MINOR-1 PROJECT

SYNOPSIS

For

Implementation and comparative study between K-means and G-means clustering algorithm

Submitted By

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

Department of Informatics

School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007. Uttarakhand

Prof. Deepak Kumar Sharma — **Project Guide**

Dr. Thipendra Pal Singh — **Cluster Head**

School of Computer Science

University of Petroleum & Energy Studies, Dehradun

Synopsis Report

Project Title

Implementation and comparative study between K-means and G-means clustering algorithms

Abstract

Nowadays, healthcare data received from multiple resources and service providers are implementing algorithms to generate meaningful patterns to identify, analyze and cluster various diseases. Data mining techniques are generally one of the most crucial research areas in identifying meaningful information from massive datasets. Heart disease data is a huge heterogeneous health data, as cardiovascular diseases account for 32% of all deaths worldwide. Thus, we'll be applying various data mining techniques which help us or can be used to detect unknown diseases, cause of diseases, identification of the treatment, cures, drug recommendations, or track the health profile of any individual. Data mining techniques such as Classification, clustering, and association rule mining are used by healthcare service providers in recent times. As per our study, clustering is an unsupervised learning technique. In clustering, large datasets are partitioned into different clusters based on the similarity measure.. The K-means clustering algorithm has a principal value estimation on the datasets to create meaningful clusters which makes it extremely costly to utilize. Also, the G means clustering algorithm facilitates the calculation of the initial centroids using a greedy approach. We'll be presenting K means and G-means clustering algorithms on our datasets to improve and understand various techniques used to manipulate data and scratch profitable information to help healthcare services in a more convenient way.

Keywords

Clustering, K-means Clustering, G-means Clustering

1. Introduction

We as individuals rely entirely on data, knowingly or unknowingly. We receive inputs in the form of verbal communication, instructions, feedback, etc., and analyse them to make appropriate decisions. Most of the useful inferences from the data collected are hidden, and can only be accessed once we find a pattern within it. This is where we benefit from data mining algorithms. We have taken the healthcare industry as our primary data set, as it uses many data manipulation models to receive accurate predictions.

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 the 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.

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. A centroid is randomly allocated to each of the clusters. Furthermore, we check the distance of each item from the centroid by calculating the Euclidean distance between them. The centroid keeps shifting to the average of the items in the cluster after every iteration. One of the major disadvantages of K means is that with the different representations of the data, the results achieved are also different. Euclidean distance can unequally weigh the factors. 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 centre is gaussian.

2. Literature Review

1. An enhanced k-means clustering algorithm for pattern discovery in health care [1]

Health care is a vast data set, containing details about patients, medical tests, and treatment. Data mining techniques are one of the very important research areas in identifying meaningful information from these huge datasets. Unsupervised learning techniques such as clustering has an upper hand over supervised techniques like classification. This is because clustering breaks the data set into smaller groups, without having to mention their predefined classes. This means we need no prior information to analyse the data. One of the widely used data mining techniques is k-means which takes the ‘k’ number of inputs, to form the ‘k’ number of clusters. An extended version of k-means is G-means, ‘G’ being from the greedy algorithm and ‘means derived from K-means, which is to facilitate the calculation of the initial centroid using a greedy approach.

2. Prediction of heart disease by clustering and classification techniques [2]

The research paper emphasizes the following sequence of steps that are very important in a data mining process: 1. Data web, 2. Resource discovery, 3. Pre-processing, 4. Generalization/ Pattern recognition, 5. Analysis, 6. Knowledge. Taking into context our heart disease database, prediction can be done by analysing a spectrum of attributes like cholesterol, smoking, age, and many more. Data mining methodologies embrace methods such as neural networks, naive Bayes, clustering mechanisms, classification, big data, etc.

3. Problem Statement

As we know that heart diseases are very prominent globally, and the prediction of cardiovascular diseases based on one’s health profile before time, makes them treatable in an early stage. Since clustering is done through k-means to date, we are figuring out an enhanced approach, ie G-means to implement the same. We’ll be comparing both the clustering methodologies, to find the most optimum result.

4. Objectives

To cluster Patients based on their health profile using clustering.

Sub-objective 1: To apply K-means and G-means clustering on the data set.

Sub-objective 2: Comparative analysis between K means and G means algorithm.

5. Methodology

The K-means algorithm is mentioned down here [4]-

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

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

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

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

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

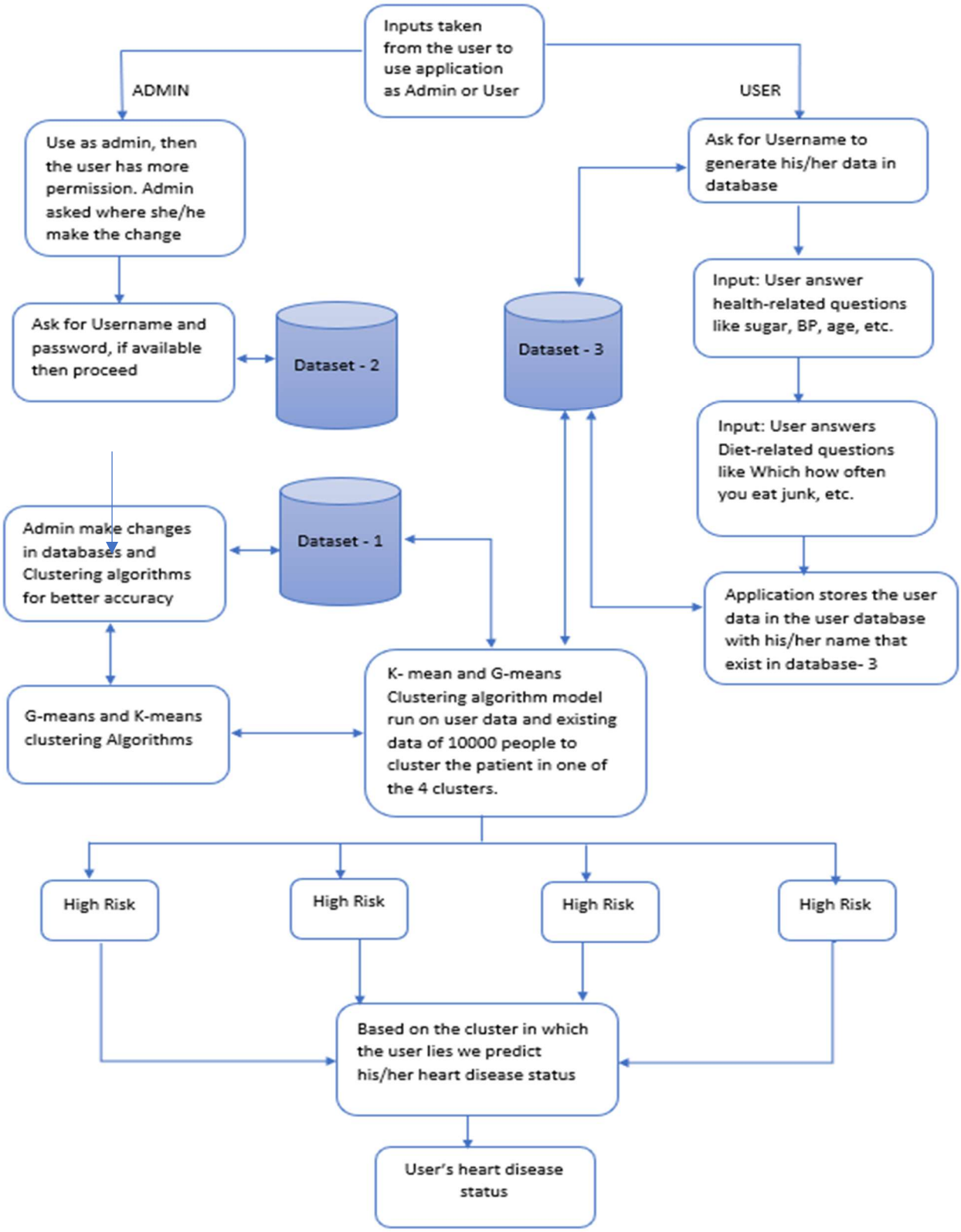
4: Calculate the variance and place a new centroid of each cluster.

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

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

QUOTIENT IN CONSIDERATION: The principal issue with K-means calculations is that it makes one probe over the whole dataset on each cycle, and it needs many such cycles before focalizing on a quality result. This makes it extremely costly to utilize, especially for substantially huge local disk datasets. G-means facilitates the calculation of the initial centroids using a greedy approach. The idea of k-means is to characterize k number of centroids for each group. the algorithm aims to reduce the squared error in this function:

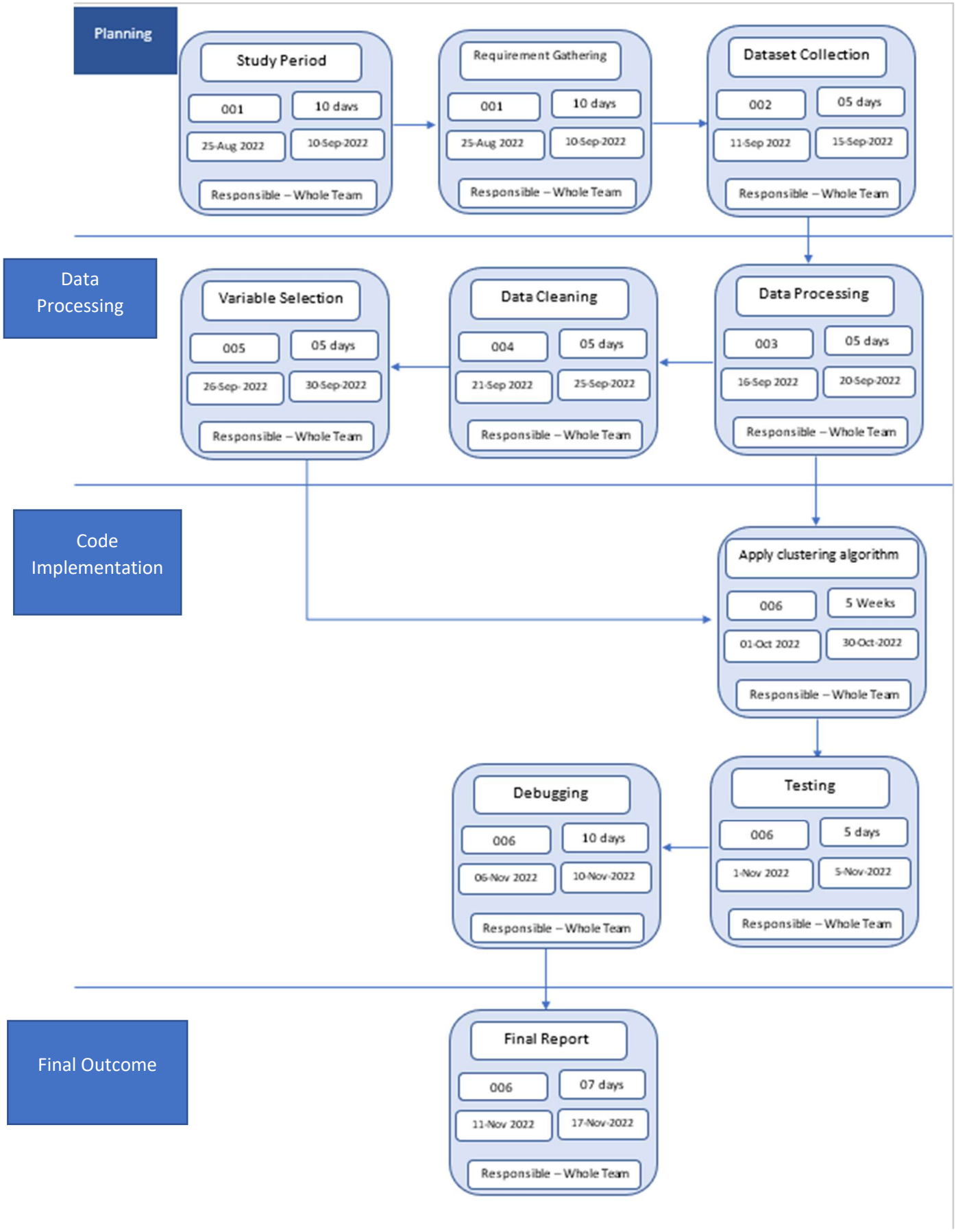
$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 .$$



Dataset 1 - Database comprising 10000 people's health profiles describing BP, Sugar, and other various parameters. Generally applying algorithm used in prediction of user's heart status, also clustering people on the basis of various heart diseases.

Dataset 3 - This is the User data entered by the patient used for the prediction of various related diseases.

PERT Chart


Planning
Study Period
001
10 days
25-Aug 2022
10-Sep-2022
Responsible – Whole Team
Requirement Gathering
001
10 days
25-Aug 2022
10-Sep-2022
Responsible – Whole Team
Dataset Collection
002
05 days
11-Sep 2022
15-Sep-2022
Responsible – Whole Team
Data Processing
Variable Selection
005
05 days
26-Sep- 2022
30-Sep-2022
Responsible – Whole Team
Data Cleaning
004
05 days
21-Sep 2022
25-Sep-2022
Responsible – Whole Team
Data Processing
003
05 days
16-Sep 2022
20-Sep-2022
Responsible – Whole Team
Code Implementation
Apply clustering algorithm
006
5 Weeks
01-Oct 2022
30-Oct-2022
Responsible – Whole Team
Debugging
006
10 days
06-Nov 2022
10-Nov-2022
Responsible – Whole Team
Testing
006
5 days
1-Nov 2022
5-Nov-2022
Responsible – Whole Team
Final Outcome
Final Report
006
07 days
11-Nov 2022
17-Nov-2022
Responsible – Whole Team

References

1. 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.
2. Singh, R., & 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*.
3. https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/ last visited on 15/09/2022
4. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning last visited on 15/09/2022