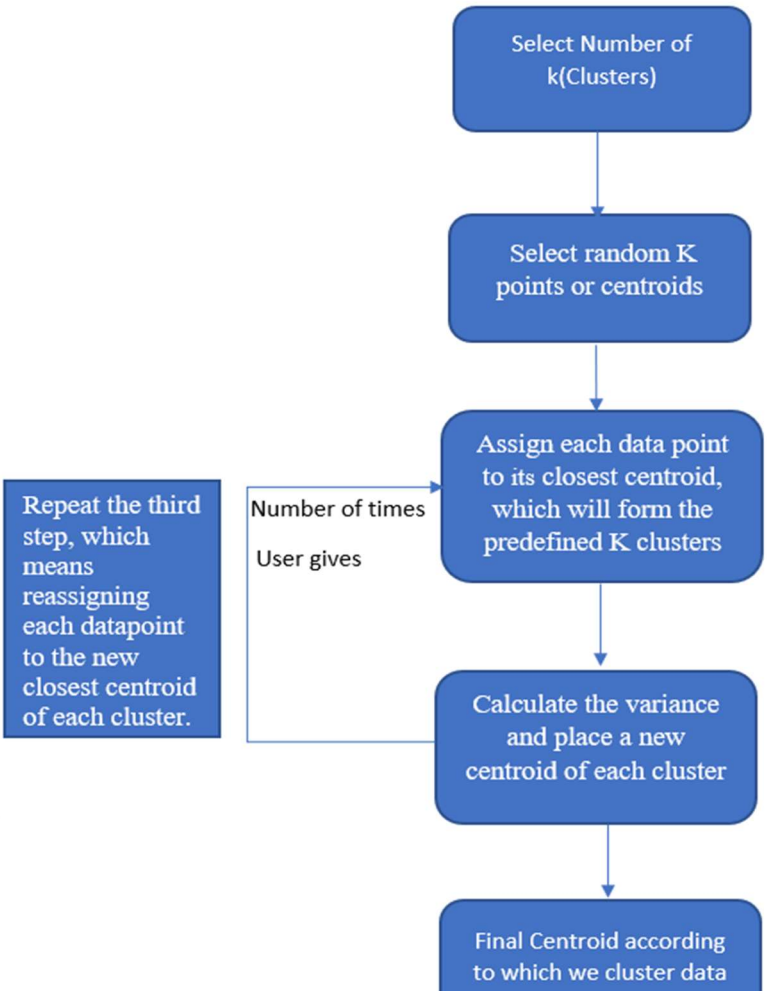
Euclidean distance- It is used for finding the distance between points and centroids

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

p, q = two points in Euclidean n-space

q_i, p_i = Euclidean vectors, starting from the origin of the space (initial point)

n = n-space



S

Figure 1.1 Algorithm Diagram

The essence behind our algorithm, G-means (with the G from the Greedy algorithm and the means from k-means since we use the same distance and similarity functions), is to facilitate the calculation of the initial centroids using a greedy approach; we believe that the k-means' original algorithm needs improvement with the initial random selection of the centroids array. Hence, our initial step is to calculate all of the existing elements that have the highest degree in the space; from there we can have an initial configuration of what the clusters should look like.

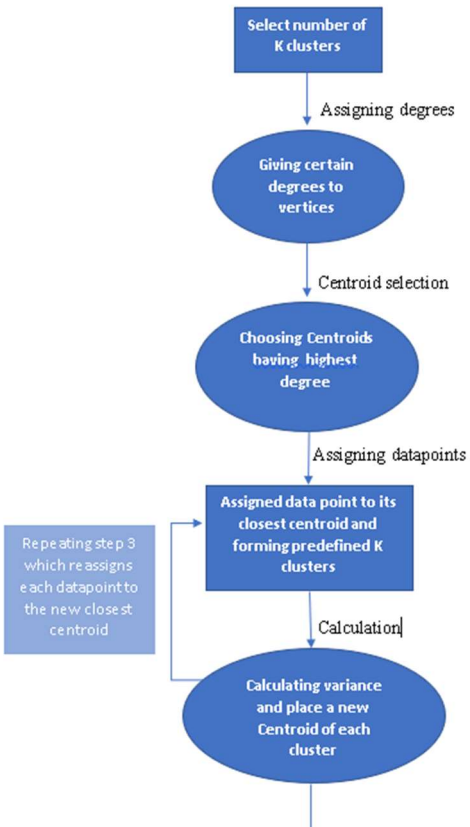


Figure 1.2

STEP 1: The initial step is to pass through the entire dataset and identify the points with the highest degrees (i.e., the points that are the closest to the neighboring points). This constitutes the greedy part of the algorithm

STEP 2: Compare these elements and consider how they are according to k number of clusters. Then, choose the k highest number of vertices

STEP 3**:** Read through the entire database, without going over the elements of the centroid array, and check the similarity and distance functions from each centroid in the clusters to all its elements.

STEP 4: Each element that is read has either one of these options (with either the distance or similarity functions):

A. to be placed in the same cluster as the centroid array, thus dropping it, and the algorithm will never pick it up again;

B. to be kept for later lookup for a potential centroid (at a later run) since the distance function is indicating that it should be in another cluster;

C. to replace the current centroid and, thus, will win both the similarity and distance vector (since we are taking the highest degree element; this is a very rare case where both elements could be centroids, yet the k integer is smaller than the number of clusters).

STEP 5: The algorithm will keep on iterating and keep taking into account the (4(b)) part of the algorithm where these points could end up to be what is called “Boundary Points” of each k cluster.

STEP 6: If the run does not change anything in the centroids array, then we declare that it has successfully converged and displayed the centroids array.

STEP 7: To display each cluster, loop once through the centroids array and match the centroids color with the

elements colors to display the entire group (as shown in figure 4)

2.5. Data flow Diagram-

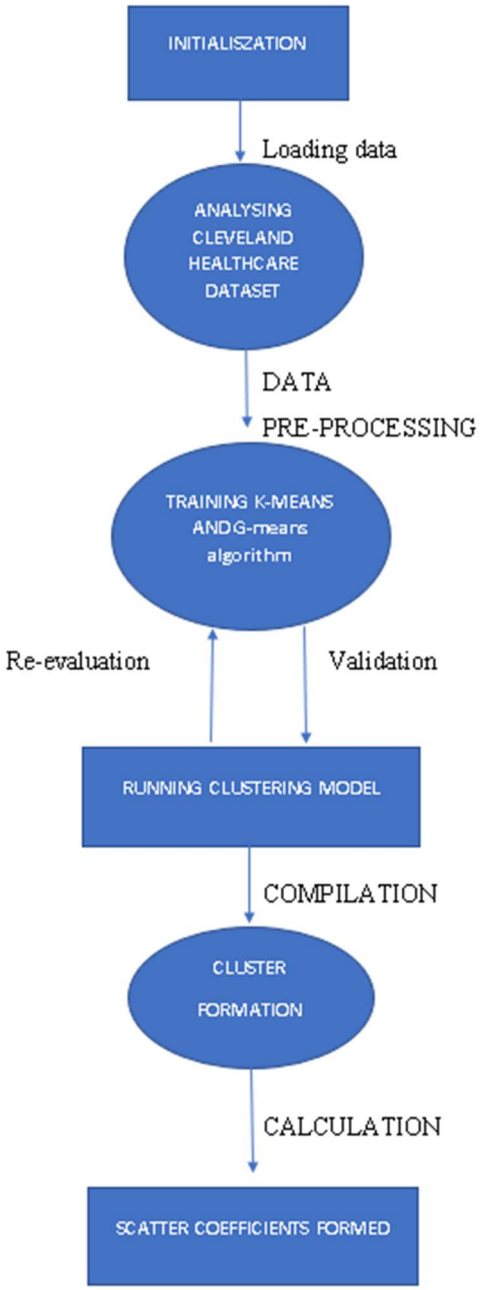


Figure-2 Data Flow Diagram

RESULTS-

For K-means clustering for 5 epochs

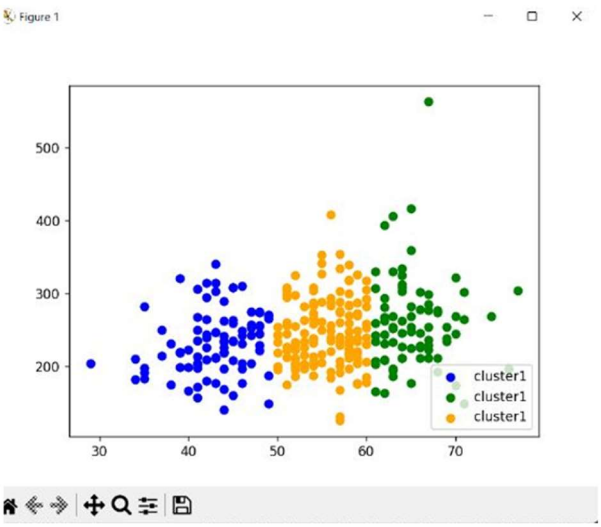


Figure-3.1

For G-means clustering for 5 epochs

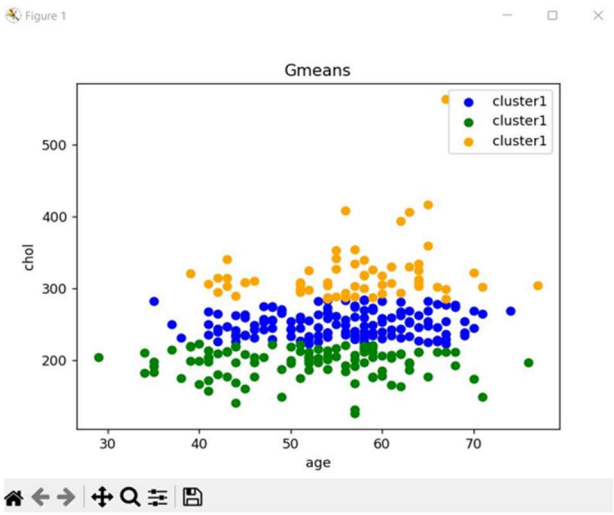


Figure-3.2

For K-means clustering

```
Scatter cofficient- 0.35701703289160264
values      cluster1    cluster2    cluster3
age          56.82        55.48        52.59
chol          329.53          260.39          203.29
```

Figure-3.3

For G-means clustering

```
Scatter cofficient- 0.4514894093095158
values      cluster1    cluster2    cluster3
age          56.18        52.93        63.0
chol          286.6          212.75          425.17
```

figure-3.4

First, we did 5 epochs and we can observe (from fig-3.1 and 3.2) how the clusters formed from these 2 algorithms are different. The cluster from the G-mean algorithm is having high scatter coefficient then K-means

For K-means clustering for 35 epochs

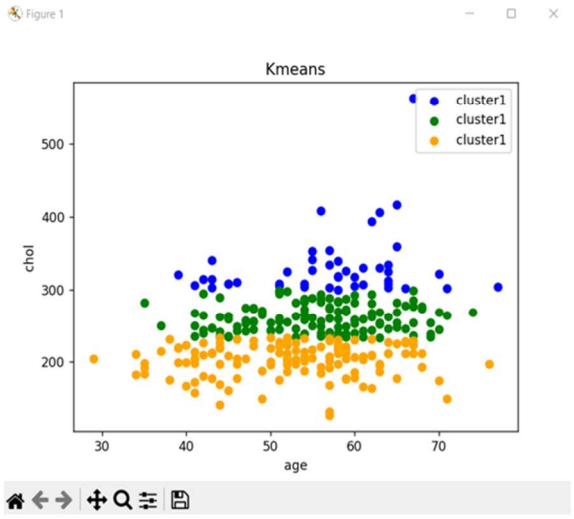


Figure – 3.5

For K-means clustering

For G-means clustering for 35 epochs

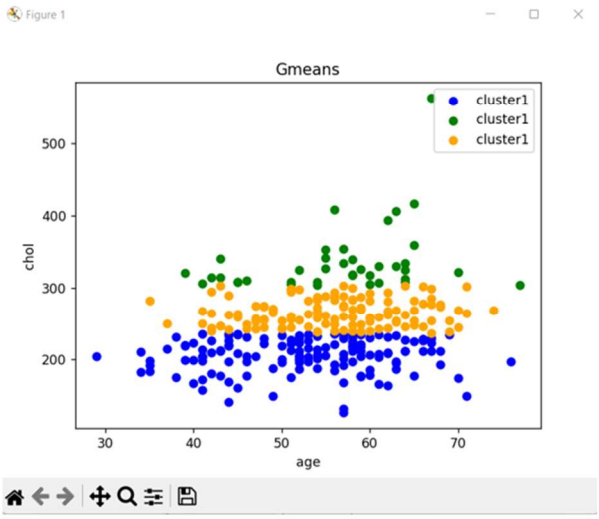


Figure-3.6

For G-means clustering

```
Scatter cofficient- 2.851961606286814
values      cluster1    cluster2    cluster3
age         55.12       56.37       52.02
chol        247.1       316.39      194.01
```

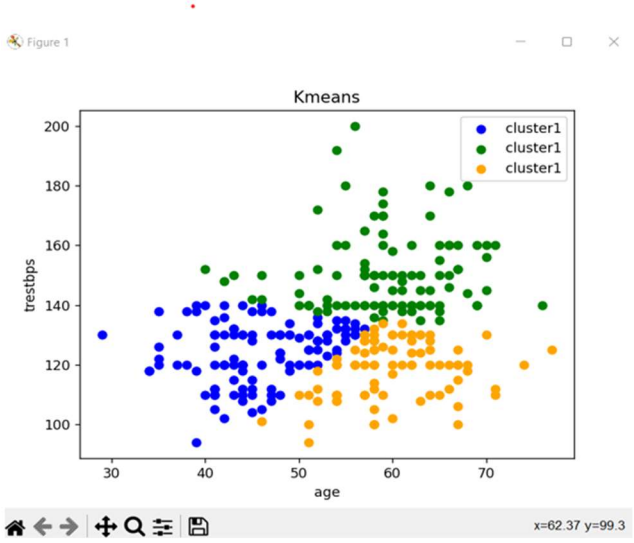
Figure 3.7

```
Scatter cofficient- 3.4514894093095156
values      cluster1    cluster2    cluster3
age         56.57       55.49       52.15
chol        326.26      254.99      198.68
```

Figure 3.8

Now for 35 epochs, we get approx. the same clusters from both the algorithms still G means having a high Scatter Coefficient. (From Figure-3.5,3.6,3.7 and 3.8)

For K-means clustering for 5 epochs



For G-means clustering for 5 epochs

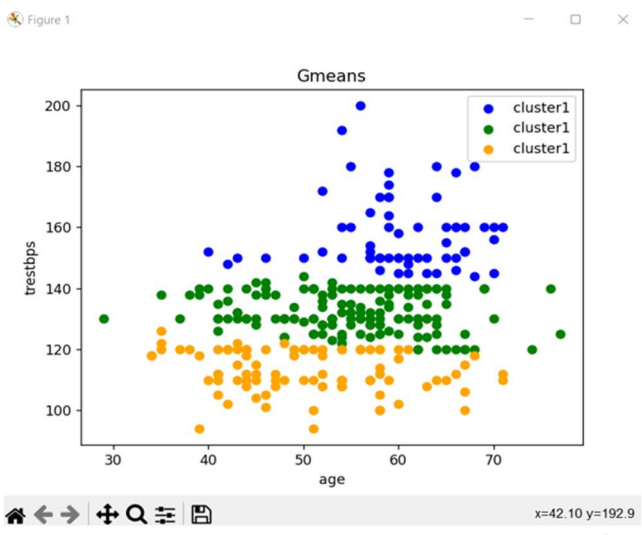


Figure 3.9

Figure 3.10

For K-means clustering

For G-means clustering

```
Scatter cofficient- 1.9262260202983645
values       cluster1    cluster2    cluster3
age          58.54         43.78       58.78
trestbps     150.64        119.63      123.42
```

Figure 3.11

```
Scatter cofficient- 2.1722096508255078
values       cluster1    cluster2    cluster3
age          59.95         54.97       49.69
trestbps     157.95        131.9       113.03
```

Figure 3.11

Here also, we run 5 epochs but on age vs Trestbps and we can observe (from fig-3.9, 3.10, 3.11, and 3.12) how the clusters formed from these 2 algorithms are different. The cluster from the G-mean algorithm is having high scatter coefficient than K-means

For K-means clustering for 35 epochs

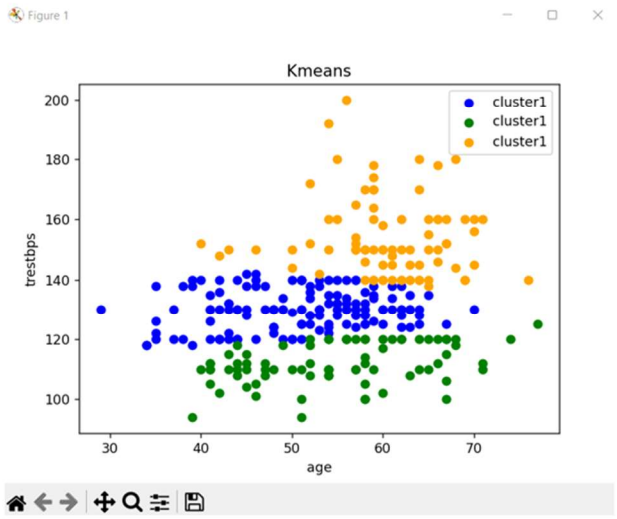


Figure 3.13

For G-means clustering for 35 epochs

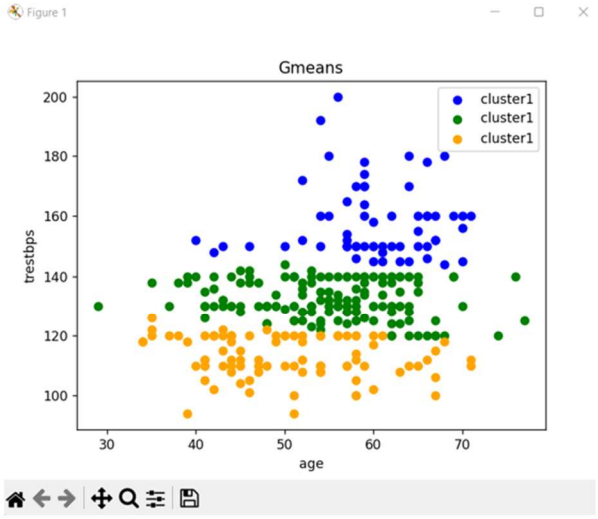


Figure 3.14

For K-means clustering

Figure 3.15

For G-means clustering

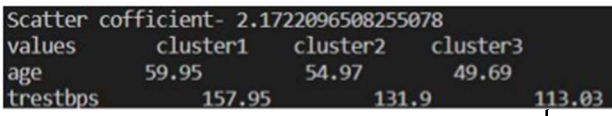


Figure 3.16

Now for 35 epochs, we get approx. the same clusters from both the algorithms still G means having a high Scatter Coefficient. (From Figure-3.13,3.14,3.15 and 3.16)

3. SYSTEM REQUIREMENTS:

3.1. Hardware Requirement-

- Minimum 8 GB RAM
- Minimum 500 GB internal storage
- Minimum Intel core i5

3.2. Software Requirement-

- System with Python installed
- IDE to run the program

4. User Interface

The user interface of our project is a limited-scope design where the user works on a Command Line Interface (CLI) to input the data and perform the following clustering algorithms.

The constraint of this model is that the user has to manually enter the data file and has to remember the command.

4.1. Software Interface

Implementation is in Python programming language that will execute in any system which will require SDE's compilers like Python.exe and an IDE for executing the code.

6. OTHER REQUIREMENTS

No other requirements for the project.

APPENDIX A: GLOSSARY

Educational Data Mining (EDM) is a research field concerned with the application of data mining and is focused on improving learning outcomes by mining and analyzing data collected as we teach

APPENDIX B: ANALYSIS MODEL

The iterative model has helped us derive a systemic approach toward our motto.

APPENDIX C: ISSUES

Not-Applicable.